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

Lübeck, 5.12.2009

Outline

- Akaike Information Criterion
- Linear Mixed Models
- Marginal Akaike Information Criterion
- Conditional Akaike Information Criterion
- Application: Childhood Malnutrition in Nigeria

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

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

(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.

- Let $\tilde{\boldsymbol{\psi}}$ denote the parameter vector minimising the Kullback-Leibler distance.
- Then

$$\begin{aligned} \text{AIC} &= -2\log(f_{\hat{\boldsymbol{\psi}}(\boldsymbol{y})}(\boldsymbol{y})) + 2\,\text{E}_{\boldsymbol{y}}[\log(f_{\hat{\boldsymbol{\psi}}(\boldsymbol{y})}(\boldsymbol{y})) - \log(f_{\tilde{\boldsymbol{\psi}}}(\boldsymbol{y}))] \\ &\quad + 2\,\text{E}_{\boldsymbol{y}}[\text{E}_{\boldsymbol{z}}[\log(f_{\tilde{\boldsymbol{\psi}}}(\boldsymbol{z})) - \log(f_{\hat{\boldsymbol{\psi}}(\boldsymbol{y})}(\boldsymbol{z}))]] \end{aligned}$$

  is unbiased for (1).

- Consider the regularity conditions
  - ψ is a k-dimensional parameter with parameter space $\mathbf{\Psi} = \mathbb{R}^k$ (possibly achieved by a change of coordinates).
  - $\boldsymbol{y}$ consists of independent and identically distributed replications $y_1, \ldots, y_n$.
- In this case, the AIC simplifies since

$$\begin{aligned} 2\left[\log(f_{\hat{\boldsymbol{\psi}}(\boldsymbol{y})}(\boldsymbol{y})) - \log(f_{\tilde{\boldsymbol{\psi}}}(\boldsymbol{y}))\right] &\overset{a}{\sim} \chi^2_k, \\ 2\,\mathrm{E}_{\boldsymbol{z}}\left[\log(f_{\tilde{\boldsymbol{\psi}}}(\boldsymbol{z})) - \log(f_{\hat{\boldsymbol{\psi}}(\boldsymbol{y})}(\boldsymbol{z}))\right] &\overset{a}{\sim} \chi^2_k \end{aligned}$$

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

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

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

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

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

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

with $\boldsymbol{\Sigma}$ known.

- Special case I: Random intercept model for longitudinal data

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

  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.

- 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.

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

- 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.

$$zscore_i = \frac{cheight_i - med}{s},$$

where med and s are the median and standard deviation of (age-stratified) height in a reference population.

- Additive mixed model for stunting:

$$\begin{aligned} zscore_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + m_1(cage_i) + m_2(cfeed_i) + m_3(mage_i) + m_4(mbmi_i) \\ &\quad + m_5(mheight_i) + b_{s_i} + \varepsilon_i, \end{aligned}$$

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:

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

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

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

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

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

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).

- 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).

Marginal AIC

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

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

- 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 likelihood ratio test statistic is no longer χ^2-distributed but (approximately) follows a mixture of a point mass in zero and a scaled χ^2_1 variable.
- The point mass in zero corresponds to the probability

