Diploma Thesis

Conditional Akaike Information Criteria in Generalized Linear Mixed Models

Jona Cilia Rachel Cederbaum

Supervisors

Dr. Sonja Greven
Prof. Dr. Thomas Kneib

Institut für Statistik
Ludwig-Maximilians-Universität München

11. August 2011

Abstract

This thesis focuses on the selection of random effects based on Akaike information criteria (AIC) in mixed models. Conventionally, the AIC based on the marginal distribution is used. However, Greven and Kneib (2010) showed that this is not an appropriate selection criterion in this framework. Therefore, this thesis concentrates on the AIC based on the conditional distribution (cAIC) for which a correction is needed to take the estimation uncertainty in the random effects into account.

For the case of linear mixed models, an analytic representation of a corrected version of the cAIC exists. It is an unbiased estimator for the conditional Akaike information. Although so far no analogue has been derived for generalized linear mixed models, an asymptotically unbiased estimator has recently been proposed by Yu and Yau (2011). This is one of the criteria which has been analyzed in the scope of this thesis. Secondly, we considered the usage of a covariance based penalty as correction term in the generalized case which has been suggested in the context of general prediction problems. We demonstrated that two bootstrap versions are possible to estimate the covariance based measure and studied in this context the influence of the error variance. We investigated the behavior of the new generalized correction approaches in two simulation studies for linear mixed models. We compared these results to the results of the analytic criterion and of the uncorrected cAIC. This permitted us to assess the performance of the new corrections in the important special case of normal errors which is an essential step towards the examination in the generalized setting. In addition, we applied all criteria in a case study on childhood malnutrition in Zambia in order to illustrate the practical relevance of model selection via AICs.

The simulations showed that the cAIC of Yu and Yau is almost identical to the analytic cAIC under maximum likelihood estimation, but differs in the restricted maximum likelihood case. We found that the implementation of this measure is rather complex due to numerical problems. For the covariance based correction term, it turned out that the consideration of the error variance is more important than expected and that further modifications will be needed in order to fully assess this approach.

Zusammenfassung

Diese Arbeit befasst sich mit der Selektion von zufälligen Effekten basierend auf Akaike Informationskriterien (AIC) in gemischten Modellen. Herkömmlicherweise wird hierfür das AIC basierend auf der marginalen Verteilung der Zielgrößen verwendet. Greven and Kneib (2010) zeigten jedoch, dass das marginale AIC kein geeignetes Selektionskriterium für die Selektion von zufälligen Effekten darstellt. Aus diesem Grund konzentrierten wir uns in dieser Arbeit auf das AIC basierend auf der konditionalen Verteilung. Dieses bedarf einer Bias-Korrektur um die Unsicherheit in der Schätzung der zufälligen Effekte zu berücksichtigen.

Für den Spezialfall von linearen gemischten Modellen existiert bereits eine analytische Darstellung einer korrigierten Version des cAICs. Diese ist ein unverzerrter Schätzer der Akaike Information. Bisher wurde kein Analogon für den Fall von generalisierten linearen gemischten Modellen hergeleitet. Allerdings entwickelten Yu and Yau (2011) kürzlich einen asymptotisch unverzerrten Schätzer. Dieser stellt eines der beiden Kriterien dar, die wir im Rahmen dieser Arbeit genauer untersuchten. Weiterhin betrachteten wir die Verwendung eines kovarianzbasierten Penaltyterms, welcher im Kontext allgemeiner Prädiktionsprobleme vorgeschlagen wurde. Wir zeigten, dass es zwei Bootstrap-basierte Methoden gibt um den kovarianzbasierten Penaltyterm zu schätzen. In diesem Zusammenhang analysierten wir auch den Einfluss der Fehlervarianz. In Rahmen zweier Simulationen für lineare gemischte Modelle untersuchten wir das Verhalten der beiden neuen generalisierten Korrekturansätze. Wir verglichen diese Ergebnisse mit denen des analytischen und des unkorrigierten cAICs. Dies ermöglichte uns, die Performance der neuen Ansätze in dem wichtigen Spezialfall von linearen gemischten Modellen zu ermitteln, was einen essentiellen Schritt in Richtung einer Untersuchung für den generalisierten Fall darstellt. Darüberhinaus wendeten wir alle Kriterien in einer Fallstudie zu Unterernährung in Zambia an, um die praktische Relevanz von Modellselektion via AICs zu illustrieren.

Die Simulationen zeigten, dass das cAIC von Yu und Yau unter Maximum-Likelihood-Schätzung beinahe identisch zu dem analytischen cAIC ist, sich jedoch unter restringierter Maximum-Likelihood-Schätzung von diesem unterscheidet. Außerdem erwies sich die Implementation des cAICs von Yu und Yau aufgrund von numerischen Schwierigkeiten als relativ komplex. Bei den Untersuchungen des kovarianzbasierten cAICs zeigte sich, dass die Betrachtung der Fehlervarianz einen größeren Einfluss auf die Ergebnisse hat als erwartet und dass es weiterer Modifikationen bedarf, um diesen Ansatz vollständig bewerten zu können.

Acknowledgments

First of all I would like to thank Sonja Greven and Thomas Kneib for their constant support and encouragement. Sonja I would like to thank especially for her time and for showing me how exciting research can be. I am grateful to Thomas for inviting us to Oldenburg where we spent a very nice and interesting time and for his good advices on my future plans.

My mother, Kai and my sister I would like to thank for carefully proof-reading this thesis, although coming from another background and probably wondering sometimes what I was actually doing. I am grateful to Matthias Hunger who proof-read parts of this thesis through the eyes of a statistician.

My special thanks go to my mother for showing me the advertising for the statistics study program of the LMU some years ago and to her and my father for their moral and financial support throughout my studies.

Finally, I would like to thank all my friends who are always there for me and who keep me happy. My deepest and dearest thanks belong to Kai, for asking good questions and for inspiring conversations and especially for his understanding, encouragement and for just being there.

Contents

Chapter 1

Introduction

Mixed models are widely used regression models which find application in many statistical areas. They are not only commonly employed in the analysis of longitudinal and cluster data, but also serve as an important inferential tool for penalized spline smoothing and have numerous applications beyond. As they offer computational simplifications for complex models and enable flexible modeling at the same time, mixed models have become a popular instrument in various disciplines such as biometrics, physics, biology and social sciences.

When using mixed models, there is no upper limit to model complexity. This is why model selection is indispensable. In particular the selection of random effects plays an important role as they constitute a major characteristic of mixed models.

In general, one possibility to perform model selection is to compare the regression models via their Akaike information criteria (AIC) (Akaike, 1973). The AIC has proven useful in practice for many classes of models and has a theoretical justification. It is more flexible than hypothesis testing as it allows comparing even non-nested models.

For mixed models, two versions of the AIC can be considered, based on either the marginal or the conditional distribution of the response variable. However, the usage of the AIC remains difficult in the context of mixed models as two main challenges result from their special structure. First, the observations in mixed models are not independent due to the correlation induced by the random effects and second, for the selection of random effects one has to deal with a non-open parameter space because of the restrictions on the variance parameters.

Greven and Kneib (2010) showed that the AIC which is based on the marginal model formulation is not an asymptotically unbiased estimator for the Akaike information. As no bias correction can be made, the marginal AIC (mAIC) is not an appropriate criterion for the selection of random effects in mixed models.

For the linear mixed model (LMM), Vaida and Blanchard (2005) proposed an estimator based on the conditional model formulation for the case of known variance parameters. Given that in practice the variance components are unknown, they suggested using a plug-in estimator of the covariances of the random effects. However, Greven and Kneib (2010) showed that ignoring the uncertainty in the estimation of the covariances of the random effects leads to a particular bias, i.e. the more complex model is always favored unless the covariance of the random effect is estimated to be exactly zero. A numerical correction of the conditional AIC (cAIC) has been proposed by Liang et al. (2008). It accounts for the estimation of the random effects components by adjusting the penalty term of the conditional AIC of Vaida and Blanchard (2005). Yet, this *approximate* cAIC turned out to be computationally very expensive and the costs even increase with sample size. In order to avoid this drawback, Greven and Kneib (2010) developed an analytic representation of the corrected version of the cAIC.

All these estimators (the uncorrected, the approximate and the analytic cAIC) are only applicable in the case of normal errors. The considerations become more complex for generalized linear mixed models (GLMMs) as inference in the GLMM is more challenging. This is due to the fact that the marginal likelihood is generally not analytically accessible and approximations have to be made.

The objective of this thesis is to compare two different approaches on an extension to generalized linear mixed models. We examined a criterion of Yu and Yau (2011) who provided an asymptotically unbiased estimator of the conditional Akaike information. This criterion has been constructed only under maximum likelihood estimation and not under restricted maximum likelihood estimation. Furthermore, we considered a bias correction term based on a covariance penalty which has been suggested in the context of the estimation of prediction errors by Efron (2004) and we applied it to the mixed model framework.
We conducted two simulation studies in order to investigate the behavior of these two generalized approaches in the special case of linear mixed models. Comparing the covariance based cAIC and the cAIC of Yu and Yau to the uncorrected, the approximate and the analytic cAIC allowed us to asses the performance of the cAIC of Yu and Yau (2011) and the covariance based cAIC of Efron (2004). The first simulation study is based on penalized spline smoothing, the second uses random intercept models. In addition, all criteria were applied in a case study on childhood malnutrition in order to illustrate the practical relevance of the topic.

This work is structured as follows. In a first part, comprising of Chapter 2-4, the theoretical background for this work will be provided. Specifically, Chapter 2 will give an introduction to model selection and conclude by the derivation of the Akaike information criterion. Linear mixed models and generalized linear mixed models will be the subject of Chapter 3, including inferential properties and implementational aspects. Chapter 4 will cover penalized spline smoothing and will relate it to the topic of mixed models.
Chapter 5 – as a second part – will then bring together Chapter 2 and Chapter 3 by elaborating on the AIC in mixed models. In this context, we will introduce all Akaike information criteria which will be considered in the simulation studies and relate them to each other. Moreover, different representations of the cAIC of Yu and Yau and details on the estimation of the covariance based cAIC will be provided.
Building on this, the third part – consisting of Chapters 6 and 7 – will cover the main work of this thesis. The two simulation studies on the behavior of the various cAICs will be presented in Chapter 6, followed by the application of the cAICs to real data in Chapter 7.
The thesis will finish with further considerations in Chapter 8 and a conclusion in Chapter 9.

Note that complete results of the two simulation studies and of the case study can be found in the appendix. Furthermore, many proofs and derivations are given there as well. Descriptions of the most important estimation algorithms and the explanation of the bootstrap algorithms used for the computation of the covariance based cAIC are also included. The appendix comprises in addition descriptions of the main `R`-functions used in the simulations and an overview of the attached `R`-code on disc.

Chapter 2

Model Selection

Model selection comprises several aspects. First, a class of models has to be chosen. This includes making assumptions on the response variable (e.g. distribution) as well as specifying the type of influence which the covariates are assumed to exert on the response. Second, building a model requires variable selection (for a given model class).

Regarding this, theoretically two alternative perceptions are possible: For model selection one can either assume that the "truth", i.e. the "reality", can only be described by an infinite number of parameters. One would therefore carry out model selection by comparing models using their *relative* goodness. Alternatively, one assumes that the "reality" can be reflected by a finite number of parameters which would make it possible to consider their *absolute* performances. The first approach does not aim to find the "truth" as this is not thought possible[1]. Instead, one intends to develop the best approximating model, keeping in mind the concept of parsimony (lat. parsimonia, to save, see Section 2.1). In contrast, the second perspective assumes it to be principally possible to detect the "true model".

It should be kept in mind that in real data analysis usually a set of candidate models is available which can be compared (relative perspective) by the investigator in order to find the best approximation to the "truth" among these candidates. Thus, models not being in the set remain unconsidered in the selection of the best approximating model.

There are various possibilities to accomplish model selection, ranging from testing, shrinkage approaches (e.g. Lasso (Tibshirani, 1996)) and the selection based on (estimated) prediction errors (e.g. Cross-Validation (Kurtz, 1948)) to the selection on the basis of information criteria.[2] The focus in this work will be on the latter, more precisely, on model selection based on the Akaike information criterion.

[1] "Truth is elusive" (DeLeeuw, 1988).

[2] For an overview, see Heumann et al. (2010).

2.1 Principle of Parsimony

As mentioned in the previous section, the objective of model selection is to find the best approximating model with due regard to the principle of parsimony. More precisely, as any model can be improved (in the sense of being closer to "reality") by taking additional parameters into account, the question arises when to stop making the model more complex (in practice). Therefore, model selection is always a question of model complexity, and is thus a matter of bias-variance trade-off which is the "statistical principle of parsimony" (Burnham and Anderson, 2002).

"Everything should be made as simple as possible, but no simpler"[3]

Introducing too large a number of parameters into a model will result in a large-sized variance, but a small bias. On the contrary, if a model is of too low complexity, it tends to have a great bias, although a small variance. It is therefore essential to find a compromise between these two scenarios and thus to prevent under- as well as overfitting.

"Parsimony lies between the evils of under- and overfitting"[4]

2.2 Information Theory and The Kullback-Leibler Distance

The following section will give an introduction on information theory and in particular on the Kullback-Leibler distance which is an essential component in the derivation of the Akaike information criterion.

Information theory is a mathematical discipline dealing with the quantification of information in general. Modern information theory was initiated by Shannon (1948) whose paper "A Mathematical Theory of Communication" started the field in the middle of the 20th century. Since its inception, the list of applications of the concepts and methods of information theory has become endless and represents a point of intersection of many scientific disciplines such as physics, economics, communication theory, and statistics (Cover and Thomas, 1991).

Motivated to provide a rigorous definition of "information" (in relation to Fisher's criterion of *sufficiency*[5] (Fisher, 1922)), Kullback and Leibler (1951) introduced a measure of the discrepancy between two probability distributions. This measure will be presented

[3] Attributed to Albert Einstein

[4] Burnham and Anderson (2002)

[5] Fisher's criterion required that "the statistic chosen should summarize the whole of the relevant information supplied by the sample" (Fisher, 1922).

in the following based on Chapter 2 and Chapter 6 in Burnham and Anderson (2002), as it forms the basis of the definition of the Akaike information criterion.

Consider two models f and g. In the following, g will denote the *"truth"* – meaning the true underlying (possibly very complex) process which generates the data z. Model f is the approximating model in terms of a probability distribution.

In the case of continuous functions, the Kullback-Leibler distance (KLD) is defined as follows:

Definition 1. *Kullback-Leibler Distance (Kullback-Leibler Information)*

$$KLD(g,f) = \int_{\mathbb{R}} g(z)\ log \left\{ \frac{g(z)}{f(z)} \right\}\ dz. \tag{2.1}$$

Here, and in the rest of this thesis, $log(\cdot)$ denotes the natural logarithm function (compare the list of abbreviations and symbols in Appendix F). In this work, we will only consider the case of continuous functions. For the definition of the Kullback-Leibler distance for discrete functions and for examples of Kullback-Leibler distances for different distributions, see Burnham and Anderson (2002).

The Kullback-Leibler distance between the models g and f measures the **directed distance from the approximation $\boldsymbol{f}$ to the "truth" $\boldsymbol{g}$**. Note that this *directed* distance does not satisfy the symmetry assumption of an ordinary distance function as $KLD(g,f)$ is not equal to $KLD(f,g)$. The roles of the "truth" g and its approximation f are thus not the same. Alternatively, the KLD can be interpreted as the **loss of information when model $\boldsymbol{f}$ is used to approximate $\boldsymbol{g}$**, which is why it is often denoted as Kullback-Leibler *information.*

Some important properties of the Kullback-Leibler distance should be noted:

1. The KLD is always non-negative: $KLD(g,f) \geq 0$.
2. The KLD is zero iff the approximating model corresponds to the truth: $KLD(g,f) = 0 \Leftrightarrow f = g$ (almost everywhere).
3. The KLD is not only based on the first two moments of a distribution (mean and variance), but on the entire distribution.
4. Adding parameters to the model f will always decrease the distance to the true underlying process (Burnham and Anderson, 2002).

For model selection, the aim clearly is to find an approximating model for which the loss of information is the smallest possible. Thus, one seeks to minimize the $KLD(g,f)$ over f which varies over the space of models indexed by ψ, whereas the "truth" is assumed to be given (fixed).

It can easily be seen that calculating the KLD involves knowing both the truth g as well as the probability distribution f (including their parameters ψ). However, this requirement is reduced when only the *relative* directed distances are used, since the KLD of g and f can be rewritten as

$$\begin{aligned} KLD(g, f) &= \int_{\mathbb{R}} g(z)\ log \left\{ \frac{g(z)}{f(z)} \right\}\ dz \\ &= \underbrace{\int_{\mathbb{R}} g(z)\ log(g(z))\ dz\}}_{constant} - \int_{\mathbb{R}} g(z)\ log(f(z))\ dz. \end{aligned} \tag{2.2}$$

The first term on the right of the expression is a constant depending only on the unknown "truth". As the constant is the same across all candidate models, no assumptions have to made for g and the interest lies in the second term which can be expressed as

$$\int_{\mathbb{R}} g(z)\ log(f(z))\ dz = E_g\left[log(f(z|\psi))\right]. \tag{2.3}$$

It is thus a statistical expectation with respect to g.

Note that – in contrast to the KLD itself – the quantity of interest here, $E_g\left[log(f(z|\psi))\right]$, is on an interval scale which lacks a true zero. This implies that the "difference, ..., means the same thing anywhere on the scale"[6].

So far, no parameter estimation has been introduced into the concept of selecting an approximating model. However, in real data analysis, the parameters ψ are unknown and have to be estimated from the data. Thus, one needs **estimates** of the relative distances between the unknown "truth" that generated the data and the candidate models $f_i(z|\hat{\psi})$, $i = 1, \ldots, M$, with M being the number of approximating models available and $\hat{\psi}$ denoting the estimator of ψ. (Note that the hat notation for estimated quantities will be used throughout this work.)
Knowing the estimated relative directed distances, the "best" (in terms of closest to the "truth") candidate model can be chosen without knowing the "truth" g. This is where Akaike (1983) comes into play. He found a way to estimate the relative KLD, based on the log-likelihood function at its maximum point which allowed "major practical and theoretical advances in model selection and the analysis of complex data sets"[7]. This will be the subject of the following section.

[6] Burnham and Anderson (2002)

[7] See Stone (1982), DeLeeuw (1992), and Bozdogan (1987).

2.3 The Akaike Information Criterion

The Akaike information criterion (AIC) is a model selection criterion based on information theory (see Section 2.2), more precisely, based on the Kullback-Leibler distance (Definition 1). It will be shown in the following sections that the AIC does not only have an interpretation in the context of the trade-off between bias and variance or the trade-off between under- and overfitting, but also provides a theoretical basis for model selection. Akaike (1973) succeeded in finding a relationship between the (relative) Kullback-Leibler distance and the maximum likelihood function (denoted as $\mathcal{L}(\cdot)$) and therefore in relating information theory with the maximum likelihood principle.

As mentioned in the previous section, the parameters ψ are usually not known in real data analysis, which is why one needs estimates for the (relative) directed distances between the underlying "truth" g and the candidate models $f_i(z|\hat{\psi})$, $i = 1, \ldots, M$ in order to select the "best" model. Based on Chapter 2 in Burnham and Anderson (2002), it will be described in the following how Akaike (1983) found an applied Kullback-Leibler model selection criterion.

Consider a parametric model $f(z|\psi)$ and denote the unique minimizer of the Kullback-Leibler distance as

$$\psi_0 = \arg\min_{\psi} KLD(g, f). \tag{2.4}$$

As the KLD-minimizer depends on the "truth" g, ψ_0 is an unknown quantity. It can be seen as the absolutely best value of ψ for the approximating model f. If ψ_0 was known, the maximum likelihood estimator $\hat{\psi}$ would estimate ψ_0, i.e. it is the "true" value of underlying maximum likelihood estimation. This is an important characteristic feature of $f(z|\psi_0)$ in the derivation of the AIC. Burnham and Anderson especially pointed out that, due to the fact that in reality models are based on estimated parameters rather than on known parameters, the model selection criterion is to minimize the *expected* estimated KLD instead of the *known* KLD over the set of candidate models (see Subsection 2.3.1).

Let y and z be two independent random samples from the same distribution (the "truth"). The critical issue for deriving an applicable model selection criterion based on the KLD (an issue which Burnham and Anderson called the *selection target*) is to find an (asymptotically unbiased) estimator of

$$E_y E_z \left[log \left(f(z|\hat{\psi}(y)) \right) \right]. \tag{2.5}$$

Note that -2 this quantity is often referred to as the *Akaike information*:

Definition 2. *Akaike Information*

$$-2 \; E_y E_z \left[log \left(f(z|\hat{\psi}(y)) \right) \right]. \tag{2.6}$$

Burnham and Anderson called it “tempting” to just estimate the quantity (2.5) by the maximized log-likelihood, but made clear that this would lead to an upwards biased estimator of the Akaike information (AI). Therefore, in order to obtain an asymptotically unbiased estimator of the AI, a bias correction (BC) is needed. Akaike showed that under certain conditions (see 2.3.1) the bias is approximately equal to the number of estimable parameters in the candidate model f. Thus, an asymptotically unbiased estimator for the quantity (2.5) is

$$log\left\{\mathcal{L}(\hat{\psi}|data)\right\} - k, \tag{2.7}$$

which is equivalent to

$$constant - \hat{E}_{\hat{\psi}}\left[KLD(g,\hat{f})\right],$$

where $\mathcal{L}(\hat{\psi}|data)$ denotes the likelihood function at its maximum point, $\hat{f}$ abbreviates $f(\cdot|\hat{\psi})$, k is the number of parameters in the model f and $\hat{E}_{\hat{\psi}}\left[KLD(g,\hat{f})\right]$ is the estimate of the expected relative KLD.

What makes Akaike’s work so important for model selection in statistical analysis is the new-found relation between the expected relative Kullback-Leibler distance and the maximized log-likelihood. The close connection of the AIC to maximum likelihood methods is “to many statisticians [...] still the ultimate in terms of rigor and precision”[8].
For historical reasons[9], Akaike multiplied the whole expression (2.7) by -2. This finally leads to the model selection criterion known as the AIC[10]:

Definition 3. *Akaike Information Criterion*

$$AIC = -2\ log\left(f(y|\hat{\psi}(y))\right) + 2k \tag{2.8}$$

The model with the smallest AIC among the candidate models is chosen.

2.3.1 Formal Derivation of the AIC

Although a brief outline of the derivation of the AIC has been given in the previous section, a more formal illustration will be supplied now. It is based on Chapter 7 in Burnham and Anderson (2002). This will inter alia allow to better understand the origin of the selection target (2.5). It should be noted that “there is no unique path from K-L [Kullback-Leibler] to AIC”[11] and it has been motivated, justified and derived in a variety of ways.

[8] DeLeeuw (1992)

[9] E.g. that -2 the logarithm of the ratio of two maximized likelihood values is asymptotically chi-squared.

[10] AIC was originally the abbreviation for an information criterion (Burnham and Anderson, 2002).

[11] Burnham and Anderson (2002)

The notation in this section will stay the same as before, all expectations are taken with respect to the underlying "truth" g. z and y denote independent random samples arising from the underlying "truth".

Consider again the parametric model $f(z|\psi)$ and denote ψ_0 as the minimizer of the $KLD(g, f(z|\psi))$. Therefore, $f(\cdot|\psi_0)$ is the best approximating model to the "truth".
The Kullback-Leibler distance itself does not involve any data, as z is integrated out. Given the data y, a natural possibility to estimate the $KLD(g, f(\cdot|\psi_0))$ is the computation of

$$KLD(g, f(\cdot|\hat{\psi}(y))) = \int_{\mathbb{R}} g(z)\ log \left\{ \frac{g(z)}{f(z|\hat{\psi}(y))} \right\} dz, \tag{2.9}$$

with $\hat{\psi}(y)$ being the maximum likelihood estimator of ψ based on the data y.

If the minimizer ψ_0 was known,

$$KLD(g, f) = 0 \tag{2.10}$$

would be satisfied and it would be possible to compare the performance of alternative models to this absolute value of zero. However, since ψ_0 is an unknown quantity, only the estimate $\hat{\psi}(y)$ is available and it holds that

$$KLD(g, f(\cdot|\hat{\psi}(y))) > KLD(g, f(\cdot|\psi_0)), \tag{2.11}$$

unless $\hat{\psi}(y) = \psi_0$.

Because the Kullback-Leibler minimizer ψ_0 is not known in reality, the idea of what the target should be has to be revised. One would expect (in the frequentistic context of repeated sample properties) that the estimated KLD has on average a value of $E_y \left[KLD(g, f(\cdot|\hat{\psi}(y))) \right]$.
Thus, instead of minimizing the (unknown) quantity $KLD(g, f(\cdot|\psi_0))$, the aim is now to minimize the (slightly larger value) $E_y \left[KLD(g, f(\cdot|\hat{\psi}(y))) \right]$. Note that the large-sample difference

$$E_y \left[KLD(g, f(\cdot|\hat{\psi}(y))) \right] - KLD(g, f(\cdot|\psi_0)) = \frac{1}{2}\ tr \left\{ \boldsymbol{J}(\psi_0) \boldsymbol{I}(\psi_0)^{-1} \right\} \tag{2.12}$$

is independent of the sample size n. Here, and in the rest of this work, $tr(\cdot)$ denotes the trace of a matrix. $\boldsymbol{J}(\psi_0)$ and $\boldsymbol{I}(\psi_0)$ are given as

$$\boldsymbol{J}(\psi_0) = E_g \left[\left[\frac{\partial}{\partial \psi} log\left(f(z|\psi)\right) \right] \left[\frac{\partial}{\partial \psi} log\left(f(z|\psi)\right) \right]^T \right] \Bigg|_{\psi=\psi_0} \tag{2.13}$$

$$\boldsymbol{I}(\psi_0) = E_g \left[-\frac{\partial log\left(f(z|\psi)\right)}{\partial \psi_i \partial \psi_j} \right] \Bigg|_{\psi=\psi_0}. \tag{2.14}$$

Thus, given that ψ must be estimated, the target is now

$$\text{“to select model } f \text{ to minimize } E_y\left[KLD(g, f(\cdot|\hat{\psi}(y)))\right]\text{”}^{\,12}.$$

One can show that $E_y\left[KLD(g, f(\cdot|\hat{\psi}(y))\right]$ can be expressed as[13]

$$E_y\left[KLD(g, f(\cdot|\hat{\psi}(y)))\right] = constant - E_y E_z\left[log\left(\hat{f}(z)\right)\right]. \tag{2.15}$$

One concentrates on this double expectation which has already been introduced as the selection target in the previous section (see (2.5)). The new quantity of interest will further be denoted as

$$T := \int_{\mathbb{R}} g(y)\left[\int_{\mathbb{R}} g(z)\; log\left(f(z|\hat{\psi}(y))\right)dz\right]dy. \tag{2.16}$$

The target is to unbiasedly estimate T in order to obtain an applicable selection criterion. Note that only relative values can be obtained for $E_y\left[KLD(g, f(\cdot|\hat{\psi}(y)))\right]$ as the constant cannot be determined (Heumann et al., 2010).

