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

Thomas Kneib

Institute of Statistics and Econometrics
Georg-August-University Göttingen

Department of Statistics
Ludwig-Maximilians-University Munich

joint work with Sonja Greven
Department of Biostatistics, Johns Hopkins University

10.2.2009
Overview

• Linear and additive mixed models.

• Akaike’s information criterion (AIC).

• Marginal AIC

• Conditional AIC

• Application: Childhood malnutrition in Zambia
Linear and Additive Mixed Models

• Mixed models form a very useful class of regression models with general form

\[ y = X\beta + Zb + \varepsilon \]

where \( \beta \) are usual regression coefficients while \( b \) are random effects with distributional assumption

\[
\begin{bmatrix} \varepsilon \\ b \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 I & 0 \\ 0 & D \end{bmatrix} \right).
\]

• Denote the vector of all unknown variance parameters as \( \theta \).

• In the following, we will concentrate on mixed models with only one variance component where

\[
b \sim \mathcal{N}(0, \tau^2 I) \quad \text{or} \quad b \sim \mathcal{N}(0, \tau^2 \Sigma)
\]

with \( \Sigma \) known and therefore \( \theta = (\sigma^2, \tau^2) \).
Special case I: Random intercept model for longitudinal data

\[ y_{ij} = \mathbf{x}_{ij}' \beta + b_i + \varepsilon_{ij}, \quad j = 1, \ldots, J_i, \quad i = 1, \ldots, I, \]

where \( i \) indexes individuals while \( j \) indexes repeated observations on the same individual.

The random intercept \( b_i \) accounts for shifts in the individual level of response trajectories and therefore also for intra-subject correlations.

Extended models include further random (covariate) effects, leading to random slopes.
• Special case II: Penalised spline smoothing for nonparametric function estimation

\[ y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \]

where \( m(x) \) is a smooth, unspecified function.

• Approximating \( m(x) \) in terms of a spline basis of degree \( d \) leads (for example) to the truncated power series representation

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

where \( \kappa_1, \ldots, \kappa_K \) denotes a sequence of knots.

• The spline approximation leads to a piecewise polynomial fit of degree \( d \) on the intervals defined by the knots under appropriate smoothness restrictions.
● Penalised estimation to avoid overly wiggly function estimates:

\[(y - X\beta - Zb)'(y - X\beta - Zb) + \lambda b'b \rightarrow \min_{\beta, b}\]

where \(X\) and \(Z\) correspond to design matrices obtained from the truncated power series representation.

● The smoothness of the curve is determined by the smoothing parameter \(\lambda\).

● Equivalent to assuming the random effect distribution \(b \sim N(0, \tau^2 I)\) and setting the smoothing parameter to

\[\lambda = \frac{\sigma^2}{\tau^2}.
\]

● Works also for other basis choices (e.g. B-splines) and other types of flexible modelling components (varying coefficients, surfaces, spatial effects, etc.).
• Additive mixed models consist of a combination of random effects and flexible modelling components such as penalised splines.

• Example: Childhood malnutrition in Zambia.

• Determine the nutritional status of a child in terms of a Z-score.

• We consider chronic malnutrition measured in terms of insufficient height for age (stunting), i.e.

\[ z_{\text{score}}_i = \frac{\text{cheight}_i - \text{med}}{s}, \]

where \( \text{med} \) and \( s \) are the median and standard deviation of (age-stratified) height in a reference population.
Additive mixed model for stunting:

\[ z_{score_i} = x_i' \beta + m_1(cage_i) + m_2(cfeed_i) + m_3(mage_i) + m_4(mbmi_i) \]

\[ + m_5(mheight_i) + b_{s_i} + \varepsilon_i, \]

with covariates

- **csex**: gender of the child (1 = male, 0 = female)
- **cfeed**: duration of breastfeeding (in months)
- **cage**: age of the child (in months)
- **mage**: age of the mother (at birth, in years)
- **mheight**: height of the mother (in cm)
- **mbmi**: body mass index of the mother
- **medu**: education of the mother (1 = no education, 2 = primary school, 3 = elementary school, 4 = higher)
- **mwork**: employment status of the mother (1 = employed, 0 = unemployed)
- **s**: residential district (54 districts in total)

The random effect \( b_{s_i} \) captures spatial variability induced by unobserved spatially varying covariates.
• Marginal perspective on a mixed model:

\[ y \sim N(X\beta, V) \]

where

\[ V = \sigma^2 I + ZDZ' \]

• Interpretation: The random effects induce a correlation structure and therefore enable a proper statistical analysis of correlated data.

