Biometrika (2005), **92,** 2, *pp.* 351–370

Printed in Great Britain

Conditional Akaike information for mixed-effects models

BY FLORIN VAIDA

Division of Biostatistics, Department of Family and Preventive Medicine, University of California at San Diego School of Medicine, La Jolla, California 92093, U.S.A.

vaida@ucsd.edu

AND SUZETTE BLANCHARD

Frontier Science and Technology Research Foundation Inc., Boston, Massachusetts 02215, U.S.A.

suzette@sdac.harvard.edu

SUMMARY

This paper focuses on the Akaike information criterion, AIC, for linear mixed-effects models in the analysis of clustered data. We make the distinction between questions regarding the population and questions regarding the particular clusters in the data. We show that the AIC in current use is not appropriate for the focus on clusters, and we propose instead the conditional Akaike information and its corresponding criterion, the conditional AIC, cAIC. The penalty term in cAIC is related to the effective degrees of freedom ρ for a linear mixed model proposed by Hodges & Sargent (2001); ρ reflects an intermediate level of complexity between a fixed-effects model with no cluster effect and a corresponding model with fixed cluster effects. The cAIC is defined for both maximum likelihood and residual maximum likelihood estimation. A pharmacokinetics data application is used to illuminate the distinction between the two inference settings, and to illustrate the use of the conditional AIC in model selection.

Some key words: Akaike information; AIC; Effective degrees of freedom; Linear mixed model.

1. INTRODUCTION

This paper focuses on model selection for linear mixed-effects models using the Akaike information criterion, AIC (Akaike, 1973). We make a distinction between questions with a focus on population and on clusters; we show that the AIC in current use is not appropriate for conditional inference, and we propose a remedy in the form of the conditional Akaike information and a corresponding criterion. For the conditional AIC, the penalty term is related to the effective number of parameters of a linear mixed model proposed by Hodges & Sargent (2001).

The results of a statistical analysis are usually reported based on the final model choice, which is treated as the 'true model'. The notion of 'true model' is elusive, since the reality producing the data is always complex, while a statistical model is at best a simplified approximation of this reality. The likelihood ratio test is commonly used for model selection between nested models, but hypothesis testing may not be the most suitable framework for model selection (Burnham & Anderson, 2002, p. 36). Alternatively, model

selection was viewed by Fan & Li (2001) as maximising a 'penalised likelihood', where a penalty term is introduced for model complexity, and all models considered are nested in a larger, comprehensive model.

The AIC is technically a penalised likelihood, but it has a sound theoretical basis in model-based prediction using the Kullback–Leibler distance. While strictly speaking it requires the distribution generating the data to belong to the class of fitted models, in practice it only needs to be 'close' to this class (Burnham & Anderson, 2002, p. 65). In addition, the AIC can compare nonnested models.

When the model under consideration contains random effects, the definition of the AIC is not straightforward. What likelihood should be used? Should the random effects be counted as parameters or not? In this paper we argue that the answer to these questions depends on the focus of the research. We identify two kinds of research question: first, inference concerning the population parameters and, secondly, inference about the parameters specific to the clusters. The AIC will be different in the two cases. The formulae are similar: AIC $= -2$ log likelihood $+ 2K$, where K is the 'degrees of freedom' correction, or the number of parameters in the model. However, for the population, or marginal, focus, the likelihood is the marginal likelihood, and K is the number of fixed parameters, counting mean parameters and variance components. In contrast, for the cluster, or conditional, focus, the likelihood is the conditional likelihood, and $K = \rho + 1$, where ρ is the effective number of parameters for the mean model defined by Hodges & Sargent (2001), and 1 stands for σ^2. In practice, we show that a small-sample correction is needed for K.

The distinction between conditional and marginal inference for mixed-effects models was made as early as Harville (1977). Burnham & Anderson (2002) survey the AIC literature, and promote the use of AIC for model selection. Following Burnham & White (2002), Burnham & Anderson (2002) use a formula similar to our cAIC but with a different correction for degrees of freedom, justified by them by analogy with the notion of degrees of freedom used in the smoothing literature (Hastie & Tibshirani, 1990, p. 158). Incidentally, the same justification led Hodges & Sargent (2001) to their definition of ρ. Our derivation of the conditional AIC from first principles gives a theoretical justification of ρ, and an insight into the analogy of a linear mixed model with a linear model where the random effects have fewer than full degrees of freedom. Spiegelhalter et al. (2002) make an implicit distinction between conditional and marginal inference using the idea of focus of inference for hierarchical models. Their DIC criterion, based on Bayesian arguments, is also closely related to our cAIC.

2. Model focus and the conditional Akaike information

2·1. *Akaike information*

Akaike based his information (Akaike, 1973; deLeeuw, 1992) on the Kullback–Leibler distance $I(f, g) = E_f \log f(y) - E_f \log g_\theta(y)$ between the true density f of the distribution generating the data y, and the approximating model for fitting the data $g_\theta = g(.|\theta)$, for $\theta \in \Theta$; E_f is the expectation with respect to the probability density f. Smaller values of $I(f, g_\theta)$ correspond to a better approximation of f by g_θ, and the minimum is obtained for some $\theta_0 \in \Theta$. If the true distribution f belongs to the fitted class of models $\mathcal{G} = \{g_\theta, \theta \in \Theta\}$ then $g_{\theta_0} = f$ and $I(f, g_{\theta_0}) = 0$. In general, f may not be in $\mathcal{G}$, and $I(f, g_\theta) \geqslant 0$. In practice θ is estimated from the data y, and $I(f, g_{\theta_0})$ is approximated by $I(f, g_{\hat{\theta}})$, where $\hat{\theta} = \hat{\theta}(y)$ is usually the maximum likelihood estimator. The quality of the approximation of the true f

by the class $\mathscr{G}$ is assessed, on average, by the quantity

$$E_f I(f, g_{\hat{\theta}}) = E_{f(y^*)} \log f(y^*) - E_{f(y)} E_{f(y^*)} \log g\{y^*|\hat{\theta}(y)\},$$

where y^* is independent of y. When we are comparing different classes of models, the constant $E_{f(y^*)} \log f(y^*)$ can be ignored, and the relative fit of competing models can be assessed using the Akaike information,

$$\text{AI} = -2E_{f(y)} E_{f(y^*)} \log g\{y^*|\hat{\theta}(y)\}. \tag{1}$$

If we consider the fit of future data y^* in (1), AI incorporates a prediction aspect, similar to crossvalidation. The AIC is an estimator of AI, given by

$$\text{AIC} = -2 \log g\{y|\hat{\theta}(y)\} + 2K, \tag{2}$$

where $K = d$, the number of free parameters in the model $\mathscr{G}$. When $\hat{\theta}(y)$ is the maximum likelihood estimator and the approximating class of models $\mathscr{G}$ includes f, $\text{AI} = E(\text{AIC}) + o(1)$ as the sample size $N \to \infty$; that is AIC is unbiased for AI to a first order of N (Akaike, 1973; Burnham & Anderson, 2002). A second-order approximation yields $\text{AI} = E(\text{AIC}_c) + o(N^{-1})$, where AIC_c is as in (2), but with K given by

$$K_c = N(N - d - 1)^{-1} d \tag{3}$$

(Sugiura, 1978; Hurvich & Tsai, 1989; Burnham & Anderson, 2002, pp. 66, 374). For linear models, AIC_c is unbiased for finite sample sizes. More precisely, if f and g are in the class of linear models,

$$y = X\beta + \varepsilon, \tag{4}$$

where X is $N \times p$ of full rank and $\varepsilon \sim N(0, \sigma^2 I_N)$, then $d = p + 1$ and AIC with $K_c = N(N - p - 2)^{-1}(p + 1)$ is unbiased for (1).

2·2. *Marginal versus conditional focus in the mixed-effects model*

Consider a vector y of data from m clusters, modelled as in Laird & Ware (1982) by

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i \tag{5}$$

with $b_i \sim N(0, G)$, independently for each i, where $i = 1, \ldots, m$ is the cluster index, y_i is the vector of n_i responses for cluster i, β is the p-vector of fixed effects, b_i is the q-vector of random effects for cluster i, X_i and Z_i are the $n_i \times p$ and $n_i \times q$ matrices of covariates for the fixed and random effects of full rank, $\varepsilon_i \sim N(0, \sigma^2 I_{n_i})$, independently of b_i, and G is $q \times q$ positive definite. Let $N = \sum_{i=1}^m n_i$ be the total number of observations. Furthermore, let θ be the vector of parameters in the model, including β, σ^2 and the parameters in G. The clusters may represent, for example, different subjects for whom several observations are recorded at different time points. We can write (5) in the condensed form

$$y = X\beta + Zb + \varepsilon, \quad b \sim N(0, G_0), \tag{6}$$

where $X = (X_1^{\mathrm{T}}, \ldots, X_m^{\mathrm{T}})^{\mathrm{T}}$ is $N \times p$ of rank p, $Z = \text{diag}(Z_1, \ldots, Z_m)$ is $N \times r$ block-diagonal of rank $r = mq$, $b = (b_1^{\mathrm{T}}, \ldots, b_m^{\mathrm{T}})^{\mathrm{T}}$, $\varepsilon = (\varepsilon_1^{\mathrm{T}}, \ldots, \varepsilon_m^{\mathrm{T}})^{\mathrm{T}}$, and $G_0 = \text{diag}_m(G)$ is block-diagonal with m blocks G on the diagonal. Model (6) without restrictions on X, Z and G_0 is more general than the Laird–Ware model, and our subsequent developments apply to it as well. Conditionally on b, the likelihood of the model is $g(y|\theta, b)$, and the marginal likelihood is $g(y|\theta) = \int g(y|\theta, b)p(b|G)db$, where $p(b|G) = \prod_{i=1}^m p(b_i|G)$ is the distribution of the random effects.

