Biometrika (2011), **98**, 3, *pp.* 685–700
© 2011 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/asr023
Advance Access publication 13 July 2011

Conditional Akaike information under generalized linear and proportional hazards mixed models

By M. C. DONOHUE

Division of Biostatistics and Bioinformatics, Department of Family and Preventive Medicine, University of California, San Diego, CA 92093, U.S.A.
mdonohue@ucsd.edu

R. OVERHOLSER

Department of Mathematics, University of California, San Diego, CA 92093, U.S.A.
rhaut@ucsd.edu

R. XU and F. VAIDA

Division of Biostatistics and Bioinformatics, Department of Family and Preventive Medicine, University of California, San Diego, CA 92093, U.S.A.
rxu@ucsd.edu vaida@ucsd.edu

Summary

We study model selection for clustered data, when the focus is on cluster specific inference. Such data are often modelled using random effects, and conditional Akaike information was proposed in Vaida & Blanchard (2005) and used to derive an information criterion under linear mixed models. Here we extend the approach to generalized linear and proportional hazards mixed models. Outside the normal linear mixed models, exact calculations are not available and we resort to asymptotic approximations. In the presence of nuisance parameters, a profile conditional Akaike information is proposed. Bootstrap methods are considered for their potential advantage in finite samples. Simulations show that the performance of the bootstrap and the analytic criteria are comparable, with bootstrap demonstrating some advantages for larger cluster sizes. The proposed criteria are applied to two cancer datasets to select models when the cluster-specific inference is of interest.

Some key words: Akaike information; Conditional likelihood; Effective degrees of freedom.

1. Introduction

Mixed effects models have been widely used to analyse clustered data, which arise in applications such as longitudinal studies, familial studies and multicentre clinical trials. The focus of inference under such models can be on the population parameters, such as the fixed regression effects, on the variance parameters or on the cluster-level parameters, often represented by the random effects themselves. As an example of the latter, in a multicentre clinical trial where the treatment effect is found to be heterogeneous among the trial centres, it is of scientific interest to estimate the treatment effects from individual centres, and to investigate the cause of the treatment differences. Other examples of cluster-level focus occur in ecology, small-area estimation, and animal husbandry.

Specifically, assume that the data consist of outcomes from m clusters, with n_i observations in cluster i $(i = 1, \ldots, m)$. Within a cluster, the outcomes are dependent, but conditional on the cluster-specific $d \times 1$ vector of random effects b_i, the outcomes y_{ij} are independent and follow a generalized linear mixed model with mean

$$\mu_{ij} = E(y_{ij} \mid \beta, b_i) = g^{-1}(\beta^{\mathrm{T}} x_{ij} + b_i^{\mathrm{T}} z_{ij}), \tag{1}$$

where x_{ij} and z_{ij} are the covariate vectors for the fixed effects β and the random effects b_i of cluster i, $b_i \sim p(b_i)$ and g is the link function. For cluster-level inference, any future prediction takes place in the same clusters as the observed data, and the random effects for these clusters are held constant (Vaida & Blanchard, 2005). More specifically, let y^0 be independently replicated outcomes from the same conditional distribution as the original data y given the same random effects b. Here y, y^0 and b are random vectors consisting of elements y_{ij}, y_{ij}^0 and b_i, respectively. For model selection purposes, Vaida & Blanchard (2005) defined the conditional Akaike information,

$$\text{cAI} = -2E_{(y,b)} E_{y^0|b}[l\{y^0 \mid \hat{\beta}(y), \hat{b}(y)\}], \tag{2}$$

where $l(\cdot \mid \cdot)$ is the conditional loglikelihood given the random effects, and $\hat{\beta}(y)$, $\hat{b}(y)$ are estimators of β and b based on the data y, for example, the maximum likelihood and the empirical Bayes estimators. The expectations are taken with respect to the true model generating the data. They proceeded to show that for the linear mixed model with known variance components, an unbiased estimator of (2) is

$$\text{cAIC} = -2l\{y \mid \hat{\beta}(y), \hat{b}(y)\} + 2\rho\ , \tag{3}$$

where the bias correction factor ρ is the effective degrees of freedom of the linear mixed model of Hodges & Sargent (2001) and Ye (1998). Expression (3) is referred to as the conditional Akaike information criterion. Vaida & Blanchard (2005) and Liang et al. (2008) give formulae for ρ in the more general case of unknown variance parameters in finite samples. The theory of the Akaike information criterion and its basis in model prediction are well understood (Akaike, 1973; Linhart & Zucchini, 1986; Burnham & Anderson, 2002). In this paper we develop the conditional Akaike information and its criterion under generalized linear and proportional hazards mixed models. Exact calculation is not available outside normal linear mixed models, and asymptotic approximations are necessary. An additional concern is the presence of nuisance baseline hazard function in the proportional hazards mixed models.

Generalized linear mixed models have been studied for the past two decades (Jiang, 2007; McCulloch et al., 2008). In contrast, the proportional hazards mixed model has only recently been proposed to model complex correlated time-to-events data (Gustafson, 1997; Sargent, 1998; Vaida & Xu, 2000; Ripatti & Palmgren, 2000; Ripatti et al., 2002); it includes the univariate frailty model (Hougaard, 2000) as a special case. The asymptotics were considered in Gamst et al. (2009). Xu et al. (2009) consider the marginal AIC which focuses on the fixed effects and the variance parameters.

From a different perspective and not specifically for clustered data, the issue of focus of model selection was addressed in Claeskens & Hjort (2003) and Hjort & Claeskens (2006). For an overview of the AIC see the Burnham & Anderson (2002) and Claeskens & Hjort (2008).

2. Generalized linear mixed models

Consider the model given by (1). To set the notation, write $D^{-1} = -\partial^2 \log p(b)/\partial b \partial b'$, where $p(b) = \prod_{i=1}^{m} p(b_i)$ is the distribution of the independent random effects; when $b_i \sim N(0, \Sigma)$, $D = \text{var}(b) = \text{diag}_m(\Sigma)$, the block-diagonal matrix with m blocks equal to Σ. Let X_i and Z_i be the matrices with rows x_{ij}^{T} and z_{ij}^{T}, and let $X = \text{stack}(X_1, \ldots, X_m)$ and $Z = \text{diag}(Z_1, \ldots, Z_m)$ be the $N \times p$ and $N \times q$ model matrices, where p , d and $q = md$ are the lengths of β, b_i and b, $N = n_1 + \cdots + n_m$, and the stack function stacks matrices on top of each other. Further, let $w_i = (w_{i1}, \ldots, w_{in_i})^{\mathrm{T}}$ be the vector of weights given by $w_{ij} = [\text{var}(y_{ij} \mid b_i)\{g'(\mu_{ij})\}^2]^{-1}$, and $W = \text{diag}(w_1, \ldots, w_m)$. The usual estimator for β is the maximizer of the marginal likelihood, $L(\beta) = \int \exp\{l_J(y, b \mid \beta)\}\, db$, where

$$l_J(y, b \mid \beta) = l(y \mid \beta, b) + \log p(b) \tag{4}$$

is the joint loglikelihood. Alternatively, (β, b) are estimated jointly as the maximizer of the joint loglikelihood l_J. The joint loglikelihood (4) and its maximizer $(\hat{\beta}, \hat{b})$ have been variously justified. Breslow & Clayton (1993), Wolfinger (1993) and Vonesh (1996) show that under suitable conditions $\hat{\beta}$ is a first-order Laplace approximation to the maximum likelihood estimator. Given the fixed effects, the joint likelihood is proportional to the posterior distribution of the random effects, maximized by $\hat{b}$ (Jiang, 2001). Lee & Nelder (1996) call (4) the hierarchical loglikelihood, and consider it as a basis of inference; see also Lee et al. (2006). In the smoothing literature l_J is seen as a penalized loglikelihood (see, e.g., Wager et al., 2007). The variance matrix Σ that is suppressed in $p(b)$ is estimated by maximum likelihood or residual maximum likelihood (Breslow & Clayton, 1993).

