Johns Hopkins University, Dept. of Biostatistics Working Papers

11-24-2009

On the Behaviour of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models

Sonja Greven
Johns Hopkins University, sgreven@jhsph.edu

Thomas Kneib
Carl von Ossietzky University Oldenburg, thomas.kneib@uni-oldenburg.de

Suggested Citation

Greven, Sonja and Kneib, Thomas, "On the Behaviour of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models" (November 2009). *Johns Hopkins University, Dept. of Biostatistics Working Papers.* Working Paper 202.
http://biostats.bepress.com/jhubiostat/paper202

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.
Copyright © 2011 by the authors

(2009), **xx**, x, *pp*. 1–17

On the Behaviour of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models

BY SONJA GREVEN

Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205, USA,
sgreven@jhsph.edu

THOMAS KNEIB

Department of Mathematics, Carl von Ossietzky University Oldenburg, 26111 Oldenburg, Germany,
thomas.kneib@uni-oldenburg.de

SUMMARY

In linear mixed models, model selection frequently includes the selection of random effects. Two versions of the Akaike information criterion (AIC) have been used, based either on the marginal or on the conditional distribution. We show that the marginal AIC is no longer an asymptotically unbiased estimator of the Akaike information, and in fact favours smaller models without random effects. For the conditional AIC, we show that ignoring estimation uncertainty in the random effects covariance matrix, as is common practice, induces a bias that leads to the selection of any random effect not predicted to be exactly zero. We derive an analytic representation of a corrected version of the conditional AIC, which avoids the high computational cost and imprecision of available numerical approximations. An implementation in an R package is provided. All theoretical results are illustrated in simulation studies, and their impact in practice is investigated in an analysis of childhood malnutrition in Zambia.

Some key words: information criterion, Kullback-Leibler information, model selection, penalized splines, random effect, variance component

Note: This technical report is a reworked and updated version of Johns Hopkins University, Department of Biostatistics Working Papers, Paper 179 (2009), including new results.

1. INTRODUCTION

Linear mixed models are a powerful inferential tool used in a wide range of statistical areas from longitudinal data analysis (Laird & Ware, 1982) to penalized spline smoothing (Ruppert et al., 2003), to functional data analysis (Di et al., 2008). They offer flexibility in modelling and computationally attractive implementations of complex models for large data sets. The resulting flexibility and complexity of models make the question of model choice increasingly important. This includes the selection of random effects, such as those modelling heterogeneity between subjects, or deviations of a curve from linearity.

We focus on properties of the Akaike information criterion (AIC, Akaike, 1973) for the selection of random effects. The AIC has been argued to be better suited to model selection than hypothesis testing, is not limited to nested models, and has an approximate justification even

Hosted by The Berkeley Electronic Press

when the candidate models do not contain the true model (Burnham & Anderson, 2002, pp. 36-37, 65). While tests for random effects or their variances have gained a lot of interest in recent years (Stram & Lee, 1994; Crainiceanu & Ruppert, 2004; Molenberghs & Verbeke, 2007; Greven et al., 2008; Scheipl et al., 2008; Giampaoli & Singer, 2009) due to the violation of typical regularity conditions in linear mixed models, potential implications for information criteria such as the AIC are often not explicitly addressed (e.g. Robert-Granié et al., 2004; Wager et al., 2007).

In mixed models, an AIC based on the marginal likelihood is typically used (mAIC), which is returned by standard statistical software. Vaida & Blanchard (2005) propose an AIC derived from the conditional model formulation (cAIC), with the effective degrees of freedom accounting for shrinkage in the random effects. In practice, the authors recommend using a plug-in estimator for the unknown random effects covariance matrix, arguing that the effect is negligible asymptotically. Liang et al. (2008) propose a corrected cAIC that accounts for the estimation of the variance parameters. However, for a sample size of n, they require n or even $2n$ additional model fits to numerically approximate their cAIC. The use of the corrected cAIC thus is computationally prohibitive in settings with larger sample sizes and number of potential models. For example, our application on childhood malnutrition with 1600 observations and 64 potential models would require an estimated 110 days computation time. This makes the approximation proposed by Vaida & Blanchard (2005) tempting, and their version of the cAIC indeed seems to be used in practice.

In this paper, we study the theoretical properties of both mAIC and cAIC for the selection of random effects in linear mixed models. We find that the mAIC is a biased estimator of the Akaike information due to the non-open parameter space and lacking independence between observations in linear mixed models. In consequence, it favours smaller models without random effects. For the cAIC, we show that ignoring the uncertainty in the estimate of the random effects covariance matrix induces a very specific bias with an interesting effect on model selection behaviour: the corresponding cAIC always selects an additional random effect into the model unless that random effect is predicted to be exactly zero, in which case there is a tie. This behaviour is independent of the sample size and does not disappear asymptotically. As accounting for the estimation uncertainty in the covariance matrix is crucial, we derive an analytic representation of the corrected version of the cAIC proposed by Liang et al. (2008). This formulation avoids the high computational cost and imprecision inherent in available numerical approximations.

All theoretical results are illustrated in simulation studies, and their impact in practice is investigated in an analysis of childhood malnutrition in Zambia. Outlines of proofs are given in the appendix. Detailed proofs, extended simulation and application results as well as an R package implementing the corrected cAIC are available in a web appendix at `http://www.biostat.jhsph.edu/~sgreven/research/appendix_AIC.zip`.

2. THE AIC IN THE LINEAR MIXED MODEL

2·1. *The Linear Mixed Model*

In the following, we consider the linear mixed model

$$y = X\beta + Zb + \varepsilon, \tag{1}$$

where X and Z are known design matrices of full column ranks p and r, β contains fixed parameters, and b and ε are assumed to be independent and normally distributed,

$$\begin{pmatrix} b \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D & 0 \\ 0 & \sigma^2 I_n \end{pmatrix} \right),$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

I_n being the $n \times n$ identity matrix. Let $D_* = \sigma^{-2} D$, and define the covariance matrix

$$V := \text{cov}(y) = \sigma^2 I_n + ZDZ^T = \sigma^2 (I_n + ZD_*Z^T) =: \sigma^2 V_*.$$

Denote by θ_* the q parameters in D_*, and let $\theta = (\sigma^2, \theta_*)$ contain all variance parameters. We suppress dependence of $V = V(\theta), V_* = V_*(\theta_*)$, etc. on the parameters when no confusion can arise. We use hat-notation for estimated quantities, such as $\widehat{V} = V(\widehat{\theta})$. We interchangeably use the notation $\widehat{\theta}$ or $\widehat{\theta}(y)$ etc. when emphasizing the dependance on the data y.

Inference in model (1) is usually based on the implied marginal likelihood, integrating over the random effects. For a given θ, the fixed effects β and the random effects b can be estimated and predicted by the best linear unbiased estimator and predictor, respectively:

$$\begin{aligned} \widehat{\beta} &= \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} y = \left(X^T V_*^{-1} X\right)^{-1} X^T V_*^{-1} y, \\ \widehat{b} &= DZ^T V^{-1}(y - X\widehat{\beta}) = D_* Z^T V_*^{-1}(y - X\widehat{\beta}), \end{aligned} \tag{2}$$

where $\widehat{\beta}$ is also the maximum likelihood (ML) estimator. The profile log-likelihood for all variance parameters θ, profiling out over β, is, up to a constant,

$$l(\theta) = \log f(y \mid \theta, \widehat{\beta}(\theta)) = -\frac{1}{2}\log\{\det(V)\} - \frac{1}{2}(y - X\widehat{\beta})^T V^{-1}(y - X\widehat{\beta}). \tag{3}$$

The corresponding restricted log-likelihood for θ is up to a constant (Harville, 1974)

$$\ell(\theta) = \log f(A^T y \mid \theta) = -\frac{1}{2}\log\{\det(V)\} - \frac{1}{2}\log\{\det(X^T V^{-1} X)\} - \frac{1}{2}(y - X\widehat{\beta})^T V^{-1}(y - X\widehat{\beta}), \tag{4}$$

where $A^T y$ are $n - p$ linearly independent error contrasts with $\text{E}(A^T y) = 0$.

