Received: 24 September 2018 | Revised: 15 November 2019 | Accepted: 18 November 2019

DOI: 10.1111/sjos.12437

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Conditional covariance penalties for mixed models

Benjamin Säfken[1,2] | **Thomas Kneib**[2]

[1]Department of Statistics, Ludwig Maximilian University

[2]Chair of Statistics, Georg August University

Correspondence
Benjamin Säfken, Chair of Statistics, Georg August University, Humboldtallee 3, 37073 Göttingen, Germany.
Email: bsaefke@uni-goettingen.de

Funding information
German Research Association (DFG) Research Training Group Scaling Problems in Statistics, (RTG 1644)

Abstract

The prediction error for mixed models can have a conditional or a marginal perspective depending on the research focus. We introduce a novel conditional version of the optimism theorem for mixed models linking the conditional prediction error to covariance penalties for mixed models. Different possibilities for estimating these conditional covariance penalties are introduced. These are bootstrap methods, cross-validation, and a direct approach called *Steinian*. The behavior of the different estimation techniques is assessed in a simulation study for the binomial-, the t-, and the gamma distribution and for different kinds of prediction error. Furthermore, the impact of the estimation techniques on the prediction error is discussed based on an application to undernutrition in Zambia.

KEYWORDS

additive models, conditional Akaike information criterion, covariance penalties, mixed models, optimism, prediction error

1 | INTRODUCTION

We discuss general methods for estimating the conditional prediction error in mixed models. Mixed models (Laird & Ware, 1982) are a common statistical tool for analyzing clustered or longitudinal data and any kind of hierarchical modeling. Modern implementations for estimation (Bates, Mächler, Bolker, & Walker, 2015) allow for fast and reliable inference in these kind of

models. Moreover, the mixed model framework can be employed for the estimation of a wide class of statistical models such as smoothing splines and (generalized) additive models, see Anderssen and Bloomfield (1974) and Wahba (1985) for an early reference. These models are very popular and widely used (see, e.g., Fahrmeir, Kneib, Lang, & Marx, 2013; Ruppert, Wand, & Carroll, 2003; Wood, 2017).

The framework of mixed models with normal random effects as any model with quadratic penalties can be extended to distributions beyond the exponential family, for example, the beta or scaled-t distribution (Shun & McCullagh, 1995). For a discussion on a general framework for inference in such models see Wood, Pya, and Säfken (2016).

The estimation of prediction error is not only interesting when using a statistical model for predicting future values but is also of major interest for model choice and variable selection. Efron (2004) distinguishes between methods based on Stein (1972), cross-validation, parametric, and nonparametric bootstrap.

For mixed models, Müller, Scealy, and Welsh (2013) give a detailed overview of existing methods for model selection. On the one hand, methods for deriving tests, especially likelihood ratio tests, are getting some attention (see, e.g., Crainiceanu & Ruppert, 2004; Greven, Crainiceanu, Küchenhoff, & Peters 2008; Self & Liang 1987). On the other hand, attention focuses on the Akaike information criterion (Akaike, 1973). Vaida and Blanchard (2005) propose to use the marginal and the conditional akaike information criterion (AIC) depending on the underlying research question. Liang, Wu, and Zou (2008) use the aforementioned method from Stein (1972) to derive the conditional AIC. While Greven and Kneib (2010) show that the marginal AIC is biased and give an analytical formula on how to calculate the conditional AIC for models with Gaussian responses. These results are not directly applicable for generalized mixed models (Saefken, Kneib, van Waveren, & Greven, 2014). However Yu, Zhang, and Yau (2018) propose a conditional generalized information criterion based on the conditional Kullback–Leibler divergence for possibly misspecified data modeled by a generalized linear mixed model. In a recent contribution, Sakamoto (2019) introduces a bias reduction for the marginal AIC.

In terms of prediction error one may also distinguish between a marginal and a conditional perspective. This paper focuses on the conditional perspective. Conditionality here refers to the perspective of prediction, meaning that the prediction is conditioned on the random effects, that is, future data are assumed to share the same random effects as the observed data. Different measures to assess prediction error, the so-called q-class of prediction errors, are presented and their representation as conditional covariance penalties with the help of the so-called optimism theorem (Efron, 2004) are discussed. A conditional version of the optimism theorem for mixed models is then presented.

These conditional covariance penalties can be estimated with methods that are broadly applicable such as bootstrap methods and cross-validation. For certain distributions, however, it is possible to derive criteria that sometimes are preferable in terms of accuracy and computational burden. This is demonstrated for the scaled t-, the Bernoulli-, and the gamma-distribution. An analytical formula for the representation of covariance penalties is derived, plug-in estimators are investigated, and a link to bootstrap-based methods is ascertained.

A special focus in the simulation study is on the model choice behavior of these estimation techniques and the different error functions, especially when comparing a complex model incorporating random effects with simpler models, that exclude the random effects, as in the simulation study in Section 4.

Furthermore, the use of the methods for practical statistical modeling is demonstrated in an application on predicting stunting in Zambia.

2 | CONDITIONAL PREDICTION ERROR IN MIXED MODELS

Consider a probability mechanism for data $y_1, \dots, y_n$, with conditional density or probability function

$$f(y_i|\mu_i, \phi), \tag{1}$$

with mean μ_i and scale parameter ϕ. The mean is linked to a predictor by a component-wise response function $h(\cdot)$, that is,

$$\mu_i = h\left(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{u}_i'\boldsymbol{\gamma}\right),$$

and the scale parameter ϕ is constant for all y_i, $i = 1, \dots, n$. The predictor is split up into fixed parameters $\boldsymbol{\beta}$ and random parameters $\boldsymbol{\gamma}$ with corresponding covariate vectors $\boldsymbol{x}_i$ and $\boldsymbol{u}_i$. We do not depend on a certain distribution for the random effects $\boldsymbol{\gamma}$. A common choice, however, is to assume normality

$$\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}),$$

with positive semi-definite covariance matrix $\boldsymbol{D}$. The covariance matrix depends on a parameter, $\boldsymbol{\tau}^2$. The parameter may be multivariate, that is, vector-valued. In the simulation study, we focus on $\boldsymbol{D} = \tau^2\boldsymbol{I}$. The normality assumption allows us to extend the type of model considered in this framework from longitudinal and cluster models to penalized spline smoothing, surface estimation, spatial models, or functional data analysis, see for instance Wood et al. (2016) for an overview of possible models. However, other choices of the random effects distribution allow for even further extension of the models under consideration.

2.1 | q-class of error measures

The error of real valued outcomes y and given prediction $\widehat{\mu}$ can be measured in different ways. A wide class of error measures, called q-class of error measures, can be constructed with the help of a concave function $q(\cdot)$ by

$$Q(y, \widehat{\mu}) = q(\widehat{\mu}) + q'(\widehat{\mu})\left(y - \widehat{\mu}\right) - q(y). \tag{2}$$

This q-class of error measures was introduced by Efron (1986). In the following, we give some examples for common choices of error measures.

2.1.1 | Example I:

The squared error is given via the concave function

$$q(\mu) = \mu(1 - \mu) \text{ or } q(\mu) = -\mu^2,$$

resulting in the corresponding error measure, that is,

$$Q(y, \widehat{\mu}) = (y - \widehat{\mu})^2.$$

2.1.2 | Example II:

For binary data, a natural and common choice is the counting error

$$Q(y, \widehat{\mu}) = \begin{cases} 0, & \text{if } y = 0 \text{ and } \widehat{\mu} < \frac{1}{2} \text{ or } y = 1 \text{ and } \widehat{\mu} > \frac{1}{2} \\ 1, & \text{else,} \end{cases}$$

which results from the triangular function on the unit interval

$$q(\mu) = \min(\mu, 1 - \mu).$$

Another choice that is applicable for a large class of probability distributions is the deviance function. For exponential family distributions with natural parameter ϑ, mean $\mu = b^{'}(\vartheta)$, scale parameter ϕ and with the function $b(\cdot)$, the logarithm of the conditional density of y_i is given by

$$\log\left(f(y_i|\vartheta_i, \phi)\right) = \frac{y_i\vartheta_i - b(\vartheta_i)}{\phi} + c(y_i, \phi). \tag{3}$$

2.1.3 | Example III:

The deviance error for exponential family distributions is defined by

$$q(\mu) = \frac{2}{\phi}\left(b(\vartheta) - y\vartheta\right),$$

and thus

$$\begin{aligned} Q(y, \widehat{\mu}) &= \frac{2}{\phi}\left(\log(f_y(y)) - \log(f_\mu(y))\right), \\ &= \frac{2}{\phi}\left(y\widehat{\vartheta}_y - b(\widehat{\vartheta}_y) - y\widehat{\vartheta} + b(\widehat{\vartheta})\right), \end{aligned} \tag{4}$$

with the log-likelihood $\log(f_\mu(y))$, saturated model $\log(f_y(y))$, and $\widehat{\vartheta}_y$ the estimated natural parameter evaluated at y. This is proportional to twice the negative relative Kullback–Leibler distance and therefore results in the Akaike information.

