Statistica Sinica **19** (2009), 819-842

USING PROFILE LIKELIHOOD FOR SEMIPARAMETRIC MODEL SELECTION WITH APPLICATION TO PROPORTIONAL HAZARDS MIXED MODELS

Ronghui Xu[1], Florin Vaida[1] and David P. Harrington[2]

[1] *University of California, San Diego and*
[2] *Dana-Farber Cancer Institute and Harvard School of Public Health*

Abstract: We consider selection of nested and non-nested semiparametric models. Using profile likelihood we can define both a likelihood ratio statistic and an Akaike information for 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. Our work was motivated by the proportional hazards mixed effects model (PHMM), which incorporates general random effects of arbitrary covariates and includes the frailty model as a special case. The asymptotic properties of its parameter estimate has recently been established, which enables the quadratic expansion of the log profile likelihood. For computation of the (profile) 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 and phrases: Akaike information, bridge sampling, Laplace approximation, likelihood ratio test, profile likelihood, reciprocal importance sampling, testing on the boundary.

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, developed under the more general models with nuisance parameters, was 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, Varnell, and Blitstein (2004) and Sylvester, van Glabbeke, Collette, Suciu, Baron, Legrand, Gorlia, Collins, Coens, Declerck and Therasse (2002). Liu, Blacker, Xu, Fitzmaurice, Lyons and Tsuang (2004a,b), 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 treatment effects in a multi-center clinical trial that depend on the trial centers (see Section 7). The model is of the form

$$\lambda_{ij}(t) = \lambda_0(t) \exp(\beta' Z_{ij} + b_i' W_{ij}), \tag{1.1}$$

where $\lambda_{ij}(t)$ is the hazard function of the j-th observation from the i-th cluster, b_i is a vector of random effects for the i-th cluster, Z_{ij} and W_{ij} are the covariate vectors for the fixed and random effects, and $\lambda_0(t)$ is the baseline hazard function. 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 of the parameters in this model, computed using the *EM* algorithm and Markov Chain Monte Carlo (*MCMC*) methods. Xu, Gamst, Donohue, Vaida and Harrington (2006) established the asymptotic properties of the nonparametric maximum likelihood estimator under the model.

As in any regression settings, model selection is an important aspect of data analysis. In particular, in the applications of model (1.1), 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. In this paper we develop a likelihood ratio test that allows arbitrary testing on the mixed model, so that 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. The

Akaike information criterion (*AIC*; Akaike (1973), deLeeuw (1992) and Burnham and Anderson (2002)) is the most commonly used in practice. It has a simple interpretation as a penalized log-likelihood, and it has 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 (1.1). During the revision of this paper, a referee alerted us to the forthcoming paper of Claeskens and Carroll (2007) which also uses a profile AIC in the semiparametric context.

For model selection concerning right-censored failure time data, Cai, Fan, Li and Zhou (2005) provided a rather complete review; see references therein. Selection methods for modeling multivariate failure time data are still underdeveloped. Fan and Li (2002) applied non-concave penalized likelihood to Cox model with frailties. Cai (1999) developed generalized likelihood ratio to test marginal models in multivariate failure time data, and Cai et al. (2005) proposed a penalized pseudo-partial likelihood method for marginal models in multivariate failure time data.

In this paper we consider both the likelihood ratio test and an Akaike information criterion (*AIC*) for model selection in the presence of nuisance parameters. They turn out to be derived from the same asymptotic expansion of a log profile likelihood. They also share the same computational algorithm. In the next section we review the proportional hazards mixed model and the profile likelihood function. We consider the profile likelihood ratio test in Section 3, including testing on the boundary. In Section 4 we develop an *AIC* using the profile likelihood. We consider three algorithms to compute the maximized (profile) likelihood under PHMM in Section 5. Simulation studies are carried out in Section 6, and an application is shown in Section 7 to illustrate the methods. Section 8 contains some further discussion.

2. PHMM and Profile Likelihood

2.1. 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 b_i, the survival times T_{ij} are independent and follow the proportional hazards model (1.1). In (1.1), W_{ij} is often a subset of Z_{ij} apart from possibly a '1' which represents the cluster effect on the baseline hazard. To insure identifiability, we assume that $E(b_i) = 0$; for distribution of the random effects we assume that

$$b_i \overset{iid}{\sim} N(0, \Sigma),$$

as in Vaida and Xu (2000). The multinormal distribution has attractive statistical properties: it is symmetric, scale invariant, and consistent with the usual setup in other typical mixed-effects scenarios (linear, non-linear, and generalized linear mixed models). In contrast, the commonly used gamma distribution, although computationally attractive, is not scale-invariant. This means that it is not suitable as a distribution for the random effects of arbitrary (continuous) covariates: if e^b is gamma-distributed, e^{bW} no longer belongs to the gamma family in general.

The data from subject j in cluster i can be written $y_{ij} = (X_{ij}, \delta_{ij}, Z_{ij}, W_{ij})$, where X_{ij} is the possibly right-censored failure time and δ_{ij} is the failure-event indicator. Let $y_i = (y_{i1}, \ldots, y_{in_i})$ be the data for cluster i. Conditional on the random effects, the observations from the same cluster are assumed to be independent. The clusters are assumed to be *i.i.d.* (Xu et al. (2006)).