In our examples, we focus on two special linear mixed models. One of the simplest linear mixed models is the random intercept model, used to account for variability between sampling units such as subjects or clusters. This model is written on the observational level as

$$y_{ij} = x_{ij}^T \beta + b_i + \varepsilon_{ij}, \quad j = 1, \ldots, J_i,\ i = 1, \ldots, I, \tag{5}$$

with I the number of clusters and J_i the number of observations from cluster i. Appropriate stacking gives a matrix-vector formulation as in (1).

The second case is penalized spline smoothing. Consider univariate smoothing

$$y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{6}$$

where $m(\cdot)$ is an unknown smooth function. $m(\cdot)$ is modeled using splines, such as truncated polynomials

$$m(x) = \sum_{j=0}^{d} \beta_j x^j + \sum_{j=1}^{K} b_j (x - \kappa_j)_+^d$$

for some $d \in \mathbb{N}_0$ and $K \in \mathbb{N}$, where $\kappa_1 < \cdots < \kappa_K$ are K knots, and $(u)_+^d = u^d$ if $u > 0$ and $= 0$ else. To avoid overfitting and knot-dependence, and to impose smoothness on the estimated function, one considers the penalized least squares criterion

$$\min_{\beta, b} \|y - X\beta - Zb\|^2 + \frac{1}{\lambda} b^T b, \tag{7}$$

where $\beta = (\beta_0, \ldots, \beta_d)$, $b = (b_1, \ldots, b_K)$, and X and Z contain the rows $(1, x_i, \ldots, x_i^d)$ and $((x_i - \kappa_1)_+^d, \ldots, (x_i - \kappa_K)_+^d)$, $i = 1, \ldots, n$, respectively. This formulation penalizes deviations from a dth degree polynomial, e.g. linearity if $d = 1$.

Hosted by The Berkeley Electronic Press

The smoothing parameter λ controls the trade-off between fit to the data and smoothness. As (7) is equivalent to determination of the best linear unbiased estimator and predictor for β and b in the linear mixed model (1) with $D = \tau^2 I_K$ and fixed variance $\tau^2 = \lambda\sigma^2$ (Brumback et al., 1999; Ruppert et al., 2003), the mixed model formulation can be used to estimate λ as τ^2/σ^2. In this framework, fixed effects model the subspace of polynomials of degree d, while random effects model any deviation. In our examples, we use a similar mixed model representation for B-splines with a difference penalty (Eilers & Marx, 1996; Fahrmeir et al., 2004).

2·2. *The Akaike Information Criterion*

We now recapitulate the definition of the AIC. Suppose that $y = (y_1, \ldots, y_n)$ is a vector of observations, generated from a true underlying distribution with joint density $g(\cdot)$, and that $f_\psi(\cdot) = f(\cdot \mid \psi)$ is a family of approximating models with unknown parameters $\psi \in \Psi$. The Kullback-Leibler divergence is defined as

$$K(f_\psi, g) = \int \log\left(\frac{g(z)}{f_\psi(z)}\right) g(z)dz = \mathrm{E}_z[\log\{g(z)\} - \log\{f_\psi(z)\}], \tag{8}$$

where E_z denotes the expectation with regard to the distribution of another realization z. $K(f_\psi, g)$ can be viewed as a measure of distance between $g(\cdot)$ and $f_\psi(\cdot)$.

In practice, ψ is estimated by the ML estimator $\widehat{\psi}(y)$ based on data y independent of z, and one would like to minimize $\mathrm{E}_y\{K(f_{\widehat{\psi}(y)}, g)\}$, or equivalently $-2\mathrm{E}_y(\mathrm{E}_z[\log\{f_{\widehat{\psi}(y)}(z)\}])$. This Akaike information, or twice the expected relative Kullback-Leibler distance, is a predictive quantity, depending on independent replications z and y. The maximized log-likelihood $\log\{f_{\widehat{\psi}(y)}(y)\}$ can be used for estimation of it, but is biased, as it only depends on y. Denote by ψ_K the parameter vector which minimizes the Kullback-Leibler distance in (8). Then, an unbiased estimator is

$$-2\log\{f_{\widehat{\psi}(y)}(y)\} + 2\mathrm{E}_y[\log\{f_{\widehat{\psi}(y)}(y)\} - \log\{f_{\psi_K}(y)\}] + \mathrm{E}_y(2\mathrm{E}_z[\log\{f_{\psi_K}(z)\} - \log\{f_{\widehat{\psi}(y)}(z)\}]).$$

In standard settings, certain regularity conditions are fulfilled, including the following. First, the parameter space for ψ, up to a change of coordinates, is $\Psi = \mathbb{R}^k$, with k the number of estimable parameters in ψ. Second, observations $y_1, \ldots, y_n$ are independent and identically distributed. If one further assumes that $f_{\psi_K}(\cdot) = g(\cdot)$, such that consistency ensures convergence of $\widehat{\psi}(y)$ to ψ_K, standard asymptotic theory gives an asymptotic χ^2_k distribution for both $2\mathrm{E}_z[\log\{f_{\psi_K}(z)\} - \log\{f_{\widehat{\psi}(y)}(z)\}]$ and $2[\log\{f_{\widehat{\psi}(y)}(y)\} - \log\{f_{\psi_K}(y)\}]$. Then,

$$\mathrm{AIC} = -2\ \log\{f(y \mid \widehat{\psi}(y))\}\ + 2k$$

is asymptotically unbiased for the Akaike information. Minimizing the AIC over a set of possible, nested or non-nested, models can thus be seen as minimizing the average distance of an approximating model to the underlying truth.

2·3. *The* AIC *in the Linear Mixed Model*

In the linear mixed model, we focus on model selection for the random effects b. Examples include the selection of a random intercept in (5), or of a random effect modelling deviations of $m(\cdot)$ from a low-order polynomial in (6). For ease of presentation, we focus on the asymptotic versions of the marginal and conventional conditional AIC criteria. Analogous results hold straightforwardly for the finite sample versions (Sugiura, 1978; Vaida & Blanchard, 2005).

For extension of the AIC to the linear mixed model, two different approaches exist. The first approach uses the marginal likelihood arising from the marginal model $y \sim N(X\beta, V)$. The

http://biostats.bepress.com/jhubiostat/paper202

number of parameters in this model is $p+q+1$. The marginal AIC (mAIC) is then defined as

$$m\text{AIC} = -2\log\{f(y \mid \widehat{\beta}, \widehat{\theta})\} + 2(p+q+1), \tag{9}$$

where $f(y \mid \widehat{\beta}, \widehat{\theta})$ is the maximized marginal likelihood. For restricted maximum likelihood (REML) estimation, the maximized restricted likelihood $f(A^T y \mid \widehat{\theta})$ is used, and the number of parameters is $q+1$.

Use of the marginal likelihood implies that in the definition of the AIC, the two independent replications z and y arising from the true underlying distribution do not share the same random effects. This is appropriate, for example, in a longitudinal study with subject-specific random effects, where interest is in the fixed (population) effects. However, the marginal AIC is typically used for model selection in all contexts in the linear mixed model, as it is routinely returned by statistical software, such as R `lme()` or SAS `PROC MIXED`.

Vaida & Blanchard (2005) argue that a second approach based on the likelihood for the conditional model $y \mid b \sim N(X\beta + Zb, \sigma^2 I_n)$ is more appropriate when the focus is on random effects shared by z and y. For example, in penalized spline smoothing, random effects are used as a tool to model the non-linear part of an underlying smooth function common to z and y.

In this setting, the Akaike information is replaced by the conditional Akaike information,

$$c\text{AI} = -2\text{E}_{y,b}(\text{E}_{z|b}[\log\{f(z \mid \widehat{\theta}(y), \widehat{b}(y))\}]) = -\int 2\log\{f(z \mid \widehat{\theta}(y), \widehat{b}(y))\} g(z \mid b) g(y,b) dz dy db,$$

where $g(y,b) = g(y \mid b)g(b)$ is the joint distribution of y and b.