For the preceding part, the main parameter of interest that plays a major part in the derivation of the conditional covariance penalties is

$$\widehat{\theta} = -\frac{q'(\widehat{\mu})}{2}. \tag{5}$$

For the squared error in Example I, the main parameter of interest is $\widehat{\theta} = \widehat{\mu} - \frac{1}{2}$ and for the counting error in Example II

$$\widehat{\theta} = \begin{cases} -\frac{1}{2}, & \text{if } \widehat{\mu} < \frac{1}{2} \\ \frac{1}{2}, & \text{if } \widehat{\mu} > \frac{1}{2}. \end{cases}$$

In the case of an exponential family in Example III, the derivative of $q(\cdot)$ is twice the negative natural parameter of the exponential family, and hence the main parameter of interest is obviously the natural parameter of the exponential family, that is, $-\frac{q'(\widehat{\mu})}{2} = \widehat{\theta} = \widehat{\vartheta}$.

For data $\boldsymbol{y} = (y_1, \dots, y_n)$ with predictions $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \dots, \widehat{\mu}_n)$, the total error is defined as the sum of the component errors, that is,

$$Q\left(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}\right) = \sum_{i=1}^{n} Q(y_i, \widehat{\mu}_i). \tag{6}$$

2.2 | Conditional covariance penalties

In order to assess the true prediction error, the quantity (Equation 6) is too optimistic since the predicted mean $\widehat{\boldsymbol{\mu}}$ depends on the observed data $\boldsymbol{y}$. The obvious interest is how well the model will fit future data from the same underlying data generating process. Hence, the quantity of interest is the expected prediction error w.r.t. future data $\boldsymbol{z}$, that is, $\mathbb{E}_{\boldsymbol{z}}\left(Q\left(\boldsymbol{z}, \widehat{\boldsymbol{\mu}}\right)\right)$. If, however, the regression model under consideration contains more than one source of randomness, such as random effects, the type of prediction is not unique. In mixed models, future values may not share the same random effects as the ones that were used for fitting the model. The prediction should then be based on the marginal mean $\widehat{\boldsymbol{\mu}}_m = \mathbb{E}(\boldsymbol{y})$, corresponding to the mean of the marginal distribution of the data $\boldsymbol{y}$. On the other hand, the future values at which the prediction is targeted can hold the same random effects as the observed data. Thus, only one source of randomness is considered for the prediction. In this case, the appropriate mean is $\widehat{\boldsymbol{\mu}}_c = \mathbb{E}(\boldsymbol{y}|\boldsymbol{\gamma})$, which is known as the conditional mean and corresponds to the mean of the conditional distribution of the data $\boldsymbol{y}|\boldsymbol{\gamma}$. For instance, in a Gaussian model, the conditional and marginal means correspond to the predictors with or without the predicted random effects, $\widehat{\boldsymbol{\mu}}_m = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\mu}}_c = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{U}\widehat{\boldsymbol{\gamma}}$. For other distributions, the distinction is not as obvious. The densities of the marginal distributions are often not analytically accessible. Thus, in a mixed model, the appropriate mean that needs to be plugged into the error function depends on the focus of the prediction. If the prediction focus lies on the population and future values may have any even unobserved random effects, the marginal mean is suitable. If, on the other hand the prediction focus lies on a cluster or individual associated with a random effect, the mean of choice should be the conditional mean. While for mixed models, the appropriate mean depends on the prediction focus, in many applications of the mixed model in which the mixed model framework is a vehicle for estimation, such as penalized regression, the prediction is always assumed to share the same random effects.

In the following, we will only concentrate on the conditional prediction. Along these lines the conditional expected prediction error w.r.t. future data $\boldsymbol{z}|\boldsymbol{\gamma}$ is

$$\mathbb{E}_{\boldsymbol{z}|\boldsymbol{\gamma}}\left(Q\left(\boldsymbol{z}, \widehat{\boldsymbol{\mu}}\right)\right). \tag{7}$$

The optimism theorem, see Efron (2004), links the observed or apparent prediction error with the expected prediction error. The adaptation to the conditional optimism theorem for the ith component is straightforward:

Theorem 1. *Let $y_i|\boldsymbol{\gamma}$ be defined as in Equation (1) with random effects $\boldsymbol{\gamma}$, $\widehat{\theta}_i$ as in Equation (5) and let $z|\boldsymbol{\gamma}$ follow the same distribution as $y_i|\boldsymbol{\gamma}$. With error measure (Equation 2) we have*

$$\mathbb{E}_{y_i,\boldsymbol{\gamma}}\left(\mathbb{E}_{z|\boldsymbol{\gamma}}\left(Q\left(z,\widehat{\mu}_i\right)\right)\right) = \mathbb{E}_{y_i,\boldsymbol{\gamma}}\left(Q\left(y_i,\widehat{\mu}_i\right)\right) + 2\mathrm{cov}_{y_i,\boldsymbol{\gamma}}\left(\widehat{\theta}_i, y_i\right). \tag{8}$$

Proof. For the ith conditionally expected error component, we have

$$\mathbb{E}_{z|\boldsymbol{\gamma}}\left(Q(z,\widehat{\mu}_i)\right) = q(\widehat{\mu}_i) + q'(\widehat{\mu}_i)(\mu_i - \widehat{\mu}_i) - \mathbb{E}_{z|\boldsymbol{\gamma}}\left(q(z)\right),$$

and the observed error is

$$Q(y_i,\widehat{\mu}_i) = q(\widehat{\mu}_i) + q'(\widehat{\mu}_i)(y_i - \widehat{\mu}_i) - q(y_i).$$

Thus, the difference between observed and expected error is

$$\mathbb{E}_{z|\boldsymbol{\gamma}}\left(Q(z,\widehat{\mu}_i)\right) - Q(y_i,\widehat{\mu}_i) = q'(\widehat{\mu}_i)(\mu_i - y_i) + q(y_i) - \mathbb{E}_{z|\boldsymbol{\gamma}}\left(q(z)\right).$$

Taking expectations w.r.t. the joint distribution of $y_i, \boldsymbol{\gamma}$ gives

$$\mathbb{E}_{y_i,\boldsymbol{\gamma}}\{\mathbb{E}_{z|\boldsymbol{\gamma}}\left(Q(z,\widehat{\mu}_i)\right) - Q(y_i,\widehat{\mu}_i)\} = 2\,\mathbb{E}_{y_i,\boldsymbol{\gamma}}\widehat{\theta}_i(y_i - \mu_i) = 2\,\mathrm{cov}_{y_i,\boldsymbol{\gamma}}(\widehat{\theta}_i, y_i).$$

■

Hence, the total conditional prediction error is

$$\mathbb{E}_{\boldsymbol{y},\boldsymbol{\gamma}}\left(\mathbb{E}_{\boldsymbol{z}|\boldsymbol{\gamma}}\left(Q\left(\boldsymbol{z},\widehat{\boldsymbol{\mu}}\right)\right)\right) = \mathbb{E}_{\boldsymbol{y},\boldsymbol{\gamma}}\left(Q\left(\boldsymbol{y},\widehat{\boldsymbol{\mu}}\right)\right) + 2\sum_{i=1}^{n}\mathrm{cov}_{y_i,\boldsymbol{\gamma}}\left(\widehat{\theta}_i, y_i\right). \tag{9}$$