• Conditional perspective on a mixed model:

\[ y|b \sim N(X\beta + Zb, \sigma^2 I). \]

• Interpretation: Random effects are additional regression coefficients (for example subject-specific effects in longitudinal data) that are estimated subject to a regularisation penalty.
• Best linear unbiased estimates / predictions in the linear mixed model:

\[
\hat{\beta} = \left( X'V^{-1}X \right)^{-1} X'V^{-1}y, \quad \hat{b} = DZ'V^{-1}(y - X\hat{\beta}).
\]

• Unknown variance parameters \( \theta \) are estimated using maximum likelihood (ML) or restricted maximum likelihood (REML).

• Interest in the following is on model choice in linear mixed models with the special form

\[
D = \text{blockdiag}(\tau_1^2 \Sigma_1, \ldots, \tau_q^2 \Sigma_q)
\]

\((q \ \text{independent random effects})\) for known correlation matrices \( \Sigma_1, \ldots, \Sigma_q \) and in particular in models with only one variance component such as

\[
D = \tau^2 I.
\]
• Without loss of generality, we consider the comparison of

\[ M_1 : D = \text{blockdiag}(\tau_1^2 \Sigma_1, \ldots, \tau_q^2 \Sigma_q) \]

and

\[ M_2 : D = \text{blockdiag}(\tau_1^2 \Sigma_1, \ldots, \tau_{q-1}^2 \Sigma_{q-1}). \]

• The two models are nested since \( M_1 \) reduces to \( M_2 \) when \( \tau_q^2 = 0 \).

• Random Intercept: \( \tau_q^2 > 0 \) versus \( \tau_q^2 = 0 \) corresponds to the inclusion and exclusion of the random intercept and therefore to the presence or absence of intra-individual correlations.

• Penalised splines: \( \tau_q^2 > 0 \) versus \( \tau_q^2 = 0 \) differentiates between a spline model and a simple polynomial model. In particular, we can compare linear versus nonlinear models.
Akaike Information Criterion

- Data $y$ generated from a true underlying model described in terms of density $g(\cdot)$.

- Approximate the true model by a parametric class of models $f_\psi(\cdot) = f(\cdot; \psi)$.

- Measure the discrepancy between a model $f_\psi(\cdot)$ and the truth $g(\cdot)$ by the Kullback-Leibler distance

$$K(f_\psi, g) = \int [\log(g(z)) - \log(f_\psi(z))] g(z) dz$$

$$= E_z [\log(g(z)) - \log(f_\psi(z))].$$

where $z$ is an independent replicate following the same distribution as $y$.

- Note that $K(f_\psi, g) \geq 0$ and $K(f_\psi, g) = 0$ iff $f_\psi = g$ almost everywhere.
• Decision rule: Out of a sequence of models, choose the one that minimises \( K(f_\psi, g) \).

• In practice, the parameter \( \psi \) will have to be estimated as \( \hat{\psi}(y) \) for the different models.

• To focus on average properties not depending on a specific data realisation, minimise the expected Kullback-Leibler distance

\[
E_y[K(f_{\hat{\psi}(y)}, g)] = E_y[E_z[\log(g(z)) - \log(f_{\hat{\psi}(y)}(z))]]
\]

• Since \( g(\cdot) \) does not depend on the data, this is equivalent to minimising

\[
-2 E_y[E_z[\log(f_{\hat{\psi}(y)}(z))]] \quad (1)
\]

(the expected relative Kullback-Leibler distance).
The best available estimate for (1) is given by

$$-2\log(f_{\hat{\psi}(y)}(y)).$$

While (1) is a predictive quantity depending on both the data \(y\) and an independent replication \(z\), the density and the parameter estimate are evaluated for the same data \(y\).

⇒ Introduce a correction term.

Let \(\tilde{\psi}\) denote the parameter vector minimising the Kullback-Leibler distance.

Then

$$AIC = -2\log(f_{\hat{\psi}(y)}(y)) + 2E_y[\log(f_{\hat{\psi}(y)}(y)) - \log(f_{\tilde{\psi}(y)})]$$

$$+ 2E_y[E_z[\log(f_{\tilde{\psi}(z)}) - \log(f_{\hat{\psi}(y)}(z))]]$$

is unbiased for (1).
• Consider the regularity conditions

  – \( \psi \) is a \( k \)-dimensional parameter with parameter space \( \Psi = \mathbb{R}^k \) (possibly achieved by a change of coordinates).

  – \( y \) consists of independent and identically distributed replications \( y_1, \ldots, y_n \).