For the case where D_* and thus θ_* is known, Vaida & Blanchard (2005) show that an asymptotically unbiased estimator of cAI is their conditional AIC (cAIC),

$$c\text{AIC} = -2\ \log f(y \mid \widehat{\beta}, \widehat{b}, \widehat{\theta})\ + 2(\rho+1), \quad \text{where}$$
$$\log f(y \mid \widehat{\beta}, \widehat{b}, \widehat{\theta}) = -\frac{1}{2}n\log(2\pi) - \frac{1}{2}n\log(\widehat{\sigma}^2) - \frac{1}{2\widehat{\sigma}^2}(y - X\widehat{\beta} - Z\widehat{b})^T(y - X\widehat{\beta} - Z\widehat{b})$$

is the conditional log-likelihood for y, conditioning on b as well as on β and θ, evaluated at the estimated or predicted quantities $(\widehat{\beta}, \widehat{b}, \widehat{\theta})$ based on ML or REML estimation, and

$$\rho = \text{tr}\left\{ \begin{pmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z + D_*^{-1} \end{pmatrix}^{-1} \begin{pmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z \end{pmatrix} \right\} \tag{10}$$

is the trace of the hat matrix projecting y onto $\widehat{y} = X\widehat{\beta} + Z\widehat{b}$. Vaida & Blanchard (2005) note the connection of the effective degrees of freedom ρ, which lie between those of a linear model without b, and those of a linear model with fixed effects b, to the effective degrees of freedom known from smoothing (p. 53, Hastie & Tibshirani, 1990; Hodges & Sargent, 2001).

Vaida and Blanchard assume that D_* and thus θ_* is known. In practice, they recommend using the cAIC with estimated D_* when it is unknown, argueing that the difference between estimated $\widehat{\rho}$ and true ρ is negligible asymptotically. We call this the conventional cAIC in the following.

Liang et al. (2008) propose a corrected cAIC that takes into account the estimation of θ. For known error variance σ^2, they replace the effective degrees of freedom by

$$\Phi_0 = \sum_{i=1}^{n} \frac{\partial \widehat{y}_i}{\partial y_i} = \text{tr}\left(\frac{\partial \widehat{y}}{\partial y}\right). \tag{11}$$

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

For known θ_*, Φ_0 reduces to the effective degrees of freedom ρ. In an accompanying technical report, they extend the idea to the case when σ^2 has to be estimated. Then,

$$\Phi_1 = \frac{\widetilde{\sigma}^2}{\widehat{\sigma}^2} \operatorname{tr}\left(\frac{\partial \widehat{y}}{\partial y}\right) + \widetilde{\sigma}^2 (\widehat{y} - y)^T \frac{\partial \widehat{\sigma}^{-2}}{\partial y} + \frac{1}{2} \widetilde{\sigma}^4 \operatorname{tr}\left(\frac{\partial^2 \widehat{\sigma}^{-2}}{\partial y \partial y^T}\right), \qquad (12)$$

is substituted for the total number of parameters $\rho + 1$. $\widetilde{\sigma}^2$ is the unknown true error variance and has to be replaced by an estimate, such as $\widehat{\sigma}^2$ based on ML or REML estimation. (11) and (12) involve derivatives of estimated or predicted quantities with respect to the data, for which Liang et al. (2008) do not provide closed form expressions. They propose numerical approximations based on small disturbances of the observed data. However, the implementation requires n and $2n$ additional model fits for evaluating (11) and (12), respectively. As a consequence, the evaluation of the corrected cAIC quickly becomes prohibitive for moderate sample sizes n. In their simulation, Liang et al. (2008) conclude that the estimated effective degrees of freedom are similar between corrected and conventional cAIC, which matches Vaida and Blanchard's recommendation to use the estimated effective degrees of freedom $\widehat{\rho}$.

3. THE MARGINAL AIC

In this section, we show that the marginal AIC is no longer an unbiased estimator of the Akaike information under the marginal model. This is due to the fact that a) the parameter space for the marginal model is not a transformation of $\mathbb{R}^k$ due to the restrictions on the variance parameters, and b) observations in the linear mixed model are not independent due to the correlation induced by the random effects. The resulting bias in the marginal AIC is closely related to results for the distribution of (restricted) likelihood ratio tests for variance components in linear mixed models (Crainiceanu & Ruppert, 2004).

We focus on the model with one unknown variance component and maximum likelihood estimation for simplicity. It is straightforward to see that analogous arguments hold for more complex models and for the restricted log-likelihood. We have the following result.

THEOREM 1. *Consider the linear mixed model* (1) *with one unknown random effects variance component, $D = \tau^2 \Sigma$ with Σ known. Then, the marginal Akaike Information Criterion (mAIC) defined in* (9) *is positively biased for the Akaike information,*

$$E_y(m\mathrm{AIC}) > -2E_y(E_z[\log\{f_{\widehat{\psi}(y)}(z)\}]),$$

where $f_\psi(\cdot) = f(\cdot \mid \psi)$ denotes the marginal likelihood with $\psi = (\beta^T, \sigma^2, \lambda)^T$, $\lambda = \tau^2/\sigma^2$. The bias is dependent on the true unknown τ^2, and does not vanish asymptotically if $\tau^2 = 0$.

All proof outlines are in the appendix, and detailed proofs can be found in the web appendix. Compared to an unbiased criterion, the mAIC favours smaller models excluding random effects.

Using the mAIC to compare two nested models with $\tau^2 = 0$ (linear model, M_1) and $\tau^2 \geq 0$ (linear mixed model, M_2) in the notation of Theorem 1, is closely related to testing for a random effects variance. The mAIC selects the larger model M_2 iff

$$\begin{aligned} &-2\log\{f_{\widehat{\psi}(y)}(y)\} + 2(p+2) < -2\log\{f_{\bar{\psi}(y)}(y)\} + 2(p+1) \\ \Leftrightarrow\ &2\log\{f_{\widehat{\psi}(y)}(y)\} - 2\log\{f_{\bar{\psi}(y)}(y)\} > 2, \end{aligned}$$

where bar-notation indicates estimation under the restriction $\lambda = \tau^2/\sigma^2 = 0$. Thus, a comparison of the mAIC is equivalent to a likelihood ratio test with the critical value 2. In standard cases, when the log-likelihood ratio is asymptotically χ^2_1-distributed, the nominal level of such a test

COBRA A BEPRESS REPOSITORY Collection of Biostatistics Research Archive

http://biostats.bepress.com/jhubiostat/paper202

would be approximately 0.157. However, for variance component testing, where the distribution has a point mass at zero, such a nominal level can be far smaller than 0.05.

Our result is related to findings by Hughes & King (2003), who propose a one-sided AIC for settings with inequality-constrained parameters. However, they assume independent and identically distributed responses, and their result is thus not applicable in most linear mixed models. Even in the independent and identically distributed case, their AIC is only unbiased if all inequality constrained parameters lie on the boundary of the parameter space.

4. The Conditional AIC

4·1. *The Conventional c*AIC

We now investigate the theoretical properties of the conventional cAIC, which substitutes $\widehat{D_*} = D(\widehat{\theta_*})$ for the unknown D_* in the calculation of the effective degrees of freedom ρ, and does not account for the resulting estimation uncertainty. For ease of presentation, we focus on the case of one unknown variance component, $D = \tau^2\Sigma$ with Σ known. The following theorem characterizes the behaviour of the conventional cAIC for the selection of random effects.

THEOREM 2. *Consider the two models*

$$M_1 : y = X\beta + \varepsilon, \quad M_2 : y = X\beta + Zb + \varepsilon, \quad (b, \varepsilon) \sim \mathcal{N}(0, \mathrm{diag}(\tau^2\Sigma, \sigma^2 I_n)),$$

*with known Σ, but unknown τ^2. For the conventional c*AIC *with estimated $\widehat{\rho}$,*

$$\widehat{\tau}^2 > 0 \;\Leftrightarrow\; c\mathrm{AIC}(M_1) > c\mathrm{AIC}(M_2) \quad \textit{and} \quad \widehat{\tau}^2 = 0 \;\Leftrightarrow\; c\mathrm{AIC}(M_1) = c\mathrm{AIC}(M_2).$$

Thus, the conventional cAIC always chooses the inclusion of the random effect b into the model, unless b is predicted to be exactly zero ($\widehat{\tau}^2 = 0$), in which case the cAIC does not distinguish between the two models. This is in contrast with the AIC, say in the linear model, where a regression coefficient estimated to be zero would still be counted in the number of estimable parameters k. The conventional cAIC does not distinguish when a random effect that is predicted to be small, but not exactly zero, should be included in the model. Remark 1 in the web appendix shows that the gist of this result carries over also to more complex models.

