Harvard University
Harvard University Biostatistics Working Paper Series

Year 2006 *Paper* 43

Using Profile Likelihood for Semiparametric Model Selection with Application to Proportional Hazards Mixed Models

Ronghui Xu* Anthony Gamst† Michael Donohue‡
Florin Vaida** David P. Harrington††

*University of California, San Diego, rxu@math.ucsd.edu
†University of California, San Diego, agamst@ucsd.edu
‡University of California, San Diego, mdonohue@ucsd.edu
**University of California, San Diego, vaida@ucsd.edu
††Dana-Farber Cancer Institute and Harvard School of Public Health, dph@hsph.harvard.edu

Using Profile Likelihood for Semiparametric Model Selection with Application to Proportional Hazards Mixed Models

Ronghui Xu[1,2,*], Anthony Gamst[1], Michael Donohue[1], Florin Vaida[1] and David P. Harrington[3]

[1]Division of Biostatistics and Bioinformatics

Department of Family and Preventive Medicine and [2]Mathematics

University of California, San Diego

[3]Department of Biostatistics and Computational Biology

Dana-Farber Cancer Institute and Harvard School of Public Health

*Correspondence: rxu@ucsd.edu

May 28, 2006

Abstract

Proportional hazards mixed effects model (PHMM) was recently proposed, which incorporates general random effects of arbitrary covariates and includes the univariate frailty model as a special case. In this paper we establish the asymptotic properties of the nonparametric maximum likelihood estimator under PHMM. The asymptotic results allow us to use the profile likelihood for selection of both nested and non-nested PHMMs. We define both a profile likelihood ratio test and a profile Akaike information for general models with nuisance parameters. Asymptotic quadratic expansion of the log profile likelihood allows derivation of the asymptotic null distribution of the likelihood ratio statistic including the boundary cases, as well as unbiased estimation of the Akaike information by an Akaike information criterion. For computation of the likelihood under PHMM we apply three algorithms: Laplace approximation, reciprocal importance sampling and bridge sampling. We compare the three algorithms under different data structures, and apply the methods to a multi-center lung cancer clinical trial.

Key words: Akaike information, asymptotic efficiency, consistency, profile likelihood, likelihood ratio test, testing on the boundary, Laplace approximation, reciprocal importance sampling, bridge sampling.

1 Motivation

In recent years random effects models for failure time data have been applied in various areas, for unobserved heterogeneity, for dependence induced by clustering in, for instance, familial studies, and in settings where some effects, such as center effects in a multi-center trial, are best thought of as sampled from a wider population. The work in this paper, although developed under the more general semiparametric models, has been motivated by the random effects models for failure time data. Like linear and generalized linear models, these random effects models have provided a natural way to model many within-cluster correlations. For example, Vaida and Xu (2000) showed how such models can be used to understand institutional variation in outcomes of a multi-center lung cancer trial conducted by the Eastern Cooperative Oncology Group. The use of random effects survival models in clinical trials was also advocated in Glidden and Vittinghoff (2004), Murray *et al.* (2004) and Sylvester *et al.* (2002). Liu *et al.* (2004ab), on the other hand, used variance components to identify the genetic contribution to the age of onset of alcohol dependence and alcohol abuse. The full power and flexibility of the random effects models, however, has not yet been extended to regression methods for right-censored data.

Vaida and Xu (2000) studied the proportional hazards model with mixed effects (PHMM). It includes the more classical 'frailty' models with random effects on the baseline hazard, but also allows random covariate effects. In this way it is able to model covariate by cluster interactions, such as varying treatment effects in a multi-center clinical trial. The model is of the form

$$\lambda_{ij}(t) = \lambda_0(t) \exp(\boldsymbol{\beta}'\mathbf{Z}_{ij} + \mathbf{b}_i'\mathbf{W}_{ij}), \tag{1}$$

where $\lambda_{ij}(t)$ is the hazard function of the j-th observation from the i-th cluster, $\mathbf{b}_i$ is a vector of random effects for the i-th cluster, and $\mathbf{Z}_{ij}, \mathbf{W}_{ij}$ are the covariate vectors for the fixed and random effects. This model contains a multivariate random effect with arbitrary design matrix in the log relative risk, in a way similar to the linear, generalized linear and nonlinear mixed models. Vaida and Xu developed the nonparametric maximum likelihood estimator (*NPMLE*) of the parameters in this model, computed using the *EM* algorithm and Markov Chain Monte Carlo (*MCMC*) methods. However, the asymptotic properties of the *NPMLE* remain unproven under the PHMM.

As in any regression setting, model selection is an important aspect of data analysis. In particular, in the application of model (1), it often needs to be decided whether a random effect term should be incorporated into the model. From the testing point of view, the null hypothesis is that the corresponding variance component is zero. Although the standard errors of the estimated variance components are obtained in Vaida and Xu (2000), they cannot be used directly for testing zero variance components, because the null hypothesis lies on the boundary of the parameter space. Gray (1995) and Commenges and Andersen (1995) proposed a score test of homogeneity for this purpose. The score test, however, is restricted to the null hypothesis of no random effects. In addition, no tests are readily available for testing more than one parameter at a time, such as for testing the significance of a categorical covariate with more than two categories. In this paper we develop a likelihood ratio test in the general semiparametric setting that, under PHMM, allows arbitrary testing on the mixed model, so a data analyst could test for the significance of a specified subset of the random and/or fixed effects.

Another approach to model selection is via information criteria (Linhart and Zucchini, 1986), which easily handles the comparison of non-nested models, and avoids the boundary problem in the case of selection of random effects. The Akaike information criterion (*AIC*; Akaika, 1973; deLeuw, 1992; Burnham and Anderson, 2002) is among the most commonly used in practice. It has a simple interpretation as penalized log-likelihood, as well as an information-theoretic foundation. Under the Cox model with no random effects, an *AIC* has been used in association with the partial likelihood

(Verweij and van Houwelingen, 1995). However, partial likelihoods do not universally exist for semiparametric models; in particular, strictly speaking it does not apply to PHMM (1). Here we aim to give a meaningful derivation of the *AIC* for general models with nuisance parameters, and in particular to semiparametric models where only the finite dimensional parameters are of interest.

In the next section we prove the consistency and asymptotic normality of the *NPMLE* under PHMM. In Section 3 we study the profile likelihood for general semiparametric models, and use it to derive the profile likelihood ratio test including the boundary case; we also develop an *AIC* using the profile likelihood. In Section 4 we apply the profile likelihood ratio test and the profile *AIC* to PHMM, and consider three algorithms to compute the maximized likelihood under PHMM. Simulation studies are carried out in Section 5 and an example is given in Section 6 to illustrate the methods. Section 7 contains some further discussion. But first, we review the proportional hazards mixed model in some detail below.

1.1 Proportional hazards mixed model

Assume that the data consist of possibly right-censored event time observations from n clusters, with n_i observations in each cluster, $i = 1 \dots n$. Within a cluster the observations are dependent, but conditional on the cluster-specific $d \times 1$ vector of random effects $\mathbf{b}_i$, the survival times T_{ij} are independent and their hazard functions follow PHMM (1). In (1) $\mathbf{W}_{ij}$ is often a subset of $\mathbf{Z}_{ij}$, apart from possibly a '1' which represents the cluster effect on the baseline hazard. To insure identifiability, we assume that $E(\mathbf{b}_i) = \mathbf{0}$. For distribution of the random effects we also assume that

$$\mathbf{b}_i \overset{iid}{\sim} N(\mathbf{0}, \mathbf{\Sigma}) \tag{2}$$

as in Vaida and Xu (2000). Note that the other commonly used frailty distribution, the gamma distribution, is not suitable under the general random effects model (1).

This is because it is not scale-invariant so that the inference is not invariant under a change of measuring unit for covariates of the random effects.

The data from subject j in cluster i can be written $\mathbf{y}_{ij} = (X_{ij}, \delta_{ij}, \mathbf{Z}_{ij}, \mathbf{W}_{ij})$, where X_{ij} is the possibly right-censored failure time and δ_{ij} is the failure-event indicator. Let $\mathbf{y}_i = (\mathbf{y}_{i1}, \ldots, \mathbf{y}_{in_i})$ be the data for cluster i. For cluster i, conditional on the random effect $\mathbf{b}_i$, the log-likelihood is

$$l_i = l_i(\boldsymbol{\beta}, \lambda_0; \mathbf{y}_i | \mathbf{b}_i) = \sum_{j=1}^{n_i} \{\delta_{ij} \log \lambda_0(X_{ij}) + \delta_{ij}(\boldsymbol{\beta}'\mathbf{Z}_{ij} + \mathbf{b}_i'\mathbf{W}_{ij}) - \Lambda_0(X_{ij}) e^{\boldsymbol{\beta}'\mathbf{Z}_{ij} + \mathbf{b}_i'\mathbf{W}_{ij}}\}, \quad (3)$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s)\, ds$. We rewrite the parameter for the baseline hazard in the following as λ, to be consistent with the general semiparametric model framework that we will use. The likelihood of the observed data is then

$$L(\theta) = \prod_{i=1}^{n} \int \exp(l_i) \mathrm{p}(\mathbf{b}_i; \boldsymbol{\Sigma})\, d\mathbf{b}_i, \quad (4)$$

where $\theta = (\boldsymbol{\beta}, \boldsymbol{\Sigma}, \lambda)$ and $\mathrm{p}(\cdot)$ is the multivariate normal distribution. Usually no closed-form expression is available for $L(\theta)$ and its calculation involves d-dimensional integration.

2 Asymptotic theory under PHMM

We assume the following conditions on the data.

C1. Conditional on the covariates $\mathbf{Z}_{ij}$ and $\mathbf{W}_{ij}$, the latent censoring time C_{ij}^* is independent of the failure time T_{ij} and random effects $\mathbf{b}_i$.

C2. There exists some positive constant ϵ such that $\mathrm{P}(C_{ij}^* \geq \tau | \mathbf{Z}_{ij}, \mathbf{W}_{ij}) \geq \epsilon$ almost surely.

C3. $\mathbf{Z}_{ij}$ and $\mathbf{W}_{ij}$ are bounded. In addition, if there exists a constant vector $\mathbf{c}$ and a symmetric matrix $\boldsymbol{\Sigma}$ such that

$$\mathbf{c}'[1, \mathbf{Z}_{ij}']' + \mathbf{W}_{ij}'\boldsymbol{\Sigma}\mathbf{W}_{ij} = 0, \quad j = 1, \ldots, n_i$$

and

$$\mathbf{W}'_{ij}\boldsymbol{\Sigma}\mathbf{W}_{ij'} = 0, \quad j \neq j'; j, j' = 1, \ldots, n_i$$

almost surely, then $\mathbf{c} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{0}$.

C4. The true cumulative hazard $\Lambda_0(t)$ is strictly increasing and continuously differentiable in $[0, \tau]$. Also, $\Lambda_0(\tau) < \infty$.

C5. The true values of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, $\boldsymbol{\beta}_0$ and $\boldsymbol{\Sigma}_0$, belong to the interior of a known compact set,

$$\begin{aligned}\Theta = \{(\boldsymbol{\beta}, \boldsymbol{\Sigma}) : |\boldsymbol{\beta}| \leq \mathcal{B} \text{ for some constant } \mathcal{B},\\ \boldsymbol{\Sigma} \text{ is positive definite and its eigenvalues}\\ \text{are bounded away from 0 and } \infty\}\end{aligned}$$

C6. The cluster sizes n_i are iid bounded random variables and $\mathrm{P}(n_i \geq 2) > 0$ for all i.

Theorem 1 *Under conditions C1–C6,* $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| \to 0$, $\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_0\| \to 0$ *and* $\sup_{t\in[0,\tau]} |\hat{\Lambda}_n(t) - \Lambda_0(t)| \to 0$ *almost surely where* $\|\cdot\|$ *is the Euclidean norm.*

Theorem 2 *Under conditions C1–C6*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}'_n - \boldsymbol{\beta}'_0, \hat{\boldsymbol{\Sigma}}'_n - \boldsymbol{\Sigma}'_0, \hat{\Lambda}_n(\cdot) - \Lambda_0(\cdot))'$$

converges to a zero mean Gaussian process in $\mathbf{R}^{d_1} \times \mathbf{R}^{d_2(d_2+1)/2} \times l^\infty[0,\tau]$ *where* $\hat{\boldsymbol{\Sigma}}_n$ *and* $\boldsymbol{\Sigma}_0$ *are treated as extended column vectors consisting of the upper triangle elements and* $l^\infty[0,\tau]$ *is the space of all bounded functions on* $[0,\tau]$ *with the* sup *norm on* $[0,\tau]$. *Furthermore,* $\hat{\boldsymbol{\beta}}_n$ *and* $\hat{\boldsymbol{\Sigma}}_n$ *are asymptotically efficient.*

Theorem 3 *Let* $V(\mathbf{h}_1, \mathbf{h}_2, h_3)$ *be the asymptotic variance of*

$$\sqrt{n}\{\mathbf{h}'_1(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + \mathbf{h}'_2(\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_0) + \int_0^\tau h_3(t)\, d(\hat{\Lambda}_n(t) - \Lambda_0(t))\};$$

$\mathbf{h}_n$ *be the vector* $\mathbf{h}_1$, $\mathbf{h}_2$, *and* $h_3(X_{ij})$ *for which* $\delta_{ij} = 1$*; and* $\boldsymbol{J}_n$ *be the negative Hessian matrix of* $\log L_n(\hat{\theta})$ *with respect to* $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ *and the jump sizes of* Λ *at* X_{ij} *for which* $\delta_{ij} = 1$. *Then under C1-C6, the variance estimator* $n\mathbf{h}_n' \boldsymbol{J}_n^{-1} \mathbf{h}_n \to V(\mathbf{h}_1, \mathbf{h}_2, h_3)$ *uniformly in probability.*

The proofs of the above theorems are given in the Appendix.