• In this case, the AIC simplifies since

\[
2 E_z \left[ \log(f_{\tilde{\psi}}(z)) - \log(f_{\hat{\psi}}(y)(z)) \right] \overset{a}{\sim} \chi^2_k,
\]

\[
2 \left[ \log(f_{\hat{\psi}}(y)(y)) - \log(f_{\tilde{\psi}}(y)) \right] \overset{a}{\sim} \chi^2_k
\]

and therefore an (asymptotically) unbiased estimate for (1) is given by

\[
AIC = -2 \log(f_{\hat{\psi}}(y)(y)) + 2k.
\]

• In linear mixed models, two variants of AIC are conceivable based on either the marginal or the conditional distribution.
The marginal AIC relies on the marginal model

$$y \sim N(X\beta, V)$$

and is defined as

$$mAIC = -2l(y|\hat{\beta}, \hat{\theta}) + 2(p + q),$$

where the marginal likelihood is given by

$$l(y|\hat{\beta}, \hat{\theta}) = -\frac{1}{2} \log(|\hat{V}|) - \frac{1}{2}(y - X\hat{\beta})'\hat{V}^{-1}(y - X\hat{\beta})$$

and $p = \dim(\beta), q = \dim(\theta)$. 

On the Behavior of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models
The conditional AIC relies on the conditional model

\[ y|b \sim N(\mathbf{X}\beta + \mathbf{Z}b, \sigma^2 \mathbf{I}) \]

and is defined as

\[ cAIC = -2l(y|\hat{\beta}, \hat{b}, \hat{\theta}) + 2(\rho + 1), \]

where

\[ l(y|\hat{\beta}, \hat{b}, \hat{\theta}) = -\frac{n}{2} \log(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(y - \mathbf{X}\hat{\beta} - \mathbf{Z}\hat{b})'(y - \mathbf{X}\hat{\beta} - \mathbf{Z}\hat{b}) \]

is the conditional likelihood and

\[ \rho = \text{trace} \left( \left( \begin{array}{cc} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \sigma^2 \mathbf{D} \end{array} \right)^{-1} \left( \begin{array}{cc} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{array} \right) \right) \]

are the effective degrees of freedom (trace of the hat matrix).
• The conditional AIC seems to be recommended when the model shall be used for predictions with the same set of random effects (for example in penalised spline smoothing).

• The marginal AIC is more plausible when observations with new random effects shall be predicted (e.g. new individuals in longitudinal studies).

• Still, both variants have been considered in both situations and seem to work reasonably well (see for example Wager, Vaida & Kauermann, 2007).
Marginal AIC

• Consider the special case of comparing

\[ M_1 : y = X\beta + Zb + \epsilon, \quad b \sim N(0, \tau^2 I) \]

deciding on the inclusion of a random effect.

• Corresponds to the decision \( \tau^2 > 0 \) (\( M_1 \)) versus \( \tau^2 = 0 \) (\( M_2 \)).
• Model $M_1$ is preferred over $M_2$ when

$$mAIC_1 < mAIC_2 \iff -2l(y|\hat{\beta}_1, \hat{\tau}^2, \hat{\sigma}_1^2) + 2(p + 2) < -2l(y|\hat{\beta}_2, 0, \hat{\sigma}_2^2) + 2(p + 1)$$

$$\iff 2l(y|\hat{\beta}_1, \hat{\tau}^2, \hat{\sigma}_1^2) - 2l(y|\hat{\beta}_2, 0, \hat{\sigma}_2^2) > 2.$$ 

• The left hand side is simply the test statistic for a likelihood ratio test on $\tau^2 = 0$ versus $\tau^2 > 0$.

• Under standard asymptotics, we would have

$$2l(y|\hat{\beta}_1, \hat{\tau}^2, \hat{\sigma}_1^2) - 2l(y|\hat{\beta}_2, 0, \hat{\sigma}_2^2) \sim_{H_0} \chi^2_1$$

and the marginal AIC would have a type 1 error of

$$P(\chi^2_1 > 2) \approx 0.1572992$$

• Common interpretation: AIC selects rather too many than too few effects.
• In contrast to the regularity conditions for likelihood ratio tests, we are testing on the boundary of the parameter space!

• The likelihood ratio test statistic is no longer $\chi^2$-distributed but (approximately) follows a mixture of a point mass in zero and a scaled $\chi^2_1$ variable.

• The point mass in zero corresponds to the probability

$$P(\hat{\tau}^2 = 0)$$

that is typically larger than 50%.

• Similar difficulties appear in more complex models with several variance components when deciding on zero variances.
• The classical assumptions underlying the derivation of AIC are also not fulfilled.

• The high probability of estimating a zero variance yields a bias towards simpler models:
  – The marginal AIC is positively biased for twice the expected relative Kullback-Leibler-Distance.
  – The bias is dependent on the true unknown parameters in the random effects covariance matrix $D$ and this dependence does not vanish asymptotically.
  – Compared to an unbiased criterion, the marginal AIC favors smaller models excluding random effects.

