Conditional Model Selection in Mixed-Effects Models with cAIC4

Benjamin Säfken [iD]
Georg-August Universität Göttingen

David Rügamer [iD]
Ludwig-Maximilians-Universität München

Thomas Kneib [iD]
Georg-August Universität Göttingen

Sonja Greven [iD]
Ludwig-Maximilians-Universität München

Abstract

Model selection in mixed models based on the conditional distribution is appropriate for many practical applications and has been a focus of recent statistical research. In this paper we introduce the R package **cAIC4** that allows for the computation of the conditional Akaike information criterion (cAIC). Computation of the conditional AIC needs to take into account the uncertainty of the random effects variance and is therefore not straightforward. We introduce a fast and stable implementation for the calculation of the cAIC for (generalized) linear mixed models estimated with **lme4** and (generalized) additive mixed models estimated with **gamm4**. Furthermore, **cAIC4** offers a stepwise function that allows for an automated stepwise selection scheme for mixed models based on the cAIC. Examples of many possible applications are presented to illustrate the practical impact and easy handling of the package.

Keywords: conditional AIC, **lme4**, mixed-effects models, penalized splines.

1. Introduction

The (generalized) linear mixed model is a flexible and broadly applicable statistical model. It is naturally used for analyzing longitudinal or clustered data. Furthermore, any regularized regression model incorporating a quadratic penalty can be written in terms of a mixed model. This incorporates smoothing spline models, spatial models and more general additive models (Wood 2017). Thus efficient and reliable estimation of such models is of major interest for applied statisticians. The package **lme4** (Bates, Mächler, Bolker, and Walker 2015; Bates, Maechler, Bolker, and Walker 2021) for the statistical computing software R (R Core Team

2021) offers such an exceptionally fast and generic implementation for mixed models (see Bates *et al.* 2015). The package has a modular framework allowing for the profile restricted maximum likelihood (REML) criterion as a function of the model parameters to be optimized using any constrained optimization function in R and uses rapid techniques for solving penalized least squares problems based on sparse matrix methods.

The fact that mixed models are widely used popular statistical tools makes model selection an indispensable necessity. Consequently research regarding model choice, variable selection and hypothesis testing in mixed models has flourished in recent years.

Hypothesis testing on random effects is well established, although for likelihood ratio tests boundary issues arise (Crainiceanu and Ruppert 2004; Greven, Crainiceanu, Küchenhoff, and Peters 2008; Wood 2013). In model selection for linear mixed models using the Akaike information criterion (AIC; Akaike 1973), Vaida and Blanchard (2005) suggest to use different criteria depending on the focus of the underlying research question. They make a distinction between questions with a focus on the population and on clusters, respectively. For the latter, they introduce a conditional AIC accounting for the shrinkage in the random effects. Based on this conditional AIC, Liang, Wu, and Zou (2008) propose a criterion that corrects for the estimation uncertainty of the random effects variance parameters based on a numerical approximation. Greven and Kneib (2010) show that ignoring this estimation uncertainty induces a bias and derive an analytical representation for the conditional AIC.

For certain generalized mixed models, analytical representations of the conditional AIC exist, for instance for Poisson responses (see Lian 2012). Although there is no general unbiased criterion in analytical form for all exponential family distributions as argued in Säfken, Kneib, van Waveren, and Greven (2014), bootstrap-based methods can often be applied as we will show for those presented in Efron (2004). An asymptotic criterion for a wider class of distributions is described in Wood, Pya, and Säfken (2016). The R package **lmerTest** (see Kuznetsova, Brockhoff, and Christensen 2017) offers some tests in linear mixed effects models.

In this paper, we describe an add-on package to **lme4** that facilitates model selection based on the conditional AIC and illustrate it with several examples. For the conditional AIC proposed by Greven and Kneib (2010) for linear mixed models, the computation of the criterion is not as simple as it is for other common AIC criteria. This article focuses on techniques for fast and stable computation of the conditional AIC in mixed models estimated with **lme4**, as they are implemented in the R package **cAIC4** (Säfken, Rügamer, Baumann, and Kruse 2021), available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=cAIC4`. The amount of possible models increases substantially with the R package **gamm4** (Wood and Scheipl 2020) allowing for the estimation of a wide class of models with quadratic penalty such as spline smoothing and additive models. The presented conditional AIC applies to any of these models.

In addition to translating the findings of Greven and Kneib (2010) to the model formulations used in Bates *et al.* (2015), we present the implementation of conditional AICs proposed for non-Gaussian settings in Säfken *et al.* (2014) and a new version for binary responses based on Efron (2004). With these results, a new scheme for stepwise conditional variable selection in mixed models is introduced. This allows for automatic choice of fixed and random effects based on the optimal conditional AIC. All methods are accompanied by examples, mainly taken from **lme4**, see Bates *et al.* (2015). The rest of this paper is structured as follows: In Section 2 the mixed model formulations are introduced based on one example with random

intercepts and random slopes and a second example on penalized spline smoothing. The conditional AIC for Gaussian, Poisson and Bernoulli responses is introduced in Section 3. Section 4 gives a hands-on introduction to **cAIC4** with specific examples for the `sleepstudy` and the `grouseticks` data from **lme4**. The new scheme for stepwise conditional variable selection in mixed models is presented in Section 5 and applied to the `Zambia` data set that is included in the package. After the conclusion in Section 6, Appendix A describes how **cAIC4** automatically deals with boundary issues. Furthermore the underlying code for the rapid computation of the conditional AIC is presented in Appendix B.

2. The mixed model

In a linear mixed model, the conditional distribution of the response $\boldsymbol{y}$ given the random effects $\boldsymbol{u}$ has the form

$$\boldsymbol{y}|\boldsymbol{u} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma^2 \boldsymbol{I}_n\right), \tag{1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ is the n-dimensional vector of responses, $\boldsymbol{\beta}$ is the p-dimensional vector of fixed effects and $\boldsymbol{u}$ is the q-dimensional vector of random effects. The matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the $(n \times p)$ and $(n \times q)$ design matrices for fixed and random effects, respectively, and σ^2 refers to the variance of the error terms.

The unconditional distribution of the random effects $\boldsymbol{u}$ is assumed to be a multivariate Gaussian with mean $\boldsymbol{0}$ and positive semidefinite $(q \times q)$ covariance matrix $\boldsymbol{D}_{\boldsymbol{\theta}}$, i.e.,

$$\boldsymbol{u} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{D}_{\boldsymbol{\theta}}\right).$$

The symmetric covariance matrix $\boldsymbol{D}_{\boldsymbol{\theta}}$ depends on the covariance parameters $\boldsymbol{\theta}$ and may be decomposed as

$$\boldsymbol{D}_{\boldsymbol{\theta}} = \sigma^2 \boldsymbol{\Lambda}_{\boldsymbol{\theta}} \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^\top, \tag{2}$$

with the lower triangular covariance factor $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$ and the variance parameter σ^2 of the conditional response distribution. In analogy to generalized linear models, the generalized linear mixed model extends the distributional assumption in (1) to a distribution $\mathcal{F}$ from the exponential family,

$$\boldsymbol{y}|\boldsymbol{u} \sim \mathcal{F}(\boldsymbol{\mu}, \phi),$$

where ϕ is a scale parameter and the mean has the form

$$\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{y}|\boldsymbol{u}) = h\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}\right), \tag{3}$$

with h being the response function applied componentwise and natural parameter $\boldsymbol{\eta} = h^{-1}(\boldsymbol{\mu})$. As the hereinafter presented results are limited to the Poisson and binomial distributions we can assume $\phi = 1$. The symmetric covariance matrix in (2) then is the same as for Gaussian responses except that σ^2 is omitted, i.e., $\boldsymbol{D}_{\boldsymbol{\theta}} = \boldsymbol{\Lambda}_{\boldsymbol{\theta}} \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^\top$.

The given conditional formulations of (generalized) linear mixed models imply marginal models, which can (conceptually) be obtained by integrating the random effects out of the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{u}$, i.e.,

$$f(\boldsymbol{y}) = \int f(\boldsymbol{y} \mid \boldsymbol{u}) f(\boldsymbol{u}) d\boldsymbol{u}.$$

However, there is typically no closed form solution for this integral. While the marginal model formulation is usually used for estimation, an analytic representation of $f(\boldsymbol{y})$ is only available

for the linear mixed model (1). The marginal distribution $f(\boldsymbol{y})$ for Gaussian responses $\boldsymbol{y}$ is given by

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{I}_n + \boldsymbol{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{\top}\boldsymbol{Z}^{\top}\right)\right).$$

Further extensions of linear mixed models can be obtained by, for example, relaxing the assumption $\mathrm{Cov}(\boldsymbol{y}|\boldsymbol{u}) = \sigma^2\boldsymbol{I}_n$.

2.1. Example I: Random intercepts and random slopes

Some special cases of mixed models are commonly used in applications, including the random intercept model and the random slope model. In the random intercept model, the responses differ in an individual- or cluster-specific intercept for m individuals or clusters. In this case the individual-specific intercept is modeled as random effect $\boldsymbol{u} = (u_{1,1}, u_{1,2}, \ldots, u_{1,m})$, yielding the (generalized) linear mixed model

$$\mathbb{E}(y_{ij}|u_{1,i}) = h(\boldsymbol{x}_{ij}\boldsymbol{\beta} + u_{1,i}), \quad u_{1,i} \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau_1^2)$$

for the jth observation from an individual or cluster i.