3 Profile likelihood for model selection

In this section we discuss the profile likelihood in the general context of semiparametric models, using the quadratic expansion of Murphy and van der Vaart (2000). Assume that the data consists of a random sample of n observations, $\mathbf{y}_1, \ldots, \mathbf{y}_n$, from a distribution depending on parameters $\boldsymbol{\phi}$ and λ. We assume that $\boldsymbol{\phi} \in \Phi$, a subset of $\mathbf{R}^p$, and λ is a nuisance parameter, possibly of infinite dimension. The log-likelihood of the data is $l(\boldsymbol{\phi}, \lambda) = \sum_{i=1}^n l_i(\boldsymbol{\phi}, \lambda)$, and l_i is the log-likelihood for $\mathbf{y}_i$. The log profile likelihood function for $\boldsymbol{\phi}$, with the nuisance parameter λ 'profiled out', is

$$\mathrm{pl}(\boldsymbol{\phi}) = \sup_{\lambda} l(\boldsymbol{\phi}, \lambda). \tag{5}$$

Following Murphy and van der Vaart (2000), under suitable conditions the log profile likelihood behaves as a quadratic function asymptotically; i.e. for any random sequence $\boldsymbol{\phi}_n$ such that $\|\boldsymbol{\phi}_n - \boldsymbol{\phi}_0\| = O_p(1/\sqrt{n})$ where $\boldsymbol{\phi}_0$ is the true parameter value,

$$\frac{1}{n}\{\mathrm{pl}(\boldsymbol{\phi}_n) - \mathrm{pl}(\boldsymbol{\phi}_0)\} = (\boldsymbol{\phi}_n - \boldsymbol{\phi}_0)'\mathbf{A} - \frac{1}{2}(\boldsymbol{\phi}_n - \boldsymbol{\phi}_0)'\mathbf{I}(\boldsymbol{\phi}_n - \boldsymbol{\phi}_0) + o_p\left(\frac{1}{n}\right), \tag{6}$$

where $\mathbf{A} = \sum_1^n \mathbf{s}(\mathbf{y}_i)/n$, $\mathbf{s}$ is the efficient score for $\boldsymbol{\phi}$, i.e. the ordinary observed score function minus its orthogonal projection onto the closed linear span of the score functions for the nuisance parameter λ, and $\mathbf{I}$, its covariance matrix, is the efficient Fisher information matrix (Murphy and van der Vaart, 2000; Severini and Wong, 1992). We will derive the results of this section for semiparametric models that satisfy (6).

3.1 Profile likelihood ratio test

The likelihood ratio statistic for two nested parametric models, when the parameter space of the smaller model lies entirely in the interior of that of the larger model, has a chi-squared null distribution with the number of degrees of freedom equal to the difference of those of the two models. For a semiparametric model such as (1), the number of degrees of freedom of the model itself is not well defined, since there is at least one infinite dimensional parameter. However, if the infinite dimensional parameter is a nuisance parameter, then under certain conditions the likelihood ratio statistic can be defined via the profile likelihoods, with the number of degrees of freedom calculated using the finite dimensional parameters.

For two nested models let Θ be the parameter space under the larger model, and Θ_0 the parameter space under the smaller model, or equivalently, under the null hypothesis H_0. We assume that H_0 places no additional restrictions on the nuisance parameter λ. Denote L the likelihood, and let

$$LR = \frac{\sup_{\Theta_0} L(\boldsymbol{\phi}, \lambda)}{\sup_{\Theta} L(\boldsymbol{\phi}, \lambda)}. \tag{7}$$

Then LR is the ratio of the maximized likelihoods under the two models. The above can also be viewed as the ratio of the maximized profile likelihoods, with the nuisance parameter λ 'profiled out'. So

$$-2\log LR = -2\{\sup_{\Phi_0} \mathrm{pl}(\boldsymbol{\phi}) - \sup_{\Phi} \mathrm{pl}(\boldsymbol{\phi})\}, \tag{8}$$

where Φ_0 and Φ are the corresponding parameter spaces for $\boldsymbol{\phi}$ under the two models. Murphy and van der Vaart (2000) showed that as result of the quadratic expansion (6), when $\boldsymbol{\phi}_0$ lies in the interior of the parameter space, the profile likelihood ratio test for $H_0 : \boldsymbol{\phi} = \boldsymbol{\phi}_0$ has asymptotically chi-squared null distribution with the number of degrees of freedom equal to the dimension of $\boldsymbol{\phi}$.

Testing on the boundary

As mentioned in Section 1, the challenging problem in hypothesis testing under model (1) is when the null hypothesis lies on the boundary of the parameter space, such as testing against zero variances of the random effects. We show in the following that the asymptotic expansion (6) enables us to obtain results on the null distribution of the profile likelihood ratio statistic similar to those in Self and Liang (1987). First we obtain a result similar to that of Theorem 1 in Self and Liang (1987), on the $\sqrt{n}$-consistency of the maximum (profile) likelihood estimator when $\boldsymbol{\phi}_0$ is on the boundary of Φ, given the $\sqrt{n}$-consistency when $\boldsymbol{\phi}_0$ lies in the interior of Φ.

Theorem 4 *Given the quadratic expansion (6), with probability tending to 1 as $n \to \infty$ there exists a sequence of points in Φ, $\hat{\boldsymbol{\phi}}_n$, at which local maxima of $pl_n(\boldsymbol{\phi})$ occur, that converges to $\boldsymbol{\phi}_0$ in probability. Moreover, $\sqrt{n}(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0) = O_p(1)$.*

See Appendix for proof.

Notice that (6) is equal to

$$\frac{1}{2}\mathbf{A}'\mathbf{I}^{-1}\mathbf{A} - \frac{1}{2}\{\mathbf{z}_n - (\boldsymbol{\phi}_n - \boldsymbol{\phi}_0)\}'\mathbf{I}\{\mathbf{z}_n - (\boldsymbol{\phi}_n - \boldsymbol{\phi}_0)\} + o_p\left(\frac{1}{n}\right), \tag{9}$$

where $\mathbf{z}_n = \mathbf{I}^{-1}\mathbf{A}$. Therefore the same representation of the asymptotic distribution of $-2\log LR$ as that from Chernoff (1954) and Self and Liang (1987) is obtained, which can then be used to calculate the null distribution of the likelihood ratio statistics. Specifically, assume that Φ and Φ_0 are regular enough to be approximated by cones with vertices at $\boldsymbol{\phi}_0$ (for definition see Self and Liang (1987) or Chernoff (1954)), we have

Theorem 5 *Let $\boldsymbol{Z}$ be a random variable with a multivariate Gaussian distribution of mean $\boldsymbol{\phi}$ and covariance matrix $\mathbf{I}^{-1}(\boldsymbol{\phi}_0)$, and let C_Φ and C_{Φ_0} be non-empty cones approximating Φ and Φ_0 at $\boldsymbol{\phi}_0$, respectively. Then the asymptotic distribution of the likelihood ratio statistic, $-2\log LR$, is the same as the distribution of the likelihood ratio test of $\boldsymbol{\phi} \in C_{\Phi_0}$ versus $\boldsymbol{\phi} \in C_\Phi$ based on a single realization of $\boldsymbol{Z}$ when $\boldsymbol{\phi} = \boldsymbol{\phi}_0$.*

3.2 Profile Akaike information

In this subsection we construct the Akaike information and its associated criterion, *AIC*, for models with nuisance parameters. Since the relevant quantity is the profile likelihood, we term the criterion profile *AIC*.

Consider a family of models $\mathcal{M}$ parameterized by $\theta = (\boldsymbol{\phi}, \lambda)$, where $\boldsymbol{\phi} \in \Phi$ is the parameter of interest, and $\lambda \in \boldsymbol{\Lambda}$ is the nuisance parameter, possibly of infinite dimension. The view we take here, similar to Claeskens and Hjort (2003), is that we are interested in selecting the '$\boldsymbol{\phi}$ part' of the modelling, while leaving the parameter space $\boldsymbol{\Lambda}$ the same across all competing models. In this way, for model selection purposes $\mathcal{M}$ is really indexed by $\boldsymbol{\phi}$ alone. Assume that the data vector $\mathbf{y}$, consisting of n independent observations $\mathbf{y}_1, ..., \mathbf{y}_n$, is generated by a distribution with density f. The classical 'distance' from the true distribution f to a member $g_\theta = g(\cdot|\boldsymbol{\phi}, \lambda)$ of $\mathcal{M}$ is given by the Kullback-Leibler information (*KL*), $I(f, g_\theta) = E_f\{\log f(\mathbf{y}) - \log g_\theta(\mathbf{y})\}$. When the focus is on $\boldsymbol{\phi}$ alone, the relevant distance is that between f and the subfamily of models $\{g_{\boldsymbol{\phi},\lambda} : \lambda \in \boldsymbol{\Lambda}\}$: $\min_{\lambda\in\boldsymbol{\Lambda}} I(f, g_{\boldsymbol{\phi},\lambda})$. Suppose that the minimum is attained at some $\lambda = \tilde{\lambda}(\boldsymbol{\phi})$ for each $\boldsymbol{\phi}$. Following Severini and Wong (1992), $\tilde{\lambda}(\boldsymbol{\phi})$ is in fact a least favorable curve under smoothness conditions (see also Fan and Wong, 2000). We denote $g_{\boldsymbol{\phi}} = g(\cdot|\boldsymbol{\phi}, \tilde{\lambda}(\boldsymbol{\phi}))$. Ignoring the constant term $E\{\log f(\mathbf{y})\}$ in $I(f, \cdot)$, we have that

$$E\{\log g_{\boldsymbol{\phi}}(\mathbf{y})\} = \max_{\lambda} E\{\log g_{\boldsymbol{\phi},\lambda}(\mathbf{y})\};$$

the expectations here and in the rest of this section are with respect to the true distribution f. Therefore $g_{\boldsymbol{\phi}}$ is the theoretical equivalent of the profile likelihood.

Minimum *KL* is attained at $\boldsymbol{\phi}_0$ such that $I(f, g_{\boldsymbol{\phi}_0}) = \min_{\boldsymbol{\phi}} I(f, g_{\boldsymbol{\phi}})$, or, equivalently,

$$E\{\log g_{\boldsymbol{\phi}_0}(\mathbf{y})\} = \max_{\boldsymbol{\phi}} E\{\log g_{\boldsymbol{\phi}}(\mathbf{y})\}.$$

Then $g_{\boldsymbol{\phi}_0}$ is the best approximation to f within the family of models $\mathcal{M}$. When the model is correct, i.e., $f \in \mathcal{M}$, we have clearly that $f = g_{\boldsymbol{\phi}_0}$. In practice $\boldsymbol{\phi}_0$ is estimated by $\hat{\boldsymbol{\phi}}(\mathbf{y})$ which maximizes the profile likelihood:

$$\mathrm{pl}(\mathbf{y}|\hat{\boldsymbol{\phi}}) = \max_{\boldsymbol{\phi}} \mathrm{pl}(\mathbf{y}|\boldsymbol{\phi}) = \max_{\boldsymbol{\phi},\lambda} \log g(\mathbf{y}|\boldsymbol{\phi}, \lambda).$$

Note that $(\hat{\boldsymbol{\phi}}, \hat{\lambda})$ is the *MLE* for $(\boldsymbol{\phi}, \lambda)$. The predictive value of $\mathrm{pl}(\cdot|\hat{\boldsymbol{\phi}})$ is given by the expected *KL* for predicting new data $\mathbf{y}^*$, independent of but from the same distribution as $\mathbf{y}$. Ignoring the constant term, we define the profile Akaike Information

$$\mathrm{pAI} = -2E_{f(\mathbf{y})}E_{f(\mathbf{y}^*)}\mathrm{pl}(\mathbf{y}^*|\hat{\boldsymbol{\phi}}(\mathbf{y})). \tag{10}$$

It is important to note that $\mathrm{pl}(\mathbf{y}^*|\hat{\boldsymbol{\phi}}(\mathbf{y}))$ is different from the log-likelihood function computed at the *MLE* $(\hat{\boldsymbol{\phi}}, \hat{\lambda})$, since it allows maximizing the likelihood over λ based on the new data $\mathbf{y}^*$. The following result shows that pAI can be estimated by a corresponding profile *AIC*, where the number in the correction term is p, the dimension of $\boldsymbol{\phi}$.

Theorem 6 *Assume that (6) holds. Assume also that $f \in \mathcal{M}$, i.e. $f = g(.|\theta_0)$, with θ_0 in the interior of the parameter space. Further, assume that $\boldsymbol{y}, \boldsymbol{y}^*$ consist of n i.i.d. vectors, and $\hat{\boldsymbol{\phi}}$ is consistent for $\boldsymbol{\phi}_0$. Then the profile AIC*

$$\mathrm{pAIC} = -2pl(\boldsymbol{y}|\hat{\boldsymbol{\phi}}(\boldsymbol{y})) + 2p \tag{11}$$

is an approximately unbiased estimator of pAI, in the sense that

$$\mathrm{pAI} = E(\mathrm{pAIC}) + E(r),$$

where $r = o_p(1)$ as $n \to \infty$. If in addition r is uniformly integrable, then $E(r) = o(1)$, and pAIC *is asymptotically unbiased for* pAI.

See Appendix for proof.

Note that in proving the above we assume that the family of models under consideration contains the operating model f, so that the parameters lie in the interior of the parameter space. This is generally the case in the theory of *AIC*. Incidentally, for model selection this avoids the boundary problem encountered in likelihood ratio testing for nested models, since the *AIC* is computed assuming that the model in each case holds. We also noted earlier that with new data $\mathbf{y}^*$ the profile likelihood function at $\hat{\boldsymbol{\phi}}(\mathbf{y})$ is not the same as the likelihood function at the *MLE* based on data $\mathbf{y}$. However, when computing the pAIC, the observed profile likelihood in (11) is the same as the maximized likelihood at $\hat{\theta}$. The correction term, $2p$, depends on the definition of the parameter of interest. In particular, if λ has finite dimension q, the classic *AIC* for $\theta = (\boldsymbol{\phi}, \lambda)$ is $-2l(\hat{\theta}) + 2(p+q)$, while the profile *AIC* for $\boldsymbol{\phi}$ is $-2l(\hat{\theta}) + 2p$.

4 Application to PHMM

Under PHMM our parameter of interest is $\boldsymbol{\phi} = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$, whereas the baseline hazard λ is seen as a nuisance parameter. Asymptotic normality of the *MLE* established in Section 2 implies that the likelihood surface is asymptotically quadratic near the true parameter values, which in turn implies that the same holds for the profile likelihood (Murphy and van der Vaart, 2000; Li, 2000) . The asymptotic properties of the *MLE* have also been established for the gamma frailty models (Murphy, 1994, 1995; Parner, 1998), and Maple *et al.* (2002) verified empirically that the contours of the profile likelihood under PHMM are elliptic.

4.1 Profile likelihood ratio test under PHMM

The representation given in Theorem 5 only involves the finite dimensional parameter $\boldsymbol{\phi}$ under the PHMM, so for the cases of null distributions considered by Self and Liang, or by Stram and Lee (1994, 1995) for linear mixed effects model, the results are exactly the same.

In the following we list the cases which are the most likely to be encountered in practice, and correct an error in the existing literature. Denote in the following d as the dimension of $\mathbf{b}$.

<u>*Case 1*</u>: $d = q + 1$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix},$$

where $\boldsymbol{\Sigma}_{11}$ is $q \times q$ and $q \geq 0$. The asymptotic null distribution of $-2 \log LR$ for testing $H_0 : \sigma_{22} = 0$ (and therefore $\sigma_{12} = 0$) against $\boldsymbol{\Sigma}$ positive semidefinite is $(\chi_q^2 + \chi_{q+1}^2)/2$. When $q = 0$, the above distribution is a 50:50 mixture of a point mass at 0 and χ_1^2; note that in this case the maximum likelihood estimator of the variance components has a positive probability of being zero. Our *Case 1* corresponds to cases 1-3 of Stram and Lee (1994).

