Electronic Journal of Statistics
Vol. 8 (2014) 201–225
ISSN: 1935-7524
DOI: 10.1214/14-EJS881

A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models

Benjamin Saefken, Thomas Kneib

Chair of Statistics, Georg-August University Goettingen
Platz der Goettinger Sieben 5
37073 Goettingen, Germany
e-mail: bsaefke@uni-goettingen.de; tkneib@uni-goettingen.de

Clara-Sophie van Waveren

Department of Ecosystem Modelling, Georg-August University Goettingen
Buesgenweg 4
37077 Goettingen, Germany
e-mail: c.vanwaveren@stud.uni-goettingen.de

and

Sonja Greven

Department of Statistics, Ludwig-Maximilians-University Munich
80539 Munich, Germany
e-mail: sonja.greven@stat.uni.muenchen.de

Abstract: The conditional Akaike information criterion, AIC, has been frequently used for model selection in linear mixed models. We develop a general framework for the calculation of the conditional AIC for different exponential family distributions. This unified framework incorporates the conditional AIC for the Gaussian case, gives a new justification for Poisson distributed data and yields a new conditional AIC for exponentially distributed responses but cannot be applied to the binomial and gamma distributions. The proposed conditional Akaike information criteria are unbiased for finite samples, do not rely on a particular estimation method and do not assume that the variance-covariance matrix of the random effects is known. The theoretical results are investigated in a simulation study. The practical use of the method is illustrated by application to a data set on tree growth.

AMS 2000 subject classifications: Primary 62J12; secondary 62J07.
Keywords and phrases: Conditional Akaike information criterion, Kullback-Leibler distance, model selection, random effects, generalized linear mixed models.

Received June 2013.

Contents

1. Introduction

Generalized linear mixed models (Breslow & Clayton (1993)), GLMMs, provide a broad range of models for the analysis of grouped data. They extend the idea of linear mixed models to non-normal data. In recent years, GLMMs have also been used as a representation of generalized additive models (e.g. Ruppert et al. (2003)). This increase in flexibility and complexity leads to extended need for model selection.

The Akaike information criterion, AIC (Akaike (1973)), is a well known information based criterion for model selection. There have been several extensions to the AIC. For example in the case of small sample size or highly overparameterized models Hurvich & Tsai (1989) proposed a corrected criterion called AIC_C. In linear mixed models, a natural choice would be to base the AIC on the marginal model, i.e. the model with the random effects integrated out. This leads to a biased criterion (Greven & Kneib (2010)). An AIC based on the conditional likelihood was introduced by Vaida & Blanchard (2005) but was derived assuming the variance parameters of the random effects to be known. Plugging in estimated variance-covariance matrices induces a bias that leads to a preference for larger models with more random effects (Greven & Kneib (2010)).

A correction to avoid that bias was proposed by Liang et al. (2008) by use of an identity known from Stein (1972).

An extension of the conditional AIC to GLMMs has for example been proposed by Yu & Yau (2012). They suggested an asymptotically unbiased conditional AIC, where the estimation of the variance parameters of the random effects is based on maximum likelihood and that of the fixed and random effects on maximizing the joint likelihood. Another conditional AIC was proposed by Donohue et al. (2011). It is also asymptotically unbiased and in addition requires that the covariance structure of the random effects is known. In this report, we suggest a method for deriving unbiased estimates of the conditional Akaike information for exponential family distributions even if the sample size is finite and the covariance structure of the random effects is unknown. This unified framework for the conditional AIC in GLMMs contains the known estimators for the normal and Poisson distribution as special cases and provides a more general derivation for the Poisson case than previously given (Lian (2011)), which highlights the connection to the normal case. We also extend this idea to the exponential distribution. In addition to the theoretical results, we illustrate the performance of the new estimator in a simulation study and in an application to tree growth data. Proofs of new results are given in the appendix.

2. Bias correction for the conditional AIC in GLMMs

2.1. Generalized linear mixed models

Consider a GLMM with predictor

$$\eta = X\beta + Zu$$

with the full column rank $(n \times p)$ and $(n \times r)$ design matrices X and Z, the fixed effects β and random effects u. The random effects are assumed to be normally distributed, i.e. $u \sim \mathcal{N}(0, G(\vartheta))$, where ϑ contains all q variance parameters in the covariance matrix G. The responses $y_1, \ldots, y_n$ have conditional expectation

$$\mu_i = \mathbb{E}(y_i|u) = h(\eta_i)$$

with response function $h(\cdot)$. Moreover the responses conditioned on the random effects u follow an exponential family distribution, i.e. the conditional density of y_i is given by

$$\log\left(f(y_i|\beta, u)\right) = \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \tag{2.1}$$

where $b(\cdot)$ only depends on θ, $c(\cdot)$ only on y_i and ϕ, ϕ is a scale parameter, and θ is the canonical parameter of the distribution as in the generalized linear model context (Nelder & Wedderburn (1972)). In the marginal density, the random effects are integrated out

$$f(y|\beta, \vartheta) = \int f(y|\beta, u) f(u|\vartheta) du \propto |G(\vartheta)|^{-\frac{1}{2}} \int \exp\left(f(y|\beta, u) - \frac{1}{2} u^t G(\vartheta)^{-1} u \right) du$$

where $f(u|\vartheta)$ is the density of the random effects. In the following, we denote by $\hat{\beta}$, $\hat{\theta}$ and $\hat{u}$ estimators of β, θ and u, respectively, e.g. the maximum likelihood estimator, the restricted maximum likelihood estimator and the empirical Bayes estimator. If we want to emphasize the dependence on the data y, we write $\hat{\beta}(y)$ and so forth.

2.2. Akaike information criterion

The Akaike information is defined as twice the expected relative Kullback-Leibler distance $-2E_y(E_z(\log f(z|\hat{\gamma}(y))))$, with independent replications z and y from the underlying model and parameter vector $\hat{\gamma}$. In standard regression settings, if certain regularity conditions are fulfilled, the Akaike information criterion

$$AIC = -2\log\left(f(y|\hat{\gamma}(y))\right) + 2\nu \tag{2.2}$$

with $\nu = dim(\gamma)$ is an asymptotically unbiased estimator for the Akaike information. A direct extension of the AIC to GLMMs based on the marginal model would be the marginal AIC,

$$mAIC = -2\log\left(f(y|\hat{\beta},\hat{\vartheta})\right) + 2(p+q) \tag{2.3}$$

where $f(y|\hat{\beta},\hat{\vartheta})$ is the maximized marginal likelihood. If the dispersion parameter ϕ is estimated, the bias correction in (2.3) changes to $2(p+q+1)$. Using the marginal model implies that the focus is on the fixed effects and that new data z does not share the random effects of y. However, the marginal AIC may be inappropriate for variable selection in linear mixed effect models if the focus is on clusters rather than on the population, as stated in Vaida & Blanchard (2005). Even under the marginal model it is not an (asymptotically) unbiased estimator of the Akaike information as shown for the Gaussian case by Greven & Kneib (2010).

Use of the conditional model formulation focuses on the random effects and implies shared random effects between y and z. The conditional Akaike information is

$$\begin{aligned} cAI &= -2\mathbb{E}_{y,u}\left[\mathbb{E}_{z|u}\left[\log\left(f(z|\hat{\beta}(y),\hat{u}(y))\right)\right]\right] \\ &= -\int 2\log\left(f(z|\hat{\beta}(y),\hat{u}(y))\right) g(z|u)g(y,u)dz\ dy\ du, \end{aligned}$$

where $g(y,u) = g(y|u)g(u)$ is the (true) joint density of y and u (Vaida & Blanchard (2005)). For (conditionally) Gaussian responses and known random effects variance parameters ϑ they show that an asymptotically unbiased estimator of the conditional Akaike information is

$$cAIC = -2\log f(y|\hat{\beta},\hat{u}) + 2(\rho+1),$$

where

$$\rho = tr\left[\begin{pmatrix} X^tX & X^tZ \\ Z^tX & Z^tZ + \sigma^2 G(\vartheta)^{-1} \end{pmatrix}^{-1} \begin{pmatrix} X^tX & X^tZ \\ Z^tX & Z^tZ \end{pmatrix}\right]$$

are the effective degrees of freedom (Hodges & Sargent (2001)). Liang et al. (2008) introduced a bias correction that takes the estimation uncertainty of ϑ into account. For known error variance σ^2 they replace 2ρ by

$$2\Phi_0(y) = 2\sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i} = 2tr\left(\frac{\partial \hat{y}}{\partial y}\right). \tag{2.4}$$

They propose a similar but lengthy formula for unknown error variance. Following the findings of Greven & Kneib (2010), the estimation uncertainty of the error variance can be ignored.