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

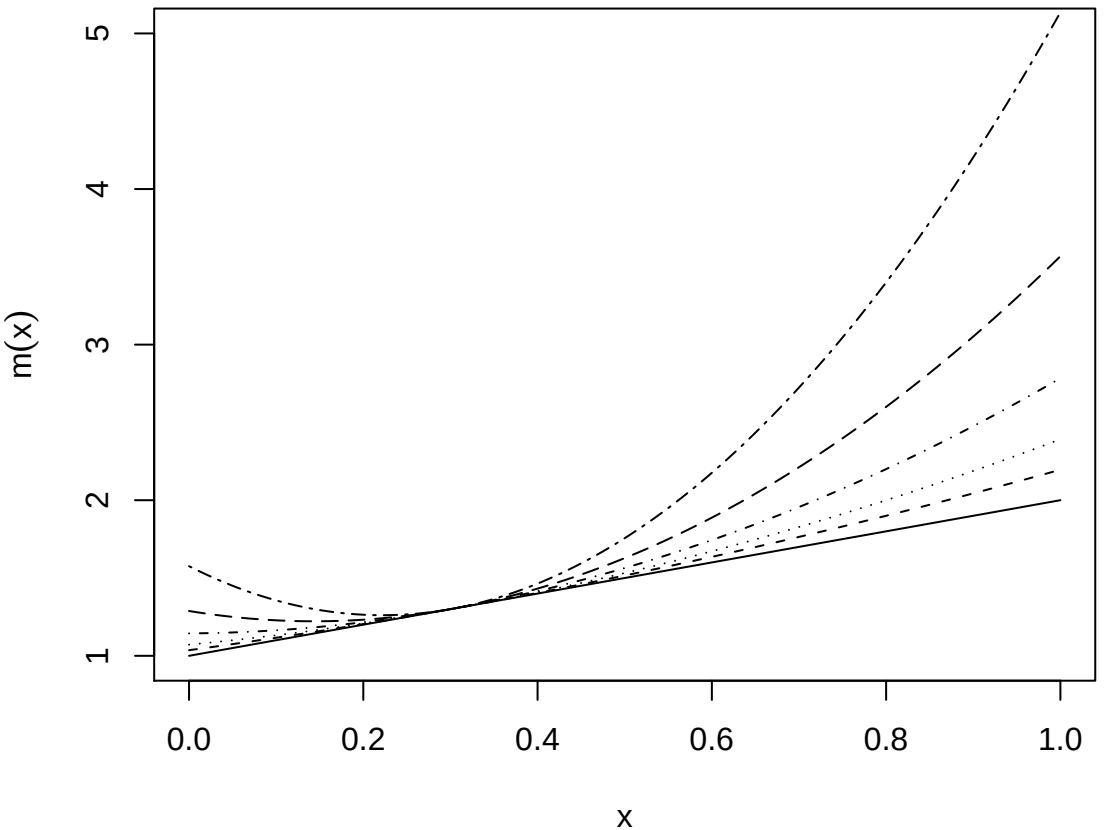
  that is typically larger than 50%.
- Similar difficulties appear in more complex models with several variance components when deciding on zero variances.