Let $U = (X, Z)$ and $\theta = \text{stack}(\beta, b)$, so that $U\theta = X\beta + Zb$. Let $s_J = \partial l_J/\partial\theta$ be the score function for the joint loglikelihood. Standard derivations show that

$$G = \text{var}\{s_J(y) \mid \theta\} = U^{\mathrm{T}} W U, \tag{5}$$

$$\Omega = E\{s_J(y)s_J^{\mathrm{T}}(y) \mid \theta\} = E\{-\partial^2 l_J(y)/\partial\theta\partial\theta^{\mathrm{T}} \mid \theta\} = U^{\mathrm{T}} W U + \text{diag}(0, D^{-1}). \tag{6}$$

Further, let

$$\rho = \text{tr}(G\,\Omega^{-1}) = \text{tr}[U^{\mathrm{T}} W U\, \{U^{\mathrm{T}} W U + \text{diag}(0, D^{-1})\}^{-1}]. \tag{7}$$

For the linear mixed model, $W = I$. In this case, ρ is the effective degrees of freedom of Hodges & Sargent (2001), as well as the correction factor for conditional AIC (3) in Theorem 1 of Vaida & Blanchard (2005). Lu et al. (2007) use a form similar to ρ as the effective degrees of freedom for the generalized linear mixed models. The following result shows that ρ is asymptotically the relevant correction factor for the conditional AIC (3).

THEOREM 1. *Assume that the data y are generated from the generalized linear mixed model* (1). *Let $\hat{\beta}$ be the maximum likelihood estimator, and $\hat{b}$ the maximizer of the joint likelihood given $\hat{\beta}$ and the maximum likelihood estimate of* Σ. *Under Conditions* A1–A11 *given in the Appendix, an asymptotically unbiased estimator of the conditional Akaike information* (2) *is given by the conditional* AIC (3)*, with $\rho = tr(G\Omega^{-1})$ as in* (7)*. That is, E*(CAIC) = CAI + $o(1)$ *for large m and n_i.*

In addition, the effective degrees of freedom ρ satisfies $p \leqslant \rho \leqslant p + q$.

The proofs for this and for the following results are given in the Appendix.

In practice, W is computed at $\theta = \hat{\theta}$. Using formulae for explicit computation of the inverse matrix in (7) (Harville, 1996, p.99) we get

$$\rho = (p+q) - \text{tr}[\{Z^{\mathrm{T}}WZ - Z^{\mathrm{T}}WX(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}WZ + D^{-1}\}^{-1}D^{-1}]. \tag{8}$$

Formulae (5) and (6) are a form of the Bartlett identities for the joint likelihood. The more general form is given below.

Proposition 1. *Bartlett identities for joint or penalized likelihood. Assume that the data y have loglikelihood $l(y \mid \theta)$, satisfying the standard regularity conditions which ensure differentiation with respect to parameter θ under the integral sign, as well as the first two Bartlett identities: $E\{s(y \mid \theta)\} = 0$ and $E\{-\ddot{l}(y \mid \theta)\} = E\{s(y \mid \theta)s(y \mid \theta)^{\mathrm{T}}\}$, where s and $\ddot{l}$ denote the first two derivatives of l with respect to θ. Further, assume that $p(\theta)$ is a nonnegative function of θ not depending on y, twice differentiable with respect to θ and with continuous second derivatives. Let $l_J(y \mid \theta) = l(y \mid \theta) + \log p(\theta)$. Let s_J and $\ddot{l}_J$ be the first two derivatives of l_J. Then l_J satisfies the modified Bartlett identities*:

$$\begin{aligned} E(s_J \mid \theta) &= \partial \log p(\theta)/\partial\theta, \\ \text{var}(s_J \mid \theta) &= \text{var}(s \mid \theta), \\ E(-\ddot{l}_J \mid \theta) &= E(-\ddot{l} \mid \theta) - \partial^2 \log p(\theta)/\partial\theta\partial\theta^{\mathrm{T}} = \text{var}(s_J \mid \theta) - \partial^2 \log p(\theta)/\partial\theta\partial\theta^{\mathrm{T}}. \end{aligned}$$

In particular, if $\theta = (\beta, b)$ and $p(\theta) = p(b)$ is the $N(0, D)$ density, then $\partial \log p(\theta)/\partial\theta = A\theta$ and $-\partial^2 \log p(\theta)/\partial\theta\partial\theta^{\mathrm{T}} = A = \text{diag}(0, D^{-1})$.

The proof follows directly from the definition of l_J and the Bartlett identities for l.

3. Proportional hazards mixed models

3·1. *Direct derivation of conditional* AIC

In the proportional hazards mixed model for clustered right-censored data, the hazard function for the jth observation of the ith cluster is

$$\lambda_{ij}(t) = \lambda_0(t)\exp(\beta^{\mathrm{T}}x_{ij} + b_i^{\mathrm{T}}z_{ij}) \qquad (i = 1, \ldots, m;\ j = 1, \ldots, n_i), \tag{9}$$

where $b_i \sim p(b_i)$ independently of each other, and x_{ij}, z_{ij} are the covariate vectors associated with the fixed and the random effects β and b_i, as before. The fixed intercept is absorbed in the baseline hazard function $\lambda_0(t)$, while the random cluster effect on the hazard needs to be included as a 1 in z_{ij}. Vaida & Xu (2000) developed the nonparametric maximum likelihood estimator of the parameters in this model, computed using a Monte Carlo EM algorithm, with the random effects estimated by the posterior empirical Bayes expectation. The program is available in the R package phmm.

The outcome data corresponding to $\lambda_{ij}(t)$ is $y_{ij} = (Y_{ij}, \delta_{ij})$, where Y_{ij} is the possibly right-censored failure time and δ_{ij} is the failure-event indicator. Put $y_i = \text{stack}(y_{i1}, \ldots, y_{in_i})$, $y = \text{stack}(y_1, \ldots, y_m)$, and $\eta_{ij} = \beta^{\mathrm{T}}x_{ij} + b_i^{\mathrm{T}}z_{ij}$. The conditional loglikelihood is

$$l(y \mid \beta, b, \lambda_0) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\{\delta_{ij}\log\lambda_0(Y_{ij}) + \delta_{ij}\eta_{ij} - \Lambda_0(Y_{ij})e^{\eta_{ij}}\}, \tag{10}$$

with $\Lambda_0(t) = \int_0^t \lambda_0(s)\, ds$ the cumulative baseline hazard function. Profiling the baseline hazard out of (10) we get the conditional profile loglikelihood

$$pl(y \mid \beta, b) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} \log \frac{\exp(\eta_{ij})}{\sum_{i'j'} \exp(\eta_{i'j'}) I(Y_{i'j'} \geqslant Y_{ij})}. \tag{11}$$

This is also the conditional partial loglikelihood; the rationale for using the profile likelihood to define an Akaike information in the presence of nuisance parameters is explained in Xu et al. (2009). For model (9) we define the conditional profile Akaike information as

$$\textsc{cai} = -2E_{(y,b)} E_{(y^0 \mid b)}[pl\{y^0 \mid \hat{\beta}(y), \hat{b}(y)\}]. \tag{12}$$

The joint conditional profile loglikelihood is

$$pl_J(y \mid \theta) = pl(y \mid \beta, b) + \log p(b), \tag{13}$$

where $\theta = \text{stack}(\beta, b)$. This is also called the penalized partial loglikelihood (Ripatti & Palmgren, 2000), or the h-likelihood (Ha & Lee, 2003), and its maximizer can be seen as an approximation of the maximum likelihood estimator. Let $s_J = \partial pl_J / \partial\theta$. Calculation in Ha & Lee (2003) shows that

$$s_J(y \mid \theta) = U^{\mathrm{T}}(\delta - \hat{\mu}) + A\theta,$$

where $\delta = \text{stack}(\delta_{ij})$, $\mu = \text{stack}(\mu_{ij})$, $\mu_{ij} = \exp(\eta_{ij} + \log \Lambda_0(Y_{ij}))$, and $\hat{\mu} = \mu(\hat{\lambda}_0)$ where $\hat{\lambda}_0 = \hat{\lambda}_0(\theta)$ is the usual Breslow's estimate of the discretized baseline hazard for a given θ; $U = (X, Z)$, and $A = \text{diag}(0, D^{-1})$, as before. In addition $-\partial^2 pl/\partial\theta\partial\theta^{\mathrm{T}} = U^{\mathrm{T}} W^* U$, where $W^* = W_1 - W_2$, with $W_1 = \text{diag}(\hat{\mu})$ and W_2 nondiagonal, given in the Appendix. Therefore,

$$-\partial^2 pl_J/\partial\theta\partial\theta^{\mathrm{T}} = -\partial s_J/\partial\theta = U^{\mathrm{T}} W^* U + A,$$

and $\Omega = E\{-\partial^2 pl_J(y)/\partial\theta\partial\theta^{\mathrm{T}} \mid \theta\} = U^{\mathrm{T}} W U + A$ with $W = E(W^* \mid \theta)$. In addition, as the Barlett identities hold asymptotically for the partial likelihood, following Proposition 1 we have the asymptotic variance needed in Theorem 2 below: $G = \text{var}\{s_J(y) \mid \theta\} = U^{\mathrm{T}} W U$. The conditional profile AIC corresponding to (12) is given by

$$\textsc{caic} = -2pl\{y \mid \hat{\beta}(y), \hat{b}(y)\} + 2\rho, \tag{14}$$

with $\rho = \text{tr}(G\,\Omega^{-1})$. The development here is parallel to the generalized linear mixed model case in § 2. Since W is not available in closed form, in practice, we use

$$\rho = \text{tr}\{\, U^{\mathrm{T}} W^* U\, (U^{\mathrm{T}} W^* U + A)^{-1}\, \}. \tag{15}$$

The following result shows that under suitable conditions (14) is an asymptotically unbiased estimator of the conditional Akaike information in (12).