<u>*Case 2*</u>: Same as in *Case 1*, but the test also includes a r-dimensional subvector of fixed effects, $\boldsymbol{\beta}_2$, i.e., $H_0 : \sigma_{22} = 0, \sigma_{12} = 0, \boldsymbol{\beta}_2 = \mathbf{0}$ against $\boldsymbol{\Sigma}$ positive semidefinite and general $\boldsymbol{\beta}_2$. The asymptotic distribution of $-2 \log LR$ is $(\chi_{q+r} + \chi_{q+r+1})/2$.

<u>*Case 3*</u>: $d = q + k$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where $\boldsymbol{\Sigma}_{11}$ is $q \times q$ and $\boldsymbol{\Sigma}_{22}$ is $k \times k$. The asymptotic null distribution of $-2 \log LR$ for testing $H_0 : \boldsymbol{\Sigma}_{22} = \mathbf{0}$ (and therefore $\boldsymbol{\Sigma}_{12} = \mathbf{0}$) against $\boldsymbol{\Sigma}$ positive semidefinite is a mixture of χ^2 distributions with degrees of freedom $s, s+1, \ldots, s+k$, where $s = kq + k(k-1)/2$.

This corresponds to Case 4 of Stram and Lee (1994). Note, however, that the degrees of freedom for the mixture indicated in their paper was in error. In Stram and Lee (1995) they corrected the maximum degrees of freedom to $s + k$, but not the minimum degrees of freedom. To see why the correct mixture is the one we stated above, reparameterize $\boldsymbol{\Sigma} = \text{diag}(\sigma)\mathbf{R}\text{diag}(\sigma)$, where σ is the vector of standard deviations, i.e. the square roots of the diagonal values of $\mathbf{\Sigma}$, and $\mathbf{R} = (\rho_{ij})$ is the correlation matrix. Testing $\mathbf{\Sigma}_{22} = \mathbf{0}$ and $\mathbf{\Sigma}_{12} = \mathbf{0}$ is equivalent to testing $\sigma_{q+1} = \ldots = \sigma_{q+k} = 0$, and $\rho_{ij} = 0, i > j > k$; that is, k variance parameters tested on the boundary and s unconstrained correlation parameters. The result then follows along the same lines as in Case 7 of Self and Liang (1987). The mixing probabilities, however, are not directly available in general, and simulation methods may be used to estimate the mixing probabilities, or to estimate the null distribution itself. See Self and Liang (1987) and Stram and Lee (1994) for further discussion.

If, in addition, the condition $\boldsymbol{\beta}_1 = \mathbf{0}$ is part of the null hypothesis, then the asymptotic distribution of $-2 \log LR$ is a χ^2 mixture with degrees of freedom $s+r, \ldots, s+r+k$.

Case 4: Another situation of interest is when in the full model $\mathbf{\Sigma}_{12} = \mathbf{0}$ and $\mathbf{\Sigma}_{22}$ is diagonal. Similarly to *Case 3*, the asymptotic null distribution for testing $\mathbf{\Sigma}_{22} = \mathbf{0}$ is a χ^2 mixture with degrees of freedom 0 through k.

Remark The above asymptotic results are obtained under the assumption that the number of clusters, n, goes to infinite. For small n, the approximation by the mixture distributions given above may not be accurate. Crainiceanu and Ruppert (2004) showed that, for balanced linear one-way ANOVA with a single variance component, the mass at zero is larger than 0.5 when n is finite. We further discuss this issue in the simulation section.

4.2 Profile *AIC* under PHMM

The PHMM was our original motivation for developing the profile *AIC*. When the focus is on the fixed effects $\boldsymbol{\beta}$ and the variance components $\mathbf{\Sigma}$, the pAIC is given by (11), where p is the number of parameters in $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$. Computation of the likelihood term in (11) is addressed in the next subsection.

As a special case, when there are only fixed effects in the proportional hazards model, the profile AIC is also given by (11), where p is the dimension of the regression parameter $\boldsymbol{\beta}$. The profile likelihood in this case is the partial likelihood (Cox, 1975; Murphy and van der Vaart, 2000). This AIC has been previously used, for example, by Verweij and van Houwelingen (1995), although no formal justification has been given as an unbiased estimate of a defined Akaike information. Murphy and van der Vaart (2000) verified the conditions for the quadratic expansion (6) in this case. The validity of this AIC as an unbiased estimate of an Akaika information can also be shown directly, using the facts that asymptotically the partial likelihood score has zero expectation, and the second derivative of the log partial likelihood gives the observed information for $\hat{\boldsymbol{\beta}}$ (Andersen and Gill, 1982).

4.3 Computing the likelihood under PHMM

For the PHMM we computed $\hat{\theta}$ using an *EM*-type algorithm, see Vaida and Xu (2000). To compute the likelihood ratio statistic and the pAIC, only the maximum of the full likelihood function given in (4) is needed, since $\mathrm{pl}(\hat{\boldsymbol{\phi}}) = \log L(\hat{\theta})$. The likelihood function (4) is, in general, an intractable integral of dimension d. Here we consider three methods for computing $l(\hat{\theta}) = \log L(\hat{\theta})$: Laplace approximation, reciprocal importance sampling (*RIS*, Gelfand and Day, 1994), and bridge sampling (*BS*, Meng and Wong, 1996). Laplace approximation is computationally simple, but it is less accurate when n_i, the number of observations per cluster, is small. *RIS* and *BS* provide a numerically unbiased estimator for $l(\hat{\theta})$ regardless of n_i, at an additional computational expense. We will compare the performance of the three methods in simulations and data analysis.

In the following we denote $\mathbf{b} = (\mathbf{b}_1', ..., \mathbf{b}_n')'$ and $\mathbf{y} = (\mathbf{y}_1', ..., \mathbf{y}_n')'$.

Laplace approximation. This general method of computing integrals (see, e.g., Tierney and Kadane, 1986) is based on a normal approximation to the posterior distribution of the non-normalized integrand in (4), $p(\mathbf{y}_i)p(\mathbf{b}_i|\mathbf{y}_i)$, and is justified asymptotically, as $n_i \to \infty$. The approxmation for cluster i is given by the formula:

$$l_L^{(i)} = (d/2)\ \log(2\pi) + (1/2)\ \log|\hat{V}_i| + \log p(\mathbf{y}_i|\hat{\mathbf{b}}_i, \hat{\theta}) + \log p(\hat{\mathbf{b}}_i|\hat{\Sigma}), \tag{12}$$

where $\hat{\mathbf{b}}_i = \mathrm{E}(\mathbf{b}_i|\mathbf{y}_i, \hat{\theta})$, $\hat{V}_i = \mathrm{Var}(\mathbf{b}_i|\mathbf{y}_i, \hat{\theta})$ are the posterior mean and variance of the random effects (DiCiccio *et al.*, 1997). We compute $\hat{\mathbf{b}}_i$ and $\hat{V}_i$ using *MCMC* sample averages after convergence of the *EM* algorithm. Alternatively, $\hat{\mathbf{b}}_i, \hat{V}_i$ can be taken as the posterior mode and inverse negative curvature of $p(\mathbf{b}_i|\mathbf{y}_i, \hat{\theta})$, respectively. We compute the Laplace approximation separately for each cluster, and let

$$l_L = \sum_{i=1}^{n} l_L^{(i)} = (nd/2)\ \log(2\pi) + (1/2)\ \log|\hat{V}| + \log p(\mathbf{y}|\hat{\mathbf{b}}, \hat{\theta}) + \log p(\hat{\mathbf{b}}|\hat{\Sigma}), \tag{13}$$

where $\hat{\mathbf{b}} = \mathrm{E}(\mathbf{b}|\mathbf{y}, \hat{\theta})$ and $\hat{V} = \mathrm{Var}(\mathbf{b}|\mathbf{y}, \hat{\theta})$. Note that Ripatti and Palmgren (2000) and Therneau and Grambsch (2000) used Laplace approximation for estimation of θ in PHMM.

Reciprocal importance sampling. Let $p_0(\mathbf{b})$ be the density of a fully specified approximating distribution to $p(\mathbf{b}|\mathbf{y}, \hat{\theta})$, for example, the normal density $p_0(\mathbf{b})$ from $N(\hat{\mathbf{b}}, \hat{V})$. If $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(M)}$ is a *MCMC* sample from $p(\mathbf{b}|\mathbf{y}, \hat{\theta})$, then the reciprocal importance sampling estimator of $l(\hat{\theta})$ is

$$l_R = l_L - \log A, \tag{14}$$

where

$$A = \frac{1}{M} \sum_{k=1}^{M} \exp\{v(\mathbf{b}^{(k)})\} \tag{15}$$

and

$$v(\mathbf{b}) = l_L + \log p_0(\mathbf{b}) - \log p(\mathbf{y}, \mathbf{b}|\hat{\theta}). \tag{16}$$

For numerical accuracy, the computations are done on the logarithmic scale as in (16). Theoretically, l_L can be omitted in (16), in which case $l_R = -\log A$. However, using the Laplace approximation l_L as a "point of reference" in (16) greatly improves the numerical accuracy of l_R. A simple probabilistic argument shows that indeed A in (15) is a Monte Carlo unbiased estimator of $\exp\{l_L - l(\hat{\theta})\}$; see Gelfand and Day (1994) for details.

The sampling and computation for l_R are straightforward to implement. The following result shows that in practice it is more efficient to compute l_R separately for each cluster.

Proposition 1 *Assume that l_R is computed as in (14) over the whole dataset, and $\tilde{l}_R$ is the same except computed cluster-by-cluster. More precisely, $\tilde{l}_R = \sum_{i=1}^n l_R^{(i)}$, where $l_R^{(i)} = l_L^{(i)} - \log A_i$, $l_L^{(i)}$ is given by (12), and $A_i = \sum_k \exp\{v(\mathbf{b}_i^{(k)})\}/M$. Put $\tilde{A} = \prod_{i=1}^n A_i$, so that $\tilde{l}_R = l_L - \log \tilde{A}$. Then both $\tilde{l}_R$ and l_R converge to $l(\hat{\theta})$ with probability one, and the sampling variance of A is at least as large as the sampling variance of $\tilde{A}$.*

See Appendix for proof.

Bridge sampling. Assume that the Monte Carlo samples $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(M)}$ from $p(\mathbf{b}|\mathbf{y}, \hat{\theta})$ and $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(M_0)}$ from $p_0(\mathbf{b})$ are both available, where $p_0(\mathbf{b})$ is a fully specified approximation to $p(\mathbf{b}|\mathbf{y}, \hat{\theta})$, as described for *RIS* above. The bridge sampling (Meng and Wong, 1996) estimator for $l(\hat{\theta})$ is given by

$$l_B = \log(B) - \log(C) + l_L, \tag{17}$$

where

$$B = \frac{1}{M_0} \sum_{k=1}^{M_0} [1 + \exp\{v(\mathbf{u}^{(k)})\}]^{-1} \tag{18}$$

$$C = \frac{1}{M} \sum_{k=1}^{M} [1 + \exp\{-v(\mathbf{b}^{(k)})\}]^{-1}. \tag{19}$$

It is again more efficient to compute l_B separately for each cluster and then combine the results, as in Proposition 1.

The three methods will be compared using simulation experiments in the next section.

5 Simulation experiments

In this section we carry out simulations to compare the accuracy of the three methods described above for calculating the likelihood values, and to study the finite sample distribution of the likelihood ratio statistic.

We simulate data under model (1) with a single binary covariate Z, $\beta = 1.5$, $\lambda_0(t) = 1$, and no random effects. The censoring distribution is Uniform $(0, \tau)$, where τ is chosen to achieve about 15% censoring. We then fit model (1) with a random intercept, i.e. $\lambda_{ij}(t) = \lambda_0(t)\exp(\beta Z_{ij} + b_i)$. Different combinations of numbers of clusters and cluster sizes $(n \times n_i)$ are used. In Figure 1 the likelihood ratios are computed using the three methods described in the last section. We see that reciprocal importance sampling (*RIS*) and bridge sampling (*BS*) have extremely close agreement in computing the likelihood (ratio) for all cases. For the number of observations per cluster $n_i = 20$ Laplace approximation also gives similar results to *RIS* and *BS*. For $n_i = 2$, however, there are discrepancies between Laplace approximation and *RIS* or *BS*. The discrepancies increase with the number of clusters n since the log likelihood is the sum of that from each cluster, and the overall discrepancies are the sums of the discrepancies from each cluster.

In Figure 1 the ordered likelihood ratio statisitcs from 100 simulations are plotted against the theoretical mixture distribution quantiles. The asymptotic results for the null distribution of the likelihood ratio statistic requires that the number of clusters $n \to \infty$. For $n = 100$ (lower panels) we compare the distribution of the likelihood ratio statistic with its asymptotic distribution given in *Case 1* of Section 4.1, i.e. a 50:50 mixture of point mass at zero and χ^2_1. In Figure 1 'p0' denotes the probability of point mass at zero. For $n = 10$ (upper panels) the asymptotic distribution does not appear to provide good approximation, and we use the result of Crainiceanu and Ruppert (2004) on linear mixed models (balanced one-way ANOVA) as a guideline, i.e. a 65:35 mixture of point mass at zero and χ_1^2. Note that their result requires the cluster size $n_i \to \infty$ while keeping the number of clusters n fixed.

There is a clear effect of the number of observations per cluster on the null distribution of the likelihood ratio. For $n_i = 20$ the empirical distributions of the computed likelihood ratio statistics agree reasonably well with their theoretical distributions according to the plots, for both $n = 100$ and $n = 10$. But for $n_i = 2$ even the distributions of the likelihood ratio values computed using *RIS* and *BS* have a clear departure from the theoretical mixtures. As mentioned before, for $n = 10$ Crainiceanu and Ruppert's result requires that n_i be reasonably large. It is interesting to note that the departure also exists for $n_i = 2$ and $n = 100$. The asymptotic mixture of 50:50 is theorectically valid for any cluster size n_i although it requires that the number of clusters $n \to \infty$. The asymptotic distribution does seem to provide a reasonable approximation for $n = 100$ and $n_i = 20$. For $n_i = 2$ we noticed (data not shown here) that the distribution of the likelihood ratio statisitcs (computed using *RIS* and *BS*) is much better approximated by the 50:50 mixture when n is as large as 250.

6 An example

In this section we consider the multi-center non-small cell lung cancer trial that was used as an example in Vaida and Xu (2000). The trial enrolled 579 patients from 31 institutions. The primary endpoint was patient death. There were two randomized treatment arms in the trial, a standard chemotherapy (CAV) arm and an alternating regimen (CAV-HEM) arm. Other important covariates that affected patient survival were: presence or absence of bone metastases, presence or absence of liver metastases, performance status at study entry and whether there was weight loss prior to entry. Gray (1995) used a score test for the existence of random treatment effect, and found it to be significant.