In the mixed model (5) the interest is either in the population parameters β, population focus, or in the individual clusters, with the associated random effects b_i, cluster focus. For example, in a clinical trial testing the effect of a new treatment versus the standard-of-care, the subjects are enrolled within hospitals, or clinical units, which determine the clusters. In the population focus the interest is in the overall treatment effect, β say, whereas in the cluster focus we want to know the treatment effect at hospital i, $\beta + b_i$ say, which may differ from hospital to hospital.

In the population focus the random effect b_i is a device for modelling the correlation of the responses within cluster i, and model (5) is equivalent to the linear model with correlation errors,

$$y_i = X_i\beta + \gamma_i, \quad \gamma_i = Z_i b_i + \varepsilon_i \sim N(0, Z_i G Z_i^{\mathrm{T}} + \sigma^2 I_{n_i}). \tag{7}$$

The prediction in this case refers to a new cluster, with new random effects. No prediction or inference can be made about the random effects b_i. The AIC in current use, which we call the marginal AIC, mAIC, is appropriate here: $\mathrm{mAIC} = -2\log g(y|\hat{\theta}) + 2K$, where the likelihood is the marginal observed likelihood, and K is the number of parameters in θ, the fixed mean parameters and variance components, perhaps adjusted with the second-order correction (3). To see this, note that mAIC is derived with the marginal likelihood used in (1) and with a new, predictive dataset y^* which contains new clusters with new random effects. In other words, the mAIC is simply the AIC for model (7).

In contrast, in the cluster focus, the random effects b_i are themselves of interest. In this more modern definition, b_i are parameters to be estimated, tied by the distributional property that $b_i \sim p(.|G)$, independently for each i. The joint parameter estimation approach of Harville (1977) and Hodges (1998), involving data cases and constraint cases, implicitly uses this notion of random effects; see also Appendix 1. In this case the choice of fixed versus random effects b_i is a legitimate modelling choice. As we show next, the cluster focus induces a different type of prediction from the population focus, and this calls for a different AIC measure.

2·3. *Conditional Akaike information*

We now define the conditional Akaike information criterion, cAIC, to be used in model selection for the cluster focus. Prediction at the cluster level is conditional on the clusters, and the b_i act as parameters. Therefore, the relevant likelihood is the conditional likelihood $\log g\{y^*|\hat{\theta}(y), \hat{b}\}$, where $\hat{\theta}(y)$ is the maximum likelihood estimator of θ, and $\hat{b} = E(b|\hat{\theta}, y)$ is the empirical Bayes estimator. The second key difference from (1) concerns the prediction dataset y^*. Assume that the true, as opposed to the modelling, distribution of y is $f(y|u)$, and that u is the true random effects vector, with distribution $p(u)$. For the moment assume that $f(y|u)$ belongs to the class of models (6), that is $f(y|u) = g(y|\theta_0, u)$ for some θ_0; u plays the role of b in (6). The prediction dataset is y^* such that y^* and y are independent conditional on u and from the same distribution $f(.|u)$. In other words, y and y^* share the same random effects u, but differ in their error terms.

DEFINITION. *The conditional Akaike information is defined to be*

$$\begin{aligned}\mathrm{cAI} &= -2E_{f(y,u)}E_{f(y^*|u)}\log g\{y^*|\hat{\theta}(y), \hat{b}(y)\} \\ &= \int -2\log g\{y^*|\hat{\theta}(y), \hat{b}(y)\} f(y^*|u) f(y, u)\, dy^*\, dy\, du,\end{aligned} \tag{8}$$

where $f(y, u) = f(y|u)p(u)$ is the joint distribution of y and u.

The CAI is justified along the same lines as (1), with the two key differences noted above. The outer expectation in (8) is also with respect to $p(u)$, to account for the random-effects assumption. The definition (8) covers the following more general settings.

Case 1. The distributions of u and b are not necessarily normal.

Case 2. The true distribution of y, $f(.|u)$, may lie outside the class of models (6). As a simple example, assume that $f(y|u)$ is given by $y = P\alpha + Qu + \varepsilon$, with $u \sim N(0, S)$, $\varepsilon \sim N(0, \sigma_0^2 I_N)$, and P and Q containing covariates different from X and Z.

Case 3. The estimators $\hat{\theta}(y)$ and $\hat{b}(y)$ can be arbitrary estimators of θ and b, such as the residual maximum likelihood estimator and the posterior mode, respectively. Thus, in theory (8) can be used to compare not only different classes of models but also different estimation methods for the same class of models. This point is further explored in §§ 3·2 and 4.

Case 4. The setting is more general than that of linear mixed models, and includes, for example, nonlinear or generalised linear mixed models.

In practice, CAI needs to be estimated from the data. In the next section, by analogy with AIC we develop an estimator for the CAI, the conditional AIC, for the case when f and g belong to the class of models (6).

3. Conditional Akaike information criterion

3·1. *Main results*

As we will see shortly, CAIC is similar in form to AIC as in (2) with two important distinctions: the model loglikelihood is conditional on $\hat{b}$, and the number of parameters is related to ρ, the 'effective degrees of freedom' of Hodges & Sargent (2001). Briefly, $\rho = \mathrm{tr}(H_1)$ where H_1 is the 'hat' matrix mapping the observed data vector y into the fitted vector $\hat{y} = X\hat{\beta} + Z\hat{b}$, that is $\hat{y} = H_1 y$; see Appendix 1 for details.

Theorem 1. *Assume that data y have true density $f(y|u) = g(y|\theta_0, u)$ for some θ_0 and for some random effect u with distribution $p(u)$. The data are modelled by* (6)*, with densities denoted by $g(y|\theta, b)$ and $p(b)$. Assume that σ^2 and G_0 are known. Let $\hat{\theta}(y)$ and $\hat{b}(y)$ be the maximum likelihood and the empirical Bayes estimators for θ and b, respectively.*

Then, an unbiased estimator of the CAI *in* (8) *is*

$$\text{CAIC} = -2 \log g\{y|\hat{\theta}(y), \hat{b}(y)\} + 2\rho. \tag{9}$$

The proof is given in Appendix 2.

Theorem 2. *Under the set-up of Theorem* 1*, assume further that σ^2 is unknown, but that the scaled variance of b, $D_0 = \sigma^{-2} G_0$, is known. Then, an unbiased estimator of the* CAI *in* (8) *is*

$$\text{CAIC} = -2 \log g\{y|\hat{\theta}(y), \hat{b}(y)\} + 2K, \tag{10}$$

where K is given by

$$K = K_{\text{MLE}} = \frac{N(N-p-1)}{(N-p)(N-p-2)}(\rho+1) + \frac{N(p+1)}{(N-p)(N-p-2)} \tag{11}$$

and p is the number of columns in X, the number of fixed effects.

The proof is given in Appendix 2.

In both theorems, cAIC is unbiased for cAI not only asymptotically but also for finite sample size N; K_{MLE} can be interpreted, by analogy with (3), as a small sample correction. The properties of K_{MLE} are summarised by the following result.

PROPOSITION 1. (i) *An alternative formula for* K_{MLE} *in* (11) *is*

$$K_{\mathrm{MLE}} = \frac{N}{N-p-2}\left\{(\rho+1) - \frac{\rho-p}{N-p}\right\}; \tag{12}$$

(ii) $$\rho + 1 < \frac{N(N-p-1)}{(N-p)(N-p-2)}(\rho+1) \leqslant K_{\mathrm{MLE}} \leqslant \frac{N}{N-p-2}(\rho+1);$$

(iii) *as* $N \to \infty$, $K_{\mathrm{MLE}}/(\rho+1) \to 1$.

The proof is included in Appendix 2. Part (iii) states that for large sample sizes an alternative to (11) in the definition of cAIC is the approximation

$$K_a = \rho + 1; \tag{13}$$

in K_a we count the degrees of freedom as ρ for the mean term and 1 for σ^2.

3·2. *Residual maximum likelihood estimation*

In § 3·1 we have assumed that $\hat{\theta}$ is the maximum likelihood estimator. The bias-corrected residual maximum likelihood estimator for σ^2, $\hat{\sigma}^2_R$, is often preferred to the maximum likelihood estimator $\hat{\sigma}^2$ in the linear model (4) or in the linear mixed model (6). In each case, the two estimators for β are identical; as in Theorem 2, we assume that D_0 is known for the linear mixed model. Let $\hat{\theta}_R = (\hat{\beta}, \hat{\sigma}^2_R)$ for either model. Then the simple relationship $\hat{\sigma}^2_R = \{N/(N-p)\}\hat{\sigma}^2$ holds. Just as for maximum likelihood, the AI and cAI for residual maximum likelihood estimators are also defined by (1) and (8) and their estimators by (2) and (10). However, as it turns out, the correction term K is different.

THEOREM 3. *Under the conditions of Theorem* 2, *an unbiased estimator of the* cAI (8) *using the residual maximum likelihood estimator* $\hat{\theta}_R$ *is*

$$\mathrm{cAIC} = -2\log g\{y|\hat{\theta}_R(y), \hat{b}(y)\} + 2K_R, \tag{14}$$

where

$$K_R = \frac{N-p-1}{N-p-2}(\rho+1) + \frac{p+1}{N-p-2}. \tag{15}$$

In other words, $K_R = \{(N-p)/N\}K_{\mathrm{MLE}}$. Incidentally, it has not been noted, to our knowledge, that the same relationship holds in the linear model (4) for AIC_c: if we use $\hat{\theta}_R$, the unbiased estimator of AI (1) is

$$\mathrm{AIC}_c = -2\log g\{y|\hat{\theta}_R(y)\} + 2\frac{(N-p)(p+1)}{N-p-2}. \tag{16}$$

Set the last term in (16) equal to $2k_R$. Then $k_R = \{(N-p)/N\}K_c$, where K_c is defined as in (3). The proofs for Theorem 3 and for (16) are given in Appendix 2. The various criteria discussed in the paper are contrasted and summarised in Table 1.