Theorem 2. *Assume that the data y are generated under* (9)*. Let $\hat{\beta}$ be the maximum likelihood estimator and $\hat{b}$ the maximizer of the profile joint likelihood* (13) *given $\hat{\beta}$ and the maximum likelihood estimate of Σ. Under analogous conditions to Theorem* 1*, an asymptotically unbiased estimator of the conditional Akaike information* (2) *is given by the conditional* AIC *in* (14)*, i.e. $E(\textsc{caic}) = \textsc{cai} + o(1)$ for large m and n_is. Moreover, $p \leqslant \rho \leqslant p + q$.*

An alternative formula for ρ is given by (8) with W^* in place of W. We call ρ the effective degrees of freedom for the proportional hazards mixed model.

3·2. *Poisson formulation*

Let $t_1 < \cdots < t_K$ be the K distinct event times among the Y_{ij}s. It is well known that the loglikelihood of the proportional hazards model is equivalent, up to a constant, to that of a Poisson model with outcomes $\xi_{ij,k} = \delta_{ij} I(Y_{ij} = t_k)$ and mean $\exp\{\log \lambda_0(t_k) + \eta_{ij}\}$, for all i, j and k such that $t_k \leqslant Y_{ij}$. This connection was observed early in Whitehead (1980), and extended to the mixed model in Ma et al. (2003), Ha & Lee (2003) and Kauermann et al. (2008), among others. The following result shows that the conditional AIC based on the Poisson generalized linear mixed model formulation is identical to that in (14) and (15) up to a constant.

Proposition 2. *For the proportional hazards mixed model let conditional* AIC$_P$ *be the conditional Akaike information criterion corresponding to the Poisson formulation, i.e.,* cAIC$_P = -2l(\xi \mid \hat{\theta}) + 2\rho_P$, *where* $l(\xi \mid \hat{\theta})$ *is the Poisson loglikelihood conditional on* b, *with data* $\xi = (\xi_{ij,k})$, *and* ρ_P *is the corresponding Poisson degrees of freedom defined in* (7)*. Then* $\rho_P = \rho + a_1$, $l(\xi \mid \hat{\theta}) = pl(y \mid \hat{\theta}) + a_2$, *and* cAIC$_P =$ cAIC $+ a$, *where* a_1, a_2 *and* $a = 2(a_1 - a_2)$ *are constants depending only on the data* y, *and* cAIC *and* ρ *are given by* (14) *and* (15)*.*

For computational efficiency and numerical accuracy, it is advantageous to fit the proportional hazards mixed model directly, rather than using the equivalent Poisson formulation.

4. Alternative estimation using the bootstrap

The bootstrap has been used in the estimation of prediction errors (Efron, 1983, 1986; Xu & Gamst, 2008) and for Akaike-type criteria (Shibata, 1997), and has shown less bias in finite samples (Cavanaugh & Shumway, 1997; Pan, 1999; Shang & Cavanaugh, 2008). Our proposed estimate is

$$\text{cAIC}_b = -2l(y \mid \hat{\beta}, \hat{b}) + 2\rho_b.$$

The correction factor ρ_b is given by

$$\rho_b = E^*\{l(y^* \mid \hat{\beta}^*, \hat{b}^*) - l(y \mid \hat{\beta}^*, \hat{b}^*)\}, \tag{16}$$

where E^* denotes the bootstrap expectation, i.e., the average over the bootstrap datasets; the bootstrap datasets y^* are obtained by resampling first the clusters, then the observations within cluster, and $(\hat{\beta}^*, \hat{b}^*) = \{\hat{\beta}(y^*), \hat{b}(y^*)\}$. For each cluster in y^*, the data y from the same cluster are used in calculating ρ_b. For applications with extremely small clusters, such as in the lung cancer example below where some clusters have only one observation, resampling within the clusters might be infeasible. In this case, parametric or model-based bootstrap can be used, where the bootstrap data are generated under a fitted large model, with estimated fixed and random effects. The formula (16) is derived similar to Yafune et al. (2005), but adjusted to our conditional setting.

5. Simulation

We carried out simulation experiments to evaluate the proposed criteria under the generalized linear and proportional hazards mixed models. Here we report a few representative results. The emphasis is two-fold: on the criteria as estimators of the underlying Akaike information as well as on their success in selecting the correct model. We computed the conditional Akaike information by simulation, and its criteria with correction factors ρ and ρ_{b}. The results are reported in Tables 1–3. The numbers of fixed and random effects in each model are indicated as a pair at the top of each column; for example, (2,1) indicates that two fixed effects and one random effect are included in the model. In each case we used a combination of small and large numbers of clusters m and observations within cluster n_i. We used two model selection rules: (i) the rule of two, in which one selects the smallest model whose criterion value is within 2 of the minimum criterion value; or (ii) select the model with the minimum criterion value. The rule of two acknowledges the variation in a criterion as an estimate of the underlying information, so that for models with close criterion values there is not enough evidence for a preference; in this situation a parsimony principle is applied. The estimation used the lme4 and phmm packages in (R Development Core Team, 2011).

Table 1 reports the simulation under a log-link Poisson generalized linear mixed model. Overall conditional AIC provides a good estimate of the conditional Akaike information, within the statistical error range. In the first scenario of 10 clusters of 5 observations each, although the average of the conditional AIC is minimized at the larger model (3,3), it is not significantly different from that of the true model (3,2). The rule of two chooses the correct model most often, and chooses the larger model (3,3) between 8 and 30% of the cases. Note, however, that the conditional Akaike information criteria for models (3,2) and (3,3) are very close to each other, so it is not surprising that they often erroneously prefer the larger model. The nonparametric bootstrap works well when the cluster sizes are reasonably large, e.g. $n_i \geqslant 10$, and when the model is not too far from the truth. While the true model is (3,2), under model (2,1) conditional AIC$_b$ is typically more than twice the standard error away from conditional Akaike information except when $n_i = 40$. Such inaccuracy seems to be attributable to the model fitting procedure by lmer and the high probability of duplicated data with small cluster resampling in the bootstrap. In the survival data case below, fitted by Markov chain Monte Carlo EM which has shown good numerical stability and accuracy (Gamst et al., 2009), the results are extremely good in comparison.

The proportional hazards mixed model simulations are summarized in Tables 2 and 3. Since models (2,2) and (3,1) in Table 2 are not nested, for the rule of two, the size of a model is determined by ρ or ρ_b. The analytic and the bootstrap methods are comparable according to this table, both in terms of bias and the selection of the correct model. The bootstrap CAIC$_b$ shows a slight advantage over the conditional AIC in picking the correct model. The rule of two consistently meets or outperforms the simple minimum rule. The simulations in Table 3 have smaller between-cluster variability compared with Table 2, and as expected the model selection criteria are less effective in correctly selecting the random effects. On average the conditional AIC is minimized at the largest model; this might be expected since the Laplace approximation used in the derivation is more accurate when the cluster sizes are large. On the other hand, when standard errors from the simulation are taken into account, the conditional AIC values are practically identical for models (2,1) and (2,2). The bootstrap CAIC$_b$ is minimized at the true model on average. For selecting the true model with smaller cluster sizes, the conditional AIC outperforms the bootstrap CAIC$_b$.