In the following we mainly consider the three nested models of Vaida and Xu (2000); they are named Models 1-3 in Table 1. They all include the fixed effects of the five covariates. Model 1 includes no random effect; Model 2 includes a random treatment effect; and Model 3 includes random treatment and random bone metastases effects. The estimate of the other variance components corresponding to potential random effects for the rest three of the covariates, as well as random center effect on the baseline hazard function (see also Gray, 1995), converged to zero during the *EM* algorithm (Vaida and Xu, 2000). The parameter estimates under the three models were given in Table 1 of Vaida and Xu (2000). Table 1 here gives minus twice the log likelihood values for the models, computed using Laplace approximation, reciprocal importance sampling and bridge sampling for models 2 and 3. Note that the likelihood can be computed directly when there are no random effects, and such is the case for Models 1 and 0 (see below). The likelihood values for Models 2 and 3 are computed after 50 *EM* steps where the maximum likelihood estimate has converged; the sample sizes for Gibbs sampler during *MCEM* are 100 initially and increased to 1000 for the last 10 *EM* steps. The Monte Carlo sample sizes for *RIS* and *BS* are 1000, respectively. ¿From the table we see that the values of the log likelihoods agree well among the three computational methods.

As seen in the table, if we are to test Model 2 versus Model 1 using the likelihood ratio statistic, its sampling distribution under Model 1 is asymptotically $(\chi^2_0 + \chi^2_1)/2$, according to *Case 1* of Section 4.1, with critical value of 2.71 at .05 significance level. Model 1 is then rejected in favor of Model 2. Similarly, to test Model 3 versus Model 2, the likelihood ratio statistic is again asymptotically $(\chi^2_0 + \chi^2_1)/2$ under Model 2. This is a special case of *Case 4*, and the mixing probabilities can be derived directly as in *Case 1*. Therefore Model 2 is rejected in favor of Model 3. Note that the finite sample distribution we considered in Section 5 puts more point mass at zero, leading to even smaller critical values for the likelihood ratio statistic.

We can also compare Models 1 and 3 directly. Under Model 1 the asymptotic distribution of the likelihood ratio statistic is a mixture of χ_0^2, χ_1^2 and χ_2^2. This is againn *Case 4* in Section 4.1. The mixing probabilities are not straightforward to compute; however, given that the 0.95 quantile of χ_2^2 is 5.99, and that the same quantile for the mixture is smaller, Model 1 is therefore rejected in favor of Model 3.

Finally, Model 0 is the Cox model with only fixed effects for the 4 covariates other than treatment. The comparison of Model 0 versus Model 2 provides an illustration for *Case 2* of Section 4.1, i.e. neither the fixed nor the random treatment effect is significant. Here $q = 0$ and $r = 1$, so the null asymptotic distribution of the likelihood ratio statistic is $(\chi_1^2 + \chi_2^2)/2$. It is again easy to see that Model 0 is rejected in favor of Model 2 at 0.05 signficance level.

Alternatively, we can use the profile *AIC* to compare the nested models. From the table it is also clear that the larger models are chosen by the criterion.

7 Discussion

In this paper we established the asymptotic properties of the nonparametric maximum likelihood estimator under the proportional hazards mixed effects model. Motivated by model selection problems under PHMM, we developed the profile likelihood ratio test and a profile Akaike information criterion that are generally applicable to models with nuisance parameters. The development was based on the asymptotic quadratic expansion of the log profile likelihood function. The profile likelihood ratio test for the null hypothesis that lies in the interior of the parameter space was given in Murphy and van der Vaart (2000); here we further developed it for testing on the boundary. The profile *AIC* has not been previously proposed in the literature, to our best knowledge. It applies to both parametric and semiparamtric models, and for the latter type of models the focus is on the finite dimenstional parameter. The *AIC* approach does not encounter the boundary problem as in hypothesis testing. The profile *AIC* also provides a theoretical justification for the use of the partial likelihood in the *AIC* under the classic Cox model.

Model selection has been an area of growing interest in the recent years. In this paper we restricted our attention to the classic derivation of the Akaike information criterion. However we acknowledge, as Longford (2005) pointed out, that, whatever the selection criterion, single-model based inference can be inherently biased. Alternatives may include the use of a mixture of plausible models, and the focused information criteria of Claeskens and Hjort (2003). The associated new challenges of such improvements in practice are model interpretability and variability of inferences following the model averaging or selection.

For computation of the maximized likelihood, the Laplace approximation is the most straightforward but is only accurate when the cluster sizes are reasonably large. In view of the *MCEM* algorithm that is used to fit the PHMM, the additional computation of *RIS* or *BS* is often comparable to one step of the *MCEM*. Therefore we include *RIS* and *BS* as default in our computational program.

Finally, under linear mixed models when the interest lies in the inference of the random effects themselves, Vaida and Blanchard (2005) propose a conditional *AIC* using the notion of effective degrees of freedom. The usefulness of conditional inference carries over to PHMM, and it is our future work to develop a conditional *AIC* under the PHMM. Additionally, the finite sample distribution of the likelihood ratio statistic for testing zero variance components is another area that requires further work.

APPENDIX

PROOF OF THEOREM 1. To prove consistency we follow methods used by Murphy (1994) and Zeng *et al.* (2005). First prove $\hat{\Lambda}_n(\cdot)$ is bounded on $[0, \tau]$. We then invoke the compactness of the parameter space and Helly's selection theorem to conclude the existence of convergent subsequence of $\{\theta_n\}$. Finally we show the limit of this subsequence must be θ_0.

Step 1. We show $\hat{\Lambda}_n(\cdot)$ has an upper bound int $[0, \tau]$. First let

$$\bar{\Lambda}_n(t) = \sum_{ij} \frac{\delta_{ij}(1 - Y_{ij}(t))}{\sum_{kl} Y_{kl}(X_{ij}) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{kl}} \mathrm{E}_\theta(e^{\mathbf{b}_k' \mathbf{W}_{kl}} | \mathbf{y}_k)},$$

$$a_i(t) = \sum_{j=1}^{n_i} \int_{u=0}^{t} \{dN_{ij}(u) - Y_{ij}(u) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij}} \mathrm{E}_\theta(e^{\mathbf{b}_i' \mathbf{W}_{ij}} | \mathbf{y}_i)\, d\Lambda_0(u)\},$$

$$f_n(u) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n_i} Y_{ij}(u) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij}} \mathrm{E}_\theta(e^{\mathbf{b}_i' \mathbf{W}_{ij}} | \mathbf{y}_i).$$

We show $\sup_{t \in [0,\tau]} |\bar{\Lambda}_n(t) - \Lambda_0(t)| \to 0$ almost surely.

Note that $\{a_i(t) : i = 1, 2, \dots\}$ is a mean zero independent sequence for fixed t. Also, by the boundedness assumptions on $\mathbf{W}_{ij}$ and n_i:

$$\mathrm{E}_\theta(e^{\mathbf{b}_i' \mathbf{W}_{ij}} | \mathbf{y}_i) < B_{\boldsymbol{\Sigma}_0} < \infty$$

for some constant $B_{\boldsymbol{\Sigma}_0}$. Similarly $e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij}} < B_{\boldsymbol{\beta}_0} < \infty$, and since $a_i(t)$ is bounded for any $t \in [0, \tau]$ we have $\mathrm{Var}(a_i(t))$ is bounded and by the *SLLN* $n^{-1} \sum_i a_i(t) \to 0$ almost surely.

Similarly, $f_n(u) - \mathrm{E}(f_n(u)) \to 0$ almost surely. Since $\mathrm{E}[Y_{ij}(u) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij}} \mathrm{E}_\theta(e^{\mathbf{b}_i' \mathbf{W}_{ij}} | \mathbf{y}_i)]$ is bounded away from zero, there exists some $c_1 > 0$ such that eventually $f_n(u) \geq c_1$ almost surely. Likewise, since n_i is bounded, there exist some $c_2 > 0$ such that $f_n(u) \leq c_2$.

Now consider

$$\sum_{ij} \int_{u=0}^{t} \left\{ dN_{ij}(u) - Y_{ij}(u) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij}} \mathrm{E}_\theta(e^{\mathbf{b}_i' \mathbf{W}_{ij}} | \mathbf{y}_i)\, d\bar{\Lambda}_n(u) \right\} = 0, \quad (20)$$

since by switching the order of summation,

$$
\begin{aligned}
LHS &= \sum_{ij} \left\{ \delta_{ij}(1 - Y_{ij}(t)) - \sum_{kl} \frac{Y_{ij}(X_{kl}) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij}} \mathrm{E}_\theta(e^{\mathbf{b}_i' \mathbf{W}_{ij}} | \mathbf{y}_i) \delta_{kl}(1 - Y_{kl}(t))}{\sum_{rs} Y_{rs}(X_{kl}) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{rs}} \mathrm{E}_\theta(e^{\hat{\mathbf{r}}' \mathbf{W}_{rs}} | \mathbf{y}_r)} \right\} \\
&= \sum_{ij} \delta_{ij}(1 - Y_{ij}(t)) - \sum_{kl} \left\{ \frac{\sum_{ij} Y_{ij}(X_{kl}) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij}} \mathrm{E}_\theta(e^{\mathbf{b}_i' \mathbf{W}_{ij}} | \mathbf{y}_i) \delta_{kl}(1 - Y_{kl}(t))}{\sum_{rs} Y_{rs}(X_{kl}) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{rs}} \mathrm{E}_\theta(e^{\hat{\mathbf{r}}' \mathbf{W}_{rs}} | \mathbf{y}_r)} \right\} \\
&= 0.
\end{aligned}
$$

Now by adding and subtracting $dN_{ij}(u)$ in (20) we have for fixed t

$$
\begin{aligned}
\int_{u=0}^{t} f_n(u)\, d(\Lambda_0 - \bar{\Lambda}_n)(u) &= n^{-1} \sum_{ij} \int_{u=0}^{t} Y_{ij}(u) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij}} \mathrm{E}_\theta(e^{\mathbf{b}_i' \mathbf{W}_{ij}} | \mathbf{y}_i)\, d(\Lambda_0 - \bar{\Lambda}_n)(u) \\
&= n^{-1} \sum_{ij} \int_{u=0}^{t} \{ dN_{ij}(u) - Y_{ij}(u) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij}} \mathrm{E}_\theta(e^{\mathbf{b}_i' \mathbf{W}_{ij}} | \mathbf{y}_i)\, d\Lambda_0(u) \} \\
&= n^{-1} \sum_{i=1}^{n} a_i(t) \\
&\to 0 \text{ a.s.},
\end{aligned}
$$

by SLLN. Futhermore

$$
c_1 \int_{u=0}^{t} d(\Lambda_0 - \bar{\Lambda}_n)(u) \leq \int_{u=0}^{t} f_n(u)\, d(\Lambda_0 - \bar{\Lambda}_n)(u) \to 0 \text{ a.s.}
$$

and

$$
c_2 \int_{u=0}^{t} d(\Lambda_0 - \bar{\Lambda}_n)(u) \geq \int_{u=0}^{t} f_n(u)\, d(\Lambda_0 - \bar{\Lambda}_n)(u) \to 0 \text{ a.s.}.
$$

Since c_1 and c_2 are both positive, we must have

$$
\int_{u=0}^{t} d(\Lambda_0 - \bar{\Lambda}_n)(u) \to 0 \text{ a.s.},
$$

which implies $\bar{\Lambda}_n(t) \to \Lambda_0(t)$ a.s. for all $t \in [0, \tau]$. Pointwise convergence of non-decreasing functions to a continuous limit implies local (on $[0, \tau]$ in particular) uniform continuity.

Since $\hat{\boldsymbol{\beta}}_n$, $\hat{\boldsymbol{\Sigma}}_n$, $\mathbf{Z}_{kl}$, and $\mathbf{W}_{kl}$ are in compact sets, there exists some finite, possibly negative C such that

$$\hat{\boldsymbol{\beta}}_n'\mathbf{Z}_{kl} + \log \mathrm{E}_{\hat{\theta}_n}[e^{\mathbf{b}_k'\mathbf{W}_{kl}}|\mathbf{y}_k] \geq \boldsymbol{\beta}_0'\mathbf{Z}_{kl} + \log \mathrm{E}_{\theta_0}[e^{\mathbf{b}_k'\mathbf{W}_{kl}}|\mathbf{y}_k] + C.$$

Therefore

$$\begin{aligned}\hat{\Lambda}_n(\tau) &= \sum_{ij} \frac{\delta_{ij}(1 - Y_{ij}(\tau))}{\sum_{kl} Y_{kl}(X_{ij}) \exp\{\hat{\boldsymbol{\beta}}_n'\mathbf{Z}_{kl} + \log \mathrm{E}_{\hat{\theta}_n}[e^{\mathbf{b}_k'\mathbf{W}_{kl}}|\mathbf{y}_k]\}} \\ &\leq \sum_{ij} \frac{\delta_{ij}(1 - Y_{ij}(\tau))}{\sum_{kl} Y_{kl}(X_{ij}) \exp\{\boldsymbol{\beta}_0'\mathbf{Z}_{kl} + \log \mathrm{E}_{\theta_0}[e^{\mathbf{b}_k'\mathbf{W}_{kl}}|\mathbf{y}_k] + C\}} \\ &= e^{-C}\bar{\Lambda}_n(\tau) \to e^{-C}\Lambda_0(\tau).\end{aligned}$$

Step 2. Since $\hat{\Lambda}$ has an upper bound almost surely, and $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{\Sigma}}_n$ are in compact sets, we can use Helly's selection theorem to establish a convergent subsequence which we now denote by $\hat{\theta}_n = (\hat{\Lambda}_n, \hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\Sigma}}_n)$ with limit θ^*.