For cluster i, conditional on the random effect b_i, the log-likelihood is

$$l_i = l_i(\beta, \lambda_0; y_i|b_i) = \sum_{j=1}^{n_i} \Big\{ \delta_{ij} \log \lambda_0(X_{ij}) + \delta_{ij}(\beta' Z_{ij} + b_i' W_{ij}) - \Lambda_0(X_{ij}) e^{\beta' Z_{ij} + b_i' W_{ij}} \Big\},$$

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 use. The likelihood of the observed data is then

$$L(\theta) = \prod_{i=1}^{n} \int \exp(l_i) p(b_i|\Sigma) db_i, \tag{2.1}$$

where $\theta = (\beta, \Sigma, \lambda)$ and $p(\cdot)$ is the distribution of the random effects, assumed multinormal. Usually no closed-form expression is available for $L(\theta)$ and its calculation involves d-dimensional integration.

2.2. Profile likelihood

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, $y_1, \ldots, y_n$, from a distribution depending on parameters ϕ and λ. We assume that $\phi \in \Phi$, a subset of $\mathbf{R}^p$, and λ is a nuisance parameter, possibly of infinite dimension. The log-likelihood of the data is $l(\phi, \lambda) = \sum_{i=1}^n l_i(\phi, \lambda)$, and l_i is the log-likelihood for y_i. The log profile likelihood function for ϕ, with the nuisance parameter λ 'profiled out', is

$$\mathrm{pl}(\phi) = \sup_{\lambda} l(\phi, \lambda). \tag{2.2}$$

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 ϕ_n such that $\|\phi_n - \phi_0\| = O_p(1/\sqrt{n})$, where ϕ_0 is the true parameter value,

$$\frac{1}{n}\Big\{\mathrm{pl}(\phi_n) - \mathrm{pl}(\phi_0)\Big\} = (\phi_n - \phi_0)'A - \frac{1}{2}(\phi_n - \phi_0)'I(\phi_n - \phi_0) + o_p\Big(\frac{1}{n}\Big). \quad (2.3)$$

Here $A = \sum_1^n s(y_i)/n$, s is the efficient score for ϕ, 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 I, its covariance matrix, is the efficient Fisher information matrix (Murphy and van der Vaart (2000) and Severini and Wong (1992)). We derive the results of this paper for models that satisfy (2.3). Note that (2.3) is used for theoretical purposes only: the computation of the likelihood ratio test is discussed separately in Section 5.

Under PHMM the parameter of interest is $\phi = (\beta, \Sigma)$, whereas the baseline hazard λ is seen as a nuisance parameter. Asymptotic normality, recently established in Xu et al. (2006), of the nonparametric maximum likelihood estimator implies that the likelihood surface is asymptotically quadratic near the true parameter value, which in turn implies that the same holds for the profile likelihood (Murphy and van der Vaart (2000) and Li (2000)). The asymptotic properties of the maximum likelihood estimate have also been established for the gamma frailty models (Murphy (1994, 1995) and Parner (1998)). Maple, Murphy and Axinn (2002) verified empirically that the contours of the profile likelihood under the multinormal PHMM are elliptic.

3. Semiparametric 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 distribution with the number of degrees of freedom equal to the difference of those of the two models. For a semiparametric model such as (1.1), the number of degrees of freedom of the model itself is not well defined, since there is at least one infinite dimensional parameter. Also the maximum likelihood ratio statistic in general may not exist in nonparametric and semiparametric setting (Fan, Zhang and Zhang (2001) and Fan and Huang (2005)). 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 restrictions on the nuisance

parameter λ. Write L for the likelihood, and let

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

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

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

where Φ_0 and Φ are the corresponding parameter spaces for ϕ under the two models. Murphy and van der Vaart (2000) showed that as result of the quadratic expansion (2.3), when ϕ_0 lies in the interior of the parameter space, the maximum likelihood estimator of ϕ is asymptotically normal, and the profile likelihood ratio test for $H_0 : \phi = \phi_0$ has asymptotically a chi-squared distribution with degrees of freedom equal to the dimension of ϕ under the null hypothesis H_0. We note that for a class of varying-coefficient partially linear models, Fan and Huang (2005) developed a profile likelihood ratio test using a profile least-squares estimator which is not the maximum likelihood estimator.

3.1. Testing on the boundary

The challenging problem in hypothesis testing under model (1.1) 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 (2.3) enables us to obtain results on the null distribution of the profile likelihood ratio statistic similar to those in Self and Liang (1987) on maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. We note that Vu and Zhou (1997) generalized Self and Liang (1987) results to nonidentically distributed random variables and a class of estimating functions.

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 ϕ_0 is on the boundary of Φ, given that the same holds when ϕ_0 lies in the interior of Φ.

Theorem 1. *Given the quadratic expansion* (2.3), *with probability tending to 1 as $n \to \infty$, there exists a sequence of points in Φ, $\hat{\phi}_n$, at which local maxima of $pl_n(\phi)$ occur, that converges to ϕ_0 in probability. Moreover, $\sqrt{n}(\hat{\phi}_n - \phi_0) = O_p(1)$.*

See the Appendix for a proof. Note that the proof only requires (2.3) to hold in the interior of the parameter space.

Expression (2.3) is equal to

$$\frac{1}{2}A'I^{-1}A - \frac{1}{2}\{z_n - (\phi_n - \phi_0)\}'I\{z_n - (\phi_n - \phi_0)\} + o_p\left(\frac{1}{n}\right),$$

where $z_n = I^{-1}A$. Therefore the same representation of the asymptotic distribution of $-2\log LR$ as those of 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 ϕ_0 (for definitions, see Self and Liang (1987) or Chernoff (1954)).

Theorem 2. *Let Z be a random variable with a multivariate Gaussian distribution with mean ϕ and covariance matrix $I^{-1}(\phi_0)$, and let C_Φ and C_{Φ_0} be non-empty cones approximating Φ and Φ_0 at ϕ_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 $\phi \in C_{\Phi_0}$ versus $\phi \in C_\Phi$ based on a single realization of Z when $\phi = \phi_0$.*

3.2. Likelihood ratio test under PHMM

The above representation only involves the finite dimensional parameter ϕ 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 models, the results are exactly the same.