6. Case studies

6·1. *The skin cancer prevention study*

The Skin Cancer Prevention Study was a randomized, double-blinded, placebo-controlled clinical trial of non-melanoma skin cancer prevention in 1805 high-risk subjects randomized to either 50 mg of beta-carotene daily or placebo, for up to five years (Greenberg et al., 1990; Fitzmaurice et al., 2004). The dataset consisted of the $m = 1683$ subjects with complete covariate information, with $N = 7081$ observations. We fitted Poisson mixed models with log-link for the main outcome, the number of new skin cancers y_{ij} for subject i in year j. The covariates

Table 1. *Comparison of model selection procedures based on simulations from a Poisson generalized linear mixed model with log-link. Data are generated from model* $(3, 2)$*: covariates* x_1, x_2, x_3 *are independent Bernoulli* $(0{\cdot}5)$*,* $\beta = (1, 1, 1)^{\mathrm{T}}$*,* $z_1 = x_1, z_2 = x_2$*, and* $b_i \sim N(0, 0{\cdot}25 I_2)$*. Model* $(2, 1)$ *includes* x_1, x_2 *and* z_1*; model* $(3, 1)$ *includes also* x_3*, and* $(3, 3)$ *includes also* z_2 *and* $z_3 = x_3$*. Averages over* 100 *simulations and simulation standard errors, in parentheses, are reported for conditional Akaike information,* cAIC*,* cAIC$_b$ *with* 400 *bootstrap samples, and* $-2l(y \mid \hat{\beta}, \hat{b})$*;* a/b *gives the number of times out of* 100 *a model was chosen using the rule of two* (a) *or simple minimum* (b)

(fixed, random)	(2,1)	(3,1)	(3,2)	(3,3)
$m = 10, n_i = 5$				
cAI	318 (3·8)	244 (1·9)	230 (1·1)	232 (1·1)
	0/0	9/2	89/74	2/24
cAIC	321 (4·4)	244 (2·1)	229 (1·3)	227 (1·8)
	0/0	14/10	73/65	13/25
cAIC$_b$	369 (7·1)	252 (3·0)	235 (2·3)	242 (3·7)
	0/0	17/14	72/66	11/20
$-2pl(y \mid \hat{\beta}, \hat{b})$	302 (4·4)	224 (2·1)	203 (1·2)	200 (1·2)
$m = 10, n_i = 10$				
cAI	674 (7·0)	494 (3·7)	448 (1·6)	450 (1·6)
	0/0	0/0	99/85	1/15
cAIC	675 (7·7)	495 (4·1)	447 (2·4)	448 (3·7)
	0/0	5/5	84/72	11/23
cAIC$_b$	746 (9·7)	509 (5·0)	450 (2·8)	454 (3·0)
	0/0	3/2	86/83	11/15
$-2pl(y \mid \hat{\beta}, \hat{b})$	654 (7·7)	473 (4·1)	414 (2·2)	411 (2·2)
$m = 10, n_i = 40$				
cAI	2732 (28·1)	1929 (14·3)	1727 (6·1)	1729 (6·1)
	0/0	0/0	100/97	0/3
cAIC	2733 (29·0)	1933 (14·7)	1730 (6·9)	1741 (13·6)
	0/0	0/0	92/83	8/17
cAIC$_b$	2810 (31·0)	1943 (15·2)	1729 (7·0)	1733 (7·0)
	0/0	0/0	95/89	5/11
$-2pl(y \mid \hat{\beta}, \hat{b})$	2711 (29·0)	1909 (14·7)	1691 (6·9)	1688 (6·8)
$m = 50, n_i = 5$				
cAI	1612 (8·1)	1240 (4·9)	1137 (2·4)	1140 (2·4)
	0/0	0/0	100/98	0/2
cAIC	1606 (10·0)	1238 (5·5)	1135 (2·8)	1135 (3·5)
	0/0	0/0	70/60	30/40
cAIC$_b$	1821 (14·2)	1276 (7·3)	1131 (3·8)	1137 (3·8)
	0/0	0/0	85/80	15/20
$-2pl(y \mid \hat{\beta}, \hat{b})$	1519 (10·0)	1151 (5·4)	1001 (2·7)	996 (2·5)
$m = 50, n_i = 10$				
cAI	3352 (14·9)	2475 (9·1)	2219 (3·6)	2221 (3·6)
	0/0	0/0	100/96	0/4
cAIC	3354 (17·3)	2479 (10·7)	2218 (4·5)	2222 (6·3)
	0/0	0/0	79/69	21/31
cAIC$_b$	3670 (20·8)	2535 (12·9)	2210 (5·5)	2225 (5·6)
	0/0	0/0	99/99	1/1
$-2pl(y \mid \hat{\beta}, \hat{b})$	3258 (17·3)	2382 (10·7)	2052 (4·4)	2046 (4·2)

cAI, conditional Akaike information; cAIC, conditional Akaike information criterion; cAIC$_b$, bootstrap estimated cAIC.

Table 2. *Comparison of model selection procedures based on simulations from a proportional hazards mixed model. Data are generated from model* $(2, 2)$*: covariates* x_1, x_2, x_3 *are independent Bernoulli* $(0{\cdot}5)$*,* $\beta = (1, 2)^{\mathrm{T}}$*,* $b_i \sim N(0, I_2)$*;* $\lambda_0(t) = 1$*, with* 20% *censoring. Model* $(1, 1)$ *includes* x_1 *and* $z_1 = x_1$*; model* $(2, 1)$ *includes* x_1, x_2 *and* $z_1 = x_1$*, etc. Mean over* 100 *simulations are reported for* cAI*,* cAIC*,* cAIC$_b$ *with* 400 *bootstrap samples, and* $-2l(y \mid \hat\beta, \hat b)$*;* $(a; b)$ *gives the number of times out of* 100 *a model was chosen using the rule of two* (a) *or simple minimum* (b)*. Simulation standard errors:* 2·2–4·3 $(n_i = 20)$*;* 4·4–6·8 $(n_i = 40)$

(fixed, random)	(1,0)	(1,1)	(2,1)	(2,2)	(3,1)	(3,2)	(3,3)
$m = 5, n_i = 20$							
cAI	578 (0;0)	565 (0;0)	527 (11;4)	513 (88;69)	528 (0;0)	514 (1;14)	515 (0;13)
cAIC	579 (0;0)	567 (1; 0)	527 (12; 3)	511 (69; 61)	527 (0;0)	512 (6; 4)	511 (12; 32)
cAIC$_b$	577 (0;0)	566 (1;1)	524 (18;15)	510 (71;63)	525 (0;0)	512 (7;16)	515 (3;5)
$-2pl(y \mid \hat\beta, \hat b)$	577	560	517	496	516	495	493
$m = 5, n_i = 40$							
cAI	1372 (0;0)	1345 (0;0)	1261 (4;1)	1229 (95;77)	1261 (1;0)	1230 (0;12)	1230 (0;10)
cAIC	1370 (0;0)	1344 (0;0)	1258 (5;1)	1225 (86;77)	1260 (1;1)	1227 (1;5)	1226 (7;16)
cAIC$_b$	1367 (0;0)	1341 (0;0)	1254 (5;4)	1221 (88;83)	1255 (1;2)	1222 (2;6)	1225 (4;5)
$-2pl(y \mid \hat\beta, \hat b)$	1368	1336	1247	1208	1247	1207	1206

cAI, conditional Akaike information; cAIC, conditional Akaike information criterion ; cAIC$_b$, bootstrap estimated cAIC.

Table 3. *Comparison of model selection procedures based on simulations from a proportional hazards mixed model, continued. Data are generated from model* $(2, 1)$*: covariates* x_1, x_2 *are independent,* $x_1 \sim \mathrm{Ber}(0{\cdot}5)$*,* $x_2 \sim \mathrm{Un}(0, 1)$*,* $\beta = (1, 1)^{\mathrm{T}}$*,* $b_i \sim N(0, \sigma^2)$*;* $\lambda_0(t) = 1$*, with* 20% *censoring. Simulation standard errors:* 0·8–3·3 $(m = 5)$*;* 2·4–8·7 $(m = 20)$