Taking limits of both sides of

$$\hat{\Lambda}_n(t) = \int_0^t \frac{\sum_{kl} Y_{kl}(u) \exp\{\boldsymbol{\beta}_0'\mathbf{Z}_{kl} + \log \mathrm{E}_{\theta_0}[e^{\mathbf{b}_k'\mathbf{W}_{kl}}|\mathbf{y}_k]\}}{\sum_{kl} Y_{kl}(u) \exp\{\hat{\boldsymbol{\beta}}_n'\mathbf{Z}_{kl} + \log \mathrm{E}_{\hat{\theta}_n}[e^{\mathbf{b}_k'\mathbf{W}_{kl}}|\mathbf{y}_k]\}}\, d\bar{\Lambda}_n(u) \qquad (21)$$

we see that Λ^* is absolutely continuous with respect to Λ_0. Furthermore, $\Lambda^*(t)$ is differentiable with respect to t and $d\hat{\Lambda}_n(t)/\, d\bar{\Lambda}_n(t)$ converges to $d\Lambda^*(t)/\, d\Lambda_0(t)$. Note that the finite sample likelihood as expressed via (3) has no finite maximum, since λ is free to go to infinity at any X_{ij}. We restrict Λ to be right continuous with jumps at X_{ij}; and for cluster i, conditional on the random effect $\mathbf{b}_i$, we let the log-likelihood be

$$l_i = l_i(\boldsymbol{\beta}, \lambda; \mathbf{y}_i|\mathbf{b}_i) = \sum_{j=1}^{n_i} \{\delta_{ij} \log \Lambda\{X_{ij}\} + \delta_{ij}(\boldsymbol{\beta}'\mathbf{Z}_{ij} + \mathbf{b}_i'\mathbf{W}_{ij}) - \Lambda(X_{ij})e^{\boldsymbol{\beta}'\mathbf{Z}_{ij} + \mathbf{b}_i'\mathbf{W}_{ij}}\}, (22)$$

where $\Lambda\{t\}$ is the size of the jump in Λ at t. The likelihood of the observed data, $L_n(\theta)$, is still as defined in (3) and we let $l_n(\theta) = \log L_n(\theta)$. In place of Λ_0, which is continuous at X_{ij}, we use $\bar{\Lambda}_n$. In particular we have:

$$
\begin{aligned}
0 &\leq n^{-1}\{l_n(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\Sigma}}_n, \hat{\Lambda}_n) - l_n(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, \bar{\Lambda}_n)\} \\
&= n^{-1}\sum_{i=1}^{n} \log\left\{\int_{\mathbf{b}} R_i(\hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n, \mathbf{b})\mathrm{p}(\mathbf{b}, \hat{\boldsymbol{\Sigma}}_n)d\mathbf{b}\right\} \\
&\quad - n^{-1}\sum_{i=1}^{n} \log\left\{\int_{\mathbf{b}} R_i(\boldsymbol{\beta}_0, \bar{\Lambda}_n, \mathbf{b})\mathrm{p}(\mathbf{b}, \boldsymbol{\Sigma}_0)d\mathbf{b}\right\} \\
&\quad + n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i} \delta_{ij}\log(\hat{\Lambda}_n\{X_{ij}\}/\bar{\Lambda}_n\{X_{ij}\})
\end{aligned}
$$

where

$$
R_i(\boldsymbol{\beta}, \Lambda, \mathbf{b}) = \prod_{j=1}^{n_i} \exp[\delta_{ij}(\boldsymbol{\beta}'\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij}) - \Lambda(X_{ij})\exp(\boldsymbol{\beta}'\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij})].
$$

Letting $n \to \infty$ we have

$$
\begin{aligned}
0 \leq \mathrm{E}\log &\left\{\int_{\mathbf{b}} R_i(\boldsymbol{\beta}^*, \Lambda^*, \mathbf{b})\mathrm{p}(\mathbf{b}, \boldsymbol{\Sigma}^*)d\mathbf{b}\prod_{j=1}^{n_i}\lambda^*(X_{ij})^{\delta_{ij}}\right. \\
&\left.\times\left(\int_{\mathbf{b}} R_i(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b})\mathrm{p}(\mathbf{b}, \boldsymbol{\Sigma}_0)d\mathbf{b}\prod_{j=1}^{n_i}\lambda_0(X_{ij})^{\delta_{ij}}\right)^{-1}\right\}.
\end{aligned}
$$

Because the right side is negative the Kullback-Leibler information we have

$$
\int_{\mathbf{b}} R_i(\boldsymbol{\beta}^*, \Lambda^*, \mathbf{b})\mathrm{p}(\mathbf{b}, \boldsymbol{\Sigma}^*)d\mathbf{b}\prod_{j=1}^{n_i}\lambda^*(X_{ij})^{\delta_{ij}} = \int_{\mathbf{b}} R_i(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b})\mathrm{p}(\mathbf{b}, \boldsymbol{\Sigma}_0)d\mathbf{b}\prod_{j=1}^{n_i}\lambda_0(X_{ij})^{\delta_{ij}}
$$

or

$$
\begin{aligned}
&\int_{\mathbf{b}}\prod_{j=1}^{n_i}\lambda^*(X_{ij})^{\delta_{ij}}\exp[\delta_{ij}(\boldsymbol{\beta}^{*\prime}\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij}) - \Lambda^*(X_{ij})\exp(\boldsymbol{\beta}^{*\prime}\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij})]\mathrm{p}(\mathbf{b}, \boldsymbol{\Sigma}^*)d\mathbf{b} \\
&\quad = \int_{\mathbf{b}}\prod_{j=1}^{n_i}\lambda_0(X_{ij})^{\delta_{ij}}\exp[\delta_{ij}(\boldsymbol{\beta}_0'\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij}) - \Lambda_0(X_{ij})\exp(\boldsymbol{\beta}_0'\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij})]\mathrm{p}(\mathbf{b}, \boldsymbol{\Sigma}_0)d\mathbf{b}
\end{aligned} \tag{23}
$$

Now we use techniques adapted from Zeng *et al.* (2005) to conclude $\theta^* = \theta_0$. Fix some k in $1, \ldots, n_i$. For $j = 1, \ldots, k$, let $\delta_{ij} = 1, X_{ij} = 0$ in (23) and note that we assume $\Lambda^*(0) = \Lambda_0(0) = 0$. If $j = k+1, \ldots, n_i$ and $\delta_{ij} = 0$, we replace X_{ij} with τ. Otherwise, if $j = k+1, \ldots, n_i$ and $\delta_{ij} = 1$, we integrate X_{ij} from 0 to τ. We get:

$$
\begin{aligned}
&\int_{\mathbf{b}} \prod_{j=1}^{k} \lambda^*(0) \exp[\boldsymbol{\beta}^{*\prime}\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij}] \\
&\quad \times \prod_{j=k+1}^{n_i} \left\{\exp[-\Lambda^*(\tau)\exp(\boldsymbol{\beta}^{*\prime}\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij})]\right\}^{1-\delta_{ij}} \\
&\quad \times \prod_{j=k+1}^{n_i} \left\{\int_{y=0}^{\tau} \lambda^*(y) \exp[\boldsymbol{\beta}^{*\prime}\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij} - \Lambda^*(y)\exp(\boldsymbol{\beta}^{*\prime}\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij})]dy\right\}^{\delta_{ij}} \mathrm{p}(\mathbf{b}, \boldsymbol{\Sigma}^*)d\mathbf{b} \\
&= \int_{\mathbf{b}} \prod_{j=1}^{k} \lambda_0(0) \exp[\boldsymbol{\beta}_0'\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij}] \\
&\quad \times \prod_{j=k+1}^{n_i} \left\{\exp[-\Lambda_0(\tau)\exp(\boldsymbol{\beta}_0{}'\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij})]\right\}^{1-\delta_{ij}} \\
&\quad \times \prod_{j=k+1}^{n_i} \left\{\int_{y=0}^{\tau} \lambda_0(y) \exp[\boldsymbol{\beta}_0{}'\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij} - \Lambda_0(y)\exp(\boldsymbol{\beta}_0{}'\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij})]dy\right\}^{\delta_{ij}} \mathrm{p}(\mathbf{b}, \boldsymbol{\Sigma}_0)d\mathbf{b}
\end{aligned}
$$

or

$$
\begin{aligned}
&\int_{\mathbf{b}} \prod_{j=1}^{k} \lambda^*(0) \exp[\boldsymbol{\beta}^{*\prime}\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij}] \\
&\quad \times \prod_{j=k+1}^{n_i} \left\{\exp[-\Lambda^*(\tau)\exp(\boldsymbol{\beta}^{*\prime}\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij})]\right\}^{1-\delta_{ij}} \\
&\quad \times \prod_{j=k+1}^{n_i} \left\{1 - \exp[-\Lambda^*(\tau)\exp(\boldsymbol{\beta}^{*\prime}\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij})]\right\}^{\delta_{ij}} \mathrm{p}(\mathbf{b}, \boldsymbol{\Sigma}^*)d\mathbf{b} \\
&= \int_{\mathbf{b}} \prod_{j=1}^{k} \lambda_0(0) \exp[\boldsymbol{\beta}_0'\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij}] \\
&\quad \times \prod_{j=k+1}^{n_i} \left\{\exp[-\Lambda_0(\tau)\exp(\boldsymbol{\beta}_0{}'\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij})]\right\}^{1-\delta_{ij}} \\
&\quad \times \prod_{j=k+1}^{n_i} \left\{1 - \exp[-\Lambda_0(\tau)\exp(\boldsymbol{\beta}_0{}'\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij})]\right\}^{\delta_{ij}} \mathrm{p}(\mathbf{b}, \boldsymbol{\Sigma}_0)d\mathbf{b}. \qquad (24)
\end{aligned}
$$

Because δ_{ij} are arbitrary, we sum the two sides of (24) over all possible δ_{ij} to yield:

$$\int_{\mathbf{b}} \prod_{j=1}^{k} \lambda^*(0) \exp[\boldsymbol{\beta}^{*\prime}\mathbf{Z}_{ij}+\mathbf{b}'\mathbf{W}_{ij}]\mathrm{p}(\mathbf{b},\boldsymbol{\Sigma}^*)d\mathbf{b} = \int_{\mathbf{b}} \prod_{j=1}^{k} \lambda_0(0) \exp[\boldsymbol{\beta}_0'\mathbf{Z}_{ij}+\mathbf{b}'\mathbf{W}_{ij}]\mathrm{p}(\mathbf{b},\boldsymbol{\Sigma}_0)d\mathbf{b}$$

and

$$\exp\left\{\sum_{j=1}^{k} \boldsymbol{\beta}^{*\prime}\mathbf{Z}_{ij} + \frac{(\sum_{j=1}^{k}\mathbf{W}_{ij})'\boldsymbol{\Sigma}^*(\sum_{j=1}^{k}\mathbf{W}_{ij})}{2}\right\}\lambda^*(0)^k$$
$$= \exp\left\{\sum_{j=1}^{k} \boldsymbol{\beta}_0{}'\mathbf{Z}_{ij} + \frac{(\sum_{j=1}^{k}\mathbf{W}_{ij})'\boldsymbol{\Sigma}_0(\sum_{j=1}^{k}\mathbf{W}_{ij})}{2}\right\}\lambda_0(0)^k$$

We assume $\lambda^*(0) > 0$. Since the index set can be replaced by any subset of $1, \ldots, n_i$ we have

$$\mathbf{W}_{ij}'\boldsymbol{\Sigma}^*\mathbf{W}_{ij'} = \mathbf{W}_{ij}'\boldsymbol{\Sigma}_0\mathbf{W}_{ij'},\ j \neq j' : j, j' = 1, \ldots, n_i,$$

and

$$\boldsymbol{\beta}^{*\prime}\mathbf{Z}_{ij} + \frac{\mathbf{W}_{ij}'\boldsymbol{\Sigma}^*\mathbf{W}_{ij}}{2} + \log\lambda^*(0)$$
$$= \boldsymbol{\beta}_0{}'\mathbf{Z}_{ij} + \frac{\mathbf{W}_{ij}'\boldsymbol{\Sigma}_0\mathbf{W}_{ij}}{2} + \log\lambda_0(0),\ j = 1, \ldots, n_i$$

Therefore, under C3, $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_0$, $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$, and $\lambda^*(0) = \lambda_0(0)$.

To show $\Lambda^* = \Lambda_0$, we manipulate the terms of (23) again. Let $\delta_{i1} = 1$ and integrate X_{i1} from 0 to t. Also for $j = 2, \ldots, n_i$, if $\delta_{ij} = 0$, replace X_{ij} with τ and if $\delta_{ij} = 1$ integrate X_{ij} from 0 to τ. Summing the result over all possible $\{\delta_{ij} : j = 2, \ldots, n_i\}$, this time we get

$$\int_{\mathbf{b}} 1 - \exp[-\Lambda^*(t)\exp(\boldsymbol{\beta}_0{}'Z_{i1} + \mathbf{b}'\mathbf{W}_{i1})]\mathrm{p}(\mathbf{b},\boldsymbol{\Sigma}_0)d\mathbf{b}$$
$$= \int_{\mathbf{b}} 1 - \exp[-\Lambda_0(t)\exp(\boldsymbol{\beta}_0{}'Z_{i1} + \mathbf{b}'\mathbf{W}_{i1})]\mathrm{p}(\mathbf{b},\boldsymbol{\Sigma}_0)d\mathbf{b}. \qquad (25)$$

Because both sides of (25) are strictly monotone in $\Lambda^*(t)$ and $\Lambda_0(t)$, we have $\Lambda^*(t) = \Lambda_0(t)$. Since Λ_0 is non-decreasing and continuous, the pointwise convergence can be extended to uniform convergence on $[0, \tau]$.

PROOF OF THEOREM 2. To prove asymptotic normality and efficiency we invoke methods of Murphy (1995) and Zeng *et al.* (2005). Consider the set

$$\mathcal{H} = \big\{(\mathbf{h}_1, \mathbf{h}_2, h_3) : \mathbf{h}_1 \in \mathbf{R}^{d_1}, \mathbf{h}_2 \in \mathbf{R}^{d_2(d_2+1)/2}, \\ h_3(\cdot) \text{ is a function on } [0,\tau]; \|\mathbf{h}_1\|, \|\mathbf{h}_2\|, \|h_3\|_V \leq 1\big\} \quad (26)$$

where $\|h_3\|_V$ denotes the total variation of $h_3(\cdot)$ in $[0,\tau]$. We define a sequence of maps S_n mapping a neighborhood of $(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, \Lambda_0)$, denoted by $\mathcal{U}$, in the parameter space for $(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \Lambda)$ into $l^\infty(\mathcal{H})$ as:

$$\begin{aligned} & S_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \Lambda)[\mathbf{h}_1, \mathbf{h}_2, h_3] \\ & \equiv n^{-1} \frac{d}{d\epsilon} l_n \left(\boldsymbol{\beta} + \epsilon \mathbf{h}_1, \boldsymbol{\Sigma} + \epsilon \mathbf{h}_2, \Lambda(t) + \epsilon \int_0^t h_3(s)\, d\Lambda(s)\right) \Big|_{\epsilon=0} \\ & \equiv A_{n1}[\mathbf{h}_1] + A_{n2}[\mathbf{h}_2] + A_{n3}[h_3] \end{aligned}$$

where $\boldsymbol{\Sigma}$ is treated as extended column vector consisting of the upper triangle elements; and A_{np}, $p = 1,2,3$, are linear functionals on $\mathbf{R}^{d_1}, \mathbf{R}^{d_2(d_2+1)/2}$ and $BV[0,\tau]$ (the space of functions with finite total variation in $[0,\tau]$). If we let $l_{\boldsymbol{\beta}}$, $l_{\boldsymbol{\Sigma}}$ and l_Λ be the score functions for $\boldsymbol{\beta}, \boldsymbol{\Sigma}$, and Λ (along $\int_0^t 1 + \epsilon h_3(s)\, d\Lambda(s)$) for a single cluster, then

$$A_{n1}[\mathbf{h}_1] = \mathcal{P}_n[\mathbf{h}_1' l_{\boldsymbol{\beta}}],\ A_{n2}[\mathbf{h}_2] = \mathcal{P}_n[\mathbf{h}_2' l_{\boldsymbol{\Sigma}}],\ \text{and } A_{n3}[h_3] = \mathcal{P}_n\left[l_\Lambda[h3]\right]$$

where $\mathcal{P}_n$ denotes the empirical measure based on n independent clusters. We now seek explicit expression for A_{np}. Recall the log likelihood

$$n^{-1} l_n(\theta) = n^{-1} \sum_{i=1}^n \log \left\{ \int_{\mathbf{b}} R_i(\boldsymbol{\beta}, \Lambda, \mathbf{b}) \mathrm{p}(\mathbf{b}, \boldsymbol{\Sigma}) d\mathbf{b} \right\} + n^{-1} \sum_{i=1}^n \sum_{j=1}^{n_i} \delta_{ij} \log \Lambda\{X_{ij}\}$$

where

$$R_i(\boldsymbol{\beta}, \Lambda, \mathbf{b}) = \exp\left\{ \sum_{j=1}^{n_i} \delta_{ij}(\boldsymbol{\beta}'\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij}) - \Lambda(X_{ij}) \exp(\boldsymbol{\beta}'\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij}) \right\}.$$