In the following we list the cases which are the most likely to be encountered in practice. Recall that d is the dimension of b.

Case 1. $d = q + 1$ and

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

where Σ_{11} is $q \times q$ and $q \geq 0$. The asymptotic null distribution of $-2\log LR$ for testing $H_0 : \sigma_{22} = 0$ against Σ positive semidefinite is $(\chi^2_q + \chi^2_{q+1})/2$. When $q = 0$, the distribution is a 50:50 mixture of a point mass at 0 and χ^2_1; 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).

Case 2. Same as in *Case 1*, but the test also includes a r-dimensional subvector of fixed effects, β_1, i.e., $H_0 : \sigma_{22} = 0$, $\sigma_{12} = 0$, $\beta_1 = 0$ against Σ positive semidefinite and $\beta_1 \neq 0$. The asymptotic distribution of $-2\log LR$ is $(\chi_{q+r} + \chi_{q+r+1})/2$.

Case 3. $d = q + k$ and

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

where Σ_{11} is $q \times q$ and Σ_{22} is $k \times k$. The asymptotic null distribution of $-2 \log LR$ for testing $H_0 : \Sigma_{22} = 0$ against Σ 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, 1995). The minimum number of degrees of freedom of Stram and Lee (1994, 1995) is, however, in error. To see why the correct mixture is the one we stated above, write $\Sigma = \mathrm{diag}(\sigma) R \,\mathrm{diag}(\sigma)$, where σ is the vector of standard deviations and $R = (\rho_{ij})$ is the correlation matrix. Testing $\Sigma_{22} = 0$ is equivalent to testing $\sigma_{q+1} = \cdots = \sigma_{q+k} = 0$. The result then follows along the lines of 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 $\beta_1 = 0$ 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 model of interest has $\Sigma_{12} = 0$ and Σ_{22} is diagonal. Similarly to Case 3, the asymptotic null distribution for testing $\Sigma_{22} = 0$ 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 infinity. 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. See also Greven, Crainiceanu, Peters and Kuechenhoff (2007). We further discuss this issue in the simulation section.

4. Profile Akaike Information

In this section 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*. The development parallels that of the standard AIC.

Consider a family of models $\mathcal{M}$ parameterized by $\theta = (\phi, \lambda)$, where $\phi \in \Phi$ is again the parameter of interest, and $\lambda \in \mathbf{\Lambda}$ is the nuisance parameter. The view we take here, similar to Claeskens and Hjort (2003), is that we are interested in selecting the 'ϕ part' of the modeling, while leaving the parameter space $\mathbf{\Lambda}$ the same across all competing models. In this way, for model selection purposes, $\mathcal{M}$ is really indexed by ϕ alone. Assume that the data vector y, consisting of n independent observations $y_1, \ldots, y_n$, is generated by a distribution with density f.

The classical 'distance' from the true distribution f to a member $g_\theta = g(\cdot|\phi,\lambda)$ of $\mathcal{M}$ is given by the Kullback-Leibler information (KL), $\mathrm{KL}(f, g_\theta) = E_f\{\log f(y) - \log g_\theta(y)\}$. When the focus is on ϕ alone, the relevant distance is that between f and the subfamily of models $\{g_{\phi,\lambda} : \lambda \in \mathbf{\Lambda}\} : \min_{\lambda\in\mathbf{\Lambda}} \mathrm{KL}(f, g_{\phi,\lambda})$. Suppose that the minimum is attained at some $\lambda = \tilde{\lambda}(\phi)$ for each ϕ. Following Severini and Wong (1992), $\tilde{\lambda}(\phi)$ is in fact a least favorable curve under smoothness conditions (see also Fan and Wong (2000)). We write $g_\phi = g(\cdot|\phi, \tilde{\lambda}(\phi))$. Ignoring the constant term $E\{\log f(y)\}$ in $\mathrm{KL}(f,\cdot)$, we have that

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

the expectations here and in the rest of this section are with respect to the true distribution f. Therefore g_ϕ is the theoretical equivalent of the profile likelihood.

Minimum KL is attained at ϕ_0 such that $\mathrm{KL}(f, g_{\phi_0}) = \min_\phi \mathrm{KL}(f, g_\phi)$ or, equivalently,

$$E\{\log g_{\phi_0}(y)\} = \max_\phi E\{\log g_\phi(y)\}.$$

Then g_{ϕ_0} is the best approximation to f within the family of models $\mathcal{M}$. When the model is correct, we have $f = g_{\phi_0}$. In practice ϕ_0 is estimated by $\hat{\phi}(y)$ which maximizes the profile likelihood:

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

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

$$\mathrm{pAI} = -2E_{f(y)}E_{f(y^*)}\{\mathrm{pl}(y^*|\hat{\phi}(y))\}. \tag{4.1}$$

It is important to note that $\mathrm{pl}(y^*|\hat{\phi}(y))$ in (4.1) is different from the log-likelihood function computed at the MLE $(\hat{\phi},\hat{\lambda})$, since it allows maximizing the likelihood over λ based on the new data y^*.

It is well-known that the 'apparent' estimate $-2\mathrm{pl}(y|\hat{\phi}(y))$ is biased for pAI. The following result shows that a profile AIC provides an approximately unbiased estimator, where the number in the bias correction term is p, the dimension of ϕ.

