Marginal and Conditional Akaike Information Criteria in Linear Mixed Models

Sonja Greven[1] and Thomas Kneib[2] - Oral Presentation

[1] Department of Biostatistics, Johns Hopkins University, USA; sgreven@jhsph.edu
[2] Department of Mathematics, Carl-von-Ossietzky-Universität Oldenburg, Germany; thomas.kneib@uni-oldenburg.de

Abstract: In linear mixed models, the Akaike information criterion (AIC) is often used to decide on the inclusion of a random effect. An important special case is the choice between linear and nonparametric regression models estimated using mixed model penalized splines. We investigate the behavior of two commonly used versions of the AIC, derived either from the implied marginal model or the conditional model formulation. We find that the marginal AIC is not asymptotically unbiased for twice the expected relative Kullback-Leibler distance, and favors smaller models without random effects. For the conditional AIC, it is computationally costly for large sample sizes to correct for estimation uncertainty. However, ignoring it, as is common practice, induces a bias that yields the following behavior: Whenever the random effects variance estimate is positive (even if small), the more complex model is preferred. We illustrate our results in a simulation study, and investigate their impact in modeling childhood malnutrition in Zambia.

Keywords: Kullback-Leibler information; model selection; penalized splines; random effect; variance component.

1 Introduction

Linear mixed models are increasingly used to model complex data structures. Using penalized splines, they can combine model components such as non-linear or spatial effects, interaction surfaces or varying coefficients with cluster-specific random effects. The growing flexibility of such regression models then makes the question of model selection increasingly important. The Akaike information criterion (Akaike, 1973) is often used to decide on the inclusion of random effects in linear mixed models. A common special case when using penalized splines is the decision between a linear and a nonparametric function for a covariate effect. An AIC based on the implied marginal likelihood is typically used (mAIC). Vaida & Blanchard (2005) proposed an AIC derived from the conditional model formulation (cAIC). They argue that the cAIC is more appropriate when focus is on the random effects, such as in the case of penalized splines, as the random effects are then additional parameters that are estimated subject to a distributional constraint rather than a tool for modeling the correlation structure. However, both AIC versions are commonly used. We investigate the behavior of both for the selection of random effects in the linear mixed model

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

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are known design matrices, $\boldsymbol{\beta}$ is a fixed parameter vector, $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$ are assumed to be independent, $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{D})$ and $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I_n})$.

2 The marginal AIC

The AIC can be generally defined as

$$\begin{aligned} AIC = &-2\log f(\boldsymbol{y}|\widehat{\boldsymbol{\psi}}(\boldsymbol{y})) + 2\mathrm{E}_{\boldsymbol{y}}[\log f(\boldsymbol{y}|\widehat{\boldsymbol{\psi}}(\boldsymbol{y})) - \log f(\boldsymbol{y}|\boldsymbol{\psi_K})] \\ &+2\mathrm{E}_{\boldsymbol{y}}[\mathrm{E}_{\boldsymbol{z}}[\log f(\boldsymbol{z}|\boldsymbol{\psi_K}) - \log f(\boldsymbol{z}|\widehat{\boldsymbol{\psi}}(\boldsymbol{y}))]], \end{aligned} \tag{2}$$

where $f(\boldsymbol{y}|\widehat{\boldsymbol{\psi}}\boldsymbol{(y)})$ is the maximized likelihood, and $\boldsymbol{\psi}$ are k unknown parameters with values $\boldsymbol{\psi_K}$ minimizing the Kullback-Leibler distance (Kullback & Leibler, 1951) between the true underlying joint density $g(\cdot)$ and the family of approximating candidate models $f(\cdot|\boldsymbol{\psi}), \boldsymbol{\psi} \in \boldsymbol{\Psi}$,

$$K(f_\psi, g) = \int \{\log(g(z)) - \log(f_\psi(z))\}g(z)dz = \mathrm{E}_z[\log(g(z)) - \log(f_\psi(z))].$$