Having specified the model selection target T, two steps have to be taken in order to obtain the relationship to the maximized log-likelihood.

Step 1 First, a second-order Taylor expansion is applied to $log\left(f(z|\hat{\psi})\right)$ around ψ_0 (for any given z)

$$\begin{aligned} log\left(f(z|\hat{\psi})\right) \approx{}& log\left(f(z|\psi_0)\right) + \left[\frac{\partial log\left(f(z|\psi)\right)}{\partial\psi}\right]^T\Bigg|_{\psi=\psi_0}\left[\hat{\psi}-\psi_0\right] \\ &+\frac{1}{2}\left[\hat{\psi}-\psi_0\right]^T\left[\frac{\partial^2 log\left(f(z|\psi)\right)}{\partial\psi^2}\right]\Bigg|_{\psi=\psi_0}\left[\hat{\psi}-\psi_0\right]. \end{aligned} \tag{2.17}$$

In order to relate the result to the target T (2.16), the first expectation with respect to z is taken. Because of

$$E_z\left[\frac{\partial log(f(z|\psi))}{\partial\psi}\right]\Bigg|_{\psi=\psi_0} = 0, \tag{2.18}$$

the linear term of the expansion vanishes. Then, the second expectation is taken with respect to y, yielding

$$T = E_y E_z\left[log\left(f(z|\hat{\psi})\right)\right] \tag{2.19}$$

$$\approx E_z\left[log(f(z|\psi_0))\right] - \frac{1}{2}\,tr\left\{\boldsymbol{I}(\psi_0)E_y\left[\left[\hat{\psi}-\psi_0\right]\left[\hat{\psi}-\psi_0\right]^T\right]\right\} \tag{2.20}$$

$$= E_z\left[log(f(z|\psi_0))\right] - \frac{1}{2}\,tr\left\{\boldsymbol{I}(\psi_0)\boldsymbol{\Sigma}\right\}, \tag{2.21}$$

[12] Burnham and Anderson (2002)

[13] See for the proof Appendix A.

with $\mathbf{\Sigma}$ the correct large-sample theoretical sampling variance of the maximum likelihood estimator.

Step 2 As Step 1 still not establishes a relation between T and the expected maximized log-likelihood $E_z\left[log\left(f(z|\hat{\psi}(z))\right)\right]$, a second Taylor expansion is carried out, this time of $log(f(z|\psi_0))$ around $\hat{\psi}(z)$, where z is treated as sample data. Note that $\hat{\psi}$ abbreviates $\hat{\psi}(z)$ in the following.

Since the aim is to obtain an expectation, it is possible to switch between z and y and the expectations from above can be interchanged due to the independence of z and y.

This leads to

$$log(f(z|\psi_0)) \approx log\left(f(z|\hat{\psi})\right) + \left[\frac{\partial log(f(z|\psi))}{\partial \psi}\right]^T \Bigg|_{\psi=\hat{\psi}} \left[\psi_0 - \hat{\psi}\right] \tag{2.22}$$

$$+\frac{1}{2}\left[\psi_0 - \hat{\psi}\right]^T \left[\frac{\partial^2 log(f(z|\psi))}{\partial \psi^2}\right] \Bigg|_{\psi=\hat{\psi}} \left[\psi_0 - \hat{\psi}\right]. \tag{2.23}$$

Because the maximum likelihood estimator $\hat{\psi}$ satisfies

$$\frac{\partial log(f(z|\psi))}{\partial \psi}\Bigg|_{\psi=\hat{\psi}} = 0, \tag{2.24}$$

the linear term of the expansion vanishes. Taking the expectation with respect to z then yields

$$E_z\left[log(f(z|\psi_0))\right] \approx E_z\left[log\left(f(z|\hat{\psi})\right)\right] - \frac{1}{2}\, tr\left\{E_z\left[\hat{\boldsymbol{I}}(\hat{\psi})\right]\left[\psi_0 - \hat{\psi}\right]\left[\psi_0 - \hat{\psi}\right]^T\right\}, \tag{2.25}$$

where $\hat{\boldsymbol{I}}(\hat{\psi})$ is the Hessian of the log-likelihood evaluated at the maximum likelihood estimator

$$\hat{\boldsymbol{I}}(\hat{\psi}) = -\frac{\partial^2 log(f(z|\psi))}{\partial \psi^2}\Bigg|_{\psi=\hat{\psi}}. \tag{2.26}$$

In the following, several approximations are made, which will be presented here without many details. For more details see Burnham and Anderson (2002).

First, $\hat{\boldsymbol{I}}(\hat{\psi})$ is approximated by $\boldsymbol{I}(\psi_0)$ (this approximation improves with growing sample size) in order to make analytical progress. This leads to

$$E_z\left[\hat{\boldsymbol{I}}(\hat{\psi})\right]\left[\psi_0 - \hat{\psi}\right]\left[\psi_0 - \hat{\psi}\right]^T \approx \boldsymbol{I}(\psi_0)\mathbf{\Sigma}. \tag{2.27}$$

Substitution of the result of Step 1 into the resulting

$$E_z\left[log(f(z|\psi_0))\right] \approx E_z\left[log\left(f(z|\hat{\psi}(z))\right)\right] - \frac{1}{2}\, tr\left\{\boldsymbol{I}(\psi_0)\boldsymbol{\Sigma}\right\} \tag{2.28}$$

gives

$$T \approx E_z\left[log\left(f(z|\hat{\psi}(z))\right)\right] - tr\left\{\boldsymbol{I}(\psi_0)\boldsymbol{\Sigma}\right\}.^{14} \tag{2.29}$$

Therefore, an asymptotically unbiased estimator of the target T is provided by

$$\hat{T} \approx log\left(f(z|\hat{\psi})\right) - \hat{tr}\left\{\boldsymbol{I}(\psi_0)\boldsymbol{\Sigma}\right\}. \tag{2.30}$$

The first term of this approximation is an unbiased estimator of its own expectation $E_z\left[log\left(f(z|\hat{\psi})\right)\right]$ (but a biased estimator for T. It thus needs the second term as a bias correction). $\boldsymbol{\Sigma}$ is unknown and cannot be directly[15] estimated from one sample, because only one $\hat{\psi}$ is available. Thus, it remains to find an estimator of the trace term which possibly has no or low bias.

If the "truth" g is equal to f or nested in f, than the trace term simplifies to

$$tr\left\{\boldsymbol{I}(\psi_0)\boldsymbol{\Sigma}\right\} = k, \tag{2.31}$$

with k the number of parameters to be estimated in the approximating model. Even if f is just a good approximation for g, it is advised to take

$$\widehat{tr}\left\{\boldsymbol{I}(\psi_0)\boldsymbol{\Sigma}\right\} = k \tag{2.32}$$

as approximator for the trace term (for more information on the estimation of the trace term see Burnham and Anderson (2002)).

With these two approximations and the multiplication of all terms by -2, this finally yields the so-called Akaike information criterion

$$AIC = -2\; log\left\{\mathcal{L}(\hat{\psi}|data)\right\} + 2k.$$

Other approaches have been made for the estimation of the trace term. For example, Takeuchi (1976) generalized the Akaike information criterion for cases where g is not a subset of f by suggesting bootstrap methods for the estimation of the trace terms.

[14] In the literature the alternative trace term $tr\left\{\boldsymbol{J}(\psi_0)\boldsymbol{I}(\psi_0)^{-1}\right\}$ is often presented.

[15] Bootstrapping (invented by Efron (1979)) would be a solution.

2.3.2 Properties of the AIC

Some important properties of the AIC should be mentioned. First, it should be pointed out that the AIC is a relative criterion, meaning that candidate models can be compared via their AICs but no absolute AIC value has a reasonable interpretation. Second, the AIC strongly depends on sample size as the bias correction term k is an asymptotic correction which tends to be closer to the approximated trace term (in equation (2.28)) in the case of large sample sizes. Third, it should be noted that the response variable has to be the same in all candidate models. No transformations of the response are admitted for the comparison of the AICs of different models because the inference is conditional on the data ("Data must be fixed"[16]). Fourth, the comparison of models with different probability distributions requires that all components of the log-likelihoods are retained.

2.3.3 The AIC and hypothesis testing

Although hypothesis testing will not be introduced and further discussed in this work, it seems to be of great importance to briefly point out the differences of comparing models via their AICs and using tests in order to perform model selection. For more details on hypothesis testing and especially on the likelihood ratio test and its applicability in mixed models see Greven (2008) and Burnham and Anderson (2002).

It is important to make clear that an information criterion is not a test, thus does not provide p-levels and does not allow significance conclusions. The main advantages of the AIC compared to hypothesis tests are[17]:

1. The AIC is free from arbitrary choices of α-levels and from multiple testing problems.
2. The AIC allows ranking of models whereas hypothesis testing does not provide a general way to rank models, even not for nested models.
3. The AIC can be used to compare non-nested models and can be applied to the comparison of different distributions.
4. The AIC has a theoretical basis whereas the likelihood ratio test does not.

[16] Burnham and Anderson (2002)

[17] Burnham and Anderson (2002)

2.3.4 Heuristical interpretation

Akaike's information criterion allows for an interesting heuristical interpretation.[18] Before it will be given here, it should be noted that although this explanation is quite common among users, there is a deeper theoretical basis for the AIC as shown above. However, the "heuristical" approach is very intuitive and emphasizes clearly the bias-variance trade-off.

The first term of the AIC, $-2\ log\left(f(y|\hat{\psi}(y))\right)$, can be interpreted as a measurement of the lack of model fit. It tends to decrease as more parameters are added to the approximating model f, while he second term, $2k$, gets larger as more parameters are added. The latter constitutes a "penalty" for increasing the size of the model, i.e. taking more parameters into account. This penalty leads to the compliance with the principle of parsimony (Section 2.1).

[18] Burnham and Anderson (2002)

Chapter 3

Mixed Models

3.1 The Linear Mixed Model

3.1.1 The Linear Model

Consider the standard linear model (LM) in which the relation between the metric response variable y and the covariates $x_1, \ldots, x_p$ is assumed as follows

$$y = x^T \beta + \varepsilon, \tag{3.1}$$

with $x = (1, x_1, \ldots, x_p)^T$, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ and ε a probabilistic error term.
The response variable can therefore be decomposed into a deterministic part $x^T\beta$ and some kind of stochastic dispersion around this conditional mean, ε. The deterministic part is called the *linear predictor* η which equals for the linear model the conditional mean of y for given covariates $x_1, \ldots, x_p$, denoted as $E(y|x)$.

In order to estimate the regression parameters $\beta_0, \beta_1, ..., \beta_p$ and thus to specify the influence of the covariates on the response, n independent measurements are taken, leading to the data $y_i, x_{i1}, ..., x_{ip}$ $(i = 1, \ldots, n)$.

Altogether, the model can be formulated as

$$y_i = x_i^T \beta + \varepsilon_i, \text{ for } i = 1, \ldots, n. \tag{3.2}$$

Alternatively, the linear model can be written in matrix formulation as

$$y = \mathbf{X}\beta + \varepsilon, \tag{3.3}$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}, \ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{3.4}$$

The model relies on the following assumptions:

1. The model describes the "true" relationship between the design matrix $\boldsymbol{X}$ and the response variable y, except for the error term. This means the relationship is of linear nature.

2. The expectations of the probabilistic error terms are zero. This implies that there is no systematic error in the model.

$$E\left(\varepsilon\right) = E\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix} = \begin{pmatrix}0\\ 0\\ \vdots\\ 0\end{pmatrix} \tag{3.5}$$

3. The covariance of the error terms is

$$Cov(\varepsilon) = \sigma^2\boldsymbol{I}_n,\ \ (\sigma \geq 0), \tag{3.6}$$

   with $\boldsymbol{I}_n$ denoting the $n \times n$ identity matrix. The error terms are thus independent and identically distributed (i.i.d.).

4. An optional assumption concerns the distribution of the error terms. It can be necessary to specify the distribution of the error terms, e.g. in order to use maximum likelihood methods, to conduct hypothesis testing, or to construct confidence intervals.

   One usually assumes (in the case of metric response variables)

$$\varepsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2\boldsymbol{I}_n). \tag{3.7}$$

For more details see Fahrmeir et al. (2007) and Kneib (2003)

3.1.2 Motivation of the Linear Mixed Model

In many situations, the assumptions of the standard linear model are too restrictive and generalizations are needed. One way to extend the linear model is to allow for random effects besides the fixed effects $\beta_0, \ldots, \beta_p$. The resulting model is referred to as the *linear mixed model*(LMM) (or *linear mixed effects model*). It will be motivated and introduced in the following.

There are several ways to motivate the linear mixed model. One is to consider the case of longitudinal or cluster data which will be illustrated in the following based on Konrath (2009).

Longitudinal studies are a widely used study design in e.g. medical research. The basic concept is that repeated measurements are taken of the same subjects over a period of time. The resulting data for each subject or individual has the form

$$y_{i1}, \dots, y_{ij}, \dots, y_{iJ_i},\ x_{i1}, \dots, x_{ij}, \dots, x_{iJ_i},\ \text{for } i = 1, \dots, N,\ j = 1, \dots, J_i,$$

with J_i the number of observations for individual i and N the number of individuals.

To give an example, consider a medical study where the blood pressure of $N = 100$ patients is measured under differing conditions over time. Let y_{ij} be the blood pressure of patient i at measure point j (time t_{ij}) ($i = 1, \dots, N$, $j = 1, \dots, J_i$).
The design may be unbalanced, i.e. the measurements are not necessarily taken at the same points of time and even the number of measurements can differ from subject to subject.

If instead observations are made along a cross-sectional design, where subjects are chosen from clusters – in the given example for instance hospitals – and observed only once, the resulting data is referred to as cluster data. Cluster data formally has the same structure as longitudinal data with the difference that y_{ij} denotes the value of the response variable (e.g. blood pressure) for subject j from cluster i.

It seems to be obvious that repeated measurements of one and the same subject, or the observations of subjects from the same cluster, are more alike than those between different subjects/clusters. Thus, the interesting aspect of these kinds of data is the **correlation** which is implied.

In order to analyze longitudinal/cluster data one has to be aware of the fact that there are **two sources of variability** in the data. First, due to the repeated measurements variability arises **within** the data corresponding to one subject/cluster. Second, there is variability **between** different subjects/clusters, i.e. the discrepancy from the population mean.

The aim of using mixed models is to estimate the effects of the covariates on the response variable y with respect to the contemplated correlation structure in the data. Depending on the question, the interest lies either more in the subject-specific effects or in the population-specific effects. In medical studies, for example, the subject-specific effects are often of great interest, as one aim is to make predictions for the development (of e.g. blood pressure) for each patient. Apart from the effects, the correlation structure gives insight into the data and is therefore also an object of interest.

In order to demonstrate why the standard linear model as described above (Section 3.1.1) is not adequate for the analysis of longitudinal/cluster data, the possibilities to apply the LM in such a situation are considered in the following.

Recall the longitudinal data example from above, where the blood pressure of N patients is measured over a period of time. Let the patients now be partitioned into m groups of different treatments. The focus then lies on:

1. the treatment-specific effects,
2. the subject-specific effects, and
3. the correlation structure.

The first possibility consists of applying **separate linear models for each treatment group**. In this case, the regression parameters only vary with the different treatments. Yet, this does not allow any insight neither into the subject-specific effects nor in the correlation structure. By fitting m separate models, it is only possible to learn something about the **effect of the treatments**.

A second option would be to fit N **separate linear regression models – one for each individual**. Here, the parameters vary for each individual but not for treatment groups. However, besides the expense of estimating N models and the fact that the number of observations may be too small to get reliable estimations, the regression model parameters only describe the **subject-specific effects** and do not cover any population-specific aspects. Moreover, the correlation sprouting from the repeated measurements is still not taken into account.

In order to incorporate the correlation structure, a general linear model for all individuals with special assumptions on the error term is possible.

Such a model can be written as

$$y_{ij} = \eta_{ij} + \varepsilon_{ij},\ i = 1, \ldots, N,\ j = 1, \ldots, J_i. \tag{3.8}$$

One assumes **independent** ε_i $(i = 1, \ldots, N)$, i.e. the individuals are assumed to be independent, but allows dependence within each individual:

$$\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iJ_i})^T \sim \mathcal{N}(0, \mathbf{\Sigma}_i)\ i = 1, \ldots, N.$$

The estimation of the model parameters is carried out by applying a generalized (weighted) least-squares criterion (see Fahrmeir et al. (2007)).
Here, the correlation within each individual is taken into consideration by dropping the assumption of i.i.d. error terms. However, without any further specification of $\mathbf{\Sigma}_i$ $(i = 1, \ldots, N)$, the number of parameters that have to be estimated is very high and increases with the number of observations $n = \sum_{i=1}^{N} J_i$. Furthermore, the linear predictor η_{ij} can either be specified to provide individual **or** treatment effects (not both at the same time).

These approaches show the need to extend the linear model in order to achieve a comprehensive analysis of the given data.
A further approach consists in treating the data with a two-stage analysis consisting of one stage specifying separate linear models for each subject in order to describe the individual profiles and a second stage in which knowledge from Stage 1 is used to explain the

variability between the different subjects. This approach will lead us to the linear mixed model.

In the case of m treatment groups and one covariate x_{ij} (e.g. age), the model has the following form:

Stage 1

$$y_{ij} = \beta_{0i} + \beta_{1i} \cdot x_{ij} + \varepsilon_{ij}, \text{ with } \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; (i = 1, \ldots, N).$$

Stage 2

$$\begin{aligned} \beta_{0i} &= \beta_0 + b_{0i} \\ \beta_{1i} &= \beta_1 \cdot Gr_{1i} + \ldots + \beta_m \cdot Gr_{mi} + b_{1i}, \end{aligned}$$

with $b_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{D} = \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{10} & \tau_1^2 \end{pmatrix}\right)$, with $\tau_0,\ \tau_{01},\ \tau_{10},\ \tau_1$ all ≥ 0 and

Gr_{gi}: Indicator variable for the treatment group g for subject i, $g = 1, \ldots, m$.

Thus, in the second stage the subject-specific coefficients are linked to the treatment groups which allows:

1. the estimation of the mean population-specific response

$$\beta_0 \text{ at time } t_{ij} = 0 \; (i = 1, \ldots, N,\ j = 1, \ldots, J_i),$$

2. the estimation of the mean treatment-specific slopes

$$\beta_1, \ldots, \beta_m,$$

3. the estimation of the individual discrepancies of the population mean

$$\beta_{0i} = \beta_0 + b_{0i} \; (i = 1, \ldots, N),$$

4. the estimation of the individual discrepancies of the treatment slopes

$$\beta_{1i} = \beta_1 \cdot Gr_{1i} + \ldots + \beta_m \cdot Gr_{mi} + b_{1i} \; (i = 1, \ldots, N), \text{ and}$$

5. to take the covariances between the individual effects into account by specifying the components τ_{01} and τ_{10} of the covariance $Cov(b_{0i}, b_{1i})$ $(i = 1, \ldots, N)$.

While the population- and treatment-specific effects are modeled as deterministic (fixed) unknown parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_m)^T$ like in the linear regression model (3.1), the main difference lies in the assumption of **random subject-specific effects** $\beta_i = (\beta_{0i}, \beta_{1i})^T$ $(i = 1, \ldots, N)$.
The assumption $b_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{D})$ $(i = 1, \ldots, N)$ implies that the population means are already included in the fixed effects. The variances τ_0^2 and τ_1^2 indicate how much the individual specific effects disperse around the population constant β_0 and the global slope.

Having set the two-stage formulation of the model, the following task will concern the estimation of the parameters therein. A rather naive approach would be to estimate the effects of Stage 1 in the first place and then to use them for the evaluation of the population- and treatment-specific effects. However, this entails several sources of failure. First, by using the estimated effects of Stage 1 ($\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$) for the estimation of $\beta_0, \beta_1, \ldots, \beta_m$, the variation of $\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$ is ignored. This leads to imprecision. The second disadvantage is the loss of information by pooling in the estimation of β_i. Third, the problem may arise that there are not enough observations for each subject to carry out an estimation, as has already been mentioned in the discussion about fitting separate linear regressions models for each subject.

Instead of this naive approach, a better way to combine the two stages will be described in the following. This will lead us to the definition of linear mixed models – models whose linear predictor η_{ij} includes **fixed as well as random effects** which explains the name *mixed models*.

The model in the example can be rewritten as

$$y_{ij} = \beta_0 + b_{0i} + \beta_1 \cdot Gr_{1i} \cdot x_{ij} + \ldots + \beta_m \cdot Gr_{mi} \cdot x_{ij} + b_{1i} \cdot x_{ij} + \varepsilon_{ij}$$

with

$$b_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{D} = \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{10} & \tau_1^2 \end{pmatrix}\right),$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{J_i}),$$

for $i = 1, \ldots, N$, $j = 1, \ldots, J_i$, and with $b_1, \ldots, b_N, \varepsilon_1, \ldots, \varepsilon_N$ independent.

Note that the assumption $Cov(\varepsilon_i) = \sigma^2 \boldsymbol{I}_{J_i}$ implies that the correlation between the repeated measurements on each subject are only produced by the vector of random effects b_i (which is common for these observations). Note that in general, this assumption can be relaxed and the model can be more flexible as will be shown in the following definition of the linear mixed models (Definition 4).

3.1.3 Definition of the Linear Mixed Model

A linear mixed model is given as[1]:

Definition 4. *Linear Mixed Model*

$$y = \boldsymbol{X}\beta + \boldsymbol{Z}b + \varepsilon \tag{3.9}$$

with

$$\begin{pmatrix} b \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{pmatrix}\right). \tag{3.10}$$

The matrices $\boldsymbol{X}(n \times p)$ and $\boldsymbol{Z}(n \times \nu)$ thereby denote the known design matrices, β is a vector of fixed effects and b a vector of random effects which is assumed to be independent of the unobservable and random error term ε. It is furthermore assumed that the covariance matrix of ε is positive (semi-) definite (and therefore nonsingular). Frequently, conditional independence of the response variables is assumed by setting the covariance matrix of the error term as $\boldsymbol{R} = \sigma^2 \boldsymbol{I}_n$. However, if the random effects do not seem to suffice to explain the covariance, a more general form of $\boldsymbol{R}$ should be used.

The normality assumption is – similar to the LM case – not necessary for all inferential conclusions in linear mixed models. However, as the usual estimation of the unknown components in the covariance matrices $\boldsymbol{G}$ and $\boldsymbol{R}$ is based on maximum likelihood methods, an assumption on the distribution is generally made. In analogy to the linear model, a multivariate normal distribution is used. Alternative distributions for the random effects are possible. However, this usually complicates the inference (Konrath, 2009).

The correlation structure of y is implied by the design matrix $\boldsymbol{Z}$, the covariance of the random effects $\boldsymbol{G}$ and the error variance $\boldsymbol{R}$ as

$$\boldsymbol{V} := Cov(y) = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{R}. \tag{3.11}$$

The covariance matrix of the error terms $\boldsymbol{R}$ thereby accounts for serial correlation not explained by $\boldsymbol{Z}b$, as well as measurement error. For more details see Fahrmeir et al. (2007); Konrath (2009) and Greven (2009).

3.1.4 The marginal and the conditional perspective

There are two possible – *non-equivalent* – ways to look at a mixed model. First, there is the **marginal** perspective in which the marginal distribution of the response is considered. And second, one can look at a mixed model as a hierarchical model based on the

[1] See Konrath (2009).

conditional distribution of the response given the random effects and on the marginal distribution of the random effects. The two perspectives will be introduced in the following.

Conditional perspective

Consider the conditional distribution of the response y given the random effects b in the first step of the hierarchical formulation

Step 1

$$y|b \sim \mathcal{N}\left(\boldsymbol{X}\beta + \boldsymbol{Z}b, \boldsymbol{R}\right), \tag{3.12}$$

and the marginal distribution of the random effects in the second step

Step 2

$$b \sim \mathcal{N}(0, \boldsymbol{G}). \tag{3.13}$$

Thus, for the first step one obtains a standard LM (conditional on the random effects b). For longitudinal or cluster data, the random effects b_i ($i = 1, \ldots, n$) can be interpreted as subject-specific effects on the mean that vary within the population. Thus, the subject-specific mean of y_i is modeled as a function of population-specific and subject-specific effects in the conditional model (Konrath, 2009).

The marginal point of view

For the marginal model consider the marginal distribution of y

$$y \sim \mathcal{N}\left(\boldsymbol{X}\beta, \boldsymbol{V}\right). \tag{3.14}$$

For the marginal model one thus obtains a general linear model, i.e. a model for which the assumption $\varepsilon \sim \mathcal{N}(0, \sigma^2\boldsymbol{I}_n)$ of the LM is replaced by the assumption $\varepsilon \sim \mathcal{N}(0, \sigma^2\boldsymbol{V})$ (Kneib, 2003). Here, the marginal, i.e. population-averaged mean of the response y_i is modeled as a function of only population-specific effects and no random effects are explicitly assumed in order to cater for the inter-subject variability. The random effects rather affect the correlation structure and therefore take the correlation in the data into consideration.

Comparison of the two perspectives

The two formulations are not equivalent, although the conditional model can be converted into the marginal model (not the other way round) in the case of linear mixed models (in contrast to generalizations, see 3.2.4) by integrating out the random effects b. For the proof see Appendix A.

It should be pointed out that this conversion is **restricted to the case of Gaussianity**, i.e. the case of a LMM. In more general cases, where the conditional response $y|b$ does not follow a Gaussian distribution but some distribution of the exponential family, the integral can usually not be analytically solved as will be discussed in the following.

Note that with the marginal model as a starting point, it is not possible to obtain the form of the conditional model. This is due to the fact that the marginal perspective does not contain random effects and therefore no distribution is designed for the random effects which are used in the conditional formulation. For more details see Greven (2009).

Although the two formulations are not equivalent, the interpretation of the fixed regression coefficients β stays the same[2]. This again **only holds for linear mixed models** (see 3.2.4).

3.1.5 Inference in the Linear Mixed Model

Both Likelihood and Bayesian inference methods can be applied to linear mixed models in order to draw conclusions from the data. In this work, the focus will be restricted to likelihood methods. For further details on both inferential types see Chapter 6 in Fahrmeir et al. (2007) on which the following is based.