Theorem 3. *Assume that* (2.3) *holds and that* $f = g(.|\theta_0)$*, with* θ_0 *in the interior of the parameter space. Further, assume that* y*,* y^* *consist of* n *i.i.d. vectors, and* $\hat{\phi}$ *is consistent for* ϕ_0*. Then*

$$\mathrm{pAIC} = -2\mathrm{pl}(y|\hat{\phi}(y)) + 2p \tag{4.2}$$

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 the Appendix for a proof.

Note that 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 y^* the profile likelihood function at $\hat{\phi}(y)$ is not the same as the likelihood function at the *MLE* based on data y. However, the observed profile likelihood in (4.2) 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 $\theta = (\phi, \lambda)$ is $-2l(\hat{\theta}) + 2(p+q)$, while the profile *AIC* for ϕ is $-2l(\hat{\theta}) + 2p$.

As with the standard AIC, the model with the smaller pAIC is preferred. Note however that pAIC (as well as the AIC) is subject to statistical error, and the ranking of pAIC's may not reflect the ranking of the pAI's for a set of models under consideration. In practice, a difference of greater than 2 is considered evidence in favor of the model with the smaller AIC (Burnham and Anderson (2002)). For a rigorous theory on the difference in AIC's, see Vuong (1989).

4.1. Profile AIC for PHMM

The PHMM was our original motivation for developing the profile *AIC*. When the focus is on the fixed effects β and the variance components Σ, the pAIC is given by (4.2), where p counts the number of parameters in β and Σ. Computation of the likelihood term in (4.2) is addressed in the next section.

As a special case, when there are only fixed effects in the proportional hazards model, the profile *AIC* is also given by (4.2), where p is the dimension of the regression parameter β. The profile likelihood in this case is equivalent to the partial likelihood (Cox (1975) and 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 (2.3) in this case. The validity of this *AIC* as an unbiased estimate of an Akaike 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 $\hat{\beta}$ (Andersen and Gill (1982)).

5. Computing the Likelihood under PHMM

For the PHMM we computed $\hat{\theta}$ using an *EM*-type algorithm (Vaida and Xu (2000)). Briefly, at the *E*-step the expectation of the full (not partial) log-likelihood, conditional on the random effects, is computed via Monte Carlo simulation using a Gibbs sampler. Then the *M*-step maximization proceeds as in the standard Cox model with an offset, and the baseline hazard function is replaced by Breslow's estimator. The standard errors of the parameters are computed via Louis' formula. Several alternative fitting algorithms are discussed in Cortiñas-Abrahantes, Legrand, Burzykowski, Janssen, Ducrocq, and Duchateau (2007) and the references therein.

To compute the likelihood ratio statistic and the pAIC, only the maximum of the full likelihood function given in (2.1) is needed, since $\mathrm{pl}(\hat{\phi}) = \log L(\hat{\theta})$. The likelihood function (2.1) 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 a computational price. We compare the performance of the three methods in simulations and data analysis.

In the following we write $b = (b_1', \ldots, b_n')'$ and $y = (y_1', \ldots, y_n')'$.

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 (2.1), $p(y_i)p(b_i|y_i)$, and is justified asymptotically, as $n_i \to \infty$. The approximation for cluster i is given by

$$l_L^{(i)} = \left(\frac{d}{2}\right)\log(2\pi) + \left(\frac{1}{2}\right)\log|\hat{V}_i| + \log p(y_i|\hat{b}_i, \hat{\theta}) + \log p(\hat{b}_i|\hat{\Sigma}), \tag{5.1}$$

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

$$l_L = \sum_{i=1}^{n} l_L^{(i)} = \left(\frac{nd}{2}\right)\log(2\pi) + \left(\frac{1}{2}\right)\log|\hat{V}| + \log p(y|\hat{b}, \hat{\theta}) + \log p(\hat{b}|\hat{\Sigma}),$$

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

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

$$l_R = l_L - \log A, \tag{5.2}$$

where

$$A = \frac{1}{M} \sum_{k=1}^{M} \exp\{v(b^{(k)})\}, \tag{5.3}$$

$$v(b) = l_L + \log p_0(b) - \log p(y, b|\hat{\theta}). \tag{5.4}$$

For numerical accuracy, the computations are done on the logarithmic scale as in (5.4). Theoretically, l_L can be omitted in (5.4), in which case $l_R = -\log A$. However, using the Laplace approximation l_L as a "point of reference" in (5.4) greatly improves the numerical accuracy of l_R. A simple probabilistic argument shows that indeed A in (5.3) 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* (5.2) *over the whole dataset, and $\tilde{l}_R = \sum_{i=1}^{n} l_R^{(i)}$, where $l_R^{(i)} = l_L^{(i)} - \log A_i$, $l_L^{(i)}$ is given by* (5.1), *and $A_i = \sum_k \exp\{v(b_i^{(k)})\}/M$. 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 the Appendix for a proof.

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

$$l_B = \log(B) - \log(C) + l_L,$$

where

$$B = \frac{1}{M_0}\sum_{k=1}^{M_0}\left[1+\exp\{v(u^{(k)})\}\right]^{-1}, \tag{5.5}$$

$$C = \frac{1}{M}\sum_{k=1}^{M}\left[1+\exp\{-v(b^{(k)})\}\right]^{-1}. \tag{5.6}$$

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

6. Simulation Experiments

In this section we report on simulations to compare the accuracy of the three methods described in Section 5 for calculating the likelihood values, and we study the finite sample distribution of the likelihood ratio statistic.