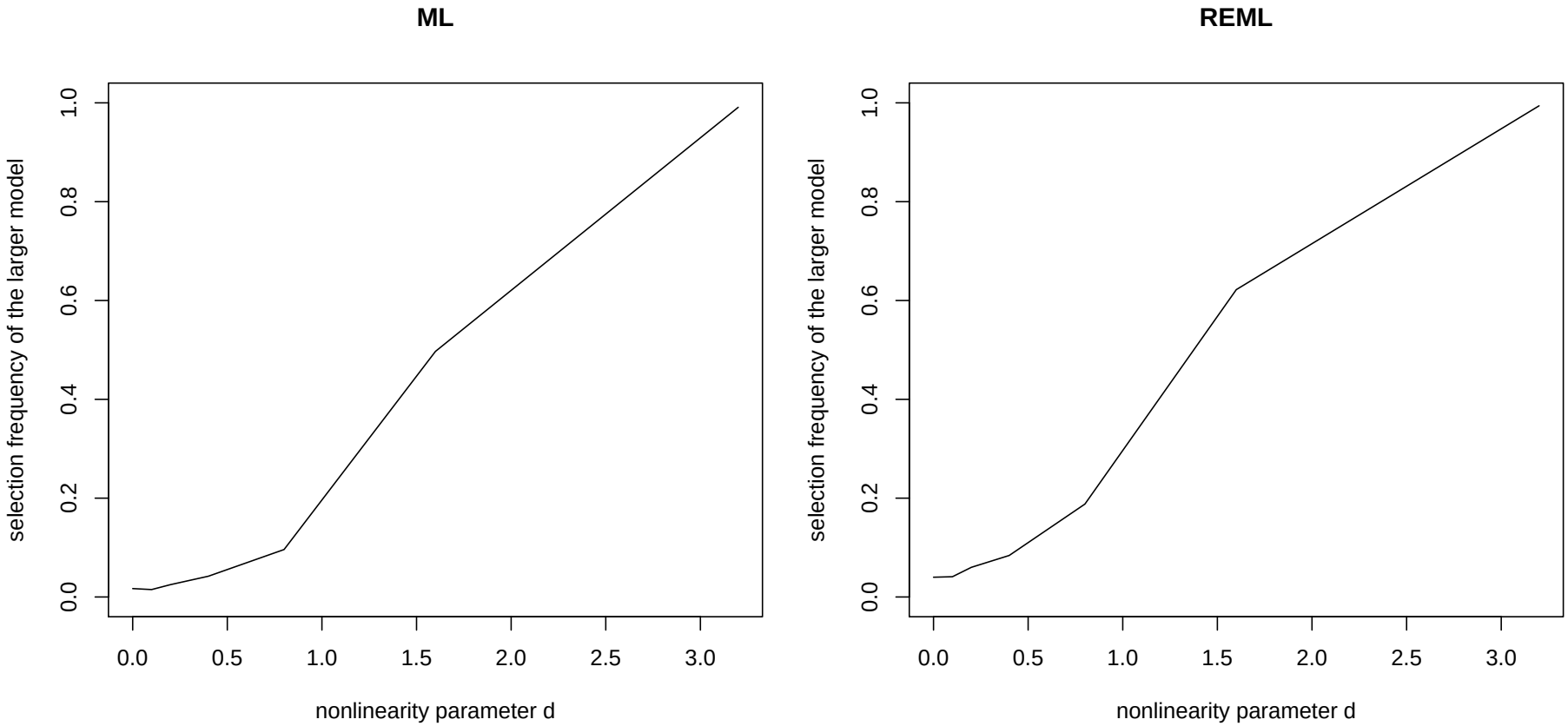
- The classical assumptions underlying the derivation of AIC are also not fulfilled.
- The high probability of estimating a zero variance yields a bias towards simpler models:
  - The marginal AIC is positively biased for twice the expected relative Kullback-Leibler-Distance.
  - The bias is dependent on the true unknown parameters in the random effects covariance matrix 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