Depending on the aim of the user, different aspects of statistical inference for mixed models can be brought into focus. If, for example, the interest lies in the population-specific effects only, the estimation of the fixed effects becomes the central objective. However, if a prediction, e.g. for each patient of a longitudinal study, is the target, then the estimation of the random effects becomes more important.
In the likelihood context, the estimation of fixed as well as random effects is based on generalized least-squares and generalized maximum likelihood approaches. The first question to be asked using likelihood inference is what the likelihood looks like – or rather which likelihood to use – for the linear mixed model. As shown before, the linear mixed model can be displayed in two ways – the conditional and the marginal form. If the fixed effects are of interest, one usually employs the marginal distribution for likelihood inference, thus one uses the fact that

$$y \sim \mathcal{N}\left(\boldsymbol{X}\beta, \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{R}\right).$$

[2] Under the condition that the canonical link function. i.e. the identical link function $g(\cdot) = h(\cdot)$ is used.

If on the other hand the focus lies on the random effects, the hierarchical formulation is used.
In the following, two situations will be distinguished. First, the case of known variance parameters ill be considered, i.e. one assumes that the matrices $\boldsymbol{G}$ and $\boldsymbol{R}$ are known. As this turns out to be a quite unrealistic assumption in real applications, the situation with unknown and therefore to be estimated covariance matrices $\boldsymbol{G}$ and $\boldsymbol{R}$ will also be considered. This will lead us to the distinction between maximum likelihood (ML) and *restricted* maximum likelihood (REML) estimation.

Estimation assuming *known* covariance matrices

1. Estimation of the fixed effects:

   The transformation

$$\begin{aligned} \boldsymbol{X}^* &= \boldsymbol{V}^{-1/2}\boldsymbol{X} \\ y^* &= \boldsymbol{V}^{-1/2}y \\ \varepsilon^* &= \boldsymbol{V}^{-1/2}\varepsilon, \end{aligned} \tag{3.15}$$

   with $\boldsymbol{V}^{1/2}$ being a square root[3] of matrix $\boldsymbol{V}$ shows, that the marginal model $y \sim \mathcal{N}(\boldsymbol{X}\beta, \boldsymbol{V})$ can be reduced to the linear model by writing

$$y^* = \boldsymbol{X}^*\beta + \varepsilon^*$$

   with $\varepsilon^* \sim \mathcal{N}(0, \boldsymbol{I_n})$ fulfilling the assumptions of the linear model.

   This allows to perform the estimation of the fixed effects vector β by using the generalized (weighted) least-squares criterion

$$GLS(\beta) = (y - \boldsymbol{X}\beta)^T\boldsymbol{V}^{-1}(y - \boldsymbol{X}\beta) \underset{\beta}{\rightarrow} \min \tag{3.16}$$

   which leads to the estimator[4]

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}y. \tag{3.17}$$

   Let in the following again $\mathcal{L}(\cdot)$ denote the likelihood function and $l(\cdot)$ the log-likelihood. Under the (optional) assumption of Gaussianity (see 3.1.3), this estimator $\hat{\beta}$ coincides with the maximum likelihood estimator which is obtained by maximizing the marginal log-likelihood with respect to β, namely

$$l(\beta) = log\,\{\mathcal{L}(\beta)\} \propto -\frac{1}{2}\,log\,(|\boldsymbol{V}|) - \frac{1}{2}\,(y - \boldsymbol{X}\beta)^T\boldsymbol{V}^{-1}(y - \boldsymbol{X}\beta) \underset{\beta}{\rightarrow} \max, \tag{3.18}$$

   with $|\boldsymbol{V}|$ denoting the determinant of matrix $\boldsymbol{V}$.

[3] obtained e.g. via Cholesky decomposition.

[4] Assuming that the inverses of $\boldsymbol{V}$ and of $\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}$ exist.

Implied by the Gauss-Markov Theorem (see Fahrmeir et al. (2007)), $\hat{\beta}$ has the following optimality properties (for known $\boldsymbol{G}$ and $\boldsymbol{R}$):

- **Unbiasedness**: $\hat{\beta}$ is an unbiased estimator for β, i.e. $E(\hat{\beta}) = \beta$.
- **Minimal variance**: $\hat{\beta}$ has minimal variance among all other linear estimators $\tilde{\beta} = H\beta$, with H any $N \times p$ matrix.

$\Rightarrow$ the estimator $\hat{\beta}$ is the **BLUE (Best linear unbiased estimator)**.

2. Estimation of the random effects:

There are several ways to derive the **best linear unbiased predictor** (**BLUP**) for the random effects vector b. As the marginal formulation does not involve random effects, one has to use the conditional model formulation in order to obtain an estimator for b. Note that the term "predictor" is used in order to point out that b is a vector of *random* effects, but has been seen as misleading by some authors (compare Kneib (2003)). Unbiasedness for random parameters requires that $E(\hat{b}) = E(b) = 0$ instead of the requirement $E(\hat{\beta}) = \beta$ which needs to hold for fixed parameters. Note that an unbiased random parameter does not have to fulfill $E(\hat{b}|b) = b$ for all b (see Greven (2009)).
The best linear unbiased predictor for b is the conditional expectation of b given the data

$$E(b|y) = \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(y - \boldsymbol{X}\beta). \tag{3.19}$$

One approach that leads to this estimator is to consider the joint density of y and b

$$\begin{pmatrix} y \\ b \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{X}\beta \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{V} & \boldsymbol{Z}\boldsymbol{G} \\ \boldsymbol{G}\boldsymbol{Z}^T & G \end{pmatrix} \right) \tag{3.20}$$

and then to use the properties of marginal and conditional probability distributions (see e.g. Theorem B.4 in Fahrmeir et al. (2007)). The same estimator for b (and also the same estimator for $\hat{\beta}$) arises by maximizing the joint density of y and b which will be described in the following paragraph.
By the replacement of the unknown vector β with the BLUE $\hat{\beta}$ from the precedent paragraph, one obtains the estimator

$$\hat{b} = \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(y - \boldsymbol{X}\hat{\beta}) \text{ for the random effects vector.} \tag{3.21}$$

As its name implies, one can show that the BLUP is the "best" estimator – in the sense of minimizing the mean squared error $E\left[(\hat{b} - b)^T(\hat{b} - b)\right]$ – in the class of all unbiased linear estimators for b.

3. Simultaneous estimation of fixed and random effects:

As mentioned above, it is possible to derive the same estimators for β and b as above by maximizing the joint density of y and b simultaneously with respect to β and b. Note that the estimator $\begin{pmatrix}\hat{\beta}\\ \hat{b}\end{pmatrix}$ is also referred to as BLUP (not only $\hat{b}$).
The joint log-likelihood

$$l(y,b) = log\left\{\mathcal{L}\right\} \propto -\frac{1}{2}(y - \boldsymbol{X}\beta - \boldsymbol{Z}b)^T\boldsymbol{R}^{-1}(y - \boldsymbol{X}\beta - \boldsymbol{Z}b) - \frac{1}{2}\, b^T\boldsymbol{G}^{-1}b \quad (3.22)$$

can be interpreted as a *penalized* log-likelihood for the random effects vector b with the *penalization term* $b^T\boldsymbol{G}^{-1}b$.

Maximizing the log-likelihood is equivalent to minimizing the penalized least-squares criterion

$$GLS_{pen}(\beta,b) = (y - \boldsymbol{X}\beta - \boldsymbol{Z}b)^T\boldsymbol{R}^{-1}(y - \boldsymbol{X}\beta - \boldsymbol{Z}b) + b^T\boldsymbol{G}^{-1}b \underset{\beta,b}{\longrightarrow} \min, \quad (3.23)$$

where the first term corresponds to the generalized (weighted) least-squares criterion from above and the second term $b^T\boldsymbol{G}^{-1}b$ accounts for the fact that b arises from a distribution.

Without the second term, the random effects vector b would – like β – be estimated like a fixed effect. Due to the assumption $b \sim \mathcal{N}(0,\boldsymbol{G})$, the term $b^T\boldsymbol{G}^{-1}b$ penalizes the discrepancy to zero and this all the more the “smaller” $\boldsymbol{G}$ is. For $\boldsymbol{G} \to \infty$, the penalization term vanishes and b is treated like a fixed effect.
Differentiating $GLS_{pen}(\beta,b)$ with respect to β and b and setting the derivatives to zero leads to the estimating equations:

Henderson's mixed model equations

$$\begin{pmatrix}\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}\end{pmatrix}\begin{pmatrix}\hat{\beta}\\ \hat{b}\end{pmatrix} = \begin{pmatrix}\boldsymbol{X}^T\boldsymbol{R}^{-1}y\\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}y\end{pmatrix}. \quad (3.24)$$

The derivation of these equations can be found in Appendix A.

Matrix conversions show that the solution of Henderson's mixed model equations is equivalent to the estimators derived in the preceding paragraphs (see Fahrmeir et al. (2007)).

The simultaneous estimation of β and b is strongly related to the empirical Bayesian estimation.

Estimation assuming *unknown* covariance matrices

1. Estimation of the covariance structure:

   There are two common ways to estimate unknown parameters in $\boldsymbol{G}$, $\boldsymbol{R}$, and $\boldsymbol{V}$: Maximum likelihood (ML) and restricted maximum likelihood (REML) methods. Let in the following θ denote these unknown parameters. To emphasize the dependency on θ, $\boldsymbol{G}$, $\boldsymbol{R}$, and $\boldsymbol{V}$ will sometimes be noted $\boldsymbol{G}(\theta)$, $\boldsymbol{R}(\theta)$, and $\boldsymbol{V}(\theta)$, respectively and thus the covariance of y can be written as

   $$\boldsymbol{V} = \boldsymbol{V}(\theta) = \boldsymbol{Z}\boldsymbol{G}(\theta)\boldsymbol{Z}^T + \boldsymbol{R}(\theta). \tag{3.25}$$

   If however it becomes clear from the context that $\boldsymbol{G}(\theta)$, $\boldsymbol{R}(\theta)$, and $\boldsymbol{V}(\theta)$ are meant the dependence on θ will be suppressed. Note that both notations, $\hat{\theta}$ and $\hat{\theta}(y)$, will be used depending on whether the dependence on the data is emphasized or not.

   In the linear model the maximum likelihood estimator of the variance σ^2 is biased due to the fact that the estimation of σ^2 involves an estimator of β but does not take into account the loss of degrees of freedom resulting from the estimation of parameter β. Similarly, it can be shown that the ML estimator for the covariance structure in the linear mixed model is biased (Fahrmeir et al., 2007). Hence, the restricted maximum likelihood estimation is usually preferred as it reduces the bias of the ML estimator $\hat{\theta}_{ML}$. However, it is not ensured that the mean squared error of $\hat{\theta}_{REML}$ also becomes smaller (Fahrmeir et al., 2007). Note that in contrast to the linear model, where the REML estimator for σ^2 is unbiased, this is not generally the case in linear mixed models, but the bias is reduced (Fahrmeir et al., 2007).

   The ML estimator can be derived as follows:
   Proceeding from the log-likelihood of the marginal formulation of the mixed model

   $$l(\beta,\theta) \propto -\frac{1}{2}\left\{log|\boldsymbol{V}(\theta)| + (y-\boldsymbol{X}\beta)^T\boldsymbol{V}(\theta)^{-1}(y-\boldsymbol{X}\beta)\right\}, \tag{3.26}$$

   with $|\boldsymbol{V}(\theta)|$ denoting the determinant of $\boldsymbol{V}(\theta)$, the *profile* log-likelihood for θ is calculated by maximizing $l(\beta,\theta)$ for fixed θ with respect to β and then plugging in the obtained estimator for β,

   $$\tilde{\beta}(\theta) = (\boldsymbol{X}^T\boldsymbol{V}(\theta)^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}(\theta)^{-1}y, \tag{3.27}$$

   into the marginal log-likelihood $l(\beta,\theta)$. This yields the

   Profile log-likelihood

   $$l_P(\theta) \propto -\frac{1}{2}\left\{log|\boldsymbol{V}(\theta)| + (y-\boldsymbol{X}\beta(\tilde{\theta}))^T\boldsymbol{V}(\theta)^{-1}(y-\boldsymbol{X}\tilde{\beta}(\theta))\right\}. \tag{3.28}$$

   Maximizing the profile log-likelihood of θ with respect to θ then yields the ML estimator $\hat{\theta}_{ML}$.

For the restricted maximum likelihood estimation of θ, the marginal or *restricted* log-likelihood

$$l_R(\theta) = log \left\{ \int \mathcal{L}(\beta, \theta) d\beta \right\} \tag{3.29}$$

is maximized instead of the profile log-likelihood $l_P(\theta)$. It is obtained by integrating out β from the likelihood of the marginal formulation of the linear mixed model and can be alternatively derived as a restricted log-likelihood in the context of linear models (Fahrmeir et al., 2007).
Relating $l_R(\theta)$ to $l_P(\theta)$ yields

$$l_R(\theta) = l_P(\theta) - \frac{1}{2}\, log|\boldsymbol{X}^T \boldsymbol{V}(\theta)\boldsymbol{X}|, \tag{3.30}$$

with $|\boldsymbol{X}^T \boldsymbol{V}(\theta)\boldsymbol{X}|$ denoting the determinant of $\boldsymbol{X}^T \boldsymbol{V}(\theta)\boldsymbol{X}$.

Again, several ways lead to the same estimator. One way to derive $\hat{\theta}_{REML}$ makes use of a linear contrast matrix $\boldsymbol{A} \neq \boldsymbol{0}$ which is constructed such that $E(\boldsymbol{A}y) = \boldsymbol{A}\boldsymbol{X}\beta = 0$ and that the resultant log-likelihood for the transformed vector $\tilde{y} = \boldsymbol{A}y$ no longer depends on the fixed effects β. It can be shown that the resultant log-likelihood is independent (up to an additive constant) of the contrast matrix used (Verbeke and Molenberghs, 2000). As one possibility for the choice of $\boldsymbol{A}$ is

$$\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{V}^{-1}(\theta)\boldsymbol{X})^{-1}\boldsymbol{X}^T \boldsymbol{V}^{-1}(\theta), \tag{3.31}$$

the restricted log-likelihood is also called *residual* log-likelihood (Fahrmeir et al., 2007; Greven, 2008). Alternatively, $\hat{\theta}_{REML}$ can be derived from the Bayesian point of view as the posterior mode estimator with the use of a non-informative prior $p(\beta) \propto constant$.

Since $\hat{\theta}_{ML}$ and $\hat{\theta}_{REML}$ are not linear in θ, the numerical calculation of $\hat{\theta}_{ML}$ and $\hat{\theta}_{REML}$ is carried out iteratively, e.g using a Newton-Raphson- or a Fisher-Scoring algorithm (for details see Fahrmeir et al. (2007) and Konrath (2009)).

The parameters β and θ can be estimated simultaneously by maximizing

$$l(\beta, \theta) - \frac{1}{2}\, log|\boldsymbol{X}^T \boldsymbol{V}(\theta)^{-1}\boldsymbol{X}|. \tag{3.32}$$

Alternatively, $\hat{\beta}$ and $\hat{\theta}$ are obtained from the mixed model equations (3.24).

Plugging in the resultant $\hat{\theta}$ after convergence leads to the estimated covariance matrices

$$\hat{\boldsymbol{R}} = \boldsymbol{R}(\hat{\theta}),\ \hat{\boldsymbol{G}} = \boldsymbol{G}(\hat{\theta}),\ \text{and } \hat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\theta}),\ \text{respectively.} \tag{3.33}$$

2. Estimation of the fixed and random effects:

   In the case of unknown covariance structure, the estimated covariance matrices $\hat{\boldsymbol{R}} = \boldsymbol{R}(\hat{\theta})$, $\hat{\boldsymbol{G}} = \boldsymbol{G}(\hat{\theta})$, and $\hat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\theta})$ from the previous paragraph are used to obtain estimators for β and b. Note that by plugging in the covariance matrices, the covariances of the estimators are no longer analytically accessible and the optimality properties do no longer hold exactly. One obtains the so called **empirical best linear unbiased predictor** (**EBLUP**) $\begin{pmatrix}\hat{\beta}\\ \hat{b}\end{pmatrix}$ with

$$\hat{\beta} = (\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}y \tag{3.34}$$

$$\hat{b} = \hat{\boldsymbol{G}}\boldsymbol{Z}^T\hat{\boldsymbol{V}}^{-1}(y - \boldsymbol{X}\hat{\beta}), \tag{3.35}$$

   or equivalently,

$$\begin{pmatrix}\hat{\beta}\\ \hat{b}\end{pmatrix} = (\boldsymbol{C}^T\hat{\boldsymbol{R}}^{-1}\boldsymbol{C} + \hat{\boldsymbol{B}})^{-1}\boldsymbol{C}^T\hat{\boldsymbol{R}}^{-1}y \tag{3.36}$$

$$\text{with } \boldsymbol{C} = (\boldsymbol{X}, \boldsymbol{Z}) \text{ and } \hat{\boldsymbol{B}} = \begin{pmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \hat{\boldsymbol{G}}^{-1}\end{pmatrix}.$$

   In contrast to the linear model where $\beta(\hat{\theta}_{ML})$ is equal to $\beta(\hat{\theta}_{REML})$, this is not the case for the linear mixed model, since the estimator of the fixed effects β depends on the covariance matrix $\boldsymbol{V}$ (see (3.34)).

Hypothesis testing

The matter of hypothesis testing in linear mixed models will be only briefly treated in this paragraph as it is not in the focus. However, there is a strong link between hypothesis testing and model selection based on information criteria and the problems arising can be traced back to the same properties of mixed models (see Greven (2008)).
Often, hypotheses about fixed effects are of central interest. In this case, standard hypothesis testing can be applied, such as Wald tests and likelihood-ratio tests using approximate covariance matrices of $\hat{\beta}$ (Fahrmeir et al., 2007).
Yet, if the interest lies in hypotheses about random effects b, one is confronted with the problem of a non-open parameter space. This implies that the classical asymptotic likelihood theory cannot be applied any more.

Consider for example the longitudinal linear mixed model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + b_{0i} + \varepsilon_{ij}, \text{ with } i = 1, \ldots, N, \ j = 1, \ldots, J_i, \tag{3.37}$$

with $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, $b_{0i} \sim \mathcal{N}(0, \tau_0^2)$ and the hypotheses pair

$$H_0 : \tau_0^2 = 0 \text{ versus } H_1 : \tau_0^2 > 0. \tag{3.38}$$

Thus, the interest lies in answering the question whether the linear model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, \text{ with } i = 1, \ldots, N, \; j = 1, \ldots, J_i \tag{3.39}$$

is valid or not. In this context, one has to deal with a non-open parameter space, since τ_0^2 is a variance and therefore assumed non-negative ($\tau_0^2 \geq 0$). Thus, the null hypothesis lies on the border of the parameter space whereas in classical asymptotic likelihood theory it is assumed to be in the interior (Fahrmeir et al., 2007). This results in a point mass at zero as under the null hypothesis there is a 50:50 chance of τ_0^2 being estimated to be zero. In contrast to the standard cases (with no random effects), the statistic in this situation is no longer asymptotically chi-squared distributed with one degree of freedom (compare Greven (2008)). Several approaches have been considered to deal with the problem in order to enable the testing for zero variance components.

One suggestion is to use parametric bootstrap. The idea here is to re-use the estimated parameters of the simpler model in order to generate new data. This data is then evaluated under both models, i.e. the simpler and the more complex model, in order to compute the likelihood-ratio test. One obtains an approximate distribution of the statistic of interest under the simpler model. The generated data is then compared to the actual value of the test statistic (see Mansmann (2009), Crainiceanu and Ruppert (2004)).

Alternatively, Self and Liang (1987) show that the asymptotic distribution is an equal mixture of chi-squared distributions. In the special situation in (3.38), it is an equal mixture between a point mass at zero and a chi-squared distribution with one degree of freedom.[5] For detailed information see Greven (2008).

3.1.6 LMM for Longitudinal and Cluster Data

In the motivation for the linear mixed model (3.1.2), one important special case of mixed models has already been introduced – the analysis of longitudinal or cluster data. These kind of data arises when, for example, a medical survey with multiple waves is executed, producing repeated measurements for each patient or whenever the observed subjects are grouped in some way (e.g. subjects belonging to the same family, school, etc.). The wide use of longitudinal and cluster data (especially in medical fields) makes it important to take a closer look at mixed models for longitudinal or cluster data. This section can also serve as an illustration of how these models arise as a special case from general mixed models.

For longitudinal or cluster data, the linear mixed model is given as:

Definition 5. *LMM for Longitudinal or Cluster Data*

$$y_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i b_i + \varepsilon_i, \; \textit{for } i = 1, \ldots, N, \tag{3.40}$$

where N is the number of individuals or clusters, and y_i is the J_i-dimensional vector of response variables for individual/cluster i.

[5] Greven (2008)

For longitudinal data, y_{ij} denotes the observation of individual i at time t_{ij}, whereas for cluster data, y_{ij} indicates the observation for object j in cluster i. The design matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are of dimension $(J_i \times p)$ and $(J_i \times q)$, respectively. β is the the p-dimensional vector of fixed effects and b_i the q-dimensional vector of random effects, where $b_i \sim \mathcal{N}(0, \boldsymbol{D})$. For the error term ε_i, one assumes $\varepsilon_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_i)$ $(i = 1, \dots, N)$ and additionally that $b_1, \dots, b_N, \varepsilon_1, \dots, \varepsilon_N$ are independent.

Alternatively, the model can be written more compactly as

$$y = \boldsymbol{X}\beta + \boldsymbol{Z}b + \varepsilon, \tag{3.41}$$

with $y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$, $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix}$, $b = \begin{pmatrix} b_1 \\ \vdots \\ b_N \end{pmatrix}$, and the design matrices $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_N \end{pmatrix}$ and $\boldsymbol{Z} = diag(\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_N) = \begin{pmatrix} Z_1 & & \\ & \ddots & \\ & & Z_N \end{pmatrix}$.

This notation allows to see that the longitudinal/cluster model results from the general linear mixed model by choosing a block-diagonal matrix $\boldsymbol{Z}$ and the covariance matrices of the general linear mixed model, $Cov(\varepsilon) = \boldsymbol{R}$ and $Cov(b) = \boldsymbol{G}$, as the block-diagonal matrices

$$\boldsymbol{R} = diag(\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_N) \tag{3.42}$$

$$\boldsymbol{G} = diag(\boldsymbol{D}_1, \dots, \boldsymbol{D}_N), \text{ where } \boldsymbol{D}_i = \boldsymbol{D}. \tag{3.43}$$

The block-diagonal structure results from the assumption that the individuals/clusters are independent but the repeated measurements at the same subject (in the same cluster) are not.

The assumption of independence is not made in the general linear mixed model. The relaxation of this assumption permits the construction of more flexible models, comprising e.g. nested structures or smooth components modeled by penalized splines (see Chapter 4).

Often, the design matrix $\boldsymbol{Z}$ contains covariates which are also included in $\boldsymbol{X}$. Thus, with the random effects b_i and the assumption $E(b_i) = 0$, the individual discrepancy of the respective population mean is modeled.

Usually an intercept is included in the model by adding a 1 as the first component to the vectors x_{ij} and z_{ij}.

Furthermore, an interesting interpretation exists for the longitudinal linear mixed model. Namely, the best linear unbiased predictor for y_i (i.e. applying the *BLUP* from Section 3.1.5)

$$\begin{aligned}\hat{y}_i &= \boldsymbol{X}_i\hat{\beta} + \boldsymbol{Z}_i\hat{b}_i \\ &= \boldsymbol{X}_i\hat{\beta} + \boldsymbol{Z}_i\hat{\boldsymbol{D}}_i\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}(y - \boldsymbol{X}_i\hat{\beta}) \\ &= (\boldsymbol{I}_{J_i} - \boldsymbol{Z}_i\hat{\boldsymbol{D}}_i\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1})\boldsymbol{X}_i\hat{\beta} + \boldsymbol{Z}_i\hat{\boldsymbol{D}}_i\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}y_i \\ &= (\hat{\boldsymbol{V}}_i^{-1} - \boldsymbol{Z}_i\hat{\boldsymbol{D}}_i\boldsymbol{Z}_i^T)\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{X}_i\hat{\beta} + \boldsymbol{Z}_i\hat{\boldsymbol{D}}_i\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}y_i \\ &= \hat{\boldsymbol{\Sigma}}_i\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{X}_i\hat{\beta} + \boldsymbol{Z}_i\hat{\boldsymbol{D}}_i\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}y_i\end{aligned} \tag{3.44}$$

is a weighted average of the population mean $\boldsymbol{X}_i\hat{\beta}$ and the observed data y_i. Recall, that $\hat{\boldsymbol{V}}_i = \hat{\boldsymbol{\Sigma}}_i + \boldsymbol{Z}_i\hat{\boldsymbol{D}}_i\boldsymbol{Z}_i^T$. The ith subject response profile is thus shrunk to the population average mean profile ("borrowing of strength"[6]). The amount of shrinkage depends on the magnitude of $\hat{\boldsymbol{\Sigma}}_i$ and $\hat{\boldsymbol{V}}_i$. If $\hat{\boldsymbol{\Sigma}}_i\hat{\boldsymbol{V}}_i^{-1}$ is large, i.e. the residual variability is large compared to the between-subject variability $\boldsymbol{Z}_i\hat{\boldsymbol{D}}_i\boldsymbol{Z}_i^T$, the population-averaged profile is given much weight. In contrast, when the residual variance $\hat{\boldsymbol{\Sigma}}_i$ is small compared to $\boldsymbol{Z}_i\hat{\boldsymbol{D}}_i\boldsymbol{Z}_i^T$, the opposite is the case (Greven, 2009).

The Random Intercept Model

One important special case of the linear mixed model for longitudinal or cluster data is a model which contains fixed effects and a random intercept, called the *random intercept model*. It will be quickly introduced here as it is applied in the second part of the simulation studies (compare Section 6.2). In the example from above (Section 3.1.2), a random intercept model would be adequate if it was assumed that the blood pressure curve of the patients differed due to subject specific intercepts, but that the trend stayed the same. The following definition is based on Konrath (2009).

Definition 6. *Random Intercept Model*

$$y_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i b_{0i} + \varepsilon_i, \ \textit{for } i = 1, \ldots, N,$$

with

$$\boldsymbol{Z}_i = \mathbb{1}_i = (1, \ldots, 1)^T, \ b_{0i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tau^2).$$

For each observation it has the form:

$$y_{ij} = x_{ij}^T\beta + b_{0i} + \varepsilon_i, \ \text{for } i = 1, \ldots, N, \ \text{and } j = 1, \ldots, J_i.$$

In combination with the assumption

$$\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \tag{3.45}$$

[6] Greven (2009)

one obtains a model with a marginal covariance structure that implies a constant correlation structure (*compound symmetry*), i.e.

$$Cor(y_{ij}, y_{ij'}) = \rho = \frac{\tau^2}{\sigma^2 + \tau^2}, \text{ for } j \neq j'. \tag{3.46}$$

For each observation of an individual/within a cluster, the variance is

$$Var(y_i) = \begin{pmatrix} \sigma^2 + \tau^2 & \tau^2 & \dots & \tau^2 \\ \tau^2 & \sigma^2 + \tau^2 & \dots & \tau^2 \\ \vdots & & \ddots & \vdots \\ \tau^2 & \tau^2 & \dots & \sigma^2 + \tau^2 \end{pmatrix}. \tag{3.47}$$

Here, the correlations in the random intercept model with compound symmetry always have to be positive (or zero) – in contrast to a general marginal model – as they correspond to the random effects variance τ^2.