We first simulated data under model (1.1) with a single binary covariate Z, $\beta = 1.5$, $\lambda_0(t) = 1$, and no random effects. The censoring distribution was Uniform $(0, \tau)$, where τ was chosen to achieve about 15% censoring. We then fit model (1.1) 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 ($n \times n_i$) were used; for each case 100 simulations were carried out. 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. As suggested by a referee, we also compared the three computational methods with a direct numerical solution, namely, Gaussian-Hermite quadrature. In all cases, quadrature gave overlaying plots on top of *RIS* and *BS*; for illustration purposes we only plot them for the case of 10×20 in Figure 1. For the number of observations per cluster $n_i = 20$ Laplace approximation gave similar results to *RIS* and *BS*. For $n_i = 2$, however, there were discrepancies between Laplace approximation and *RIS* or *BS*. The discrepancies increased 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 we plot the ordered likelihood ratio statistics from 100 simulations versus 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 $n = 100$ (second row) we compare the empirical distribution of the likelihood ratio statistic with its asymptotic distribution given in Case 1 of Section 3.2, i.e., a 50:50 mixture of point mass at zero and χ^2_1. In Figure 1 'p0'

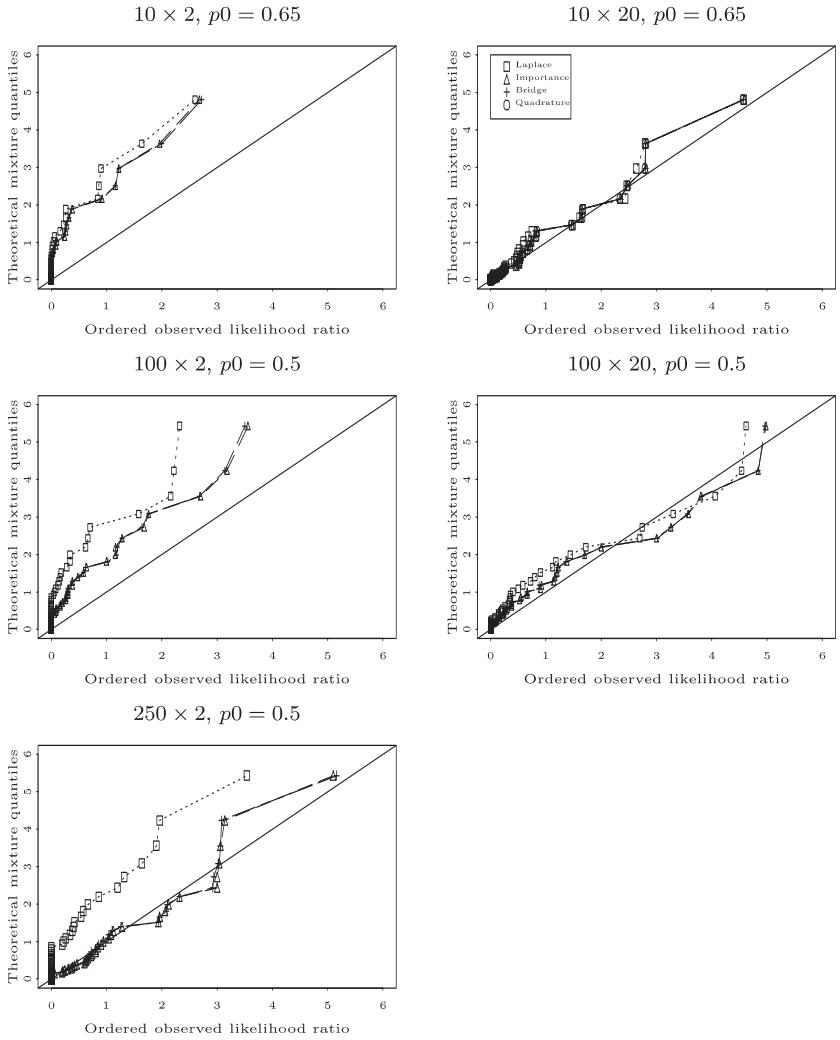

Figure 1. Q-Q plots of likelihood ratio statistics from simulated data for testing H_0: $\lambda_{ij}(t) = \exp(\beta Z_{ij})$ versus $H_1 : \lambda_{ij}(t) = \exp(\beta Z_{ij} + b_i)$. The theoretical mixture distribution is $p_0 : (1-p_0)$ mixture of point mass at zero and χ^2_1.

denotes the probability of point mass at zero. For $n = 10$ (first row) 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^2_1. 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

Table 1. Error rates of nominal 0.05-level likelihood ratio test and AIC.

	Sample size	Likelihood ratio			pAIC$^+$	
		Laplace	*RIS*	*BS*	Smallest	Difference $> 2^1$
$d = 1$	$10 \times 2^*$	1%	1%	1%	1%	0%
	100×2	0%	2%	2%	3%	0%
	250×2	1%	6%	6%	9%	1%
	$10 \times 20^*$	5%	5%	5%	5%	1%
	100×20	5%	6%	6%	6%	2%
$d = 2$	20×5	3%	3%	3%	6%	3%
	20×20	4%	4%	4%	5%	4%
	100×5	2%	3%	3%	4%	2%
	100×20	3%	4%	5%	8%	3%

$d = 1$ corresponds to the simulation setup in Figure 1, $d = 2$ corresponds to the setup in Figure 2.

* the reference distribution is the 65:35 mixture of point mass at zero and χ^2_1.

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

$^+$ computed using *RIS*.

[1] Choose the smaller model unless the difference in pAIC's > 2.

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 $n_i = 2$ the asymptotic 50:50 mixture is not a good approximation unless n is as large as 250 (bottom plot, computed using *RIS* and *BS*). As mentioned before, for $n = 10$ the 65:35 mixture of Crainiceanu and Ruppert's requires that n_i be reasonably large, and clearly $n_i = 2$ is not sufficient.