$K(f_\psi, g)$ can be viewed as a measure of distance between $g(\cdot)$ and $f(\cdot|\boldsymbol{\psi})$, $\boldsymbol{\psi} \in \boldsymbol{\Psi}$. As the AIC is unbiased for twice the expected relative Kullback-Leibler distance, minimizing (2) can be seen as minimizing the average distance of an approximating model to the underlying truth.
In standard cases, certain regularity conditions are fulfilled, including that observations are independent and identically distributed, and the parameter space (up to a change of coordinates) is R^k. Then, the last two terms in (2) reduce to $2k$ asymptotically. This is the AIC commonly used.
The marginal AIC (mAIC) in the linear mixed model uses the likelihood of the implied marginal model $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V})$ with $\boldsymbol{V} = \boldsymbol{I_n} + \boldsymbol{ZDZ'}$. The number of estimable parameters then is $p+q$, with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and q the number of unknown parameters $\boldsymbol{\theta}$ in $\boldsymbol{V}$. Thus, the mAIC is defined as

$$mAIC = -2\log(f(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})) + 2(p+q).$$

Now, we can show (Greven & Kneib, 2008) that due to the marginal correlation structure in $\boldsymbol{y}$ in (1) and the constraints on $\boldsymbol{\theta}$ (variances have to be non-negative, and more generally, $\boldsymbol{D}$ has to be positive semi-definite), the last two terms in (2) are smaller than $2(p+q)$ as well as not independent of the true values in $\boldsymbol{\theta}$. Consequently, the mAIC is positively biased, and favors smaller models without random effects.

3 The conditional AIC

Vaida & Blanchard (2005) define the conditional AIC (cAIC) as

$$cAIC = -2\log(f(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{b}}, \widehat{\boldsymbol{\theta}})) + 2(\rho + 1),$$

where $f(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{b}}, \widehat{\boldsymbol{\theta}})$ is the maximized conditional likelihood (conditioning on $\boldsymbol{b}$ as well as on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$), $\widehat{\boldsymbol{b}}$ is the best linear unbiased predictor of $\boldsymbol{b}$, $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$ are the maximum likelihood (ML) or restricted maximum likelihood (REML) estimates of $(\boldsymbol{\beta}, \boldsymbol{\theta})$, and

$$\rho = \text{trace}\left(\begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{D}_*{}^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{Z} \end{pmatrix} \right).$$

This definition of ρ corresponds to the trace of the hat matrix, and is connected to the effective degrees of freedom definition known from smoothing. The authors assume $\boldsymbol{D}_* = \sigma^{-2}\boldsymbol{D}$ to be known, but suggest using $\widehat{\rho}$ with estimated $\boldsymbol{D}_*$ otherwise, arguing that the difference is negligible for large n. We will call this the conventional or simplified cAIC in the following. Liang et al. (2008) propose a corrected cAIC, accounting for estimation of $\boldsymbol{D}_*$. For known σ^2, they replace ρ by $\Phi_0 = \text{trace}\,(\partial\widehat{\boldsymbol{y}}/\partial\boldsymbol{y})$, where $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}\widehat{\boldsymbol{b}}$. For unknown σ^2, the effective degrees of freedom Φ_1 involve even second derivatives,

$$\Phi_1 = \frac{\widetilde{\sigma}^2}{\widehat{\sigma}^2}\text{trace}\left(\frac{\partial\widehat{\boldsymbol{y}}}{\partial\boldsymbol{y}}\right) + \widetilde{\sigma}^2(\widehat{\boldsymbol{y}} - \boldsymbol{y})'\frac{\partial\widehat{\sigma}^{-2}}{\partial\boldsymbol{y}} + \frac{1}{2}\widetilde{\sigma}^4\text{trace}\left(\frac{\partial^2\widehat{\sigma}^{-2}}{\partial\boldsymbol{y}\partial\boldsymbol{y}'}\right),$$

where $\widetilde{\sigma}^2$ is an estimate for the true error variance. As these derivatives are not available in closed form, numerical approximations using n respectively $2n$ additional model fits have to be used. This can be prohibitive in large samples. In our application ($n = 1600$, 64 models to compare), we estimated the necessary computation time to be about 110 days. As the authors in their simulations find only small differences between conventional and corrected cAIC, we investigate whether the often used simplified cAIC is a computationally feasible alternative - especially when n is large. In this case, the computational cost of the corrected cAIC can be too high, and consistent estimators should yield precise variance estimates. Unlike Liang et al. (2008a), who concentrated on estimating the effective degrees of freedom, we focus on the performance for differentiating between zero and non-zero random effects variances.