Notice that the conditional covariance penalty in the total conditional prediction error is additionally conditional on the observed responses excluding the ith datum. Thus, if

$$\boldsymbol{y}_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$$

indicates the data vector in which the ith component is excluded, the prediction error of the ith component is also conditioned on $\boldsymbol{y}_{-i}$. Hence, the data conditioned version or "fixed data" version of Equation (8) is

$$\mathbb{E}_{y_i,\boldsymbol{\gamma}|\boldsymbol{y}_{-i}}\left(\mathbb{E}_{z|\boldsymbol{\gamma}}\left(Q\left(z,\widehat{\mu}_i\right)\right)\right) = \mathbb{E}_{y_i,\boldsymbol{\gamma}|\boldsymbol{y}_{-i}}\left(Q\left(y_i,\widehat{\mu}_i\right)\right) + 2\mathrm{cov}_{y_i,\boldsymbol{\gamma}|\boldsymbol{y}_{-i}}\left(\widehat{\theta}_i, y_i\right). \tag{10}$$

3 | ESTIMATING CONDITIONAL COVARIANCE PENALTIES

The covariance penalties are in general not observable and therefore need to be estimated. There are several possible approaches that will be discussed here. Very general methods that can be applied to any distributional and parametric setting are the bootstrap and cross-validation that are presented in Sections 3.2 and 3.3. The latter can even be applied to nonparametric settings. Nevertheless, for certain distributions it is possible to derive estimators that are

preferable. These can be seen as generalizations of the *Steinian*-type estimators as presented in Efron (2004). Those presented in this section are not unbiased but are reasonable approximations, as shown in the subsequent simulation study. Such estimators are proposed for the gamma-, the Bernoulli-, and the scaled t-distribution. With the help of Theorem 2, it is furthermore possible to gain insight into the circumstances under which these estimators are available. Furthermore, there is an interesting connection between the *Steinian*-type estimators that are presented here and the conditional "fixed data" bootstrap, thereby the *Steinian*-type estimators appear to be large sample approximations of the conditional bootstrap estimators.

3.1 | Steinian-type estimators

There are attempts to generalize such kind of *Steinian* formulas to further distributions for mixed models, see Saefken et al. (2014), although there is up to date no generalization that leads to unbiased estimates of the conditional covariance penalties for all distributions. One such generalized *Steinian* formula for a large class of distributions is given in Shen and Huang (2006):

Theorem 2. *Let y be a continuous random variable with probability density function (Equation 1) and $\widehat{\theta} = \widehat{\theta}(y)$ a differentiable function such that $\mathbb{E}\left(\widehat{\theta}(y)(y-\mu)\right) < \infty$ with $\mu = \mathbb{E}(y)$, then*

$$\operatorname{cov}(\widehat{\theta}(y), y) = \mathbb{E}\left(\widehat{\theta}'(y)V(y,\mu)\right), \tag{11}$$

with $V(y,\mu) = \frac{1}{f(y)}\int_{-\infty}^{y}(\mu - t)f(t)dt$.

The function $V(y,\mu) = \frac{1}{f(y)}\int_{-\infty}^{y}(\mu - t)f(t)dt$ is not self-explanatory, but its expectation is the variance, that is, $\mathbb{E}\left(V(y,\mu)\right) = \operatorname{Var}(y)$.

A similar identity also holds for discrete random variables with probability function $p(\cdot)$ with support $\mathcal{S}$ and $V(y,\mu) = \frac{1}{p(y)}\sum_{t\in\mathcal{S}, t\leq y}(\mu - t)p(t)$. The derivative $\widehat{\theta}'(y)$ is replaced by $\Delta\widehat{\theta}(y) = \widehat{\theta}(y^+) - \widehat{\theta}(y)$, where y^+ is the smallest number that is larger then y.

In the following, the "variance" function $V(y,\mu)$ is explicitly stated for a number of distributions in order to give an intuition on its appearance.

3.1.1 | Example I:

For the Gaussian distribution with mean μ and variance σ^2

$$V(y,\mu) = \sigma^2.$$

3.1.2 | Example II:

For the Poisson distribution

$$V(y,\mu) = y.$$

3.1.3 | Example III:

For the gamma distribution with parameters μ and ν

$$V(y,\mu) = \frac{\mu}{\nu} y.$$

The formula (11), and the same holds for its discrete analogue, does not automatically leads to an observable quantity because in general we cannot plug in an estimate of $V(y,\mu)$, since

$$\mathbb{E}\left(\widehat{\theta}'(y)V(y,\mu)\right) \neq \mathbb{E}\left(\widehat{\theta}'(y)\right)\mathbb{E}\left(V(y,\mu)\right) = \mathbb{E}\left(\widehat{\theta}'(y)\right)\mathrm{Var}(y).$$

We therefore need to distinguish between two cases: either the "variance" function is independent of y, that is, $V(y,\mu) = V(\mu)$ or the "variance" function somehow depends on y. The first case is unproblematic since we have $\mathbb{E}\left(\widehat{\theta}'(y)V(y,\mu)\right) = \mathbb{E}\left(\widehat{\theta}'(y)\right)\mathrm{Var}(y)$ and we can plug in an estimate of $\mathrm{Var}(y)$ as is done in the bias correction for the normal distribution, see Efron (2004). However, we can make some progress in the second case for the gamma distribution or more generally for all distributions, for which

$$V(y,\mu) = y \cdot \Psi(\mu), \tag{12}$$

with $\Psi(\mu)$ only depending on μ not on y.

Theorem 3. *Let y and $\widehat{\theta} = \widehat{\theta}(y)$ be defined as in Equation (2) and $\widehat{\theta}(y)$ additionally be s-times continuously differentiable with $s \in \mathbb{N}$ and $\widehat{\theta}^{(s+1)}(y) = 0$. If Equation (12) is fulfilled the conditional covariance penalty is*

$$\mathrm{cov}(\widehat{\theta}, y) = \mathrm{Var}(y) \sum_{i=1}^{s} \mathbb{E}(\widehat{\theta}^{(i)}) \cdot \Psi(\mu)^{i-1}. \tag{13}$$

Proof. The theorem basically only needs the covariance formula, that is, for an arbitrary function $h(y)$, for which the expectation $\mathbb{E}\left(h(y)\right)$ exists, it holds

$$\mathbb{E}\left(h(y)y\right) = \mathbb{E}\left(h(y)\right)\mathbb{E}\left(y\right) + \mathrm{cov}(h(y), y),$$

in combination with Theorem 11. Thus, the covariance can be rewritten:

$$\begin{aligned}
\mathrm{cov}(\widehat{\theta}(y), y) &= \mathbb{E}\left(\widehat{\theta}'(y)V(y,\mu)\right) \\
&= \mathbb{E}\left(\widehat{\theta}'(y)\cdot y\right)\Psi\left(\mu\right) \\
&= \left[\mathbb{E}\left(\widehat{\theta}'(y)\right)\mathbb{E}\left(y\right) + \underbrace{\mathrm{cov}(\widehat{\theta}'(y), y)}_{=\mathbb{E}\left(\widehat{\theta}''(y)V(y,\mu)\right)}\right]\Psi\left(\mu\right) \\
&= \mathbb{E}\left(\widehat{\theta}'(y)\right)\underbrace{\mathbb{E}\left(y\right)\Psi\left(\mu\right)}_{=\mathrm{Var}(y)} + \mathbb{E}\left(\widehat{\theta}''(y)\underbrace{V(y,\mu)}_{=y\Psi(\mu)}\right)\Psi\left(\mu\right)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left(\widehat{\theta}'(y)\right)\mathrm{Var}(y) + \mathbb{E}\left(\widehat{\theta}''(y)y\right)\Psi(\mu)^2 \\
&= \mathbb{E}\left(\widehat{\theta}'(y)\right)\mathrm{Var}(y) + \mathbb{E}\left(\widehat{\theta}''(y)\right)\mathbb{E}\,(y)\,\Psi(\mu)^2 + \mathrm{cov}(\widehat{\theta}''(y), y)\Psi(\mu)^2 \\
&= \mathrm{Var}(y)\mathbb{E}\left(\widehat{\theta}'(y)\right) + \mathrm{Var}(y)\mathbb{E}\left(\widehat{\theta}''(y)\right)\Psi\,(\mu) + \mathrm{cov}(\widehat{\theta}''(y), y)\Psi(\mu)^2 \\
&= \ldots \\
&= \mathrm{Var}(y)\sum_{i=1}^{s}\mathbb{E}\left(\widehat{\theta}^{(i)}(y)\right)\Psi(\mu)^{i-1}.
\end{aligned}
$$