This built-in preference of the conventional cAIC for larger models has an intuitive explanation. If one were to use the maximized log-likelihood for model selection, the choice would always be the largest model under consideration. This over-optimism in the model fit is due to the parameters being estimated from the same y that is the argument of the log-likelihood. The AIC, on the other hand, is a predictive quantity and corrects this bias using a suitable bias correction term. However, the conventional cAIC estimates the bias correction term again from y. In a sense, it does not sufficiently correct, resulting in a similar preference for lager models.

4·2. *The Corrected c*AIC

The corrected cAIC of Liang et al. (2008) remedies the problems of the conventional cAIC. However, the available numerical approximation, similarly to other predictive criteria such as cross validation, can be computationally prohibitive. We derive an analytic representation with an efficient implementation. We focus on an analytic representation of Φ_0. A representation of Φ_1 could be obtained along the same lines, but would be lengthy and cumbersome, whereas simulations in Section 5 show the close agreement between Φ_1 and $\Phi_0 + 1$ for model selection.

Denote the parameter space for $\theta_* = (\theta_{*,1}, \ldots, \theta_{*,q})$ by $\Theta \subseteq \mathbb{R}^q$. Denote by $\widehat{\theta}_*$ the maximum likelihood or restricted maximum likelihood estimator of θ_*.

Hosted by The Berkeley Electronic Press

THEOREM 3. *For the conditional* AIC *in the linear mixed model* (1) *with unknown θ, the bias correction term* (11) *can be written as*

$$\Phi_0 = \widehat{\rho} + \sum_{j=1}^{s} e_j^T \widehat{B}_*^{-1} \widehat{G}_* \widehat{A}_* \widehat{W}_{*,j} \widehat{A}_* y,$$

where we assume that after potential reordering, we can write $\theta_* = (\theta_s^T, \theta_t^T, \theta_{q-s-t}^T)^T$ for some $0 \le s \le q$, $0 \le t \le q-s$, such that $\Theta = \{\theta_* | \theta_s \in \Theta_s \subseteq \mathbb{R}^s, \theta_t \in [0,\infty)^t, \theta_{q-s-t} \in F(\theta_s, \theta_t) \subset \mathbb{R}^{q-s-t}\}$, $\widehat{\theta}_s$ lies in the interior of Θ_s, $F(\theta_s, 0) = 0$ for all θ_s, and $(\widehat{\theta}_t^T, \widehat{\theta}_{q-s-t}^T)^T = 0$. Furthermore, e_j denotes the $s \times 1$ unit vector for component j, $A_* = V_*^{-1} - V_*^{-1} X (X^T V_*^{-1} X)^{-1} X^T V_*^{-1}$, $W_{*,j} = \frac{\partial}{\partial \theta_{*,j}} V_*$, $U_{*,jl} = \frac{\partial^2}{\partial \theta_{*,l} \partial \theta_{*,j}} V_*$, $j, l = 1, \ldots, s$, are $n \times n$ matrices, the jth row of the $s \times n$ matrix G_*, $j = 1, \ldots, s$, is $2\{(y^T A_* y) y^T A_* W_{*,j} A_* - (y^T A_* W_{*,j} A_* y) y^T A_*\}$, and B_* is the negative definite $s \times s$ Hessian matrix for θ_s with jlth entry

$$b_{jl} - y^T A_* W_{*,j} A_* y y^T A_* W_{*,l} A_* y - y^T (A_* U_{*,jl} A_* - 2 A_* W_{*,l} A_* W_{*,j} A_*) y y^T A_* y,$$

where $b_{jl} = \{(y^T A_* y)^2 \operatorname{tr}(U_{*,jl} A_* - W_{*,j} A_* W_{*,l} A_*)/(n-p)\}$ for restricted maximum likelihood estimation, and $b_{jl} = \{(y^T A_* y)^2 \operatorname{tr}(U_{*,jl} V_*^{-1} - W_{*,j} V_*^{-1} W_{*,l} V_*^{-1})/n\}$ for maximum likelihood estimation, $j, l = 1, \ldots, s$.

To give an intuition for the assumptions in Theorem 3, consider the case of a block-diagonal D_* with blocks

$$\frac{1}{\sigma^2} \begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{pmatrix} = \begin{pmatrix} \lambda_1 & \lambda_{12} \\ \lambda_{12} & \lambda_2 \end{pmatrix},$$

such as in a random intercept and random slope model. We need a partition of the parameter space, similarly to Self & Liang (1987); Stram & Lee (1994), to account for potential parameters on the boundary of the parameter space. After potential reordering, either a) $\widehat{\lambda}_1 = \widehat{\lambda}_2 = 0$ b) $\widehat{\lambda}_1 > 0$, $\widehat{\lambda}_2 = 0$, or c) $\widehat{\lambda}_1 > 0$, $\widehat{\lambda}_2 > 0$; $\lambda_2 = 0$ also implies $\lambda_{12} = 0$. Thus, we can write $\theta_* = (\lambda_1, \lambda_2, \lambda_{12})^T = (\theta_s^T, \theta_t^T, \theta_{q-s-t}^T)^T$, with a) $s = 0$, $t = 2$ and $q - s - t = 1$, $(\widehat{\lambda}_1, \widehat{\lambda}_2) = (0, 0) \in [0, \infty)^2$ and $F(0, 0) = 0 = \widehat{\lambda}_{12}$; b) $s = 1$, $t = 1$ and $q - s - t = 1$, $\widehat{\lambda}_1$ in the interior of $[0, \infty)$, $\widehat{\lambda}_2 = 0 \in [0, \infty)$ and $F(\lambda_1, 0) = 0 = \widehat{\lambda}_{12}$ for all λ_1; c) $s = 3$, $t = q - s - t = 0$, $\widehat{\theta}_s = (\widehat{\lambda}_1, \widehat{\lambda}_2, \widehat{\lambda}_{12})^T$ in the interior of Θ_s, which restricts θ_s to ensure positive semi-definiteness of D. Analogous considerations hold for larger blocks.

As $\widehat{\rho} = n - \operatorname{tr}(\widehat{A}_*)$ are the estimated effective degrees of freedom from the conventional cAIC, the second term is a correction term for estimation of the unknown θ_*. The Φ_0 is equal to the Φ_0 one would obtain in the reduced model where $(\theta_t, \theta_{q-s-t}) = 0$ is known. In an implementation, the cAIC can thus be computed in a suitable sub-model. In determining a suitable sub-model, increases of maximized likelihoods should be used in addition to parameter estimates due to numerical imprecisions. We give an implementation as an R package in the web appendix.

Typically, $W_{*,j}$ and $U_{*,jl}$ can be derived explicitly. For example, if D_* is block-diagonal with blocks $\tau_j^2 \Sigma_j$ and known Σ_j, such that $\theta_* = (\theta_{*,1}, \ldots, \theta_{*,q}) = (\lambda_1, \ldots, \lambda_q) = (\tau_1^2, \ldots, \tau_q^2)/\sigma^2$, we have $W_{*,j} = Z_j \Sigma_j Z_j^T$, and $U_{*,jl} = 0_{n \times n}$, $j, l = 1, \ldots q$, where Z_j denotes the corresponding columns of Z. Furthermore, using the Woodbury formula, we can write $V_*^{-1} = I_n - Z(Z^T Z + D_*^{-1})^{-1} Z^T$, and thus only $r \times r$ and $s \times s$ matrices need to be inverted to compute Φ_0.