3.1.7 Implementation of the Linear Mixed Model in `R`

The implementation of linear mixed models in `R` can be conducted with the function `lme` from package `nlme`, which has been used in the second part of the simulation study in Chapter 6 for the estimation of the random intercept models (Section 6.2). Both approaches – maximum likelihood and restricted maximum likelihood – are implemented in this package and can be specified by the argument `method` in function `lme`. Note that the function `lme` maximizes the (restricted) log-likelihood with respect to the scaled logarithm of the variances, and thus can never find a maximum at zero (see Pinheiro and Bates (2000) who give a detailed description of their package). Various specifications of correlation structures, such as compound symmetry and unspecified correlation, are available in `lme`. The iterative optimization algorithm is a hybrid of an EM-algorithm and a Newton-Raphson algorithm (Konrath, 2009; Greven, 2009).
The iterations of the EM-algorithm are fast and easy to compute and one usually quickly reaches the regions of the optima of the parameters. However, it often takes long until the EM-algorithm converges once one is in a close neighborhood of the optimum. On the other hand, the iterations of the Newton-Raphson algorithm are computationally very expensive as the score-function and the Hessian matrix have to be recalculated for the actual values of the estimators in each step. Moreover, the Newton-Raphson algorithm turns out to be instable in regions at longer ranges of the optimum. However, having reached a close neighborhood of the optimum, the Newton-Raphson algorithm converges very fast. It is therefore convenient to start off with several EM-iterations and then to switch over to iterations of the Newton-Raphson algorithm (compare Konrath (2009)). The number of EM-iterations can be specified in the function `lme` by the argument `control = list(niterEM)` and has a default of 25 iterations. For a brief documentation of this function see Appendix E.1.1.

3.2 The Generalized Linear Mixed Model

3.2.1 The Generalized Linear Model

In analogy to the introduction of the linear mixed model, where the standard linear model served as starting point, its generalization, the generalized linear model (GLM), will be used in order to introduce the generalized linear mixed model. As the concept of the *exponential family* is crucial for the definition of the GLM, it will be introduced first.

The exponential family is a family of distributions which can all be written in the same form. This is very useful, as it allows to show properties in general and one does not have to conduct the proofs for every single distribution.

Definition 7. *One-parametric Exponential Family*

A random variable y_i follows a distribution from the one-parametric exponential family, if the density or probability mass function (pmf) is of the form

$$f(y_i|\vartheta_i, \phi) = exp\left\{\frac{y_i\vartheta_i - b(\vartheta_i)}{\phi} + c(y_i, \phi)\right\}, \tag{3.48}$$

with ϑ_i denoting the canonical parameter, ϕ is the dispersion parameter, $b(\cdot)$ (for which the first and second derivative have to exist), and $c(\cdot)$ are known functions. The term $c(y_i, \phi)$ is a scaling.

It can be shown that the important relationships

$$E(y_i) = \mu = b'(\vartheta_i) \tag{3.49}$$

$$Var(y_i) = \sigma_i^2 = \phi v(\mu_i) = \phi b''(\vartheta_i) \tag{3.50}$$

hold for the exponential family (for the proof see McCullagh and Nelder (1989)). The relation of the mean to the variance is specified by the *variance function* $v(\cdot)$, which is a function of μ_i.

The following three distributions rank among the most important examples of the exponential family:

1. Gaussian distribution: $f(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\, exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$
2. Bernoulli distribution: $f(y|\pi) = \pi^y(1-\pi)^{1-y}$
3. Poisson distribution: $f(y|\lambda) = \frac{\lambda^y}{y!}\, exp(-\lambda)$.

The corresponding parameters of the representation as a member of the one-parametric exponential family of these three distributions can bee seen in Table 3.1.

Distribution	$\boldsymbol{\vartheta(\mu)}$	$\boldsymbol{b(\vartheta)}$	$\boldsymbol{v(\mu)}$	$\boldsymbol{\phi}$
Gaussian	μ	$\frac{1}{2}\vartheta^2$	1	σ^2
Bernoulli	$log\left\{\frac{\pi}{(1-\pi)}\right\}$	$log\,(1+exp\,(\vartheta))$	$\pi(1-\pi)$	1
Poisson	$log\,(\lambda)$	$exp\,(\vartheta)$	λ	1

Table 3.1: *Some members of the one-parametric exponential family.*

One can see that the mean and the variance are independent for the Gaussian distribution, since the variance function is equal to 1. In contrast, this is not the case or the Bernoulli and the Poisson distribution as they are one-parameter distributions.

The one-parametric exponential family comes into play in the definition of the generalized linear model. This definition consists of two aspects. First, the assumption about the distribution of the response variable and second the assumption about the structure (or systematic component) which answers the question of how the covariates affect the response variable.

Definition 8. *Generalized Linear Model (GLM)*

Distribution

For given covariates x_i, the response variables y_i $(i = 1, \dots, n)$ are (conditionally) independent and the conditional density (or pmf) is a member of the one-parametric exponential family.

Structure

The conditional mean $E(y_i|x_i)$ is linked to the linear predictor $\eta_i = x_i^T\beta$ through

$$\mu_i = h(\eta_i) \text{ or respectively } \eta_i = g(\mu_i) \tag{3.51}$$

with $h(\cdot)$ the bijective and twice continuously differentiable response function and $g(\cdot) = h^{-1}(\cdot)$ its inverse function, called the link function.

If the equality $\vartheta_i = \eta_i = x_i^T\beta$ holds, the link function $g(\cdot)$ is called the *canonical link function.* In this case, many components of the inference in the GLM can be simplified. Thanks to the formulation of the exponential family, it is possible to express the inferential components in a general way for all members of the exponential family. The estimation in the general linear model is usually conducted by using maximum likelihood estimation. Since the observations $y_1, \ldots, y_n$ are independent (for given covariates), the log-likelihood can be written as

$$log\left\{\mathcal{L}(\beta,\phi)\right\} = \frac{1}{\phi}\sum_{i=1}^{n}\left\{y_i\vartheta_i - b(\vartheta_i)\right\} + \sum_{i=1}^{n} c(y_i,\phi). \tag{3.52}$$

The derivation with respect to β yields the score equations

$$\mathcal{S}(\beta) = \sum_{i=1}^{n} x_i \frac{\partial h(\eta_i)}{\partial \eta_i}(y_i - \mu_i) \overset{!}{=} 0 \tag{3.53}$$

which have to be solved in order to obtain an estimator for β. This is usually done numerically by either using the Newton Raphson algorithm or Fisher-Scoring in form of an Iteratively Reweighted least-squares (IRLS) estimation (see Fahrmeir et al. (2007)). Note that the two algorithms coincide in the case of a canonical link function. The dispersion parameter ϕ is usually estimated by a methods-of-moment estimator.

3.2.2 Motivation of the Generalized Linear Mixed Model

Similarly to the linear case, where the introduction of random effects in the linear predictor was motivated by the longitudinal study example on blood pressure, it can also be reasonable to allow random effects in the case of non-Gaussian, e.g. binary, response variables. Just as the GLM is a generalization of the LM, allowing y to follow any member of the one-parametric exponential family, the generalized linear mixed model (GLMM) extends the linear mixed model. The GLMM is thus an extension to the generalized linear model as well as to the linear mixed model which are themselves generalizations of the linear model (see Figure 3.1).

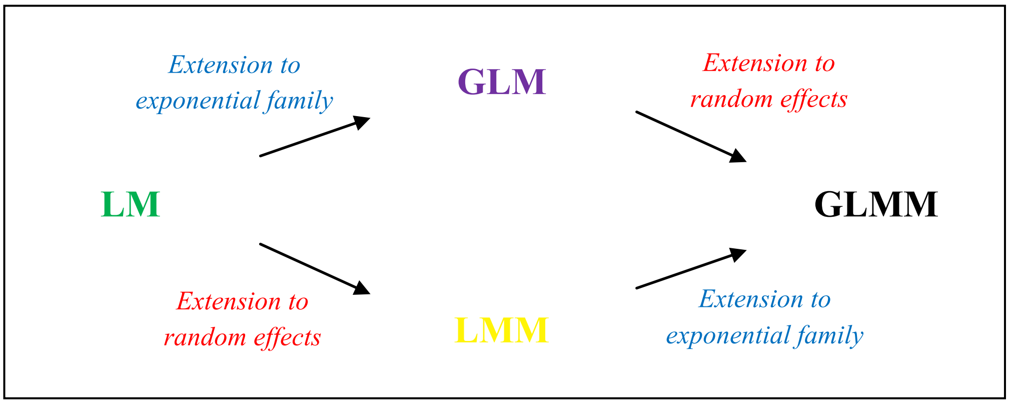

Figure 3.1: *Connection between the linear model (LM), the generalized linear model (GLM), the linear mixed model (LMM) and the generalized linear mixed model (GLMM).*

3.2.3 Definition of the Generalized Linear Mixed Model

Three assumptions are made for the definition of the generalized linear mixed model. First, like in the GLM, an assumption on the distribution of the response variables is made. Second, the structure has to be specified and third, one has to make an assumption on the distribution of the random effects.

Definition 9. *Generalized Linear Mixed Model (GLMM)*

Distribution of $\boldsymbol{y}$

Given the random effects b and the covariates x_i, the response variables y_i $(i = 1, \ldots, n)$ are assumed to be conditionally independent and the conditional density (or pmf) $f(y_i|b_i, x_i)$ is a member of the one-parametric exponential family.
Note that the assumption of conditional independence corresponds to the assumption of independent errors $\varepsilon \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$, i.e $\boldsymbol{R} = \sigma^2 \boldsymbol{I}_n$, in the linear mixed model and can in principle be relaxed. However, as the dilution makes the model much more complicated than it is the case in the LMM, conditional independence is assumed in general and dependencies are modeled via random effects in the linear predictor η (Fahrmeir et al., 2007).

Structure

The conditional mean $E(y_i|b_i, x_i)$ is linked to the extended linear predictor

$$\eta_i = x_i^T \beta + z_i^T b_i$$

through

$$\mu_i = h(\eta_i) \text{ or respectively } \eta_i = g(\mu_i) \tag{3.54}$$

where $h(\cdot)$ is the bijective, twice differentiable response function.

Distribution of $\boldsymbol{b}$

The random effects b are usually assumed to follow a multivariate Gaussian distribution

$$b \sim \mathcal{N}(0, \boldsymbol{G}). \tag{3.55}$$

In matrix notation the GLMM can be written as

$$f(y|b, \vartheta, \phi) = exp\left\{\frac{y\vartheta - b(\vartheta)}{\phi} + c(y, \phi)\right\} \tag{3.56}$$

$$h(\eta) = h(\boldsymbol{X}\beta + \boldsymbol{Z}b) = \mu = E(y|b, x) \tag{3.57}$$

$$b \sim \mathcal{N}(0, \boldsymbol{G}). \tag{3.58}$$

For more details see Fahrmeir et al. (2007).

3.2.4 The marginal and the conditional perspective

In analogy to the linear mixed model, it is possible to represent the GLMM in two different and *non-equivalent* ways, the marginal and the conditional formulation. Theoretically, the marginal model, which is based on the marginal distribution of the response, $f(y)$, can be deduced from the conditional distribution (the member of the exponential family) by integrating out the random effects b,

$$f(y) = \int f(y|b) f(b)\, db. \tag{3.59}$$

However, in general, when the conditional response does not necessarily follow a Gaussian distribution, the integral cannot be solved analytically what makes inference technically more demanding than it is in the linear case. Using the rules for conditional expectations, it can be shown that also the marginal mean,

$$E(y_i) = E\left[E(y_i|b_i)\right] = E(\mu_i) = E\left[h(\eta_i)\right] = E\left[h(\boldsymbol{X}_i\beta + \boldsymbol{Z}_i b_i)\right], \tag{3.60}$$

the marginal variance

$$\begin{aligned} Var(y_i) &= Var\left[E(y_i|b_i)\right] + E\left[Var(y_i|b_i)\right] = Var(\mu_i) + E\left[\phi\ v(\mu_i)\right] \\ &= Var\left[h(\eta_i)\right] + \phi\ E\left[v(h(\eta_i))\right] \\ &= Var\left[h(\boldsymbol{X}_i\beta + \boldsymbol{Z}_i b_i)\right] + \phi\ E\left[v(h(\boldsymbol{X}_i\beta + \boldsymbol{Z}_i b_i))\right], \end{aligned} \tag{3.61}$$

and the marginal covariance of the response,

$$\begin{aligned} Cov(y_i, y_j) &= Cov\left[E(y_i|b), E(y_j|b)\right] + E\left[Cov(y_i, y_j|b)\right] \\ &= Cov\left[h(\eta_i),\ h(\eta_j)\right] = Cov\left[h(\boldsymbol{X}_i\beta + \boldsymbol{Z}_i b_i),\ h(\boldsymbol{X}_j\beta + \boldsymbol{Z}_j b_j)\right], \end{aligned} \tag{3.62}$$

can in general not be computed analytically.[7] This property can be traced back to the non-linearity of the link function $g(\cdot)$ (Fahrmeir et al., 2007).

Note that due to the fact that the marginal expectation (3.60) is in general not equal to the conditional expectation, i.e.

$$E(y_i) \neq \boldsymbol{X}_i\beta = E(y_i|b_i), \tag{3.63}$$

the interpretation of the fixed regression coefficients β in the two perspectives is **not** the same. An exception is the case of Gaussianity with the use of the canonical link function, $g(\cdot) = id(\cdot)$, as in this special case it holds that

$$\begin{aligned} E(y_i) &= E\left[E(y_i|b_i)\right] = E(\mu_i) = E(\eta_i) = E\left[\boldsymbol{X}_i\beta + \boldsymbol{Z}_i b_i\right], \\ &= \boldsymbol{X}_i\beta + \underbrace{E\left[\boldsymbol{Z}_i b_i\right]}_{=\,0} \\ &= \boldsymbol{X}_i\beta. \end{aligned} \tag{3.64}$$

[7] The term $E\left[Cov(y_i, y_j|b)\right]$ in the marginal covariance vanishes due to the conditional independence of the response variables.

3.2.5 Inference in the Generalized Linear Mixed Model

The main idea of the inference in the GLMM stays the same as in the linear case. However, due to the non-linearity of the link function, inference in the GLMM cannot be carried out analytically, but numerical procedures or approximations are needed (Fahrmeir et al., 2007).
Different approaches exist to estimate the quantities of interest and new algorithms are still developed as this is an active field of research. Three approaches will be introduced in the following. All of them are based on some kind of approximation in order to compute the inaccessible marginal likelihood. The first one approximates the integrand, the second the data and in the third, the integral is approximated (for more details see Greven (2009) and Fahrmeir et al. (2007)). The implementation of GLMMs will be the subject of the following section.

The Laplace Approximation (LA)

Consider the case of a canonical link function and let θ_* denote the vector of all unknown components of $\boldsymbol{G} = Cov(b)$. The marginal likelihood is given by

$$\begin{aligned} \mathcal{L}(\beta, \theta_*, \phi) &= f(y|\beta, \theta_*, \phi) = \int f(y|b, \beta, \phi) f(b|\theta_*)\ db \\ &\propto \int \prod_{i=1}^{n} exp\left\{\frac{y_i\eta_i - b(\eta_i)}{\phi}\right\} exp\left\{-\frac{1}{2}\ b^T\boldsymbol{G}^{-1}b\right\} db \\ &= \int \prod_{i=1}^{n} exp\left\{\frac{y_i\eta_i - b(\eta_i)}{\phi} - \frac{1}{2}\ b^T\boldsymbol{G}^{-1}b\right\} db. \end{aligned} \tag{3.65}$$

Because the application of the Laplace approximation requires that b is known, one usually conducts a swing algorithm consisting of two steps:

Step 1

Prediction of b for given β, θ_*, and ϕ through a penalized Iteratively Reweighted least-squares algorithm (PIRLS) (see for details Appendix B):

$$\hat{b} = \underset{b}{argmax}\ \mathcal{L}(\beta, \phi, b, \theta_*). \tag{3.66}$$

The PIRLS is an extension of the Iteratively Reweighted least-squares algorithm used for the inference in GLMs (compare 3.2.1. See for details Fahrmeir et al. (2007)).

Step 2

The Laplace approximation $\hat{\mathcal{L}}(\beta, \theta_*, \phi)$ of $\mathcal{L}(\beta, \theta_*, \phi)$ is determined in $\hat{b}$, followed by the maximization of $\hat{\mathcal{L}}(\beta, \theta_*, \phi)$ with respect to β, θ_*, and ϕ via a pseudo-Newton algorithm (see for details Scheipl (2009)).

The two steps are iterated until convergence of the deviance, $-\mathcal{L}(\beta, \theta_*, \phi)$, is attained. For a detailed explanation of the Laplace approximation see Appendix B.

The Penalized Quasi-Likelihood (PQL)

The idea of the second method for inference in the GLMM – the Penalized Quasi-Likelihood approach – is to approximate the data such that the model can be displayed as a linear mixed model for pseudo-data. In a first step, the data y are approximated by their mean $E(y) = \mu$ and a random error term ε, with variance equal to $Var(y|b)$:

$$y \approx \mu + \varepsilon = h(\boldsymbol{X}\beta + \boldsymbol{Z}b) + \varepsilon. \tag{3.67}$$

Then, a first order Taylor expansion of the mean around $\boldsymbol{X}\hat{\beta} + \boldsymbol{Z}\hat{b}$ is carried out

$$\mu \approx h(\boldsymbol{X}\hat{\beta} + \boldsymbol{Z}\hat{b}) + h'(\boldsymbol{X}\hat{\beta} + \boldsymbol{Z}\hat{b})\ \boldsymbol{X}(\beta - \hat{\beta}) + h'(\boldsymbol{X}\hat{\beta} + \boldsymbol{Z}\hat{b})\ \boldsymbol{Z}(b - \hat{b}). \tag{3.68}$$

Thus, it follows

$$y \approx h(\boldsymbol{X}\hat{\beta} + \boldsymbol{Z}\hat{b}) + h'(\boldsymbol{X}\hat{\beta} + \boldsymbol{Z}\hat{b})\ \boldsymbol{X}(\beta - \hat{\beta}) + h'(\boldsymbol{X}\hat{\beta} + \boldsymbol{Z}\hat{b})\ \boldsymbol{Z}(b - \hat{b}) + \varepsilon \tag{3.69}$$

for the response. Considering the case of a canonical link function (i.e. $v(\cdot) = h'(\cdot)$) this yields

$$\begin{aligned} y &\approx \hat{\mu} + v(\hat{\mu})\ \boldsymbol{X}(\beta - \hat{\beta}) + v(\hat{\mu})\ \boldsymbol{Z}(b - \hat{b}) + \varepsilon \text{ or rather} \\ y &\approx \hat{\mu} + \hat{\boldsymbol{V}}\ \boldsymbol{X}(\beta - \hat{\beta}) + \hat{\boldsymbol{V}}\ \boldsymbol{Z}(b - \hat{b}) + \varepsilon, \end{aligned} \tag{3.70}$$

with $\hat{\boldsymbol{V}}$ denoting the diagonal matrix with elements $v(\hat{\mu}_i) = {}^{\partial h(\eta_i)}/_{\partial \eta}$ $(i = 1, \ldots, n)$. Consequently, multiplication of equation (3.70) by $\hat{\boldsymbol{V}}^{-1}$ from the left leads to the pseudo-data

$$\begin{aligned} \tilde{y} &\approx \hat{\boldsymbol{V}}^{-1}(y - \hat{\mu}) + \boldsymbol{X}\hat{\beta} + \boldsymbol{Z}\hat{b} \\ &\approx \boldsymbol{X}\beta + \boldsymbol{Z}b + \tilde{\varepsilon}, \end{aligned} \tag{3.71}$$

with $\tilde{\varepsilon} = \hat{\boldsymbol{V}}^{-1}\varepsilon$. Thus, the result is a linear mixed model for pseudo-data $\tilde{y}$ and it is now possible to apply the usual estimation methods for LMMs. It should be noted that, as the method uses an approximate likelihood (except for the LMMs), it leads to better results the closer the responses are to normal (Greven, 2009). The complete algorithm to estimate the interesting components via PQL is as follows[8]:

Initialization
Initial values $\hat{\beta}^{(0)}$, $\hat{\theta}_*^{(0)}$, and $\hat{b}^{(0)}$ are chosen.

Step 1
For given $\hat{\beta}$ and $\hat{\theta}_*$, the BLUP $\hat{b}$ and the resulting pseudo-data are computed.

Step 2
Having obtained the pseudo-data $\tilde{y}$, the linear mixed model (3.71) is fitted and the estimates for β and θ_* are updated.

Step 1 and Step 2 are iterated until convergence occurs.

Note that the name Penalized Quasi-Likelihood stems from the fact that it is based on a quasi-likelihood involving only the first and second (conditional) moments, plus a penalty term for the random effects (Greven, 2009). Other justifications exists for using PQL (see Greven (2009)).

[8] Greven (2009); Fahrmeir et al. (2007)

The (Adaptive) Gaussian Quadrature ((A)GQ)

The third method consists in approximating the integral of interest by a weighted sum:

$$\int \zeta(b)f(b)\,db \approx \sum_{q=1}^{Q} w_q \zeta(b_q). \tag{3.72}$$

Here, $\zeta(b)$ denotes $\zeta(b) := f(y|\beta, b, \phi)$, Q is the number of quadrature points b_q ($q = 1, \ldots, Q$) and w_q are appropriate weights. $f(b)$ is the density of the random effects, i.e. a Gaussian distribution. It is assumed that $\boldsymbol{G} = Cov(b)$ is the identity matrix, i.e. orthonormal random effects are used. Gaussian quadrature with quadrature points b_q that are solutions to the Qth order Hermite polynomial is not optimal for the densities (or pmfs) from the exponential family. Here, *adaptive* Gaussian quadrature (AGQ) is more appropriate. For AGQ, quadrature points are chosen more suitably and usually fewer points are required. However, the adaptive method is more time-consuming, as – in contrast to the Gaussian quadrature – the weights are not determined by the quadrature points. Instead, both b_q and w_q ($q = 1, \ldots, Q$) have to be calculated. As both depend on β and θ, they have to be updated in every step of the iteration.[9] The accuracy can be improved by increasing the number of quadrature points Q. Note that AGQ reduces to the Laplace approximation (3.2.5) for $Q = 1$. For further details see Greven (2009) and Scheipl (2009).

In addition to the presented approximation methods, it is possible to treat the random effects as missing data and to use the Expectation Maximization (EM)- algorithm for the estimation (Dempster et al., 1977). However, while the maximization steps are analytically accessible, the computation of the expectation step involves difficulties (see Greven (2009)). One possibility is to evaluate the E-steps using Monte-Carlo integration. Note that the algorithm depends on the specification of the type of missing data (Walker, 1996).

Another way of inference in the GLMM is to apply Bayesian inferential methods for which all parameters are assumed to be random variables and priors are put on each of them. The quantity of interest then becomes the posterior distribution which is accessed by Markov chain Monte Carlo (MCMC) methods (compare Fahrmeir et al. (2007); Greven (2009)).

[9] Again, a swing algorithm is used which iteratively estimates the random effects b and β and θ.

3.2.6 Implementation of the GLMM in R

Different R-packages include functions which allow the estimation of generalized linear mixed models. Particularly noteworthy are the two packages MASS and lme4.
The former provides the function glmmPQL which uses (as the name indicates) the PQL approach in order to fit a GLMM with multivariate normal random effects. It iteratively calls the lme-function of package nlme (see 3.1.7) and returns the fitted lme-model object for the working model at convergence (Wood, 2006). Note that the estimation of the variance components is (even asymptotically) downwardly biased and that the function works rather slowly (Scheipl, 2009). The PQL approach is moreover the default for the generalized case in function gamm {mgcv}, which is based on function gammPQL, a modification of glmmPQL {MASS} (compare Appendix E.1.2).
The latter package (lme4) provides a function glmer which uses the first approach – the swing algorithm consisting of PIRLS and the Laplace approximation (see 3.2.5). It is possible to use the adaptive Gauss-Hermite approximation (instead of the Laplacian approximation) by setting the parameter nAGQ – which specifies the number of quadrature points Q – greater than one[10]. This improves the approximation at the expense of speed as the Laplace approximation uses sparse matrix algorithms (Scheipl, 2009).
It should be remarked that function glmer does not allow anything else than unstructured or diagonal covariances $Cov(b_i)$ in contrast to the function glmmPQL {MASS} where – as for the function lme {nlme} – wide classes of covariance structures are available (Scheipl, 2009). Moreover, the function glmer assumes that the errors are independent and homoscedastic, i.e. $Cov(\varepsilon) = \sigma^2 \boldsymbol{I}_n$. In return, it allows the usage of nested and crossed data structures and large samples sizes which can impose problems for the function glmmPQL. For more details see Scheipl (2009).

[10] One standing for the Laplace approximation ($Q = 1$) which is a special case of AGQ.

Chapter 4

Penalized Splines

4.1 The Idea of Penalized Splines in General

In this section, the idea of non-parametric regression and in particular the conception of penalized spline smoothing will be concisely introduced (mainly) based on Chapter 7 in Fahrmeir et al. (2007). In this context, only univariate non-parametric regression, i.e. one metric scaled covariate x_i effecting the response variable y_i $(i = 1, \ldots, n)$, will be considered as this suffices to establish the connection between penalized splines and mixed models. The special case of Gaussianity will be considered separately as it will subsequently serve for the representation of penalized splines as mixed models (in Section 4.3). For more details on univariate as well as multivariate non-parametric regression, see Fahrmeir et al. (2007) and Heumann et al. (2010).