Whereas for the random intercept model all covariates modeled with fixed effects are assumed to have the same influence on the response variable across individuals, the random slope model is suitable when an independent variable x_s is assumed to have an individual-specific effect on the dependent variable. The random intercept model is extended to

$$\mathbb{E}(y_{ij}|\boldsymbol{u}_i) = h(\boldsymbol{x}_{ij}\boldsymbol{\beta} + u_{1,i} + x_{s,ij}u_{2,i}),$$

where $u_{2,i}$ is the individual-specific slope, which can be regarded as the deviation from the population slope β_s corresponding to the sth covariate $x_{s,ij}$ in $\boldsymbol{x}_{ij}$. In most cases, there is no reason to suppose $u_{1,i}$ and $u_{2,i}$ to be uncorrelated and the distributional assumption thus is

$$\begin{pmatrix} u_{1,i} \\ u_{2,i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{21} & \tau_2^2 \end{pmatrix}\right). \tag{4}$$

2.2. Example II: Penalized spline smoothing

In addition to many possibilities to extend these simple random effect models, (generalized) linear mixed models can also be utilized to fit semi-parametric regression models (see, e.g., Ruppert, Wand, and Carroll 2003). For univariate smoothing, consider the model

$$\mathbb{E}(y_i) = f(x_i), \tag{5}$$

for $i = 1, \ldots, n$, where $f(\cdot)$ is a deterministic function of the covariate x_i, which shall be approximated using splines. For illustrative purposes, we consider the truncated polynomial basis representation

$$f(x) = \sum_{j=0}^{g} \beta_j x^j + \sum_{j=1}^{k} u_j (x - \kappa_j)_+^g, \tag{6}$$

in the following, where $\kappa_1 < \ldots < \kappa_k$ are $k \in \mathbb{N}$ knots, partitioning the domain of x, $g \in \mathbb{N}$ and

$$(z)_+^g = z^g \cdot I(z > 0) = \begin{cases} z^g & \text{if } z > 0 \\ 0 & \text{if } z \le 0 \end{cases}. \tag{7}$$

As the truncated part $u_j(x - \kappa_j)_+^g$ is non-zero for $x > \kappa_j$, u_j can be seen as a gradient change of the two consecutive function segments defined on $(\kappa_{j-1}, \kappa_j]$ and $(\kappa_j, \kappa_{j+1}]$. In order to estimate $\beta_j, j = 0, \ldots, g$ and $u_j, j = 1, \ldots, k$, the method of ordinary least squares (OLS) could in principle be applied. In most cases, however, this yields a rather rough estimate of f for suitably large k as the gradient changes of function segments have a large impact. Therefore estimation methods for linear mixed models can be utilized in order to obtain a smooth function. Representing the untruncated polynomial part in (6) as the fixed effects and $\sum_{j=1}^k u_j(x - \kappa_j)_+^g$ as the random effects part, the well known shrinkage effect of mixed models is transferred to the estimation of the u_js, shrinking the changes in the gradient of the fitted polynomials. The random effects assumption corresponds to a quadratic penalty on the u_js, with the smoothing parameter estimated from the data.

This approach also works analogously for various other basis functions including the frequently used B-spline basis and generalized linear models (see, e.g., Fahrmeir, Kneib, Lang, and Marx 2021). Moreover, a rich variety of models that can be represented as reduced rank basis smoothers with quadratic penalties allow for this kind of representation. The estimation via **lme4** can be employed by the use of **gamm4**. For an overview of possible model components see Wood (2017). An example is also given in Section 5.

3. The conditional AIC

3.1. The Akaike information criterion

Originally proposed by Hirotogu Akaike (Akaike 1973) the AIC was one of the first model selection approaches to attract special attention among users of statistics. In some way, the AIC extends the maximum likelihood paradigm by making available a framework, in which both parameter estimation and model selection can be accomplished. The principle idea of the AIC can be traced back to the Kullback-Leibler distance (KLD; Kullback and Leibler 1951), which can be used to measure the distance between a true (but normally unknown) density $g(\boldsymbol{y})$ and a parametric model $f(\boldsymbol{y} \mid \boldsymbol{\nu})$. The unknown parameters $\boldsymbol{\nu}$ are commonly estimated by their maximum likelihood estimator $\hat{\boldsymbol{\nu}}(\boldsymbol{y})$. As minimizing the expected Kullback-Leibler distance is equivalent to minimizing the so called Akaike information

$$\text{AI} = -2 \, \mathbb{E}_{g(\boldsymbol{y})} \mathbb{E}_{g(\tilde{\boldsymbol{y}})} \log f(\tilde{\boldsymbol{y}} \mid \hat{\boldsymbol{\nu}}(\boldsymbol{y})), \tag{8}$$

with $\tilde{\boldsymbol{y}}$ a set of independent new observations from g, minus twice the maximized log-likelihood $\log f(\boldsymbol{y} \mid \hat{\boldsymbol{\nu}}(\boldsymbol{y}))$ as a natural measure of goodness-of-fit is an obvious estimator of the AI. However, this approach induces a bias as the maximized log-likelihood only depends on $\boldsymbol{y}$ whereas (8) is defined as a predictive measure of two independent replications $\tilde{\boldsymbol{y}}$ and $\boldsymbol{y}$ from the same underlying distribution. Therefore the bias correction is defined by

$$\text{BC} = 2 \left(\mathbb{E}_{g(\boldsymbol{y})} \log f(\boldsymbol{y} \mid \hat{\boldsymbol{\nu}}(\boldsymbol{y})) - \mathbb{E}_{g(\boldsymbol{y})} \mathbb{E}_{g(\tilde{\boldsymbol{y}})} \log f(\tilde{\boldsymbol{y}} \mid \hat{\boldsymbol{\nu}}(\boldsymbol{y})) \right). \tag{9}$$

Akaike derived the bias correction, which under certain regularity conditions can be estimated asymptotically by two times the dimension of $\boldsymbol{\nu}$. This yields the well-known AI estimator

$$\text{AIC}(\boldsymbol{y}) = -2 \log f(\boldsymbol{y} \mid \hat{\boldsymbol{\nu}}(\boldsymbol{y})) + 2 \dim(\boldsymbol{\nu}).$$

Hence, as the statistical model $f(\cdot|\boldsymbol{\nu})$ with the smallest AI aims at finding the model which is closest to the true model, the AIC can be seen as a relative measure of goodness-of-fit for different models of one model class. Notice that the bias correction is equivalent to twice the (effective) degrees of freedom and the covariance penalty, see Efron (2004).

3.2. The marginal and the conditional perspective on the AIC

Adopting this principle for the class of mixed models to select amongst different random effects is not straightforward. First of all, the question arises on the basis of which likelihood to define this AIC. For the class of mixed models, two common criteria exist, namely the marginal AIC (mAIC) based on the marginal log-likelihood and the conditional AIC (cAIC) based on the conditional log-likelihood. The justification of both approaches therefore corresponds to the purpose of the marginal and the conditional mixed model perspective, respectively. Depending on the question of interest, the intention of both perspectives differs, as for example described in Vaida and Blanchard (2005) or Greven and Kneib (2010).

The marginal perspective of mixed models is suitable when the main interest is to model fixed population effects with a reasonable correlation structure. The conditional perspective, by contrast, can be used to make statements based on the fit of the predicted random effects. In longitudinal studies, for example, the latter point of view seems to be more appropriate if the focus is on subject- or cluster-specific random effects. Another crucial difference in both approaches lies in the model's use for prediction. On the one hand, the marginal model seems to be more plausible if the outcome for new observations comes from new individuals or clusters, i.e., observations having new random effects. The conditional model on the other hand is recommended if predictions are based on the same individuals or clusters, thereby predicting on the basis of already modeled random effects.

The corresponding AI criteria have closely related intentions. The conditional AIC estimates the optimism of the estimated log-likelihood for a new data set $\tilde{\boldsymbol{y}}$ by leaving the random effects unchanged. This can be understood as a predictive measure based on a new data set originating from the same clusters or individuals as $\boldsymbol{y}$. On the contrary, the marginal approach evaluates the log-likelihood using a new predictive data set $\tilde{\boldsymbol{y}}$, which is not necessarily associated with the cluster(s) or individual(s) of $\boldsymbol{y}$.

In particular for the use of mixed models in penalized spline smoothing, the cAIC usually represents a more plausible choice. As demonstrated in Example II of Section 2, the representation of penalized spline smoothing via mixed models divides certain parts of the spline basis into fixed and random effects. Using the marginal perspective in Example II, predictions would therefore be based only on the polynomial coefficients of f. If the fitted nonlinear function is believed to represent a general relationship of x and y, predictions as well as the predictive measure in terms of the Akaike information, however, make more sense if the truncated parts of the basis are also taken into account.

Vaida and Blanchard (2005) propose the cAIC, an estimator of the conditional Akaike information

$$\text{cAI} = -2\,\mathbb{E}_{g(\boldsymbol{y},\boldsymbol{u})}\mathbb{E}_{g(\tilde{\boldsymbol{y}}|\boldsymbol{u})}\log f(\tilde{\boldsymbol{y}}\mid\hat{\boldsymbol{\nu}}(\boldsymbol{y}),\hat{\boldsymbol{u}}(\boldsymbol{y})) \tag{10}$$

as an alternative to the mAIC, where $\boldsymbol{\nu}$ includes the fixed effects and covariance parameters $\boldsymbol{\theta}$. The cAIC may be more appropriate when the AIC is used for the selection of random effects. In addition, Greven and Kneib (2010) investigate the difference of both criteria from a mathematical point of view. Since the mAIC is intended for the use in settings

where the observations are independent and the parameter space $\boldsymbol{\mathcal{V}}$ can be transformed to $\mathbb{R}^{\dim(\boldsymbol{\nu})}$, the corresponding bias correction $2\dim(\boldsymbol{\nu})$ is biased for mixed models for which these conditions do not apply. In particular, Greven and Kneib (2010) show that the mAIC leads to a preference for the selection of smaller models without random effects.