Note that

$$\frac{\partial}{\partial \epsilon} R_i(\boldsymbol{\beta} + \epsilon \mathbf{h}_1, \Lambda, \mathbf{b})\big|_{\epsilon=0} = R_i(\boldsymbol{\beta}, \Lambda, \mathbf{b}) \sum_{j=1}^{n_i} \mathbf{h}_1' \mathbf{Z}_{ij} \left(\delta_{ij} - \Lambda(X_{ij}) \exp(\boldsymbol{\beta}'\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij})\right).$$

Furthermore let $\Lambda_\epsilon(t) = \int_0^t 1 + \epsilon h_3 \, d\Lambda$, then $\frac{\partial}{\partial \epsilon}\Lambda_\epsilon(t) = \int_0^t h_3(s)\, d\Lambda(s)$ and

$$\frac{\partial}{\partial \epsilon} R_i\left(\boldsymbol{\beta}, \Lambda_\epsilon, \mathbf{b}\right)\big|_{\epsilon=0} = -R_i\left(\boldsymbol{\beta}, \Lambda, \mathbf{b}\right) \sum_{j=1}^{n_i} \int_0^{X_{ij}} h_3(s)\, d\Lambda(s) \exp(\boldsymbol{\beta}'\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij}).$$

Also $\Lambda_\epsilon\{t\} = (1 + \epsilon h_3(t))\, \Lambda\{t\}$, so

$$\frac{d}{d\epsilon} \log \Lambda_\epsilon\{t\}\big|_{\epsilon=0} = \left.\frac{h_3(t)\Lambda\{t\}}{\Lambda_\epsilon\{t\}}\right|_{\epsilon=0} = h_3(t).$$

If we let $\mathcal{D}(\mathbf{h}_2)$ denote the matrix corresponding to the extended vector $\mathbf{h}_2$ and define the " $\cdot$ " operation on two matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ to be trace$(\mathbf{M}_1\mathbf{M}_1')$, then

$$\begin{aligned}\frac{\partial}{\partial \epsilon} \mathrm{p}(\mathbf{b}; \boldsymbol{\Sigma} + \epsilon \mathbf{h}_2)\big|_{\epsilon=0} &= \frac{\partial}{\partial \epsilon} |\boldsymbol{\Sigma} + \epsilon \mathbf{h}_2|^{-1/2} e^{-\mathbf{b}'(\boldsymbol{\Sigma}+\epsilon\mathbf{h}_2)^{-1}\mathbf{b}/2}\big|_{\epsilon=0} \\ &= \left\{\mathbf{b}'\boldsymbol{\Sigma}^{-1}\mathcal{D}(\mathbf{h}_2)\boldsymbol{\Sigma}^{-1}\mathbf{b}/2 - \boldsymbol{\Sigma}^{-1} \cdot \mathcal{D}(\mathbf{h}_2)/2\right\} e^{-\mathbf{b}'\boldsymbol{\Sigma}^{-1}\mathbf{b}/2}.\end{aligned}$$

Finally, we can explicitly write A_{np} as

$$\begin{aligned}A_{n1}[\mathbf{h}_1] =& n^{-1} \sum_{i=1}^{n} \left(\int_{\mathbf{b}} \sum_{j=1}^{n_i} \mathbf{h}_1' \mathbf{Z}_{ij} \left(\delta_{ij} - \Lambda(X_{ij}) e^{\boldsymbol{\beta}'\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij}}\right)\right. \\ & \left. \times R_i(\boldsymbol{\beta}, \Lambda, \mathbf{b}) e^{-\mathbf{b}'\boldsymbol{\Sigma}^{-1}\mathbf{b}/2} d\mathbf{b}\right) \\ & \times \left(\int_{\mathbf{b}} R_i(\boldsymbol{\beta}, \Lambda, \mathbf{b}) e^{-\mathbf{b}'\boldsymbol{\Sigma}^{-1}\mathbf{b}/2} d\mathbf{b}\right)^{-1} \\ A_{n2}[\mathbf{h}_2] =& n^{-1} \sum_{i=1}^{n} \left(\int_{\mathbf{b}} \left\{\mathbf{b}'\boldsymbol{\Sigma}^{-1}\mathcal{D}(\mathbf{h}_2)\boldsymbol{\Sigma}^{-1}\mathbf{b}/2 - \boldsymbol{\Sigma}^{-1} \cdot \mathcal{D}(\mathbf{h}_2)/2\right\}\right. \\ & \left. \times R_i(\boldsymbol{\beta}, \Lambda, \mathbf{b}) e^{-\mathbf{b}'\boldsymbol{\Sigma}^{-1}\mathbf{b}/2} d\mathbf{b}\right) \\ & \times \left(\int_{\mathbf{b}} R_i(\boldsymbol{\beta}, \Lambda, \mathbf{b}) e^{-\mathbf{b}'\boldsymbol{\Sigma}^{-1}\mathbf{b}/2} d\mathbf{b}\right)^{-1} \\ A_{n3}[h_3] =& n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \delta_{ij} h_3(X_{ij}) - \int_0^{X_{ij}} h_3(s)\, d\Lambda(s) \\ & \times \int_{\mathbf{b}} e^{\boldsymbol{\beta}'\mathbf{Z}_{ij} + \mathbf{b}'\mathbf{W}_{ij}} R_i(\boldsymbol{\beta}, \Lambda, \mathbf{b}) e^{-\mathbf{b}'\boldsymbol{\Sigma}^{-1}\mathbf{b}/2} d\mathbf{b} \\ & \times \left(\int_{\mathbf{b}} R_i(\boldsymbol{\beta}, \Lambda, \mathbf{b}) e^{-\mathbf{b}'\boldsymbol{\Sigma}^{-1}\mathbf{b}/2} d\mathbf{b}\right)^{-1}\end{aligned}$$

or

$$
\begin{aligned}
A_{n1}[\mathbf{h}_1] =& n^{-1}\sum_{i=1}^{n}\int_{\mathbf{b}}\sum_{j=1}^{n_i}\mathbf{h}_1'\mathbf{Z}_{ij}\left(\delta_{ij}-\Lambda(X_{ij})e^{\boldsymbol{\beta}'\mathbf{Z}_{ij}+\mathbf{b}'\mathbf{W}_{ij}}\right)d\mu_i(\mathbf{b})\\
A_{n2}[\mathbf{h}_2] =& n^{-1}\sum_{i=1}^{n}\int_{\mathbf{b}}\left\{\mathbf{b}'\boldsymbol{\Sigma}^{-1}\mathcal{D}(\mathbf{h}_2)\boldsymbol{\Sigma}^{-1}\mathbf{b}/2-\boldsymbol{\Sigma}^{-1}\cdot\mathcal{D}(\mathbf{h}_2)/2\right\}d\mu_i(\mathbf{b})\\
A_{n3}[h_3] =& n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\delta_{ij}h_3(X_{ij})-\int_0^{X_{ij}}h_3(s)\,d\Lambda(s)\int_{\mathbf{b}}e^{\boldsymbol{\beta}'\mathbf{Z}_{ij}+\mathbf{b}'\mathbf{W}_{ij}}d\mu_i(\mathbf{b})
\end{aligned}
$$

where

$$
d\mu_i(\mathbf{b})=\frac{R_i(\boldsymbol{\beta},\Lambda,\mathbf{b})e^{-\mathbf{b}'\boldsymbol{\Sigma}^{-1}\mathbf{b}/2}d\mathbf{b}}{\int_{\mathbf{b}}R_i(\boldsymbol{\beta},\Lambda,\mathbf{b})e^{-\mathbf{b}'\boldsymbol{\Sigma}^{-1}\mathbf{b}/2}d\mathbf{b}}
$$

We define the limit map $S:(\boldsymbol{\beta},\boldsymbol{\Sigma},\Lambda)[\mathbf{h}_1,\mathbf{h}_2,h_3]\to l^\infty(\mathcal{H})$ as

$$
S(\boldsymbol{\beta},\boldsymbol{\Sigma},\Lambda)[\mathbf{h}_1,\mathbf{h}_2,h_3]=A_1[\mathbf{h}_1]+A_2[\mathbf{h}_2]+A_3[h_3]
$$

where the linear functionals A_p are obtained by replacing the empirical sum in A_{np} by the expectation. By construction, $S_n(\hat{\boldsymbol{\beta}}_n,\hat{\boldsymbol{\Sigma}}_n,\hat{\Lambda}_n)=0$ and $S(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0,\Lambda_0)=0$.

Asymptotic normality will follow as desired by verifying the four conditions of Theorem 2 in Murphy (1995). First, $\sqrt{n}(S_n(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0,\Lambda_0)-S(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0,\Lambda_0))$ weakly converges to a tight Gaussian process on $l^\infty(\mathcal{H})$, because $\mathcal{H}$ is a Donsker class and the functionals A_{np} are bounded Lipschitz functionals with respect to $\mathcal{H}$. The approximation condition that

$$
\begin{aligned}
&\sup_{(\mathbf{h}_1,\mathbf{h}_2,h_3)\in\mathcal{H}}\left|(S_n-S)(\hat{\boldsymbol{\beta}}_n,\hat{\boldsymbol{\Sigma}}_n,\hat{\Lambda}_n)-(S_n-S)(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0,\Lambda_0)\right|\\
&\quad=o_p\left(n^{-1/2}\vee\left\{\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|+\|\hat{\boldsymbol{\Sigma}}_n-\boldsymbol{\Sigma}_0\|+\sup_{t\in[0,\tau]}|\hat{\Lambda}_n(t)-\Lambda_0(t)|\right\}\right)
\end{aligned}
$$

can be proved in a manner similar to Lemma 1 in the appendix of Murphy (1995). By the smoothness of $S(\boldsymbol{\beta},\boldsymbol{\Sigma},\Lambda)$, the Fréchet differentiability condition holds and the derivative of $S(\boldsymbol{\beta},\boldsymbol{\Sigma},\Lambda)$ at $(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0,\Lambda_0)$ by $\dot{S}(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0,\Lambda_0)$. We consider $\dot{S}(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0,\Lambda_0)$ to be a map, T, from the space

$$
\{(\boldsymbol{\beta}-\boldsymbol{\beta}_0,\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0,\Lambda-\Lambda_0):(\boldsymbol{\beta},\boldsymbol{\Sigma},\Lambda)\text{ is in the neighborhood }\mathcal{U}\text{ of }(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0,\Lambda_0)\}
$$

to $l^{\infty}(\mathcal{H})$. Lastly, we need to show the linear map, T, is continuously invertible on its range.

Now we can write

$$T(\boldsymbol{\beta}-\boldsymbol{\beta}_0, \boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0, \Lambda-\Lambda_0) = (\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\mathcal{Q}_1(\mathbf{h}_1, \mathbf{h}_2, h_3) + (\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0)'\mathcal{Q}_2(\mathbf{h}_1, \mathbf{h}_2, h_3) + \int_0^{\tau} \mathcal{Q}_3(\mathbf{h}_1, \mathbf{h}_2, h_3)\, d(\Lambda-\Lambda_0)$$

where the $\mathcal{Q}_i$ are the respective partial derivatives of S with respect to $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, and Λ. The $\mathcal{Q}_i$ are of the form

$$\mathcal{Q}_1(\mathbf{h}_1, \mathbf{h}_2, h_3) = B_1 \begin{pmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{pmatrix} + \int_0^{\tau} h_3(t) D_1(t)\, dt,$$
$$\mathcal{Q}_2(\mathbf{h}_1, \mathbf{h}_2, h_3) = B_2 \begin{pmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{pmatrix} + \int_0^{\tau} h_3(t) D_2(t)\, dt,$$

and

$$\mathcal{Q}_3(\mathbf{h}_1, \mathbf{h}_2, h_3) = B_3 \begin{pmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{pmatrix} + b_4 h_3(t) + \int_0^{\tau} h_3(t) D_3(t)\, dt;$$

where B_1, B_2, and B_3 are constant matrices; $D_1(t)$, $D_2(t)$, $D_3(t)$ are continuously differentiable functions; and $b_4 > 0$; each of which depends on θ_0. Therefore the operator $\mathcal{Q} = (\mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_3)'$ can be considered the sum of a continuously invertible operator and a compact operator from $\mathcal{H}$ to itself.