As seen in Subsection 3.2.1, covariates in the GLM (and therefore in particular in the LM) are assumed to take effect via a linear predictor $\eta = x^T\beta$. This can be very restrictive and is often not sufficient as the underlying function cannot always be approximated by polynomials, even in cases where the structure of the function is identifiable from a scatter plot.
The idea of non-parametric regression is to overcome this restriction by providing a more flexible class of models. These models do not assume a linear predictor, but extend this idea to the presumption of an *unknown smooth function* $s(x)$ which effects the mean of the response variable.
Whereas in classical parametric inference, families of densities or probability mass functions of the form

$$\{f(y|\theta), \theta \in \Theta \subseteq \mathbb{R}^p\}, \text{ with } p \text{ the number of covariates,}$$

are considered, in the non-parametric framework, the statistical model contains unknown functions which cannot be parameterized by a fixed number of parameters. Instead, one can think of an unknown "infinite dimensional" parameter s, which is an element of a function space (see Heumann et al. (2010)).
An important trade-off always goes along with the estimation of a regression function in non-parametric regression, namely the **bias-variance trade-off**, or rather the **conflict of under- versus overfitting** (compare Chapter 2).

This conflict results from the fact that, on the one hand, one aims to obtain a rather smooth function, coming along with a low variance, but a high bias. On the other hand, one seeks to model the data well and does not want to have too a great bias. Therefore, a compromise has to be found in order to adequately accomplish the estimation of s.

Consider a univariate non-parametric regression model. Let y_i denote the observations of the response variables and x_i those of the metric scaled covariates, $i = 1, \dots, n$. Similar to the GLM, two assumptions are made to define the model.

Definition 10. *Univariate Non-Parametric Regression Model*

Distribution

For given covariates x_i, the response variables y_i $(i = 1, \dots, n)$ are (conditionally) independent and the conditional density (or pmf) is a member of the one-parametric exponential family, thus

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

Structure

The conditional mean $E(y_i|x_i) = \mu_i$ is linked to the unknown smooth function s through

$$\mu_i = h(s(x_i)) \text{ or respectively } g(\mu_i) = s(x_i), \tag{4.1}$$

with $h(\cdot)$ the twice continuously differentiable response function and $g(\cdot) = h^{-1}(\cdot)$ its inverse function, the link function.

For a Gaussian response variable this corresponds to the definition:

$$y_i = s(x_i) + \varepsilon_i, \text{ with } \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \text{ for } i = 1, \dots, n. \tag{4.2}$$

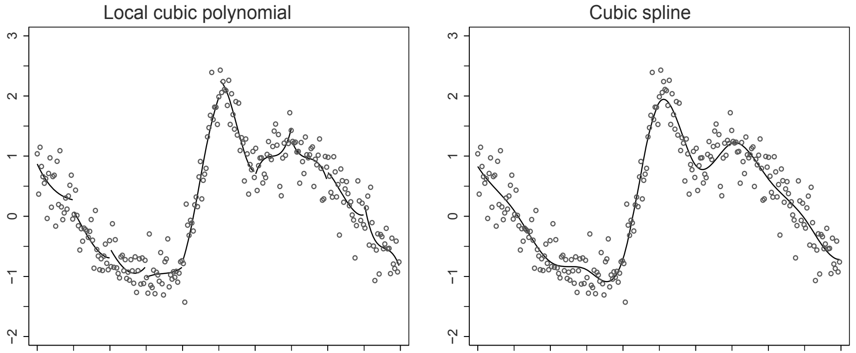

Figure 4.1: *Cubic Spline. The left figure shows piecewise polynomial regression. The domain is divided into 10 intervals of width 0.1 and to each interval a cubic polynomial is fitted. In the right figure, additional assumptions of global smoothing are added, yielding a cubic polynomial spline. Source: Fahrmeir et al. (2007)*

One idea to make the estimation more flexible than in polynomial regression is to decompose the co-domain of the covariate into intervals on which separate polynomials are estimated. Thus, instead of using a global model, the function $s(x)$ is approximated by locally defined polynomials. This proceeding is illustrated in the left graphic of Figure 4.1. In order to account for the requested smoothness, an assumption of global smoothness is added (see right graphic in Figure 4.1). This yields the definition of *polynomial splines* or *regression splines*.[1]

Definition 11. *Polynomial Spline*

A function $s : [a,b] \to \mathbb{R}$ is called polynomial spline of degree $d \geq 0$ to the knots $a = \kappa_1 < \ldots < \kappa_m = b$, if the following assumptions are fulfilled:

1. *$s(x)$ is $(d-1)$-times continuously differentiable. For d equal to 1 this corresponds to the condition that $s(x)$ is continuous, for $d = 0$ no smoothness requirements are imposed.*
2. *$s(x)$ is a polynomial of degree d on the intervals given by the knots $[\kappa_j, \kappa_{j+1})$ $\forall j$.*

It can be shown, that the set of all polynomials of degree d to the knots $\kappa_1 < \ldots < \kappa_m$ spans a $(l = m + d - 1)$-dimensional vector subspace of the vector space of all $(d-1)$-times continuously differentiable functions (for a proof see Hämmerlin and Hoffmann (1994)). Therefore, the polynomial spline $s(x)$ can be uniquely expressed through a linear combination of basis functions

$$s(x) = \sum_{j=1}^{l} \gamma_j B_j(x), \tag{4.3}$$

where $B_j(x)$ denote the basis functions and γ_j are coefficients $(j = 1, \ldots, l)$. In the Gaussian case this allows to display the model as a linear model of the form

$$y = \boldsymbol{U}\gamma + \varepsilon, \tag{4.4}$$

with parameter vector $\gamma = (\gamma_1, \ldots, \gamma_l)^T$ and design matrix $\boldsymbol{U}$, the matrix of basis functions evaluated in $x_1, \ldots, x_n$:

$$\boldsymbol{U} = \begin{pmatrix} B_1(x_1) & \ldots & B_l(x_1) \\ \vdots & & \vdots \\ B_1(x_n) & \ldots & B_l(x_n) \end{pmatrix}. \tag{4.5}$$

The concrete form of the design matrix depends on the choice of basis functions and will be given in the following. Due to the representation as a LM, the least-squares criterion can be minimized in order to estimate the parameter vector γ

$$LS(\gamma) = (y - \boldsymbol{U}\gamma)^T(y - \boldsymbol{U}\gamma) \underset{\gamma}{\to} \min. \tag{4.6}$$

[1] This definition is taken from Heumann et al. (2010).

In the more general case, the constructive form via basis functions enables one to construct a linear predictor as in GLMs via

$$\eta = \boldsymbol{U}\gamma, \tag{4.7}$$

with γ and $\boldsymbol{U}$ as in (4.4). Thus, for the general case, it is possible to estimate γ by maximizing the log-likelihood with respect to γ.

Some choices have to be made in order to specify the model. First of all, the degree of the regression spline can be specified. Second, the number and the location of the knots have to be chosen. And third, the kind of basis functions $B(x)$ has to be specified. All this has to be done, keeping in mind the bias-variance trade-off.
In practice, cubic polynomial splines are often chosen, as this yields a twice continuously differentiable function. The location of the knots is usually chosen either (a) visually (scatter plot), (b) chosen equidistantly, or (c) based on the quantiles of the observed covariate. The two most frequently employed basis functions will be introduced in Section 4.2.
Most important for the motivation of *penalized splines* is the difficulty to assign an adequate number of knots. The choice of the quantity of knots directly affects the diversity of displayable functions and the bias-variance trade-off, as the use of more knots leads to higher data fidelity, but holds a greater variance.
The idea of penalized splines is to deal with the uncertain choice of the number of knots by using many (∼20-40) equidistant knots to allow for modeling highly varying functions and adding a penalization term, which penalizes the variability. Note that in the Bayesian framework – which will not be discussed here –, penalization terms are replaced by smoothing priors.[2]
Thus, penalized splines can be seen as polynomial splines which account for the compromise of under- versus overfitting by preserving flexibility while penalizing data fidelity.
The penalty term is quadratic in the parameters γ and has the form

$$pen(\gamma, \boldsymbol{K}) = \lambda^{-1}\, \gamma^T \boldsymbol{K}\gamma, \tag{4.8}$$

where, matrix $\boldsymbol{K}$ denotes a *penalty matrix* and λ is referred to as the *smoothing parameter*. The concrete form of the penalty matrix $\boldsymbol{K}$ depends on the choice of basis functions (see Section 4.2). Thus, the degree of data fidelity is not controlled anymore by the choice of the quantity and the position of the knots, but instead by the smoothing parameter λ.

For a Gaussian distribution, the addition of the penalty term to the least-squares criterion yields the penalized least-squares criterion

$$LS_{pen}(\gamma, \lambda) = (y - \boldsymbol{U})^T(y - \boldsymbol{U}) + \lambda^{-1}\gamma^T\boldsymbol{K}\gamma \underset{\gamma}{\rightarrow} \min. \tag{4.9}$$

For the estimation of the parameters, one obtains (for given λ)

$$\hat{\gamma}_{pen} = (\boldsymbol{U}^T\boldsymbol{U} + \lambda^{-1}\boldsymbol{K})^{-1}\boldsymbol{U}y, \tag{4.10}$$

yielding the estimator

$$\hat{s}(x)_{pen} = \boldsymbol{U}\hat{\gamma}_{pen}. \tag{4.11}$$

[2] The interested reader is referred to Fahrmeir et al. (2007) and Heumann et al. (2010).

The estimator $\hat{\gamma}_{pen}$ is normally distributed with mean $(\boldsymbol{U}^T\boldsymbol{U}+\lambda^{-1}\boldsymbol{K})^{-1}\gamma$ and covariance $\sigma^2(\boldsymbol{U}^T\boldsymbol{U}+\lambda^{-1}\boldsymbol{K})^{-1}\boldsymbol{U}^T\boldsymbol{U}(\boldsymbol{U}^T\boldsymbol{U}+\lambda^{-1}\boldsymbol{K})^{-1}$. It is thus a biased estimator.

In the general case, the log-likelihood criterion is extended to a penalized log-likelihood criterion, given by

$$l_{pen}(\gamma,\lambda)=l(\gamma)-\frac{1}{2}\lambda^{-1}\ \gamma^T\boldsymbol{K}\gamma\underset{\gamma}{\rightarrow}\max, \tag{4.12}$$

with $l(\gamma)$ denoting the (unpenalized) log-likelihood. This criterion is composed by the usual log-likelihood, extended by $-1/2$ the penalty term. The negative sign stems from the fact that the penalized log-likelihood is to be maximized, in contrast to the penalized least-squares criterion which is minimized in the special case of Gaussianity. The factor $1/2$ is a scaling which is introduced as it disappears in the derivative of the penalized log-likelihood and eases further calculations. The derivation of the penalized log-likelihood yields the penalized score-function and the penalized Fisher matrix:

$$\mathcal{S}_{pen}(\gamma)=\mathcal{S}(\gamma)-\lambda^{-1}\boldsymbol{K}\gamma, \tag{4.13}$$

$$\mathcal{F}_{pen}(\gamma)=\mathcal{F}(\gamma)+\lambda^{-1}\boldsymbol{K}. \tag{4.14}$$

Here, $\mathcal{S}(\gamma)$ denotes the (unpenalized) score-function and $\mathcal{F}(\gamma)$ is the (unpenalized) Fisher matrix.

Similarly to the estimation in the GLM, the basis coefficients γ_j $(j=1,\ldots,l)$ are estimated numerically, e.g. via a penalized Fisher-Scoring algorithm (for given λ). Note that in general the distribution of the estimator is inaccessible (Heumann et al., 2010).

In order to obtain an estimator for the basis coefficients – and thus for the regression function $s(x)$ – the smoothing parameter λ which controls the amount of smoothing has to be chosen as well.

The influence of λ is as follows:

$\lambda\rightarrow 0$: The penalized least-squares or rather the penalized log-likelihood criterion is fully dominated by the penalty term.

$\lambda\rightarrow\infty$: The penalty term has a very small influence on the estimation, i.e. the penalized least-squares criterion almost corresponds to the least-squares criterion used in the linear model. The same holds for the penalized log-likelihood, which almost equates the log-likelihood criterion for GLMs.

The smoothing parameter λ can be chosen in various ways. First, an “optimal” smoothing parameter can be obtained by minimizing the mean squared error (MSE), which is a compromise itself of the bias and the variance. A second option is to minimize the (Generalized) Cross-validation criterion ((G)CV) (for details see Fahrmeir et al. (2007); Heumann et al. (2010)). And third, the smoothing parameter can be determined on the basis of the representation of penalized splines as mixed models. This will be elaborated on in the following as this method establishes the connection between the mixed models, the AIC, and penalized splines and will be used in the simulations in Chapter 6.

4.2 Basis functions

As seen in the previous section, the choice of the basis used for the representation of the regression spline $s(x)$ has an influence on the penalty matrix $\boldsymbol{K}$ – and thus on the entire penalty term – and on the design matrix $\boldsymbol{U}$.
Two frequently applied bases will be introduced in the following. The *truncated power series (TP-) basis* and the *B-Spline basis.*

4.2.1 The TP-basis

Definition 12. *Truncated Power Series Basis of Degree l*

The $l = m + d - 1$ linearly independent basis functions of the TP-basis of degree d to the set of knots $\{\kappa_1, \ldots, \kappa_m\}$ are given by

$$B_1(x) = 1,\ B_2(x) = x, \ldots,\ B_{d+1}(x) = x^d,$$
$$B_{d+2}(x) = (x - \kappa_2)_+^d, \ldots,\ B_l(x) = (x - \kappa_{m-1})_+^d,$$

with $(x - \kappa_i)_+^d = \begin{cases} (x - \kappa_i)^d, & x \geq \kappa_i \\ 0, & \textit{otherwise.} \end{cases}$

Thus, the basis is constructed of two parts, modeling a global polynomial form through the first $d + 1$ basis functions and deviations of these polynomials through the $m - 2$ truncated powers. This allows to modify the coefficients of the highest polynomial in each knot in order to make the function more flexible. The parameters can be interpreted as the modification of the slope in the knots. Figure 4.2 illustrates the construction of TP-basis functions for an example of a polynomial spline of degree $d = 1$.
Yet, as the function should not be too coarse, the idea is to penalize the coefficients of the basis functions of the truncated powers, which allows for high variability, yielding the penalization matrix

$$\boldsymbol{K} = diag(\underbrace{0, \ldots, 0}_{(d+1)}, \underbrace{1, \ldots, 1}_{(m-2)}). \tag{4.15}$$

In the case of truncated power series basis, the design matrix $\boldsymbol{U}$ has the form

$$\boldsymbol{U} = \begin{pmatrix} 1 & x_1 & \ldots & x_1^d & (x_1 - \kappa_2)_+^d & \ldots & (x_1 - \kappa_{m-1})_+^d \\ \vdots & & & & \vdots & & \\ 1 & x_n & \ldots & x_n^d & (x_n - \kappa_2)_+^d & \ldots & (x_n - \kappa_{m-1})_+^d \end{pmatrix}. \tag{4.16}$$

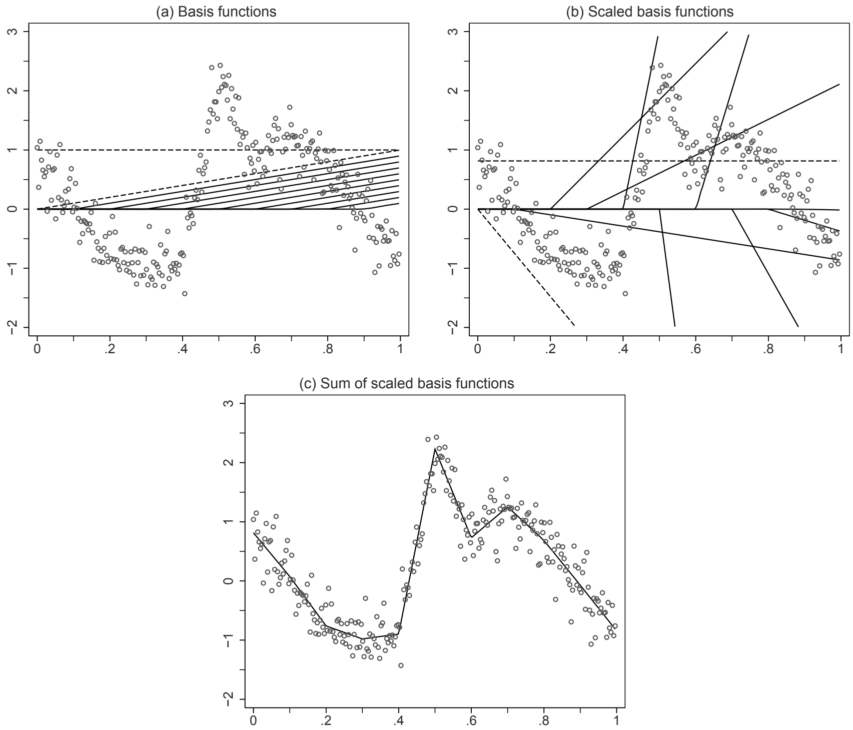

Figure 4.2: *Construction of TP-basis functions for linear polynomials ($d = 1$). The broken lines in figure (a) show the functions of a global polynomial of degree 1. The solid lines illustrate the truncated polynomials. These functions are scaled by the coefficients γ, yielding (b) and then added up resulting in (c). The horizontal line at $y \approx 0.8$ in (b) corresponds to the global constant γ_1. In these figures, equidistant knots with width 0.1 were used. Source: Fahrmeir et al. (2007)*

4.2.2 The B-Spine basis

Definition 13. *B-Spline Basis of Degree d*

The $l = m + d - 1$ linearly independent basis functions of the B-Spline basis of degree d to the set of knots $\{\kappa_1, \dots, \kappa_m\}$ are recursively given by

$$d = 0 : B_j^0(x) = 1_{[\kappa_j,\kappa_{j+1})}(x) = \begin{cases} 1, & \kappa_j \leq x \leq \kappa_{j+1}, \\ 0, & \textit{elsewhere}, \end{cases} \quad j = 1, \dots, l-1,$$

$$d > 0 : B_j^d(x) = \frac{x - \kappa_j}{\kappa_{j+d} - \kappa_j} B_j^{d-1}(x) + \frac{\kappa_{j+d+1} - x}{\kappa_{j+d+1} - \kappa_{j+1}} B_{j+1}^{d-1}(x), \; j = -d+1, \dots, m-1.$$

Note that 2 d additional knots outside of the domain are required for the calculation. A suitable change of indices yields the $l = m + d - 1$ linearly independent basis functions $B_j(x) = B^d_{j+d}(x)$ $(j = 1, \dots, l)$ (Konrath, 2009).
In words, each basis function is a piecewise $(d-1)$-times continuously differentiable, non-negative polynomial of degree d reaching over $d + 2$ knots and overlapping with $2d$ adjoining basis functions. Hence, the B-Spline basis represents a **local basis** consisting of polynomial pieces composed sufficiently smooth. For equidistant knots, all basis functions have the same shape and are only shifted on the x-axis. The shape of B-spline bases with equidistant and unevenly distributed knots is shown in Figure 4.3.

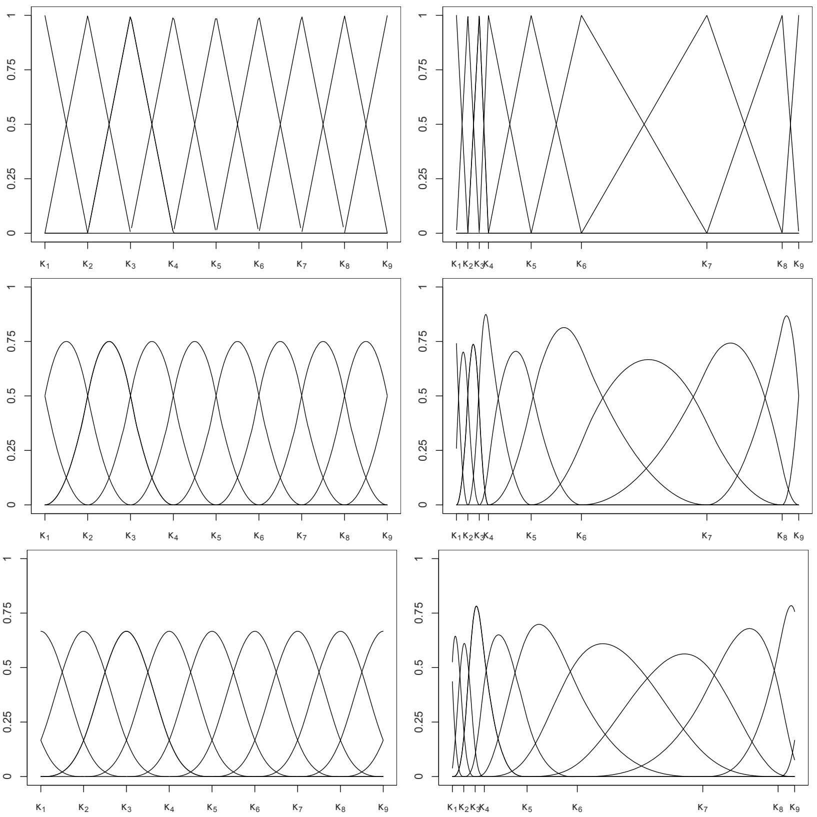

Figure 4.3: *B-Spline bases of degree l=1,2,3 for equidistant knots (left) and unevenly distributed knots (right). Source: Fahrmeir et al. (2007)*

Using a B-Spline basis, the design matrix has the form

$$\boldsymbol{U} = \begin{pmatrix} B^d_{-d+1}(x_1) & \dots & B^d_{m-1}(x_1) \\ \vdots & & \vdots \\ B^d_{-d+1}(x_1) & \dots & B^d_{m-1}(x_n) \end{pmatrix}. \tag{4.17}$$

As the B-Spline basis is a local basis, the quantity $\boldsymbol{U}^T\boldsymbol{U}$ is a banded matrix of bandwidth d, which makes calculations with it numerically more efficient than the use of a TP-basis. Its numerical properties are the reason why the B-Spline basis is often preferred over the TP-basis and implemented in statistical programs, such as R. Figure 4.4 shows schematically the estimation of a B-spline based on simulated data.

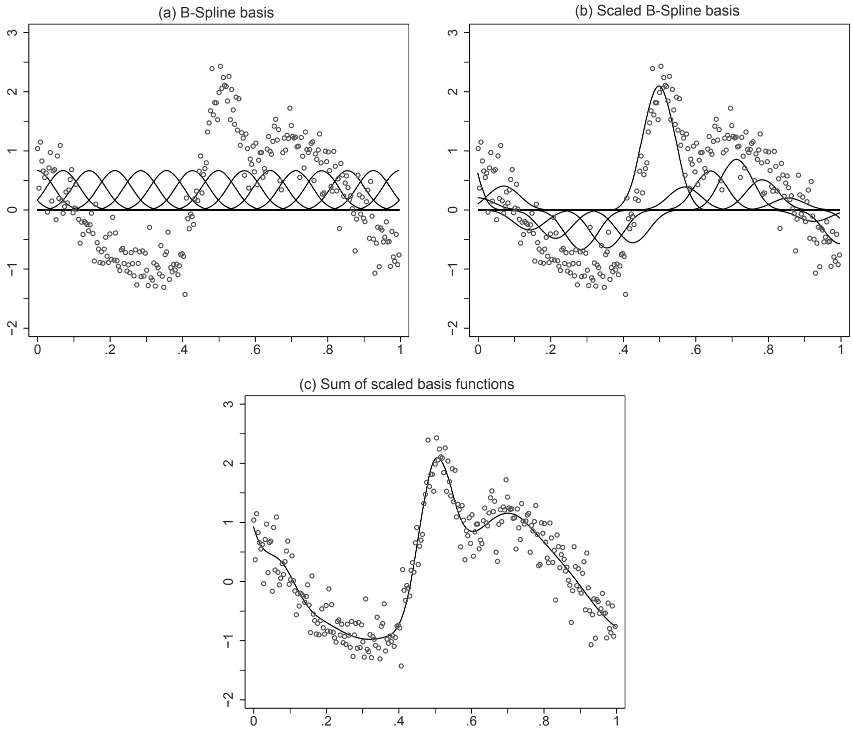

Figure 4.4: *Estimation of a non-parametric effect via B-Splines. In figure (a), a B-Spline basis of degree 3 is computed to a given number of knots. The basis functions are then scaled (figure (b)) by using the least-squares estimator $\hat{\gamma}$. Figure (c) shows the final estimation resulting from added scaled basis functions. Source: Fahrmeir et al. (2007)*

In general, the integral of the kth derivative of a function can be seen as a measure for its variability. This can be used in order to define the penalty term for the representation with B-Splines. Especially the squared derivative is frequently used. For a B-Spline basis, a penalty term based on the integral of the squared derivative has the form

$$\lambda^{-1} \int (s''(x))^2 \, dx = \lambda^{-1} \sum_{i=1}^{l} \sum_{j=1}^{l} \gamma_i \gamma_j \int B_i''(x) B_j''(x) \, dx = \lambda^{-1} \gamma^T \boldsymbol{K} \gamma, \tag{4.18}$$

with $s''(x)$ the second derivative of $s(x)$ and $B_i''(x)$ the second derivative if $B_i(x)$. The entries of the penalty matrix $\boldsymbol{K}$ are determined from the derivatives of the basis functions. For equidistantly chosen knots, the kth derivatives can be represented by the kth order differences Δ^k of the parameters γ. The differences are recursively defined as

$$\begin{aligned} \Delta^1 \gamma_j &= \gamma_j - \gamma_{j-1} \\ &\vdots \\ \Delta^k \gamma_j &= \Delta^{k-1} \gamma_j - \Delta^{k-1} \gamma_{j-1}. \end{aligned} \tag{4.19}$$

The penalty then has the form

$$pen(\gamma, \boldsymbol{K}) = \lambda^{-1} \sum_{j=k+1}^{l} (\Delta^k \gamma_j)^2 = \lambda^{-1} \gamma^T \boldsymbol{K} \gamma, \tag{4.20}$$

with the penalty matrix

$$\boldsymbol{K} = \boldsymbol{D}^T \boldsymbol{D}, \tag{4.21}$$

and $\boldsymbol{D}$ denoting the difference operator matrix which is recursively defined as

$$\begin{aligned} \underbrace{\boldsymbol{D}_1}_{((l-1)\times l)} &= \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}, \\ \boldsymbol{D}_k &= \boldsymbol{D}_1 \boldsymbol{D}_{k-1}. \end{aligned} \tag{4.22}$$

For $k = 1$, the penalty matrix $\boldsymbol{K}$ has the form

$$\underbrace{\boldsymbol{K}}_{(l\times l)} = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}.$$

The idea of this penalty term is that neighboring, weighted basis functions should not differ much in their mean in order to obtain a global function which is not too coarse. Therefore, the corresponding coefficients are penalized. Matrices of kth order differences penalize deviations of degree $k - 1$, i.e., for $\lambda \to \infty$ one obtains a polynomial of degree $k - 1$ if the degree of the spline is at least as great as d. Typically, second or third order differences are used.

Apart from the numerical properties, one advantage using B-Splines is that the order of differences k and the degree of the polynomial spline d can be chosen separately. This provides more flexibility.