3.3. Conditional AIC for Gaussian responses

Depending on the distribution of $\boldsymbol{y}$, different bias corrections of the maximized conditional log-likelihood exist to obtain the cAIC. For the Gaussian case, Liang *et al.* (2008) derive a corrected version of the initially proposed cAIC by Vaida and Blanchard (2005) for known error variance, taking into account the estimation of the covariance parameters $\boldsymbol{\theta}$:

$$\text{cAIC}(\boldsymbol{y}) = -2\log f(\boldsymbol{y} \mid \hat{\boldsymbol{\nu}}(\boldsymbol{y}), \hat{\boldsymbol{u}}(\boldsymbol{y})) + 2\operatorname{tr}\left(\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{y}}\right). \tag{11}$$

Evaluating the bias correction $BC = 2\operatorname{tr}(\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{y}})$ in expression (11) via numerical approximation, or a similar formula for unknown error variance, is however computationally expensive. Greven and Kneib (2010) develop an analytic version of the corrected cAIC making the calculation of the corrected cAIC feasible. We adapt their efficient implementation originally written for 'lme'-objects (returned by the **nlme** package; Pinheiro, Bates, DebRoy, Sarkar, and R Core Team 2021) and reimplement their algorithm for 'lmerMod'-objects (returned by **lme4**). A more detailed description on the calculation of several terms in the proposed formula of Greven and Kneib (2010) is given in Appendix B. Furthermore, a partition of the parameter space is needed in order to account for potential parameters on the boundary of the parameter space, as presented in Theorem 3 in Greven and Kneib (2010). This process can be very unwieldy. Therefore, a fully automated correction algorithm is implemented in **cAIC4** and presented in Appendix A.

3.4. Conditional AIC for Poisson responses

As for the Gaussian case, note that for the Poisson and the binomial distribution the bias correction (9) can be rewritten as twice the sum of the covariances between $\hat{\eta}_i$ and y_i,

$$BC = 2\sum_{i=1}^{n} \mathbb{E}\left(\hat{\eta}_i \left(y_i - \mu_i\right)\right), \tag{12}$$

with true but unobserved mean $\mu_i = \mathbb{E}(y_i)$ and the estimator of the natural parameter $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \ldots, \hat{\eta}_n)^\top$ depending on $\boldsymbol{y}$. For the Poisson distribution an analytic reformulation of the bias correction term (12) has to be utilized to make it analytically accessible as in Säfken *et al.* (2014). Using results from Hudson (1978) and an identity due to Chen (1975), the bias correction (12) for Poisson distributed responses can be reformulated to

$$BC = 2\sum_{i=1}^{n} \mathbb{E}\left(y_i \left(\log \hat{\mu}_i(\boldsymbol{y}) - \log \hat{\mu}_i(\boldsymbol{y}_{-i}, y_i - 1)\right)\right), \tag{13}$$

for observations $i = 1, \ldots, n$ and mean estimator $\hat{\mu}_i$. The ith component of $\boldsymbol{y}$ in $(\boldsymbol{y}_{-i}, y_i - 1)$ is substituted by $y_i - 1$ along with the convention $y_i \log \hat{\mu}_i(\boldsymbol{y}_{-i}, y_i - 1) = 0$ if $y_i = 0$. The computational implementation of the cAIC in this case requires $n - d$ model fits, where

d corresponds to the number of Poisson responses being equal to zero (see Section 4.2 for details). The resulting cAIC was first derived by Lian (2012).

3.5. Conditional AIC for Bernoulli responses

For binary responses there is no analytical representation for the bias correction (12), see Säfken *et al.* (2014). Nevertheless a bootstrap estimate for the bias correction can be based on Efron (2004). The bias correction is equal to the sum over the covariances of the estimators of the natural parameter $\widehat{\eta}_i$ and the data y_i. To estimate this quantity, we could in principle draw a parametric bootstrap sample z_i of size B for the ith data point – keeping all other observations fixed at their observed values – to estimate the ith component $\mathbb{E}\left(\widehat{\eta}_i\left(y_i - \mu_i\right)\right)$ of the bias correction (12) for binary responses by

$$\frac{1}{B-1}\sum_{j=1}^{B}\hat{\eta}_i(z_{ij})\left(z_{ij} - \overline{z}_{i\cdot}\right) = \frac{B_1}{B-1}\hat{\eta}_i(1)\left(1 - \overline{z}_{i\cdot}\right) + \frac{B_0}{B-1}\hat{\eta}_i(0)\left(-\overline{z}_{i\cdot}\right),$$

where B_0 is the number of zeros in the bootstrap sample, B_1 is the number of ones in the bootstrap sample, $\hat{\eta}_i(1) = \log\left(\frac{\hat{\mu}_i(1)}{1-\hat{\mu}_i(1)}\right)$ is the estimated logit (the natural parameter) with $z_{ij} = 1$, $\hat{\eta}_i(0) = \log\left(\frac{\hat{\mu}_i(0)}{1-\hat{\mu}_i(0)}\right)$ is the estimated logit with $z_{ij} = 0$ and $\overline{z}_{i\cdot}$ is the mean of the bootstrap sample z_i. Letting the number of bootstrap samples tend to infinity, i.e., $B \to \infty$, the mean of the bootstrap sample $\overline{z}_{i\cdot} = \frac{1}{B}\sum_{j=1}^{B}z_{ij} = B_1/B$ (as well as $B_1/(B-1)$) converges to the estimate from the data, which corresponds to the true mean in the bootstrap $\hat{\mu}_i$, and therefore

$$\frac{B_1}{B-1}\hat{\eta}_i(1)\left(1 - \overline{z}_{i\cdot}\right) - \frac{B_0}{B-1}\hat{\eta}_i(0)\left(\overline{z}_{i\cdot}\right) \to \hat{\mu}_i\hat{\eta}_i(1)\left(1 - \hat{\mu}_i\right) - \left(1 - \hat{\mu}_i\right)\hat{\eta}_i(0)\left(\hat{\mu}_i\right)$$

$$= \hat{\mu}_i\left(1 - \hat{\mu}_i\right)\left(\hat{\eta}_i(1) - \hat{\eta}_i(0)\right) \text{ for } B \to \infty.$$

Since the bootstrap estimates are optimal if the number of bootstrap samples B tends to infinity, this estimator can be seen as the optimal bootstrap estimator. The resulting estimator of the bias correction

$$\widehat{BC} = 2\sum_{i=1}^{n}\hat{\mu}_i\left(1 - \hat{\mu}_i\right)\left(\hat{\eta}_i(1) - \hat{\eta}_i(0)\right), \tag{14}$$

which we use in the following, avoids a full bootstrap but requires n model refits.

Notice that for generalized mixed models especially for binary data, and Poisson data with small counts, the Laplace approximation that is used in **lme4** for models with more than one grouping variable is not always optimal, see Ogden (2017). However the estimator $\hat{\mu}_i\left(\cdot\right)$ used for Poisson and Bernoulli responses does not need to fulfill any specific conditions in order for the cAIC to be applicable. Thus the conclusions based on the cAIC are trustworthy even if the parameter estimates are biased in the sense that the model with the proposed minimal conditional Akaike information is chosen.

4. Introduction to cAIC4

4.1. Example for linear mixed models

An example that is often used in connection with the R package **lme4** is the `sleepstudy` data.

Thus we use this data set to give a brief introduction on the technical usage of the R package **cAIC4**. The data set is from a study on the daytime performance changes of the reaction time during chronic sleep restriction, see Belenky, Wesensten, Thorne, Thomas, Sing, Redmond, Russo, and Balkin (2003). Eighteen volunteers were only allowed to spend three hours of their daily time in bed for one week. The speed (mean and fastest 10% of responses) and lapses (reaction times greater than 500 ms) on a psychomotor vigilance task where measured several times. The averages of the reaction times are saved as response variable `Reaction` in the data set. Each volunteer has an identifier `Subject`. Additionally the number of days of sleep restriction at each measurement is listed in the covariate `Days`.

The conditional AIC can be used to find the model that best predicts future observations, assuming that future observations share the same random effects as the ones used for the model fitting. In case of this data set, using the cAIC for model choice corresponds to finding the model that best predicts future reaction times of the volunteers that took part in the study.

A first model that could be applied is a model with a random intercept and a random slope for `Days` within each volunteer (`Subject`):

$$y_{ij} = \beta_0 + \beta_1 \cdot \text{day}_{ij} + u_{i1} + u_{i2} \cdot \text{day}_{ij} + \epsilon_{ij} \tag{15}$$

for $i = 1, \ldots, 18$ and $j = 1, \ldots 10$, with

$$\begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{pmatrix} \right)$$

and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. In the preceding notation $\tau_1^2/\sigma^2 = \theta_1$, $\tau_2^2/\sigma^2 = \theta_2$ and $\tau_{12}/\sigma^2 = \theta_3$. That τ_{12} is not necessarily zero indicates, that the random intercept and the random slope are allowed to be correlated.

```
R> (m1 <- lmer(Reaction ~ 1 + Days + (1 + Days|Subject), sleepstudy))


Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ 1 + Days + (1 + Days | Subject)
   Data: sleepstudy
REML criterion at convergence: 1743.628
Random effects:
 Groups    Name        Std.Dev. Corr
 Subject   (Intercept) 24.740
           Days         5.922   0.07
 Residual              25.592
Number of obs: 180, groups:  Subject, 18
Fixed Effects:
(Intercept)          Days
     251.41         10.47
```

Calling the `cAIC` function from the **cAIC4** package returns a 'cAIC' object:

```
R> cAIC(m1)
```

```
        Conditional log-likelihood:  -824.51
                 Degrees of freedom:    31.30
Conditional Akaike information criterion:  1711.62
```

The conditional log-likelihood and the corrected degrees of freedom, i.e., one half the bias correction, are printed together with the conditional AIC as proposed in Greven and Kneib (2010). The 'cAIC' object is a list additionally containing an element called `reducedModel` that is the model without the random effects covariance parameters that were estimated to lie on the boundary of the parameter space, see Appendix A and Greven and Kneib (2010), and `NULL` if there were none on the boundary.