(fixed, random)	(1,0)	(1,1)	(2,1)	(2,2)
$m = 5, n_i = 20, \sigma^2 = 0{\cdot}05$				
cAI	576 (9;3)	576 (1;1)	571 (90;58)	572 (0;38)
cAIC	573 (18;4)	572 (7;8)	567 (66;32)	566 (9;56)
cAIC$_b$	569 (39;26)	571 (9;6)	566 (49;62)	569 (3;6)
$-2pl(y \mid \hat\beta, \hat b)$	571	568	561	560
$m = 5, n_i = 20, \sigma^2 = 0{\cdot}5$				
cAI	574 (1;0)	565 (7;3)	561 (92;60)	561 (0;37)
cAIC	570 (8;2)	560 (23;15)	555 (59;32)	555 (10;51)
cAIC$_b$	567 (14;7)	558 (22;15)	553 (63;74)	556 (1;4)
$-2pl(y \mid \hat\beta, \hat b)$	568	554	547	545
$m = 20, n_i = 20, \sigma^2 = 0{\cdot}05$				
cAI	3181 (0;0)	3180 (0;0)	3157 (100;64)	3157 (0;36)
cAIC	3178 (0;0)	3175 (0;0)	3154 (88;50)	3153 (12;50)
cAIC$_b$	3174 (21;19)	3183 (0;0)	3161 (76;78)	3172 (3;3)
$-2pl(y \mid \hat\beta, \hat b)$	3176	3165	3142	3138
$m = 20, n_i = 20, \sigma^2 = 0{\cdot}5$				
cAI	3171 (0;0)	3125 (0;0)	3103 (100;69)	3104 (0;31)
cAIC	3169 (0;0)	3122 (1;0)	3101 (86;45)	3100 (13;55)
cAIC$_b$	3165 (0;0)	3120 (1;0)	3099 (98;98)	3112 (1;2)
$-2pl(y \mid \hat\beta, \hat b)$	3167	3092	3068	3064

cAI, conditional Akaike information; cAIC, conditional Akaike information criterion; cAIC$_b$, bootstrap estimated cAIC.

included skin type and gender as binary variables, and age at entry and number of skin cancers at entry as continuous variables. The year of follow-up was either omitted, or included as

Table 4. *Skin cancer prevention study. Estimates of fixed effects and* caic *from five Poisson mixed models differing in the fixed and random effects for year of follow-up. Age and number of skin cancers at baseline are continuous. The best models, marked with a* *, *cannot be ranked based on* caic.

(fixed, random)	(5,1)	(6,1)	(7,1)	(6,2)	(7,2)
β Intercept	−4·28	−4·35	−4·18	−4·79	−5·36
	(0·37)	(0·37)	(0·39)	(0·50)	(0·54)
Age (years)	0·02	0·02	0·02	0·02	0·02
	(0·01)	(0·01)	(0·01)	(0·01)	(0·01)
Skin that burns	0·33	0·33	0·33	0·33	0·32
	(0·10)	(0·12)	(0·11)	(0·14)	(0·15)
Male gender	0·63	0·63	0·63	0·69	0·70
	(0·12)	(0·12)	(0·12)	(0·16)	(0·17)
Baseline cancers	0·18	0·18	0·18	0·19	0·19
	(0·01)	(0·01)	(0·01)	(0·02)	(0·02)
Year	–	0·02	−0·13	−0·03	0·38
	–	(0·02)	(0·08)	(0·04)	(0·11)
Year^2	–	–	0·03	–	−0·08
	–	–	(0·01)	–	(0·02)
σ^2 Intercept	2·37	2·38	2·39	9·97	13·12
Year	–	–	–	0·85	1·22
$-2l(y \mid \hat{\beta}, \hat{b})$	6072·10	6068·57	6063·50	4942·28	4825·32
caic	7824·69*	7824·28*	7823·13*	8023·39	8038·86

caic, conditional Akaike information criterion.

a linear or quadratic effect. In some models, a subject-specific year effect was fitted. All models included a subject-specific random intercept. The treatment effect was proven not significant in earlier analyses and was not included. The results in Table 4 show that the random year effect should not be included in the model; the three models with the year omitted, in linear, or quadratic form yield comparable conditional Akaike criteria and cannot be distinguished. On parsimony grounds, the model without year effect can be chosen. To determine whether the difference of conditional Akaike information between models was significant, a 95% confidence interval of this difference was computed by bootstrap for each pair. The 95% confidence intervals for this difference were as follows: (5,1) versus $(6, 1) = (-2, 19, 18, 55)$; (6,1) versus $(7, 1) = (-1 \cdot 96, 22 \cdot 54)$; (5,1) versus $(7, 1) = (-3 \cdot 74, 28 \cdot 61)$. This analysis confirms that the three models cannot be ranked on conditional aic alone, and that the simpler (5,1) model may be chosen.

6·2. *The E*1582 *lung cancer trial*

The E1582 multicentre non-small cell lung cancer trial including $N = 579$ subjects was discussed and analysed using a proportional hazards mixed model in Vaida & Xu (2000) and Xu et al. (2009), with observations clustered by the $m = 31$ institutions. The number of subjects per institution, n_i, ranged from 1 to 50. The primary endpoint was time to death. The subjects were randomized to either standard chemotherapy or an alternative regimen. Other important covariates related to survival were the presence of bone metastases, presence of liver metastases, performance status at study entry and weight loss, prior to entry, all binary. Gray (1995) found a significant difference in treatment effects across institutions, using a score test. We consider three models here, all of them including the fixed effects for the five important covariates. The first model (5,0) includes no random effects,

Table 5. *E1582 lung cancer trial data. Estimates of fixed effects and variance components and* cAIC$_b$ *from three proportional hazards mixed models. Best model is marked by* *

(fixed, random)	(5,0)	(5,1)	(5,2)
β			
Treatment	−0·254 (0·085)	−0·250 (0·104)	−0·247 (0·119)
Bone metastases	0·223 (0·093)	0·212 (0·095)	0·230 (0·144)
Liver metastases	0·429 (0·090)	0·423 (0·091)	0·393 (0·094)
Performance status	−0·602 (0·104)	−0·641 (0·109)	−0·649 (0·131)
Weight loss	0·200 (0·087)	0·218 (0·089)	0·208 (0·092)
Estimates of σ^2			
Treatment	–	0·071 (0·069)	0·046 (0·184)
Bone met.	–	–	0·129 (0·083)
cAIC	6107	6098	6088*
cAIC$_b$	6107	6088	6071*

cAIC, conditional Akaike information criterion.

the second model (5,1) includes the random treatment effect and the third model (5,2) includes the random treatment and bone metastases effects. The random components are assumed independent, i.e., Σ is diagonal. The results are given in Table 5. Model selection is done using both the conditional AIC and cAIC$_b$. The model-based bootstrap samples are generated under the fitted model (5,2), including the estimated b_i. The favoured model is (5,2), including the random effects for both treatment and presence of bone metastases.

7. Discussion

The analytic derivations and resampling methods used in this paper follow a general approach that can readily be applied to other models, such as nonlinear mixed models. More importantly, the results apply to general distributions for the random effects, although in practice the normal distribution is often used. Our setting is of independent clusters. This is not the most general model for random effects. Much of the theory applies to the general setting. However, care is needed in establishing the asymptotic results, since they require adequate convergence for the random effects.

For generalized linear hierarchical models with only random intercepts, the ρ in (7) turns out to be equal to the effective degrees of freedom obtained in equation (11) of Lu et al. (2007), using their quasi-exact method; such a connection was in fact conjectured in their paper. In addition Spiegelhalter et al. (2002) gave an approximation to their Bayesian measure of model complexity, which is the same as (7). See also Ruppert et al. (2003, Ch. 8 and 11). In the case of linear mixed models, Cui et al. (2010) show that the part of ρ due to the random effects can be interpreted as the ratio of the random effects variance to the total variance. For frailty models based on the hierarchical likelihood, Ha et al. (2007) also proposed an AIC; the models they considered included only random intercepts. We note the close connection between the joint and the hierarchical likelihood. Therneau & Grambsch (2000, Ch. 5) define the degrees of freedom in the penalized partial likelihood formulation of the proportional hazards mixed models. Our derivation provides a theoretical basis both for the information criteria and for the model degrees of freedom under generalized linear and proportional hazards mixed models, that is, as an approximately unbiased estimate of the conditional Akaike information defined in (2) and (12), respectively.

Although the bootstrap has been applied to risk estimation and has been shown to have good finite sample performance for independent and identically distributed data, in our investigation it did not substantially outperform the analytic approximation. This may be because the cluster sizes are typically not large, and the resampling is done within the clusters. Given the wide range of possible implementations allowed by the bootstrap, further improvements are possible, and the topic deserves further exploration. As an alternative to the bootstrap, the numeric methods of Liang et al. (2008) may be extended to our setting for evaluating ρ.

One limitation of the Akaike information is that it is model inconsistent; that is, even in large samples it can select, with nonzero probability, a wrong model which is usually too large in dimension. This is the case when there is a fixed true model. It is now understood (Shao, 1997) that for linear model selection various procedures including the AIC and the BIC fall into three classes: valid if there exist fixed-dimension correct models, valid if no fixed-dimension correct model exists or a compromise of the above two. Our work here on the conditional Akaike information, adjusted to mixed models, does not address these issues. Where the more classical case is concerned and there is a fixed true model, a possible remedy is to include considerations of parsimony. One would choose the smaller model, unless the larger one has an AIC that is significantly better. Our rule of two used in the simulation is a simple step in this direction.