To prove T is invertible, we need only show the invertibility of the linear operator $\mathcal{Q}(\mathbf{h}_1, \mathbf{h}_2, h_3)$; or equivalently that $\mathcal{Q}$ is one-to-one (Zeng et al. 2005; Rudin 1973, pp. 99-103). Suppose $\mathcal{Q}(\mathbf{h}_1, \mathbf{h}_2, h_3) = \mathbf{0}$, then $T(\boldsymbol{\beta}-\boldsymbol{\beta}_0, \boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0, \Lambda-\Lambda_0)[\mathbf{h}_1, \mathbf{h}_2, h_3] = \mathbf{0}$ for any $(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \Lambda)$ in the neighborhood $\mathcal{U}$. In particular, fix some small constant ϵ and let

$$\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \epsilon \mathbf{h}_1, \quad \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 + \epsilon \mathbf{h}_2,$$
$$\Lambda(t) = \Lambda_0(t) + \epsilon \int_0^t h_3(t)\, d\Lambda_0(t).$$

By definition of T, we have

$$\begin{aligned} 0 =& T(\boldsymbol{\beta}-\boldsymbol{\beta}_0, \boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0, \Lambda-\Lambda_0)[\mathbf{h}_1, \mathbf{h}_2, h_3] \\ =& \epsilon \mathrm{E}\{(l_{\boldsymbol{\beta}_0}[\mathbf{h}_1] + l_{\boldsymbol{\Sigma}_0}[\mathbf{h}_2] + l_{\Lambda_0}[h_3])^2\}, \end{aligned}$$

so that $l_{\boldsymbol\beta_0}[\mathbf{h}_1] + l_{\boldsymbol\Sigma_0}[\mathbf{h}_2] + l_{\Lambda_0}[h_3] = 0$ almost surely. Expanding this expression we get

$$
\begin{aligned}
0 = &\sum_{j=1}^{n_i} \int_{\mathbf{b}} \mathbf{h}_1' \mathbf{Z}_{ij} \left(\delta_{ij} - \Lambda_0(X_{ij}) e^{\boldsymbol\beta_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij}} \right) R_{2i}(\boldsymbol\beta_0, \Lambda_0, \mathbf{b})\, d_{\mathbf{b}} N(\mathbf{0}, \boldsymbol\Sigma_0) \\
&+ \int_{\mathbf{b}} \left\{ \mathbf{b}' \boldsymbol\Sigma_0^{-1} \mathcal{D}(\mathbf{h}_2) \boldsymbol\Sigma_0^{-1} \mathbf{b}/2 - \boldsymbol\Sigma_0^{-1} \cdot \mathcal{D}(\mathbf{h}_2)/2 \right\} R_{2i}(\boldsymbol\beta_0, \Lambda_0, \mathbf{b})\, d_{\mathbf{b}} N(\mathbf{0}, \boldsymbol\Sigma_0) \\
&+ \sum_{j=1}^{n_i} \int_{\mathbf{b}} \left(\delta_{ij} h_3(X_{ij}) - \int_0^{X_{ij}} h_3(s)\, d\Lambda_0(s) e^{\boldsymbol\beta_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij}} \right) R_{2i}(\boldsymbol\beta_0, \Lambda_0, \mathbf{b})\, d_{\mathbf{b}} N(\mathbf{0}, \boldsymbol\Sigma_0)
\end{aligned}
\tag{27}
$$

where

$$
\begin{aligned}
R_{2i}(\boldsymbol\beta_0, \Lambda_0, \mathbf{b}) &= R_i(\boldsymbol\beta_0, \Lambda_0, \mathbf{b}) \prod_{j=1}^{n_i} \{\lambda_0(X_{ij})\}^{\delta_{ij}} \\
&= \prod_{j=1}^{n_i} \exp[\delta_{ij}(\boldsymbol\beta_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij}) - \Lambda_0(X_{ij}) \exp(\boldsymbol\beta_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij})] \{\lambda_0(X_{ij})\}^{\delta_{ij}}.
\end{aligned}
$$

Using techniques from Zeng *et al.* (2005) similar to the identifiability step of the consistency proof, we show that (27) implies $\mathbf{h}_1 = \mathbf{0}$, $\mathbf{h}_2 = \mathbf{0}$, and $h_3 = 0$. Let $\mathbf{Z}_{ij}$ and $\mathbf{W}_{ij}$ be fixed. Then for fixed integer k in $1, \ldots, n_i$, we define measures $\mu_1, \ldots, \mu_{n_i}$ on the set $\{0,1\} \times [0, \tau]$ as follows:

$$\mu_m(\{0\} \times A) = 0, \quad \mu_m(\{1\} \times A) = I(0 \in A), \quad m \le k,$$

and

$$\mu_m(\{0\} \times A) = I(\tau \in A), \quad \mu_m(\{1\} \times A) = \int I_A dx, \quad m > k,$$

where A is any Borel set in $[0, \tau]$. We integrate both sides of (27) with respect to $\{(\delta_{i1}, X_{i1}), \ldots, (\delta_{in_i}, X_{in_i})\}$ and the product measure $d\mu_1, \ldots, d\mu_{n_i}$. That is, we let $\delta_{im} = 1$ and $X_{im} = 0$ for all $m \le k$. Where $m > k$, we choose $X_{im} = \tau$ if $\delta_{im} = 0$, integrate X_{im} from 0 to τ if $\delta_{im} = 1$, then sum over $\delta_{ij} \in \{0,1\}$. Then we sum all of the equalities of (27) for all possible combinations of $\{\delta_{i1}, \ldots, \delta_{in_i}\} \in \{0,1\}^{n_i - k}$.

We compute the integral of each term on the right side of (27) with respect to the product measure, $\prod_{m=1}^{n_i} \mu_m$, the sum of which must be 0. First note, for any $\mathbf{b}$,

$$
\begin{aligned}
&\int R_{2i}(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b})\, d\left(\prod_{m=1}^{n_i} \mu_m\right) \\
&= \prod_{m \le k} \{\lambda_0(0) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im}}\} \\
&\quad \times \sum_{\substack{\delta_{im} \in \{0,1\} \\ m>k}} \prod_{m>k} \left(\exp[-\Lambda_0(\tau) \exp(\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im})]\right)^{1-\delta_{im}} \\
&\quad \times \left\{ \int_{y=0}^{\tau} \exp[\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im} - \Lambda_0(y) \exp(\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im})] \lambda_0(y) dy \right\}^{\delta_{im}} \\
&= \prod_{m \le k} \{\lambda_0(0) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im}}\} \\
&\quad \times \sum_{\substack{\delta_{im} \in \{0,1\} \\ m>k}} \prod_{m>k} \left(\exp[-\Lambda_0(\tau) \exp(\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im})]\right)^{1-\delta_{im}} \\
&\quad \times \left(1 - \exp[-\Lambda_0(\tau) \exp(\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im})]\right)^{\delta_{im}} \\
&= \prod_{m \le k} \{\lambda_0(0) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im}}\}
\end{aligned}
$$

For the first term of (27), if $j \le k$, then for any $\mathbf{b}$:

$$
\begin{aligned}
&\int \mathbf{h}_1' \mathbf{Z}_{ij} \left(\delta_{ij} - \Lambda_0(X_{ij}) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij}}\right) R_{2i}(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b})\, d\left(\prod_{m=1}^{n_i} \mu_m\right) \\
&= \int \mathbf{h}_1' \mathbf{Z}_{ij} R_{2i}(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b})\, d\left(\prod_{m=1}^{n_i} \mu_m\right) \\
&= \mathbf{h}_1' \mathbf{Z}_{ij} \prod_{m \le k} \{\lambda_0(0) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im}}\}
\end{aligned}
$$

If $j > k$, then

$$\int \mathbf{h}_1' \mathbf{Z}_{ij} \left(\delta_{ij} - \Lambda_0(X_{ij}) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij}} \right) R_{2i}(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b}) \, d \left(\prod_{m=1}^{n_i} \mu_m \right)$$

$$\begin{aligned}
=& \mathbf{h}_1' \mathbf{Z}_{ij} \prod_{m \le k} \{ \lambda_0(0) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im}} \} \\
& \times \sum_{\substack{\delta_{im} \in \{0,1\} \\ m>k, m \neq j}} \prod_{\substack{m>k \\ m \neq j}} \left(\exp[-\Lambda_0(\tau) \exp(\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im})] \right)^{1-\delta_{im}} \\
& \times \left(1 - \exp[-\Lambda_0(\tau) \exp(\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im})] \right)^{\delta_{im}} \\
& \times \sum_{\delta_{ij} \in \{0,1\}} (1 - \delta_{ij}) \left(-\Lambda_0(\tau) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij}} \right) \exp[-\Lambda_0(\tau) \exp(\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij})] \\
& + \delta_{ij} \int_{y=0}^{\tau} \left(1 - \Lambda_0(y) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + b' \mathbf{W}_{ij}} \right) \exp[\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij} - \Lambda_0(y) \exp(\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij})] \lambda_0(y) dy \\
=& \mathbf{h}_1' \mathbf{Z}_{ij} \prod_{m \le k} \{ \lambda_0(0) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im}} \} \\
& \times \sum_{\delta_{ij} \in \{0,1\}} (1 - \delta_{ij}) \left(-\Lambda_0(\tau) \exp[\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij} - \Lambda_0(\tau) \exp(\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij})] \right) \\
& + \delta_{ij} \Lambda_0(\tau) \exp[\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij} - \Lambda_0(\tau) \exp(\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij})] \\
=& 0
\end{aligned}$$

Therefore

$$\begin{aligned}
& \int \sum_{j=1}^{n_i} \int_{\mathbf{b}} \mathbf{h}_1' \mathbf{Z}_{ij} \left(\delta_{ij} - \Lambda_0(X_{ij}) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij}} \right) R_{2i}(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b}) \, d_{\mathbf{b}} N(\mathbf{0}, \boldsymbol{\Sigma}_0) \, d \left(\prod_{m=1}^{n_i} \mu_m \right) \\
& = \sum_{j \le k} \mathbf{h}_1' \mathbf{Z}_{ij} \int_{\mathbf{b}} \prod_{m \le k} \{ \lambda_0(0) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im}} \} \, d_{\mathbf{b}} N(\mathbf{0}, \boldsymbol{\Sigma}_0). \qquad (28)
\end{aligned}$$

Likewise, from the second term of (27):

$$\begin{aligned}
& \int \int_{\mathbf{b}} \left\{ \mathbf{b}' \boldsymbol{\Sigma}_0^{-1} \mathcal{D}(\mathbf{h}_2) \boldsymbol{\Sigma}_0^{-1} \mathbf{b}/2 - \boldsymbol{\Sigma}_0^{-1} \cdot \mathcal{D}(\mathbf{h}_2)/2 \right\} R_{2i}(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b}) \, d_{\mathbf{b}} N(\mathbf{0}, \boldsymbol{\Sigma}_0) \, d \left(\prod_{m=1}^{n_i} \mu_m \right) \\
& = \int_{\mathbf{b}} \left\{ \mathbf{b}' \boldsymbol{\Sigma}_0^{-1} \mathcal{D}(\mathbf{h}_2) \boldsymbol{\Sigma}_0^{-1} \mathbf{b}/2 - \boldsymbol{\Sigma}_0^{-1} \cdot \mathcal{D}(\mathbf{h}_2)/2 \right\} \prod_{m \le k} \{ \lambda_0(0) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im}} \} \, d_{\mathbf{b}} N(\mathbf{0}, \boldsymbol{\Sigma}_0). \\
& \qquad (29)
\end{aligned}$$

Furthermore, from the third term of (27), if $j \leq k$ then

$$\int \left(\delta_{ij} h_3(X_{ij}) - \int_0^{X_{ij}} h_3(s)\, d\Lambda_0(s) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij}} \right) R_{2i}(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b})\, d \left(\prod_{m=1}^{n_i} \mu_m \right)$$
$$= h_3(0) \prod_{m \leq k} \{ \lambda_0(0) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im}} \}; \qquad (30)$$

if $j > k$, then

$$\int \left(\delta_{ij} h_3(X_{ij}) - \int_0^{X_{ij}} h_3(s)\, d\Lambda_0(s) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij}} \right) R_{2i}(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b})\, d \left(\prod_{m=1}^{n_i} \mu_m \right)$$
$$= \prod_{m \leq k} \{ \lambda_0(0) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im}} \}$$
$$\times \sum_{\delta_{ij} \in \{0,1\}} \Big\{ -(1-\delta_{ij}) \int_0^{\tau} h_3(s)\, d\Lambda_0(s)$$
$$\times \exp[\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij} - \Lambda_0(t) \exp(\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij})]$$
$$+ \delta_{ij} \int_{y=0}^{\tau} \left(h_3(y) - \int_{s=0}^{y} h_3(s)\, d\Lambda_0(s) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij}} \right)$$
$$\times \exp[\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij} - \Lambda_0(t) \exp(\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij})] \lambda_0(y) dy \Big\}$$
$$= \prod_{m \leq k} \{ \lambda_0(0) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im}} \}$$
$$\times \sum_{\delta_{ij} \in \{0,1\}} \Big\{ -(1-\delta_{ij}) \int_0^{\tau} h_3(s)\, d\Lambda_0(s)$$
$$\times \exp[\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij} - \Lambda_0(t) \exp(\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij})]$$
$$+ \delta_{ij} \int_{s=0}^{\tau} h_3(s)\, d\Lambda_0(s) \exp[\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij} - \Lambda_0(t) \exp(\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij})] = 0.$$

Thus,

$$\int \sum_{j=1}^{n_i} \int_{\mathbf{b}} \left(\delta_{ij} h_3(X_{ij}) - \int_0^{X_{ij}} h_3(s)\, d\Lambda_0(s) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{ij} + \mathbf{b}' \mathbf{W}_{ij}} \right) R_{2i}(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b})\, d_{\mathbf{b}} N(\mathbf{0}, \boldsymbol{\Sigma}_0)\, d \left(\prod_{m=1}^{n_i} \mu_m \right)$$
$$= \sum_{j \leq k} h_3(0) \int_{\mathbf{b}} \prod_{m \leq k} \{ \lambda_0(0) e^{\boldsymbol{\beta}_0' \mathbf{Z}_{im} + \mathbf{b}' \mathbf{W}_{im}} \}\, d_{\mathbf{b}} N(\mathbf{0}, \boldsymbol{\Sigma}_0) \qquad (31)$$

Combining (28), (29), and (31) and integrating over $\mathbf{b}$, we obtain

$$\sum_{j=1}^{k} \mathbf{h}_1' \mathbf{Z}_{ij} + \frac{1}{2} \left(\sum_{j=1}^{k} \mathbf{W}_{ij} \right)' \mathcal{D}(\mathbf{h}_2) \left(\sum_{j=1}^{k} \mathbf{W}_{ij} \right) + k h_3(0) = 0.$$

Since the index set $j = 1, \ldots, k$ is arbitrary, we conclude

$$\sum_{j=k_1+1}^{k_2} \mathbf{h}_1' \mathbf{Z}_{ij} + \frac{1}{2} \left(\sum_{j=k_1+1}^{k_2} \mathbf{W}_{ij} \right)' \mathcal{D}(\mathbf{h}_2) \left(\sum_{j=k_1+1}^{k_2} \mathbf{W}_{ij} \right) + (k_2 - k_1) h_3(0) = 0.$$

for any $1 \leq k_1 < k_2 \leq n_i$. Therefore $\mathbf{W}_{ij}' \mathcal{D}(\mathbf{h}_2) \mathbf{W}_{ij'} = 0$ for $j \neq j'$ and $\mathbf{Z}_{ij}' \mathbf{h}_1 + \mathbf{W}_{ij}' \mathcal{D}(\mathbf{h}_2) \mathbf{W}_{ij}/2 + h_3(0) = 0$. Condition C3 yields $\mathcal{D}(\mathbf{h}_2) = \mathbf{0}$, and it follows that $\mathbf{h}_1 = \mathbf{0}$, $\mathbf{h}_2 = \mathbf{0}$, and $h_3(0) = 0$.