■

With Theorem 3, we have the following remarks:

(i) The condition $\widehat{\theta}^{(s+1)}(y) = 0$ is a strong condition that in many cases may not be fulfilled. However, if the function $\widehat{\theta}(y)$ is approximated by a Taylor expansion of order s around the mean μ, then the condition is fulfilled. Hence, with the Taylor approximation

$$\widehat{\theta}(y) \approx \tilde{\theta}(y) = \sum_{i=0}^{s}\frac{\widehat{\theta}^{(i)}(\mu)}{i!}(y-\mu)^i,$$

the covariance penalty can be approximated by

$$\mathrm{cov}(\widehat{\theta}(y), y) \approx \mathrm{cov}(\tilde{\theta}(y), y) = \mathrm{Var}(y)\sum_{i=1}^{s}\mathbb{E}\left(\tilde{\theta}^{(i)}(y)\right)\Psi(\mu)^{i-1}.$$

(ii) For the mean μ in $\tilde{\theta}(y)$, a plug-in estimator can be used. Either the estimated mean $\widehat{\mu}$ or the estimator of the saturated model, that is, the observed values y, are applicable. Thus, using a first order Taylor expansion and the observed values as estimators of the mean and substituting an estimator for the variance, formula (13) can be estimated by:

$$\mathrm{cov}(\widehat{\theta}(y), y) \approx \widehat{\mathrm{Var}(y)}\mathbb{E}\left(\widehat{\theta}'(y)\right). \tag{14}$$

(iii) Note that the subsequent result in Equation (13) can be generalized by allowing for linear translations of the random variable to be separated from the "variance" function. Under the same conditions that need to hold for the formula (13) and, additionally, with $a, d \in \mathbb{R}$ and $V(y, \mu) = (a + dy)\cdot\Psi(\mu)$, the covariance penalty is

$$\mathrm{cov}(\widehat{\theta}, y) = \mathrm{Var}(y)\sum_{i=1}^{s}\mathbb{E}(\widehat{\theta}^{(i)})\cdot d^{i-1}\Psi(\mu)^{i-1}. \tag{15}$$

(iv) Other distribution-specific methods to derive observable (conditional) covariance penalties are available for certain distributions. For instance for Bernoulli models no unbiased estimator of the conditional covariance penalty is available. Nevertheless, the covariance can be written as

$$\mathrm{cov}_{y_i,\gamma|\boldsymbol{y}_{-i}}(\widehat{\theta}_i, y_i) = \mu_i(1-\mu_i)\left(\widehat{\theta}_i(1) - \widehat{\theta}_i(0)\right). \tag{16}$$

since

$$\begin{aligned}\mathrm{cov}_{y_i,\boldsymbol{\gamma}|\boldsymbol{y}_{-i}}(\widehat{\theta}_i, y_i) &= \mathbb{E}_{y_i,\boldsymbol{\gamma}|\boldsymbol{y}_{-i}}\left[\widehat{\theta}_i(y_i)\,(y_i-\mu_i)\right] \\ &= \mu_i\widehat{\theta}_i(1)\,(1-\mu_i) + (1-\mu_i)\widehat{\theta}_i(0)\,(0-\mu_i) \\ &= \mu_i(1-\mu_i)\left(\widehat{\theta}_i(1)-\widehat{\theta}_i(0)\right).\end{aligned}$$

Notice that $\widehat{\theta}_i$ is defined as in Equation (5). For the deviance error q for instance this is $\widehat{\theta}_i = \log\left(\frac{\widehat{\mu}_i}{1-\widehat{\mu}_i}\right)$. Thus, substituting an estimator for the variance $\mu_i(1-\mu_i)$ leads to the estimate

$$\sum_{i=1}^{n}\widehat{\mathrm{cov}}_{y_i,\boldsymbol{\gamma}|\boldsymbol{y}_{-i}}(\widehat{\theta}_i, y_i) = \sum_{i=1}^{n}\widehat{\mu}_i(1-\widehat{\mu}_i)\left(\widehat{\theta}_i(1)-\widehat{\theta}_i(0)\right). \tag{17}$$

In Section 3.2, we will show the close connection between this estimator and the bootstrap estimator.

3.2 | Conditional parametric bootstrap

A direct way of estimating the conditional covariance penalty is the parametric bootstrap with conditional random effects. This means that every bootstrap sample is taken from the conditional distribution with conditional mean $\widehat{\boldsymbol{\mu}}_c$. For bootstrap simulations $\boldsymbol{z}$ of size B from the originally fitted model $\widehat{\boldsymbol{\theta}}|\boldsymbol{\gamma} = \widehat{\boldsymbol{\theta}}_c$, the covariance penalty is calculated by

$$\sum_{i=1}^{n}\widehat{\mathrm{cov}}_i = \sum_{i=1}^{n}\frac{1}{B-1}\sum_{j=1}^{B}\widehat{\theta}_{ij}(z_{ij})\left(z_{ij}-\overline{z}_{i\cdot}\right), \tag{18}$$

with the mean over all bootstrap samples $\overline{z}_{i\cdot} = \frac{1}{B}\sum_{j=1}^{B} z_{ij}$ for each data point i. The conditional parametric bootstrap assumes that the underlying model is true and is thus a model-based approach. On the other hand, as presented here, the method is global as it changes all cases in each simulation step in contrast to the plug-in estimates that only vary the ith data point when estimating $\widehat{\mathrm{cov}}_i$.

With the conditional parametric bootstrap, the simulation error of the conditional covariance penalty estimation can be assessed by

$$\mathrm{sd}\left(\sum_{i=1}^{n}\widehat{\mathrm{cov}}_i\right) = \left(\frac{\sum_{j=1}^{B}(c_j-\overline{c})^2}{B(B-1)}\right)^{\frac{1}{2}},$$

with

$$c_j = \sum_{i=1}^{n}\widehat{\theta}_{ij}(z_{ij})\left(z_{ij}-\overline{z}_{i\cdot}\right) \text{ and } \overline{c} = \frac{1}{B}\sum_{j=1}^{B}c_j.$$

Since in every bootstrap sample all data points are resampled the bootstrap in (Equation 18) needs B model refits.

Another possibility for a bootstrap estimate would be a so-called "fixed data" bootstrap, in which for each bootstrap sample of each individual observation the other $n-1$ data points are fixed and only the ith datum is resampled. This corresponds to the estimate in Equation (10). However, this approach is computationally burdensome since for each observed response a whole set of bootstrap samples and model fits must be computed and thus the total error estimation requires $n \cdot B$ model fits. Nonetheless, there is an approximation of the bootstrap estimate that makes it less computationally expensive. Therefore, consider the bootstrap estimator for the ith covariance penalty with all cases but the ith fixed. Instead of evaluating the main parameter of interest, we use a Taylor approximation around the estimated mean $\widehat{\mu}_i$ yielding

$$\widehat{\theta}_i(z) \approx \widehat{\theta}_i(\widehat{\mu}_i) + \left.\frac{\partial \widehat{\theta}_i}{\partial z}\right|_{\widehat{\mu}_i} \left(z - \widehat{\mu}_i\right).$$

Accordingly, the bootstrap estimator of the ith conditional covariance penalty can be approximated by

$$\begin{aligned}
\widehat{\mathrm{cov}}_i &= \frac{1}{B-1} \sum_{j=1}^{B} \widehat{\theta}_i(z_{ij}) \left(z_{ij} - \bar{z}_{i\cdot}\right) \\
&\approx \frac{1}{B-1} \sum_{j=1}^{B} \left(\widehat{\theta}_i(\widehat{\mu}_i) + \left.\frac{\partial \widehat{\theta}_i}{\partial z}\right|_{\widehat{\mu}_i} \left(z_{ij} - \widehat{\mu}_i\right) \right) \left(z_{ij} - \bar{z}_{i\cdot}\right) \\
&\approx \left.\frac{\partial \widehat{\theta}_i}{\partial z}\right|_{\widehat{\mu}_i} \frac{1}{B-1} \sum_{j=1}^{B} \left(z_{ij} - \widehat{\mu}_i\right) \left(z_{ij} - \bar{z}_{i\cdot}\right) \\
&\approx \left.\frac{\partial \widehat{\theta}_i}{\partial z}\right|_{\widehat{\mu}_i} \widehat{\mathrm{Var}(y_i)},
\end{aligned} \tag{19}$$

with the last approximation holding if the number of bootstrap samples tends to infinity, $B \to \infty$. A similar link between the bootstrap and the plug-in estimator for Bernoulli conditional covariance penalties is derived in Section 4.2.