There are several further possible models for these data. One possibility might be to reduce the complexity in the model by omitting the correlation between the random intercept u_{i1} and the random slope u_{i2} for each `Subject`. Since this correlation is needed in order to preserve translational invariance (see Bates *et al.* 2015) such a comparison can be ill-advised. Otherwise the complexity of the model could be reduced by excluding the random slope from the model. This results in a simple random intercept model:

$$y_{ij} = \beta_0 + \beta_1 \cdot \text{day}_{ij} + u_i + \epsilon_{ij} \tag{16}$$

for $i = 1, \ldots, 18$ and $j = 1, \ldots 10$, with

$$u_i \sim \mathcal{N}\left(0, \tau^2\right).$$

In this model the pace of increasing reaction time does not systematically vary between the volunteers. This model is estimated by

```
R> (m2 <- lmer(Reaction ~ 1 + Days + (1 | Subject), sleepstudy))

Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ 1 + Days + (1 | Subject)
   Data: sleepstudy
REML criterion at convergence: 1786.465
Random effects:
 Groups   Name        Std.Dev.
 Subject  (Intercept) 37.12
 Residual             30.99
Number of obs: 180, groups:  Subject, 18
Fixed Effects:
(Intercept)         Days
     251.41        10.47
```

The **cAIC4** package offers an `anocAIC()` function along the lines of the `anova()` generic which allows to compare several 'lmer' or 'glmer' objects via the cAIC:

```
R> anocAIC(m1, m2)

                                          cll     df    cAIC Refit
Reaction ~ 1 + Days + (1 + Days | Subject) -824.51 31.30 1711.62 FALSE
Reaction ~ 1 + Days + (1 | Subject)        -864.53 19.03 1767.12 FALSE
```

Variable	Description
INDEX	Identifier of the chick
TICKS	Number of ticks sampled
BROOD	Brood number
HEIGHT	Height above sea level in meters
YEAR	Year as 95, 96 or 97
LOCATION	Geographic location code

Table 1: The variables and response of the `grouseticks` data set.

Thus the decrease in the conditional likelihood is much larger than the reduction of the degrees of freedom by omitting the random slopes. The strong evidence of subject-specific (random) slopes found in Bates *et al.* (2015) is also reflected by the cAIC.

The conditional AIC is also appropriate for choosing between a simple null model without any random effects and a complex model incorporating random effects, as has been noticed by Greven and Kneib (2010). Thus it is possible to compare the cAIC of the two previous mixed models with the standard AIC for a linear model, here including three parameters (intercept, linear effect for `Days` and error variance). This can be called via a `cAIC()` wrapper function

```
R> cAIC(lm(Reaction ~ 1 + Days, sleepstudy))
```

```
            Conditional log-likelihood:  -950.15
                   Degrees of freedom:     3.00
 Conditional Akaike information criterion:  1906.29
```

In this case, however, the mixed model structure is evident, reflected by the large AIC for the linear model.

4.2. Example for generalized linear mixed models

The **cAIC4** package additionally offers a conditional AIC for conditionally Poisson distributed responses and an approximate conditional AIC for binary data. The Poisson cAIC uses the bias correction (13) and the bias correction term for the binary data is (14).

Making use of the fast `refit()` function of the **lme4** package, both cAICs can be computed moderately fast, since $n - d$ and n model refits are required, respectively, with n being the number of observations and d the number of responses that are zero for the Poisson responses. In the following, the cAIC for Poisson response is computed for the `grouseticks` data set from the **lme4** package as an illustration.

The `grouseticks` data set was originally published in Elston, Moss, Boulinier, Arrowsmith, and Lambin (2001). It contains information about the aggregation of parasites, so-called sheep ticks, on red grouse chicks. The variables in the data set are given in Table 1. Every chick, identified by `INDEX`, is of a certain `BROOD` and every `BROOD`, in turn, corresponds to a specific `YEAR`.

The number of ticks is the response variable. Following the authors in a first model the expected number of ticks λ_j with `INDEX` (j) is modeled depending on the year and the height as fixed effects and for each of the grouping variables `BROOD` (i), `INDEX` (j) and `LOCATION` (k) a random intercept is incorporated.

The full model is

$$\log\left(\mathbb{E}\left(\texttt{TICKS}_j\right)\right) = \log\left(\lambda_j\right) = \beta_0 + \beta_1 \cdot \texttt{YEAR}_j + \beta_2 \cdot \texttt{HEIGHT}_j + u_{1,i} + u_{2,j} + u_{3,k} \qquad (17)$$

with three independent univariate random variables

$$u_{1,i} \sim \mathcal{N}\left(0, \tau_1^2\right), \ u_{2,j} \sim \mathcal{N}\left(0, \tau_2^2\right), \ u_{3,k} \sim \mathcal{N}\left(0, \tau_3^2\right).$$

Notice that in this model the `INDEX` is an observation-level random effect that addresses over-dispersion in the conditional Poisson distribution given the random effects. Before fitting the model the covariates `HEIGHT` and `YEAR` are centered for numerical reasons and stored in the data set `grouseticks_cen`.

```
R> f1 <- TICKS ~ YEAR + HEIGHT + (1 | BROOD) + (1 | INDEX) + (1 | LOCATION)
R> p1 <- glmer(f1, family = "poisson", data = grouseticks_cen)
```

A summary of the estimated model is given below. Notice that the reported AIC in the automated summary of **lme4** is not appropriate for conditional model selection.

```
Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: TICKS ~ YEAR + HEIGHT + (1 | BROOD) + (1 | INDEX) + (1 | LOCATION)
   Data: grouseticks_cen

     AIC      BIC   logLik deviance df.resid
  1845.5   1869.5   -916.7   1833.5      397

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.6507 -0.5609 -0.1348  0.2895  1.8518

Random effects:
 Groups   Name        Variance  Std.Dev.
 INDEX    (Intercept) 2.979e-01 5.458e-01
 BROOD    (Intercept) 1.466e+00 1.211e+00
 LOCATION (Intercept) 5.411e-10 2.326e-05
Number of obs: 403, groups:  INDEX, 403; BROOD, 118; LOCATION, 63

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.472353   0.134712   3.506 0.000454 ***
YEAR        -0.480261   0.166128  -2.891 0.003841 **
HEIGHT      -0.025715   0.003772  -6.817 9.32e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 .

Correlation of Fixed Effects:
```

```
        (Intr) YEAR
YEAR    0.089
HEIGHT 0.096  0.061
```

The conditional log-likelihood and the degrees of freedom for the conditional AIC with conditionally Poisson distributed responses as in (13) for model (17) are obtained by the call of the `cAIC` function:

```
R> cAIC(p1)
```

```
The original model was refitted due to zero variance components.
Refitted model:  TICKS ~ YEAR + HEIGHT + (1 | INDEX) + (1 | BROOD)

                Conditional log-likelihood:  -572.01
                        Degrees of freedom:   205.59
 Conditional Akaike information criterion:  1555.22
```

The output is the same as for Gaussian linear mixed models. It becomes apparent that there is a substantial difference between the conditional and the marginal AIC: In the output of the model the marginal AIC is reported to be 1845.5. Note that the marginal AIC is biased, see Greven and Kneib (2010), and based on a different likelihood.

The `cAIC` function calls the `deleteZeroComponents` function that reduces the model by random effects components that are estimated to have variances on the boundary of the parameter space, see Appendix A for the technical details and Greven and Kneib (2010) for the methodological background. In this example this is the case for the random intercepts of the `LOCATION` grouping variable. Thus the resulting cAIC is the same (modulo rounding) as if the random intercept associated with `LOCATION` is excluded from the model:

```
R> f2 <- TICKS ~ YEAR + HEIGHT + (1 | BROOD) + (1 | INDEX)
R> p2 <- glmer(f2, family = "poisson", data = grouseticks_cen)
R> cAIC(p2)
```

```
    Conditional log-likelihood:  -572.01
            Degrees of freedom:   205.59
 Conditional Akaike information:  1555.21
```

Another possible model for the comparison omits random intercepts associated with the `BROOD` grouping. This is equivalent to setting the associated random intercepts variance to zero, i.e., $\tau_2^2 = 0$.

```
R> f3 <- TICKS ~ YEAR + HEIGHT + (1 | INDEX) + (1 | LOCATION)
R> p3 <- glmer(f3, family = "poisson", data = grouseticks_cen)
R> cAIC(p3)$caic
```

```
[1] 1594.424
```

Following the models in Elston *et al.* (2001) also the selection of the fixed effects parameters is of interest. Starting with model `p2` four different combinations of the fixed effects variables `YEAR` and `HEIGHT` are possible. The models can be conveniently compared by the call of the `anocAIC()` function:

```
R> f4 <- TICKS ~ YEAR + (1 | BROOD) + (1 | INDEX)
R> p4 <- glmer(f4, family = "poisson", data = grouseticks_cen)
R> f5 <- TICKS ~  HEIGHT + (1 | BROOD) + (1 | INDEX)
R> p5 <- glmer(f5, family = "poisson", data = grouseticks_cen)
R> f6 <- TICKS ~ (1 | BROOD) + (1 | INDEX)
R> p6 <- glmer(f6, family = "poisson", data = grouseticks_cen)
R> anocAIC(p2, p4, p5, p6)
```

	cll	df	cAIC	Refit
TICKS ~ YEAR + HEIGHT + (1 | BROOD) + (1 | INDEX)	-572.01	205.59	1555.22	FALSE
TICKS ~ YEAR + (1 | BROOD) + (1 | INDEX)	-570.13	208.06	1556.39	FALSE
TICKS ~ HEIGHT + (1 | BROOD) + (1 | INDEX)	-571.63	206.46	1556.17	FALSE
TICKS ~ (1 | BROOD) + (1 | INDEX)	-569.97	208.37	1556.68	FALSE

In this case the cAIC chooses the full model `p2`.