$$m(x) = 1 + x + 2d(0.3 - x)^2.$$

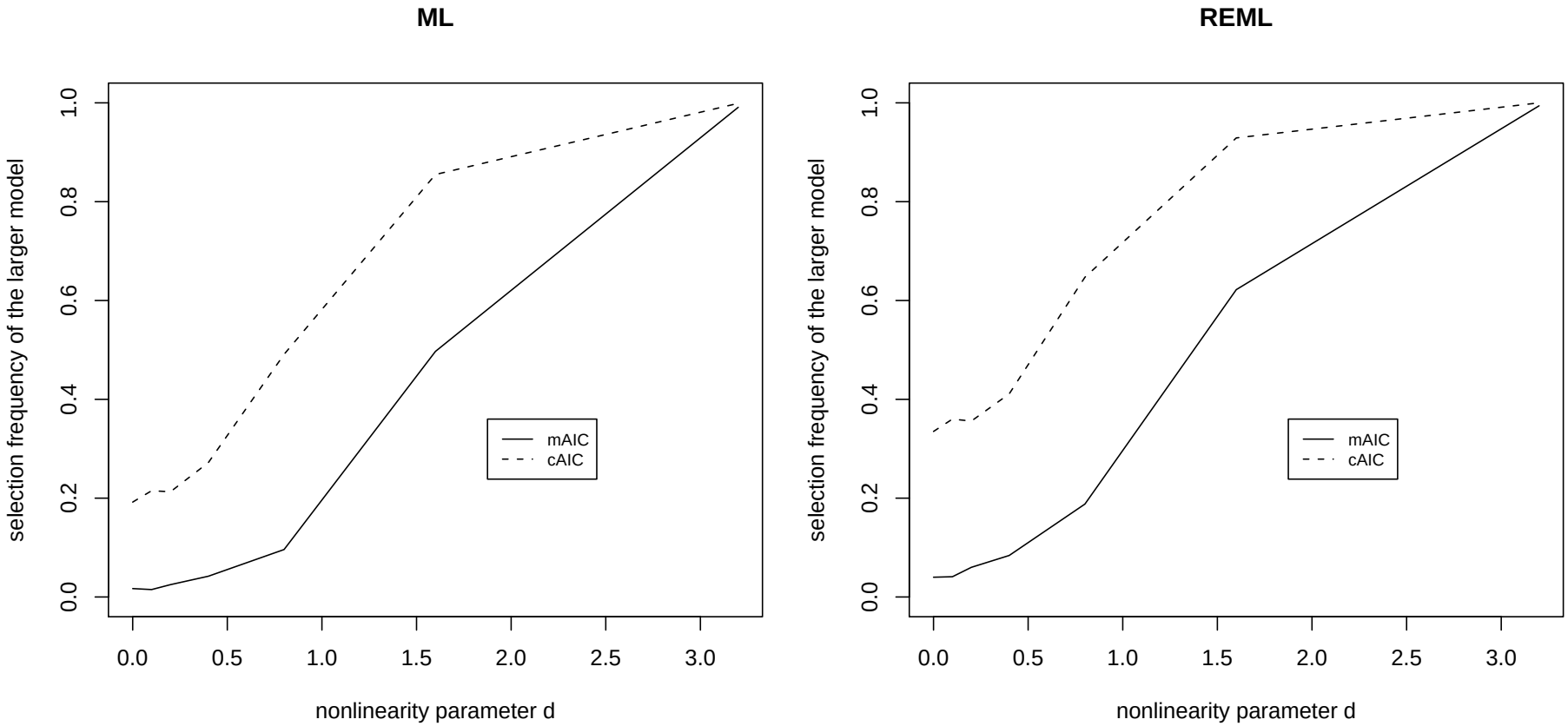
- The parameter d determines the amount of nonlinearity.




ML
REML
selection frequency of the larger model
nonlinearity parameter d
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.5
2.0
2.5
3.0

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.
- Simulation results indicate that this is not the case.


ML
REML
selection frequency of the larger model
nonlinearity parameter d
mAIC
cAIC
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

- 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 can be shown to be a general property of the conditional AIC:

$$\begin{aligned} \hat{\tau}^2 > 0 \quad &\Leftrightarrow \quad \text{cAIC}(\hat{\tau}^2) < \text{cAIC}(0) \\ \hat{\tau}^2 = 0 \quad &\Leftrightarrow \quad \text{cAIC}(\hat{\tau}^2) = \text{cAIC}(0). \end{aligned}$$