Table 1: *Summary of the* AIC*'s. Mixed-effects model*: $\text{cAIC} = -2 \times$ *conditional loglikelihood* $+ 2K$; $\text{mAIC} = -2 \times$ *marginal loglikelihood* $+ 2K$. *Linear model*: $\text{AIC} = -2 \times$ *loglikelihood* $+ 2K$. *For* mAIC, d *is the total number of parameters in* β, σ^2 *and* D_0

Criterion	Method	Type	K	Formula for K
			Mixed-effects model	
cAIC	ML	Asymptotic	K_a	$\hat{\rho} + 1$
		Finite-sample	K_{MLE}	$\frac{N(N-p-1)}{(N-p)(N-p-2)}(\hat{\rho}+1) + \frac{N(p+1)}{(N-p)(N-p-2)}$
	REML	Asymptotic	K_a	$\hat{\rho} + 1$
		Finite-sample	K_R	$\frac{N-p-1}{N-p-2}(\hat{\rho}+1) + \frac{p+1}{N-p-2}$
mAIC	ML	Asymptotic	K	d
		Finite-sample	K_c	$\frac{N}{N-d-1}d$
			Linear model	
AIC	ML	Asymptotic	K	$p + 1$
		Finite-sample	K_c	$\frac{N}{N-p-2}(p+1)$
	REML	Asymptotic	K	$p + 1$
		Finite-sample	K_R	$\frac{N-p}{N-p-2}(p+1)$

ML, maximum likelihood; REML, residual maximum likelihood.

The likelihood in (14) and (16) is the 'standard', not the residual, or restricted, likelihood. The latter is often reported by software, for example by `gls()` or `lme()` in R and S-Plus or `proc mixed` in SAS, with the proviso that it should only be used for comparing models with the same mean structure and different variance structures. Furthermore, the residual-likelihood-based AIC cannot be compared with a likelihood-based AIC. In contrast, (14) and (16) are derived as estimators of AI as in (1), and can be used for comparing models with different mean and variance structures, and indeed for comparing the residual and standard maximum likelihood estimators.

3·3. *The case of unknown variance* G_0

We now turn to the case when D_0, and therefore $G_0 = \sigma^2 D_0$, is unknown. Let $D = \sigma^{-2} G$ be the scaled variance of the random effects b_i, that is $D_0 = \text{diag}_m(D)$. The 'hat' matrix H_1 depends on D: $H_1 = H_1(D)$. If D is estimated by its maximum likelihood estimator $\hat{D}$, then $\hat{y} = \hat{H}_1 y$, where $\hat{H}_1 = H_1(\hat{D})$. Accordingly, we distinguish between the observed $\hat{\rho} = \text{tr}(\hat{H}_1)$ and the true $\rho = \text{tr}(H_1)$, where H_1 is computed at the true D. Strictly speaking, H_1 is not a hat matrix, since it does not map y on to $\hat{y}$; however, ρ, the effective degrees of freedom if D were known, does reflect the degrees of freedom associated with the model design without the additional variability induced by not knowing D.

The following simple case shows that the correction for the degrees of freedom in (10) due to G_0 is negligible asymptotically; we consider cluster focus asymptotics with m fixed and $n_i \to \infty$ uniformly.

Example: *The one-way random-effects model.* Let the data y follow the model

$$y_{ij} = b_i + \varepsilon_{ij}, \tag{17}$$

with $i = 1, \ldots, m$ and $j = 1, \ldots, n$, $b_i \sim N(0, \tau^2)$, independently for each i, and $\varepsilon_{ij} \sim N(0, \sigma^2)$, independently for each i and j. For simplicity we assume that σ^2 is known. The cAIC has the form (10), where K is an estimator of the 'bias correction' BC, which is defined by

$$\text{cAI} = -2E_{f(y,u)} \log g\{y|\hat{\theta}(y), \hat{b}(y)\} + 2\text{BC}. \tag{18}$$

Let $\lambda = \tau^2/\sigma^2$ be the scaled variance of b_i, playing the role of D, and let $\hat{\lambda}$ be its maximum likelihood estimator, that is, $\hat{D}$. As shown in Appendix 3, a Taylor expansion for BC yields

$$\text{BC} = \rho + 2/(n\lambda + 1) + o(n^{-1}). \tag{19}$$

The term $2/(n\lambda + 1)$ is the degrees-of-freedom correction due to the unknown variance τ^2. This correction is of $O(n^{-1})$. It is also less than 2, since $\lambda > 0$. It is small for large numbers of observations per subject n and/or for large values of λ, that is, for a large ratio of between to within subject variance. The correction does not depend on the number of clusters m.

Table 2 compares the bias correction BC with its estimator $K = \hat{\rho} + 2/(n\hat{\lambda} + 1)$, for different values of n, σ^2 and λ. In all cases BC is very close to the true effective degrees of freedom ρ. Not knowing D affects K in two ways, through the extra term $2/(n\hat{\lambda} + 1)$ in (19) and through using $\hat{\rho}$ instead of ρ. In Table 2 $\hat{\rho}$ is negatively biased for ρ when $n = 6$; for $n = 26$ the bias is smaller. In § 5 we show a simulation where the effect of the unknown variance D is not negligible for small cluster size.

Table 2: *Example. Simulation to study* cAIC *and the degrees of freedom due to the scaled variance* λ *in the one-way mixed-effects model. Bias correction* BC *is the expected value of the difference between the observed loglikelihood and* cAI/2; $K = \hat{\rho} + 2/(n\hat{\lambda} + 1)$ *is the first-order estimator of* BC; $\hat{\rho}$ *is the estimated number of degrees of freedom for the random effects, as the maximum likelihood estimator of* $\rho = m(1 + n^{-1}\lambda^{-1})^{-1}$. *Mean values are reported for* K *and* $\hat{\rho}$ *based on* 10 000 *simulations. The Monte Carlo standard errors are less than* 1·5% *and* 0·25% *of* BC *and* $\hat{\rho}$, *respectively. In all cases* $m = 10$ *clusters, and* $\tau^2 = 0{\cdot}0367$

n	σ	BC	K	$\hat{\rho}$	ρ
6	0·0705	9·80	9·78	9·73	9·78
26	0·0705	9·94	9·95	9·94	9·95
6	0·141	9·19	9·18	8·97	9·17
26	0·141	9·79	9·80	9·75	9·80
6	0·282	7·37	7·37	6·74	7·35
26	0·282	9·22	9·23	9·04	9·23

Burnham & White (2002) and Burnham & Anderson (2002) have recently proposed using

$$-2 \log g\{y|\hat{\beta}(y), \hat{b}(y)\} + 2(\hat{\rho} + 1 + c) \tag{20}$$

as AIC, where c is the number of unknown parameters in the scaled variance matrix D. Our analysis does not support this way of accounting for the unknown variances D: since

in (19) the second-order term goes to zero as n or λ grows large, a fixed correction c for the number of parameters in D cannot be correct.

The general case of unknown D is analytically complex: no unbiased estimator for cAI such as (9) or (10) exists and no simple correction for D seems to be available. We have computed a first-order correction when D depends on a scalar parameter; the formulae are complex and not immediately transparent, and therefore they were not included here. The general case may require a separate investigation. Based on the evidence of the case study above and of the simulation in the next section, we propose the use in practice of the cAIC (10) with the correction (11) for maximum likelihood, or (14) with the correction (15) for residual maximum likelihood estimation.

It seems inconsistent, at first sight, not to count the parameters of D in (11). The reason for this becomes apparent in (8): the conditional likelihood $g\{y^*|\hat{\theta}(y), \hat{b}(y)\}$ does not depend on D, and therefore D need not be accounted for. However, there is a small price to pay for not knowing D, in the additional variability in $\hat{\theta}(y)$ and $\hat{b}(y)$. The less precise the estimation of D, the higher the price.

4. Case study: Cadralazine data

4·1. *The data analysis*

To illustrate the use of cAIC and the distinction from mAIC, we analysed as a case study a pharmacokinetics dataset, the Cadralazine data (Lunn et al., 1999; Wakefield et al., 1994). The dataset consists of plasma drug concentrations from 10 cardiac failure patients who were given a single intravenous dose of 30 mg of cadralazine, an anti-hypertensive drug. Each subject has the plasma drug concentration, in mg/l, measured at 2, 4, 6, 8, 10 and 24 hours, for a total of 6 observations per subject. In the original dataset, two of the ten subjects had observations at 28 and 32 hours; we removed these, in order to work with a simple balanced dataset. Also, three concentrations of 0 were replaced with 0·005 mg/l.

The data for a given subject are well described by the one-compartment model

$$\text{concentration} = (\text{dose}/V_d) \times \exp(-kt),$$

where 'concentration' is the drug concentration at time t, 'dose' is the dose of the drug, i.e. 30 mg, V_d is the volume of distribution, a scaling factor simplistically interpretable as the volume of blood over which the drug is distributed, k is the elimination rate constant, measured in hours^{-1}, and t is the time since drug administration, also in hours; V_d and k are the unknown parameters. This corresponds to the linear model $\log(\text{concentration}) - \log(\text{dose}) = -\log(V_d) - kt + \text{error}$, or, in statistical notation,

$$y_{ij} = \beta_{0i} + \beta_{1i} t_j + \varepsilon_{ij}, \tag{21}$$

where $i = 1, \ldots, 10$ stands for the subject, and $j = 1, \ldots, 6$ is the measurement index for subject i. The plot of the response versus time is given in Fig. 1. The data for each patient are well described by a straight line, but the slopes and intercepts of the ten regression lines differ from subject to subject. A main interest of the analysis is to determine the values of the volume, $\exp(-\beta_{0i})$, and elimination rate constants, $-\beta_{1i}$, of the 10 subjects in the study, and their population-level averages.