5. A scheme for stepwise conditional variable selection

Now having the possibility to compare different (generalized) linear mixed models via the conditional AIC, we introduce a model selection procedure in this section, searching the space of possible model candidates in a stepwise manner. Inspired by commonly used `step` functions as for example given by the `stepAIC` function in the **MASS** package (Venables and Ripley 2002), our `stepcAIC` function provides an automatic model selection applicable to some models of the class 'merMod' (produced by `[g]lmer`) or objects resulting from a **gamm4** call. A general strategy could be to automatically select the random effect structure in a forward fashion first and afterwards select the fixed effects structure for given random effects using the stepwise cAIC (or other common selection criteria for fixed effects). Alternative methods to also select both the fixed and random effect structure are available, e.g., by the function `step` in the package **lmerTest**. Our method has the advantage to be solely based on a predictive measure, the cAIC, and models are compared based on a simple comparison of this value. This circumvents problems associated with post-selection inference, where tests are conducted under the assumption of a null distribution that does not hold when the corresponding null hypothesis is the result of a preceding model selection. Although based on a predictive measure, we want to emphasize that our intention behind the `stepcAIC` function is not to provide a black-box algorithm for model selection, but rather to provide a convenience function for users who would otherwise manually select the best random effect structure in the same fashion. In addition, the function facilitates the comparison of different additive models when applied to a **gamm4** call, allowing the user to assess which nonlinearities are necessary for a given data set and which would instead overfit the data. The step procedure can therefore be used to successively extend or reduce the model in order to check whether a fixed covariate has a constant, linear or nonlinear effect.

Whereas the backward procedure has a straightforward mechanism and does not need any further mandatory arguments, the `stepcAIC` function provides several optional and obligatory arguments for the `forward` and `both` procedure in order to limit the possibly large number of model extensions. Regarding the required parameters, the user must specify the variables, which may be added with fixed or random effects as they are referred to in the '`data.frame`' given by the argument `data`. For the fixed effects, this is done by specifying the `fixEf` argument, which expects a character vector with the names of the covariates, e.g., `fixEf = c("x1", "x2")`. Variables listed in the `fixEf` argument are firstly included in the model as linear terms and, if the linear effect leads to an improvement of the cAIC, checked for their nonlinearity by evaluating the cAIC of the corresponding model(s). Model extensions resulting from additional random effects are created in two different ways. A new model may, on the one hand, include a random intercept for a variable forming a grouping structure or, on the other hand, a random slope for a variable. These two types are specified using the arguments `groupCandidates` for grouping variable candidates or `slopeCandidates` for candidates for variables with random slope, again by referring to the variable names in `data` using strings.

Further optional arguments determine the way random effects are treated in the step procedure:

- `allowUseAcross`: logical value whether slope variables, which are already in use with a grouping variable can also be used with other grouping variables;

- `maxSlopes`: maximum number of slopes for one grouping variable.

Following the `stepAIC` function, the `stepcAIC` function also provides an argument for printing interim results (`trace`) and allows for the remaining terms of the initial model to be unaffected by the procedure (`keep`: list with entries `fixed` and `random`, each either `NULL` or a `formula`). In addition, the user may choose whether the cAIC is calculated for models, for which the fitting procedure in (g)`lmer` could not find an optimum (`calcNonOptimMod` with default `FALSE`) and might choose the type of smoothing terms added in forward steps (`bsType`).

If the step function is used for large data sets or in the presence of highly complex models the fitting procedures as well as the calculations of the cAIC can be parallelized by defining the number of cores (`numCores`) being used if more than one model has to be fitted and evaluated in any step (by passing the `numCores` argument to a `mclapply` function implemented in the base package **parallel**).

Due to the variety of additive model definitions in **gamm4**, the `stepcAIC` function is however limited in its generic step functionality for GAMMs. On the one hand, extensions with nonlinear effects are restricted to one smooth class given by `bsType`, on the other hand, the step-procedure is not able to deal with further arguments passed in smooth terms. The latter point is a current limitation, since the default basis dimension of the smooth term (i.e., the number of knots and the order of the penalty) is essentially arbitrary.

An additional current limitation of the `stepcAIC` function in its applications with GAMMs is the handling of zero variance components occurring during the function call. As a meaningful handling of zero variance smoothing terms would depend on the exact specification of the nonlinear term, the stepwise procedure is stopped and returns the result of the previous step. After removing the term associated with the zero variance manually, e.g., by specifying a

linear instead of a nonlinear effect for the corresponding covariate, the user may call the step function again.

5.1. Examples for stepwise conditional variable selection

In order to demonstrate some functionalities of the `stepcAIC` function, we use a data set on childhood malnutrition in Zambia, which has also been analyzed by Kandala, Lang, Klasen, and Fahrmeir (2000) and also by Greven and Kneib (2010) to demonstrate the applicability of the cAIC for model selection in additive models. The subset of the original data set, which is used here for illustrative purposes, is available in our R package and can be loaded into the environment via `data("Zambia", package = "cAIC4)`. The dependent variable is the Z-score, a standardized measure for the height of a child, which is an indicator for chronic undernutrition or stunting in children. We estimate a linear mixed model using four covariates, the duration of breastfeeding (`c.bf`) in months, the age (`c.age`) of the child in months, the height (`m.ht`) as well as the body mass index (`m.bmi`) of the mother, included in the model as fixed effects, and account for spatial heterogeneity by including a random intercept for the region (`reg`) and the district in the region (`dr`) the child lives in.

Starting with the following random effects model, we additionally check whether the two continuous covariates `m.ht` and `m.bmi` also should be used as a random slope for either of the random intercepts.

```
R> data("Zambia", package = "cAIC4")
R> initmixmod <- lmer(z ~ c.age + c.bf + m.ht + m.bmi + (1 | reg) + (1 | dr),
+     data = Zambia)
```

We therefore employ the `stepcAIC` function using `direction = "both"` and supplying the two possible candidates in `slopeCandidates`.

```
R> mod <- stepcAIC(initmixmod, direction = "both",
+     slopeCandidates = c("m.ht", "m.bmi"), data = Zambia, trace = TRUE)

Starting stepwise procedure...
----------------------------------------------
----------------------------------------------

Step 1 (forward):   cAIC=1827.891
Best model so far:
~ c.age + c.bf + m.ht + m.bmi + (1 | dr) + (1 | reg)
New Candidates:

Calculating cAIC for 4 model(s) ...

                                                          models caic
  ~c.age + c.bf + m.ht + m.bmi + (1 + m.ht | dr) + (1 | reg) 1836
  ~c.age + c.bf + m.ht + m.bmi + (1 | dr) + (1 + m.ht | reg) 1832
 ~c.age + c.bf + m.ht + m.bmi + (1 | dr) + (1 + m.bmi | reg) 1832
 ~c.age + c.bf + m.ht + m.bmi + (1 + m.bmi | dr) + (1 | reg) 1832
```

```
------------------------------------------------
------------------------------------------------

Step 2 (backward):  cAIC=1827.891
Best model so far:
~ c.age + c.bf + m.ht + m.bmi + (1 | dr) + (1 | reg)
New Candidates:

Calculating cAIC for 2 model(s) ...

                                      models caic
 ~c.age + c.bf + m.ht + m.bmi + (1 | reg) 1831
  ~c.age + c.bf + m.ht + m.bmi + (1 | dr) 1827


------------------------------------------------
------------------------------------------------

Step 3 (forward):  cAIC=1827.239
Best model so far:
~ c.age + c.bf + m.ht + m.bmi + (1 | dr)
New Candidates:

Calculating cAIC for 2 model(s) ...

                                            models caic
  ~c.age + c.bf + m.ht + m.bmi + (1 + m.ht | dr) 1836
 ~c.age + c.bf + m.ht + m.bmi + (1 + m.bmi | dr) 1830


------------------------------------------------
------------------------------------------------

Step 4 (backward):  cAIC=1827.239
Best model so far:
~ c.age + c.bf + m.ht + m.bmi + (1 | dr)
New Candidates:

Calculating cAIC for 1 model(s) ...

                        models caic
 ~c.age + c.bf + m.ht + m.bmi 1837


------------------------------------------------
------------------------------------------------

Best model:
 ~ c.age + c.bf + m.ht + m.bmi + (1 | dr) ,
 cAIC: 1827.239
------------------------------------------------
```

The step function starts with automatically calculating the four possible models, which result from extending either one of the intercepts by one of the two slope variables. Per default the scope of possible models is also restricted to random effect structures, which assume a non-zero correlation between each random effect.

In a first step, the model cannot be improved by extending the model. The next step is a backward step in which the random intercept for region is dropped. The resulting model is not altered until the final step 4 and chosen as final model.