2.3. Bias correction

For GLMMs with responses following an exponential family distribution as in (2.1) and unknown random effects variance parameters ϑ, we derive the following bias correction.

Proposition 2.1. *In GLMMs with responses following an exponential family distribution and unknown ϑ, the bias correction for $-2\log f(y|\hat{\beta}, \hat{u})$ as an estimator of cAI is*

$$\begin{aligned} 2\Psi &= 2\sum_{i=1}^{n} \mathbb{E}_{y,u}\left[\frac{y_i - \mu_i}{\phi}\hat{\theta}_i(y)\right] \\ &= \frac{2}{\phi}\sum_{i=1}^{n}\left\{\mathbb{E}_{y,u}\left[y_i\hat{\theta}_i(y)\right] - \mu_i\mathbb{E}_{y,u}\left[\hat{\theta}_i(y)\right]\right\}. \end{aligned} \tag{2.5}$$

If ϕ is estimated, ϕ in the first expression is replaced by $\hat{\phi}$. A proof for this result is given in the Technical details section.

3. Stein's method for exponential families

3.1. Continuous distributions

The proposed bias correction in (2.5) suffers from the use of the true but unknown mean μ and therefore cannot be applied directly. Liang et al. (2008) solved this problem by the use of a formula known from Stein (1972) which turns (2.5) into (2.4). The following result extends the idea of Stein to continuous exponential family distributions and is a slight modification of Hudson (1978).

Theorem 3.1. *Let y be continuous and have density given by (2.1). For a differentiable function $m : \mathbb{R} \to \mathbb{R}$ that vanishes on the limits of the support of y if the limits of the support are finite and $\mathbb{E}[|m'(y)|] < \infty$ if the limits are infinite it holds that*

$$\mathbb{E}\left[m'(y)\right] = \mathbb{E}\left[-\left(\frac{\theta}{\phi} + \frac{\partial}{\partial y}c(y,\phi)\right)m(y)\right]. \tag{3.1}$$

If y is Gaussian, formula (3.1) simplifies to

$$\mathbb{E}\left[m'(y)\right] = \mathbb{E}\left(\frac{y-\mu}{\sigma^2}m(y)\right),$$

the formula known from Stein. Applied to the bias correction (2.5) this yields the bias correction $2\Phi_0$ known from Liang et al. (2008). The theorem can also be applied to obtain bias corrections for other exponential family distributions as stated in the following. For y exponentially distributed with mean μ, $y \sim \mathcal{E}(\frac{1}{\mu})$, and letting $m(y) = \int_0^y g(x)\,dx$, equation (3.1) becomes

$$\mu\mathbb{E}\left[g(y)\right] = E\left[\int_0^y g(x)dx\right]. \tag{3.2}$$

We use this equation to derive an analytically accessible version of (2.5).

Corollary 3.1. *Let $y_i|u \sim \mathcal{E}(\frac{1}{\mu_i})$. Then an unbiased estimator of the cAI is*

$$cAIC = -2\log f(y|\hat{\beta},\hat{u}) + 2\Psi,$$

with

$$\Psi = \sum_{i=1}^{n} y_i\hat{\theta}_i(y) - \int_0^{y_i}\hat{\theta}_i(y_{-i},x)dx \tag{3.3}$$

where y_{-i} is the vector of observed responses without the $i-th$ observation and hence $\hat{\theta}_i(y_{-i},x)$ is the estimator based on $(y_1,\ldots,y_{i-1},x,y_{i+1},\ldots,y_n)$.

A proof of this result is outlined in the appendix. In Section 5 numerical integration is used to evaluate 3.3.

3.2. Discrete distributions

A similar identity to Theorem 3.1 also holds for discrete random variables from an exponential family distribution. The following Theorem is also due to Hudson (1978).

Theorem 3.2. *Let y be a discrete random variable taking values in $\mathbb{N}_0 = \{0,1,2,\ldots\}$ and let y have probability function given by (2.1). For $m : \mathbb{N} \to \mathbb{R}$ with $\mathbb{E}[|m(y)|] < \infty$ it holds that*

$$\exp(\theta)\mathbb{E}\left(m(y)\right) = \mathbb{E}\left[t(y)m(y-1)\right] \tag{3.4}$$

where

$$t(x) := \begin{cases} 0, & \text{for } x = 0 \\ \exp\left(c(x-1,\phi) - c(x,\phi)\right), & \text{for } x = 1, 2, \ldots \end{cases}$$

For y Poisson distributed with parameter λ, $y \sim \mathcal{P}(\lambda)$, equation (3.4) simplifies to

$$\lambda \mathbb{E}\left[m(y)\right] = \mathbb{E}\left[y m(y-1)\right], \tag{3.5}$$

with $ym(y-1) = 0$ if $y = 0$ by convention. This is an identity due to Chen (1975). With the help of this identity the bias correction (2.5) can be made analytically accessible.

Corollary 3.2. *Let $y_i|u \sim \mathcal{P}(\lambda_i)$. Then an unbiased estimator of the cAI is*

$$cAIC = -2 \log f(y|\hat{\beta}, \hat{u}) + 2\Psi,$$

with

$$\Psi = \sum_{i=1}^{n} y_i \left(\hat{\theta}_i(y) - \hat{\theta}_i(y_{-i}, y_i - 1)\right), \tag{3.6}$$

where y_{-i} is the vector of observed responses without the $i-th$ observation and y_i is the $i-th$ observation with $y_i\hat{\theta}_i(y_{-i}, y_i - 1) = 0$ if $y_i = 0$ by convention.

Corollary 3.2 gives an alternative derivation of the result in Lian (2011), which highlights the connection to the normal case.

4. Limits of the approach

Theorem 3.1 and Theorem 3.2 can be extended to further distributions. For instance the generalized SURE formula (Lemma 2) in Shen & Huang (2006) is a generalisation of Theorem 3.1 and Theorem 3.2 to distributions not necessarily from the exponential family. Although the formula has been obtained in a different context, it is closely related to the findings in Section 3 and gives further insight on how identities for further distributions could potentially be derived. On the other hand, formulas as in Theorems 3.1 and 3.2 do not necessarily lead to bias correction terms computable from observable quantities for all distributions, as discussed in the following.

4.1. Continuous distributions

For example if y follows a gamma distribution with mean μ and scale parameter ν, i.e. $y \sim \mathcal{G}(\mu, \nu)$ identity (3.1) is

$$\mathbb{E}\left(m'(y)\right) = \mathbb{E}\left[\left(\frac{\nu}{\mu} - \left(\frac{\nu}{y} - \frac{1}{y}\right)\right) m(y)\right].$$

This can be rewritten in terms of μ

$$\mu \mathbb{E}\left[m'(y) + \left(\frac{\nu}{y} - \frac{1}{y}\right) m(y)\right] = \nu \mathbb{E}\left[m(y)\right].$$

In contrast to the $\nu = 1$ case, this identity cannot be used to remove the true but unknown parameter μ_i in the bias correction term (2.5) unless we could rewrite the estimator of the canonical parameter in (2.5) by

$$\hat{\theta}_i(y_{-i}, y_i) = m'(y_i) + \left(\frac{\nu}{y_i} - \frac{1}{y_i}\right) m(y_i)$$

for a function $m(\cdot)$ fulfilling the requirements in Theorem 3.1. Since this seems to be not possible, Theorem 3.1 cannot be used to rewrite the bias correction term (2.5) for a gamma distribution with $\nu \neq 1$.

4.2. Discrete distributions

Similarly, applying Theorem 3.2 to the negative binomial distribution where y has the probability function

$$f(y|\mu, \lambda) = \frac{\Gamma(\lambda + y)}{\Gamma(\lambda) y!} \frac{\mu^y \lambda^\lambda}{(\mu + \lambda)^{(\lambda + y)}},$$

identity (3.1) becomes

$$\frac{\mu}{\mu + \lambda} \mathbb{E}\left(m(y)\right) = \mathbb{E}\left(\frac{y}{y + \lambda - 1} m(y - 1)\right)$$

with $m(y - 1) = 0$ for $y = 0$. In terms of the mean μ, the identity above is

$$\mu \left(\mathbb{E}\left(m(y) - \frac{y}{y + \lambda - 1} m(y - 1)\right)\right) = \lambda \mathbb{E}\left(\frac{y}{y + \lambda - 1} m(y - 1)\right).$$

The second part of the bias correction (2.5), i.e. $\mu_i \mathbb{E}_{y,u}(\hat{\theta}_i(y))$ could therefore only be replaced if the estimator for the canonical parameter $\hat{\theta}_i(\cdot)$ can be written as

$$\hat{\theta}_i(y) = m(y) - \frac{y}{y + \lambda} m(y - 1)$$

for some arbitrary function $m(\cdot)$ as in Theorem 3.2. This is not possible.