• This contradicts the usual intuition that the AIC picks rather too many than too few effects.
• Simulated example: \( y_i = m(x) + \varepsilon \) where

\[
m(x) = 1 + x + 2d(0.3 - x)^2.
\]

• The parameter \( d \) determines the amount of nonlinearity.
On the Behavior of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models
Vaida & Blanchard (2005) have shown that the conditional AIC is asymptotically unbiased for the expected relative Kullback Leibler distance for given random effects covariance matrix $D$.

If $D$ is estimated consistently, one would hope that their result carries over to the case of estimated $\hat{D}$.

Simulation results seem to indicate that this is not the case.
On the Behavior of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models
• Surprising result of the simulation study: The complex model including the random effect is chosen whenever $\hat{\tau}^2 > 0$.

• If $\hat{\tau}^2 = 0$, the conditional AICs of the simple and the complex model coincide (despite the additional parameters included in the complex model).

• The observed phenomenon could be shown to be a general property of the conditional AIC:

\[
\begin{align*}
\hat{\tau}^2 > 0 & \iff cAIC(\hat{\tau}^2) < cAIC(0) \\
\hat{\tau}^2 = 0 & \iff cAIC(\hat{\tau}^2) = cAIC(0).
\end{align*}
\]

• Principal difficulty: The degrees of freedom in the cAIC are estimated from the same data as the model parameters.
Liang et al. (2008) propose a corrected conditional AIC, where the degrees of freedom $\rho$ are replaced by

$$\Phi_0 = \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i} = \text{trace} \left( \frac{\partial \hat{y}}{y} \right)$$

if $\sigma^2$ is known.

For unknown $\sigma^2$, they propose to replace $\rho + 1$ by

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

where $\tilde{\sigma}^2$ is an estimate for the true error variance.
On the Behavior of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models
• The corrected conditional AIC shows **satisfactory theoretical properties**.

• However, it is **computationally cumbersome**:
  
  – The first and second derivative are not available in closed form and must be approximated numerically (by adding small perturbations to the data).
  
  – Numerical approximations require $n$ and $2n$ model fits. In our example, computing the corrected conditional AICs would take about 110 days.
  
  – In addition, the numerical derivatives were found to be instable in some situations (for example the random intercept model with small cluster sizes).
Application: Childhood Malnutrition in Zambia

• Model equation:

\[ z_{\text{score}}_i = \mathbf{x}_i' \beta + m_1(\text{cage}_i) + m_2(\text{cfeed}_i) + m_3(\text{mage}_i) + m_4(\text{mbmi}_i) + m_5(\text{mheight}_i) + b_{s_i} + \varepsilon_i. \]

• Parametric effects are not subject to model selection.

\[ \Rightarrow 2^6 = 64 \text{ models to consider in the model comparison.} \]
On the Behavior of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models
• The six best fitting models:

<table>
<thead>
<tr>
<th>cfeed</th>
<th>cage</th>
<th>mage</th>
<th>mheight</th>
<th>mbmi</th>
<th>district</th>
<th>ML</th>
<th>REML</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cAIC</td>
<td>mAIC</td>
</tr>
<tr>
<td>14</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>4125.78</td>
<td>4151.10</td>
</tr>
<tr>
<td>34</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>4125.78</td>
<td>4153.10</td>
</tr>
<tr>
<td>36</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>4125.78</td>
<td>4153.10</td>
</tr>
<tr>
<td>38</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>4125.78</td>
<td>4153.10</td>
</tr>
<tr>
<td>54</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>4125.78</td>
<td>4155.10</td>
</tr>
<tr>
<td>56</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>4125.78</td>
<td>4155.10</td>
</tr>
<tr>
<td>58</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>4125.78</td>
<td>4155.10</td>
</tr>
<tr>
<td>64</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>4125.78</td>
<td>4157.10</td>
</tr>
</tbody>
</table>
Summary

- The marginal AIC suffers from the same theoretical difficulties as likelihood ratio tests on the boundary of the parameter space.

- The marginal AIC is biased towards simpler models excluding random effects.

- The conventional conditional AIC tends to select too many variables.

- Whenever a random effects variance is estimated to be positive, the corresponding effect will be included.

- The corrected conditional AIC rectifies this difficulty but comes at a high computational price.
Open questions:

- Is there a computationally advantageous version / representation of the corrected conditional AIC?
- Can the marginal AIC be corrected?
- Is there a working likelihood ratio test based on the corrected conditional AIC?
References


• A place called home:

  http://www.stat.uni-muenchen.de/~kneib