Collection of Biostatistics Research Archive

5. Simulations

5·1. *Penalized Spline Smoothing*

To illustrate our theoretical findings, we conduct a simulation study covering several settings. For penalized spline smoothing, we concentrate on univariate scatterplot smoothing (6). We consider the following three classes of non-linear functions:

$$
\begin{aligned}
m_1(x) &= 1 + x + 2d(0.3 - x)^2, \\
m_2(x) &= 1 + x + d(\log(0.1 + 5x) - x), \\
m_3(x) &= 1 + x + 0.3d(\cos(0.5\pi + 2\pi x) - 2x).
\end{aligned}
$$

For each function, increasing values of d correspond to increased non-linearity and thus a higher signal-to-noise ratio. We consider the sequence $d = 0, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2$, see the web appendix for a graphical display of the resulting functions. For $d = 0$, all functions reduce to a linear model in x. We set the error variance to $\sigma^2 = 1$, choose x equidistantly from the interval $[0, 1]$, and use a sequence of sample sizes $n = 30, 50, 100, 200$.

For each setting and function, 1000 data sets are generated, and linear and non-linear models fitted to the data based on both ML and REML estimation. The nonparametric effects are specified using cubic B-splines with ten inner knots and second order difference penalty. The mixed model representation from Section 2·1 yields a mixed model with a fixed linear effect in x, and random effects modelling the deviation from this linear effect.

To assess the performance of mAIC and cAIC in model selection, we compute the frequency of selecting the more complex, non-linear model for each value of d. For the conditional AIC, we consider the conventional as well as the corrected variants. For the latter, we compare the exact formula for Φ_0 developed in Section 4 and the numerical approximations for Φ_0 and Φ_1 suggested in Liang et al. (2008), where we insert $\widehat{\sigma}^2$ for the true variance $\widetilde{\sigma}^2$ in the latter. We intrinsically decide on the simpler, linear model whenever the cAICs of both models conincide. Results for function m_1 and sample sizes $n = 30$ and $n = 100$ are shown in Figure 1 (a), complete results can be found in the web appendix.

The conventional cAIC leads to the largest proportion of decisions for the non-linear model under either ML or REML estimation. This agrees with our theoretical findings that this cAIC will always select the non-linear model when the estimated variance is positive. Under a truly linear effect ($d = 0$), the proportion of false decisions for the non-linear model consequently equals the probability of a positive variance estimate, derived previously in Crainiceanu et al. (2003). The corrected cAIC no longer shows this deficiency in any of its variants. In fact, Fig. 1 (a) indicates that the model choice performances of the corrected cAICs are almost indistinguishable and lie between those for the conventional cAIC and the mAIC. Indeed, $\Phi_0 + 1$ and Φ_1 are very close to each other, and estimation uncertainty in the error variance thus seems to be largely ignorable.

The main difference between the analytic and the numeric corrected cAIC lies in computation times. The corrected cAIC involving second derivatives requires 18 seconds per model for $n = 30$, 46 seconds for $n = 100$, and 480 seconds for $n = 500$, while the analytic version is available almost instantaneously. The small differences observed between the analytic and the numeric version of Φ_0, especially for small values d, are due to occasional failure of the numeric computation. In these cases, spurious values in the range of 100s or even negative values may occur for Φ_0. Most cases of differing model choice decisions between numeric and analytic cAIC are due to small underestimations of Φ_0 in the numeric version, causing the cAIC to favour the more complex model although the variance has been estimated to be zero.

In Section 3, we discussed that the probability of selecting a truly zero parameter, analogous to a significance level for the AIC, converges to 0.157 in standard cases. In consequence, the AIC

Collection of Biostatistics Research Archive

Hosted by The Berkeley Electronic Press

is commonly perceived as selecting rather too many than too few variables. For the mAIC, this perception would be misleading, as the corresponding probability is much lower. Conversely, the probability can be more than 35% for the conventional cAIC. Only the corrected cAIC is close to the behavior expected from linear models.

For functions m_2 and m_3 as well as $n = 50$ and $n = 200$, the qualitative findings completely agree with the results presented, and are therefore deferred to the web appendix.

5·2. *Random Intercept Model*

We consider the balanced random intercept model (5) with $J_i = J$ for all i. The random intercepts b_i are assumed to be independent $N(0, d)$ variables such that the variance $d = \tau^2$ is again a measure of the signal-to-noise ratio. We set $\sigma^2 = 1$ and $\beta_0 = 0$ and consider varying random effects variances $d = \tau^2 = 0, 0.1, 0.2, 0.4, 0.6, 0.8$, cluster sizes $J = 3, 6, 9, 12$ and numbers of clusters $I = 10, 20, 40, 80$. All other settings remain the same as in the previous subsection.

Exemplarily for 20 clusters and 3 or 6 observations per cluster, Fig. 1 (b) displays the proportion of simulation replications where the larger random effects model was preferred. The curves are qualitatively similar to the ones for the penalized splines in Fig. 1 (a), with three main differences. First, the selection frequency of the larger model increases faster, due to the different meaning of the signal-to-noise ratio d. Second, the selection frequency of the larger model for $d = 0$ is larger for the random intercept case, both for conventional cAIC as well as mAIC. This is owing to the larger proportion of positive $\widehat{\tau}^2$ estimates for a true value of $\tau^2 = 0$, $1/2$ asymptotically compared to about $1/3$ for the penalized spline case (Stram & Lee, 1994; Crainiceanu & Ruppert, 2004). In contrast, the corrected cAIC has very comparable levels for both cases. And third, the numerical difficulties in approximating the second derivatives in Φ_1 are worse than for the penalized spline case. Several replications now yield very large or even negative degrees of freedom. As an ad hoc correction, we exclude all results with negative degrees of freedom, and those exceeding a certain threshold, chosen as 20 in case of Fig. 1 (b). Still, some deviations remain. Numerical problems are most pronounced for larger τ^2 values, where it is more likely that a single outlier has a large impact on the estimation of σ^2. These numerical problems are, of course, not present for the analytic form of the corrected cAIC.

6. CASE STUDY: CHILDHOOD MALNUTRITION IN ZAMBIA

6·1. *Background*

One of the most urgent and challenging problems in developing countries is malnutrition of large parts of the population, and in particular childhood malnutrition. To monitor the developments in malnutrition, regular demographic and health surveys (DHS) are conducted by Macro International in cooperation with the world health organization (WHO), and made publicly available at `www.measuredhs.com`. In the following, we show how the theoretical results derived for cAIC and mAIC affect the selection of sensible models for the analysis of childhood malnutrition based on a subsample of 1,600 observations chosen randomly from the 1992 Zambia Demographic and Health Survey (Gaisie et al., 1993). Our considerations are based on models developed previously in Kandala et al. (2001).

Malnutrition is generally assessed by comparing anthropometric indicators such as weight with a reference population, accounting for age. We focus on chronic undernutrition (stunting) as measured by insufficient height for age. The dependent variable in our regression models is the Z-score, $zscore_i = (cheight_i - m)/s$, where $cheight_i$ denotes the height of the ith child, m is the median height of children of the same age from a reference population, and s is the corresponding standard deviation. Available covariate information includes categorical variables (gender of the child, education of the mother, employment status of the mother), continuous covariates ($cfeed$ = duration of breastfeeding in months, $cage$ = age of the child in months, $mage$ = age of the mother, $mheight$ = height of the mother, $mbmi$ = body mass index of the mother) and a spatial factor variable that represents the residential district. The web appendix gives more details on the covariates.

We are interested in determining a model that approximates the true data generating mechanism, i.e. a model that contains the essential features driving childhood malnutrition. Our focus is on determining the best model from a set of flexible candidate models. Information criteria such as the AIC are considered more appropriate than significance testing (Burnham & Anderson, 2002, pp. 36-37) in such a setting. In the following, we focus on the selection of linear versus nonlinear functions for the continuous covariates, and of presence versus absence of a random spatial cluster effect, both corresponding to the selection of random effects. We do not focus on the selection of fixed effects, and thus include parametric effects for the categorical covariates without selection. We compare the performance of the marginal and conditional AIC. As computation time (an estimated 110 days) for the numerical approximations is prohibitive for our large sample size and model space, only the analytic representation of the corrected conditional cAIC is considered (60 minutes for all 64 models).