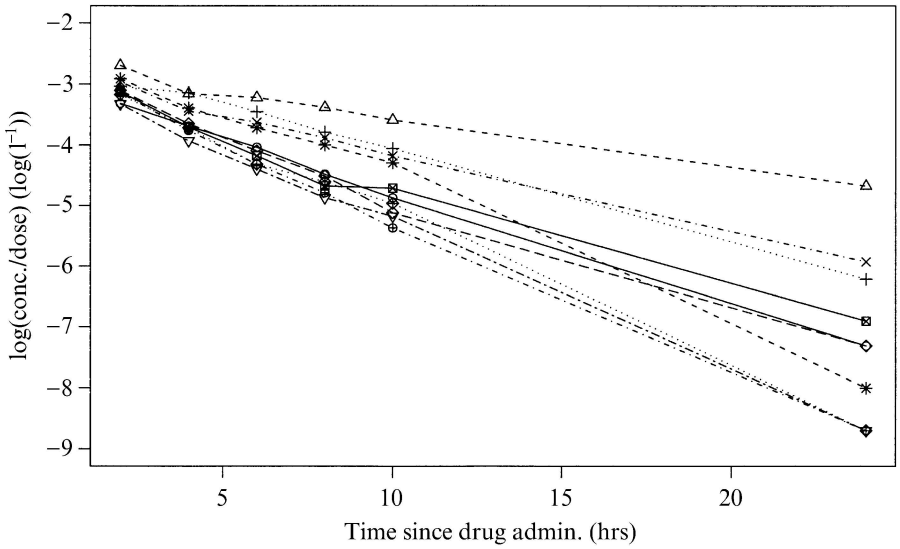

Fig. 1: Cadralazine data. Individual observations of log(concentration/dose) at six time points, for 10 subjects. The time is measured since the administration of the drug.

Two reasonable models to be considered based on (21) are the following.

Model 1: *Linear regression model.* The intercepts and slopes, β_{0i} and β_{1i}, are fixed parameters, $i = 1, \ldots, m$. Note that the effects differ from subject to subject.

Model 2: *Mixed effects model.* This has random intercepts and slopes, that is $\beta_{0i} = \beta_0 + b_{0i}$ and $\beta_{1i} = \beta_1 + b_{1i}$, where, independently for each i,

$$(b_{0i}, b_{1i}) \sim N(0, G). \tag{22}$$

Formally, the linear regression model is an extreme case of the mixed-effects models, when the variance components go to infinity: $G \to \text{diag}(\infty)$.

The estimates for the linear regression slopes and intercepts are similar to the two models; details not included. Based on these and on the residuals plots, Fig. 2, both models give very similar fits. We want to choose the better fitting model based on the AIC. In view of the similar fits, we expect the two models to have comparable AIC values. The software output from the `nlme` package in R (Pinheiro & Bates, 2000, p. 283) is surprising: the mixed-effects model has an AIC of 11·0, and the linear regression model has an AIC of −47·1; see Table 3.

While no rigorous theory is available, Burnham & Anderson (2002, p. 70) suggest that a difference of at most 2 in AIC is not reliable for ranking two models, whereas a difference of 10 is overwhelmingly in favour of the model with the smaller AIC. Based on the reported AIC, the mixed-effects model appears inferior by an enormous margin. This blatant favouritism for the linear regression model is all the more intriguing, since the linear regression model has more parameters, 21, than the mixed-effects model, which has 6. In the light of § 2, the apparent contradiction between the AIC values and the model fits is easily explained: the AIC for the mixed-effects model is what we call the marginal AIC, and it is not appropriate for this conditional model comparison. In contrast, cAIC with K_a is −44·5, comparable to the linear regression model. The comparison of the effective degrees of freedom is illuminating: in the mixed-effects model, $\rho + 1 = 19{\cdot}2$, close

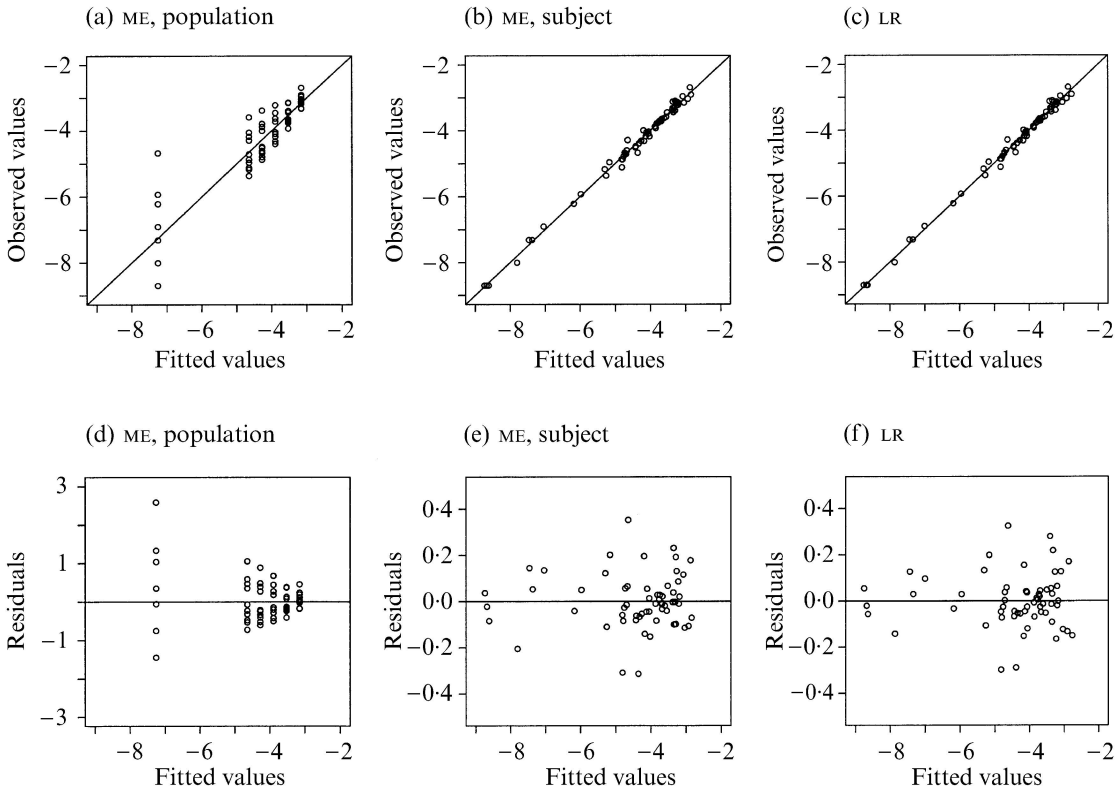

Fig. 2: Application of mixed-effects, ME, models to the Cadralazine data. (a) Fitted-value plot for population-focus approach, (b) fitted-value plot for cluster-focus, i.e. subject-specific, approach, (c) fitted value for linear regression, LR, (d) residual plot for population-focus approach, (e) residual plot for cluster-focus, i.e. subject-specific, approach, (f) residual plot for linear regression.

Table 3. *Comparison of* AIC*'s for the mixed-effects and linear regression models for the Cadralazine dataset. The finite-sample corrected* AIC *and* mAIC *include second-order correction* (3) *for* ML, *and* (16) *for* REML; *the* cAIC *is computed from* (10), *with* $K = K_a$ *for the asymptotic version, and* $K = K_{\mathrm{MLE}}$ *for* ML *or* K_R *for* REML, *for the finite-sample corrected version. The* mAIC *for* REML *was omitted, since it is based on the residual likelihood, and therefore not comparable with the other values in the table*

		Mixed-effects model		Linear regression model
Method	Type	mAIC	cAIC	AIC
ML	Asymptotic	11·0	−44·5	−47·1
	Finite-sample corrected	12·6	−42·3	−22·8
REML	Asymptotic	—	−44·5	−42·8
	Finite-sample corrected	—	−43·7	−40·6

ML, maximum likelihood; REML, residual maximum likelihood.

to the linear regression model value of $p + 1 = 21$. For practical purposes the random effects act almost as unconstrained parameters, which explains why the two model fits are so similar.

Interestingly, the finite-sample corrected AIC's using (11) and (3) make an important difference, with a large penalty for the fixed effects model. The values are cAIC $= -42{\cdot}3$ and $\text{AIC}_c = -22{\cdot}8$, suggesting a clear modelling advantage for the mixed-effects model; see Table 3. The poor AIC_c value is in fact due to the high bias of the maximum likelihood estimator for σ^2, and it can be avoided by using instead the residual maximum likelihood estimator, which is unbiased. For the residual maximum likelihood estimators the cAIC values for the mixed-effects model are $-44{\cdot}5$, using K_a, and $-43{\cdot}7$, using K_R, and the AIC values for the linear regression model are $-42{\cdot}8$, asymptotic, and $-40{\cdot}6$, for AIC_c from (16); see Table 3. The finite-sample corrected values still favour the mixed-effects model over the linear regression model, but the difference is much smaller than for maximum likelihood estimation. For both the residual and maximum likelihood estimation, the adjustment due to the finite-sample correction in the cAIC is small, since the number of fixed effects in the model is small. The comparison of the finite-sample corrected cAIC between the two estimation methods, which is possible, as discussed in § 3·2, slightly favours the residual maximum likelihood estimation.