3.3 | Conditional cross-validation

The probably most popular method for prediction error estimation is cross-validation. Compared to conditional parametric bootstrap and the plug-in estimates, the conditional cross-validation has the advantage that it is not model-dependent. On the other hand, just like the plug-in estimates, the conditional cross-validation is a local method in the sense that, for estimation of the covariance penalty, it only changes the ith data point. Let $\widehat{\mu}_{-i}$ be the estimated mean with the ith observation deleted, that is, the estimator based on the reduced data set $\boldsymbol{y}_{-i} = (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$. Then the cross-validation estimate of the conditional expected prediction error (Equation 7) is $Q(y_i, \widehat{\mu}_{-i})$. Facilitating the connection of the apparent error, $Q(y_i, \widehat{\mu}_i)$, with the expected conditional prediction error stated in the optimism theorem 1 results in

$$Q(y_i, \widehat{\mu}_{-i}) - Q(y_i, \widehat{\mu}_i) = 2\widehat{\mathrm{cov}}_i.$$

Thus, the conditional covariance penalty can by derived by

$$\sum_{i=1}^{n} \widehat{\mathrm{cov}}_i = \frac{1}{2} \sum_{i=1}^{n} \left[Q(y_i, \widehat{\mu}_{-i}) - Q(y_i, \widehat{\mu}_i) \right]. \tag{20}$$

For instance, in case of the deviance error and data from an exponential family distribution, the cross-validation estimator of the conditional covariance penalty is

$$\sum_{i=1}^{n} \widehat{\mathrm{cov}}_i = \sum_{i=1}^{n} b(\widehat{\vartheta}_{-i}) - b(\widehat{\vartheta}_i) + y_i(\widehat{\vartheta}_i - \widehat{\vartheta}_{-i}), \tag{21}$$

where $\widehat{\vartheta}_i$ is the estimated natural parameter of the exponential family, and $\widehat{\vartheta}_{-i}$ is the estimated natural parameter with the ith case deleted.

The parametric bootstrap is related to the cross-validation by a Rao–Blackwell type of relationship, see Efron (2004). That implies that the conditional bootstrap (and the proposed *Steinian* estimators) is more accurate than cross-validation, assuming that the applied model is near enough to the truth. The simulation study, though, does not reflect this behavior in the case of mixed models.

4 | SIMULATIONS

In order to assess the behavior of the different proposed estimation techniques and error classes, various simulation scenarios are presented. A particular focus will lie on the model choice behavior of the estimators if the variance parameter of the random effects lies on the boundary of the parameter space.

For the Bernoulli distribution, the deviance error is employed. Thus, when comparing two distinct models, this corresponds to a conditional AIC in an exponential family setting. The deviance error is additionally used to choose between models following a gamma distribution. In this setting, the connection of the *Steinian* to the covariance penalty for Gaussian distributions becomes apparent. Moreover, it emphasizes the close relationship between the *Steinian* and the generalized degrees of freedom defined in Ye (1998). Furthermore, the expected squared error of a random intercept model with conditionally scaled t distribution is investigated (again with an emphasis on the null model rejection rate).

4.1 | Gamma distribution

The conditional density of the data generating process for gamma distributed responses is given by

$$f(y_{ij}|\mu_{ij}, \nu) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu_{ij}} \right)^{\nu} y^{\nu-1} \exp\left(-\frac{\nu y}{\mu_{ij}} \right), \tag{22}$$

for $i = 1, \dots, n$ and $j = 1, \dots, m$, with $\mu_{ij} = \exp\left(\beta_0 + \gamma_j\right)$ and $\boldsymbol{\gamma} \sim \mathcal{N}\left(0, \tau^2 \boldsymbol{I}_m\right)$. The number of individuals is set to $n = 12$, and the number of observations is $m = 6$ per individual. An example

with more random effects is considered in the application in Section 5. The variance parameter of the random intercept τ^2 varies between 0 and 1.6. For each setting, 1,000 data sets of model (Equation 22) are generated. The scale parameter $\phi = \frac{1}{\nu} = 1$ is constant for all observations and is estimated within each model fit as $\widehat{\phi}$.

The error for the gamma distributed observations is assessed with the deviance error as in formula (4) with the natural parameter $\widehat{\theta}_{ij} = \widehat{\vartheta}_{ij} = -\frac{1}{\widehat{\mu}_{ij}}$. Based on formula (19), with $\widehat{\mathrm{Var}(y_{ij})} = \widehat{V}_{ij} = \widehat{\phi}\widehat{\mu}_{ij}^2 = \widehat{\phi}\frac{\partial\widehat{\mu}_{ij}}{\partial\widehat{\theta}_{ij}}$, we approximate the conditional covariance penalty for gamma distributed responses of the ith data point by

$$\widehat{\mathrm{cov}}_{ij} \approx \widehat{V}_{ij}\frac{\partial\widehat{\theta}_{ij}}{\partial y_{ij}} = \widehat{\phi}\frac{\partial\widehat{\mu}_{ij}}{\partial\widehat{\theta}_{ij}}\frac{\partial\widehat{\theta}_{ij}}{\partial y_{ij}} = \widehat{\phi}\frac{\partial\widehat{\mu}_{ij}}{\partial y_{ij}}. \tag{23}$$

This equation highlights that the *Steinian*-type estimator and the generalized degrees of freedom proposed in Ye (1998) coincide. Moreover, the close relationship to the findings of Liang et al. (2008) becomes apparent. Next to this estimate, the conditional covariance penalties are estimated by conditional parametric bootstrap and conditional cross-validation, see Equations (18) and (20). The estimation of the random effects variance parameter is done by REML estimation based on the R-package `mgcv` version `1.8-2`, see Wood et al. (2016). The bootstrap needs 500 model fits and takes about 13 s on a 2.9-GHz personal computer, while the cross-validation only needs $n \cdot m = 72$ model fits and about 2 s. The derivatives in Equation (23) are calculated on the basis of the algorithm in Gilbert and Varadhan (2012). This takes about 17 s.

A null model rejection rate plot similar to those considered in Greven and Kneib (2010) is displayed in Figure 1. This plot shows the frequency of favoring the more complex model (Equation 22), incorporating random effects over the simpler models with only an intercept. For each data set a simple model excluding random effects and a complex model including a random intercept is fitted. For both models, the total expected prediction error is estimated. The model with the smaller prediction error is selected and the model with the higher prediction error is rejected. The proportion of times the complex model is chosen is plotted in Figure 1. The cross-validation estimate here behaves similar to the marginal AIC in Greven and Kneib (2010). For a random effects variance of 0.5, the cross-validation only chooses to include random effects in roughly 60% cases, whereas the *Steinian* estimator does so in more than 75% of the cases. The bootstrap shows a good behavior although it incorporates random effects in a quarter of the cases although there are none in the underlying data-generating mechanism.

The estimated conditional covariance penalties are listed in Table 1. The table shows that all three methods give similar estimates for the conditional covariance penalties. The value of the *Steinian* estimator (Section 3.1) for random effects variance $\tau^2 = 0.4$ is salient. This is due to the fact that the estimates have not appropriately been corrected for numerical anomalies that arise from the use of numerical derivation. An extended table with the standard deviations and corresponding boxplots can be found in the Supporting Information.