Surprisingly, we can show (Greven & Kneib, 2008) that ignoring estimation uncertainty in $\boldsymbol{D}_*$ for the simplified cAIC results in the following interesting behavior (for simplicity, we focus on the case of one unknown variance component, i.e. $\boldsymbol{D} = \tau^2\boldsymbol{\Sigma}$ with known $\boldsymbol{\Sigma}$): When $\widehat{\tau}^2 = 0$, the cAICs of the models including and excluding $\boldsymbol{b}$ agree, i.e. there is a tie. When $\widehat{\tau}^2 > 0$, the cAIC prefers the larger model including $\boldsymbol{b}$, regardless of the size of

$\widehat{\tau}^2$. The simplified cAIC thus is not a useful decision rule, as it does not give guidance on when an estimated variance is large enough to warrant inclusion of the random effect in the model, or small enough to justify exclusion of the random effect from the model.

The principal difficulty of the simplified cAIC is that the degrees of freedom in the cAIC are estimated from the same data as the model parameters. This leads to a bias that results in a preference for larger models. This behavior has its analogy in the AIC itself. Use of the maximized log-likelihood for model choice would always result in the largest model being chosen. The underlying over-optimism in the model fit is due to the parameter estimates being obtained from the same data which is the argument of the log-likelihood. The AIC corrects for this bias and is a truly predictive quantity. A similar mechanism is at play here. While the correct bias correction term in our case cannot be derived analytically, Liang et al. (2008a) circumvent the problem using numerical derivatives. In a sense, their bias correction term is measuring the sensitivity of results to new data, similar in spirit to other predictive criteria such as generalized cross validation (GCV). Unfortunately, this comes at the price of computational complexity (and some numerical instability) that is comparable to leave-one-out cross validation. More work is clearly needed here.

4 Simulations

First, we compare a linear model $y_i = \beta_0+\beta_1 x_i+\varepsilon_i$ with the nonparametric regression model $y_i = m(x_i) + \varepsilon_i$, modeled using penalized splines in the mixed model framework. Thus, the comparison corresponds to selecting a random effect modeling deviations of $m(\cdot)$ from linearity. The true functions are chosen as (see Figure 1)

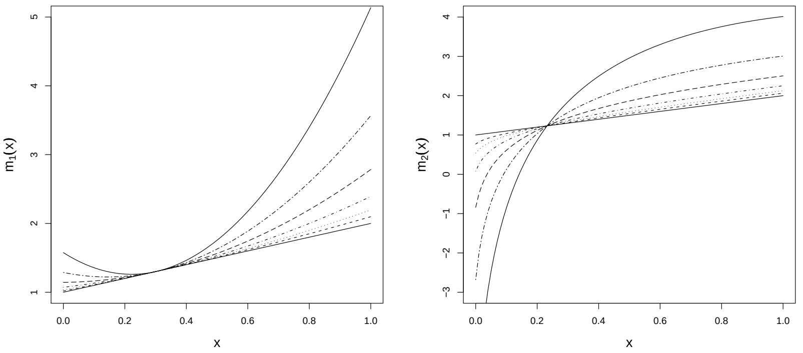
$$\begin{aligned} m_1(x) &= 1 + x + 2d(0.3 - x)^2, \\ m_2(x) &= 1 + x + d(\log(0.1 + 5x) - x), \\ m_3(x) &= 1 + x + 0.3d(\cos(0.5\pi + 2\pi x) - 2x) \end{aligned}$$

with varying non-linearity parameter $d = 0, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2$, where $d = 0$ corresponds to linearity. The sample size is taken as $n = 30, 50, 100,$ 200. The error variance is set to $\sigma^2 = 1$, and x is chosen equidistantly from the interval $[0, 1]$.

Second, we compare a random intercept and a common intercept model, varying random intercept variance, number and size of clusters. In case of a tie between models, we intrinsically decide on the smaller model.