6·2. *Univariate Smoothing*

As a first illustration, we consider the univariate smoothing problem (6) for the $zscore$ y, and the height of the mother x. The function $m(\cdot)$ is modeled using B-splines with ten knots and second order difference penalty. To decide whether nonlinear modelling is required, we estimate the mixed model corresponding to (6) based on ML or REML and compare it to a linear model. For REML estimation, a slightly non-linear curve with corresponding positive variance is estimated (see Figure 2 (a)), while the variance is estimated to be zero for ML estimation. The mAIC chooses the simpler model for either (M_1: 4542.6 (linear) versus M_2: 4544.6 (non-linear) for ML). As expected, the cAIC always gives the same value (4542.6) for either model in the case of ML estimation. As predicted from Theorem 2, the conventional cAIC chooses the more complex model (M_1: 4542.6 versus M_2: 4541.6) for REML estimation. In contrast, the corrected cAIC appropriately incorporates uncertainty in estimating the degrees of freedom and decides on the simpler model (M_1: 4542.6 versus M_2: 4543.2).

6·3. *Additive Mixed Model*

In a more realistic scenario, we consider additive mixed models. The full model contains nonparametric effects for all continuous covariates and a district-specific random intercept,

$$zscore_i = x_i^T \beta + m_1(cage_i) + m_2(cfeed_i) + m_3(mage_i) + m_4(mbmi_i) + m_5(mheight_i) + b_{s_i} + \varepsilon_i.$$

The nonparametric functions $m_1, \ldots, m_5$ are specified as before, and we consider model selection between linear and non-linear effects. The spatial heterogeneity is captured in the district-specific random intercept b_{s_i}, where s_i denotes the region observation i pertains to. The b_{s_i} are assumed to be independent and identically distributed Gaussian, and model selection addresses the question of spatial heterogeneity. We focus on the selection of random effects, and include fixed parametric effects of all categorical and binary covariates, contained in $x_i^T \beta$, in all models.

These choices give 64 possible models overall. Table 1 contains cAIC and mAIC values for the eight best-fitting models for ML and REML estimation, with a complete table in the web appendix. Minimal AIC values in each column are bolded. The eight models correspond to all possible combinations of linear and non-linear modelling for age, height, and body mass index of the mother, and have identical conventional and corrected cAIC values for both ML and REML

Hosted by The Berkeley Electronic Press

estimation. The corresponding estimated curves, given in the web appendix, show that these effects are estimated to be linear.

The effects of the age of the child and the duration of breastfeeding are estimated to be nonlinear using either ML or REML estimation (Fig. 2 (b)). The age effect indicates a steady decline from a relatively well-nourished level immediately after birth to more severe malnutrition later on. The increase for older ages is in fact related to a change in the reference standard used to determine the Z-score. The nonlinear effect of the duration of breastfeeding shows the beneficial effect of longer breastfeeding for the first 10 to 20 months, and a saturation of the effect for longer durations. The mAIC and all versions of the cAIC agree on model 14 as the best fitting model, including nonparametric effects for the age of the child and duration of breastfeeding, and a district-specific random intercept (visualized in the web appendix).

7. DISCUSSION

The class of model choice questions considered in this paper is of relevance for a wide range of models. In addition to linear mixed models for longitudinal data and penalized spline smoothing, considered as examples in this paper, surface estimation, varying coefficient models, or spatial models yield similar model choice questions that can be formulated in terms of the selection of random effects (Ruppert et al., 2003; Fahrmeir et al., 2004). Linear mixed models have also been used in other statistical areas, such as functional data analysis (Di et al., 2008, Greven et al., 2009), where the choice of the number of functional principal components corresponds to the selection of random effects. While we do not specifically focus on the selection of fixed effects in linear mixed models, we expect the corrected conditional AIC to also perform well in this setting.

In the future, it would be of interest to extend our results to generalized linear mixed models. Another interesting question is the relevance of our findings for other criteria used for model selection in mixed models, such as the Bayesian information criterion (BIC, Schwarz, 1978).

APPENDIX 1

Proofs of main results

We give outlines of all proofs here; detailed proofs can be found in a web appendix at `http://www.biostat.jhsph.edu/~sgreven/research/appendix_AIC.zip`.

Proof of Theorem 1. We can expand $2[\log\{f_{\widehat{\psi}(y)}(y)\} - \log\{f_{\psi_K}(y)\}]$ into two contributions from σ^2 and β, which as usual converge in distribution to χ^2_1 and χ^2_p variables, and a third contribution from λ, which is studied by Crainiceanu & Ruppert (2004). They show that if $\widetilde{\lambda} = 0$, this term has a point mass at zero and a second mixture component smaller or equal than χ^2_1. Analogously, $2\mathrm{E}_z[\log\{f_{\psi_K}(z)\} - \log\{f_{\widehat{\psi}(y)}(z)\}]$ can be expanded. Overall, the expectations with respect to y of the respective sums are smaller than $p+2$ and depend on the true $\widetilde{\lambda}$, with the resulting bias in the mAIC not vanishing asymptotically for $\widetilde{\lambda} = 0$. □

For the proof of Theorem 2, we need the following Lemma.

LEMMA 1. *In the linear mixed model* (1) *with $D = \tau^2\Sigma$, let $\widehat{\theta}$ and $\widehat{b}$ be the* ML *estimator and the best linear unbiased predictor for θ and b, respectively. Then, with $P_* = I_n - X(X^T V_*^{-1} X)^{-1} X^T V_*^{-1}$, the conditional log-likelihood allows the representation*

$$\log\{f(y \mid \widehat{\beta}, \widehat{b}, \widehat{\theta})\} = -\frac{1}{2} n \log(2\pi) - \frac{1}{2} n \log\left(\frac{y^T \widehat{P_*}^T \widehat{V_*}^{-1} \widehat{P_*} y}{n}\right) - \frac{1}{2}\,\mathrm{tr}\left(\widehat{V_*}^{-1}\right).$$

COBRA
Collection of Biostatistics
Research Archive

The corresponding quantity when REML *estimation is used is*

$$\log\{f(y \mid \widehat{\beta}, \widehat{b}, \widehat{\theta})\} = -\frac{1}{2} n \log(2\pi) - \frac{1}{2} n \log\left(\frac{y^T \widehat{P_*}^T \widehat{V_*}^{-1} \widehat{P_*} y}{n-p}\right) - \frac{1}{2} \operatorname{tr}\left(\widehat{P_*}^T \widehat{V_*}^{-1} \widehat{P_*}\right).$$

Proof of Lemma 1. Let $\lambda = \tau^2/\sigma^2$ and consider REML estimation. We either have $\widehat{\lambda} = 0$, or the derivative of the profile restricted log-likelihood at $\widehat{\lambda}$ is zero, giving

$$(n-p) \frac{y^T \widehat{P_*}^T \widehat{V_*}^{-1} Z \Sigma Z^T \widehat{V_*}^{-1} \widehat{P_*} y}{y^T \widehat{P_*}^T \widehat{V_*}^{-1} \widehat{P_*} y} = \operatorname{tr}(\widehat{P_*} Z \Sigma Z^T \widehat{V_*}^{-1}).$$

The result follows making additional use of equation (2) and $\widehat{\sigma}^2 = (y - X\widehat{\beta})^T (y - X\widehat{\beta} - Z\widehat{b})/(n-p)$. The result for ML estimation follows analogously using the profile log-likelihood (3). □

