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

Thomas Kneib

Department of Mathematics
Carl von Ossietzky University Oldenburg

Sonja Greven

Department of Biostatistics
Johns Hopkins University

CARL VON OSSIETZKY universität OLDENBURG
FAKULTÄT V
MATHEMATIK UND NATURWISSENSCHAFTEN

Dortmund, 23.3.2010

Akaike Information Criterion

- Most commonly used model choice criterion for comparing parametric models.
- Definition:

$$\text{AIC} = -2l(\hat{\boldsymbol{\psi}}) + 2k.$$

  where $l(\hat{\boldsymbol{\psi}})$ is the log-likelihood evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\psi}}$ for the unknown parameter vector $\boldsymbol{\psi}$ and $k = \dim(\boldsymbol{\psi})$ is the number of parameters.
- Properties:
  - Compromise between model fit and model complexity.
  - Allows to compare non-nested models.
  - Selects rather too many than too few variables in variable selection problems.

- Data $\boldsymbol{y}$ generated from a true underlying model described in terms of density $g(\cdot)$.
- Approximate the true model by a parametric class of models $f_{\boldsymbol{\psi}}(\cdot) = f(\cdot; \boldsymbol{\psi})$.
- Measure the discrepancy between a model $f_{\boldsymbol{\psi}}(\cdot)$ and the truth $g(\cdot)$ by the Kullback-Leibler distance

$$
\begin{aligned}
K(f_{\boldsymbol{\psi}}, g) &= \int \left[\log(g(\boldsymbol{z})) - \log(f_{\boldsymbol{\psi}}(\boldsymbol{z}))\right] g(\boldsymbol{z}) d\boldsymbol{z} \\
&= \mathrm{E}_{\boldsymbol{z}} \left[\log(g(\boldsymbol{z})) - \log(f_{\boldsymbol{\psi}}(\boldsymbol{z}))\right].
\end{aligned}
$$

  where $\boldsymbol{z}$ is an independent replicate following the same distribution as $\boldsymbol{y}$.
- Decision rule: Out of a sequence of models, choose the one that minimises $K(f_{\boldsymbol{\psi}}, g)$.

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

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

$$\mathrm{E}_{\boldsymbol{y}}[K(f_{\hat{\boldsymbol{\psi}}(\boldsymbol{y})}, g)] = \mathrm{E}_{\boldsymbol{y}}[\mathrm{E}_{\boldsymbol{z}}\left[\log(g(\boldsymbol{z})) - \log(f_{\hat{\boldsymbol{\psi}}(\boldsymbol{y})}(\boldsymbol{z}))\right]]$$

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

$$-2\,\mathrm{E}_{\boldsymbol{y}}[\mathrm{E}_{\boldsymbol{z}}[\log(f_{\hat{\boldsymbol{\psi}}(\boldsymbol{y})}(\boldsymbol{z}))]] \qquad (1)$$

(the expected relative Kullback-Leibler distance).

- The best available estimate for (1) is given by

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

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

  $\Rightarrow$ Introduce a correction term.

- Consider the regularity conditions
  - $\boldsymbol{\psi}$ is a k-dimensional parameter with parameter space $\boldsymbol{\Psi} = \mathbb{R}^k$ (possibly achieved by a change of coordinates).
  - $\boldsymbol{y}$ consists of independent and identically distributed replications $y_1, \dots, y_n$.

- In this case, an (asymptotically) unbiased estimate for (1) is given by

$$\text{AIC} = -2\log(f_{\hat{\boldsymbol{\psi}}(\boldsymbol{y})}(\boldsymbol{y})) + 2k.$$

Linear Mixed Models

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

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\beta}$ are usual regression coefficients while $\boldsymbol{b}$ are random effects with distributional assumption

$$\begin{bmatrix} \boldsymbol{\varepsilon} \\ \boldsymbol{b} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma^2 \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{D} \end{bmatrix} \right).$$

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

$$\boldsymbol{b} \sim \mathrm{N}(\mathbf{0}, \tau^2 \boldsymbol{I}) \quad \text{or} \quad \boldsymbol{b} \sim \mathrm{N}(\mathbf{0}, \tau^2 \boldsymbol{\Sigma})$$

with $\boldsymbol{\Sigma}$ known.

- Special case I: Random intercept model for longitudinal data

$$y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + b_i + \varepsilon_{ij}, \quad j = 1, \dots, J_i,\ i = 1, \dots, I,$$

  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.

- Special case II: Penalised spline smoothing for nonparametric function estimation

$$y_i = m(x_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

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 $\kappa_1, \dots, \kappa_K$ denotes a sequence of knots.

- Assume random effects distribution $\boldsymbol{b} \sim \mathrm{N}(\mathbf{0}, \tau^2 \boldsymbol{I})$ for the basis coefficients of truncated polynomials to enforce smoothness.

- Marginal perspective on a mixed model:

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V})$$

where

$$\boldsymbol{V} = \sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}'$$

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

- Conditional perspective on a mixed model:

$$\boldsymbol{y}|\boldsymbol{b} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \sigma^2 \boldsymbol{I}).$$

- Interpretation: Random effects are additional regression coefficients (for example subject-specific effects in longitudinal data) that are estimated subject to a regularisation penalty.