In Table 1 we give the empirical significance levels of the nominal 0.05-level profile likelihood rato test as well as the error rates of model selection by the profile *AIC*, here the true model being the null model. When compared to Figure 1, it is seen that the nominal siginficance level is better achieved when the theoretical distribution has reasonable agreement with the empirical distribution of the likelihood ratio statistic, that is, for the sample sizes 250×2 (*RIS* and *BS* only), 10×20 and 100×20. Otherwise the test tends to be conservative. Note that choosing the model with the smallest pAIC here is equivalent to a likelihood ratio statistic greater than two time the difference in the numbers of parameters under the two models. We also considered the method of choosing the smaller model unless the difference in the pAIC's is at least 2, and this is equivalent to a likelihood ratio statistic greater than two time the difference in the numbers of parameters plus 2. It is clear then the effect of different computational methods on the likelihood ratio test carries over to the pAIC, and the second approach to using pAIC gave lower error rates than the first one in this case.

Finally, we simulated data with two covariates, and test against the null hypothesis that both the fixed and the random effects for one of the covariates

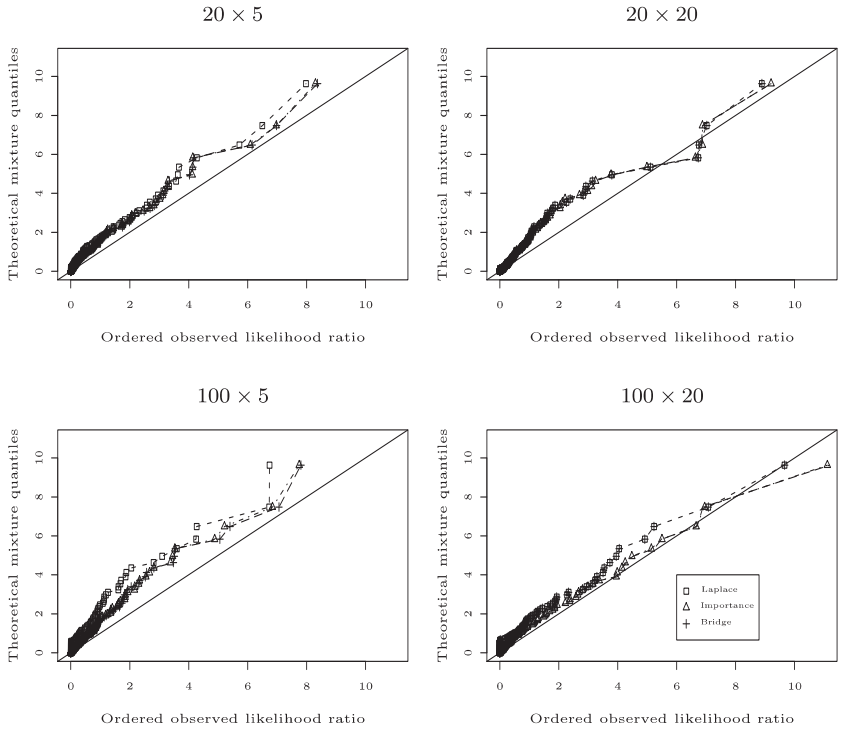

Figure 2. Q-Q plots of likelihood ratio statistics from simulated data for testing $H_0 : \lambda_{ij}(t) = \exp(\beta_1 Z_{ij1} + b_{i0})$ versus $H_1 : \lambda_{ij}(t) = \exp(\beta_1 Z_{ij1} + \beta_2 Z_{ij2} + b_{i0} + b_{i2} Z_{ij2})$. The theoretical mixture distribution is $(\chi^2_1 + \chi^2_2)/2$.

are zero. That is, under the null model, we have $\exp(\beta_1 Z_{ij1} + b_{i0})$ in the relative risk, while under the alternative model we have $\exp(\beta_1 Z_{ij1} + \beta_2 Z_{ij2} + b_{i0} + b_{i2} Z_{ij2})$. This is Case 2 of Section 3.2, and the likelihood ratio statistic has an asymptotic null distribution of $(\chi^2_1 + \chi^2_2)/2$. Figure 2 shows the Q-Q plots of the likelihood ratio statistics from 100 simulations against the theoretical mixture distribution quantiles. The empirical significance levels of the nominal 0.05-level tests are also given in Table 1, along with the error rates of pAIC. From the simulations we see that for the larger cluster size 20, for both numbers of clusters $n = 20$ and 100 the null distribution of the likelihood ratio test was well approximated by its asymptotic counterpart. For the smaller cluster size 5, the asymptotic approximation was less accurate and tends to be conservative. This tendency was also seen for $d = 1$ above.

7. Application

In this section we consider the multi-center non-small cell lung cancer trial that was discussed in Vaida and Xu (2000). The trial enrolled 579 patients

Table 2. Parameter estimates (and standard errors) from the lung cancer data.

Model		1	2	3
treatment	β	−0.25 (0.09)	−0.25 (0.10)	−0.25 (0.12)
bone		0.22 (0.09)	0.21 (0.10)	0.23 (0.14)
liver		0.43 (0.09)	0.42 (0.09)	0.39 (0.09)
ps		−0.60(0.10)	−0.64 (0.11)	−0.65 (0.13)
wt loss		0.20 (0.09)	0.22 (0.09)	0.21 (0.09)
treatment	σ	-	0.27 (0.13)	0.21 (0.43)
bone		-	-	0.36 (0.12)

from 31 institutions. The primary endpoint was patient death. There were two randomized treatment arms in the trial, a standard chemotherapy (CAV) arm, or an alternative 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 the tables. 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 other three 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 are reproduced in Table 2; for more discussion on the parameter estimates see Vaida and Xu (2000). Table 3 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 were computed after 50 *EM* steps where the maximum likelihood estimate has converged; the sample sizes for Gibbs sampler during *MCEM* were 100 initially and increased to 1,000 for the last 10 *EM* steps. The Monte Carlo sample sizes for *RIS* and *BS* were 1,000, respectively. From the table we see that the values of the log likelihoods agree well among the three computational methods.