Acknowledgement

Michael Donohue, Ronghui Xu and Florin Vaida were supported in part by grants from the U.S. National Institutes of Health.

Appendix

Setup and conditions for Theorem 1

We assume the following. Given the number of clusters m, let $\theta_m = \text{stack}(\beta, b_1, \ldots, b_m)$, and let $\theta = \text{stack}(\beta, b_1, b_2, \ldots)$ be the corresponding infinite-dimensional parameter as m increases; θ_m contains the first $m+1$ elements, or $p+q$ components, of θ. The true value of θ is θ_0, with first $m+1$ elements $\theta_{0m} = \text{stack}(\beta_0, b_{01}, \ldots, b_{0m})$. For fixed m and cluster size $n = n_1 = \cdots = n_m$, the data y is generated from model (1), with parameter $\theta_m = \theta_{m0}$. Further, θ_m is estimated by $\hat{\theta}_{nm}$, the maximizer of the joint likelihood $l_J(y \mid \theta_m)$. Note that $N = mn$.

Nie (2007) partitions β into $\beta = (\beta_1, \beta_2)$, where the covariates x_{ij1} of β_1 do not have random effects, and the covariates x_{ij2} of β_2 have random effects, i.e., $x_{ij2} = z_{ij}$. Nie (2007) shows that the maximum likelihood estimates of these two components have different rates of convergence as $m, n \to \infty$. For simplicity we will assume that $\beta = \beta_1$; the more general case follows with straightforward modifications.

In the following, unless explicitly stated, all expectations are conditional on θ, and therefore on the random effects b. Let $\Delta(\theta_m) = -E\{l_J(\theta_m; y)\}$, and $\hat{\Delta}(\theta_m) = -l_J(\theta_m; y)$. We do not include indices m, n for Δ and $\hat{\Delta}$ unless necessary. Let Δ', Δ'' denote the derivative and the Hessian of Δ with respect to θ, with a similar notation for $\hat{\Delta}$. Write $l_J = \sum_{i=1}^{m} l_{Ji}$, where l_{Ji} is the component for the ith cluster. From (6) we have that $\Delta'' = U^{\mathrm{T}} W U + \text{diag}(0, D^{-1})$. Let $\Delta''_{\beta\beta}$, $\Delta''_{\beta b_i}$, $\Delta''_{b_i b_j}$ be the corresponding matrix blocks from the Hessian matrix Δ'', and similarly for $\hat{\Delta}''$. We assume that the following conditions hold:

*Condition A*1. The true parameter θ_0 is unique, and is in the interior of a convex closed bounded set $\Theta \subset \mathcal{R}^\infty$ equipped with the sup norm.

*Condition A*2. The fixed effects component $\hat{\beta}$ of θ_{mn} satisfies $\hat{\beta} \to \beta_0$ almost surely as $m, n \to \infty$.

*Condition A*3. The random effects components $\hat{b}_1, \ldots, \hat{b}_m$ of θ_{mn} satisfy $\max_{i=1,\ldots,m} ||\hat{b}_i - b_{0i}|| \to 0$ almost surely as $m, n \to \infty$.

*Condition A*4. For any m, n, the first and second derivatives Δ', $\hat{\Delta}'$ and Δ'', $\hat{\Delta}''$ exist, and are continuous on Θ.

*Condition A*5. The ratio $n/m \to \infty$, as $m, n \to \infty$.

*Condition A*6. As $m, n \to \infty$, $\{\hat{\Delta}_{\beta\beta}(\theta)'' - \Delta_{\beta\beta}(\theta)''\}/N$, $\sum_{i=1}^{m}\{\hat{\Delta}_{b_i b_i}(\theta)'' - \Delta_{b_i b_i}(\theta)''\}/n$, and $\sum_{i=1}^{m}\{\hat{\Delta}_{\beta b_i}(\theta)'' - \Delta_{\beta b_i}(\theta)''\}/(nm^{1/2})$ converge almost surely to 0 uniformly on Θ.

*Condition A*7. As $m, n \to \infty$, $N^{1/2}(\hat{\beta} - \beta_0) \to N(0, v_1)$ in distribution, and $N||\hat{\beta} - \beta_0||_2^2$ is uniformly integrable.

*Condition A*8. As $n \to \infty$, $n^{1/2}(\hat{b}_i - b_{i0}) \to N(0, v_{bi})$ in distribution uniformly over i, and $n||\hat{b}_i - b_{0i}||_2^2$ is uniformly integrable for all i.

*Condition A*9. The quantity $\hat{\Delta}''_{\beta\beta}/N$ is bounded for all θ, m and n, and $\lim_{m,n\to\infty} \hat{\Delta}''_{\beta\beta}/N$ is positive definite.

*Condition A*10. The quantity $\hat{\Delta}''_{b_i b_i}/n$ is bounded for all θ, m and n, and $\lim_{n\to\infty} \hat{\Delta}''_{b_i b_i}/n$ is positive definite.

*Condition A*11. The quantity $\sum_{i=1}^{m} \hat{\Delta}''_{\beta b_i}/(nm^{1/2})$ is bounded for all θ, m and n.

Under the generalized linear mixed model the distributional convergences in Conditions A7 and A8 are established in Nie (2007), assuming Conditions A9–A11 and some additional conditions that are discussed in details in Nie (2007); they can be interpreted as non-collinearity among the covariates under the mixed model, for example. The uniform integrability conditions in Conditions A7 and A8, and the boundedness conditions in Conditions A9–A11 are for the uniform integrability of R_{nm} which leads to $E(R_{nm}) = o(1)$ in the proof below; these are not always easy to establish in general. However Conditions A9–A11 can be directly verified under specific models such as the generalized linear mixed model, since the derivatives can be explicitly calculated.

Proof of Theorem 1. A second order Taylor expansion of $\hat{\Delta}$ yields

$$\begin{aligned}\hat{\Delta}(\hat{\theta}_{nm}) &= \hat{\Delta}(\theta_{0m}) + (\hat{\theta}_{nm} - \theta_{0m})^{\mathrm{T}}\hat{\Delta}'(\theta_{0m}) + \frac{1}{2}(\hat{\theta}_{nm} - \theta_{0m})^{\mathrm{T}}\hat{\Delta}''(\bar{\theta}_{nm})(\hat{\theta}_{nm} - \theta_{0m}) \\ &= \hat{\Delta}(\theta_{0m}) - \frac{1}{2}(\hat{\theta}_{nm} - \theta_{0m})^{\mathrm{T}}\Delta''(\theta_{0m})(\hat{\theta}_{nm} - \theta_{0m}) + R_{nm},\end{aligned} \tag{A1}$$

where

$$R_{nm} = (\hat{\theta}_{nm} - \theta_{0m})^{\mathrm{T}}\{\hat{\Delta}''(\bar{\theta}_{nm}) + \Delta''(\theta_{0m}) - 2\hat{\Delta}''(\tilde{\theta}_{nm})\}(\hat{\theta}_{nm} - \theta_{0m})/2,$$

$\bar{\theta}_{nm}$, $\tilde{\theta}_{nm}$ are measurable functions such that $||\bar{\theta}_{nm} - \theta_{0m}|| \leqslant ||\hat{\theta}_{nm} - \theta_{0m}||$ almost surely, $||\tilde{\theta}_{nm} - \theta_{0m}|| \leqslant ||\hat{\theta}_{nm} - \theta_{0m}||$ almost surely, and $\hat{\Delta}'(\theta_{0m}) = -\hat{\Delta}''(\tilde{\theta}_{nm})(\hat{\theta}_{nm} - \theta_{0m})$. Write $\hat{Q}_{nm}(\theta) = (\hat{\theta}_{nm} - \theta_{0m})^{\mathrm{T}}\hat{\Delta}''(\theta)(\hat{\theta}_{nm} - \theta_{0m})$, $Q_{nm}(\theta) = (\hat{\theta}_{nm} - \theta_{0m})^{\mathrm{T}}\Delta''(\theta)(\hat{\theta}_{nm} - \theta_{0m})$. Then $R_{nm} = \{\hat{Q}(\bar{\theta}_{nm}) + Q(\theta_{0m}) - 2\hat{Q}(\tilde{\theta}_{nm})\}/2$. We have that