Theorem 3.2 cannot be applied to the binomial distribution $\mathcal{B}(n, p)$ since a binomially distributed random variable only takes values in $\{0, 1, \ldots, n\} \subset \mathbb{N}_0$. Extending the distribution by defining $P(y = n + k) = 0 \; \forall k \in \mathbb{N}$ does not yield an identity which could be applied to the bias correction (2.5), for the same reason as in the case of the negative binomial distribution.

5. Simulation study

In the first part of this simulation study, we concentrate on random intercept models. The bias corrections (3.3) and (3.6) are analysed in two different ways. First we compare the precision and the variability of different bias corrections as estimators of the correction term obtained by estimating the relative Kullback-Leibler distance with the log-likelihood. In a second step, the model choice

behaviour of the bias correction for exponential responses (3.3) and Poisson distributed responses (3.6) is assessed.

The second part of the simulation study is concerned with the model choice behaviour of the proposed estimators for smoothing spline models. We therefore use a common link between mixed-effects models and smoothing spline models.

5.1. Random intercept model

5.1.1. Exponential distribution

First we will focus on the precision and the variability of the proposed bias correction (3.3). We therefore consider a model with an exponentially distributed response y_{ij} and a random intercept u_i with

$$\mu_{ij} = \exp(\beta_0 + \beta_1 x_j + u_i);\ i = 1, \ldots, m;\ j = 1, \ldots, n_i, \tag{5.1}$$

where $u_i \sim \mathcal{N}(0, \tau^2)$, $\beta_0 = 0.1$, $\beta_1 = 0.2$ and $x_j = j$. Different numbers of clusters, cluster sizes and random effect variances are considered: $m = 5, 10$, $n_i = 5, 10$ for $i = 1, \ldots, m$ and the random effect variances are $\tau^2 = 0,\ 0.5,\ 1$. For each of the settings, 1,000 data sets are generated and the mean and the standard deviations of the different bias correction terms are calculated. The model is fitted by the PQL method as introduced by Breslow & Clayton (1993). We use an implementation in R based on Wood (2006).

We compare the proposed estimator for the bias correction Ψ obtained from refitting the model for each i with the true bias BC defined by (2.5), the asymptotically unbiased estimator $\hat{\rho}_{ml}$ proposed by Yu & Yau (2012) and the estimator $\hat{\rho}_{Don}$ of Donohue et al. (2011). The true bias correction BC is derived by averaging 30,000 samples of (2.5) based on model (5.1). This criterion used as a benchmark is not available in practice since for its calculation the true mean μ has to be known.

For the proposed bias correction Ψ as in (3.3), an integral needs to be evaluated. Since this can not be done analytically it is approximated by adaptive quadrature. The resulting bias correction is used to obtain the proposed cAIC.

The cAIC suggested by Yu & Yau (2012) is included to assess the performance of an asymptotically unbiased estimator of the cAI in finite sample settings. Similarly to the cAIC suggested by Vaida & Blanchard (2005) for Gaussian responses, the cAIC proposed by Donohue et al. (2011) requires known random effects variance parameters. For known random effects variance parameters, the criterion is consistent. In our simulated random intercept model, τ^2 would need to be known. Since in many applications this will not be the case, we use the proposed bias correction of Donohue et al. (2011) with the estimated variance parameter $\hat{\tau}^2$ taken as truth.

In the calculation of $\hat{\rho}_{ml}$, the bias correction proposed by Yu & Yau (2012), numerical difficulties occurred. We therefore excluded all results in which the bias correction exceeded a threshold of 200. This excluded between 0 and 5 observations per setting.

TABLE 1

Mean estimated values of four different estimators of the bias correction (2.5) and the corresponding standard deviations (indicated by σ with corresponding index) of model (5.1) for different cluster sizes (n_i), number of clusters (m) and variances of random effects (τ^2). The true bias correction BC is derived by (2.5), the estimator Ψ is directly calculated by (3.3), $\hat{\rho}_{ml}$ is the estimator proposed by Yu & Yau (2012) and $\hat{\rho}_{Don}$ is the estimator proposed by Donohue et al. (2011)

m	n_i	τ^2	BC	Ψ	$\hat{\rho}_{ml}$	$\hat{\rho}_{Don}$	σ_Ψ	$\sigma_{\hat{\rho}_{ml}}$	$\sigma_{\hat{\rho}_{Don}}$
5	5	0	3.66	3.54	3.72	2.54	1.64	9.08	0.93
5	5	0.5	5.21	5.24	5.03	3.55	2.01	9.28	1.31
5	5	1	6.72	6.77	5.44	4.73	1.81	3.59	1.19
5	10	0	3.08	3.10	3.36	2.45	1.38	7.05	0.83
5	10	0.5	5.30	5.32	5.04	4.08	1.56	4.74	1.24
5	10	1	6.21	6.27	5.65	5.22	1.17	1.13	0.85
10	5	0	4.12	4.22	4.47	3.19	2.62	8.22	1.86
10	5	0.5	8.24	8.38	7.60	6.20	3.03	6.55	2.54
10	5	1	11.58	11.80	9.59	9.09	2.06	7.43	1.51
10	10	0	3.51	3.46	4.21	2.80	2.03	12.28	1.47
10	10	0.5	9.12	9.09	8.50	7.62	2.05	1.76	2.01
10	10	1	11.18	11.28	10.16	9.87	1.25	0.68	0.85

Table 1 shows the results. They suggest, that the proposed estimator performs well although numerical integration was used. The estimator $\hat{\rho}_{ml}$ has the tendency to underestimate the true bias correction for positive true τ^2 and to overestimate it for true $\tau^2 = 0$. This may be due to the fact that a non-canonical link function was used, while the authors derive their estimator only for canonical links. Furthermore the authors do not use PQL as fitting method, see 5.3 for a short remark.

The estimator $\hat{\rho}_{Don}$ consistently underestimates BC, as it ignores variability due to the estimation of the variance components. The last four columns give the standard deviations of each estimator. The standard deviation of the proposed estimator is low, which also speaks in favour of the estimator. The standard deviation of $\hat{\rho}_{ml}$ is very high especially for low random effects variance, despite the exclusion of extreme values.

We now consider the behaviour of the proposed bias correction (3.3) when selecting random effects. Therefore consider the same settings as in model (5.1) but with the random effect variances as $\tau^2 = 0,\ 0.1,\ 0.2,\ \ldots,\ 1.8$, respectively. For each of the settings, 1000 data sets are generated and one model containing a random intercept ($\tau^2 \geq 0$) and another (generalized linear) model without random effects are fitted to each data set. The random effects model is fitted by PQL, see Breslow & Clayton (1993) and Wood (2006).

We compute the frequency of selecting the model including the random intercept ($\tau^2 > 0$), which is chosen whenever the proposed AIC is smaller than an AIC, derived from the model without a random intercept ($\tau^2 = 0$). As reference AICs for the model without random intercept we use (2.2) for the marginal AIC, Donohue's cAIC and Yu & Yau's cAIC. For the proposed cAIC we use formula (3.3) with a generalized linear model as reference. Thus, for each AIC we use as a reference the AIC it reduces to in the null model without intercept.

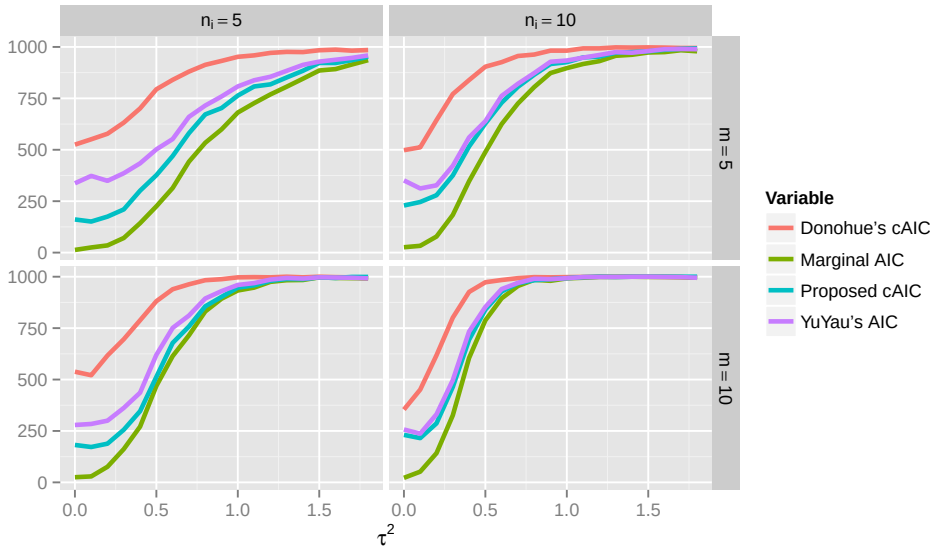

FIG 1. *Results for the random intercept model with exponentially distributed responses. The y-axis shows the number of simulation replications out of 1000 where the more complex model was favoured by the different AICs.*

The marginal AIC as defined in (2.3) requires the marginal log-likelihood, which is obtained by Laplace approximation. The results for different settings and AICs are displayed in Figure 1.