4.2 | Bernoulli distribution

For the Bernoulli distribution, the true data generating process is given by a logistic random intercept model, with the conditional probability function

$$f(y_{ij}|\mu_{ij}) = (1-\mu_{ij})^{1-y_{ij}}\mu_{ij}^{y_{ij}} \text{ for } i = 1, \dots, n \text{ and } j = 1, \dots, m, \tag{24}$$

with $\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \gamma_j$ and $\boldsymbol{\gamma} \sim \mathcal{N}\left(0, \tau^2 \boldsymbol{I}_m\right)$. The number of individuals is set to $n = 13$, and the number of observations is $m = 7$ per individual. The variance parameter of the random intercept τ^2 varies between 0 and 2.4. For each setting, 1,000 data sets of model (Equation 24) are generated, and the covariance penalties of the model are estimated by the different estimation techniques proposed in the preceding sections. The bootstrap estimate is based on 800 bootstrap samples.

The models are fitted with the REML method implemented in the R-package `lme4`, see Bates et al. (2015). The conditional bootstrap and the *Steinian* are estimated with the R-package `cAIC4`, see Säfken, Rügamer, Kneib, and Greven (2018). The bootstrap here needs 800 model fits and takes about 34 s on a 2.9-GHz personal computer, while the cross-validation which takes about 5 s and the *Steinian*, which takes about 3 s, only need $n \cdot m$ model fits. The fits needed for the *Steinian* are faster than for cross-validation, since the data set remains unchanged except for one response value in each computation. Thus, from the computational perspective the *Steinian* performs best.

Figure 1 displays the frequencies of how often the complex model (Equation 24) is favored against a simple model with only an intercept. This means we choose the model that minimizes the expected conditional prediction error $\mathbb{E}_{\boldsymbol{z}|\boldsymbol{\gamma}}\left(Q\left(\boldsymbol{z}, \widehat{\boldsymbol{\mu}}\right)\right)$. The error function Q is the deviance error as in formula (4) with the logit parameter $\widehat{\theta}_{ij} = \widehat{\vartheta}_{ij} = \log\left(\frac{\widehat{\mu}_{ij}}{1-\widehat{\mu}_{ij}}\right)$. It is, however, not clear how the covariance penalties of the simple model can be estimated. In order to stay consistent with the estimation of the conditional covariance penalties of the complex models, the covariance penalties of the simple models are estimated with bootstrap, cross-validation, and the *Steinian* applied to the generalized linear model. The bootstrap and the *Steinian* in one quarter of the cases choose the complex model although the true underlying model does not incorporate random effects. Notice that the behavior of choosing too many parameters is rather common for AIC-like criteria. For instance, the significance level of the AIC in standard settings is approximately 0.157, see Greven and Kneib (2010). The bootstrap chooses the true complex model slightly more often than the *Steinian* for increasing random effects variance. Cross-validation, on the other hand, performs worse as the chance of selecting the false complex model is almost 0.5 and increases very slowly with the increasing random effects variance.

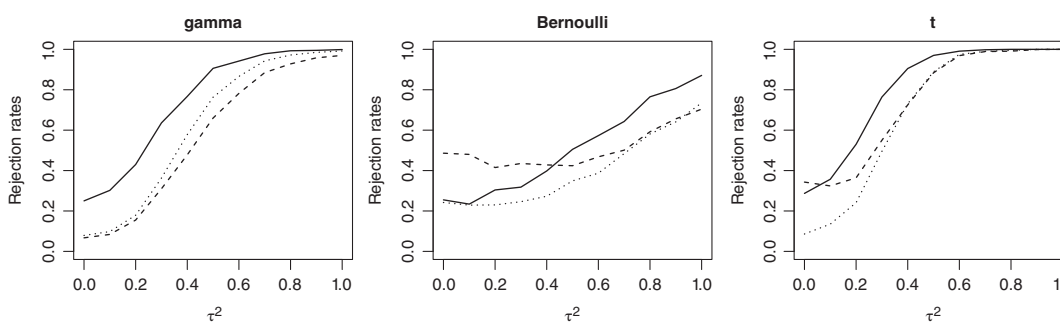

FIGURE 1 Frequency of choosing the complex model against a simple model only including an intercept for conditionally gamma, Bernoulli, and scaled t distributed responses. The conditional covariance penalties are estimated by bootstrap (Equation 18), cross-validation (Equation 20), and the *Steinian* estimator (Section 3.1)

TABLE 1 Mean and standard error of the conditional covariance penalty samples estimated by bootstrap (cov_b), cross-validation (cov_{cv}), and the Steinian estimates (cov_s) for gamma, Bernoulli, and scaled t distribution

	Gamma			Bernoulli			Scaled t		
τ^2	cov_b	cov_{cv}	cov_s	cov_b	cov_{cv}	cov_s	cov_b	cov_{cv}	cov_s
0.0	3.54	3.26	2.95	2.94	2.69	2.55	3.98	3.53	3.58
0.1	3.66	3.66	3.29	2.95	2.67	2.54	4.17	4.25	4.32
0.2	3.92	4.57	4.10	3.00	3.03	2.86	4.58	5.72	5.67
0.4	5.88	7.90	11.93	3.16	3.60	3.39	7.48	10.67	10.34
0.6	8.32	10.74	8.87	3.64	4.80	4.50	11.57	14.44	13.82
0.8	10.53	12.87	10.15	4.40	6.20	5.82	14.35	16.43	15.67

For the 1,000 estimated conditional covariance penalties the empirical means of the conditional covariance samples are listed in Table 1. The boxplots and the standard errors for different sizes of the random effects variance parameters can be found in the Supporting Information. The performance of cross-validation, bootstrap, and the *Steinian* highly depend on the random effects variance parameter. This makes a comparison difficult. The variability of the *Steinian* with the true mean plugged in is larger than the variability when the estimated mean is used.

Notice that the bootstrap estimate here is unconditional in the sense that all cases of the data set are varied in each set of covariance penalty estimates. The bootstrap that is conditioned on $\boldsymbol{y}_{-i}$, that is, in which for the estimation of the ith covariance penalty only the ith case is resampled for Bernoulli responses is approximately equal to the *Steinian* (Equation 17), see Säfken et al. (2018).

4.3 | Scaled t distribution

As stated, the proposed framework does not only work for exponential family distributions and the deviance error. Therefore, the behavior of the squared error functions and the different corresponding conditional covariance estimators are considered in this setting with conditionally scaled t distributed responses. Hence, the data are generated by the mechanism

$$f(y_{ij}|\mu_{ij}, \nu, \sigma) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}\sigma}\left(1 + \frac{1}{\nu}\left(\frac{y_{ij} - \mu_{ij}}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}, \tag{25}$$

for $i = 1, \dots, n$ and $j = 1, \dots, m$, with $\mu_{ij} = \exp\left(\beta_0 + \gamma_j\right)$ and $\boldsymbol{\gamma} \sim \mathcal{N}\left(0, \tau^2 \boldsymbol{I}_m\right)$. The number of individuals is set to $n = 17$, and the number of observations is $m = 7$ respectively. The variance parameter of the random intercept τ^2 varies between 0 and 1.6. For each setting 1,000 data sets of model (Equation 25) are generated. For ease of computation, we expect the remaining parameters to be fixed and known, that is, $\nu = 7$ and $\sigma = 1$.

The total expected prediction error is assessed by the squared error function in **Example I**. Since for the squared error the parameter of main interest is $\widehat{\mu}$, the total expected prediction error that we want to minimize is

$$\mathbb{E}_{\boldsymbol{z}|\boldsymbol{u}} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(z_{ij} - \widehat{\mu}_{ij}\right)^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(y_{ij} - \widehat{\mu}_{ij}\right)^2 + 2 \sum_{i=1}^{n} \sum_{j=1}^{m} \operatorname{cov}\left(\widehat{\mu}_{ij}, y_{ij}\right). \tag{26}$$

Thus, based on formula (19), we approximate the conditional covariance penalty for scaled t distributed responses of the ijth data point by writing the parameter of main interest as a function of the data $\widehat{\mu}_{ij}(y)$ and then taking the derivative with respect to the data y at the point of the current estimate $\widehat{\mu}_{ij} = \widehat{\mu}_{ij}(y_{ij})$, that is,

$$\widehat{\operatorname{cov}}_{ij} \approx \widehat{\operatorname{Var}(y_{ij})} \left. \frac{\partial \widehat{\mu}_{ij}(y)}{\partial y} \right|_{y=\widehat{\mu}_{ij}}. \tag{27}$$

Next to this approximate estimate, the conditional covariance penalties are estimated by conditional parametric bootstrap and conditional cross-validation, see Equations (18) and (20). The bootstrap estimate is based on 800 bootstrap samples.