The simplified cAIC gives a much larger proportion of decisions for the larger model than the mAIC, with the corrected cAIC in-between (Figure 2). While the AIC for nested models in standard settings corresponds to a likelihood ratio test with asymptotic level $\alpha = 0.157$, α is much smaller for the mAIC (as low as 0.01 in our simulations), much larger for the simplified cAIC (up to 0.49), and more similar for the corrected cAIC (0.07 to 0.40).

FIGURE 1. Functions $m_1(\cdot)$ and $m_2(\cdot)$ for different values of the non-linearity parameter d.



As predicted from our theoretical results, the simplified cAIC chooses the larger model when $\widehat{\tau}^2 > 0$, and gives a tie when $\widehat{\tau}^2 = 0$. Thus, α here simply corresponds to the proportion of non-zero variance estimates given a true zero variance, which for penalized splines is about 20% for ML estimation, and more than 35% for REML estimation, and which approaches 50% for the random intercept model.

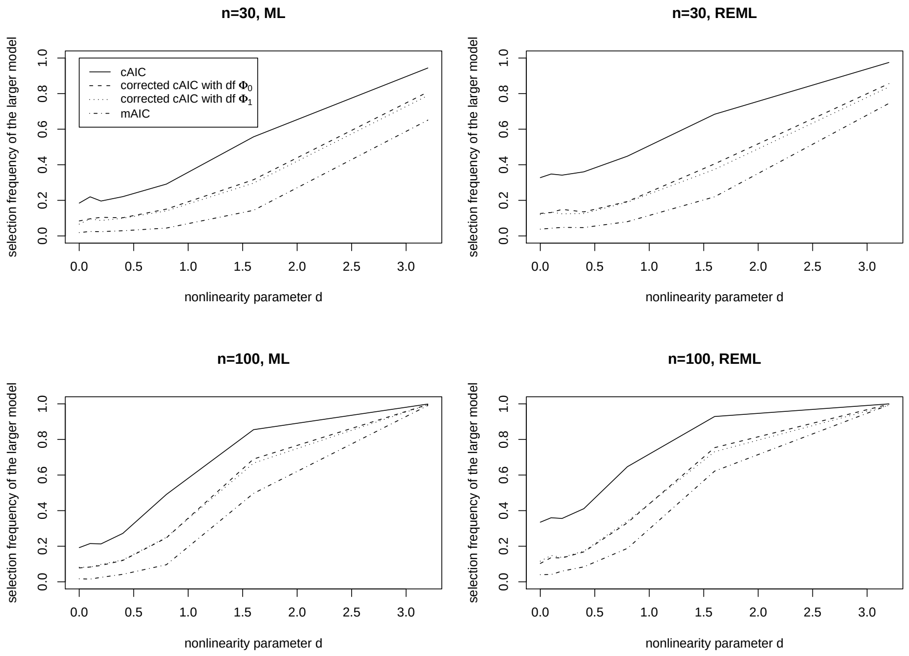
In contrast, the mAIC does not show this behavior, and in particular never yields equality under the linear and the non-linear model due to the additional parameter count for the variance parameter. The corrected cAIC with $\Phi_0 + 1$ still results in a large number of ties, which disappear when using Φ_1.

The corrected cAIC often favors the more complex model even when $\widehat{\tau}^2 = 0$ due to numerical problems. Especially for Φ_1 using second derivatives, the numerical approximation fails in some cases, resulting in spurious estimated degrees of freedom. Overall, $\Phi_0 + 1$ approximates Φ_1 rather well (see Figure 2), but is numerically much more stable.

5 Childhood malnutrition in Zambia

We investigate implications of our theoretical findings for model choice in practice. We are interested in modelling the Z-score, measuring chronic undernutrition (stunting) as insufficient height for age, for 1600 children from the 1992 Zambia Demographic and Health Survey. The available predictors were 1) categorical/binary: child's gender, mother's employment status and education 2) spatial: residential district and 3) continuous: duration of breastfeeding, child's age, mother's age, height and body mass index. Due to computational cost of the corrected cAIC in our large data set, we focused only on the mAIC and the simplified cAIC for selecting a random

FIGURE 2. Selection frequencies of the larger, non-linear model in our simulations for function $m_1(\cdot)$.