- Interest in the following is on the selection of random effects: Compare

$$M_1: \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}, \quad \boldsymbol{b} \sim \mathrm{N}(\boldsymbol{0}, \tau^2\boldsymbol{\Sigma})$$

  and

$$M_2: \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

- Equivalent: Compare model with random effects ($\tau^2 > 0$) and without random effects ($\tau^2 = 0$).
- Random Intercept: $\tau^2 > 0$ versus $\tau^2 = 0$ corresponds to the inclusion and exclusion of the random intercept and therefore to the presence or absence of intra-individual correlations.
- Penalised splines: $\tau^2 > 0$ versus $\tau^2 = 0$ differentiates between a spline model and a simple polynomial model. In particular, we can compare linear versus nonlinear models.

Akaike Information Criteria in Linear Mixed Models

- 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

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X\beta}, \boldsymbol{V})$$

and is defined as

$$\mathrm{mAIC} = -2l(\boldsymbol{y}|\hat{\boldsymbol{\beta}}, \hat{\tau}^2, \hat{\sigma}^2) + 2(p+2),$$

where the marginal likelihood is given by

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

and $p = \dim(\boldsymbol{\beta})$.

- The conditional AIC relies on the conditional model

$$\boldsymbol{y}|\boldsymbol{b} \sim \mathrm{N}(\boldsymbol{X\beta} + \boldsymbol{Zb}, \sigma^2 \boldsymbol{I})$$

and is defined as

$$\mathrm{cAIC} = -2l(\boldsymbol{y}|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{b}}, \hat{\tau}^2, \sigma^2) + 2(\rho + 1),$$

where

$$l(\boldsymbol{y}|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{b}}, \hat{\tau}^2, \sigma^2) = -\frac{n}{2}\log(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}})$$

is the conditional likelihood and

$$\rho = \mathrm{tr}\left(\begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{Z} + \sigma^2/\tau^2\boldsymbol{\Sigma} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{Z} \end{pmatrix}\right)$$

are the effective degrees of freedom (trace of the hat matrix).

Marginal AIC

- Model M_1 $(\tau^2 > 0)$ is preferred over M_2 $(\tau^2 = 0)$ when

$$\begin{aligned} \text{mAIC}_1 < \text{mAIC}_2 \quad &\Leftrightarrow \quad -2l(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_1, \hat{\tau}^2, \hat{\sigma}_1^2) + 2(p+2) < -2l(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_2, 0, \hat{\sigma}_2^2) + 2(p+1) \\ &\Leftrightarrow \quad 2l(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_1, \hat{\tau}^2, \hat{\sigma}_1^2) - 2l(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_2, 0, \hat{\sigma}_2^2) > 2. \end{aligned}$$

- 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

$$2l(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_1, \hat{\tau}^2, \hat{\sigma}_1^2) - 2l(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_2, 0, \hat{\sigma}_2^2) \overset{a,H_0}{\sim} \chi_1^2$$

  and the marginal AIC would have a type 1 error of

$$P(\chi_1^2 > 2) \approx 0.1572992$$

- Common interpretation: AIC selects rather too many than too few effects.

- In contrast to the regularity conditions for likelihood ratio tests, τ^2 is on the boundary of the parameter space for model M_2.
- The classical assumptions underlying the derivation of AIC are also not fulfilled.
- Consequences:
  - 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 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.

Conditional AIC

- Vaida & Blanchard (2005) have shown that the conditional AIC from above is asymptotically unbiased for the expected relative Kullback Leibler distance for given random effects covariance matrix.
- Intuition: Result should carry over when using a consistent estimate.
- Surprisingly, this is not the case: The complex model including the random effect is chosen whenever $\hat{\tau}^2 > 0$:

$$\hat{\tau}^2 > 0 \quad \Leftrightarrow \quad \mathrm{cAIC}(\hat{\tau}^2) < \mathrm{cAIC}(0)$$

$$\hat{\tau}^2 = 0 \quad \Leftrightarrow \quad \mathrm{cAIC}(\hat{\tau}^2) = \mathrm{cAIC}(0).$$

- 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 ρ are replaced by the estimate

$$\hat{\rho} = \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i} = \operatorname{tr}\left(\frac{\partial \hat{\boldsymbol{y}}}{\boldsymbol{y}}\right).$$

- The resulting corrected conditional AIC shows satisfactory theoretical properties.
- However, it is computationally cumbersome:
  - Liang et al. suggested to approximate the derivatives numerically (by adding small perturbations to the data).
  - Numerical approximations require n and $2n$ model fits. In an application with 1,600 Observations and 64 candidate models, computing the corrected conditional AICs would take about 110 days.
- We have developed a closed form representation of $\hat{\rho}$ that is available almost instantaneously.

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 and is now available in closed form.

- References:

  - Greven, S. & Kneib, T. (2009): On the Behavior of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models. Under Revision for *Biometrika*.
  - Liang, H., Wu, H. & Zou, G. (2008): A note on conditional AIC for linear mixed-effects models. *Biometrika* 95, 773–778.
  - Vaida, F. & Blanchard, S. (2005): Conditional Akaike information for mixed-effects models. *Biometrika* 92, 351–370.

- A place called home:

`http://www.staff.uni-oldenburg.de/thomas.kneib`