Usually, penalized splines with a B-Spline basis are referred to as *P-Splines*. Note that some authors use this term to denote penalized splines in general, not necessarily with a B-Spline basis. In this work, only penalized splines with a B-Spline basis will be termed P-Splines.

4.3 Penalized Splines as Mixed Models

In the following, it will be shown how penalized splines can be represented as mixed models. This allows to take advantage of inferential methods for mixed models and induces implementational simplifications in the estimation. It should be pointed out that, although penalty approaches can be displayed in the mixed model form, their structure is not the same. One distinction is that penalized splines do not contain any grouping structure (Konrath, 2009). At first, the representation of Gaussian penalized splines with TP-basis will be demonstrated, followed by a more general approach. The following section is based on Chapter 5 in Konrath (2009) and on Chapter 7 in Fahrmeir et al. (2007).

Consider a penalized spline with TP-basis and $y|x$ normally distributed with mean $s(x)$ and covariance $\sigma^2 \boldsymbol{I}_n$. As for the TP-bases only the coefficients of the basis functions of the truncated powers are penalized, the penalized least-squares criterion can be written as

$$LS_{pen}(\gamma, \lambda) = (y - \boldsymbol{U}\gamma)^T (y - \boldsymbol{U}\gamma) + \lambda^{-1} \sum_{j=d+2}^{l} \gamma_j^2. \tag{4.23}$$

In order to link this to mixed models, the parameter vector γ is decomposed into a first sub-vector consisting of the parameters of the polynomial which are not penalized

$$\beta = (\gamma_1, \dots, \gamma_{d+1})^T$$

and a second sub-vector comprising the parameters of the truncated powers

$$b = (\gamma_{d+2}, \dots, \gamma_l)^T.$$

Let now $\boldsymbol{X}$ and $\boldsymbol{Z}$ denote the respective design matrices, such that for the entire design matrix $\boldsymbol{U} = [\boldsymbol{X}, \boldsymbol{Z}]$ applies. Then, the penalized least-squares criterion (4.23) can be reformulated as

$$LS_{pen}(\beta, b, \lambda) = (y - \boldsymbol{X}\beta - \boldsymbol{Z}b)^T (y - \boldsymbol{X}\beta - \boldsymbol{Z}b) + \lambda^{-1} b^T b. \tag{4.24}$$

As seen in equation (3.23) in Chapter 3.1.5, the criterion to minimize in the estimation of an LMM has the form

$$GLS_{pen}(\beta, b) = (y - \boldsymbol{X}\beta - \boldsymbol{Z}b)^T \boldsymbol{R}^{-1} (y - \boldsymbol{X}\beta - \boldsymbol{Z}b) + b^T \boldsymbol{G}^{-1} b.$$

For $Cov(\varepsilon) = \boldsymbol{R} = \sigma^2 \boldsymbol{I}_n$ and $Cov(b) = \boldsymbol{G} = \tau^2 \boldsymbol{I}_m$, this reduces to

$$GLS_{pen}(\beta, b) = \sigma^{-2} (y - \boldsymbol{X}\beta - \boldsymbol{Z}b)^T (y - \boldsymbol{X}\beta - \boldsymbol{Z}b) + \tau^{-2} b^T b,$$

which is equal to

$$GLS_{pen}(\beta, b) = \sigma^{-2} \left\{ (y - \boldsymbol{X}\beta - \boldsymbol{Z}b)^T (y - \boldsymbol{X}\beta - \boldsymbol{Z}b) + \frac{\sigma^2}{\tau^2} b^T b \right\}.$$

Thus, as the minimization with respect to b and β is independent of σ^2, the penalized least-squares criterion for LMMs is equivalent to that for Gaussian penalized splines with TP-basis, by interpreting

- β, which models the subspace of polynomials of degree d, as vector of fixed effects in the LMM,
- b, which models any deviation from polynomials of degree d, as vector for random effects in the LMM,
- and by setting the smoothing parameter λ as the ratio of the variance of the random effects to the error variance, i.e. ${}^{\tau^2}/{}_{\sigma^2}$.

The choice of an optimal smoothing parameter λ can therefore be made by estimating σ^2 and τ^2 in the mixed model framework (compare Section 3.1.5), yielding $\hat{\lambda} = {}^{\hat{\tau}^2}/{}_{\hat{\sigma}^2}$.

Note that in the literature (see for example Fahrmeir et al. (2007)), the smoothing term is often alternatively defined as

$$pen(\lambda, \boldsymbol{K}) = \lambda \sum_{j=d+2}^{l} \gamma_j^2,$$

and therefore λ is estimated as $\hat{\lambda} = {}^{\hat{\sigma}^2}/{}_{\hat{\tau}^2}$. However, in this work the inverse formulation will be used, as it is advantageous for the reason that the smoothing parameter is zero, iff the random effects variance is equal to zero ($\lambda = 0 \Leftrightarrow \tau^2 = 0$).

Now, having shown that univariate Gaussian penalized splines with TP-basis can be represented as mixed models, this finding will be extended to more general penalization approaches (still for univariate smooth terms and the Gaussianity assumption).
Consider approaches for which the penalty term has the form

$$LS_{pen}(\gamma, \lambda) = (y - \boldsymbol{U}\gamma)^T(y - \boldsymbol{U}\gamma) + \lambda^{-1}\gamma^T\boldsymbol{K}\gamma. \tag{4.25}$$

In analogy to the case of the truncated power series basis, the aim is to construct a linear mixed model of the form

$$y = \boldsymbol{U}\gamma + \varepsilon,$$

with

$$\varepsilon \sim \mathcal{N}(0, \sigma^2\boldsymbol{I}_n) \text{ and } \gamma \sim \mathcal{N}(0, \tau^2\boldsymbol{K}^{-1}),\ \tau^2 = \lambda\sigma^2. \tag{4.26}$$

However, for general penalization approaches, the penalty matrix $\boldsymbol{K}$ does not necessarily have full rank, e.g. for P-Splines (B-Spline basis), where $\boldsymbol{K}$ is given by $\boldsymbol{D}^T\boldsymbol{D}$. Thus, the inverse matrix $\boldsymbol{K}$ does not always exist which implies that the resulting density of

γ is partially improper[3] and can hence not be normalized. A representation of a general penalized spline as a mixed model has thus to be done differently than for the TP-basis. In the LMM, the partial improperness dissolves into a non-informative[4] distribution for the fixed effects and a proper Gaussian distribution for the random effects. In order to achieve such a suitable decomposition for generalized penalization approaches, the parameter vector γ has to be decomposed into two sub-vectors with respect to the rank drop of $\boldsymbol{K}$. First, the $(l-\nu)$-dimensional vector β and second, the ν-dimensional vector b, such that

$$\gamma = \underbrace{\boldsymbol{X}}_{(l\times(l-\nu))} \beta + \underbrace{\boldsymbol{Z}}_{(l\times\nu)} b. \tag{4.27}$$

For $\boldsymbol{X}$ and $\boldsymbol{Z}$ chosen such that the penalty term can be written as

$$pen(\gamma, \boldsymbol{K}) = \lambda^{-1}\gamma^T\boldsymbol{K}\gamma = \lambda^{-1}b^Tb,$$

β can be interpreted as a vector of fixed and b as a vector of random effects. For details on the decomposition, see Konrath (2009) and Fahrmeir et al. (2007).

With the transformations $\widetilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{X}$ and $\widetilde{\boldsymbol{Z}} = \boldsymbol{U}\boldsymbol{Z}$, equation (4.25) can be represented as a mixed model

$$y = \boldsymbol{U}\gamma + \varepsilon = \boldsymbol{U}(\boldsymbol{X}\beta + \boldsymbol{Z}b) + \varepsilon = \widetilde{\boldsymbol{X}}\beta + \widetilde{\boldsymbol{Z}}b + \varepsilon. \tag{4.28}$$

Here, β denote the fixed effects and b the random effects with $b \sim \mathcal{N}(0, \tau^2\boldsymbol{I}_\nu)$.
Note that – strictly speaking – in the representation of penalized splines as mixed models, a part of the vector γ is transfered into random effects and thus does not (formally) represent a fixed parameter anymore. The representation should thus rather be seen as an algorithmic artifact than as a real reformulation. In the Bayesian framework, this does not pose a problem as all parameters are assumed to be random in the first place.

In the simulations in Chapter 6, mixed model representation for P-Splines (i.e B-Spline basis and difference penalty) will be considered. In this context, the fact that penalization of differences of order k penalize deviations of the fitted smooth term from a polynomial of degree $(k-1)$ will be used.
The exact representation of penalized splines with a B-Spline basis can be found in Fahrmeir et al. (2004) and Eilers and Marx (1996) and. For generalizations to the non-Gaussian case see Kneib (2003).

For the practical realization, the statistical software `R` offers the package `mgcv` which includes a function `gamm` that can be used to fit penalized splines based on the representation of mixed models (compare Appendix E.1.2).

[3] A distribution is improper if its total probability equals infinity rather than one (Ruppert et al., 2003).

[4] See Fahrmeir et al. (2007).

Chapter 5

The AIC in Mixed Models

In contrast to the linear model, for which the Akaike information criterion is uniquely defined using the maximized log-likelihood and the number of parameters k in the model (which equal the degrees of freedom), no equivalent definition for mixed models exists. This has two reasons. One reason is that two perspectives exist for mixed models (see Subsection 3.1.4) which affects the first part of the AIC. In other words, one has to decide if the AIC should be based on either the marginal or the conditional likelihood. The resulting AICs are denoted as the *marginal AIC (mAIC)* and the *conditional AIC (cAIC)*. The second reason is that there is no unique definition of the degrees of freedom for mixed models which affects the second part of the AIC. Instead, several suggestion for an extensions of the concept of degrees of freedom to mixed models were made which all simplify to the degrees of freedom under the linear model.
For the linear mixed model, Greven and Kneib (2010) showed that the AIC resulting from the marginal model is not an adequate criterion for the selection of random effects for two reasons. First, its derivation assumes independent and identically distributed observations which is not the case for mixed models. Second, the derivation of the mAIC assumes an open parameter space. The parameter space for mixed models however is non-open due to the restrictions on the variance parameters of the random effects. As the LMM is a special case of generalized linear mixed models, this clearly applies to GLMMs as well.
Despite the inadequacy of the marginal AIC, it has been – and still is – commonly used for the selection of random effects in mixed models, as it is returned by statistical software such as `R` and `SAS` (compare the results[1] of the simulation studies in Subsection 6.1.4 and Subsection 6.2.4).
Vaida and Blanchard (2005) and Greven and Kneib (2010) showed for the LMM that the conditional AIC is more adequate for the selection of random effects. Therefore, the main focus in this work lies on the construction of an AIC using the conditional log-likelihood.

In the next section, first the AIC of the LM will be defined. A brief introduction of the mAIC will be given, resulting in an motivation for “the” cAIC. It follows an introduction of the conditional Akaike information and a detailed presentation of different conditional Akaike information criteria for the LMM (in Subsection 5.1.2). Two generalizations of conditional AICs for the GLMM will be introduced in Section 5.2.

[1] The results showed that the function `logLik.gamm{mgcv}` and the function `logLik.lme{nlme}` both automatically return the marginal AIC.

5.1 The AIC in Linear Mixed Models

First consider the standard linear model (3.1). The AIC in the linear model is defined as

$$AIC = -2\ log \left\{ \mathcal{L}(\hat{\psi}|y) \right\} + 2k,$$

with the maximized likelihood

$$\mathcal{L}(\hat{\psi}|y) = \frac{1}{(2\pi\hat{\sigma}^2)^{\frac{n}{2}}} \exp\left\{ -\frac{1}{2\hat{\sigma}^2}\ (y - \boldsymbol{X}\hat{\beta})^T(y - \boldsymbol{X}\hat{\beta}) \right\} \tag{5.1}$$

and k the number of parameters which is equal to the degrees of freedom of the linear model. ψ denotes the vector of unknown parameters $(\beta^T, \sigma^2)^T$.

Thus, except for the likelihood term, which differs depending on whether maximum likelihood or restricted maximum likelihood estimation is used for the estimation of the error variance σ^2, the AIC is uniquely defined in the linear model.
When using ML estimation, the error variance is estimated as[2]

$$\hat{\sigma}^2_{ML} = \frac{(y - \boldsymbol{X}\hat{\beta})^T(y - \boldsymbol{X}\hat{\beta})}{n}, \tag{5.2}$$

and under REML it is as estimated as[3]

$$\hat{\sigma}^2_{REML} = \frac{(y - \boldsymbol{X}\hat{\beta})^T(y - \boldsymbol{X}\hat{\beta})}{n - p}. \tag{5.3}$$

For the LM, no distinction is made between a marginal and a conditional model formulation (as no random effects are assumed). In contrast, for the LMM it plays an important role whether the definition of the AIC is based on the marginal or the conditional log-likelihood. This will be the subject of the next Subsection.

5.1.1 The marginal AIC versus the conditional AIC in LMMs

The AIC arising from the marginal distribution (cf. 3.1.4)

$$y \sim \mathcal{N}(\boldsymbol{X}\beta, \boldsymbol{V}) \tag{5.4}$$

has the form[4]

$$mAIC_{ML} = -2\ log \left(f(y|\hat{\beta}, \hat{\theta}) \right) + 2(p + q + 1) \text{ for ML estimation and} \tag{5.5}$$

$$mAIC_{REML} = -2\ log \left(f(\boldsymbol{A}^T y|\hat{\theta}) \right) + 2(q + 1) \text{ for REML estimation,} \tag{5.6}$$

[2] see Fahrmeir et al. (2007)

[3] see Fahrmeir et al. (2007)

[4] Greven and Kneib (2010)

with θ again denoting the vector of unknown variance parameters as in Subsection 3.1.5 and $\hat{\theta} = \hat{\theta}(y)$ the estimator of θ. The quantity $log\left(f(y|\hat{\beta}, \hat{\theta})\right)$ is the maximized marginal log-likelihood and $log\left(f(\boldsymbol{A}^T y|\hat{\theta})\right)$ denotes the maximized restricted log-likelihood with $\boldsymbol{A}$ the linear contrast matrix (compare 3.1.5).
Note that because the error contrasts $\boldsymbol{A}^T y$ depend on the design matrix $\boldsymbol{X}$, a model comparison via the marginal AIC using REML can only be adequately accomplished when it is ensured that the fixed effects do not differ.[5,6]

Greven and Kneib (2010) showed that the mAIC is not an asymptotically unbiased estimator for the Akaike information (2.6). The mAIC is proven to be inadequate for two reasons. First, observations in the linear mixed model are not independent due to the correlation caused by the random effects. And second, the parameter space for the marginal model is not a transformation of $\mathbb{R}^k$.
Considering the case of conditional independence $\boldsymbol{R} = \sigma^2 \boldsymbol{I}_n$ and of one unknown random effects variance component $\boldsymbol{G} = \tau^2 \boldsymbol{\Sigma}$, with $\boldsymbol{\Sigma}$ known, Greven and Kneib (2010) showed that the inequality

$$E_y(mAIC) > -2E_y\left[E_x\left[log\left\{f(x|\hat{\psi}(y))\right\}\right]\right] \tag{5.7}$$

holds with $\psi = (\beta^T, \sigma^2, \lambda)^T$ and $\lambda = {}^{\tau^2}\!/_{\sigma^2}$. Thus, the mAIC favors smaller models without random effects compared to an asymptotically unbiased estimator of the Akaike information. As the bias depends on the unknown true variance parameters, no simple correction can be accomplished (Greven and Kneib, 2010).

Note that there is a close relationship between comparing a model with LMM ($\tau^2 \geq 0$) with its nested linear model ($\tau^2 = 0$) using the marginal Akaike information criterion and testing for a random effects variance. The interested reader is referred to Greven and Kneib (2010).

Vaida and Blanchard (2005) suggested the use of an AIC based on the conditional likelihood of the linear mixed model, with the number of parameters related to the effective degrees of freedom of Hodges and Sargent (2001) to account for shrinkage in the random effects. They defined a conditional version of the Akaike information and derived an (asymptotically[7]) unbiased estimator for this quantity.

As the marginal AIC is proven non-adequate, in the following the focus lies on conditional Akaike information criterion.

[5] Greven and Kneib (2010)

[6] This can be achieved by a re-parametrization of the data.

[7] Note thatVaida and Blanchard (2005) also provided a finite sample criterion, i.e. an unbiased estimator for the cAI. But for ease of presentation the asymptotic version will be considered here only.

5.1.2 Conditional AICs in LMMs

For model selection based on the conditional model formulation (cf. 3.1.4),

$$\begin{aligned} y|b &\sim \mathcal{N}(\boldsymbol{X}\beta + \boldsymbol{Z}b, \boldsymbol{R}) \\ b &\sim \mathcal{N}(0, \boldsymbol{G}), \end{aligned}$$

Vaida and Blanchard (2005) defined the conditional analogue of the Akaike information as follows.

Definition 14. *Conditional Akaike Information (cAI)*

$$\begin{aligned} cAI &= -2\ E_{y,b}\left[E_{z|b}\left[log\left(f(z|\hat{\theta}(y), \hat{b}(y))\right)\right]\right] \\ &= -\int\int\int 2\ log\left(f(z|\hat{\theta}(y), \hat{b}(y))\right) g(z|b)g(y,b)\ dz\ dy\ db, \end{aligned} \tag{5.8}$$

where $g(y, b) = g(y|b)g(b)$ denotes the joint distribution of y and the random effects vector b. θ is the vector of unknown variance parameters as before.

Like in the non-conditional case, this quantity (cAI) is unobservable and has to be estimated (Vaida and Blanchard, 2005). In the rest of this Subsection, several proposals on this estimation will be compared.

In this context, two distinctions are made:

1. Considering the case of *known* versus *unknown* covariance of the random effects $\boldsymbol{G}$.

2. Assuming the error variance to be *known* or *unknown*.

Consider in the following the linear mixed model with conditional independence, i.e. $\boldsymbol{R} = \sigma^2\boldsymbol{I}_n$. Let $\boldsymbol{G}_* := \sigma^{-2}\boldsymbol{G}$. The covariance of y thus becomes

$$Cov(y) = \boldsymbol{V} = \sigma^2\boldsymbol{I}_n + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T = \sigma^2(\boldsymbol{I}_n + \boldsymbol{Z}\boldsymbol{G}_*\boldsymbol{Z}^T) =: \sigma^2\boldsymbol{V}_*. \tag{5.9}$$

Further, θ_* will in the following denote the q parameters in $\boldsymbol{G}_*$ and $\theta = (\sigma^2, \theta_*)$ again stands for the parameter vector which contains all unknown parameters in the covariance matrices $\boldsymbol{G}$ and $\boldsymbol{R} = \sigma^2\boldsymbol{I}_n$. When emphasizing the dependence of θ and accordingly θ_* on the data y, the notation $\hat{\theta}(y)$ and $\hat{\theta}_*(y)$ is used.

The conventional cAIC in LMMs

The first suggestion for an estimator of the conditional Akaike information was contributed by its initiators, Vaida and Blanchard (2005). For the case of **known variance components**, i.e. $\boldsymbol{G}$ and thus θ_* known, and **known error variance $\boldsymbol{\sigma^2}$**, they derived an asymptotically unbiased estimator for the cAI which will be further referred to as the *conventional* cAIC (ccAIC).

Definition 15. *Conventional cAIC (ccAIC) for Known Error Variance and Known* $\boldsymbol{G}$

$$ccAIC = -2\ log\left(f(y|\hat{\beta},\hat{b},\hat{\theta})\right) + 2\rho, \tag{5.10}$$

where

$$log\left(f(y|\hat{\beta},\hat{b},\hat{\theta})\right) = -\frac{n}{2}\ log(2\pi) - \frac{n}{2}\ log\left(\hat{\sigma}^2\right) - \frac{1}{2\hat{\sigma}^2}\ (y-\boldsymbol{X}\hat{\beta}-\boldsymbol{Z}\hat{b})^T(y-\boldsymbol{X}\hat{\beta}-\boldsymbol{Z}\hat{b}) \tag{5.11}$$

is the conditional log-likelihood for y, given β, b, and θ, evaluated at the estimated/predicted quantities $(\hat{\beta},\hat{b},\hat{\theta})$ based on maximum likelihood or restricted maximum likelihood estimation. ρ are the effective degrees of freedom defined by Hodges and Sargent (2001), measured as the trace of the hat matrix which maps y onto $\hat{y} = \boldsymbol{X}\hat{\beta} + \boldsymbol{Z}\hat{b}$.
The hat matrix $\boldsymbol{H}_1$ has the form

$$\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{G}_*^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{Z} \end{pmatrix}. \tag{5.12}$$

For the derivation of the hat matrix see Appendix A.

Note that $\boldsymbol{H}_1$ itself is – unlike in the linear model – not a projection matrix, but it is the top-left of a projection matrix (Vaida and Blanchard, 2005).

An extension to the case of **unknown error variance $\boldsymbol{\sigma^2}$** can be achieved for large sample size by setting

$$ccAIC = -2\ log\left(f(y|\hat{\beta},\hat{b},\hat{\theta})\right) + 2\ (\rho+1). \tag{5.13}$$

Note that this only holds for the case of **known covariance matrix $\boldsymbol{G}$**. In real data analysis, however, $\boldsymbol{G}$ is usually unknown. In practice, Vaida and Blanchard recommended applying their ccAIC using a plug-in estimator for $\boldsymbol{G}$, arguing that the differences between an estimator of ρ and the true ρ itself is negligible asymptotically.
However, Greven and Kneib (2010) disproved this argument by showing that ignoring the uncertainty in the estimation of the covariances of the random effects, $\boldsymbol{G}$, leads to a particular bias, i.e. the more complex model is always favored unless the covariance of the random effect is estimated to be exactly zero, in which case the ccAIC does not distinguish between the two models. Thus, the conventional cAIC does not allow a distinction when a random effect that is predicted to be small, but not exactly zero, should be included into the model. This is due to the fact that the ccAIC estimates the parameters (and thus the bias correction term) from the same data y that is the argument of the log-likelihood (Greven and Kneib, 2010).

The approximate cAIC in LMMs

Liang et al. (2008) proposed a corrected version of the cAIC taking the estimation of θ_* into account. This measure will from now on be referred to as the *approximate* cAIC (acAIC) for reasons which will become clear in the following.

For **known error variance $\boldsymbol{\sigma^2}$**, the conditional AIC of Liang et al. (2008) has the form:

Definition 16. *Approximate cAI (acAIC) for Known Error Variance*

$$acAIC = -2\ log\left(f(y|\hat{\beta},\hat{b},\hat{\theta})\right) + 2\ \Phi_0, \tag{5.14}$$

where Φ_0 replaces the effective degrees of freedom ρ in the ccAIC of Vaida and Blanchard (2005) (5.8),

$$\Phi_0 = \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i} = tr\left\{\frac{\partial \hat{y}}{\partial y}\right\},\ i = 1,\ldots,n. \tag{5.15}$$

This is an unbiased[8] estimator for cAI as the bias correction satisfies[9]

$$\begin{aligned} BC = cAI - E_{g(y,b)}\left[-2\ log\left(f(y|\hat{\beta}(y),\hat{b}(y))\right)\right] &= \sum_{i=1}^{n} \frac{2}{\sigma^2}\ Cov_{g(y,b)}\left(\hat{\mu}_i, y_i\right) \\ &= \frac{2}{\sigma^2}\ E_{g(y,b)}\left[\sum_{i=1}^{n}(y_i-\mu_i)\hat{\mu}_i\right] \\ &= 2\ E_{g(y,b)}\left[\Phi_0(y)\right]. \end{aligned} \tag{5.16}$$

Note that for **known variance components** θ_*, Φ_0 reduces to ρ.

The bias correction for **unknown error variance $\boldsymbol{\sigma^2}$** has to be extended by a second term yielding[10]

$$\begin{aligned} BC &= cAI - E_{g(y,b)}\left[-2\ log\left(f(y|\hat{\beta}(y),\hat{b}(y),\hat{\sigma}^2(y))\right)\right] \\ &= 2E_{g(y,b)}\left[\sum_{i=1}^{n}(y_i-\hat{\mu}_i)\frac{\hat{\eta}_i}{\hat{\sigma}^2}\right] + 2E_{g(y,b)}\left[\sum_{i=1}^{n}\left\{c(y_i,\hat{\sigma}_i) - E_{g(y_i^*|b)}\left[c(y_i^*,\hat{\sigma}^2)\right]\right\}\right]. \end{aligned} \tag{5.17}$$

Note that for known σ^2 the second term cancels (Greven, 2011b).

Liang et al. (2008) extended their measure to the case of **unknown error variance $\boldsymbol{\sigma^2}$** by replacing Φ_0 by Φ_1 of the form

$$\Phi_1 = \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} tr\left\{\frac{\partial \hat{y}}{\partial y}\right\} + \tilde{\sigma}^2(\hat{y}-y)^T\frac{\partial \hat{\sigma}^{-2}}{\partial y} + \frac{1}{2}\ \tilde{\sigma}^4 tr\left\{\frac{\partial^2 \hat{\sigma}^{-2}}{\partial y \partial y^T}\right\}. \tag{5.18}$$

[8] Note that in contrast to the conventional degrees of freedom this result holds for finite samples.

[9] Liang et al. (2008)

[10] Liang et al. (2008)

The parameter $\tilde{\sigma}^2$ denotes the unknown true error variance which is replaced by an estimator $\hat{\sigma}^2$ based on maximum likelihood or restricted maximum likelihood estimation for practical use.

Liang et al. (2008) did not provide closed form expressions for the derivatives involved in the calculation of Φ_0 as well as of Φ_1. Instead, they proposed numerical approximations based on small disturbances of the observed data.
For **known error variance** they suggested approximating the first partial derivatives $\partial \hat{y}_i / \partial y_i$ $(i = 1, \ldots, n)$ numerically by

$$\left\{ \hat{y}_i(y + he_i) - \hat{y}_i(y) \right\} / h, \tag{5.19}$$

where h is a small number and e_i is the $n \times 1$ vector, with the ith component equal to 1 and all other components equal to 0.

The drawback of the use of this approximate cAIC lies in its high computational costs. The implementation of the acAIC (5.14) requires n – and using Φ_1 even $2n$ – additional model fits and thus becomes very time-consuming for even moderate sample size n (Greven and Kneib, 2010).

The analytic cAIC in LMMs

Based on the findings that the conventional cAIC of Vaida and Blanchard (5.10) is no more an asymptotically unbiased estimator for the cAI in the case of unknown θ_* and that the high computational costs involved in the numerical approximation of Liang et al. (5.14) can be prohibitive, Greven and Kneib (2010) derived an analytic representation with an efficient implementation, further referred to as the *analytic* cAIC.[11]
Due to close agreement between Φ_1 (5.18) and $\Phi_0 + 1$ (5.15) in their simulation studies, Greven and Kneib focused on an analytic representation of Φ_0 which will be the quantity of interest here as well.