As seen in Table 3, if we 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 3.2, with critical value of 2.71 at .05 significance

Table 3. $-2\times$ 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.).

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 $(\chi_0^2 + \chi_1^2)/2$ 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 for this dataset, there are 31 institutions and the average number of patients per institution is close to 20, so we expect the asymptotic approximation to be reasonably good. On the other hand, 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 again Case 4 in Section 3.2. 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 rejected in favor of Model 3.

Finally, Model 0 in Table 3 is the Cox model with only fixed effects for the four covariates other than treatment. The comparison of Model 0 versus Model 2 provides an illustration for Case 2 of Section 3.2, i.e., under the null Model 0 the random treatment effect is zero and a subset of the fixed effects, namely the treatment effect, is also zero. 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 significance level.

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

8. Discussion

In this paper, 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 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 semiparametric models where, for the latter type of models, the focus is on the finite-dimensional 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.

We focused on the PHMM with multinormal random effects. The development of the likelihood ratio test and the profile AIC holds in more general situations. For example, for gamma frailty PHMM, Murphy and van der Vaart (2000) show that the asymptotic expansion (2.3) holds, on which our development hinges. Alternatives to the multinormal include the multivariate t distribution, considered in Sargent (1998). Glidden (1999), Viswanathan and Manatunga (2001) and Economou and Caroni (2005) proposed methods for checking the distributional assumption of the random effects under the frailty models. O'Quigley and Stare (2002) demonstrated the robustness in the fixed effects estimation against frailty distribution misspecifications. This latter result parallels the existing literature for the linear, nonlinear and generalized linear mixed models, which shows that the estimation of the fixed effects and variance components is practically unaffected by the misspecifications of the random effect distributions (Verbeke and Lesaffre (1997), Neuhaus, Hauck and Kalbfleisch (1992), Chen, Zhang, and Davidian (2002) and Agresti, Caffo and Ohman-Strickland (2004)).

Model selection has been an area of growing interest in 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 currently we are working 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.

Acknowledgements

The authors would like to thank the editors and referees for helpful suggestions. They would also like to acknowledge the joint work with colleagues Anthony Gamst and Michael Donohue on the asymptotic theory of PHMM, on which part of this work is based. The first author's work was partially supported by NIH grant M01 RR000827, and the second author acknowledges partial support from NIH grants AI-51164, MH-22005, and AI-47033 and the third author acknowledges support from NIH grant 5R01 AI052817.

Appendix

Proof of Theorem 1. 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 (2.3).

Proof of Theorem 3. From Theorem 1 we have that $\sqrt{n}(\hat{\phi} - \phi_0) = O_p(1)$. Applying (2.3) for the sequence $\phi_n = \hat{\phi}$, we get

$$\mathrm{pl}(y^*|\hat{\phi}(y)) = \mathrm{pl}(y^*|\phi_0) + s(y^*|\phi_0)'(\hat{\phi}-\phi_0) - \frac{n}{2}(\hat{\phi}-\phi_0)'\mathcal{I}_0(\hat{\phi}-\phi_0) + r_1, \quad \text{(A.1)}$$

where $r_1 = o_p(1)$. The main result (2.2) from Murphy and van der Vaart (2000) implies that $E\{s(y^*|\phi_0)\} = 0$ (divide by $\sqrt{n}$ and take limits on both sides of (2.2), and then apply the Strong Law of Large Numbers). Therefore, taking expectations on both sides of the equality in (A.1), the first-order term vanishes and we get

$$E_{f(y^*)}\{\mathrm{pl}(y^*|\hat{\phi}(y))\} = E\{\mathrm{pl}(\phi_0)\} - \frac{n}{2}(\hat{\phi} - \phi_0)'\mathcal{I}_0(\hat{\phi} - \phi_0) + E(r_1). \quad \text{(A.2)}$$

Taking expectation with respect to y on both sides of (A.2), we have

$$\begin{aligned}\mathrm{pAI} &= -2E\{\mathrm{pl}(y|\phi_0)\} + E\{n(\hat{\phi} - \phi_0)'\mathcal{I}_0(\hat{\phi} - \phi_0)\} + E(r_1)\\ &= -2E\{\mathrm{pl}(y|\hat{\phi}(y))\} + 2E\{\mathrm{pl}(y|\hat{\phi}(y)) - \mathrm{pl}(y|\phi_0)\} + E\{n(\hat{\phi}-\phi_0)'\mathcal{I}_0(\hat{\phi}-\phi_0)\}\\ &\quad +E(r_1).\end{aligned}$$