Proof of Theorem 2. For $\widehat{\lambda} = 0$, equality of the cAICs follows directly. For $\widehat{\lambda} > 0$ and REML estimation, we make use of Lemma 1 in the representation of the cAIC. The fact that $\log(x) + 1/x$ is a strictly monotonic increasing function for $x > 1$ allows us to link the inequality $cAIC(M_1) < cAIC(M_2)$ to the inequality $\ell(\widehat{\lambda}) \geq \ell(0)$ for the restricted profile log-likelihood $\ell(\lambda)$, which is true by definition. We additionally use the spectral representation of $\ell(\lambda)$ as well as of $P_*^T V_*^{-1} P_*$ in Crainiceanu & Ruppert (2004), and equation (5) in Liang et al. (2008). This gives us $cAIC(M_1) < cAIC(M_2)$, and overall, the stated equivalence follows. The result for ML estimation is derived analogously, additionally using an inequality for the eigenvalues of the sum of two matrices (Theorem 1 in Thompson & Freede, 1971). □

Proof of Theorem 3. We can write $\widehat{y} = y - \widehat{V}_*^{-1} \widehat{P}_* y$, and thus

$$\Phi_0 = \operatorname{tr}\left(\frac{\partial \widehat{y}}{\partial y}\right) = \operatorname{tr}\left[I_n - \widehat{V}_*^{-1} \widehat{P}_* - \sum_{j=1}^{q} \frac{\partial}{\partial \theta_{*,j}} \{\widehat{V}_*^{-1} \widehat{P}_*\} y \left\{\frac{d}{dy} \widehat{\theta}_{*,j}(y)\right\}\right].$$

It is $\partial/(\partial\theta_{*,j})(V_*^{-1} P_*) = -A_* W_{*,j} A_*$ for all j. We can show that $\partial/(\partial y_i) \widehat{\theta}_{*,j} = 0$, for all i and $j = s+1, \ldots, q$. Using the score equation, and as $(\theta_{*,1}, \ldots, \theta_{*,s})$ is in the interior of Θ_s, the restricted maximum likelihood estimator of θ_* fulfills

$$0 \equiv h_j(\widehat{\theta}_*(y), y) := \operatorname{tr}(\widehat{P}_* \widehat{W}_{*,j} \widehat{V}_*^{-1}) - (n-p) \frac{y^T \widehat{P}_*^T \widehat{V}_*^{-1} \widehat{W}_{*,j} \widehat{V}_*^{-1} \widehat{P}_* y}{y^T \widehat{P}_*^T \widehat{V}_*^{-1} \widehat{P}_* y}, \quad j = 1, \ldots, s.$$

The result follows from

$$\frac{d}{dy} \widehat{\theta}_s(y) = -\left[\frac{\partial}{\partial \theta_{*,l}} h_j(\widehat{\theta}_*(y), y)\right]_{j,l=1,\ldots,s}^{-1} \frac{\partial}{\partial y} h(\widehat{\theta}_*(y), y),$$

where $\frac{\partial}{\partial y} h(\widehat{\theta}_*(y), y)$ includes rows $\frac{\partial}{\partial y} h_j(\widehat{\theta}_*(y), y)$, $j = 1, \ldots, s$, as well as lengthy matrix algebra, noting that the Hessian in the first s components of the profile restricted log-likelihood at $\widehat{\theta}_*(y)$ is negative definite, and thus invertible. The result for ML estimation follows analogously. □

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, B. N. Petrov & F. Csaki, eds. Akademiai Kiado.

BRUMBACK, B., RUPPERT, D. & WAND, M. (1999). Comment on "Variable selection and function estimation in additive nonparametric regression using a data-based prior". *Journal of the American Statistical Association* **94**, 794–797.

BURNHAM, K. & ANDERSON, D. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, 2nd ed.

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

Hosted by The Berkeley Electronic Press

Crainiceanu, C., Ruppert, D. & Vogelsang, T. (2003). Some properties of likelihood ratio tests in linear mixed models. Tech. rep., Department of Statistical Science, Cornell University. `http://legacy.orie.cornell.edu/~ddavidr/papers/zeroprob_rev01.pdf`.

Di, C., Crainiceanu, C., Caffo, B. & Punjabi, N. (2008). Multilevel functional principal component analysis. *Annals of Applied Statistics* **3**, 458–488.

Eilers, P. & Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Sciences* **11**, 89–121.

Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica* **14**, 731–761.

Gaisie, K., Cross, A. R. & Nsemukila, G. (1993). Zambia demographic and health survey. Tech. rep. `http://measuredhs.com`.

Giampaoli, V. & Singer, J. (2009). Likelihood ratio tests for variance components in linear mixed models. *Journal of Statistical Planning and Inference* **139**, 1435–1448.

Greven, S., Crainiceanu, C., Küchenhoff, H. & Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics* **17**, 870–891.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383–385.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*. CRC Press.

Hodges, J. & Sargent, D. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika* **88**, 367–379.

Hughes, A. & King, M. (2003). Model selection using AIC in the presence of one-sided information. *Journal of Statistical Planning and Inference* **115**, 397–411.

Kandala, N. B., Lang, S., Klasen, S. & Fahrmeir, L. (2001). Semi-parametric analysis of the socio-demographic and spatial determinants of undernutrition in two African countries. *Research in Official Statistics* **4**, 81–99.

Laird, N. & Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Liang, H., Wu, H. & Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika* **95**, 773–778.

Molenberghs, G. & Verbeke, G. (2007). Likelihood ratio, score, and Wald tests in a constrained parameter space. *The American Statistician* **61**, 22–27.

Robert-Granié, C., Foulley, J., Maza, E. & Rupp, R. (2004). Statistical analysis of somatic cell scores via mixed model methodology for longitudinal data. *Animal Research* **53**, 259–273.

Ruppert, D., Wand, M. & Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press.

Scheipl, F., Greven, S. & Küchenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis* **52**, 3283–3299.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

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

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

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods* **7**, 13–26.

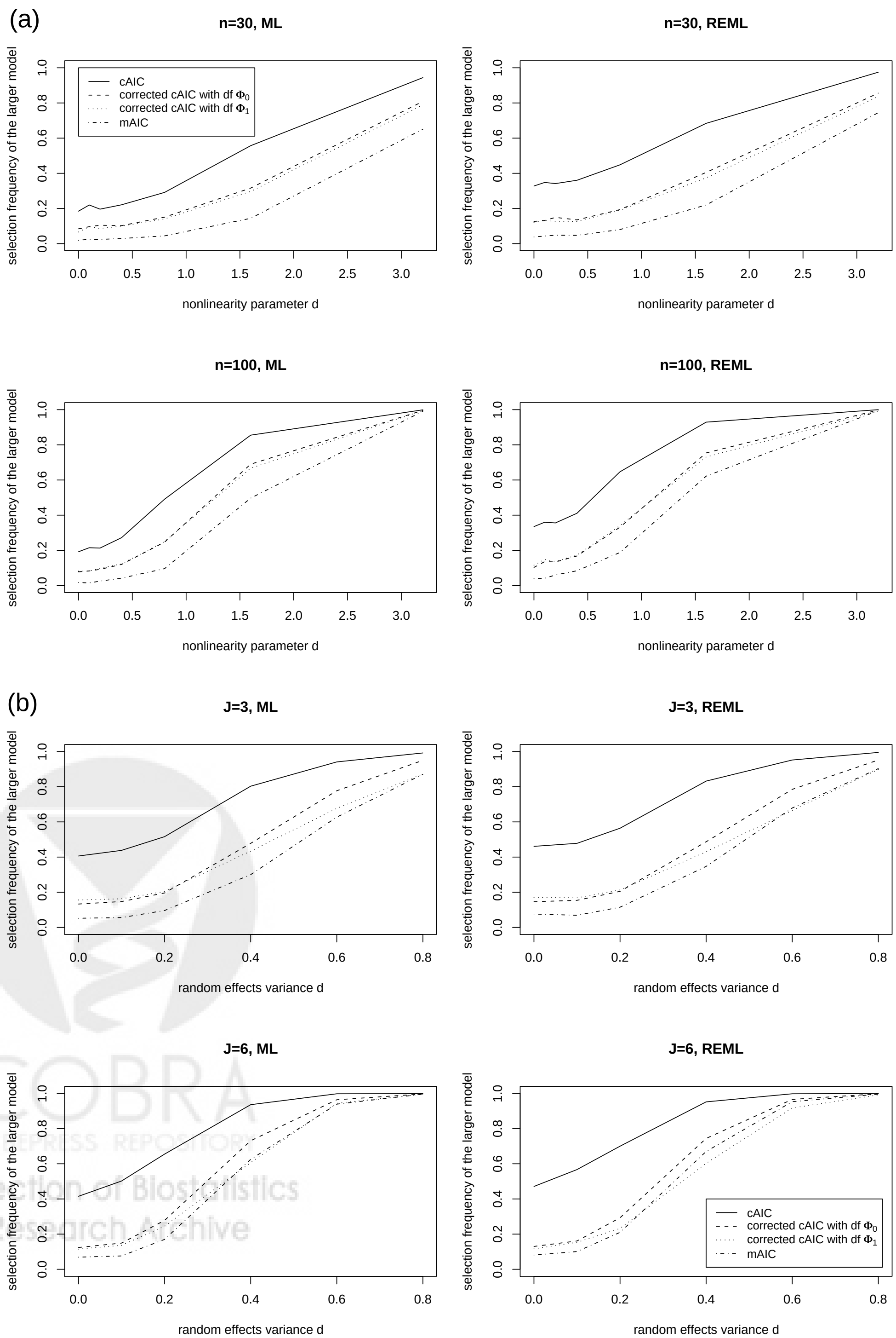
Thompson, R. & Freede, L. (1971). On the eigenvalues of sums of Hermitian matrices. *Linear Algebra and its Applications* **4**, 369–376.