The mAIC behaves similarly to the mAIC with Gaussian responses as investigated in Greven & Kneib (2010). For small τ^2 the mAIC never chooses the model including the random effects. When the sample size increases, a preference for the smaller model remains. The other AICs select the more complex model in a higher proportion of cases. Both the proposed AIC and Yu and Yau's proposal exhibit increasing sensitivity as well as specificity as sample size increases, with the asymptotic criterion showing a stronger preference for larger models when the variance is zero or small. The estimator suggested by Donohue et al. (2011) shows a behaviour similar to the cAIC of Vaida & Blanchard (2005), observed by Greven & Kneib (2010): It chooses the model including the random effects far more often than the other criteria do. This might have been expected, since similarly to the cAIC by Vaida & Blanchard (2005), this criterion needs the variance-covariance matrices of the random effects to be known and using a plug-in estimator introduces a bias.

5.1.2. Poisson distribution

Investigating the precision and variability of the bias correction (3.6), we consider a random intercept model with Poisson distributed responses and subject

TABLE 2

Mean estimated values of four different estimators of the bias correction (2.5) and the corresponding standard deviations (indicated by σ with corresponding index) of model (5.2) for different cluster sizes (n_i), number of clusters (m) and variances of random effects (τ^2). The true bias correction BC is derived by (2.5), the estimator Ψ is directly calculated by (3.6), $\hat{\rho}_{ml}$ is the estimator proposed by Yu & Yau (2012) and $\hat{\rho}_{Don}$ is the estimator proposed by Donohue et al. (2011)

m	n_i	τ^2	BC	Ψ	$\hat{\rho}_{ml}$	$\hat{\rho}_{Don}$	σ_{Ψ}	$\sigma_{\hat{\rho}_{ml}}$	$\sigma_{\hat{\rho}_{Don}}$
5	5	0	3.07	2.99	3.61	2.47	1.28	6.74	0.81
5	5	0.3	3.98	4.12	4.54	3.35	1.43	9.39	1.18
5	5	0.6	5.17	5.12	5.44	4.51	0.99	5.83	1.06
5	10	0	2.79	2.88	3.30	2.41	1.24	6.67	0.72
5	10	0.3	5.10	4.92	5.11	4.30	1.11	2.15	1.16
5	10	0.6	5.80	5.65	5.74	5.37	0.44	1.18	0.65
10	5	0	3.63	3.62	3.91	3.04	2.15	8.01	1.67
10	5	0.3	6.35	6.39	6.50	5.52	2.36	5.76	2.30
10	5	0.6	8.87	8.87	9.22	8.45	1.20	1.77	1.45
10	10	0	3.17	3.42	3.94	2.85	1.93	7.41	1.39
10	10	0.3	8.47	8.79	8.89	8.24	1.28	1.24	1.55
10	10	0.6	10.26	10.21	10.33	10.09	0.41	0.43	0.52

specific random intercept, $y_{ij}|u_i \sim \mathcal{P}(\lambda_{ij})$. A logarithmic link function is used

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 x_j + u_i;\ i = 1, \ldots, m;\ j = 1, \ldots, n_i, \tag{5.2}$$

where $u_i \sim \mathcal{N}(0, \tau^2)$, $\beta_0 = 0.1$, $\beta_1 = 0.2$ and $x_j = j$. Different numbers of clusters, cluster sizes and random effect variances are considered: $m = 5, 10$, $n_i = 5, 10$ for $i = 1, \ldots, m$ and the random effect variances are $\tau^2 = 0,\ 0.3,\ 0.6$, respectively. The differing values of τ^2, compared to the model with exponentially distributed responses, are chosen due to the changed signal-to-noise ratio. We generate 1,000 data sets for each setting and calculate the mean and the standard deviations of the different bias corrections. The true bias correction is derived the same way as for the exponential responses.

The results are shown in Table 2. The proposed estimator Ψ combines high precision with low variance. Compared to the estimates with exponentially distributed responses, $\hat{\rho}_{ml}$ performs well although it shows a tendency towards overestimation and has high variances especially for a larger number of small clusters. The estimator $\hat{\rho}_{Don}$ underestimates the true bias correction as it did in the previous setting.

As for the simulation study with exponentially distributed responses, we also assess the model choice behaviour of bias correction (3.6). The settings are the same as in model (5.2) except the random effects variance, that is $\tau^2 = 0, \ldots, 0.8$. Then 1000 data sets for each setting are generated. Two models are fit to the data, one model containing a random intercept ($\tau^2 \geq 0$) and another model without random effects ($\tau^2 = 0$). The frequency of selecting the more complex model, including the random effects is computed for different AICs. Just as for the exponential responses, PQL was used as model fitting method. The different proposed AICs are the same as in the exponential model (5.1): 1) the proposed bias correction Ψ as in (3.6); 2) the cAIC suggested by Yu & Yau (2012); 3) the

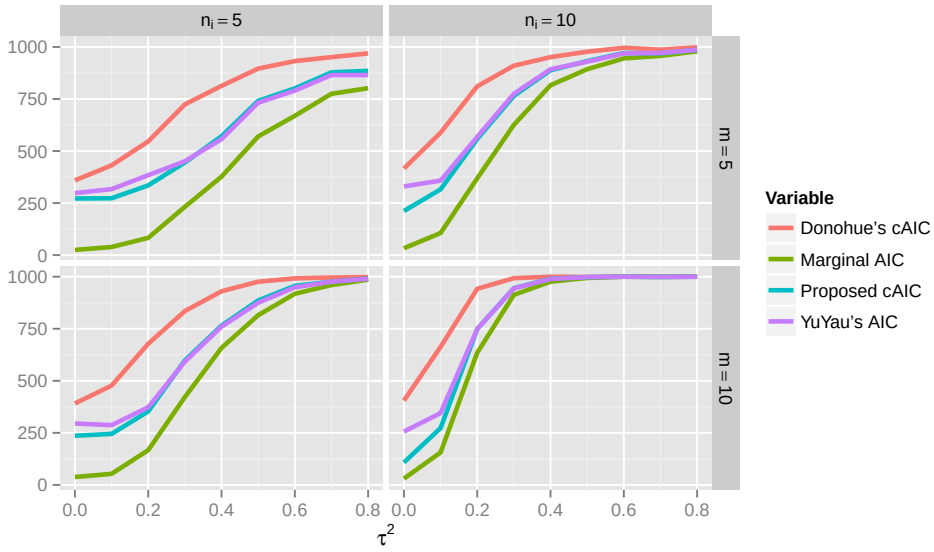

FIG 2. *Results for the random intercept model with Poisson distributed responses. The y-axis shows the number of simulation replications out of 1000 where the more complex model was favoured by the different AICs.*

cAIC proposed by Donohue et al. (2011), with the estimated variance parameter $\hat{\tau}^2$ plugged in as true τ^2; 4) the marginal AIC as defined in (2.3), which is obtained by Laplace approximation.

The results are displayed in Figure 2. They are similar to the results observed for exponential responses. The marginal AIC chooses the model including the random effects only very rarely even for random effects variances larger than zero. On the other hand, the AIC proposed by Donohue et al. (2011) chooses to include random effects very often, even if the model was simulated without random effects. The proposed criterion and Yu and Yau's asymptotic critierion behave similar, with a stronger preference for the larger model when the variance is zero or small for Yu and Yau's AIC. The asymptotically unbiased criterion proposed by Yu & Yau (2012) behaves as expected. For larger cluster sizes and increasing number of clusters the model choice behaviour gets better.

5.2. Penalized spline smoothing

It is well known that penalized spline models have a mixed model representation, see for example Wood (2006) and Ruppert et al. (2003). In this part of the simulation study, we assess the performance of different criteria for model selection in penalized spline models using the mixed model representation.

We investigate models where the mean μ is linked to a smooth function $m(\cdot)$:

$$g(\mu_i) = m(x_i), \ i = 1, \ldots, n.$$

In this setting, we choose the smooth function to be

$$m(x) = 1 + x + d\left(\frac{1}{3} - x\right)^2.$$

The x_i are chosen equidistantly from the interval $[0, 1]$. The sample sizes are $n = 25,\ 50,\ 75,\ 100$.