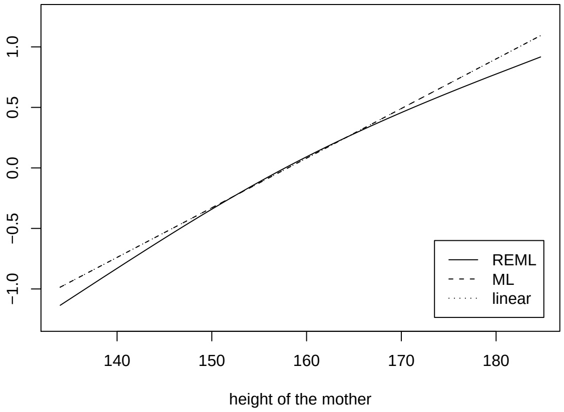


district intercept, and linear or non-linear effects for the continuous variables. Categorical and binary variables were modeled parametrically, giving a total of 64 models to choose from.

To illustrate our findings in a simple example, consider the model with height of the mother as the single predictor. Using maximum likelihood estimation, the estimated effect is linear (Figure 3). This results in a tie for the cAIC, while the mAIC clearly prefers the smaller linear model due to the additional parameter count for the variance parameter. Using REML estimation, the estimated effect is slightly non-linear. While the mAIC still prefers the smaller, linear model, the cAIC as expected chooses the larger, non-linear model, despite the estimated non-linearity being quite small.

For the overall comparison of all 64 models, let a tie in the cAIC be indicative of a choice of the simpler model. Then, cAIC and mAIC for both ML and REML agree on the overall best model including a random district intercept, linear effects for age, height and body mass index of the mother, and non-linear effects of child's age and duration of breastfeeding (Figure 4). As mAIC and simplified cAIC are biased in opposite directions, agreement between the two indicates optimality of this final model.

FIGURE 3. Estimated effect of height of the mother (in cm) on the Z-score measuring chronic undernutrition of children in Zambia.



6 Discussion

We investigated the behavior of mAIC and cAIC for selecting random effects in linear mixed models. This corresponds to interesting model choice questions, including decision on non-linearity of effects, constancy of varying coefficients, or the necessity for a random intercept. We found the mAIC to be biased towards simpler models without random effects. The bias is dependent on the setting and the true value of the random effects variance. For the cAIC, it is essential to correct for estimation uncertainty in the unknown random effects covariance matrix. Ignoring the uncertainty, while common and computationally attractive, leads to selection of the random effect whenever it is not estimated to be exactly zero. This problem is independent of the sample size and does not vanish asymptotically. More research is needed to obtain numerically feasible and robust versions of the corrected cAIC, and to extend methodology to generalized linear mixed models.

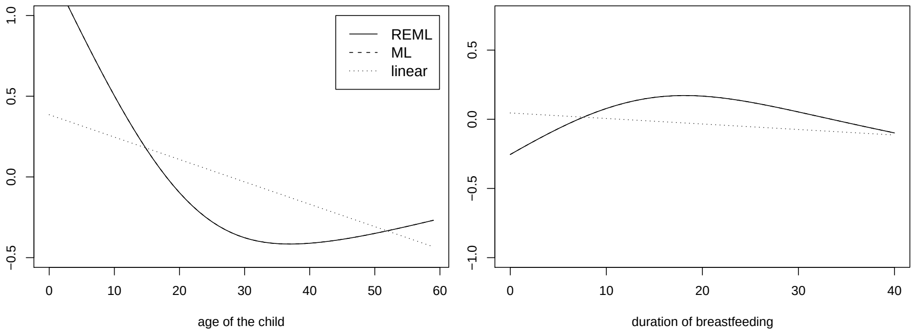
For a longer working paper on our results including proofs, please see Greven & Kneib (2008).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: *2nd International Symposium on Information Theory.* 267-281, Akademiai Kiado.

Greven, S. and Kneib, T. (2008). On the Behavior of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models. *Johns*

FIGURE 4. Estimated effects of age of the child and duration of breastfeeding (in months) on the Z-score measuring chronic undernutrition of children in Zambia.



Hopkins University, Department of Biostatistics Working Papers, Paper 179. http://www.bepress.com/jhubiostat/paper179/

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22**, 79–86.

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

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