The models are fitted with the `R`-package `mgcv` version `1.8-2`, see Wood et al. (2016). This package uses a REML criterion to find the optimal random effects variance parameter τ^2. The bootstrap here needs 800 model fits and takes about 32 s on a 2.9-GHz personal computer, while the cross-validation and the *Steinian* only need $n \cdot m = 119$ model fits. However, the cross-validation only takes 4 s while the *Steinian* also takes 32 s. The derivatives in Equation (27) are numerically approximated based on the algorithm in Gilbert and Varadhan (2012).

The estimated conditional covariance penalties are listed in Table 1. An extended table with further random effects variances and the corresponding standard deviations can be found in the Supporting Information. The means of all three estimation techniques are similar for all random effects variances. In many cases, the *Steinian* lies between the cross-validation and the bootstrap estimate. In combination with the results on the selection frequency (see Figure 1), this gives evidence for the superior behavior of the *Steinian*. Although the covariance penalty for $\tau^2 = 0$ is smaller for the *Steinian* than for the bootstrap, the *Steinian* selects the null model more often than the bootstrap. So the *Steinian* seems to penalize in the "right" situations.

The convergence to a selection rate (of the more complex model) of one with rising signal-to-noise ratio τ^2 is fast, as can be seen in Figure 1. The distribution under consideration is close to the Gaussian for which the convergence rate is also high. However, the squared error function is also a possible influencing factor. Moreover, the *Steinian* has lower variance than the cross-validation and bootstrap in all settings. The reduced variability can also be observed in the boxplots in the Supporting Information.

Summing up there is no superior estimation method in terms of the model choice behavior. In fact, the behavior depends on the distribution and the type of prediction error that is applied.

5 | APPLICATION ON PREDICTING STUNTING OF CHILDREN IN ZAMBIA

In this case study, we focus on finding the model that is most suitable for predicting childhood malnutrition. We base our analyses on a dataset from the 1992 Zambia Demographic and Health Survey. The dataset was analyzed by Fahrmeir et al. (2013) and is publicly available in the Supporting Information of the book. The data consists of 4,421 observations from 54 districts in Zambia.

For measuring childhood malnutrition, we use stunting, that is, insufficient height for age. Thus, our variable of interest is the child height (in cm) standardized with respect to all children of the same age. The standardization uses the median instead of the mean. Several covariates are available. Next to the residential district these are the gender (binary variable), the education level of the mother (three possible outcomes), the employment situation of the mother (binary variable), and some continuous covariates, that is, the duration of breastfeeding (in months), the age, height, and body mass index of the mother and the age of the child.

Other authors used a normal distribution for the response variable, see for example Greven and Kneib (2010) or Kandala, Lang, Klasen, and Fahrmeir (2001). Instead, this analysis uses a scaled t model as in Equation (25) in order to account for heavy tails in the response variable.

5.1 | Conditional prediction of stunting with a linear mixed model

As we want to distinguish between conditional and marginal prediction in mixed models, the first model is a random intercept model that contains a district-specific random intercept and all the binary and categorical variables. Hence, the predictor is given by

$$\log(\mu_i) = \boldsymbol{x}_i^t \boldsymbol{\beta} + \gamma_{d_i}, \; i = 1 \dots n, \tag{28}$$

all the binary and categorical variables are subsumed in the covariate vector $\boldsymbol{x}_i$ and γ_{d_i} is the random intercept from district d, where the ith child lives, and accounts for the spatial heterogeneity. The random intercepts are independent and identically normal distributed with mean 0 and unknown variance τ^2. We suppose that the standardized child height y_i follows a scaled t distribution as in Equation (25) with the remaining parameters ν and σ being estimated alongside the random intercept variance, that is estimated with restricted maximum likelihood, see Wood et al. (2016).

The predicted district-specific random intercepts γ_d from the estimated model are plotted in Figure 2. Notice that for three districts there are no observations. The map shows that stunting is worse in southern Zambia.

Our aim is to estimate the prediction error of the random intercept model. Thus, for a new data point with covariate vector $\boldsymbol{x}$ we want to know the expected squared difference of the standardized child height z and our estimator $\widehat{\mu}$, that is, $(z - \widehat{\mu})^2$. For the new data point, one might either know the district the child is from or one may not. If the district is known one can condition on the random effects for the prediction. The expected conditional squared prediction error is then given by $\mathbb{E}_{z|\gamma}(z - \widehat{\mu})^2$. If otherwise the district is unknown the random intercepts γ_d need to be integrated out. This corresponds to the marginal prediction error. We focus on the conditional perspective and hence think of the district as known. Thus, the expected squared prediction error can be assessed by the sum of the apparent error and twice the covariance penalty similar to formula (26). As in Section 4.3, we estimate the conditional covariance penalty by conditional parametric bootstrap, conditional cross-validation and the *Steinian* as defined in Equation (27). The bootstrap is based on 5,000 bootstrap samples.

The apparent error of the fitted model (Equation 28) is 4,148.81. The covariance penalty estimated with cross validation is 4.22 and hence the conditional prediction error using the conditional optimism Theorem 9 is 4,157.26. If the covariance penalty is estimated by bootstrap it is 6.26, and the expected conditional prediction error is 4,161.32. The *Steinian* estimate for the covariance penalty as in Equation (27) is 7.11 and the corresponding prediction error is 4,163.03. Thus, the estimated covariance penalty depends on the estimation technique and when

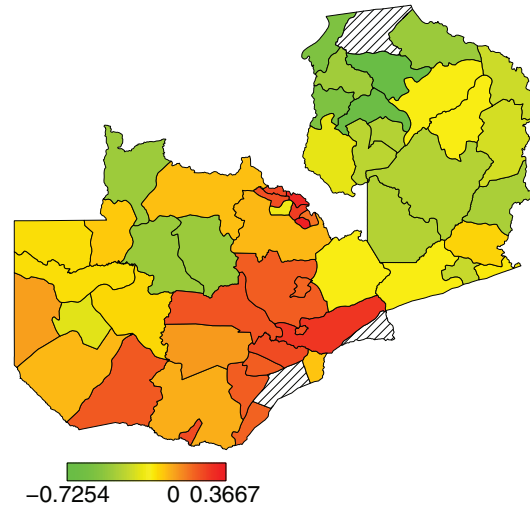

FIGURE 2 A map of Zambia showing the predicted district-specific random intercepts γ_d [Colour figure can be viewed at wileyonlinelibrary.com]

comparing these results to the simulation study there is no clear tendency of any estimation technique to give higher or lower estimates than any other.

5.2 | Conditional prediction of stunting with an additive mixed model

In a more sophisticated approach, we include the continuous covariates, that is, the duration of breastfeeding in months (*feed*), the age of the child (*cage*), height and body mass index of the mother (*hei* and *bmi*), and the age of the mother (*mage*) in our model. Hence, for the full model, the mean of the scaled t distribution is modeled as

$$\log(\mu_i) = \boldsymbol{x}_i^t \boldsymbol{\beta} + f_1(feed_i) + f_2(cage_i) + f_3(hei_i) + f_4(bmi_i) + f_5(mage_i) + \gamma_{d_i}. \tag{29}$$

The nonlinear functions $f_1(\cdot), \dots, f_5(\cdot)$ are modeled by thin-plate regression splines, and there is a penalty term associated with each function controlling for the smoothness of the function. The penalty term is an approximation of

$$\int f_j''(x)^2 dx, j = 1, \dots, 5,$$

as proposed in Wood (2017). $\boldsymbol{x}_i$ and γ_{d_i} are again the categorical covariate vector and the district specific random effect. Notice that the additive mixed model can be reformulated in terms of a mixed model. For these kinds of models a conditional approach is especially sensible since the covariate information of a new realization can be thought of as sharing the same random effects, see Greven and Kneib (2010). For the full model (Equation 29), the fitted smoothing splines and a QQ-plot of the random effects are shown in Figure 3. The effects of the age and the body mass index of the mother are already estimated to be almost linear in the full model.