The parameter d controls the nonlinearity of the function m. For increasing d the nonlinearity increases and a higher signal-to-noise ratio is obtained. For $d = 0$ the function $m(\cdot)$ is linear.

The smooth function is estimated by a penalized spline

$$\widehat{m}(x) = \sum_{j=1}^{J} b_j(x)\beta_j$$

with associated smoothness penalty $\lambda\beta^t S\beta$, where S is a positive semi-definite matrix and λ is a smoothing parameter, which is estimated via the mixed model representation. The mixed model is fitted by PQL, see Breslow & Clayton (1993) and Wood (2006). In the mixed model framework, the smoothing parameter λ is associated with the inverse random effects variance parameter $1/\tau^2$. The key idea of the mixed model representation is to separate β into a penalized and an unpenalized part, which are estimated as fixed and random effects, respectively. We choose the basis functions $b_j(x)$ from the B-Spline basis with 10 inner knots, see Eilers & Marx (1996). The penalty matrix S is a second-order difference penalty matrix. In this setting the null space of S is two-dimensional, corresponding to the coefficients describing a linear function that remains unpenalized by the penalty matrix S.

5.2.1. Exponential distribution

The model for exponentially distributed responses $y_i \sim \mathcal{E}(\mu_i)$, with logarithmic link function, is

$$\log(\mu_i) = 1 + x_i + d\left(\frac{1}{3} - x_i\right)^2,\ i = 1, \ldots, n. \tag{5.3}$$

For nonlinearity parameters $d = 0, 0.5, 1$ the averaged estimated bias corrections and corresponding standard deviations are derived from 1000 data sets simulated from model (5.3).

The results in Table 3 indicate that the bias correction Ψ in (3.3) and BC in (2.5) have the same expected value, as was shown analytically in Corollary (3.1). The high variance $\sigma_{\hat{\rho}_{ml}}$ of the estimator proposed by Yu & Yau (2012) is due to outliers that occur, caused by numerical instability. The estimator $\hat{\rho}_{Don}$ does not change a lot for differing levels of nonlinearity and underestimates the bias correction term.

The model choice behaviour of the same criteria as in (5.1) is assessed in the same way as in the random intercept model. For each setting and each value of nonlinearity $d = 0,\ 0.25,\ 0.5,\ 0.75,\ 1,\ 1.5,\ 2,\ 2.5,\ 3,\ 3.5$ and 4, 1000 datasets

TABLE 3

Mean estimated values of four different estimators of the bias correction (2.5) and the corresponding standard deviations (indicated by σ with corresponding index) of model (5.3) for different sample sizes n and different degrees of nonlinearity d. The estimator Ψ is directly calculated by (3.3), BC is derived by (2.5), $\hat{\rho}_{ml}$ is the estimator proposed by Yu & Yau (2012) and $\hat{\rho}_{Don}$ is the estimator proposed by Donohue et al. (2011)

n	d	BC	Ψ	$\hat{\rho}_{ml}$	$\hat{\rho}_{Don}$	σ_{Ψ}	$\sigma_{\hat{\rho}_{ml}}$	$\sigma_{\hat{\rho}_{Don}}$
25	0	3.00	2.97	2.76	2.21	1.18	4.41	0.44
25	0.5	3.02	3.15	2.83	2.21	1.25	2.81	0.41
25	1	3.30	3.28	3.02	2.31	1.33	2.06	0.50
50	0	2.68	2.66	2.76	2.16	0.97	7.95	0.38
50	0.5	2.77	2.80	3.13	2.21	1.04	8.35	0.42
50	1	3.09	3.11	3.39	2.34	1.05	5.78	0.48
75	0	2.55	2.63	2.81	2.14	0.84	7.23	0.32
75	0.5	2.77	2.80	2.90	2.21	0.98	4.57	0.39
75	1	3.09	3.17	3.09	2.40	1.00	4.87	0.48
100	0	2.49	2.62	2.87	2.14	0.87	7.95	0.34
100	0.5	2.68	2.80	2.65	2.21	0.89	13.46	0.39
100	1	3.25	3.29	3.55	2.49	0.99	8.83	0.51

are generated, and a linear and a nonlinear model are fitted to the data. The frequency of selecting the more complex, nonlinear model for each criterion is computed.

Figure 3 shows the results. The marginal AIC behaves as expected and chooses the nonlinear model only very rarely. The proposed cAIC based on the

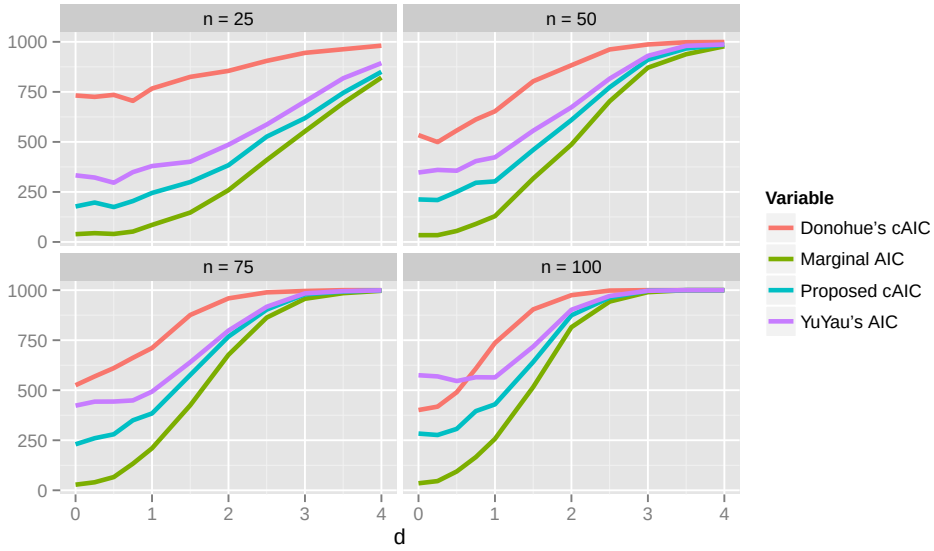

FIG 3. *Results for the spline smoothing model with exponentially distributed responses. The y-axis shows the number of simulation replications where out of 1000 the more complex model was favoured by the different AICs.*

TABLE 4

Mean estimated values of four different estimators of the bias correction (2.5) and the corresponding standard deviations (indicated by σ with corresponding index) of model (5.3) with Poisson distributed responses for different sample sizes n and different degrees of nonlinearity d. The estimator Ψ is directly calculated by (3.6), BC is derived by (2.5), $\hat{\rho}_{ml}$ is the estimator proposed by Yu & Yau (2012) and $\hat{\rho}_{Don}$ is the estimator proposed by Donohue et al. (2011)

n	d	BC	Ψ	$\hat{\rho}_{ml}$	$\hat{\rho}_{Don}$	σ_{Ψ}	$\sigma_{\hat{\rho}_{ml}}$	$\sigma_{\hat{\rho}_{Don}}$
25	0	2.47	2.61	2.78	2.20	1.07	1.88	0.44
25	0.8	3.11	3.33	3.31	2.61	1.01	3.90	0.59
25	1.6	4.18	3.93	3.77	3.33	0.49	1.62	0.45
50	0	2.29	2.66	3.09	2.18	1.21	7.80	0.37
50	0.8	3.55	3.53	3.30	2.77	0.98	3.01	0.55
50	1.6	4.06	3.99	3.86	3.63	0.63	0.36	0.30
75	0	2.25	2.48	2.87	2.14	1.05	5.01	0.34
75	0.8	3.95	3.63	3.52	2.94	0.65	3.33	0.49
75	1.6	4.62	4.05	3.96	3.77	0.31	0.29	0.24
100	0	2.37	2.45	2.87	2.13	0.90	8.60	0.32
100	0.8	3.78	3.68	3.55	3.05	0.71	1.51	0.47
100	1.6	3.73	4.13	4.06	3.90	0.30	0.27	0.20

bias correction (3.3) shows similar behaviour to the other settings. For increasing sample size, Yu & Yau (2012) show an unexpected behaviour. The cAIC by Yu & Yau (2012) selects the nonlinear model with a proportion increasing with sample size, even for zero or small variances τ^2, and for the largest sample size more often than the cAIC proposed by Donohue et al. (2011). Since this behaviour seems to contradict the findings of Yu & Yau (2012), a short discussion is given in 5.3.