- 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

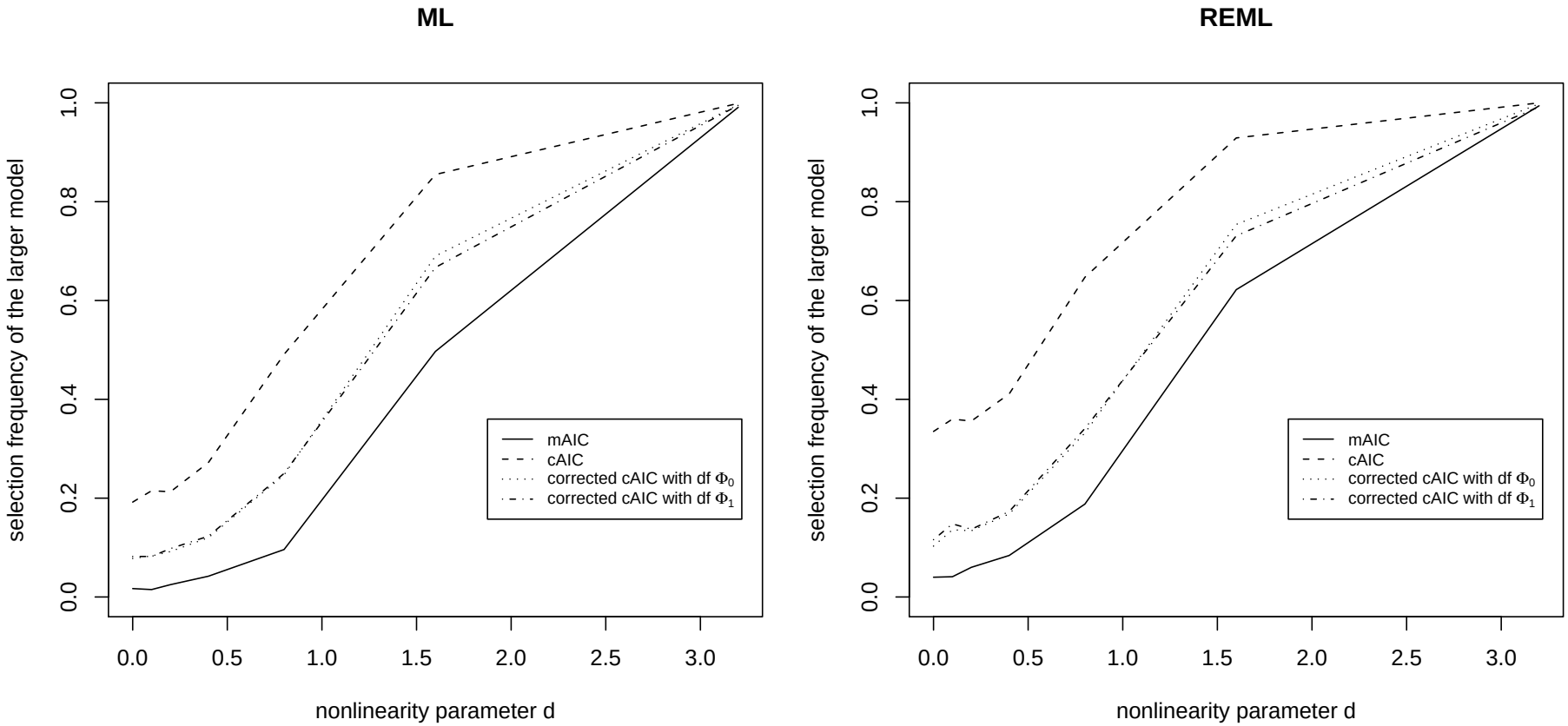
$$\Phi_0 = \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i} = \operatorname{tr}\left(\frac{\partial \hat{\boldsymbol{y}}}{\boldsymbol{y}}\right)$$

  if σ^2 is known.

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

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

  where $\tilde{\sigma}^2$ is an estimate for the true error variance.


ML
REML
selection frequency of the larger model
nonlinearity parameter d
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.5
2.0
2.5
3.0
mAIC
cAIC
corrected cAIC with df Φ_0
corrected cAIC with df Φ_1

- The 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 our application, 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).
- We have developed a closed form representation of Φ_0 that is available almost instantaneously.

Application: Childhood Malnutrition in Zambia

- Model equation:

$$\begin{aligned} zscore_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + m_1(cage_i) + m_2(cfeed_i) + m_3(mage_i) + m_4(mbmi_i) \\ &\quad + m_5(mheight_i) + b_{s_i} + \varepsilon_i. \end{aligned}$$