The main challenge in the derivation of an analytic representation of Liang et al.'s cAIC arises from the dependence of the hat matrix $\boldsymbol{H}_1$ on y. $\boldsymbol{H}_1$ depends on y due to the estimation of the covariance matrix from the data. The calculation of Φ_0 involves the derivation of $\hat{y} = \boldsymbol{H}_1 y$ with respect to y. Therefore, in addition to the product rule, the chain rule of differentiation has to be applied in order to execute the derivation.

[11] As this measure is an analytic version of the approximate degrees of freedom of Liang et al. (2008) (5.15) it is also not based on asymptotics.

This yields

$$\begin{aligned}\frac{\partial \hat{y}}{\partial y} &= \frac{\partial \boldsymbol{H}_1(y)y}{\partial y}\\ &= \boldsymbol{H_1}(y) + \frac{\partial \boldsymbol{H}_1(y)}{\partial y} \cdot y\\ &= \boldsymbol{H_1}(y) + \frac{\partial \boldsymbol{H}_1(\hat{\theta}(y))}{\partial y} \cdot y \qquad (5.20)\\ &= \boldsymbol{H_1}(y) + \frac{\partial}{\partial \theta}\boldsymbol{H}_1(\hat{\theta}(y))\frac{\partial}{\partial y}\hat{\theta}(y) \cdot y.\end{aligned}$$

Hence, the derivative of $\boldsymbol{H}_1$ involves the derivation of the estimators of the covariance parameters with respect to y. This is nontrivial due to the lack of an analytic representation of these estimators as they are determined iteratively.
Note that in the linear model this problem does not occur because the hat matrix

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \qquad (5.21)$$

is independent of the covariance matrix of y.

Except for notational differences to adapt the notation used in this work, the following theorem is an excerpt of Greven and Kneib (2010).[12]

Theorem 1 (The analytic cAIC).
Denote the parameter space for $\theta_* = (\theta_{*,1}, \dots, \theta_{*,q})$ by $\Theta \subseteq \mathbb{R}^q$. Denote by $\hat{\theta}_*$ the maximum likelihood or restricted maximum likelihood estimator of θ_*.
For the conditional AIC in the linear mixed model with unknown θ, the bias correction term can be written as

$$\Phi_0 = \hat{\rho} + \sum_{j=1}^{s} e_j^T \hat{\boldsymbol{B}}_*^{-1}\hat{\boldsymbol{\Upsilon}}_*\hat{\boldsymbol{A}}_*\hat{\boldsymbol{W}}_{*,j}\hat{\boldsymbol{A}}_*y, \qquad (5.22)$$

where it is assumed that after potential reordering, θ_* can be written as $\theta_* = (\theta_s^T, \theta_t^T, \theta_{q-s-t}^T)^T$ for some $0 \leq s \leq q, 0 \leq t \leq q-s$, such that

$$\Theta = \left\{\theta_* | \theta_s \in \Theta_s \subseteq \mathbb{R}^s, \theta_t \in [0,\infty)^t, \theta_{q-s-t} \in F(\theta_s, \theta_t) \subset \mathbb{R}^{q-s-t}\right\},$$

$\hat{\theta}_s$ lies in the interior of Θ_s, $F(\theta_s, 0) = 0$ for all θ_s, and $(\hat{\theta}_t^T, \hat{\theta}_{q-s-t})^T = 0$.

Furthermore, e_j denotes the $s \times 1$ unit vector for component j,

$$\begin{aligned}\boldsymbol{A}_* &= \boldsymbol{V}_*^{-1} - \boldsymbol{V}_*^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_*^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}_*^{-1},\\ \boldsymbol{W}_{*,j} &= (\partial/\partial\theta_{*,j})\boldsymbol{V}_*,\\ \boldsymbol{U}_{*,jl} &= (\partial^2/\partial\theta_{*,l}\partial\theta_{*,j})\boldsymbol{V}_*,\ j,\ l = 1,\dots,s \text{ are } n \times n \text{ matrices.}\end{aligned}$$

The jth row of the $s \times n$ matrix $\boldsymbol{\Upsilon}_*$, $j = 1, \dots, s$ is

$$2(y^T\boldsymbol{A}_*y)y^T\boldsymbol{A}_*\boldsymbol{W}_{*,j}\boldsymbol{A}_* - (y^T\boldsymbol{A}_*\boldsymbol{W}_{*,j}\boldsymbol{A}_*y)y^T\boldsymbol{A}_*$$

[12] See Theorem 3 in Greven and Kneib (2010).

and B_* is the negative definite $s \times s$ Hessian matrix for θ_* with jl-th entry

$$b_{jl} - y^T \boldsymbol{A}_* \boldsymbol{W}_{*,j} \boldsymbol{A}_* y y^T \boldsymbol{A}_* \boldsymbol{W}_{*,l} \boldsymbol{A}_* y - y^T (\boldsymbol{A}_* \boldsymbol{U}_{*,jl} \boldsymbol{A}_* - 2 \boldsymbol{A}_* \boldsymbol{W}_{*,l} \boldsymbol{A}_* \boldsymbol{W}_{*,j} \boldsymbol{A}_*) y y^T \boldsymbol{A}_* y,$$

where b_{jl} is

$b_{jl} = (y^T \mathbf{A}_* y)^2 \; tr \left\{ \mathbf{U}_{*,jl} \boldsymbol{A}_* - \boldsymbol{W}_{*,j} \boldsymbol{A}_* \boldsymbol{W}_{*,l} \boldsymbol{A}_* \right\} / (n-p)$ *for REML estimation and*
$b_{jl} = (y^T \boldsymbol{A}_* y)^2 \; tr \left\{ \boldsymbol{U}_{*,jl} \boldsymbol{V}_*^{-1} - \boldsymbol{W}_{*,j} \boldsymbol{V}_*^{-1} \boldsymbol{W}_{*,l} \boldsymbol{V}_*^{-1} \right\} / n$ *for ML estimation, $j, l = 1, \ldots, s$.*

Thus, the analytic cAIC can be written as follows:

Definition 17. *Analytic cAIC ($cAIC_{analyt}$) for Known Error Variance*

$$cAIC_{analyt} = -2 \; log \left(f(y|\hat{\beta}, \hat{b}, \hat{\theta}) \right) + 2 \; \left(\hat{\rho} + \sum_{j=1}^{s} e_j^T \hat{\boldsymbol{B}}_*^{-1} \hat{\boldsymbol{\Upsilon}}_* \hat{\boldsymbol{A}}_* \hat{\boldsymbol{W}}_{*,j} \hat{\boldsymbol{A}}_* y + 1 \right). \tag{5.23}$$

It holds that $\hat{\rho} = n - tr(\hat{\boldsymbol{A}}_*)$, with ρ the effective degrees of freedom from the conventional cAIC (5.10). Thus, the second term of Φ_0, $\sum_{j=1}^{s} e_j^T \hat{\boldsymbol{B}}_*^{-1} \hat{\boldsymbol{\Upsilon}}_* \hat{\boldsymbol{A}}_* \hat{\boldsymbol{W}}_{*,j} \hat{\boldsymbol{A}}_* y$, is a correction term for the estimation of the unknown θ_* which has not been taken into account in the derivation of the conditional AIC.

For simplicity and ease of implementation, in the simulation studies in Chapter 6 we considered the case of a linear mixed model with only one unknown variance component, block-diagonal $\boldsymbol{G} = \tau^2 \boldsymbol{I}_\nu$, and thus $\boldsymbol{G}_* = \lambda \boldsymbol{I}_\nu$, with $\lambda = {}^{\tau^2}/_{\sigma^2}$.

This leads to the following simplifications in the representation of the analytical cAIC:

$$\hat{\boldsymbol{W}}_{*,j} = \hat{\boldsymbol{W}}_* = \boldsymbol{Z}\boldsymbol{Z}^T \tag{5.24}$$

$$\hat{\boldsymbol{U}}_{*,jl} = \hat{\boldsymbol{U}}_* = 0 \tag{5.25}$$

$$\hat{\boldsymbol{\Upsilon}}_* = 2(y^T \hat{\boldsymbol{A}}_* y) y^T \hat{\boldsymbol{A}}_* \hat{\boldsymbol{W}}_* \hat{\boldsymbol{A}}_* - (y^T \hat{\boldsymbol{A}}_* \hat{\boldsymbol{W}}_* \hat{\boldsymbol{A}}_* y) y^T \hat{\boldsymbol{A}}_* \text{ is a vector.} \tag{5.26}$$

Thus $\hat{\boldsymbol{B}}_*$ is a scalar rather than a matrix.

Hence, the $cAIC_{analyt}$ is reduced to

$$cAIC_{analyt} = -2 \; log \left(f(y|\hat{\beta}, \hat{b}, \hat{\theta}) \right) + 2 \; \left(\hat{\rho} + \frac{1}{\hat{\boldsymbol{B}}_*} \hat{\boldsymbol{\Upsilon}}_* \hat{\boldsymbol{A}}_* \hat{\boldsymbol{W}}_* \hat{\boldsymbol{A}}_* y + 1 \right). \tag{5.27}$$

The cAIC based on a covariance penalty in LMMs

In a slightly different context, namely in the analysis of prediction problems, Efron (2004) provided an extended definition of the degrees of freedom of the linear model to more general models. He showed in this context that the minimization of an unbiased estimator for the expected true predictive error is equivalent to the minimization of the Akaike information criterion for a rather general class of models. We will demonstrate in the following that the definition of the generalized degrees of freedom can be used to construct two versions of a conditional Akaike information criterion for both the LMM and the GLMM.

In this paragraph, first the estimation of prediction errors will be introduced, followed by a presentation of Efron's definition of generalized degrees of freedom. Then, the estimation of this quantity will be outlined and linked to the (linear) mixed model framework. In the following section (Section 5.2), the generalization beyond Gaussianity will be considered.

Analysis of Prediction Errors

Two distinctions were made in Efron's analysis of the estimation of prediction errors. First, he distinguished between the case of

1. a linear model $\hat{\mu} = \boldsymbol{H}y$ (where $\boldsymbol{H}$ is not depending on y) and
2. a more general model by dropping the linearity assumption, yielding $\hat{\mu} = m(y)$.

Second, a distinction was made between the types of error measures. Efron regarded

1. the case where the prediction error $Q(y, \hat{\mu})$ is measured by the squared error
$$Q(y, \hat{\mu}) = (y - \hat{\mu})^2 \tag{5.28}$$
and
2. a generalization beyond squared error to a wider class of error measures:
$$Q(y, \hat{\mu}) = q(\hat{\mu}) + \dot{q}(\hat{\mu})(y - \hat{\mu}) - q(y), \tag{5.29}$$
with $q(\cdot)$ denoting any concave function and $\dot{q}(\hat{\mu}) = dq/d\mu|_{\hat{\mu}}$.[13]

Consider first the case of a standard linear model without random effects.[14] Let the squared error be the error measure for the prediction error. Thus,

$$\hat{\mu} = \boldsymbol{H}y$$
$$Q(y, \hat{\mu}) = (y - \hat{\mu})^2.$$

[13] The choice of $q(\mu) = \mu(1 - \mu)$ gives rise to a squared error.

[14] No normality assumption is required at this point.

Let Err denote the unobservable, true predictive error of $\hat{\mu}$. err is the apparent error, which is proved to be an optimistic quantity, as it is based on the same data y and does not permit to answer the question of how well $\hat{\mu}$ will predict a future data set, independently generated from the same mechanism that produced y (Efron, 2004).
The choice of quadratic error measure yields

$$err_i = (y_i - \hat{\mu}_i)^2 \text{ and} \tag{5.30}$$

$$Err_i = E_0 \left[y_i^0 - \hat{\mu}_i \right]^2, \tag{5.31}$$

where the expectation E_0 denotes the expectation with respect to a new data set y^0 independently drawn from the same mechanism. Thus, when

$$y_i \sim (\mu_i, \sigma^2), \tag{5.32}$$

it is $E_0(y_i^0) = \mu_i$ and $Var_0(y_i^0) = \sigma^2$.
It should be pointed out that Err itself is an expectation (see 5.31).

Efron referred to Mallows (1973), who showed for the linear case that

$$\widehat{Err} = err + 2\sigma^2 tr \left\{ \boldsymbol{H} \right\}, \tag{5.33}$$

with

$$err = \sum_{i=1}^{n} err_i, \ Err = \sum_{i=1}^{n} Err_i,$$

is an unbiased estimator for the expectation Err.[15] Efron extended this finding by dropping the linearity assumption, i.e $\hat{\mu} = m(y)$. He showed that in order to unbiasedly estimate the true predictive error Err_i, a covariance penalty must be added to the apparent error[16]

$$E\left[Err_i\right] = E\left[err_i + 2\ Cov(y_i, \hat{\mu}_i)\right]. \tag{5.34}$$

In the linear case ($\hat{\mu} = \boldsymbol{H}y$), the degrees of freedom are commonly defined as $tr(\boldsymbol{H})$. Efron suggested to analogously extend this definition to any rule $\hat{\mu} = m(y)$, by defining the *generalized degrees of freedom* (gdf) as Ye (1998):

$$gdf = \sum_{i=1}^{n} \frac{Cov(y_i, \hat{\mu}_i)}{\sigma^2}. \tag{5.35}$$

Note that twice the quantity (5.35) **corresponds to the bias correction term**[17] **(5.16) used by Liang et al. (2008)**, with the significant difference that the covariance in (5.16) is with respect to both $\boldsymbol{y}$ **and the random effects** $\boldsymbol{b}$.

[15] In practice, σ^2 has to be replaced by an estimate $\hat{\sigma}^2$ (Efron, 2004).
[16] In the linear case, this simplifies to Mallows estimator (5.33).
[17] Assuming known error variance.

It should be pointed out that the estimator (5.34) is not practicable in general, as $Cov(y_i, \hat{\mu}_i)$ is an *unobservable* quantity. For the special case of $y \sim \mathcal{N}(\mu, \sigma^2 \boldsymbol{I}_n)$, Stein (1981) showed that the estimator can be applied and displayed in the form

$$\widehat{Err} = err + 2\sigma^2 \sum_{i=1}^{n} \partial\hat{\mu}_i / \partial y_i, \tag{5.36}$$

with $\partial\hat{\mu}_i/\partial y_i$ *observable.*

For more general situations, Efron (2004) suggested to use parametric bootstrap methods to approximate the covariance penalty

$$\begin{aligned} Cov(y_i, \hat{\mu}_i) &= E\left[(y_i - E(y_i))(\hat{\mu}_i - E(\hat{\mu}_i))\right] \\ &= E\left[y_i\hat{\mu}_i - \hat{\mu}_i y_i - y_i E(\hat{\mu}_i) + \mu_i E(\hat{\mu}_i)\right] \\ &= E\left[(y_i - \mu_i)\hat{\mu}_i\right]. \end{aligned} \tag{5.37}$$

Here, a density $\hat{f}$ is assumed for the data y and a large number B of simulated observations (*bootstrap replications*) from $\hat{f}$ are generated

$$\hat{f} \to y^*,$$

followed by the estimation of the parameters as

$$\hat{\mu}^* = m(y^*).$$

Finally, the covariance is estimated from the observed bootstrap covariance[18]

$$\widehat{Cov}_i = \widehat{Cov}(y_i, \hat{\mu}_i) = \frac{1}{B-1} \sum_{\xi=1}^{B} \hat{\mu}_i^{*\xi}(y_i^{*\xi} - y_i^{*\cdot}), \tag{5.38}$$

with

$$y_i^{*\cdot} = \frac{1}{B} \sum_{\xi=1}^{B} y_i^{*\xi}.$$

It should be noted that although Efron argued that the generalized degrees of freedom apply for a general rule $\hat{\mu} = m(y)$, one has to be cautious with the transfer to mixed models, as mixed models contain random effects and variance parameters have to be estimated as well. However, Efron (2004) showed that the covariance penalty (5.34) can be generalized beyond squared error which simplifies the application to mixed models. This will be the focus in the following.

So far, a quadratic error measure for the prediction error was considered. In a next step, Efron (2004) extended his findings to a wider class of error measures, namely the q-class of error measures, with $Q(y, \hat{\mu})$ as in (5.29).

[18] Whereby the subtraction of 1 in $(B-1)$ accounts for the fact that the mean has been estimated.

Let

$$O_i = O_i(f, y) = Err_i - err_i \tag{5.39}$$

denote the *optimism* and its expectation with respect to f the *expected optimism*

$$\Omega_i = \Omega(f) = E_f\left[O_i(f, y)\right]. \tag{5.40}$$

Finally, let

$$\hat{\lambda}_i = \dot{q}(\hat{\mu}_i)/2. \tag{5.41}$$

Efron (2004) formulated the extension of the covariance penalty theory beyond squared error in the following theorem.

Theorem 2 (Optimism Theorem).
For the error measure $Q(y, \hat{\mu})$ it holds that

$$E\left\{Err_i\right\} = E\left\{err_i + \Omega_i\right\}, \tag{5.42}$$

where

$$\Omega_i = 2\ Cov(\hat{\lambda}_i, y_i). \tag{5.43}$$

the expectations and covariance being with respect to f.

For the proof see Appendix A.

Efron (2004) remarked that his optimism theorem applies to any probability mechanism and that even independence among components of y is not required which benefits the application to mixed models.

For the special case where $Q(y, \hat{\mu})$ is the deviance function of an exponential family

$$D(y|\hat{\mu}) = -2\phi\ (log\left\{\mathcal{L}(\hat{\mu}|y)\right\} - log\left\{\mathcal{L}(y|y)\right\}), \tag{5.44}$$

$\hat{\lambda}$ is the corresponding estimated natural parameter $\hat{\vartheta}$ in (3.48) (see Efron (2004)). For Gaussianity and $Q(y, \hat{\mu}) = D(y|\hat{\mu})$[19] with the canonical link function $g(\cdot) = h(\cdot)$, the parameter $\hat{\lambda}$ equals the estimated mean $\hat{\mu}$ and the correction (5.42) is equal to (5.34).[20] Other distributions of the one-parametric exponential family will be discussed in Section 5.2.

[19] In the case of Gaussianity the deviance corresponds to the squared error.

[20] Note that for the standard linear model with normally distributed error terms and the usage of the squared error as a measure for the prediction error, the covariance penalty $Cov(\hat{\lambda}_i, y_i)$ simplifies to the degrees of freedom $tr(\boldsymbol{H})$.

For practical use, parametric bootstrap can be again employed to approximate the penalty $\Omega_i = 2\ Cov(\hat{\lambda}_i, y_i)$ as in the case of the squared error measure. The covariance $Cov_i = Cov(\hat{\lambda}_i, y_i)$ is then estimated from the generated data $y_i^{*1}, \ldots, y_i^{*B}$ $(i = 1, \ldots, n)$ as[21]

$$\widehat{Cov}_i = \widehat{Cov}(\hat{\lambda}_i, y_i) = \frac{1}{B-1} \sum_{\xi=1}^{B} \hat{\lambda}_i^{*\xi} (y_i^{*\xi} - y_i^{*\cdot}), \tag{5.45}$$

with

$$y_i^{*\cdot} = \frac{1}{B} \sum_{\xi=1}^{B} y_i^{*\xi}$$

and B the number of bootstrap replications.

Application to Mixed Models

Consider now the linear mixed model to which these findings will be applied.

Assuming **known error variance $\boldsymbol{\sigma^2}$**, the covariance based conditional Akaike information criterion can be defined as

Definition 18. *cAIC Based on a Covariance Penalty (cAIC$_{Cov}$) for Known Error Variance*

$$\begin{aligned} cAIC_{Cov} &= -2\ log\left(f(y|\hat{\beta}, \hat{b}, \hat{\theta})\right) + 2 \sum_{i=1}^{n} Cov(y_i, \frac{\hat{\mu}_i}{\sigma^2}) \\ &= -2\ log\left(f(y|\hat{\beta}, \hat{b}, \hat{\theta})\right) + \frac{2}{\sigma^2} \sum_{i=1}^{n} Cov(y_i, \hat{\mu}_i), \end{aligned} \tag{5.46}$$

with $log\left(f(y|\hat{\beta}, \hat{b}, \hat{\theta})\right)$ *denoting the maximized conditional log-likelihood.*

Note that this definition changes when the error variance is unknown since in the bias correction (8.1), the error variance can no longer be pulled out of the expectation of the first term of the BC[22]

$$E_{g(y,b)} \left[2\ log\left(f(y|\hat{\beta}(y), \hat{b}(y), \hat{\sigma}^2(y))\right)\right]. \tag{5.47}$$

[21] The estimation of the mean is again taken into account through dividing by $(B-1)$.

[22] Another adjustment concerns the second term of the BC, for more information see Chapter 8.

The definition therefore has to be adjusted to

Definition 19. *cAIC Based on a Covariance Penalty ($cAIC_{Cov}$) for Unknown Error Variance*

$$cAIC_{Cov} = -2\ log\left(f(y|\hat{\beta}, \hat{b}, \hat{\theta})\right) + 2\sum_{i=1}^{n} Cov(y_i, \frac{\hat{\mu}_i}{\hat{\sigma}^2},). \tag{5.48}$$

Practical Use for Linear Mixed Models

We now demonstrate that due to the presence of random effects in LMMs, the generation of bootstrap replications $y_i^{*\xi}$ ($i = 1, \ldots, n$, $\xi = 1, \ldots, B$) can be performed in two different ways.

1. Either the random effects are kept constant (they are fixed at the estimated quantities) and replications are drawn as

$$y^{*\xi} = \boldsymbol{X}\hat{\beta} + \boldsymbol{Z}\hat{b} + \varepsilon^{*\xi},\ \xi = 1, \ldots, B, \tag{5.49}$$

   where $\hat{\beta}$ and $\hat{b}$ denote the BLUP for the mixed model $y = \boldsymbol{X}\beta + \boldsymbol{Z}b + \varepsilon$,

2. or the random effects are also drawn from a distribution and the data is generated as

$$y^{*\xi} = \boldsymbol{X}\hat{\beta} + \boldsymbol{Z}b^{*\xi} + \varepsilon^{*\xi},\ \xi = 1, \ldots, B. \tag{5.50}$$

The first method will be referred to as the **conditional** version of the covariance based penalty and the second will be named the **joint** version as both – the random error term and the random effects – are individually drawn for each bootstrap sample.
The detailed algorithms for the estimation of the covariance penalties can be found in Appendix A.

The distinction between known and unknown error variance is translated by either using a constant variance, i.e. fixing σ^2 to the estimated quantity $\hat{\sigma}^2$ (when assuming known variance) or applying re-estimated variances in each bootstrap sample, $(\hat{\sigma}^2)^{*1}, \ldots, (\hat{\sigma}^2)^{*B}$ (when assuming unknown variance).

A closer look at Efron's estimation of the covariance discloses the need for modifications for the joint version. Recall that the quantity of interest equals $E\left[(y_i - \mu_i)\hat{\mu}_i\right]$ (see (5.37)). For the linear mixed model, it is $\mu_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_ib_i$ ($i = 1, \ldots, n$).

Efron's suggestion to approximate $(y_i - \mu_i)$ by the difference

$$(y_i^{*\xi} - y_i^{*\cdot}),\ i = 1, \dots, n,\ \xi = 1, \dots, B$$

seems to be adequate in the conditional case (5.49) as for a large number of replications

$$\begin{aligned}
y_i^{*\cdot} &= \frac{1}{B} \sum_{\xi_1}^{B} y_i^{*\xi} && (5.51)\\
&= \frac{1}{B} \sum_{\xi=1}^{B} \boldsymbol{X}_i \hat{\beta} + \boldsymbol{Z}_i \hat{b}_i + \varepsilon_i^{*\xi} && (5.52)\\
&= \boldsymbol{X}_i \hat{\beta} + \boldsymbol{Z}_i \hat{b}_i + \underbrace{\frac{1}{B} \sum_{\xi=1}^{B} \varepsilon_i^{*\xi}}_{\xrightarrow{B \to \infty} 0} && (5.53)
\end{aligned}$$

averages to the ith component of $\boldsymbol{X}\hat{\beta} + \boldsymbol{Z}\hat{b}$. However, this does not apply to the joint case. Here, $y_i^{*\cdot}$ is an estimator for $\boldsymbol{X}\beta$ and not for $\boldsymbol{X}\beta + \boldsymbol{Z}b$ as

$$\begin{aligned}
y_i^{*\cdot} &= \frac{1}{B} \sum_{\xi_1}^{B} y_i^{*\xi} && (5.54)\\
&= \frac{1}{B} \sum_{\xi=1}^{B} \boldsymbol{X}_i \hat{\beta} + \boldsymbol{Z}_i b_i^{*\xi} + \varepsilon_i^{*\xi} && (5.55)\\
&= \boldsymbol{X}_i \hat{\beta} \underbrace{\frac{1}{B} \sum_{\xi=1}^{B} \boldsymbol{Z}_i b_i^{*\xi} + \varepsilon_i^{*\xi}}_{\xrightarrow{B \to \infty} 0}. && (5.56)
\end{aligned}$$

Greven (2011b) proposed to replace $y_i^{*\cdot}$ with the ith component of $\boldsymbol{X}\hat{\beta} + \boldsymbol{Z}b^{*\xi}$ and thus to directly use $\varepsilon_i^{*\xi}$ to approximate $(y_i - \mu_i)$ yielding the formula[23]

$$\widehat{Cov}_i = \widehat{Cov}(y_i, \frac{\hat{\mu}_i}{\hat{\sigma}^2}) = \frac{1}{B} \frac{1}{\hat{\sigma}^2} \sum_{\xi=1}^{B} \hat{\mu}_i^{*\xi} \varepsilon_i^{*\xi},\ i = 1, \dots, n, \tag{5.57}$$

for known error variance and

$$\widehat{Cov}_i = \widehat{Cov}(y_i, \frac{\hat{\mu}_i}{\hat{\sigma}^2}) = \frac{1}{B} \sum_{\xi=1}^{B} \hat{\mu}_i^{*\xi} \frac{\varepsilon_i^{*\xi}}{(\hat{\sigma}^2)^{*\xi}},\ i = 1, \dots, n, \tag{5.58}$$

for unknown error variance.
For a detailed description of the proceeding of the bootstrap estimation for the special case of linear mixed models see Appendix B.

[23] Here, one does not have to account for an estimated mean and thus divides by B instead of $B - 1$.

The cAIC of Yu and Yau for LMMs

Yu and Yau (2011) recently proposed an asymptotically unbiased estimator of the conditional Akaike information for generalized linear mixed models which takes the estimation uncertainty of the variance parameters into account.[24] In this section, their suggestion will be considered by means of the special case of Gaussianity. The generalization follows in Section 5.2.
For simplicity, the case of one unknown variance component, i.e. $\boldsymbol{G} = \tau^2 \boldsymbol{I}_\nu$, will be considered in the following and in the simulation studies in Chapter 6. Moreover, the **error variance $\boldsymbol{\sigma^2}$** is assumed to be **known**.