To illustrate the use of the `stepcAIC` function in the context of GAMM selection, we now investigate if all fixed effects should be kept in the model given by the previous `stepcAIC` call and/or whether any of the covariates may yield an improvement in cAIC if included as nonlinear effect. We therefore fit an initial **gamm4** model as follows.

```
R> bestmixmod <- gamm4(z ~ c.age + c.bf + m.ht + m.bmi, random = ~ (1 | dr),
+    data = Zambia)
```

In order to allow linear effects to be dropped and/or nonlinear effects to be added to the model, we again employ the `stepcAIC` function using `direction = "both"` and supply the names of the four covariates, which should be considered as nonlinear effects in `fixEfCandidates`.

```
R> admod <- stepcAIC(bestmixmod, direction = "both",
+    fixEfCandidates = c("c.bf", "c.age", "m.ht", "m.bmi"), data = Zambia,
+    trace = TRUE)

Starting stepwise procedure...
----------------------------------------------
----------------------------------------------

Step 1 (forward):  cAIC=1827.239
Best model so far:
~ c.age + c.bf + m.ht + m.bmi + (1 | dr)
New Candidates:

Calculating cAIC for 4 model(s) ...
                                                   models caic
 ~c.age + c.bf + m.bmi + s(m.ht, bs = "tp") + (1 | dr) 1829
 ~c.age + c.bf + m.ht + s(m.bmi, bs = "tp") + (1 | dr) 1828
 ~c.age + m.ht + m.bmi + s(c.bf, bs = "tp") + (1 | dr) 1793
 ~c.bf + m.ht + m.bmi + s(c.age, bs = "tp") + (1 | dr) 1774

----------------------------------------------
----------------------------------------------

Step 2 (backward):  cAIC=1774.126
Best model so far:
~ s(c.age,bs="tp") + c.bf + m.ht + m.bmi + (1 | dr)
New Candidates:

Calculating cAIC for 5 model(s) ...
```

```
                                                       models caic
      ~c.age + c.bf + m.ht + m.bmi + (1 | dr) 1827
~c.bf + m.bmi + s(c.age, bs = "tp") + (1 | dr) 1799
   ~s(c.age, bs = "tp") + c.bf + m.ht + m.bmi 1789
 ~c.bf + m.ht + s(c.age, bs = "tp") + (1 | dr) 1781
~m.ht + m.bmi + s(c.age, bs = "tp") + (1 | dr) 1780

-----------------------------------------------
-----------------------------------------------

Step 3 (forward):  cAIC=1774.126
Best model so far:
~ s(c.age,bs="tp") + c.bf + m.ht + m.bmi + (1 | dr)
New Candidates:

Calculating cAIC for 3 model(s) ...
                                                               models caic
 ~s(c.age, bs = "tp") + c.bf + m.ht + s(m.bmi, bs = "tp") + (1 | dr) 1776
 ~s(c.age, bs = "tp") + c.bf + m.bmi + s(m.ht, bs = "tp") + (1 | dr) 1775
 ~s(c.age, bs = "tp") + m.ht + m.bmi + s(c.bf, bs = "tp") + (1 | dr) 1775

-----------------------------------------------
-----------------------------------------------

Best model:
 ~ s(c.age,bs="tp") + c.bf + m.ht + m.bmi + (1 | dr) ,
 cAIC: 1774.126

-----------------------------------------------
```

In the first step, `c.age` is changed to have a nonlinear effect using a thin-plate regression spline. The type of spline in forward steps can be specified using the `bsType` argument. In a second (backward) step, the function checks whether any of the model components should be dropped from the model. As all possibilities would result in models with larger cAIC values, the current
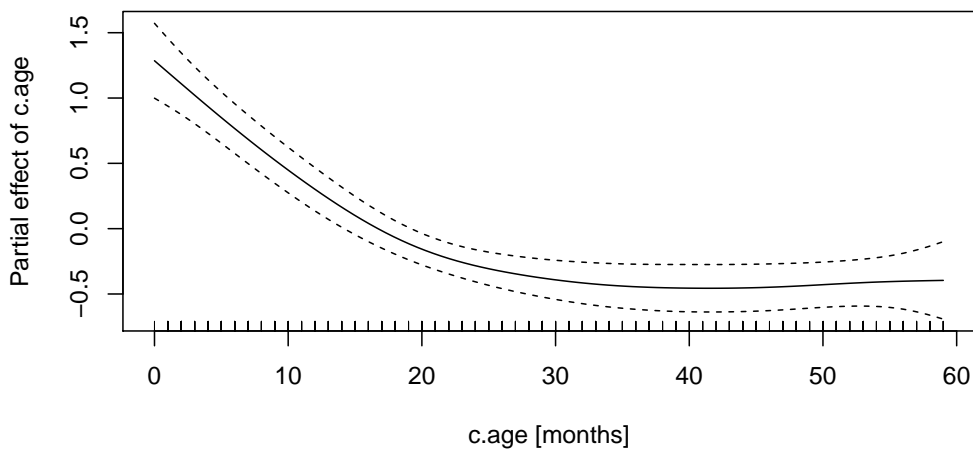

Figure 1: Plot of added smoothing spline effect.

model is checked for other nonlinearities in a third (forward) step. As all options result in a larger cAIC value the final model, within the specified scope of the step function, is given by the formula `~ s(age, bs = "tp") + c.bf + m.ht + m.bmi + (1 | dr)`. The estimated nonlinear effect of the final model can be visualized by `plot(admod$finalModel$gam)`, i.e., by calling the `plot` function on the `gam` part of the final model. The result is given in Figure 1, indicating that the age of very young children under 17 months is associated with a positive Z-score, hence a "positive" value in terms of health, which however decreases continuously for increasing age. After 2 to 2.5 years the effect of age on the Z-score is estimated to be almost constant and having a negative association with the Z-score. A similar effect of the children's age was also found by Greven and Kneib (2010).

6. Conclusion

This paper gives a hands-on introduction to the R package **cAIC4** allowing for model selection in mixed models based on the conditional AIC. The package and the paper offer a possibility for users from the empirical sciences to use the conditional AIC without having to worry about lengthy and complex calculations or mathematically sophisticated boundary issues of the parameter space. The applications presented in this paper go far beyond model selection for mixed models and extend to penalized spline smoothing and other structured additive regression models. Furthermore a stepwise algorithm for these models is introduced that allows for fast model selection.

Often statistical modeling is not about finding one "true model". In such cases it is of interest to define weighted sums of plausible models. This approach called model averaging is presented in Zhang, Zou, and Liang (2014) for weights chosen by the cAIC. We plan to implement this approach in **cAIC4**. Another future research path is to implement an appropriate version of the Bayesian information criterion (BIC) for conditional model selection. Since the Laplace approximation is not always optimal, see Ogden (2017), methods for preferable likelihood approximations are implemented in the **glmmsr** package (see Ogden 2019). Furthermore the newly developed **GLMMadaptive** package (see Rizopoulos 2021) fits mixed models with one grouping variable using the adaptive Gaussian quadrature. We plan to make **cAIC4** usable for these packages in the future.

Acknowledgments

References

Akaike H (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." In BN Petrov, F Csaki (eds.), *Proceedings of the 2nd International Symposium on Information Theory*, pp. 267–281. Akademiai Kiado.

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using **lme4**." *Journal of Statistical Software*, **67**(1), 1–48. `doi:10.18637/jss.v067.i01`.

Bates D, Maechler M (2021). **Matrix***: Sparse and Dense Matrix Classes and Methods.* R package version 1.3-4, URL `https://CRAN.R-project.org/package=Matrix`.

Bates D, Maechler M, Bolker B, Walker S (2021). **lme4***: Linear Mixed-Effects Models using* **Eigen** *and* S4. R package version 1.1-27.1, URL `https://CRAN.R-project.org/package=lme4`.

Belenky G, Wesensten NJ, Thorne DR, Thomas ML, Sing HC, Redmond DP, Russo MB, Balkin TJ (2003). "Patterns of Performance Degradation and Restoration during Sleep Restriction and Subsequent Recovery: A Sleep Dose-Response Study." *Journal of Sleep Research*, **12**(1), 1–12. `doi:10.1046/j.1365-2869.2003.00337.x`.

Chen LHY (1975). "Poisson Approximation for Dependent Trials." *The Annals of Probability*, **3**(3), 534–545. `doi:10.1214/aop/1176996359`.

Crainiceanu CM, Ruppert D (2004). "Likelihood Ratio Tests in Linear Mixed Models with One Variance Component." *Journal of the Royal Statistical Society B*, **66**(1), 165–185. `doi:10.1111/j.1467-9868.2004.00438.x`.

Efron B (2004). "The Estimation of Prediction Error." *Journal of the American Statistical Association*, **99**(467), 619–632. `doi:10.1198/016214504000000692`.

Elston DA, Moss R, Boulinier T, Arrowsmith C, Lambin X (2001). "Analysis of Aggregation, a Worked Example: Numbers of Ticks on Red Grouse Chicks." *Parasitology*, **122**(5), 563–569. `doi:10.1017/s0031182001007740`.

Fahrmeir L, Kneib T, Lang S, Marx BD (2021). *Regression: Models, Methods and Applications.* 2nd edition. Springer-Verlag. `doi:10.1007/978-3-662-63882-8`.

Greven S, Crainiceanu CM, Küchenhoff H, Peters A (2008). "Restricted Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models." *Journal of Computational and Graphical Statistics*, **17**(4), 870–891. `doi:10.1198/106186008x386599`.

Greven S, Kneib T (2010). "On the Behaviour of Marginal and Conditional AIC in Linear Mixed Models." *Biometrika*, **97**(4), 773–789. `doi:10.1093/biomet/asq042`.

Hudson HM (1978). "A Natural Identity for Exponential Families with Applications in Multiparameter Estimation." *The Annals of Statistics*, **6**(3), 473–484. `doi:10.1214/aos/1176344194`.

Kandala NB, Lang S, Klasen S, Fahrmeir L (2000). "Semiparametric Analysis of the Socio-Demographic and Spatial Determinants of Undernutrition in Two African Countries." *Research in Official Statistics*, **4**(1), 81–100.

Kullback S, Leibler RA (1951). "On Information and Sufficiency." *The Annals of Mathematical Statistics*, **22**(1), 79–86. `doi:10.1214/aoms/1177729694`.

Kuznetsova A, Brockhoff PB, Christensen RHB (2017). "**lmerTest** Package: Tests in Linear Mixed Effects Models." *Journal of Statistical Software*, **82**(13), 1–26. `doi:10.18637/jss.v082.i13`.