We use the total conditional prediction error to decide if the effects are modeled by linear or nonlinear functions. Overall, there are 32 possible combinations of linear and nonlinear functions if we include all continuous covariates and let all the models contain the district specific random effects.

Table 2 contains the estimated total conditional prediction error with the covariance penalties estimated by cross-validation, bootstrap, or the *Steinian* for a group of submodels. The table containing all submodels is available in the Supporting Information. All estimation techniques

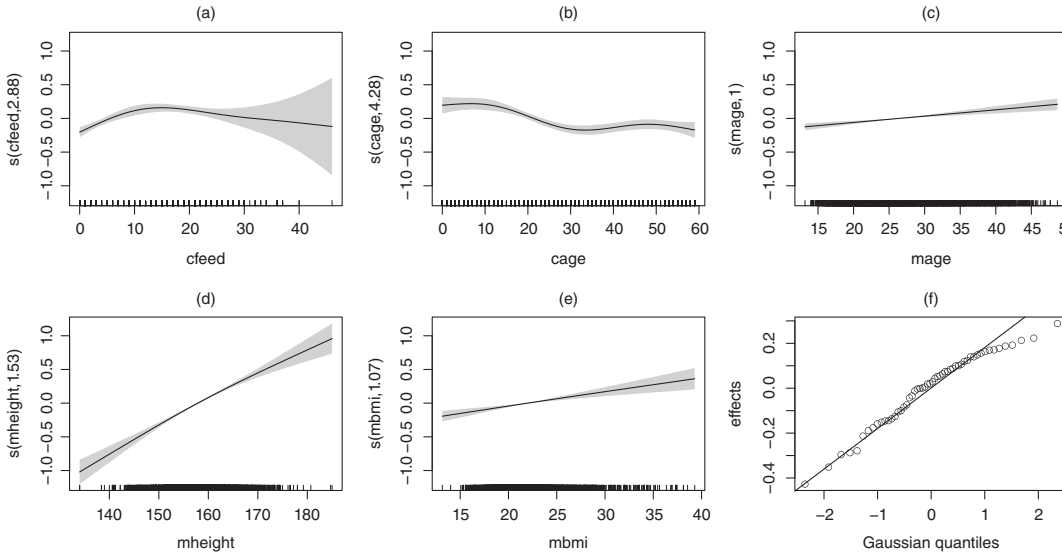

FIGURE 3 The fitted smoothing splines and spatial random effects of the full model (Equation 29) on stunting in Zambia. (a) is the effect of the duration of breastfeeding (in months), (b) is the effect of the child age, (c) is the effect of the age of the mother, (d) is the effect of the height of the mother, (e) is the effect of the body mass index of the mother, and (f) is the normal QQ-plot of the predicted spatial random effects of the districts

TABLE 2 Estimated conditional prediction error for model (Equation 29) and some submodels. The first five columns indicate if the covariates are modeled by linear (–) or nonlinear (∼) functions. The covariance penalties are estimated by cross-validation (cv), bootstrap (bs) and *Steinian*.The blue covariance penalties give the lowest total prediction error

Feed	Cage	Mage	*hei*	*bmi*	cv	bs	*Steinian*
∼	∼	∼	∼	∼	3,821	3,845	3,904
∼	∼	∼	∼	–	3,819	3,844	3,898
∼	∼	∼	–	–	3,818	3,843	3,895
∼	∼	–	∼	–	3,819	3,842	3,898
∼	∼	–	–	∼	3,820	3,844	3,900
∼	∼	–	–	–	3,818	3,842	3,895
–	–	–	–	–	3,849	3,871	3,921

choose two possible models with lowest total conditional prediction error. All of them agree on the model with the effect of the age, the height, and the body mass index as linear. However cross-validation and *Steinian* additionally prefer the effect of the age of the mother to be nonlinear while bootstrap additionally chooses the effect of the height of the mother to be nonlinear.

6 | DISCUSSION

The conditional prediction error is relevant not only for mixed models but beyond that also for regression models using the mixed model formulation as estimation vehicle (Wood et al., 2016).

The methods for estimating the conditional prediction error presented here can easily be applied to other distributional settings and are therefore broadly applicable. On the downside the estimation procedures come along with a computational burden. This is especially the case when using cross-validation and bootstrap. But also for the *Steinian* if numerical approximations are used to calculate the inherent derivatives. The most plausible way for reducing the computational burden is to approximate the derivatives as in Equation (19) directly in the fitting procedure. However, the resulting estimation techniques would be very specific in terms of response distribution and fitting procedure.

While our approach focuses on the conditional perspective on mixed models, in certain cases, it makes sense to consider a marginal prediction error in mixed models. This, for example, can be the case if in future observations the grouping structure underlying the random effects is unknown. Deriving marginal covariance penalties can be considerably more difficult taking into account the complex structure of the marginal distribution. Nevertheless, this is an interesting question for further research.

Other future adjacent fields of research are, for example, extending the results of Efron (2004) on the Rao–Blackwell type of relation between *Steinian* methods and cross-validation to conditional covariance penalties. Moreover, the grouping structure of the mixed models could be taken into account when using cross-validation. A somewhat more fundamental question is how to extend these results to non-mean regression models such as distributional regression, see Kneib (2013) for an overview.

ACKNOWLEDGEMENTS

The authors thank the reviewers and associate editor for the useful comments that have improved the quality of this paper a lot. The first author also want to thank the Department of Statistics at the Ludwig Maximilian University of Munich where parts of the research work was carried out. The research was supported by the RTG 1644—Scaling Problems in Statistics.

ORCID

Benjamin Säfken https://orcid.org/0000-0003-4702-3333

REFERENCES

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle*. In B.N. Petrov, & F. Csäki (Eds.), *2nd International Symposium on Information Theory*. Budapest: Akademiai Kiádo; 267–281.

Anderssen, R. S., & Bloomfield, P. (1974). A time series approach to numerical differentiation. *Technometrics*, *16*, 69–75.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Crainiceanu, C. M., & Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*, 165–185.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, *81*, 461–470.

Efron, B. (2004). The estimation of prediction error. *Journal of the American Statistical Association*, *99*, 619–632.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. D. (2013). *Regression models, methods and applications* (2nd ed.). Berlin, Heidelberg: Springer.

Gilbert, P., & Varadhan, R. (2012). numDeriv: accurate numerical derivatives.

Greven, S., Crainiceanu, C. M., Küchenhoff, H., & Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*, *17*, 870–891.

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

Kandala, N. B., Lang, S., Klasen, S., & Fahrmeir, L. (2001). Semiparametric analysis of the socio-demographic and spatial determinants of undernutrition in two African countries. Retrieved from http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-1626-2.

Kneib, T. (2013). Beyond mean regression. *Statistical Modelling*, *13*, 275–303.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963–974.

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

Müller, S., Scealy, J. L., & Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, *28*, 135–167.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression. Cambridge series in statistical and probabilistic mathematics*. Cambridge, UK: Cambridge University Press.

Saefken, B., Kneib, T., van Waveren, C.-S., & 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.

Säfken, B., Rügamer, D., Kneib, T., & Greven, S. (2018). Conditional model selection in mixed-effects models with cAIC4. *ArXiv e-prints*.

Sakamoto, W. (2019). Bias-reduced marginal Akaike information criteria based on a Monte Carlo method for linear mixed-effects models. *Scandinavian Journal of Statistics*, *46*, 87–115.

Self, S. G., & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*(398), 605–610.

Shen, X., & Huang, H.-C. (2006). Optimal model assessment, selection, and combination. *Journal of the American Statistical Association*, *101*, 554–568.

Shun, Z., & McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*, 749–760.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Sixth Berkeley Symposium on Mathematical Statistics and Probability*. Retrieved from http://projecteuclid.org/euclid.bsmsp/1200514239.

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

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics*, *13*(4), 1378–1402.

Wood, S. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Boca Raton, FL: Chapman and Hall.

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*, 1548–1563.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, *93*, 120–131.

Yu, D., Zhang, X., & Yau, K. K. W. (2018). Asymptotic properties and information criteria for misspecified generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*, 817–836.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Säfken B, Kneib T. Conditional covariance penalties for mixed models. *Scand J Statist*. 2020;47:990–1010. https://doi.org/10.1111/sjos.12437