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 $r = o_p(1)$, we get $\text{pAI} = -2E\{\text{pl}(y|\hat{\phi}(y))\} + 2p + E(r)$ 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 $A = \sum_k \exp\{\sum_i v(b_i^{(k)}\}/M$. Assume for simplicity that $n = 2$ (the general case follows by induction). Put $\exp\{v(b_i)^{(k)}\} = \xi_i^{(k)}$, 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 over M observations. Let μ_i, σ_i^2 denote respectively the mean and variance of ξ_i, $i = 1, 2$. Then

$$\begin{aligned}
\text{Var}\,(\overline{\xi_1\xi_2}) &= \frac{\text{Var}\,(\xi_1\xi_2)}{M} \\
&= \frac{\sigma_1^2\sigma_2^2}{M} + \frac{\mu_1^2\sigma_2^2}{M} + \frac{\mu_2^2\sigma_1^2}{M} \\
\text{Var}\,(\bar{\xi}_1\bar{\xi}_2) &= \Big(\frac{\sigma_1^2}{M}\Big)\Big(\frac{\sigma_2^2}{M}\Big) + \frac{\mu_1^2\sigma_2^2}{M} + \frac{\mu_2^2\sigma_1^2}{M}.
\end{aligned}$$

The first term in $\text{Var}\,(\overline{\xi_1\xi_2})$ is no smaller than the corresponding term in $\text{Var}\,(\bar{\xi}_1\bar{\xi}_2)$, while the other two terms are identical, so the result follows.

References

Agresti, A., Caffo, B. and Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Comput. Statist. Data Anal.* **47**, 639-653.

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

Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100-20.

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

Cai, J. (1999). Hypothesis testing of hazard ratio parameters in marginal models for multivariate failure time data. *Lifetime Data Anal.* **5**, 39-53.

Cai, J., Fan, J., Li, R. and Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika* **92**, 303-316.

Chen, J., Zhang, D. and Davidian, M. (2002). A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics* **3**, 347-360.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Statist.* **25**, 573-578.

Claeskens, G. and Carroll, R. J. (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika* **94**, 249-265.

Claeskens, G. and Hjort, N. L. (2003). Focused information criterion (with discussion). *J. Amer. Statist. Assoc.* **98**, 900-945.

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

Cortiñas-Abrahantes, J., Legrand, C., Burzykowski, T., Janssen, P., Ducrocq, V. and Duchateau, L. (2007). Comparison of different estimation procedures for proportional hazards model with random effects. *Comput. Statist. Data Anal.* **51**, 3913-3930.

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. *J. Roy. Statist. Soc. Ser. 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. Springer, New York.

DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *J. Amer. Statist. Assoc.* **92**, 903-915.

Economou, P. and Caroni, C. (2005). Graphical tests for the assumption of gamma and inverse gaussian frailty distribution. *Lifetime Data Anal.* **11**, 565-582.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11**, 1031-1057.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30**, 74-99.

Fan, J. and Wong, W. H. (2000). Discussion of 'On profile likelihood', by Murphy, S. A. and van der Vaart, A. W. *J. Amer. Statist. Assoc.* **95**, 468-471.

Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29**, 153-193.

Gelfand, A. and Day, D. (1994). Bayesian model choice: asymptotics and exact calculations. *J. Roy. Statist. Soc. Ser. B* **56**, 501-514.

Glidden, D. (1999). Checking the adequacy of the gamma frailty model for multivariate failure times. *Biometrika* **86**, 381-393.

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

Gray, R. (1995). Tests for variation over groups in survival data. *J. Amer. Statist. Assoc.* **90**, 198-203.

Greven, S., Crainiceanu, C. M., Peters, A. and Kuechenhoff, H. (2007). Likelihood ratio testing for zero variance components in linear mixed models. In *Proceedings of the 22nd International Workshop on Statistical Modelling*, 300-305.

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

Li, B. (2000). Comment on 'on profile likelihood'. *J. Amer. Statist. Assoc.* **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? *J. Roy. Statist. Soc. Ser. A* **168**, 469-472.

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

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

Murphy, S. (1994). Consistency in a proportional hazards model incorporating a random effect. *Ann. Statist.* **22**, 712-731.

Murphy, S. (1995). Asymptotic theory for the frailty model. *Ann. Statist.* **23**, 182-198.

Murphy, S. and van der Vaart, A. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **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.

Neuhaus, J. M., Hauck, W. W. and Kalbfleisch, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79**, 755-762.

O'Quigley, J. and Stare, J. (2002). Proportional hazards models with frailties and random effects. *Statist. Medicine* **21**, 3219-33.

Parner, E. (1998). Asymptotic theory for the correlated Gamma-frailty model. *Ann. Statist.* **26**, 183-214.

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

Sargent, D. J. (1998). A general framework for random effects survival analysis in the Cox proportional hazards setting. *Biometrics* **54**, 1486-1497.

Self, S. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* **82**, 398, 605-610.

Severini, T. A. and Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **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.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **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. *Statist. Medicine* **19**, 3309-3324.

Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Comput. Statist. Data Anal.* **53**, 541-556.

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

Viswanathan, B. and Manatunga, A. (2001). Diagnostic plots for assessing the frailty distribution in multivariate survival data. *Lifetime Data Anal.* **7**, 143-155.

Vu, H. T. V. and Zhou, S. (1997). Generalization of likelihood ratio tests under nonstandard conditions. *Ann. Statist.* **25**, 897-916.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 307-333.

Xu, R., Gamst, A., Donohue, M., Vaida, F. and Harrington, D. (2006). Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models. *Harvard University Biostatistics Working Paper Series.* `http://www.bepress.com/harvardbiostat/paper43`.

Division of Biostatistics and Bioinformatics, Department of Family and Preventive Medicine, and Department of Mathematics, University of California, San Diego, CA92093-0112, U.S.A.

E-mail: rxu@ucsd.edu

Division of Biostatistics and Bioinformatics, Department of Family and Preventive Medicine, University of California, San Diego, CA92093, U.S.A.

E-mail: vaida@ucsd.edu

Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA02115, U.S.A.

E-mail: dph@jimmy.harvard.edu

(Received February 2007; accepted October 2007)