Lian H (2012). "A Note on Conditional Akaike Information for Poisson Regression with Random Effects." *Electronic Journal of Statistics*, **6**, 1–9. `doi:10.1214/12-ejs665`.

Liang H, Wu H, Zou G (2008). "A Note on Conditional AIC for Linear Mixed-Effects Models." *Biometrika*, **95**(3), 773–778. `doi:10.1093/biomet/asn023`.

Ogden H (2017). "On Asymptotic Validity of Naive Inference with an Approximate Likelihood." *Biometrika*, **104**(1), 153–164. `doi:10.1093/biomet/asx002`.

Ogden H (2019). **glmmsr***: Fit a Generalized Linear Mixed Model.* R package version 0.2.3, URL `https://CRAN.R-project.org/package=glmmsr`.

Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2021). **nlme***: Linear and Nonlinear Mixed Effects Models.* R package version 3.1-152, URL `https://CRAN.R-project.org/package=nlme`.

R Core Team (2021). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Rizopoulos D (2021). **GLMMadaptive***: Generalized Linear Mixed Models Using Adaptive Gaussian Quadrature.* R package version 0.8-2, URL `https://CRAN.R-project.org/package=GLMMadaptive`.

Ruppert D, Wand MP, Carroll RJ (2003). *Semiparametric Regression.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Säfken B, Kneib T, van Waveren CS, Greven S (2014). "A Unifying Approach to the Estimation of the Conditional Akaike Information in Generalized Linear Mixed Models." *Electronic Journal of Statistics*, **8**, 201–225. `doi:10.1214/14-ejs881`.

Säfken B, Rügamer D, Baumann P, Kruse RM (2021). **cAIC4** *Conditional Akaike Information Criterion for* **lme4** *and* **nlme**. R package version 1.0, URL `https://CRAN.R-project.org/package=cAIC4`.

Vaida F, Blanchard S (2005). "Conditional Akaike Information for Mixed-Effects Models." *Biometrika*, **92**(2), 351–370. `doi:10.1093/biomet/92.2.351`.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with* **S**. 4th edition. Springer-Verlag, New York.

Wang T, Merkle EC (2018). "**merDeriv**: Derivative Computations for Linear Mixed Effects Models with Application to Robust Standard Errors." *Journal of Statistical Software, Code Snippets*, **87**(1), 1–16. `doi:10.18637/jss.v087.c01`.

Wood S, Scheipl F (2020). **gamm4***: Generalized Additive Mixed Models Using* **mgcv** *and* **lme4**. R package version 0.2-6, URL `https://CRAN.R-project.org/package=gamm4`.

Wood SN (2013). "A Simple Test for Random Effects in Regression Models." *Biometrika*, **100**(4), 1005–1010. `doi:10.1093/biomet/ast038`.

Wood SN (2017). *Generalized Additive Models: An Introduction with R.* 2nd edition. Chapman & Hall/CRC. `doi:10.1201/9781315370279`.

Wood SN, Pya N, Säfken B (2016). "Smoothing Parameter and Model Selection for General Smooth Models." *Journal of the American Statistical Association*, **111**(516), 1548–1563. `doi:10.1080/01621459.2016.1180986`.

Zhang X, Zou G, Liang H (2014). "Model Averaging and Weight Choice in Linear Mixed-Effects Models." *Biometrika*, **101**(1), 205–218. `doi:10.1093/biomet/ast052`.

A. Dealing with the boundary issues

A major issue in obtaining the conditional AIC in linear mixed models is to account for potential parameters of $\boldsymbol{\theta}$ on the boundary of the parameter space (see Greven and Kneib 2010). This needs to be done in order to ensure positive definiteness of the covariance matrix $\boldsymbol{D_\theta}$.

The restructuring of the model in order to obtain the cAIC is done automatically by **cAIC4**. To gain insight into the restructuring, an understanding of the mixed model formulas used in **lme4** is essential. For an in-depth explanation on how the formula module of **lme4** works, see Bates *et al.* (2015), Section 2.1.

Suppose we want to fit a mixed model with two grouping factors g1 and g2. Within the first grouping factor g1, there are three continuous variables v1, v2 and v3 and within the second grouping factor there is only one variable x. Thus there are not only random intercepts but also random slopes that are possibly correlated. Such a model with response y would be fitted in **lme4** using

```
R> m <- lmer(y ~ (v1 + v2 + v3 | g1) + (x | g2), exampledata)
```

```
singular fit
```

The singular fit statement here already indicates that potentially some parameters are on the boundary of the parameter space. In mixed models fitted with **lme4**, the random effects covariance matrix $\boldsymbol{D_\theta}$ always has block-diagonal structure. For instance in the example from above the Cholesky factorized blocks of the estimated $\boldsymbol{D_\theta}$ associated with each random effects term are

```
R> getME(m, "ST")
```

```
$g2
            [,1] [,2]
[1,]  1.18830318  NaN
[2,] -0.01488366    0

$g1
              [,1]          [,2] [,3] [,4]
[1,]  1.0184629027  0.00000000  NaN  NaN
[2,] -0.1438760400  0.05495793  NaN  NaN
[3,] -0.0007341786  0.19904380    0  NaN
[4,] -0.0883652162 -1.36463316 -Inf    0
```

Notice that as some of the parameters are on the boundary of the parameter space the results can be numerically instable. If any of the diagonal elements of the blocks are zero the corresponding random effects terms are deleted from the formula.In **lme4** this is done conveniently by rebuilding the formula from the component names list.

```
R> m@cnms
```

```
$g2
[1] "(Intercept)" "x"

$g1
[1] "(Intercept)" "v1"           "v2"           "v3"
```

Thus a new model formula can be obtained by designing a new components names list without the components associated with the boundary issues:

```
R> varBlockMatrices <- getME(m, "ST")
R> cnms <- m@cnms
R> for (i in 1:length(varBlockMatrices)) {
+    cnms[[i]] <- cnms[[i]][which(diag(varBlockMatrices[[i]]) != 0)]
+ }
R> cnms

$g2
[1] "(Intercept)"

$g1
[1] "(Intercept)" "v1"
```

The `cnms2formula` function from the **cAIC4** package forms a new formula from the `cnms` object above. Hence the new formula can be computed by

```
R> rhs <- cAIC4:::cnms2formula(cnms)
R> lhs <- formula(m)[[2]]
R> reformulate(rhs, lhs)

y ~ (1 | g2) + (1 + v1 | g1)
```

This code is called by function `deleteZeroComponents` in the **cAIC4** package. This function automatically deletes all zero components from the model. Function `deleteZeroComponents` is called recursively, i.e., the new model is checked again for zero components. In the example above only the random intercepts are non-zero. Hence the formula of the reduced model from which the conditional AIC is calculated is

```
R> formula(cAIC4:::deleteZeroComponents(m))

singular fit

y ~ (1 | g2) + (1 | g1)
```

With the new model the conditional AIC is computed. If there are no random effect terms left in the formula, a linear model and the conventional AIC is returned. In addition function `deleteZeroComponents` also accounts for several special cases that may occur.

Notice however that in case of using smoothing terms from **gamm4** no automated check for boundary issues can be applied and zero components have to be manually deleted. Furthermore the boundary issues are only relevant for the analytic Gaussian case. For generalized models some kind of refitting is used to calculate the cAIC and thus the boundary issues are not relevant.

B. Computational matters

B.1. Gaussian responses

The corrected conditional AIC proposed in Greven and Kneib (2010) accounts for the uncertainty induced by the estimation of the random effects covariance parameters $\boldsymbol{\theta}$. In order to adapt the findings of Greven and Kneib (2010), a number of quantities from the `lmer` model fit need to be extracted and transformed. In the following these computations are presented. They are designed to minimize the computational burden and maximize the numerical stability. Parts of the calculations needed, for instance the Hessian of the ML/REML criterion, can also be found in Bates *et al.* (2015). Notice however, that **lme4** does not explicitly calculate these quantities but uses derivative free optimizers for the profile likelihoods. Parts of the computations are similar to those performed in the **merDeriv** package, see Wang and Merkle (2018).

A core ingredient of mixed models is the covariance matrix of the marginal responses $\boldsymbol{y}$. The inverse of the scaled covariance matrix $\boldsymbol{V}_0$ will be used in the following calculations:

$$\boldsymbol{V} = \mathrm{cov}(\boldsymbol{y}) = \sigma^2 \left(\boldsymbol{I}_n + \boldsymbol{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{\top}\boldsymbol{Z}^{\top} \right) = \sigma^2 \boldsymbol{V}_0.$$

Large parts of the computational methods in **lme4** rely on a sparse Cholesky factor that satisfies

$$\boldsymbol{L}_{\boldsymbol{\theta}}\boldsymbol{L}_{\boldsymbol{\theta}}^{\top} = \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{\top}\boldsymbol{Z}^{\top}\boldsymbol{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}} + \boldsymbol{I}_q. \tag{18}$$

From this equation and keeping in mind that $\boldsymbol{I} - \boldsymbol{V}_0^{-1} = \boldsymbol{Z} \left(\boldsymbol{Z}^{\top}\boldsymbol{Z} + \left(\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{\top} \right)^{-1} \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{-1} \right)^{-1} \boldsymbol{Z}^{\top}$, see Greven and Kneib (2010), it follows that

$$\boldsymbol{\Lambda}_{\boldsymbol{\theta}} \left(\boldsymbol{L}_{\boldsymbol{\theta}}^{\top} \right)^{-1} \boldsymbol{L}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{\top} = \left(\boldsymbol{Z}^{\top}\boldsymbol{Z} + \left(\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{\top} \right)^{-1} \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{-1} \right)^{-1}$$

$$\Rightarrow \quad \boldsymbol{I} - \boldsymbol{V}_0^{-1} = \left(\boldsymbol{L}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{\top}\boldsymbol{Z}^{\top} \right)^{\top} \left(\boldsymbol{L}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{\top}\boldsymbol{Z}^{\top} \right).$$