$$\begin{aligned}\hat{Q}_{nm}(\theta) &= \begin{pmatrix} \hat{\beta} - \beta_0 \\ \hat{b}_1 - b_{01} \\ \vdots \\ \hat{b}_m - b_{0m} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \frac{\partial^2 l_J}{\partial\beta\partial\beta^{\mathrm{T}}} & \frac{\partial^2 l_{J1}}{\partial\beta\partial b_1'} & \cdots & \frac{\partial^2 l_{Jm}}{\partial\beta\partial b_m'} \\ \frac{\partial^2 l_{J1}}{\partial b_1\partial\beta^{\mathrm{T}}} & \frac{\partial^2 l_{J1}}{\partial b_1\partial b_1'} & & 0 \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 l_{Jm}}{\partial b_m\partial\beta^{\mathrm{T}}} & 0 & \cdots & \frac{\partial^2 l_{Jm}}{\partial b_m\partial b_m'} \end{pmatrix} \begin{pmatrix} \hat{\beta} - \beta_0 \\ \hat{b}_1 - b_{01} \\ \vdots \\ \hat{b}_m - b_{0m} \end{pmatrix} \\ &= (\hat{\beta} - \beta_0)^{\mathrm{T}}\hat{\Delta}''_{\beta\beta}(\hat{\beta} - \beta_0) + 2(\hat{\beta} - \beta_0)^{\mathrm{T}}\sum_{i=1}^{m}\hat{\Delta}''_{\beta b_i}(\hat{b}_i - b_{0i}) + \sum_{i=1}^{m}(\hat{b}_i - b_{0i})^{\mathrm{T}}\hat{\Delta}''_{b_i b_i}(\hat{b}_i - b_{0i}),\end{aligned}$$

with an analogous formula for $Q_{nm}(\theta)$. Under Conditions A2, A3, A6–A8 it is seen that $R_{nm} = o_p(1)$. In addition, Conditions A7–A11 imply that $E(R_{nm}) = o(1)$.

Now, take expectations conditional on b on both sides of equation (A1):

$$\begin{aligned} E(\hat{\Delta}(\hat{\theta}_{nm})) &= E\{\hat{\Delta}(\theta_{0m})\} - \frac{1}{2}E\{(\hat{\theta}_{nm}-\theta_{0m})^{\mathrm{T}}\Delta''(\theta_{0m})(\hat{\theta}_{nm}-\theta_{0m})\} + E(R_{nm}) \\ &= \Delta(\theta_{0m}) - \frac{1}{2}\mathrm{tr}\{\Omega\,\mathrm{var}(\hat{\theta}_{nm})\} + o(1). \end{aligned}$$

In a similar manner we can show that $E\{\Delta(\hat{\theta}_{nm})\} = \Delta(\theta_{0m}) + \frac{1}{2}\mathrm{tr}\{\Omega\mathrm{var}(\hat{\theta}_{nm})\} + o(1)$. Replacing $\Delta(\theta_{0m})$ in the last two equations above we get

$$E\{\Delta(\hat{\theta}_{nm})\} = E\{\hat{\Delta}(\hat{\theta}_{mn})\} - \mathrm{tr}\{\Omega\,\mathrm{var}(\hat{\theta}_{nm})\} + o(1).$$

Finally, it can be seen that $\mathrm{tr}\{\Omega\,\mathrm{var}(\hat{\theta}_{nm})\} = \mathrm{tr}(G\,\Omega^{-1}) + o(1) = \rho + o(1)$. After the simplification of $\log p(\hat{b})$ terms, and taking expectations over b, we get

$$-2E_{(y,b)}E_{y^0|b}[l\{y^0 \mid \hat{\theta}(y)\}] = -2E_y[l\{y \mid \hat{\theta}(y)\}] + 2\rho + o(1).$$

Now we show $p \leqslant \rho \leqslant p+q$. The semipositive definite matrix W admits a square root, so we can write $Z_1 = W^{1/2}Z$, $X_1 = W^{1/2}X$. In (8) we have $Z^{\mathrm{T}}WZ - Z^{\mathrm{T}}WX(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}WZ = Z_1^{\mathrm{T}}(I-P)Z_1 = Z_2^{\mathrm{T}}Z_2$, where $P = X_1(X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}$ and therefore $I-P$ are both projection matrices, and $Z_2 = (I-P)Z_1$. Then $\rho = p+q-\mathrm{tr}\{(Z_2^{\mathrm{T}}Z_2 + D^{-1})^{-1}D^{-1}\} = p+q-\mathrm{tr}\{(I_q + Z_3^{\mathrm{T}}Z_3)^{-1}\} = p+q-\sum_{i=1}^{q}(1+u_i)^{-1}$, where $Z_3 = Z_2D^{1/2}$, and $0 \leqslant u_1 \leqslant \cdots \leqslant u_q$ are the eigenvalues of $Z_3^{\mathrm{T}}Z_3$. It follows that $p \leqslant \rho \leqslant p+q$. □

Details of Theorem 2 *and Proposition* 2

Let $d_1, \ldots, d_K$ be the number of failures at the distinct failure times $t_1 < \cdots < t_K$. Further, let V be the $N \times K$ indicator matrix with element (ij, k) equal to $I(Y_{ij} \geqslant t_k)$, and $W_3 = \mathrm{diag}\{\exp(\eta_{ij}); i, j\}$. The matrix W^* in (15) is given by $W^* = W_1 - W_2$, where $W_1 = \mathrm{diag}(\hat{\mu})$ and $W_2 = W_3V\mathrm{diag}(\hat{\lambda}_{0k}^2/d_k; k = 1, \ldots, K)V'W_3$.

The conditions for Theorem 2 are identical to conditions A1–A11 for Theorem 1, where $\Delta(\theta)$ and $\hat{\Delta}(\theta)$ are defined with respect to the joint profile loglikelihood (13). The proof is then analogous to the proof of Theorem 1.

Although Conditions A9–A11 can still be directly verified under the proportional hazards mixed model, unlike the generalized linear mixed model, the distributional results as in Conditions A7 and A8 have only been established for parametric baseline hazard functions (Feng et al., 2009), with simulation results to suggest that the same likely hold for the nonparametric baseline hazard function. The more standard asymptotic results with only $m \to \infty$ and n_i bounded were given in Gamst et al. (2009).

Proof of Proposition 2. For simplicity, let the N data points be counted by $l = 1, \ldots, N$, and put $r = p+q$. For $l = 1, \ldots, N$ define K_l such that $Y_l = t_{K_l}$ if $\delta_l = 1$, and $K_l = \max\{k : t_k \leqslant Y_l\}$, if $\delta_l = 0$. Using the Poisson formulation of the proportional hazards mixed model, the effective degrees of freedom are $\rho_P = \mathrm{tr}\{\tilde{U}^{\mathrm{T}}\tilde{W}\tilde{U}(\tilde{U}^{\mathrm{T}}\tilde{W}\tilde{U} + \tilde{A})^{-1}\}$, where $\tilde{U} = (\tilde{U}_1, \tilde{U}_2)$, $\tilde{U}_1 = \mathrm{stack}(\tilde{U}_{11}, \ldots, \tilde{U}_{1N})$, $\tilde{U}_2 = \mathrm{stack}(\tilde{U}_{21}, \ldots, \tilde{U}_{2N})$, and $\tilde{W} = \mathrm{diag}(\tilde{W}_1, \ldots, \tilde{W}_N)$; $\tilde{U}_{1l}$ is $K_l \times K$, $\tilde{U}_{2l}$ is $K_l \times r$ and $\tilde{W}_l$ is $K_l \times K_l$, with $\tilde{U}_{1l} = (I_{K_l}, 0)$, $\tilde{U}_{2l} = 1_{K_l}u_l$, $\tilde{W}_l = \exp(\eta_l)\,\mathrm{diag}(\lambda_k; k = 1 \ldots K_l)$, 1_k is a k-vector with 1 of each element, and u_l is the lth row of U, $\tilde{A} = \mathrm{diag}(0, A)$; $\tilde{W}$ is computed at $\hat{\lambda}$. Since $1_{K_l}^{\mathrm{T}}\tilde{W}_l1_{K_l} = \exp(\eta_l)$ $\Lambda(Y_l)$, $\tilde{U}_2^{\mathrm{T}}\tilde{W}\tilde{U}_2 = U^{\mathrm{T}}W_1U$. Also, $T = \tilde{U}_1^{\mathrm{T}}\tilde{W}\tilde{U}_1 = \sum_{l=1}^{N}\mathrm{diag}(\tilde{W}_l, 0) = \mathrm{diag}(\alpha_k\lambda_k, k = 1 \ldots K)$, where $\alpha_k = \sum_{l=1}^{N}\exp(\eta_l)I(Y_l \geqslant t_k)$. Put $J = \mathrm{diag}(1_{K_l}, l = 1, \ldots, N)$. Note that $\tilde{U}_2 = JU$. The $N \times K$ matrix is $\Gamma = J^{\mathrm{T}}\tilde{W}\tilde{U}_1$ with generic element $\gamma_{lk} = \exp(\eta_l)I[t_k \leqslant Y_l]\,\lambda_k$. So $\Gamma = W_3V\mathrm{diag}(\lambda)$.