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

Wager, C., Vaida, F. & Kauermann, G. (2007). Model selection for penalized spline smoothing using Akaike information criteria. *Australian and New Zealand Journal of Statistics* **49**, 173–190.

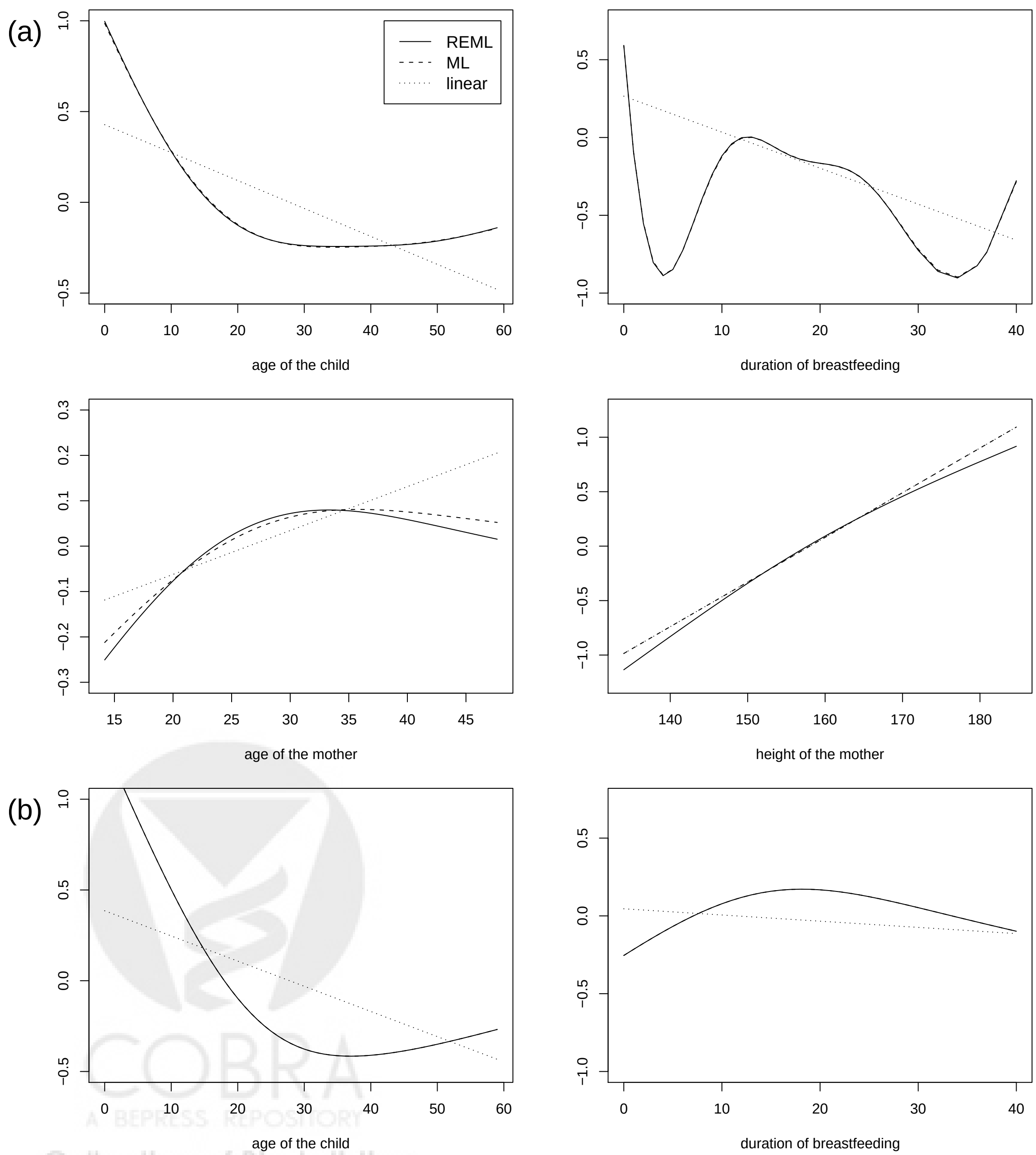
COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Fig. 1. Proportion of simulation replications where the more complex model was favoured by the AIC: (a) for penalized spline smoothing with function $m_1(\cdot)$ and sample sizes $n = 30$ and $n = 100$, and (b) for a random intercept model with twenty clusters and cluster sizes $J = 3$ and $J = 6$ ($\cdot - \cdot -$ conventional cAIC, $- - -$ corrected cAIC with numerically approximated Φ_0, - - - corrected cAIC with numerically approximated Φ_1, — corrected cAIC with analytic Φ_0, $\cdots$ mAIC).



COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Fig. 2. Results from the analysis of the Zambia data. (a) Univariate smoothing: estimated linear (dotted line) and non-linear effects of the age of the mother on the Z-score for both ML (dashed line) and REML (solid line) estimation. (b) Additive mixed model: selected estimated non-linear effects in the full model, and estimated linear effects in the simplest model without random effects, for both ML and REML estimation.



COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Table 1. *c*AIC *and m*AIC *for the eight best-fitting additive mixed models for the Zambia data. The first column contains a model identification number, the following six columns indicate non-linear (+) versus linear (−) modelling of continuous covariate effects and presence (+) versus absence (−) of a district-specific random intercept. In each column, the models with minimal* AIC *are marked in bold. A complete table is in the web appendix.*

							ML			REML		
							conventional	corrected		conventional	corrected	
	cfeed	cage	mage	mheight	mbmi	district	cAIC	cAIC	mAIC	cAIC	cAIC	mAIC
14	+	+	−	−	−	+	**4064.2**	**4068.1**	**4088.6**	**4064.2**	**4068.0**	**4111.3**
34	+	+	+	−	−	+	**4064.2**	**4068.1**	4090.6	**4064.2**	**4068.0**	4113.3
36	+	+	−	+	−	+	**4064.2**	**4068.1**	4090.6	**4064.2**	**4068.0**	4113.3
38	+	+	−	−	+	+	**4064.2**	**4068.1**	4090.6	**4064.2**	**4068.0**	4113.3
54	+	+	+	+	−	+	**4064.2**	**4068.1**	4092.6	**4064.2**	**4068.0**	4115.3
56	+	+	+	−	+	+	**4064.2**	**4068.1**	4092.6	**4064.2**	**4068.0**	4115.3
58	+	+	−	+	+	+	**4064.2**	**4068.1**	4092.6	**4064.2**	**4068.0**	4115.3
64	+	+	+	+	+	+	**4064.2**	**4068.1**	4094.6	**4064.2**	**4068.0**	4117.3

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press