5.2.2. Poisson distribution

For Poisson distributed responses $y_i \sim \mathcal{P}(\mu_i)$, model (5.3) stays the same but, due to a different signal-to-noise ratio, we choose a different sequence of nonlinearity parameters. In order to compare the precision and variability of the different bias corrections, we choose the nonlinearity parameter $d = 0, 0.8, 1.6$. For each level of nonlinearity and for the sample sizes $n = 25, 50, 75, 100$ the estimated bias corrections are listed in Table 4. The results show, that the proposed estimator Ψ is close to the bias correction BC derived by 30,000 times reestimating model (5.3) with Poisson distributed responses and calculating 2.5. The BC bias correction is not applicable in practice since the true unknown mean μ has to be known for its calculation. The high variance of the estimator $\hat{\rho}_{ml}$, proposed by Yu & Yau (2012) indicates some very large values which seem to be due to numerical instabilities.

The selection frequencies are derived for nonlinearity levels $d = 0, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6$. They are shown in Figure 4. They behave similar to the ones observed for the smoothing spline model with exponentially distributed responses. The unexpected behaviour of the cAICs proposed by Yu & Yau (2012)

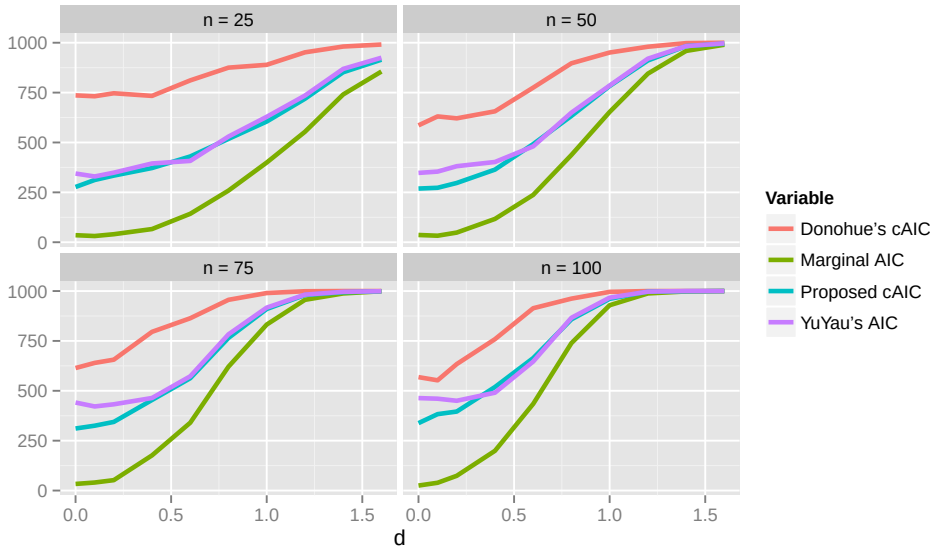

FIG 4. *Results for the spline smoothing model with Poisson distributed responses. The y-axis shows the number of simulation replications out of 1000 where the more complex model was favoured by the different AICs.*

and Donohue et al. (2011) are not as pronounced as for exponentially distributed responses. Nevertheless the bias correction of Yu & Yau (2012) is occasionally smaller than the one proposed by Donohue et al. (2011) in this setting as well.

5.3. General remarks

The cAIC proposed by Yu & Yau (2012) is used here as an ad-hoc criterion since it is one of the few available benchmarks for model selection in generalized linear mixed models. The criterion was derived for ML estimation of the variance parameters based on McGilchrist (1994), while our models were fitted using the REML based PQL method proposed by Breslow & Clayton (1993). Despite the difference between REML and ML, the two approaches are similar to each other in maximizing the joint likelihood of y and u as mentioned by McGilchrist (1994). However the main objection to the application of the cAIC proposed by Yu & Yau (2012) may be that the models (5.3) and (5.1) have a non-canonical link although the criterion of Yu & Yau (2012) requires canonical link functions.

Nevertheless the results of our simulation study do not reflect the findings of Yu & Yau (2012), even for Poisson distributed responses with canonical link, since in their simulation study the proposed cAIC can distinguish between $\tau^2 = 0$ and $\tau^2 > 0$ very well, i. e. the proportion of selecting a model with $\tau^2 > 0$, although $\tau^2 = 0$ is the true model, is zero, see Figure 1, p. 637 in Yu & Yau (2012). In our simulation, on the other hand, in at least a quarter of the

cases the more complex model ($\tau^2 > 0$) was chosen, independent of the specific settings.

Furthermore in our simulations the bias correction of Yu & Yau (2012) sometimes was smaller than the bias correction proposed by Donohue et al. (2011). This contradicts Remark 3 in Yu & Yau (2012) that says, that their bias correction is equal to the one proposed by Donohue et al. (2011) plus the trace of a positive semi-definite matrix. However in our simulation the matrix which, following Remark 3 in Yu & Yau (2012), is positive semi-definite sometimes has negative eigenvalues. This seems to be due to a boundary issue. When deriving the criterion, the derivative with respect to τ^2 needs to be calculated when $\hat{\tau}^2$ lies on the boundary of the parameter space. In these cases the trace of the matrix is sometimes negative.

The implementation of the cAIC by Yu & Yau (2012) was adapted from the MATLAB code the authors provided, but simulations were carried out in statistical programming language R. The code of the simulation study can be found in the supplementary material (Saefken et al. (2014)).

A disadvantage regarding the proposed estimator (3.3) when using numeric integration is, that for each datum the integral needs to be calculated. Therefore if for one i in (3.3) the integral can not be calculated the bias correction can not be obtained. This may be a problem particularly in large data sets and for instance, if there is collinearity in the data.

The implementation of the proposed method to derive (3.3) based on numerical integration takes 330 s for model (5.1) with random-effects variance 1 and five clusters with five observations each on a 2.80-GHz personal computer. The computational cost depends on how precise the numerical integration is and on the size of the data set.

For data from model (5.2) with random-effects variance 1 and five clusters with five observations each it takes about 3 s to compute (3.6) on a 2.80-GHz personal computer. This leave-one-out implementation is increasingly time consuming for larger data sets and less time consuming if there are many zeros in the observed responses.

6. Example: Modelling tree growth with water availability

Tree growth is of high economic importance as it determines the amount of available timber per time. As the trend is turning to a more sustainable silviculture, it becomes even more important to understand the underlying processes under close to natural conditions.

In this case study, we show how the proposed estimator of the Kullback-Leibler distance for exponential responses influences the selection of models for tree growth. The study is based on a sub-sample of 2655 trees, from a 28.5 ha large area that is located in the core zone of the Hainich National Park, Thuringia, Germany. The National Park is part of Germany's largest continuous broad-leaved forest. To estimate tree growth, in 1999 and 2007 for each tree within the study area the Diameter at Breast Height (DBH), i.e. at about

1.30 m, was mapped, see Butler-Manning (2008). The difference in diameter is the dependent variable growth. We only consider beech, which accounted for 90 % of the recorded trees. We included only trees with 10–30 cm DBH, because they can be reasonably assumed to have completed the phase of highest mortality due to competition (self-thinning), without reproducing yet themselves. Furthermore, we excluded trees for which no positive growth was recorded as these measurements seem to be erroneous.

Growth performance is highly influenced by competition for light. Thus, we assumed that neighbours that potentially overshadow the individuals are crucial for predicting growth. Neighbour-processes are included as KRAFT-class (k_i), nearest- and second nearest-neighbour distances ($nnd1_i$ and $nnd2_i$).

Water availability is a good proxy for abiotic resource availability on rich soils, because water availability, apart from light, mainly limits tree-growth, influencing the predominance of beech. To estimate spatial variation in water availability due to soil properties, we use the soil depth (sd_i) as covariate. A second available covariate, the Topographic Wetness Index (twi_i), is calculated from a Digital Elevation Model and measures water availability determined by topography, see Boehner et al. (2006).

Our aim is to find a model that best describes the tree growth with the help of the given covariates. Hence we choose the model with the lowest estimated Kullback-Leibler distance from a set of candidate models. We concentrate on the selection of linear versus nonlinear modelling of the continuous covariates. This corresponds to the selection of random effects in the mixed model framework. We model the DBH difference using an exponential distribution, as using a gamma distribution resulted in a dispersion parameter estimate of 0.98 for model 6.1 that is very close to one.

6.1. Univariate smooth function

In order to investigate the model choice behaviour of the mAIC and the proposed AIC with bias correction (3.3) in a simple model, we consider a univariate smoothing example, based on the tree growth data. We estimate the effect of soil depth on the tree growth and include the KRAFT-class to account for differing growth potentials due to light availability. For the mean of the tree growth μ, we obtain the following model:

$$\log(\mu_i) = \beta_0 + \beta_1 k_i + m(sd_i), \; i = 1, \ldots, 2655, \tag{6.1}$$

where $\mu_i = E(y_i)$ and y_i is the difference in DBH measurements between 2007 and 1999.