Using the derivation for ρ_P as in (8), it follows that

$$\begin{aligned}\rho_P &= K + r - \mathrm{tr}[\{\tilde{U}_2^{\mathrm{T}}\tilde{W}\tilde{U}_2 - \tilde{U}_2^{\mathrm{T}}\tilde{W}\tilde{U}_1(\tilde{U}_1^{\mathrm{T}}\tilde{W}\tilde{U}_1)^{-1}\tilde{U}_1^{\mathrm{T}}\tilde{W}\tilde{U}_2 + A\}^{-1}A]\\ &= K + r - \mathrm{tr}[\{U^{\mathrm{T}}W_1U - U^{\mathrm{T}}(J^{\mathrm{T}}\tilde{W}\tilde{U}_1T^{-1}\tilde{U}_1^{\mathrm{T}}\tilde{W}J)U + A\}^{-1}A]\\ &= K + r - \mathrm{tr}[\{U^{\mathrm{T}}W_1U - U^{\mathrm{T}}W_3V\mathrm{diag}(\lambda)\,T^{-1}\,\mathrm{diag}(\lambda)V'W_3U + A\}^{-1}A]\\ &= K + (p+q) - \mathrm{tr}[\{U^{\mathrm{T}}(W_1 - W_2)U + A\}^{-1}A] = K + \rho,\end{aligned}$$

after noting that $\mathrm{diag}(\lambda)\,T^{-1}\mathrm{diag}(\lambda)$, computed at $\hat{\lambda}$, has diagonal elements $\hat{\lambda}_k/\alpha_k = \hat{\lambda}_k^2/d_k$. It is well-known in the standard Cox model that $l(y \mid \theta) = pl(y \mid \theta) + a_2$, where a_2 depends on data. Therefore, for $a_1 = K$ and $a = 2(a_1 - a_2)$ we have $\textsc{caic}_P = \textsc{caic} + a$, which completes the proof. □

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Breakthroughs in Statistics (1992)*, vol. 1. pp. 610–24. New York: Springer.

Breslow, N. & Clayton, D. (1993). Approximate inference in generalized linear mixed models. *J. Am. Statist. Assoc.* **88**, 9–25.

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

Cavanaugh, J. E. & Shumway, R. H. (1997). A bootstrap variant of AIC for state-space model selection. *Statist. Sinica* **7**, 473–96.

Claeskens, G. & Hjort, N. L. (2003). Focused information criterion (with discussion). *J. Am. Statist. Assoc.* **98**, 900–45.

Claeskens, G. & Hjort, N. L. (2008). *Model Selection and Model Averaging*. New York: Cambridge University Press.

Cui, Y., Hodges, J. S., Kong, X. & Carlin, B. P. (2010). Partitioning degrees of freedom in hierarchical and other richly-parameterized models. *Technometrics* **52**, 124–36.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Statist. Assoc.* **78**, 316–31.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Am. Statist. Assoc.* **81**, 461–70.

Feng, S., Nie, L. & Wolfe, R. A. (2009). Laplace's approximation for relative risk frailty models. *Lifetime Data Anal.* **15**, 343–56.

Fitzmaurice, G. M., Laird, N. M. & H, W. J. (2004). *Applied Longitudinal Analysis*. Hoboken, NJ: Wiley.

Gamst, A., Donohue, M. & Xu, R. (2009). Asymptotic properties and empirical evaluation of the *npmle* in the proportional hazards mixed-effects model. *Statist. Sinica* **19**, 997–1011.

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

Greenberg, E. R., Baron, J. A., Stukel, T. A., Stevens, M. M., Mandel, J. S., Spencer, S. K., Elias, P. M., Lowe, N., Nierenberg, D. W., Bayrd, G., et al. (1990). A clinical trial of beta carotene to prevent basal-cell and squamous-cell cancers of the skin. *New Engl. J. Med.* **323**, 789–95.

Gustafson, P. (1997). Large hierarchical Bayesian analysis of multivariate survival data. *Biometrics* **53**, 230–42.

Ha, I. D. & Lee, Y. (2003). Estimating frailty models via Poisson hierarchical generalized linear models. *J. Comp. Graph. Statist.* **12**, 663–81.

Ha, I. D., Lee, Y. & MacKenzie, G. (2007). Model selection for multi-component frailty models. *Statist. Med.* **26**, 4790–807.

Harville, D. A. (1996). *Matrix Algebra From a Statistician's Perspective*. New York: John Wiley.

Hjort, N. L. & Claeskens, G. (2006). Focused information criterion and model averaging for the Cox hazard regression model. *J. Am. Statist. Assoc.* **101**, 1449–64.

Hodges, J. & Sargent, D. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika* **88**, 367–79.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. New York: Springer.

Jiang, J. (2001). Maximum posterior estimation of random effects in generalized linear mixed models. *Statist. Sinica* **11**, 97–120.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer.

Kauermann, G., Xu, R. & Vaida, F. (2008). Stacked Laplace-EM algorithm for duration models with time-varying and random effects. *Comp. Statist. Data Anal.* **52**, 2514–28.

Lee, Y., Nelder, J. & Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via h-likelihood*. Boca Raton, FL: Chapman & Hall/CRC.

Lee, Y. & Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc.* B **58**, 619–78.

Liang, H., Wu, H. L. & Zou, G. H. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika* **95**, 773–8.

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

Lu, H., Hodges, J. S. & Carlin, B. P. (2007). Measuring the complexity of generalized linear hierarchical models. *Can. J. Statist.* **35**, 69–87.

Ma, R., Krewski, D. & Burnett, R. T. (2003). Random effects Cox model: a Poisson modelling approach. *Biometrika* **90**, 157–69.

McCulloch, C. E., Searle, S. R. & Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*. New York: Wiley.

Nie, L. (2007). Convergence rate of mle in generalized linear and nonlinear mixed-effects models: Theory and applications. *J. Statist. Plan. Infer.* **137**, 1787–804.

Pan, W. (1999). Bootstrapping likelihood for model selection with small samples. *J. Comp. Graph. Statist.* **8**, 687–98.

Ripatti, S., Larsen, K. & Palmgren, J. (2002). Maximum likelihood inference for multivariate frailty models using an automated Monte Carlo EM algorithm. *Lifetime Data Anal.* **8**, 349–60.

Ripatti, S. & Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* **56**, 1016–22.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Sargent, D. J. (1998). A general framework for random effects survival analysis in the Cox proportional hazards setting. *Biometrics* **54**, 1486–97.

Shang, J. & Cavanaugh, J. E. (2008). Bootstrap variants of the Akaike information criterion for mixed model selection. *Comp. Statist. Data Anal.* **52**, 2004–21.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7**, 221–64.

Shibata, R. (1997). Bootstrap estimate of Kullback–Leibler information for model selection. *Statist. Sinica* **7**, 375–94.

Spiegelhalter, D. J., Best, N. G., P, C. B. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Statist. Soc.* B **64**, 583–639.

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

Vaida, F. & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92**, 351–70.

Vaida, F. & Xu, R. (2000). Proportional hazards model with random effects. *Statist. Med.* **19**, 3309–24.

Vonesh, E. F. (1996). A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika* **83**, 447–52.

Wager, C., Vaida, F. & Kauermann, G. (2007). Model selection for penalized spline smoothing using Akaike information criteria. *Aust. New Zeal. J. Statist.* **49**, 173–90.

Whitehead, J. (1980). Fitting Cox's regression model to survival data using GLIM. *J. R. Statist. Soc.* C **29**, 268–75.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika* **80**, 791–5.

Xu, R. & Gamst, A. (2008). Risk estimation. In *High Dimensional Data Analysis in Oncology*. pp. 63–88. New York: Springer.

Xu, R., Vaida, F. & Harrington, D. (2009). Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models. *Statist. Sinica* **19**, 819–42.

Yafune, A., Funatogawa, T. & Ishiguro, M. (2005). Extended information criterion approach for linear mixed effects models under restricted maximum likelihood estimation. *Statist. Med.* **24**, 3417–29.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Am. Statist. Assoc.* **93**, 120–31.

[*Received April* 2009. *Revised March* 2011]