Hence the inverse of the scaled variance matrix $\boldsymbol{V}_0^{-1}$ can be efficiently computed with the help of the R package **Matrix** (see Bates and Maechler 2021) that provides methods specifically for sparse matrices:

```
R> Lambdat <- getME(m, "Lambdat")
R> V0inv <- diag(rep(1, n)) -
+    crossprod(solve(getME(m, "L"), system = "L") %*%
+    solve(getME(m, "L"), Lambdat, system = "P") %*% t(Z))
```

Notice that `solve(getME(m, "L"), Lambdat, system = "P")` accounts for a fill-reducing permutation matrix $\boldsymbol{P}$ associated (and stored) with $\boldsymbol{L}_{\boldsymbol{\theta}}$, see Bates *et al.* (2015). Another quantity needed for the calculation of the corrected degrees of freedom in the conditional AIC are the derivatives of the scaled covariance matrix of the responses $\boldsymbol{V}_0$ with respect to the jth element of the parameter vector $\boldsymbol{\theta}$:

$$\boldsymbol{W}_j = \frac{\partial}{\partial \theta_j} \boldsymbol{V}_0 = \boldsymbol{Z}\boldsymbol{D}_{\boldsymbol{\theta}}^{(j)}\boldsymbol{Z}^{\top},$$

where the derivative of the scaled covariance matrix of the random effects with respect to the jth variance parameter is defined by

$$\boldsymbol{D}_{\boldsymbol{\theta}}^{(j)} = \frac{1}{\sigma^2} \frac{\partial}{\partial \theta_j} \boldsymbol{D}_{\boldsymbol{\theta}}.$$

Notice that $\boldsymbol{D}_{\boldsymbol{\theta}} = [d_{st}]_{s,t=1,\dots,q}$ is symmetric and block-diagonal and its scaled elements are stored in $\boldsymbol{\theta}$, hence $d_{st} = d_{ts} = \theta_j \sigma^2$, for certain t, s and j. Thus the matrix $\boldsymbol{D}_{\boldsymbol{\theta}}^{(j)} = \left[d_{st}^{(j)}\right]_{s,t=1,\dots,q}$ is sparse with

$$d_{st}^{(j)} = \begin{cases} 1, & \text{if } d_{st} = d_{ts} = \theta_j \sigma^2 \\ 0, & \text{else.} \end{cases}$$

The derivative matrices $\boldsymbol{W}_j$ can be derived as follows:

```
R> Lambda <- getME(m, "Lambda")
R> ind <- getME(m, "Lind")
R> len <- rep(0, length(Lambda@x))
R> for (j in 1:length(theta)) {
+     LambdaS <- Lambda
+     LambdaSt <- Lambdat
+     LambdaS@x <- LambdaSt@x <- len
+     LambdaS@x[which(ind == j)] <- LambdaSt@x[which(ind == j)] <- 1
+     diagonal <- diag(LambdaS)
+     diag(LambdaS) <- diag(LambdaSt) <- 0
+     Dj <- LambdaS + LambdaSt
+     diag(Dj) <- diagonal
+     Wlist[[j]] <- Z %*% Dj %*% t(Z)
+ }
```

The following matrix is essential to derive the corrected AIC of Theorem 3 in Greven and Kneib (2010). Adapting their notation, the matrix is

$$\boldsymbol{A} = \boldsymbol{V}_0^{-1} - \boldsymbol{V}_0^{-1} \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{V}_0^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{V}_0^{-1}.$$

Considering that the cross-product of the fixed effects Cholesky factor is

$$\boldsymbol{X}^\top \boldsymbol{V}_0^{-1} \boldsymbol{X} = \boldsymbol{R}_{\boldsymbol{X}}^\top \boldsymbol{R}_{\boldsymbol{X}},$$

the matrix $\boldsymbol{A}$ can be rewritten as

$$\boldsymbol{A} = \boldsymbol{V}_0^{-1} - \left(\boldsymbol{X} \boldsymbol{R}_{\boldsymbol{X}}^{-1} \boldsymbol{V}_0^{-1}\right) \left(\boldsymbol{X} \boldsymbol{R}_{\boldsymbol{X}}^{-1} \boldsymbol{V}_0^{-1}\right)^\top.$$

Accordingly the computation in R can be done as follows:

```
R> A <- V0inv - crossprod(crossprod(X %*% solve(getME(m, "RX")), V0inv))
```

With these components, the Hessian matrix

$$\boldsymbol{B} = \frac{\partial^2 \mathrm{REML}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \ \text{or} \ \boldsymbol{B} = \frac{\partial^2 \mathrm{ML}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}$$

and the matrix

$$\boldsymbol{G} = \frac{\partial^2 \mathrm{REML}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{y}^\top} \ \text{or} \ \boldsymbol{G} = \frac{\partial^2 \mathrm{ML}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{y}^\top},$$

depending on whether the restricted or the marginal profile log-likelihood $\mathrm{REML}(\boldsymbol{\theta})$ or $\mathrm{ML}(\boldsymbol{\theta})$ is used, can be computed straightforward as in Greven and Kneib (2010). Depending on the optimization, it may not even be necessary to compute the matrix $\boldsymbol{B}$. Considering that $\boldsymbol{B}$ is the Hessian of the profile (restricted) log-likelihood, the matrix can also be taken from the model fit, although this is only a numerical approximation. If the Hessian is computed it is stored in:

```
R> B <- m@optinfo$derivs$Hessian
```

The inverse of $\boldsymbol{B}$ does not need to be calculated – instead, if $\boldsymbol{B}$ is positive definite, a Cholesky decomposition and two backward solves are sufficient:

```
R> Rchol <- chol(B)
R> L1 <- backsolve(Rchol, G, transpose = TRUE)
R> Gammay <- backsolve(Rchol, L1)
```

The trace of the hat matrix, the first part of the effective degrees of freedom needed for the cAIC, can also easily be computed with the help of the residual matrix $\boldsymbol{A}$:

```
R> df <- n - sum(diag(A))
```

The correction needed to account for the uncertainty induced by the estimation of the variance parameters can be added for each random effects variance parameter separately by calculating

```
R> for (j in 1:length(theta)) {
+    df <- df + sum(Gammay[j,] %*% A %*% Wlist[[j]] %*% A %*% y)
+ }
```

B.2. Poisson responses

The computation of the bias correction for Poisson distributed responses is obtained differently. In a first step the non-zero responses need to be identified and a matrix with the responses in each column is created. Consider the `grouseticks` example in Section 4.2 with the model `p1` fitted by `glmer`.

```
R> y <- p1@resp$y
R> ind <- which(y != 0)
R> workingMatrix <- matrix(rep(y, length(y)), ncol = length(y))
```

The diagonal values of the matrix are reduced by one and only those columns of the matrix with non-zero responses are kept.

```
R> diag(workingMatrix) <- diag(workingMatrix) - 1
R> workingMatrix <- workingMatrix[, ind]
```

Now the `refit()` function can be applied to the columns of the matrix in order to obtain the estimates $\log \hat{\mu}_i(\boldsymbol{y}_{-i}, y_i - 1)$ in (13) from the reduced data.

```
R> workingEta <- diag(apply(workingMatrix, 2, function(x)
+    refit(p1, newresp = x)@resp$eta)[ind, ])
```

The computation of the bias correction is then straightforward:

```
R> sum(y[ind] * (p1@resp$eta[ind] - workingEta))
```

```
[1] 205.5785
```

and corresponds to the bias correction obtained in Section 4.2.

B.3. Bernoulli

The computation of an estimator of the bias correction for Bernoulli distributed responses as in Equation 14 is similar to the implementation for Poisson distributed responses above. Therefore consider any Bernoulli model `b1` fitted by the `glmer` function in **lme4**. For the calculation of the bias correction for each observed response variable the model needs to be refitted with corresponding other value, i.e., 0 for 1 and vice versa. This is done best by use of the `refit()` function from **lme4**.

```
R> muHat <- b1@resp$mu
R> workingEta <- numeric(length(muHat))
R> for (i in 1:length(muHat)) {
+    workingData <- b1@resp$y
+    workingData[i] <- 1 - workingData[i]
+    workingModel <- refit(b1, nresp = workingData)
+    workingEta[i] <- log(workingModel@resp$mu[i] /
+      (1 - workingModel@resp$mu[i])) - log(muHat[i] / (1 - muHat[i]))
+ }
```

The sign of the re-estimated logit (the natural parameter) in (14) which is stored in the vector `workingEta` needs to be taken into account, i.e., $\hat{\eta}_i(1)$ is positive and $\hat{\eta}_i(0)$ negative. With a simple sign correction

```
R> signCor <- - 2 * b1@resp$y + 1
```

the following returns the bias correction:

```
R> sum(muHat * (1 - muHat) * signCor * workingEta)
```

It should be pointed out that for the conditional AIC it is essential to use the conditional log-likelihood with the appropriate bias correction. Notice that the log-likelihood that by default is calculated by the S3 method `logLik` for class 'merMod' (the class of a mixed model fitted by a `lmer` call) is the marginal log-likelihood.

Affiliation:

Benjamin Säfken
RTG 1644: Scaling Problems in Statistics
Universität Göttingen
37073 Göttingen, Germany
E-mail: bsaefke@uni-goettingen.de
URL: https://www.uni-goettingen.de/de/582009.html

David Rügamer, Sonja Greven
Ludwig-Maximilians-Universität München
E-mail: david.ruegamer@stat.uni-muenchen.de, sonja.greven@stat.uni-muenchen.de

Thomas Kneib
Chair of Statistics
Universität Göttingen
37073 Göttingen, Germany
E-mail: tkneib@uni-goettingen.de
URL: http://www.uni-goettingen.de/de/264255.html