Next, we set $X_{ij} = 0$, $j = 2, \cdots, n_i$, and $\delta_{ij} = 1$, $j = 1, \ldots, n_i$ in (31) to get

$$h_3(X_{i1}) = \frac{\int_0^{X_{i1}} h_3(s)\, d\Lambda_0(s) \int_{\mathbf{b}} e^{\boldsymbol{\beta}_0' Z_{i1} + \mathbf{b}' \mathbf{W}_{i1}} R_{2i}(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b}) d_{\mathbf{b}} N(\mathbf{0}, \boldsymbol{\Sigma}_0)}{\int_{\mathbf{b}} R_{2i}(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b}) d_{\mathbf{b}} N(\mathbf{0}, \boldsymbol{\Sigma}_0)}.$$

So the expression $g(y) \equiv \int_0^y h_3(t)\, d\Lambda_0(t)$ satisfies the homogeneous equation

$$\frac{g'(y)}{\lambda_0(y)} - g(y) \frac{\int_{\mathbf{b}} e^{\boldsymbol{\beta}_0' Z_{i1} + \mathbf{b}' \mathbf{W}_{i1}} R_{2i}(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b}) d_{\mathbf{b}} N(\mathbf{0}, \boldsymbol{\Sigma}_0)}{\int_{\mathbf{b}} R_{2i}(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b}) d_{\mathbf{b}} N(\mathbf{0}, \boldsymbol{\Sigma}_0)} = 0$$

with boundary condition $g(0) = 0$. Therefore $g(y) = 0$, $h_3(y) = 0$, $\mathcal{Q}$ is one-to-one, and $\dot{S}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, \Lambda_0)$ is invertible.

Asymptotic normality follows from Theorem 2 of Murphy (1995) and the proof of asymptotic efficiency for $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{\Sigma}}_n$ is identical to Zeng *et al.* (2005).

PROOF OF THEOREM 3. The proof is analogous to the proof of Theorem 3 in Zeng *et al.* (2005).

PROOF OF THEOREM 4. The proof is similar to the proof of Theorem 1 in Self and Liang (1987), except that the Taylor series expansion cited in Lehmann (1983, pp.429-432) is now replaced by (6).

PROOF OF THEOREM 5. From Theorem 1 we have that $\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) = O_p(1)$. Applying (6) for the sequence $\boldsymbol{\phi}_n = \hat{\boldsymbol{\phi}}$, we get

$$\mathrm{pl}(\mathbf{y}^* | \hat{\boldsymbol{\phi}}(\mathbf{y})) = \mathrm{pl}(\mathbf{y}^* | \boldsymbol{\phi}_0) + s(\mathbf{y}^* | \boldsymbol{\phi}_0)' (\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) - n (\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)' \mathcal{I}_0 (\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)/2 + r_1, \quad (32)$$

where $r_1 = o_p(1)$. The main result (5) from Murphy and van der Vaart (2000) implies that $Es(\mathbf{y}^* | \boldsymbol{\phi}_0) = 0$ (divide by $\sqrt{n}$ and take limits on both sides of (5), and then apply the strong law of large numbers). Therefore, taking expectations on both sides of the equality in (32), the first-order term vanishes and we get

$$E_{f(\mathbf{y}^*)}\mathrm{pl}(\mathbf{y}^*|\hat{\boldsymbol{\phi}}(\mathbf{y})) = E\mathrm{pl}(\boldsymbol{\phi}_0) - n(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)'\mathcal{I}_0(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)/2 + Er_1. \tag{33}$$

Taking expectation one more time, with respect to $\mathbf{y}$ on both sides of (33), we have

$$\begin{aligned}\mathrm{pAI} &= -2E\mathrm{pl}(\mathbf{y}|\boldsymbol{\phi}_0) + E\{n(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)'\mathcal{I}_0(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)\} + Er_1 \\ &= -2E\mathrm{pl}(\mathbf{y}|\hat{\boldsymbol{\phi}}(\mathbf{y})) + 2E\{\mathrm{pl}(\mathbf{y}|\hat{\boldsymbol{\phi}}(\mathbf{y})) - \mathrm{pl}(\mathbf{y}|\boldsymbol{\phi}_0)\} + E\{n(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)'\mathcal{I}_0(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)\} + Er_1.\end{aligned}$$

¿From Corollary 2 and 1 of Murphy and van der Vaart (2000), the middle term and the last term under expectation signs in the last equation above have a χ^2_p distribution, except for remainder terms of $o_p(1)$. Collecting all the remainder terms in $r_2 = o_p(1)$, we get that

$$\mathrm{pAI} = -2E\mathrm{pl}(\mathbf{y}|\hat{\boldsymbol{\phi}}(\mathbf{y})) + 2p + Er$$

which proves the theorem. If r is uniformly integrable, then $E(r) = o(1)$, and pAIC is asymptotically unbiased for pAI.

PROOF OF PROPOSITION 1. The consistency part is immediate by applying the strong law of large numbers to A and A_i.

To show the variance inequality, note that $A = \sum_k \exp\{\sum_i v(\mathbf{b}_i^{(k)})\}/M$. Assume for simplicity that $n = 2$ (the general case follows by induction). Put $\exp\{v(\mathbf{b}_i)^{(k)}\} = \xi_i^{(k)}$, for $i = 1, 2$. Then $A = \overline{\xi_1\xi_2}$, and $\tilde{A} = \bar{\xi_1}\bar{\xi_2}$, where the bar denotes sample average of M observations. Let μ_i, σ_i^2 denote respectively the mean and variance of ξ_i, i=1,2. Then

$$\begin{aligned}\mathrm{Var}(\overline{\xi_1\xi_2}) &= \mathrm{Var}(\xi_1\xi_2)/M \\ &= \sigma_1^2\sigma_2^2/M + \mu_1^2\sigma_2^2/M + \mu_2^2\sigma_1^2/M \\ \mathrm{Var}(\bar{\xi_1}\bar{\xi_2}) &= (\sigma_1^2/M)(\sigma_2^2/M) + \mu_1^2\sigma_2^2/M + \mu_2^2\sigma_1^2/M\end{aligned}$$

The first term in $\mathrm{Var}(\overline{\xi_1\xi_2})$ is no smaller than the corresponding term in $\mathrm{Var}(\bar{\xi_1}\bar{\xi_2})$, while the other two terms are identical, so the result follows.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Breakthroughs in statistics (1992)*, vol. 1, 610–624. Springer-Verlag.

Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics* **10**, 1100–20.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information - Theoretic Approach.* Springer, 2nd edn.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics* **25**, 573–578.

Claeskens, G. and Hjort, N. L. (2003). Focused information criterion (with discussion). *Journal of the American Statistical Association* **98**, 900–945.

Commenges, D. and Andersen, P. (1995). Score test of homogeneity for survival data. *Lifetime Data Analysis* **1**, 145–156.

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

Crainiceanu, C. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B* **66**, 165–185.

deLeeuw, J. (1992). Introduction to Akaike (1973) 'Information theory and an extension of the maximum likelihood principle'. In *Breakthroughs in Statistics*, vol. 1, 599–609. New York: Springer.

DiCiccio, T. J., Kass, R. E., Raftery, A., and Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association* **92**, 903–915.

Fan, J. H. and Wong, W. H. (2000). Discussion of 'On profile likelihood', by Murphy, S. A. and van der Vaart, A. W. *Journal of the American Statistical Association* **95**, 468–471.

Gelfand, A. and Day, D. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B* **56**, 501–514.

Glidden, D. and Vittinghoff, E. (2004). Modelling clustered survival data from multicenter clinical trials. *Statistics in Medicine* **23**, 369–388.

Gray, R. (1995). Tests for variation over groups in survival data. *Journal of the American Statistical Association* **90**, 198–203.

Lehmann, E. L. (1983). *Theory of Point Estimation.* John Wiley, New York.

Li, B. (2000). Comment on 'on profile likelihood'. *Journal of the American Statistical Association* **95**, 472–474.

Linhart, H. and Zucchini, W. (1986). *Model Selection.* Wiley, New York.

Liu, I., Blacker, D., Xu, R., Fitzmaurice, G., Lyons, M., and Tsuang, M. (2004a). Genetic and environmental contributions to the development of alcohol dependence in male twins. *Archives of General Psychiatry* **61**, 897–903.

Liu, I., Blacker, D., Xu, R., Fitzmaurice, G., Tsuang, M., and Lyons, M. J. (2004b). Genetic and environmental contributions to age of onset of alcohol dependence symptoms in male twins. *Addiction* **99**, 1403–1409.

Longford, N. T. (2005). Model selection and efficiency – is 'Which model. . . ?' the right question? *Journal of the Royal Statistical Association, Series A* **168**, 469–472.

Maple, J., Murphy, S., and Axinn, W. (2002). Two-level proportional hazards models. *Biometrics* **58**, 754–763.

Meng, X.-L. and Wong, W. (1996). Simulating ratios of normalizing constants via a simple identity. *Statistica Sinica* **6**, 831–860.

Murphy, S. (1994). Consistency in a proportional hazards model incorporating a random effect. *Annals of Statistics* **22**, 2, 712–731.

Murphy, S. (1995). Asymptotic theory for the frailty model. *Annals of Statistics* **23**, 1, 182–198.

Murphy, S. and van der Vaart, A. (2000). On profile likelihood. *Journal of the American Statistical Association* **95**, 449–485.

Murray, D., Varnell, S., and Blitstein, J. (2004). Design and analysis of group-randomized trials: a review of recent methodological developments. *American Journal of Public Health* **94**, 423–432.

Parner, E. (1998). Asymptotic theory for the correlated Gamma-frailty model. *Annals of Statistics* **26**, 1, 183–214.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* **56**, 1016–1022.

Rudin, W. (1973). *Functional Analysis*. New York: McGraw-Hill.

Self, S. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**, 398, 605–610.

Severini, T. A. and Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *Annals of Statistics* **20**, 1768–1802.

Stram, D. and Lee, J. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**, 1171–1177.

Stram, D. and Lee, J. (1995). Correction to "Variance component testing in the longitudinal mixed effects model". *Biometrics* **51**, 1196.

Sylvester, R., van Glabbeke, M., Collette, L., Suciu, S., Baron, B., Legrand, C., Gorlia, T., Collins, G., Coens, C., Declerck, L., and Therasse, P. (2002). Statistical methodology of phase iii cancer clinical trials: advances and future perspectives. *European Journal of Cancer* **38**, S162–S168.

Therneau, T. and Grambsch, P. (2000). *Modelling Survival Data: Extending the Cox Model*. Springer Verlag, New York, USA.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82–86.

Vaida, F. and Blanchard, S. (2005). Conditional akaike information for mixed-effects models. *Biometrika* **92**, 351–370.

Vaida, F. and Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine* **19**, 3309–3324.

Verweij, P. and van Houwelingen, H. (1995). Time-dependent effects of fixed covariates in Cox regression. *Biometrics* **51**, 1550–56.

Zeng, D., Lin, D. Y., and Yin, G. (2005). Maximum likelihood estimation for proportional odds model with random effects. *Journal of the American Statistical Association* **100**, 470–483.

Table 1: $-2\times$ *Log likelihood values from the lung cancer data*

Model	Laplace	*RIS*	*BS*	pAIC^{+}
0*	7241.76	7241.76	7241.76	7249.76
1*	7232.80 (8.96)	7232.80	7232.80	7242.80
2	7228.98 (3.82)	7228.80 (4.00)	7228.78 (4.02)	7240.80
3	7222.72 (6.26)	7222.55 (6.25)	7222.60 (6.18)	7236.55

RIS - reciprocal importance sampling, *BS* - bridge sampling.

$^{+}$ computed using *RIS*.

* likelihood computed directly when there are no random effects.

In $(\cdot)$ are the likelihood ratio statistics between the model and its immediate submodel (3 vs. 2, 2 vs. 1, etc.).

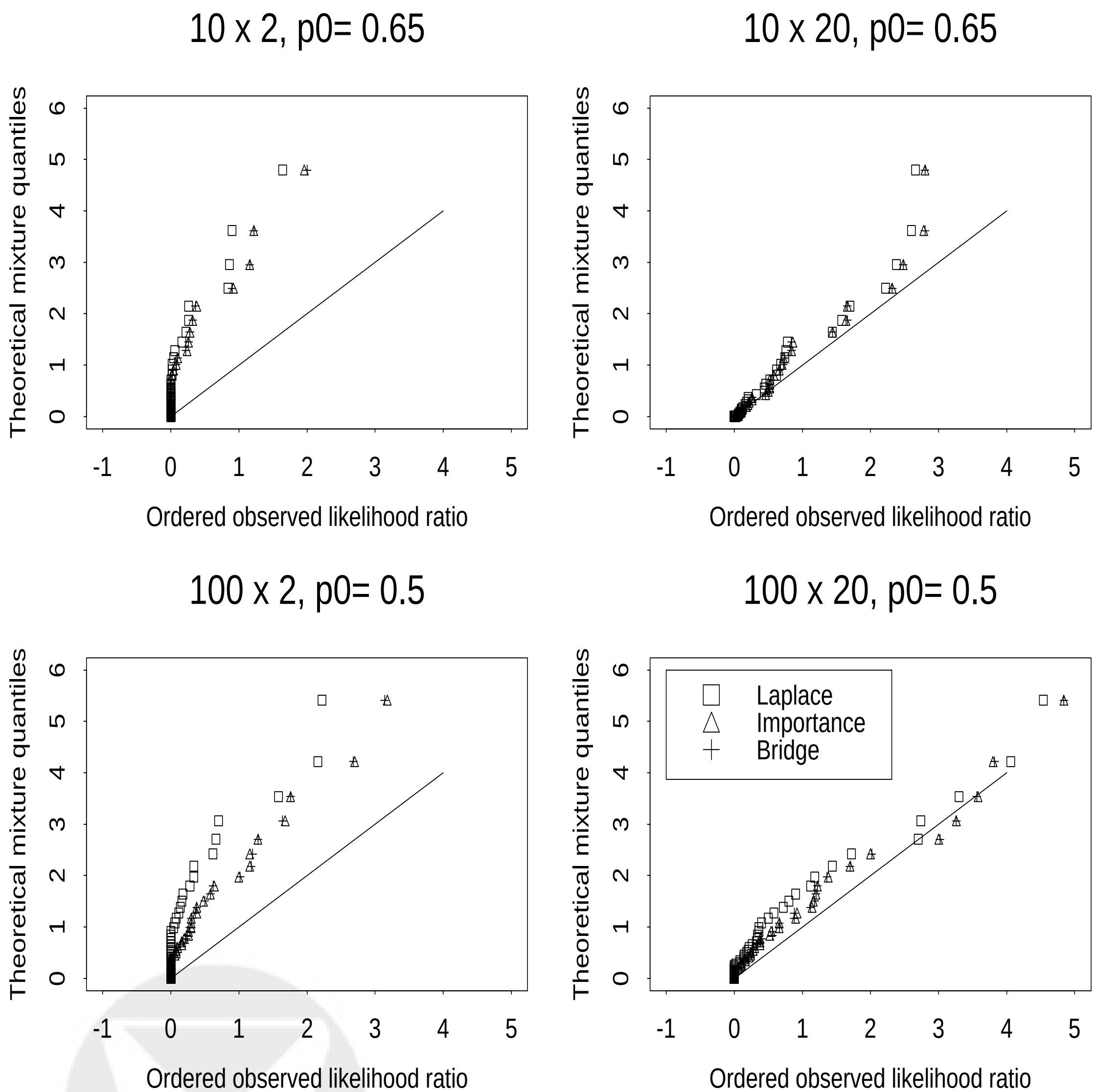

Figure 1: *Q-Q plots of likelihood ratio statistics from simulated data*