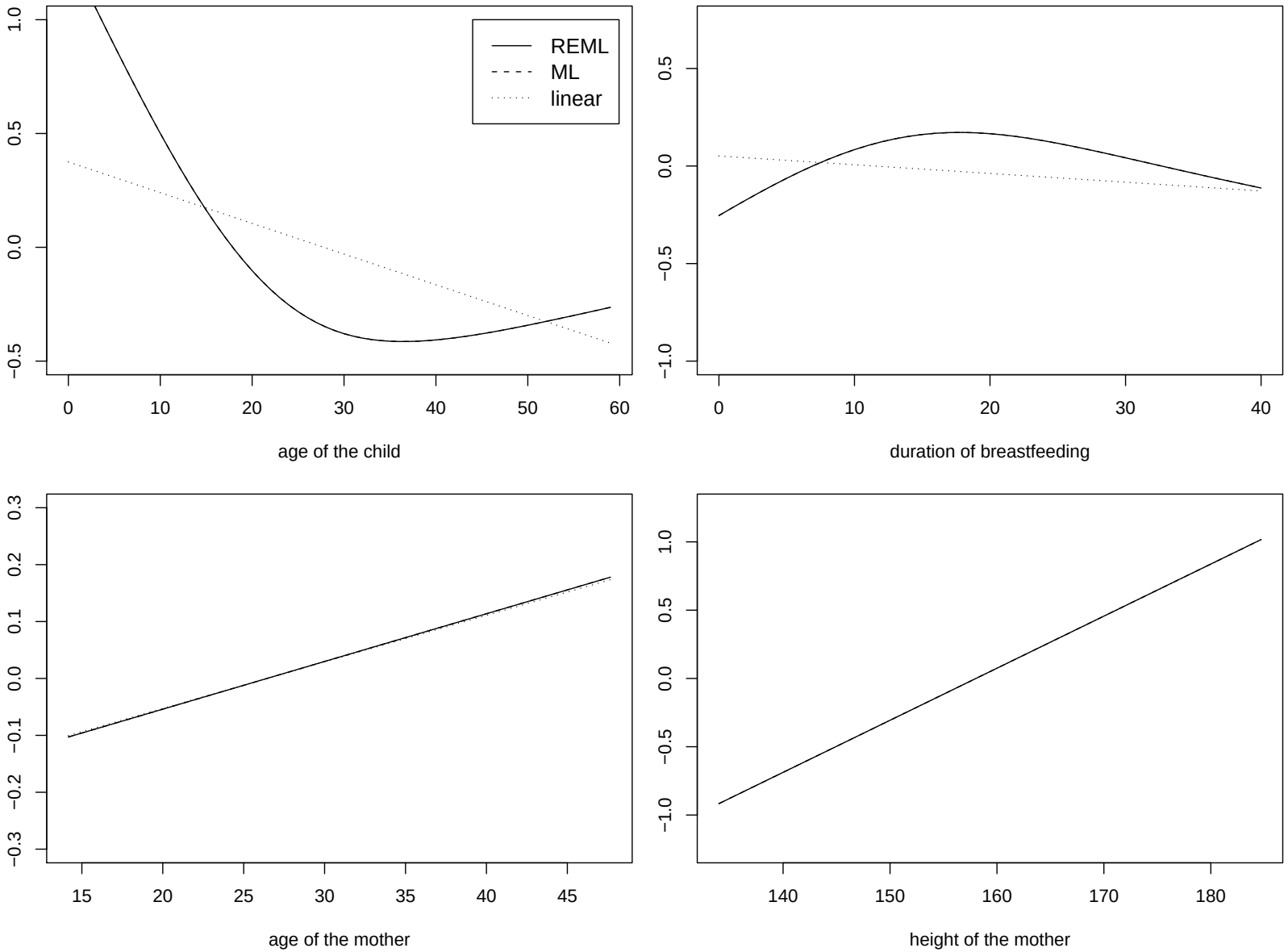
- Parametric effects in $\boldsymbol{x}'\boldsymbol{\beta}$ are not subject to model selection.

  $\Rightarrow 2^6 = 64$ models to consider in the model comparison.