The distinction between the population and cluster, i.e. subject-specific in this context, focuses is well illustrated by Fig. 2, which shows plots of the observed versus fitted values, and residuals versus fitted values, for the mixed-effects model corresponding to the two focuses, Figs 2(a), (b) and Figs 2(d), (e) and for the linear regression model, Figs 2(c), (f). For population inference the fitted values are $\bar{y}_{ij} = \beta_0 + \beta_1 t_{ij}$ and the residuals are $y_{ij} - \bar{y}_{ij}$, whereas for individual inference the fitted values are $\hat{y}_{ij} = \beta_{0i} + \beta_{1i} t_{ij}$ and the residuals are $y_{ij} - \hat{y}_{ij}$. Only Figs 2(b), (e) are directly comparable to the linear regression. The finite-sample AIC values associated with the two focuses are mAIC $= 12{\cdot}6$ and cAIC $= -42{\cdot}3$, whereas the AIC for the linear regression is AIC $= -22{\cdot}8$. Thus, the graphical comparison makes it clear that the marginal and conditional levels convey different information, and that the appropriate comparison for the regression model is at the conditional level.

4·2. *Cluster focus versus population focus*

When faced with analysing clustered data, a statistician has to decide whether the substantive questions of interest refer to the particular clusters, or subjects, in the dataset, or to the general population. Accordingly, the instruments for model selection will differ in the two cases. In some situations, both kinds of question are of interest; however, the different model selection schemes for the conditional and marginal inference may lead to different final models. The cluster versus population dichotomy is of particular interest in pharmacology. In a population study we may be interested in comparing the pharmacokinetic parameters, such as absorption and elimination rate constants, for, say, two different age groups, while for therapeutic drug monitoring we are interested in the subjects' parameters of drug exposure, for the purpose of intervention.

To consider the cluster focus in the Cadralazine study, assume first that the 10 subjects are of interest. We limited ourselves to models arising from (21) where the subject-specific parameters are possibly random. The intercepts β_{0i} and the slopes β_{1i} could be either fixed, random, of the form $\beta_{0i} = \beta_0 + b_{0i}$, $b_{0i} \sim N(0, \tau_0^2)$, or they could be common to all subjects, that is $\beta_{0i} = \beta_0$ or $\beta_{1i} = \beta_1$ for all i. Table 4 lists the possible models in the order of their cAIC values using K_{MLE}. Note that the random intercept and slope model has the best cAIC fit.

In the case of the population focus, the inference is about population parameters only, and the relevant criterion is mAIC. Table 4 lists five such models, with values in the mAIC

Table 4. *Comparison of models based on subject-specific focus*, cAIC, *and population focus*, mAIC. (i) '*common*', (ii) '*random*' *and* (iii) '*subject-specific*' *mean that the corresponding parameters* β_{ki} $(i = 1, \ldots, 10)$ *are as follows*: (i) *all equal*, $\beta_{ki} = \beta_k$; (ii) $\beta_{ki} \sim N(0, \sigma_k^2)$, *independently*; (iii) β_{ki} *are different*, *fixed parameters*; d *is the total number of parameters in the model*, *with* $d = \hat{\rho} + 1$ *for the subject-specific mixed-effects models*; $K = K_{\mathrm{MLE}}$ *for* cAIC, *and* $K = K_c$ *for* AIC *and* mAIC. *For the population model*, *there are* 2 *degrees of freedom for the mean in all cases*

	Subject-specific			Population		
Model	cAIC	d	K	mAIC	d	K
ME: intercept random, slope random	−42·3	19·2	20·2	12·6	6	6·8
LR: intercept and slope subject-specific	−22·8	21	33·2	—	—	—
ME: intercept common, slope random	−12·0	12·0	12·6	22·3	4	4·4
LR: intercept common, slope subject-specific	−7·8	12	15·3	—	—	—
GEE/GLS: intercept common, slope common	—	—	—	42·1	18	26·3
ME: intercept random, slope common	85·6	10·8	11·4	100·4	4	4·4
LR: intercept subject-specific, slope common	91·2	12	15·3	—	—	—
LR: intercept common, slope common	122·3	3	3·2	122·3	3	3·2

ME, mixed-effects model; LR, linear regression model; GEE/GLS, generalised estimated equations/generalised least squares approach, with unstructured correlation matrix.

column. The population-focus random intercept and slope model also has the smallest mAIC.

Several models appear in both columns of Table 4. The parameters in these models are either random or common to all clusters. The marginal and conditional AIC are the same for the model with common intercept and slope. For the other three models, cAIC and mAIC differ, according to the focus of the model.

The models with fixed, subject-specific parameters cannot be included in the population focus, since the population focus does not allow for subject-specific inference. In contrast, the model denoted by GEE/GLS may only be used for population focus, no inference being made for the individual effects.

5. A simulation study

We conducted a simulation study to investigate the properties of the cAIC estimator (10). Ideally, this estimator would display a small bias for small sample sizes, and would be consistent, when $n_i \to \infty$ uniformly; in cluster-focus asymptotics, m is fixed by design, similarly to the number of covariates in a regression model.

We simulated data from the model (21) with random effects (22). The variance matrix G is unstructured, with $c = 3$ free parameters. In all simulations $m = 10$, $\beta_0 = -2{\cdot}78$, $\beta_1 = -0{\cdot}186$ and $G = \mathrm{var}(b_i)$ were chosen as the estimated values from the Cadralazine data: $\mathrm{var}(b_{i1}) = 0{\cdot}0367$, $\mathrm{var}(b_{i2}) = 0{\cdot}00279$ and $\mathrm{cov}(b_{i1}, b_{i2}) = -0{\cdot}00126$. The time points in the simulation were evenly spaced at 5-hour intervals, from 0 to 25, giving $n = 6$, 0 to 125, giving $n = 26$, and 0 to 250, giving $n = 51$, respectively; in the Cadralazine data, these times were $t = 2$, 4, 6, 8, 10 and 24 hours.

To study the influence of the variance ratios, that is D, on cAIC, we considered three cases: $\sigma = 0{\cdot}141$, from the Cadralazine dataset analysis, and half and double this value, namely $\sigma = 0{\cdot}0705$, and $\sigma = 0{\cdot}282$. We present maximum likelihood estimates throughout;

we also looked at residual maximum likelihood estimates, which were very similar. We compared cAI with its estimator cAIC. The value of cAI was computed by Monte Carlo simulation, with 10 000 and 1000 iterations respectively for the outer and inner expectations in (8). We compared the mean correction term K_{MLE} with the 'bias correction' BC given by (18). The results are included in Table 5. For comparison, we also computed the large-sample correction K_a and Burnham & White's (2002) correction $(\hat{\rho}+1+c)$. In K_{MLE} and K_a, the maximum likelihood estimate $\hat{\rho}$ was used for ρ.

Table 5: *Simulation study. Comparison of various corrections for degrees of freedom for* cAIC *as in* (10): K_{MLE} *and* K_a *are computed using* $\hat{\rho}$; $\hat{\rho}+1+c$ *is Burnham & White's* (2002) *correction* (20). *The number of observations per cluster is n. In all cases,* $m=10$ *clusters and* $c=3$ *variance components. Mean values are reported for* $\hat{\rho}$, K_{MLE} *and* K_a *based on* 10 000 *simulations. The Monte Carlo standard errors are less than* 1% *and* 0·2% *of* BC *and* $\hat{\rho}$, *respectively*

n	σ	BC	K_{MLE}	K_a	$(\hat{\rho}+1+c)$
6	0·0705	21·5	21·2	20·1	23·1
26	0·0705	21·1	21·1	20·8	23·8
51	0·0705	21·1	21·0	20·9	23·9
6	0·141	20·0	19·0	18·0	21·0
26	0·141	20·6	20·3	20·1	23·1
51	0·141	20·8	20·6	20·5	23·5
6	0·282	17·4	15·4	14·6	17·6
26	0·282	18·9	18·2	18·0	21·0
51	0·282	19·7	19·3	19·2	22·2

In general K_{MLE} gives a very good approximation to BC, with a small negative bias. The bias has two sources: we do not account for the influence of the unknown G, and we use $\hat{\rho}$ instead of ρ. As the sample size n increases, the bias of K_{MLE} decreases. The approximation is more precise for smaller σ, closer to the limiting case $\sigma/|G|=0$, when $\rho=20$ which is the number of parameters of the mean in the fixed-effects model.

The asymptotic K_a is valid for $n=26$ and $n=51$, but it is inadequate for $n=6$, which is more typical in applications, proving the need for the finite-sample correction K_{MLE} instead of K_a. The correction $(\hat{\rho}+1+c)$ is clearly biased, with a bias stabilising at the value $c=3$ for small σ, as n increases. For $\sigma=0{\cdot}282$, $(\hat{\rho}+1+c)$ is more accurate for $n=6$, but then becomes increasingly biased as n grows.

6. DISCUSSION

Theorem 1 gives a theoretical justification for the use of Hodges & Sargent's (2001) definition of ρ. This investigation gives further appeal to the idea of thinking of a linear mixed model in the conditional focus as behaving similarly to a standard linear model, but with the random effects counting as a 'fraction' of the degrees of freedom, for a total

of ρ degrees of freedom for the mean. The cAIC allows for comparison of models with different random effects structures, as well as comparison of mixed-effects models with cluster-specific models where the parameters are fixed.

There is a close connection between our conditional Akaike information and the deviance information criterion proposed for Bayesian inference by Spiegelhalter et al. (2002). Under a flat prior, these authors show that the Bayesian measure of model complexity is the same as ρ, and therefore the cAIC takes the same value as a DIC for a mixed-effects model focus with known variances.