We distinguish between a linear model (M1) in which the function $m(\cdot)$ is a linear function and a semiparametric model (M2) with nonlinear function $m(\cdot)$. The semiparametric model is fit by PQL. Both estimated functions are plotted in Figure 5. The mAIC for the linear model (M1) is 6258 and for the semiparametric model (M2) the mAIC is 6276. The conditional AIC based on (3.3) for the linear model (M1) is 6257 and for the semiparametric model (M2) it is 6235. Therefore the mAIC chooses the model (M1) with $m(\cdot)$ as linear

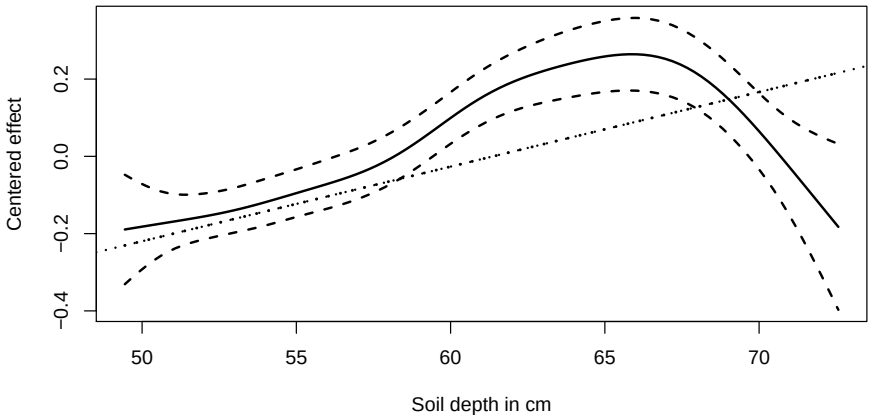

FIG 5. *The linear effect (dotted line) and the nonlinear effect (solid line) with confidence interval (dashed lines) of the soil depth on the tree growth. The pointwise confidence intervals are defined using twice the standard deviation of the estimator.*

function and the proposed conditional AIC chooses the model (M2) with $m(\cdot)$ as nonlinear function.

The model captures the positive effect of increasing soil depth for water availability. This effect levels off in very deep soils when fine root density is very low. The negative trend in very deep soils is a joint effect of soil depth and change of grain size to silt perceived as dry soils.

6.2. Generalized additive model

In a more sophisticated approach, we consider a model incorporating possibly nonlinear effects of three covariates and one linear effect of the KRAFT-class k. Accordingly we extend model (6.1) to a generalized additive model, see Wood (2006):

$$\log(\mu_i) = \beta_0 + \beta_1 k_i + m_1(sd_i) + m_2(twi_i) + m_3(nnd1_i) \tag{6.2}$$

or

$$\log(\mu_i) = \beta_0 + \beta_1 k_i + m_1(sd_i) + m_2(twi_i) + m_3(nnd2_i), \tag{6.3}$$

for $i = 1, \ldots, 2655$, depending on whether the first- or second nearest-neighbour distance is included. We only consider one of the two nearest-neighbour distances, since the two variables are collinear. Thus we use the AICs to decide which of the nearest-neighbour distances to include into the model.

The functions $m_1, \ldots, m_3$ may either be linear or nonlinear functions. In the model selection process, we choose between the two possibilities for each of the three functions in the two models. In consequence we can choose from a set of 16 candidate models. We expect all covariates to have an effect on growth and therefore do not include models into the model selection process

TABLE 5

Estimated Kullback-Leibler distance for 16 models fitted to the tree growth data. The first four columns indicate if the effects of the covariates are modelled by linear (−) or nonlinear (∼) functions, corresponding to the absence and presence of random effects. Two different estimators of the Kullback-Leibler distance are listed in the table: The AIC based on the bias correction 3.3 (cAIC) and the AIC proposed by Donohue et al. (2011) (dAIC)

$m_1\,(sd)$	$m_2\,(twi)$	$m_3\,(nnd1)$	$m_3\,(nnd2)$	*cAIC*	*dAIC*
∼	∼	∼		6189.461	6185.855
∼	∼	−		6191.001	6189.359
∼	−	∼		6200.224	6197.756
∼	−	−		6201.596	6201.107
−	∼	∼		6199.270	6197.488
−	∼	−		6203.715	6202.995
−	−	∼		6213.788	6212.881
−	−	−		6218.176	6218.333
∼	∼		∼	6180.980	6177.974
∼	∼		−	6189.084	6187.287
∼	−		∼	6190.039	6188.696
∼	−		−	6198.649	6198.123
−	∼		∼	6190.120	6188.629
−	∼		−	6201.498	6200.599
−	−		∼	6202.958	6202.775
−	−		−	6214.723	6214.844

that completely omit one of the covariates, except $nnd1$ and $nnd2$ respectively. All possible models and the corresponding AIC values can be found in Table 5.

The model selection process is based on two criteria, the proposed conditional AIC with associated bias correction (3.3) and the conditional AIC proposed by Donohue et al. (2011). The marginal AIC is omitted because we can not extract the design matrices $Z_i,\ i = 1, 2, 3$ corresponding to the random effects parametrization of the smoothing splines, that are needed to derive the Laplace approximation. This problem does not occur in univariate smoothing models since there is no need to split up the design matrix Z corresponding to the random effect. The conditional AIC proposed by Yu & Yau (2012) could not be calculated due to the need for inverting matrices that are singular.

The two criteria both choose the model including the second nearest-neighbour distance with all three effects modelled as nonlinear functions. Comparing each specific model whether to include the second or the first nearest-neighbour distance, both criteria in each case favour the model with the second nearest-neighbour distance. For all models, except the two models only containing linear effects, the proposed conditional AIC is larger and therefore penalizes more than the conditional AIC proposed by Donohue et al. (2011). This confirms the behaviour observed in the simulation study, that the criterion by Donohue et al. (2011) underestimates the bias correction.

This example additionally highlights that the various criteria for the estimation of the Kullback-Leibler distance can lead to different model choices. For instance, in the comparison of the two models

$$\log(\mu_i) = \beta_0 + \beta_1 k_i + m_1(sd_i) + \beta_2 twi_i + m_3(nnd1_i) \tag{6.4}$$

and

$$\log(\mu_i) = \beta_0 + \beta_1 k_i + \beta_2 sd_i + m_2(twi_i) + m_3(nnd2_i), \tag{6.5}$$

our proposed estimator chooses the first model (6.4), while the estimator proposed by Donohue et al. (2011) chooses the latter (6.5).

7. Discussion

The proposed class of estimators of the conditional Akaike information is unbiased for finite samples, does not require a particular estimation method for GLMMs and does not assume known variance-covariance matrices for the random effects. Therefore it improves on available asymptotic results in terms of bias, variability as well as model selection. The formulation of penalized regression as mixed models allows the model choice techniques considered in this paper to be used for penalized regression models as well. All of these characteristics make these conditional Akaike information criteria appealing to use. For the theoretical derivation, the working model has to be correctly specified in equations (3.1) and (3.4). The behaviour of the proposed methods for misspecified models needs to be investigated in future work.

For other exponential family distributions than the ones discussed above, like the gamma and the binomial distribution, the Stein type formulas do not seem to yield criteria that are computable from observable quantities. It may also be of interest to investigate extensions to distributions beyond the exponential family using the generalized SURE formula of Shen & Huang (2006). The behaviour of other information based criteria like the Bayesian information criterion, BIC, for the selection of random effects in GLMMs needs further investigation.

Appendix: Technical details

Here we give outlines of proofs of the main results. The proofs of the essential Theorem 3.1 and Theorem 3.2 can be done by integration by parts; see Hudson (1978) with small modifications.