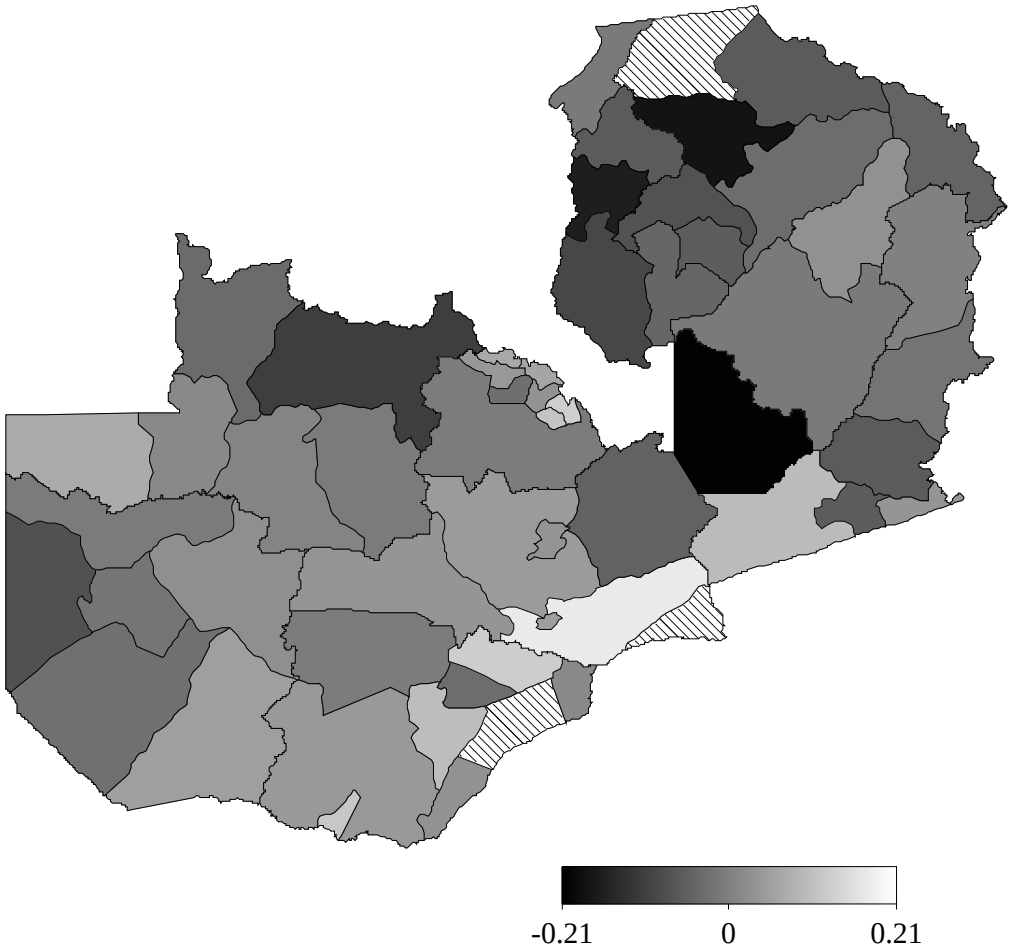
- The eight best fitting models:

						ML		$REML$	
cfeed	cage	mage	mheight	mbmi	s	$cAIC$	$mAIC$	$cAIC$	$mAIC$
+	+	−	−	−	+	**4125.78**	**4151.10**	**4125.78**	**4173.72**
+	+	+	−	−	+	**4125.78**	4153.10	**4125.78**	4175.72
+	+	−	+	−	+	**4125.78**	4153.10	**4125.78**	4175.72
+	+	−	−	+	+	**4125.78**	4153.10	**4125.78**	4175.72
+	+	+	+	−	+	**4125.78**	4155.10	**4125.78**	4177.72
+	+	+	−	+	+	**4125.78**	4155.10	**4125.78**	4177.72
+	+	−	+	+	+	**4125.78**	4155.10	**4125.78**	4177.72
+	+	+	+	+	+	**4125.78**	4157.10	**4125.78**	4179.72

- Linear effects are selected for age, height and body mass index of the mother.
- Some nonlinearity is detected for age of the child and duration of breastfeeding.


REML
ML
linear
1.0
0.5
0.0
−0.5
0
10
20
30
40
50
60
age of the child
0.5
0.0
−0.5
−1.0
0
10
20
30
40
duration of breastfeeding
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
15
20
25
30
35
40
45
age of the mother
1.0
0.5
0.0
−0.5
−1.0
140
150
160
170
180
height of the mother

- Inclusion of the region-specific random effect is required to capture spatial variation.



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. Technical Report.
  - 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`