The conditional Akaike information has an interesting connection to recent work in the smoothing literature. In *P*-spline smoothing, the observed data y are a smooth function $h(X)$ of the covariates, plus error. The unknown function h can be modelled by maximising the likelihood of the linear mixed model (6), where Z is related to the basis of a smoothing spline, with coefficients b (Eilers & Marx, 1996; Kammann & Wand, 2003). The fit of the model is measured by a mean squared error criterion, related to Mallows' C_p for a linear model, which has exactly the form (9); see Kauermann (2005).

One criticism of AIC is that it is less adept than the BIC at identifying the 'true model' generating the data, in that it tends to fit too complex a model. Our view is that the purpose of modelling is usually not to find the 'true model', which is, in practice, almost always much more complex than the statistical models we consider, but rather to find a good approximation thereof, adequate to the amount of data and the amount of information it contains. This goal is usually formalised as model prediction, and is convincingly advocated by Breiman (2001). It is in this setting that the AIC and its relatives, such as cAIC, are most effective and useful, by finding a balance between the bias and variance of model predictions.

Acknowledgement

This research was supported in part by grants from the U.S. National Institutes of Health and by Frontier Science and Technology Research Foundation, Inc. The authors thank Vincent Carey, Göran Kauermann, Carrie Wager, Ronghui Xu and the participants of the 3rd Euroworkshop for Statistical Modelling, Höhenried, November 2002, for helpful discussions. We gratefully acknowledge the constructive suggestions of the editor and two anonymous referees, which have greatly improved the paper.

Appendix 1

Hodges & Sargent's rho

Hodges & Sargent (2001) show that the maximum likelihood estimator for model (6) can be obtained as a weighted least squares solution from a linear model with added 'pseudo-data'; see also Hodges (1998) and Lee & Nelder (2001). Write (6) formally as $y = X\beta + Zb + \varepsilon$, $0 = -b + b$ or equivalently

$$Y = U\delta + e, \tag{A1}$$

where

$$Y = \begin{pmatrix} y \\ 0_r \end{pmatrix}, \quad \delta = \begin{pmatrix} \beta \\ b \end{pmatrix}, \quad U = \begin{pmatrix} X & Z \\ 0 & -I_r \end{pmatrix}, \quad e = \begin{pmatrix} \varepsilon \\ b \end{pmatrix}.$$

In (A1) the random effects b formally play the dual role of parameter, in δ, and error, in e. Then $\mathrm{var}(e) = \mathrm{diag}(\sigma^2 I_N, G_0) = \sigma^2\,\mathrm{diag}(I_N, D_0)$. Since D_0 is positive definite, for some $r \times r$ matrix Δ we have that $D_0 = (\Delta^{\mathrm{T}}\Delta)^{-1}$; put $\Gamma = \mathrm{diag}(I_N, \Delta)$. In what follows we specify dimensions for the matrices I and 0, and for the vectors 1 and 0, only when they are not clear from the context. Pre-multiply both sides in (A1) by Γ, to obtain

$$Y = M\delta + w, \tag{A2}$$

where

$$\Gamma Y = Y, \quad w = \Gamma e = \begin{pmatrix} \varepsilon \\ \Delta b \end{pmatrix}, \quad M = \Gamma U = \begin{pmatrix} X & Z \\ 0 & -\Delta \end{pmatrix}.$$

We now have $w \sim N(0, \sigma^2 I)$. Formally, the least squares estimator of δ from (A2) is

$$\hat{\delta} = \begin{pmatrix} \hat{\beta} \\ \hat{b} \end{pmatrix} = (M^{\mathrm{T}}M)^{-1}M^{\mathrm{T}}Y = (M^{\mathrm{T}}M)^{-1}\begin{pmatrix} X^{\mathrm{T}} \\ Z^{\mathrm{T}} \end{pmatrix} y. \tag{A3}$$

In (A3), $\hat{\beta}$ coincides with the estimator of Harville (1977, p. 323, formula (3·8)), and $\hat{b}$ with the empirical Bayes estimator $\hat{b} = E(b|y, \hat{\beta})$. The fitted vector is

$$\hat{y} = X\hat{\beta} + Z\hat{b} = H_1 y, \quad H_1 = (X \;\; Z)(M^{\mathrm{T}} \;\; M)^{-1}\begin{pmatrix} X^{\mathrm{T}} \\ Z^{\mathrm{T}} \end{pmatrix}. \tag{A4}$$

Hodges & Sargent (2001) define the effective number of degrees of freedom for model (A1) as $\rho = \mathrm{tr}(H_1)$.

Unlike for a linear model, H_1 is not a projection matrix, but it is the top-left submatrix of the projection matrix $H = M(M^{\mathrm{T}}M)^{-1}M^{\mathrm{T}}$, that is

$$H = \begin{pmatrix} H_1 & H_{12} \\ H_{21} & H_2 \end{pmatrix}, \tag{A5}$$

where H_1 and H_2 are square, of orders N and r respectively, and $H_{12} = H_{21}^{\mathrm{T}}$ is $N \times r$; $\mathrm{rank}(H) = \mathrm{rank}(M) = p + r$.

APPENDIX 2

Proofs of main results

Proof of Theorem 1. We want to estimate the bias correction BC in (18), with $f(.|b)$ and $g(.|\theta, b)$ in the class of models (6). Since σ^2 and D_0 are known, only β is unknown, that is $\theta = \beta$. With a slight abuse of notation, let b be the true random effect in $f(y|b)$, and let β, σ^2 and D_0 be the true parameters in $f(y|b)$. Write $y = \mu + \varepsilon$, $y^* = \mu + \varepsilon^*$, that is $\mu = E_f(y|b) = X\beta + Zb$, and ε, $\varepsilon^* \sim N(0, \sigma^2 I_N)$, independently of each other and of b. Then

$$\begin{aligned}
\mathrm{BC} &= E_y \log g\{y|\hat{\theta}(y), \hat{b}(y)\} - E_y E_{y^*} \log g\{y^*|\hat{\theta}(y), \hat{b}(y)\} \\
&= E_y\left\{-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|y - \hat{y}\|^2\right\} - E_y E_{y^*}\left\{-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|y^* - \hat{y}\|^2\right\} \\
&= E(\|y^* - \hat{y}\|^2 - \|y - \hat{y}\|^2)/(2\sigma^2) \\
&= [E\|y^* - \mu\|^2 + E\|\hat{y} - \mu\|^2 - 2E\{(y^* - \mu)^{\mathrm{T}}(\hat{y} - \mu)\}]/(2\sigma^2) \\
&\quad - [E\|y - \mu\|^2 + E\|\hat{y} - \mu\|^2 - 2E\{(y - \mu)^{\mathrm{T}}(\hat{y} - \mu)\}]/(2\sigma^2) \qquad \text{(A6)} \\
&= E\{(y - \mu)^{\mathrm{T}}(\hat{y} - \mu)\}/\sigma^2 \qquad \text{(A7)} \\
&= E\{(y - \mu)^{\mathrm{T}}\hat{y}\}/\sigma^2. \qquad \text{(A8)}
\end{aligned}$$

In (A6) we expanded the square norm $\|y-\hat{y}\|^2 = \|(y-\mu)-(\hat{y}-\mu)\|^2$, and similarly for $\|y^*-\hat{y}\|^2$; in (A7) we used the facts that, conditionally on b, $E\|y^*-\mu\|^2 = E\|y-\mu\|^2$, and $E\{(y^*-\mu)^{\mathrm{T}}(y-\mu)\} = \{E(y^*-\mu)\}^{\mathrm{T}}E(y-\mu) = 0$; and (A8) uses the fact that, conditionally on b, $E\{(y-\mu)^{\mathrm{T}}\mu\} = 0$. We dropped the distribution of the expectation when there was no risk of confusion.

From (A4) $\hat{y} = H_1 y$, where H_1 only depends on X, Z and D_0, and not on the data y. Therefore,

$$\begin{aligned} \mathrm{BC} &= E\{(y-\mu)^{\mathrm{T}}H_1 y\}/\sigma^2 = E\{(y-\mu)^{\mathrm{T}}H_1(y-\mu)\}/\sigma^2 \\ &= E[\mathrm{tr}\{H_1(y-\mu)(y-\mu)^{\mathrm{T}}\}]/\sigma^2 \\ &= \mathrm{tr}[H_1 E\{(y-\mu)(y-\mu)^{\mathrm{T}}\}]/\sigma^2 = \mathrm{tr}(H_1) \\ &= \rho, \end{aligned} \tag{A9}$$

since $E\{(y-\mu)(y-\mu)^{\mathrm{T}}|b\} = \mathrm{var}(y|b) = \sigma^2 I_N$.