Proof of Proposition 2.1. The conditional log-likelihood is then as

$$\log f(y|\beta, u) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi),$$

where ϕ is the known scaling parameter. Then the conditional Akaike Information becomes

$$\begin{aligned} cAI &= -2\mathbb{E}_{y,u}\mathbb{E}_{z|u}\left[\log f(z|\hat{\beta}(y), \hat{u}(y))\right] \\ &= -2\mathbb{E}_{y,u}\left[\sum_{i=1}^{n} \frac{\mu_i\hat{\theta}_i - b(\hat{\theta}_i)}{\phi} + \mathbb{E}_{z|u}c(z_i, \phi)\right]. \end{aligned}$$

Now the bias correction can be calculated by

$$\begin{aligned}
2\Psi &= cAI - \mathbb{E}_{y,u}\left[-2\log f(y|\hat{\beta}(y),\hat{u}(y))\right] \\
&= 2\mathbb{E}_{y,u}\left[\sum_{i=1}^{n}\frac{y_i\hat{\theta}_i(y)-b(\hat{\theta}_i(y))}{\phi}+c(y_i,\phi)\right] \\
&- 2\mathbb{E}_{y,u}\left[\sum_{i=1}^{n}\frac{\mu_i\hat{\theta}_i(y)-b(\hat{\theta}_i(y))}{\phi}+\mathbb{E}_{z,u}c(z_i,\phi)\right] \\
&= 2\mathbb{E}_{y,u}\left[\sum_{i=1}^{n}\frac{y_i-\mu_i}{\phi}\hat{\theta}_i(y)\right]+2\mathbb{E}_{y,u}\left[\sum_{i=1}^{n}c(y_i,\phi)\right]-2\mathbb{E}_{z,u}\left[\sum_{i=1}^{n}c(z_i,\phi)\right] \\
&= 2\sum_{i=1}^{n}\mathbb{E}_{y,u}\left[\frac{y_i-\mu_i}{\phi}\hat{\theta}_i(y)\right]. \qquad \square
\end{aligned}$$

Proof of Corollary 3.1. Let $y_i|u \sim \mathcal{E}(\frac{1}{\mu_i})$, then we can rewrite the bias correction (2.5) with the help of equation (3.2):

$$\begin{aligned}
2\Psi &= 2\sum_{i=1}^{n}\mathbb{E}_{y,u}\left[\frac{y_i-\mu_i}{\phi}\hat{\theta}_i(y)\right] \\
&= 2\sum_{i=1}^{n}\mathbb{E}_{y,u}\left[(y_i-\mu_i)\hat{\theta}_i(y)\right] \\
&= 2\left[\sum_{i=1}^{n}\mathbb{E}_{y,u}\left[y_i\hat{\theta}_i(y)\right]-\sum_{i=1}^{n}\mathbb{E}_{y,u}\left[\mu_i\hat{\theta}_i(y)\right]\right] \\
&= 2\left[\sum_{i=1}^{n}\mathbb{E}_{y,u}\left[y_i\hat{\theta}_i(y)\right]-\sum_{i=1}^{n}\mathbb{E}_{y,u}\left[\mu_i\hat{\theta}_i(y_{-i},y_i)\right]\right] \\
&= 2\left[\sum_{i=1}^{n}\mathbb{E}_{y,u}\left[y_i\hat{\theta}_i(y)\right]-\sum_{i=1}^{n}\mathbb{E}_{y,u}\left[\int_0^{y_i}\hat{\theta}_i(y_{-i},x)dx\right]\right] \\
&= 2\mathbb{E}_{y,u}\left[\sum_{i=1}^{n}y_i\hat{\theta}_i(y)-\int_0^{y_i}\hat{\theta}_i(y_{-i},x)dx\right]
\end{aligned}$$

Where y_{-i} is the vector of observed responses without the $i-$th observation. $\square$

Proof of Corollary 3.2. If $y_i|u \sim \mathcal{P}(\lambda_i)$ then equation (2.5) becomes with the help of equation (3.5):

$$\begin{aligned}
2\Psi &= 2\sum_{i=1}^{n}\mathbb{E}_{y,u}\left[\frac{y_i-\mu_i}{\phi}\hat{\theta}_i(y)\right] \\
&= 2\sum_{i=1}^{n}\mathbb{E}_{y,u}\left[(y_i-\lambda_i)\hat{\theta}_i(y)\right]
\end{aligned}$$

$$
\begin{aligned}
&= 2\left[\sum_{i=1}^{n}\mathbb{E}_{y,u}\left[y_i\hat{\theta}_i(y)\right]-\sum_{i=1}^{n}\mathbb{E}_{y,u}\left[\lambda_i\hat{\theta}_i(y)\right]\right]\\
&= 2\left[\sum_{i=1}^{n}\mathbb{E}_{y,u}\left[y_i\hat{\theta}_i(y)\right]-\sum_{i=1}^{n}\mathbb{E}_{y,u}\left[\lambda_i\hat{\theta}_i(y_{-i},y_i)\right]\right]\\
&= 2\left[\sum_{i=1}^{n}\mathbb{E}_{y,u}\left[y_i\hat{\theta}_i(y)\right]-\sum_{i=1}^{n}\mathbb{E}_{y,u}\left[y_i\hat{\theta}_i(y_{-i},y_i-1)\right]\right]\\
&= 2\mathbb{E}_{y,u}\left[\sum_{i=1}^{n}y_i\left(\hat{\theta}_i(y)-\hat{\theta}_i(y_{-i},y_i-1)\right)\right]
\end{aligned}
$$

Here y_{-i} is the vector of observed responses without the $i-$th observation and y_i is the $i-$th observation with $y_i\hat{\theta}_i(y_{-i},y_i-1)=0$ if $y_i=0$ by convention. □

Acknowledgements

The research of the first, second and third author was supported by the RTG 1644 - Scaling Problems in Statistics. The fourth author was supported by Emmy Noether grant GR 3793/1-1 from the German research foundation. We thank Christian Wirth, University of Leipzig, for generously sharing the Hainich tree data and Ernst-Detlef Schulze, Max-Planck-Institute for Biogeochemistry Jena, for providing the Digital Elevation Model and Julia Braun, University of Zurich, for sharing her R-code with us. We also want to thank the associated editor and two anonymous referees who have substantially improved this paper with their remarks.

Supplementary Material

The code of the simulation study. Part 1
(doi: 10.1214/14-EJS881SUPPA; .zip).

The code of the simulation study. Part 2
(doi: 10.1214/14-EJS881SUPPB; .zip).

References

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory* 267–281. MR0483125

Boehner, J., McCloy, K. R. & Strobl, J. (2006). *SAGA – Analysis and Modelling Applications. Goettinger Geographische Abhandlungen* **115**, 130.

Breslow, N. E. & Clayton, D. G. (1993). *Approximate inference in generalized linear mixed models. Journal of the American Statistical Association* **88**, 9–25.

Butler-Manning, D. (2008). *Stand structure, gap dynamics and regeneration of a semi-natural mixed beech forest on limestone in central Europe: A case study. PhD thesis, Universität Freiburg, Forstwissenschaftliche Fakultät.*

Chen, L. H. Y. (1975). *Poisson approximation for dependent trials. The Annals of Probability* **3**, 534–545. MR0428387

Donohue, M. C., Overholser, R., Xu, R. & Vaida, F. (2010). *Conditional Akaike information under generalized linear and proportional hazards mixed models. Biometrika* **98**, 685–700. MR2836414

Eilers, P. H. C. & Marx, B. D. (1996). *Flexible smoothing with B-splines and penalties. Statistical Science* **2**, 89–121. MR1435485

Greven, S. & Kneib, T. (2010). *On the behaviour of marginal and conditional AIC in linear mixed models. Biometrika* **97**, 773–789. MR2746151

Hodges, J. S. & Sargent, D. J. (2001). *Counting degrees of freedom in hierarchical and other richly-parameterised models. Biometrika* **88**, 367–379. MR1844837

Hudson, H. M. (1978). *A natural identity for exponential families with applications in multiparameter estimation. The Annals of Statistics* **6**, 473–484. MR0467991

Hurvich, C. M. & Sargent, C.-L. (1989). *Regression and time series model selection in small samples. Biometrika* **76**, 297–307. MR1016020

Lian, H. (2011). *A note on conditional Akaike information for Poisson regression with random effects. Electronic Journal of Statistics* **6**, 1–9. MR2879670

Liang, H., Wu, H. & Zou, G. (2008). *A note on conditional AIC for linear mixed-effects models. Biometrika* **95**, 773–778. MR2443190

McGilchrist, C. A. (1994). *Estimation in generalized mixed models. Journal of the Royal Statistical Society. Series B (Methodological)* **56**, 61–69. MR1257795

Nelder, J. A. & Wedderburn, R. W. M. (1972). *Generalized linear models. Journal of the Royal Statistical Society, Series A, General* **135**, 370–384.

Ruppert, D., Wand, M. & Carroll, R. (2003). *Semiparametric Regression. Cambridge University Press: New York, 2003* MR1998720

Saefken, B., Kneib, T., van Waveren, C.-S. & Greven, S. (2014). Supplements to "A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models." DOIs: 10.1214/14-EJS881SUPPA, 10.1214/14-EJS881SUPPB.

Shen, X. & Huang, H.-C. (2006). *Optimal model assessment, selection, and combination. Journal of the American Statistical Association* **474**, 554–568. MR2281243

Stein, C. (1972). *A Bound for the Error in the Normal Approximation to the Distribution of a Sum of Dependent Random Variables. Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* **2**, 586–602. MR0402873

Vaida, F. & Blanchard, S. (2005). *Conditional Akaike information for mixed-effects models. Biometrika* **92**, 351–370. MR2201364

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC* MR2206355

Yu, D. & Yau, K. K. W. (2012). *Conditional Akaike information criterion for generalized linear mixed models. Computational Statistics & Data Analysis* **56**, 629–644. MR2853760