It follows that $\mathrm{cAIC} = -2\log g\{y|\hat{\beta}(y), \hat{b}(y)\} + 2\rho$ is an unbiased estimator of the conditional Akaike information (8). □

Proof of Theorem 2. In the set-up of Theorem 1, assume now that σ^2 is unknown but D_0 is known. Using the 'pseudo-data' augmented vectors $Y = (y^{\mathrm{T}}, 0)^{\mathrm{T}}$, $Y^* = ((y^*)^{\mathrm{T}}, 0)^{\mathrm{T}}$, $\hat{Y} = (\hat{y}^{\mathrm{T}}, (\Delta\hat{b})^{\mathrm{T}})^{\mathrm{T}}$ and $\tilde{\mu} = (\mu^{\mathrm{T}}, 0)^{\mathrm{T}}$, we have, since $\hat{\sigma}^2 = \|Y-\hat{Y}\|^2/N$ (Pinheiro & Bates, 2000, p. 65, eqn (2·12)),

$$\begin{aligned} \mathrm{BC} &= E\left\{\frac{1}{2\hat{\sigma}^2}(\|Y^*-\hat{Y}\|^2 - \|Y-\hat{Y}\|^2)\right\} = E\{\|Y^*-\hat{Y}\|^2/(2\hat{\sigma}^2)\} - N/2 \\ &= -N/2 + E[\|Y^*-\tilde{\mu}\|^2/(2\hat{\sigma}^2) + \|\hat{Y}-\tilde{\mu}\|^2/(2\hat{\sigma}^2) - 2(Y^*-\tilde{\mu})^{\mathrm{T}}\{(\hat{Y}-\tilde{\mu})/(2\hat{\sigma}^2)\}] \\ &= -N/2 + N\sigma^2 E(\hat{\sigma}^{-2})/2 + E\{\|\hat{Y}-\tilde{\mu}\|^2/(2\hat{\sigma}^2)\}, \end{aligned}$$

since $E(Y^*-\tilde{\mu}|b) = 0$ and $E(\|Y^*-\tilde{\mu}\|^2|b) = N\sigma^2$. It is easy to see that $\hat{Y} = HY$. Then

$$\hat{\sigma}^2 = \|Y-\hat{Y}\|^2/N = Y^{\mathrm{T}}(I-H)Y/N.$$

Using (A2) and $(I-H)M = 0$ we obtain further that

$$\hat{\sigma}^2 = w^{\mathrm{T}}(I-H)w/N. \tag{A10}$$

Since $w \sim N(0, \sigma^2 I)$ and $(I-H)$ is a projection matrix of rank $(N-p)$, it follows that, similarly to a linear regression model, $N\hat{\sigma}^2/\sigma^2 \sim \chi^2_{N-p}$. In particular, $E(\hat{\sigma}^{-2}) = N(N-p-2)^{-1}\sigma^{-2}$ (Gelman et al., 2003, Appendix A).

Put $Q = \|\hat{Y}-\tilde{\mu}\|^2/(N\hat{\sigma}^2)$, which is the ratio of two quadratic forms. We need to compute $E(Q)$. Using (A2), we have $\hat{Y} = H(M\delta + w) = M\delta + Hw$; $\tilde{\mu} = M\delta + \mathrm{diag}(0, I)w$. Therefore,

$$\|\hat{Y}-\tilde{\mu}\|^2 = \|Hw - \mathrm{diag}(0, I)w\|^2 = w^{\mathrm{T}}Aw,$$

where $A = \mathrm{diag}(H_1, I-H_2)$ and H_2 is given by (A5). Using (A10) we obtain

$$Q = w^{\mathrm{T}}Aw/\{w^{\mathrm{T}}(I-H)w\}.$$

The spectral decomposition for the projection matrix $(I-H)$ of rank $(N-p)$ yields

$$I-H = L\,\mathrm{diag}(I_0, 0)L^{\mathrm{T}}, \tag{A11}$$

where L is an orthogonal matrix, and $I_0 = \mathrm{diag}(I_{N-p}, 0_p)$.

Put $\sigma^{-1}L^{\mathrm{T}}w = v$, and write $v^{\mathrm{T}} = (v_1^{\mathrm{T}}, v_2^{\mathrm{T}})$, where v_1 contains the first $(N-p)$ components of v, such that $w^{\mathrm{T}}(I-H)w = \sigma^2 v_1^{\mathrm{T}} v_1$. Then $v \sim N(0, I)$. Also,

$$Q = \frac{v^{\mathrm{T}}(L^{\mathrm{T}}AL)v}{v_1^{\mathrm{T}}v_1} = \frac{v_1^{\mathrm{T}}B_1 v_1}{v_1^{\mathrm{T}}v_1} + 2\frac{v_1^{\mathrm{T}}B_{12}v_2}{v_1^{\mathrm{T}}v_1} + \frac{v_2^{\mathrm{T}}B_2 v_2}{v_1^{\mathrm{T}}v_1}, \tag{A12}$$

where

$$B = (L^{\mathrm{T}}AL) = \begin{pmatrix} B_1 & B_{12} \\ B_{12}^{\mathrm{T}} & B_2 \end{pmatrix}$$

is partitioned such that B_1 is a square matrix of order $(N-p)$.

Since v_1 and v_2 are independent, with zero expectations, the middle term in the last expression for Q has expectation zero; for the last term, we have

$$E\{v_2^{\mathrm{T}}B_2v_2/(v_1^{\mathrm{T}}v_1)\} = E(v_2^{\mathrm{T}}B_2v_2)E(v_1^{\mathrm{T}}v_1)^{-1} = \mathrm{tr}(B_2)/(N-p-2). \tag{A13}$$

Finally, in the first term, the ratio is independent of its denominator (Durbin & Watson, 1950; von Neumann, 1941), and therefore

$$E(v_1^{\mathrm{T}}B_1v_1/v_1^{\mathrm{T}}v_1) = E(v_1^{\mathrm{T}}B_1v_1)/E(v_1^{\mathrm{T}}v_1) = \mathrm{tr}(B_1)/(N-p). \tag{A14}$$

Putting (A12)–(A14) together, we obtain

$$E(Q) = \frac{\mathrm{tr}(B_1)}{N-p} + \frac{\mathrm{tr}(B_2)}{N-p-2}. \tag{A15}$$

We show now that $\mathrm{tr}(B_1) = \rho - p$ and $\mathrm{tr}(B_2) = \rho$. Partition

$$L = \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix},$$

where L_{11} is $N \times N$. The $N \times N$ top-left corner of $B = L^{\mathrm{T}}AL$ is $B^{11} = L_{11}^{\mathrm{T}}H_1L_{11} + L_{21}^{\mathrm{T}}(I-H_2)L_{21}$. Then

$$\begin{aligned}
\mathrm{tr}(B_1) &= \mathrm{tr}\ \mathrm{diag}(B_1, 0_p) = \mathrm{tr}(I_0B^{11}I_0) = \mathrm{tr}(I_0L_{11}^{\mathrm{T}}H_1L_{11}I_0) + \mathrm{tr}\{I_0L_{21}^{\mathrm{T}}(I-H_2)L_{21}I_0\} \\
&= \mathrm{tr}\{H_1(L_{11}I_0L_{11}^{\mathrm{T}})\} + \mathrm{tr}\{(I-H_2)(L_{21}I_0L_{21}^{\mathrm{T}})\} \\
&= \mathrm{tr}\{H_1(I-H_1)\} + \mathrm{tr}(I-H_2)^2,
\end{aligned}$$

since from (A11) it follows that $I - H_1 = L_{11}I_0L_{11}^{\mathrm{T}}$ and $I - H_2 = L_{21}I_0L_{21}^{\mathrm{T}}$.

Writing the equation $(I-H) = (I-H)^2$ according to the partitioned H, we obtain that $I - H_1 = (I-H_1)^2 + H_{12}H_{21}$ and $I - H_2 = (I-H_2)^2 + H_{21}H_{12}$, and therefore

$$\begin{aligned}
\mathrm{tr}(B_1) &= \mathrm{tr}\ H_1(I-H_1) + \mathrm{tr}(I-H_2)^2 = \mathrm{tr}\ H_{12}H_{21} - \mathrm{tr}\ H_{21}H_{12} + \mathrm{tr}(I-H_2) \\
&= \mathrm{tr}(I-H_2) = \mathrm{tr}(I-H) - \mathrm{tr}(I-H_1) = (N-p) - (N-\rho) \\
&= \rho - p.
\end{aligned} \tag{A16}$$

Since $\mathrm{tr}(B) = \mathrm{tr}(A) = 2\rho - p$, we have that $\mathrm{tr}(B_2) = \mathrm{tr}(B) - \mathrm{tr}(B_1) = \rho$. Finally,

$$E(Q) = (\rho - p)/(N-p) + \rho/(N-p-2).$$

Replacing $E(Q)$ in BC gives $\mathrm{BC} = K_{\mathrm{MLE}}$, and therefore the cAIC in (10) is unbiased for cAI. □

Proof of Proposition 1. Part (i) is immediate. For part (ii), we have that $\mathrm{tr}(I-H_2) = \rho - p$, for example from (A16). The diagonal elements of $(I-H_2)$ are also diagonal elements of $(I-H)$. As $(I-H)$ is a projection matrix, all its diagonal elements are nonnegative, and therefore $\rho - p \geqslant 0$. Then, the right-most inequality in (ii) then follows from (i). The middle inequality follows from the definition of K_{MLE}, equation (11). Part (iii) is immediate. □

Proof of Theorem 3. With the notation of the proof of Theorem 2, we have

$$\begin{aligned}\text{BC} &= E\left[\frac{1}{2\hat{\sigma}_R^2}\{\|Y^*-\hat{Y}\|^2-\|Y-\hat{Y}\|^2\}\right]\\ &= \frac{N-p}{N}E\left[\frac{1}{2\hat{\sigma}^2}\{\|Y^*-\hat{Y}\|^2-\|Y-\hat{Y}\|^2\}\right]=\frac{N-p}{N}K_{\text{MLE}}\\ &= K_R,\end{aligned}$$

which shows that (14) is unbiased for cAI. □

Proof of (16). Sugiura (1978) and Hurvich & Tsai (1989) showed that AIC_c (2) with K_c given by (3) is unbiased for AI. Just as in the proof of Theorem 3, with obvious notation, for the residual maximum likelihood AIC_c, $\text{BC}=\{(N-p)/N\}K_c=\{(N-p)/(N-p-2)\}(p+1)$, so that (16) is unbiased for AI.

Appendix 3

The one-way analysis of variance example

We show here that the bias correction of the AIC, when $f(.|b)$ and $g(.|\theta,b)$ are in the class of models (17) and σ^2 is known, is $\text{BC}=\rho+2/(n\lambda+1)+o(n^{-1})$.

As in (A8), $\text{BC}=E\{(y-\mu)^{\mathrm{T}}\hat{y}\}/\sigma^2=E(\varepsilon^{\mathrm{T}}\hat{y})/\sigma^2$. We replace $\hat{y}$ by $\hat{H}_1 y$, where $\hat{H}_1=H_1(\hat{\lambda})$. A Taylor expansion of $H_1(\lambda)$ around the true λ yields $\hat{H}_1-H_1=(\hat{\lambda}-\lambda)(dH_1/d\lambda)+o_p(1)$. Then

$$\begin{aligned}\text{BC} &= E(\varepsilon^{\mathrm{T}}H_1 y)/\sigma^2+E\{\varepsilon^{\mathrm{T}}(\hat{H}_1-H_1)y\}/\sigma^2\\ &= \rho+E\left\{(\hat{\lambda}-\lambda)\varepsilon^{\mathrm{T}}\left(\frac{dH_1}{d\lambda}\right)y\right\}\Big/\sigma^2+o(1),\end{aligned}\tag{A17}$$

where ρ in (A17) is obtained as in (A9). The mixed-effects model matrix is $Z=\text{diag}_m(1_n)$, where 1_n is a n-vector of 1's. There is no fixed-effects model matrix X, that is $p=0$. Then $Z^{\mathrm{T}}Z=nI_m$, and $H_1=Z(Z^{\mathrm{T}}Z+\lambda^{-1}I_m^{-1})^{-1}Z^{\mathrm{T}}=(n+\lambda^{-1})^{-1}J$, where $J=ZZ^{\mathrm{T}}=\text{diag}_m(J_n)$, and J_n is the square matrix of order n with all elements equal to 1. It follows that $dH_1/d\lambda=(n\lambda+1)^{-2}J$.

We compute now the maximum likelihood estimator $\hat{\lambda}$, by maximising the marginal loglikelihood for (17). The loglikelihood is $l(\lambda)=-N/2-y^{\mathrm{T}}(I-H_1)y/(2\sigma^2)-m\log(n\lambda+1)/2$, as in Pinheiro & Bates (2000, p. 64), with score function $l'(\lambda)=y^{\mathrm{T}}Jy/\{2\sigma^2(n\lambda+1)^2\}-mn/\{2(n\lambda+1)\}$. This yields $\hat{\lambda}=y^{\mathrm{T}}Jy/(mn^2\sigma^2)-1/n$.

We can now replace the formulae for $dH_1/d\lambda$ and $\hat{\lambda}$ in the second term in the bias correction (A17), which we will denote by U:

$$\begin{aligned}U &= E\left\{(\hat{\lambda}-\lambda)\varepsilon^{\mathrm{T}}\left(\frac{dH_1}{d\lambda}\right)y\right\}\Big/\sigma^2\\ &= E(y^{\mathrm{T}}Jy\varepsilon^{\mathrm{T}}Jy)/\{Nn(n\lambda+1)^2\sigma^4\}-E(\varepsilon^{\mathrm{T}}Jy)/\{n(n\lambda+1)\sigma^2\}.\end{aligned}$$

Put $\bar{\varepsilon}_i=\sum_j\varepsilon_{ij}/n$ for all i, and $\bar{\varepsilon}=(\bar{\varepsilon}_1,\ldots,\bar{\varepsilon}_m)^{\mathrm{T}}$. Note that $\bar{\varepsilon}$ and b are independent, and that $\bar{\varepsilon}\sim N(0,\sigma^2n^{-1}I_m)$. Then $Z^{\mathrm{T}}\varepsilon=n\bar{\varepsilon}$, $Z^{\mathrm{T}}y=Z^{\mathrm{T}}(Zb+\varepsilon)=n(b+\bar{\varepsilon})$. Since $J=ZZ^{\mathrm{T}}$ we obtain $E(\varepsilon^{\mathrm{T}}Jy)=E(\varepsilon^{\mathrm{T}}ZZ^{\mathrm{T}}y)=E\{n^2\bar{\varepsilon}^{\mathrm{T}}(b+\bar{\varepsilon})\}=N\sigma^2$.

The numerator of the first term in U is

$$\begin{aligned}E\{(y^{\mathrm{T}}Jy)(\varepsilon^{\mathrm{T}}Jy)\} &= E[\{n^2(b+\bar{\varepsilon})^{\mathrm{T}}(b+\bar{\varepsilon})\}\{n^2(\bar{\varepsilon}^{\mathrm{T}}\bar{\varepsilon}+b^{\mathrm{T}}\bar{\varepsilon})\}]\\ &= n^4E\{(b^{\mathrm{T}}b)(\bar{\varepsilon}^{\mathrm{T}}\bar{\varepsilon})\}+2n^4E(b^{\mathrm{T}}\bar{\varepsilon}\bar{\varepsilon}^{\mathrm{T}}b)+n^4E(\bar{\varepsilon}^{\mathrm{T}}\bar{\varepsilon}\bar{\varepsilon}^{\mathrm{T}}\bar{\varepsilon});\end{aligned}\tag{A18}$$

the remaining terms have zero expectations. The first term in (A18) is

$$n^4E(b^{\mathrm{T}}b)E(\bar{\varepsilon}^{\mathrm{T}}\bar{\varepsilon})=n^4\,\text{tr}\{\text{var}(b)\}\,\text{tr}\{\text{var}(\bar{\varepsilon})\}=n^3m^2\lambda\sigma^4.$$

The second term in (A18) is $2n^4 E(b^{\mathrm{T}}\bar{\varepsilon}\bar{\varepsilon}^{\mathrm{T}}b) = 2n^4 E\,\mathrm{tr}(b^{\mathrm{T}}\bar{\varepsilon}\bar{\varepsilon}^{\mathrm{T}}b) = 2n^4\,\mathrm{tr}\{E(bb^{\mathrm{T}})E(\bar{\varepsilon}\bar{\varepsilon}^{\mathrm{T}})\} = 2n^3 m\lambda\sigma^4$. For the last term in (A18), first put $W = n\sigma^{-2}\bar{\varepsilon}^{\mathrm{T}}\bar{\varepsilon}$. Then $W \sim \chi^2_m$, and

$$n^4 E\{(\bar{\varepsilon}^{\mathrm{T}}\bar{\varepsilon})(\bar{\varepsilon}^{\mathrm{T}}\bar{\varepsilon})\} = n^2\sigma^4 E(W^2) = n^2\sigma^4\{\mathrm{var}(W) + E^2(W)\} = n^2\sigma^4(2m + m^2).$$

Putting everything together we obtain, after simplification, $U = 2/(n\lambda + 1)$, and, from (A17), $\mathrm{BC} = \rho + 2(n\lambda + 1)^{-1} + o(n^{-1})$.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory*, Ed. B. N. Petrov and F. Csaki, pp. 267–81. Budapest: Akademia Kiado.

BREIMAN, L. (2001). Statistical modeling: the two cultures (with Discussion). *Statist. Sci.* **16**, 199–231.

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

BURNHAM, K. P. & WHITE, G. C. (2002). Evaluation of some random effects methodology applicable to bird ringing data. *J. Appl. Statist.* **29**, 245–64.

DELEEUW, J. (1992). Introduction to Akaike (1973) 'Information theory and an extension of the maximum likelihood principle'. In *Breakthroughs in Statistics*, vol. **1**, Ed. S. Kotz and N. L. Johnson, pp. 599–609. New York: Springer.

DURBIN, J. & WATSON, G. S. (1950). Testing for serial correlation in least squares regression. *Biometrika* **37**, 409–28.

EILERS, P. H. C. & MARX, B. D. (1996). Flexible smoothing with B-splines and penalties (with Discussion). *Statist. Sci.* **11**, 89–121.

FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

GELMAN, A., CARLIN, J. B., STERN, H. S. & RUBIN, D. B. (2003). *Bayesian Data Analysis*, 2nd ed. London: CRC Press.

HARVILLE, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Statist. Assoc.* **72**, 320–38.

HASTIE, T. & TIBSHIRANI, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

HODGES, J. S. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics (with Discussion). *J. R. Statist. Soc.* B **60**, 497–536.

HODGES, J. S. & SARGENT, D. J. (2001). Counting degrees of freedom in hierarchical and other richly parameterized models. *Biometrika* **88**, 367–79.

HURVICH, C. M. & TSAI, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

KAMMANN, E. E. & WAND, M. P. (2003). Geoadditive models. *Appl. Statist.* **52**, 1–18.

KAUERMANN, G. (2005). A note on smoothing parameter selection for penalised spline smoothing. *J. Statist. Plan. Infer.* **127**, 53–69.

LAIRD, N. M. & WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–74.

LEE, Y. & NELDER, J. A. (2001). Hierarchical generalized linear models: A synthesis of generalised linear models, random effect models and structured dispersions. *Biometrika* **88**, 987–1006.

LUNN, D., WAKEFIELD, J., ANDREW, T., BEST, N. & SPIEGELHALTER, D. (1999). *PKBugs Users Guide*. London: Imperial College of Science, Technology and Medicine.

PINHEIRO, J. C. & BATES, D. M. (2000). *Mixed Effects Models in S and S-PLUS*. New York: Springer.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. & VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit (with Discussion). *J. R. Statist. Soc.* B **64**, 583–639.

SUGIURA, N. (1978). Further analysis of the data by Akaike's information criteria and the finite corrections. *Commun. Statist.* A**7**, 13–26.

VON NEUMANN, J. (1941). Distribution of the ratio of the mean square successive difference to the variance. *Ann. Math. Statist.* **12**, 367–95.

WAKEFIELD, J. C., SMITH, A. F. M., RACINE-POON, A. & GELFAND, A. E. (1994). Bayesian analysis of linear and nonlinear population models by using the Gibbs sampler. *Appl. Statist.* **43**, 201–21.

[*Received November* 2002. *Revised September* 2004]