Let h denote the sum of the conditional log-likelihood and the logarithm of the probability density function (pdf) of the random effects b

$$h = log\left\{\mathcal{L}(y|\beta, b)\right\} + log\left(f(b|\tau^2)\right). \tag{5.59}$$

Further, $\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}}$ designates the negative second derivative of h with respect to
$\tilde{\theta} = (\beta^T, b^T)^T$

$$\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}} = -\frac{\partial^2}{\partial\tilde{\theta}\partial\tilde{\theta}^T} h(y|\beta, b) = \frac{1}{\sigma^2}\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{Z} + \frac{1}{\lambda}\boldsymbol{I}_\nu \end{pmatrix} = \begin{pmatrix} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{21} & \boldsymbol{H}_{22} \end{pmatrix}, \tag{5.60}$$

with $\lambda = {}^{\tau^2}\!/_{\sigma^2}$.

Note that this matrix corresponds to σ^{-2} times the first part of the hat matrix used for the calculation of the conventional cAIC (see 5.12).

$\boldsymbol{H}_{\tilde{\theta}\tau^2}$ and $\boldsymbol{H}_{\tau^2\tilde{\theta}}$ are analogously the negative second derivatives of h with respect to $\tilde{\theta}$ and τ^2. In the considered special case they are given as

$$\boldsymbol{H}_{\tilde{\theta}\tau^2} = -\frac{\partial^2 h}{\partial\tilde{\theta}\partial\tau^2} = -\frac{1}{\tau^4}(0|b^T) \tag{5.61}$$

$$\boldsymbol{H}_{\tau^2\tilde{\theta}} = -\frac{\partial^2 h}{\partial\tau^2\partial\tilde{\theta}^T} = \boldsymbol{H}^T_{\tilde{\theta}\tau^2}. \tag{5.62}$$

Let $\boldsymbol{H}^*$ be the negative second derivative of the conditional log-likelihood of the data given the random effects, $log\left\{\mathcal{L}(y|\beta, b)\right\}$, with respect to $\tilde{\theta}$

$$\boldsymbol{H}^* = -\frac{\partial^2 \log\left\{\mathcal{L}(y|\tilde{\theta})\right\}}{\partial\tilde{\theta}\partial\tilde{\theta}^T} = \frac{1}{\sigma^2}\begin{pmatrix} \boldsymbol{X}^T X & \boldsymbol{X}^T\boldsymbol{Z} \\ \boldsymbol{Z}^T X & \boldsymbol{Z}^T\boldsymbol{Z} \end{pmatrix}. \tag{5.63}$$

This matrix corresponds to σ^{-2} times the second part of the hat matrix of the conventional cAIC (5.12).

[24] Note that in contrast to the approximate and the analytic measures, here the unbiasedness is asymptotically.

And finally, denote

$$\boldsymbol{H}_{\tau^2\tau^2} = -\frac{\partial^2 h_a}{\partial\tau^2 \partial\tau^2},$$

with

$$h_a = -\frac{1}{2}\, log\left\{det\left(\boldsymbol{H}_{22}\right)\right\} + log\left\{\mathcal{L}(y|\beta, b)\right\} + log\left(f(b|\tau^2)\right), \tag{5.64}$$

with $det(\cdot)$ denoting the determinant. For h_a we derived the specific form here as

$$h_a \propto -\frac{1}{2}\left[log\left\{det\left(\frac{1}{\sigma^2}\boldsymbol{Z}^T\boldsymbol{Z} + \frac{1}{\tau^2}\boldsymbol{I}_\nu\right)\right\} + \frac{1}{\sigma^2}(y-\eta)^T(y-\eta) + \nu\ log\left\{(\tau^2)\right\} + \frac{1}{\tau^2}b^T b\right]. \tag{5.65}$$

For $\boldsymbol{H}_{\tau^2\tau^2}$ we obtained

$$\begin{aligned}\boldsymbol{H}_{\tau^2\tau^2} &= \frac{1}{2}\ tr\left\{-\frac{\sigma^4}{\tau^8}(\boldsymbol{Z}^T\boldsymbol{Z} + \frac{\sigma^2}{\tau^2}\boldsymbol{I}_\nu)^{-2} + 2\frac{\sigma^2}{\tau^6}(\boldsymbol{Z}^T\boldsymbol{Z} + \frac{\sigma^2}{\tau^2}\boldsymbol{I}_\nu)^{-1}\right\} - \frac{\nu}{2\tau^4} + \frac{1}{\tau^6}b^T b \qquad (5.66)\\ &= \frac{1}{\tau^6}b^T b - \frac{1}{2\sigma^4}tr\left\{\left[(\boldsymbol{I}_\nu + \frac{\tau^2}{\sigma^2}\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{Z}\right]^2\right\}.\end{aligned}$$

For a detailed derivation see Appendix A.

Yu and Yau (2011) derived an asymptotic unbiased estimator of cAI for unknown variance parameter τ^2 as:

Definition 20. *cAIC of Yu and Yau ($cAIC_{YuYau}$)*

$$cAIC_{YuYau} = -2\ log\left(f(y|\hat{\beta}, \hat{b}, \hat{\tilde{\theta}})\right) + 2\ \hat{\rho}_{ml}, \tag{5.67}$$

with

$$\hat{\rho}_{ml} = tr\left\{(\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}} - \boldsymbol{H}_{\tilde{\theta}\tau^2}\boldsymbol{H}^{-1}_{\tau^2\tau^2}\boldsymbol{H}_{\tau^2\tilde{\theta}})^{-1}\boldsymbol{H}^*\right\}|_{\hat{\tilde{\theta}},\hat{b}}\ . \tag{5.68}$$

Note that the index ml of $\hat{\rho}$ is used in analogy to the notation of Yu and Yau (2011), pointing out that the estimator is constructed under ML estimation. For the proof and further details as well as the generalization to more than one random effect see Yu and Yau (2011).

By applying the Woodbury formula, the penalty term $\hat{\rho}_{ml}$ in (5.68) can be expressed dependent on the conventional cAIC of Vaida and Blanchard (2005) (5.10) (noted here as $\hat{\rho}$), yielding[25]

$$\hat{\rho}_{ml} = \hat{\rho} + \frac{\boldsymbol{H}_{\tau^2\tilde{\theta}}\boldsymbol{H}^{-1}_{\tilde{\theta}\tilde{\theta}}\boldsymbol{H}^*\boldsymbol{H}^{-1}_{\tilde{\theta}\tilde{\theta}}\boldsymbol{H}_{\tilde{\theta}\tau^2}}{\boldsymbol{H}_{\tau^2\tau^2} - \boldsymbol{H}_{\tau^2\tilde{\theta}}\boldsymbol{H}^{-1}_{\tilde{\theta}\tilde{\theta}}\boldsymbol{H}_{\tilde{\theta}\tau^2}}|_{\hat{\tilde{\theta}},\hat{\tau}^2}. \tag{5.69}$$

[25] Instead of $\tilde{\theta}$ one can also write b as β does not appear.

Note that both the numerator and the denominator of (5.69) are scalars. For the proof of the transformation, see Appendix A.

For the case of known random effects variance parameter, i.e. τ^2 known, Yu and Yau showed that their measure simplifies to Vaida and Blanchard's conventional cAIC (5.10).

By inserting the expressions from above for the matrices $\boldsymbol{H}_{\tau^2\tau^2}$ (5.66), $\boldsymbol{H}_{\tau^2\tilde{\theta}}$ (5.62), $\boldsymbol{H}_{\tilde{\theta}\tau^2}$ (5.61), $\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}}$ (5.60), and $\boldsymbol{H}^*$ (5.63) into the formula (5.69), one obtains

$$\hat{\rho}_{ml} = \hat{\rho} + \frac{\frac{1}{\tau^2\sigma^4} y^T \boldsymbol{A}_* \boldsymbol{Z} \left\{ A_1^{-1} - A_1^{-2} \right\} \boldsymbol{Z}^T \boldsymbol{A}_* y}{\frac{1}{2} tr \left\{ -\frac{\sigma^4}{\tau^8} A_2^{-2} + 2\frac{\sigma^2}{\tau^6} A_2^{-1} \right\} - \frac{\nu}{2\tau^4} + \frac{1}{\tau^2\sigma^4} y^T \boldsymbol{A}_* \boldsymbol{Z}\boldsymbol{Z}^T \boldsymbol{A}_* y - \frac{1}{\tau^4\sigma^2} y^T \boldsymbol{A}_* \boldsymbol{Z} \frac{\tau^2}{\sigma^2} A_1^{-1} \boldsymbol{Z}^T \boldsymbol{A}_* y}, \tag{5.70}$$

where

$$\boldsymbol{P}_0 = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T, \tag{5.71}$$

$$A_1 = \frac{\tau^2}{\sigma^2} \boldsymbol{Z}^T \boldsymbol{P}_0 \boldsymbol{Z} + \boldsymbol{I}_\nu, \tag{5.72}$$

$$A_2 = \boldsymbol{Z}^T \boldsymbol{Z} + \frac{\sigma^2}{\tau^2} \boldsymbol{I}_\nu. \tag{5.73}$$

Note that in formula (5.70), the random effect variance τ^2 – which can possibly be equal to zero[26] – appears in the denominator. Therefore, Greven (2011a) derived another formulation of the penalty term of Yu and Yau which seems to be more adequate, especially for implementation. This formula is not longer expressed depending on the conventional penalty term, but is based on equation (5.68). It is given by

$$\hat{\rho}_{ml} = tr \left\{ \begin{pmatrix} A_3^{-1} & -\tau^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z}A_4^{-1} \\ -\tau^2(\boldsymbol{U} + \tau^2\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}A_3^{-1} & \tau^2 A_4^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{X}^T X & \boldsymbol{X}^T\boldsymbol{Z} \\ \boldsymbol{Z}^T X & \boldsymbol{Z}^T\boldsymbol{Z} \end{pmatrix} \right\} \tag{5.74}$$

where P_0 again denotes $\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$,

$$\boldsymbol{U} = \sigma^2 \boldsymbol{I}_\nu - \frac{\sigma^2 \boldsymbol{Z}^T \boldsymbol{A}_* y y^T \boldsymbol{A}_* \boldsymbol{Z}}{y^T \boldsymbol{A}_* \boldsymbol{Z}\boldsymbol{Z}^T \boldsymbol{A}_* y - \frac{\tau^2}{2} tr \left\{ \left[(\boldsymbol{I}_\nu + \frac{\tau^2}{\sigma^2} \boldsymbol{Z}^T\boldsymbol{Z})^{-1} \boldsymbol{Z}^T \boldsymbol{Z} \right]^2 \right\}}, \tag{5.75}$$

$$\boldsymbol{T} = \boldsymbol{X}^T\boldsymbol{Z}(\tau^2\boldsymbol{Z}^T\boldsymbol{Z} + \boldsymbol{U})^{-1}\boldsymbol{Z}^T\boldsymbol{X}, \tag{5.76}$$

$$A_3 = \boldsymbol{X}^T\boldsymbol{X} - \tau^2\boldsymbol{T} \tag{5.77}$$

$$A_4 = (\tau^2\boldsymbol{Z}^T\boldsymbol{P}_0\boldsymbol{Z} + \boldsymbol{U}). \tag{5.78}$$

The derivation of this expression can be found in Appendix A.

[26] This is in fact the most interesting case.

Moreover, we derived a formulation of $\hat{\rho}_{ml}$ in which the random effects variance τ^2 only appears in the numerator based on representation (5.69). It is introduced here as it plays a role in the simulation studies in Chapter 6. It is given by

$$\hat{\rho}_{ml} = \hat{\rho} + \frac{\frac{\tau^2}{\sigma^4} y^T \boldsymbol{A}_* \boldsymbol{Z} \left\{ A_1^{-1} - A_1^{-2} \right\} \boldsymbol{Z}^T \boldsymbol{A}_* y}{\frac{1}{2}\sigma^2 tr \left\{ -\sigma^2 A_4^{-2} + 2A_4^{-1} \right\} - \frac{\nu}{2} + \frac{\tau^2}{\sigma^4} y^T \boldsymbol{A}_* \boldsymbol{Z} (\boldsymbol{I} - A_1^{-1}) \boldsymbol{Z}^T \boldsymbol{A}_* y}, \tag{5.79}$$

where again $\hat{\rho}$ denotes the conventional penalty of Vaida and Blanchard (2005), A_1 is as in (5.70) and $A_4 = \tau^2 A_2$ from (5.70).

5.2 The AIC in Generalized Linear Mixed Models

For the generalization beyond Gaussianity, the search for an appropriate Akaike information criterion poses additional challenges. This is due to the fact that – as seen in Section 3.2.4 – the marginal distribution of the generalized linear mixed model is not analytically accessible. For this reason and because it has already been shown that in the simplest special case (the case of normal distribution) the conditional AIC is more adequate than its marginal counterpart, only estimators for the conditional Akaike information will be considered in the following section. In this context, two measures will be looked at: the cAIC based on a generalized covariance penalty of Efron (2004) and the extension of the cAIC of Yu and Yau (2011) beyond Gaussianity.

The cAIC based on a covariance penalty in GLMMs

As described in the previous section, Efron (2004) developed a covariance penalty ($cAIC_{Cov}$) which is not restricted to the Gaussian distribution but applies to any probability mechanism. For members of the exponential family, he showed that using the deviance function (5.44) as a measure for the prediction error, the penalty term can be written as

$$2\sum_{i=1}^{n} Cov(\frac{\hat{\vartheta}_i}{\phi}, y_i) \tag{5.80}$$

for a **known dispersion parameter $\boldsymbol{\phi}$** and with $\hat{\phi}$ replacing ϕ in the case of **unknown dispersion**. Analogously to LMMs, this yields the conditional Akaike information criterion. Assuming that the **dispersion parameter is known**, the covariance based cAIC is defined as:

Definition 21. *cAIC Based on a Covariance Penalty ($cAIC_{Cov}$) for GLMMs for Known Dispersion Parameter*

$$\begin{aligned} cAIC_{Cov} &= -2\ log\left(f(y|\hat{\beta}, \hat{b}, \hat{\theta})\right) + 2\sum_{i=1}^{n} Cov(\frac{\hat{\vartheta}_i}{\phi}, y_i) && (5.81)\\ &= -2\ log\left(f(y|\hat{\beta}, \hat{b}, \hat{\theta})\right) + 2\ \frac{1}{\phi}\ \sum_{i=1}^{n} Cov(\hat{\vartheta}_i, y_i), && (5.82)\end{aligned}$$

with log $\left(f(y|\hat{\beta}, \hat{b}, \hat{\theta})\right)$ denoting the maximized conditional log-likelihood.

When the **dispersion parameter is unknown**[27] the cAIC is given by:

Definition 22. *cAIC Based on a Covariance Penalty for GLMMs for Unknown Dispersion Parameter*

$$cAIC_{Cov} = -2\ log\left(f(y|\hat{\beta}, \hat{b}, \hat{\theta})\right) + 2\ \sum_{i=1}^{n} Cov(\frac{\hat{\vartheta}_i}{\hat{\phi}}, y_i). \tag{5.83}$$

For canonical link functions, ϑ corresponds to $\eta = g(\mu)$.

To give an example, consider the Bernoulli distribution $y_i \sim Bin(1, \pi)$ with the canonical link function, i.e. logit link. The corresponding deviance has the form[28]

$$Q(y, \hat{\mu}) = \begin{cases} -2\ log(\mu), & \text{if } y = 1, \\ -2\ log(1-\mu), & \text{if } y = 0. \end{cases} \tag{5.84}$$

The estimated natural parameter $\hat{\lambda} = \hat{\eta} = g(\hat{\mu})$ is given by

$$\hat{\lambda} = log\left\{\frac{\hat{\mu}}{1-\hat{\mu}}\right\}, \tag{5.85}$$

and the dispersion parameter is equal to 1.

The main differences to the Gaussian case lie first in the replacement of the error variance by the dispersion parameter, and obviously second in the estimation of the models in each bootstrap replication, as for the generalized case no analytic formulations are available which complicates the proceeding.

Consider in the following a canonical link function. Let $\hat{\eta}$ denote the predictor in the joint case and $\hat{\eta}_{fixed}$ the one for the conditional version, i.e. for normally distributed errors one has

$$\hat{\eta}^{*\xi} = \boldsymbol{X}\hat{\beta} + \boldsymbol{Z}b^{*\xi},\ \text{for } \xi = 1, \dots, B, \tag{5.86}$$

$$\hat{\eta}_{fixed} = \boldsymbol{X}\hat{\beta} + \boldsymbol{Z}\hat{b}. \tag{5.87}$$

Instead of drawing new data from a normal distribution as described in (5.49) and (5.50), the generation of data has to be adjusted in the generalized case, e.g. observations in the binary case are drawn as

$$y_i \sim Bin(1, \pi)$$

with

$$\pi = \frac{exp(\hat{\eta}_i)}{1 + exp(\hat{\eta}_i)} \text{ and} \tag{5.88}$$

$$\pi = \frac{exp(\hat{\eta}_{fixed,i})}{1 + exp(\hat{\eta}_{fixed,i})}, \text{ respectively.} \tag{5.89}$$

[27] Note that for most distributions in the exponential family the dispersion parameter is a constant.

[28] Efron (2004)

In the case of a Poisson distribution, observations are drawn as

$$y_i \sim Pois(\lambda)$$

with

$$\lambda = exp(\hat{\eta}_i) \text{ and} \tag{5.90}$$
$$\lambda = exp(\hat{\eta}_{fixed,i}), \text{ respectively.} \tag{5.91}$$

It should be noted that, although the case of exponential family and canonical link function is discussed here as it represents an important special case and is the situation which has been considered for the other cAICs as well, Efron's covariance penalty is not restricted to these assumptions.[29]

As in the Gaussian case, we advise modifications for the joint version and the consideration concerning the estimation of ϕ (either global or in every bootstrap replication) stays important – unless ϕ is a constant.

The cAIC of Yu and Yau in GLMMs

As already mentioned in the previous section, Yu and Yau (2011) derived their asymptotically unbiased estimator of the cAI for the case of GLMMs, strictly speaking for GLMMs with the canonical link function and restricted to ML estimation.

As the special case of normal distribution has already been discussed in Section 5.1.2, the generalization beyond Gaussianity will now be considered.

Let us again assume the **error variance $\boldsymbol{\sigma^2}$ to be known** and consider as before the case of one unknown variance component, i.e. $\boldsymbol{G} = \tau^2 \boldsymbol{I}_\nu$.

In analogy to the normal case, the function h denotes the sum of the log-likelihood and the logarithm of the pdf of the random effects vector b (compare (5.59)). Note that the second part of h stays the same as in equation (5.59), whereas the log-likelihood clearly has to be adjusted to the distribution of the response variable. As the canonical link ($\vartheta = \eta$) is considered, it holds that

$$log\left(f(y|\hat{\beta}, \hat{b}, \hat{\theta})\right) \propto \frac{1}{\phi}\sum_{i=1}^{n}\{y_i\vartheta_i - b(\vartheta_i)\} \tag{5.92}$$
$$= \frac{1}{\phi}\sum_{i=1}^{n}\{y_i\eta_i - b(\eta_i)\}\,. \tag{5.93}$$

As in the Gaussian case, $\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}}$ denotes the negative second derivative of h with respect to $\tilde{\theta} = (\beta^T, b^T)^T$, yielding

$$\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}} = -\frac{\partial^2}{\partial\tilde{\theta}\partial\tilde{\theta}^T} h(y|\beta, b) = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{B}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{B}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{B}\boldsymbol{Z} + \frac{1}{\tau^2}\boldsymbol{I}_\nu \end{pmatrix} = \begin{pmatrix} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{21} & \boldsymbol{H}_{22} \end{pmatrix},$$

[29] In contrast to the generalized cAIC of Yu and Yau (2011) which is restricted to members of the exponential family and the use of the canonical link function.

with the matrix $\boldsymbol{B}$ being the negative second derivative of the log-likelihood of the response with respect to the linear predictor η

$$\boldsymbol{B} = -\frac{\partial^2}{\partial\eta\partial\eta^T} log\left(f(y|\hat{\beta},\hat{b},\hat{\theta})\right). \tag{5.94}$$

We derived the specific form of $\boldsymbol{B}$ here as

$$\boldsymbol{B} = \frac{1}{\phi}\, b''(\eta_k)\delta_{kl}, \tag{5.95}$$

with $b''(\cdot)$ being the second derivative of $b(\cdot)$ and δ_{kl} denoting the Kronecker delta, i.e.

$$\delta_{kl} = \begin{cases} 1, & k = l \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the resulting matrix has the form

$$\boldsymbol{B} = \frac{1}{\phi}\begin{pmatrix} b''(\eta_1) & & 0 \\ & \ddots & \\ 0 & & b''(\eta_n) \end{pmatrix}. \tag{5.96}$$

In the case of binary data (Bernoulli distribution) and logit link, $\boldsymbol{B}$ becomes

$$\boldsymbol{B} = \begin{pmatrix} \mu_1(1-\mu_1) & & 0 \\ & \ddots & \\ 0 & & \mu_n(1-\mu_n) \end{pmatrix} = \begin{pmatrix} \frac{exp(\eta_1)}{(1+exp(\eta_1))^2} & & 0 \\ & \ddots & \\ 0 & & \frac{exp(\eta_n)}{(1+exp(\eta_n))^2} \end{pmatrix} \tag{5.97}$$

as the dispersion parameter ϕ is equal to one.

For a Poisson distribution one obtains (again $\phi = 1$)

$$\boldsymbol{B} = \begin{pmatrix} \mu_1 & & 0 \\ & \ddots & \\ 0 & & \mu_n \end{pmatrix} = \begin{pmatrix} exp(\eta_1) & & 0 \\ & \ddots & \\ 0 & & exp(\eta_n) \end{pmatrix}. \tag{5.98}$$

The matrices $\boldsymbol{H}_{\tilde{\theta},\tau^2}$ and $\boldsymbol{H}_{\tau^2,\tilde{\theta}}$ stay the same as in the Gaussian case (5.61) and the negative second derivative of the conditional log-likelihood of the response with respect to $\tilde{\theta}$ is extended by $\boldsymbol{B}$ to

$$\boldsymbol{H}^* = -\frac{\partial^2 \log\left\{\mathcal{L}(y|\tilde{\theta})\right\}}{\partial\tilde{\theta}\partial\tilde{\theta}^T} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{B}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{B}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{B}\boldsymbol{Z} \end{pmatrix}. \tag{5.99}$$

The extension by the matrix $\boldsymbol{B}$ also applies to the matrix $\boldsymbol{H}_{\tau^2\tau^2}$, which is given by

$$\begin{aligned} \boldsymbol{H}_{\tau^2\tau^2} &= -\frac{\partial^2 h_a}{\partial\tau^2\partial\tau^2} \\ &= \frac{\partial^2}{\partial\tau^2\partial\tau^2}\left\{\frac{1}{2}\, log\left\{det\left(\boldsymbol{Z}^T\boldsymbol{B}\boldsymbol{Z} + \frac{1}{\tau^2}\boldsymbol{I}_\nu\right)\right\}\right\} - \frac{\nu}{2\tau^4} + \frac{1}{\tau^6}b^Tb \\ &= \frac{1}{2}tr\left\{-\frac{1}{\tau^8}(\boldsymbol{Z}^T\boldsymbol{B}\boldsymbol{Z} + \frac{1}{\tau^2}\boldsymbol{I}_\nu)^{-2} + \frac{2}{\tau^6}(\boldsymbol{Z}^T\boldsymbol{B}\boldsymbol{Z} + \frac{1}{\tau^2}\boldsymbol{I}_\nu)^{-1}\right\} - \frac{\nu}{2\tau^4} + \frac{1}{\tau^6}b^Tb, \end{aligned} \tag{5.100}$$

with

$$\begin{aligned} h_a &= -\frac{1}{2}\, log\left\{det\left(\boldsymbol{H}_{22}\right)\right\} + log\left\{\mathcal{L}(y|\beta, b)\right\} + log\left(f(b|\tau^2)\right) \\ &\propto -\frac{1}{2}\, log\left\{det\left(\boldsymbol{Z}^T\boldsymbol{B}\boldsymbol{Z} + \frac{1}{\tau^2}\boldsymbol{I}_\nu\right)\right\} + \frac{1}{\phi}\sum_{i=1}^n\left\{y_i\eta_i - b(\eta_i)\right\} - \frac{\nu}{2}\, log\left(\tau^2\right) - \frac{1}{2\tau^2}b^Tb. \end{aligned} \tag{5.101}$$

Altogether, this yields the following definition of an asymptotically unbiased estimator for the cAI by Yu and Yau (2011).

Definition 23. *cAIC of Yu and Yau for GLMMs ($cAIC_{YuYau}$) for Known Dispersion Parameter*

$$cAIC_{YuYau} = -2\, log\left(f(y|\hat{\beta}, \hat{b}, \hat{\tilde{\theta}})\right) + 2\, \hat{\rho}_{ml}, \tag{5.102}$$

with

$$\hat{\rho}_{ml} = tr\left\{(\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}} - \boldsymbol{H}_{\tilde{\theta}\tau^2}\boldsymbol{H}^{-1}_{\tau^2\tau^2}\boldsymbol{H}_{\tau^2\tilde{\theta}})^{-1}\boldsymbol{H}^*\right\}|_{\hat{\tilde{\theta}},\hat{b}}. \tag{5.103}$$

Chapter 6

Simulations

To compare the performance and the numerical efficiency of the various Akaike information criteria introduced in Section 5.1, we conducted two simulation studies covering several settings. In the first simulation study, we considered univariate penalized spline smoothing (cf. Chapter4). In the second one, we examined the behavior of the cAICs and the mAIC in balanced random intercept models with N groups of each J observations per group.

Both simulation studies were structured as follows:

1. $nrep = 250$ simulation data sets were generated for each sample size n (for the second simulation study it is $n = J \times N$) and for each d, the parameter corresponding to the signal to noise ratio.

2. In a main simulation step, a linear model (m_1) and a non-linear model (m_2) were estimated using both ML estimation and REML estimation for all settings, followed by the computation of the corresponding degrees of freedom and the AICs.

3. As a measure for the performance of the Akaike information criteria, the frequency of selecting the more complex model (m_2) for each value of d was returned and illustrated in a graphic for each estimation method and sample size. The non-linear model was considered to be selected whenever its AIC was lower than that of the linear model. If the AICs coincided, the simpler model was chosen.
   Furthermore, scatter plots for all degrees of freedom were displayed for each value of n, d and each type of estimation.

A precise description of the structure, the components and some technical details of the two simulation studies, as well as a detailed presentation of the results will be given in the following two sections.

6.1 Penalized Spline Smoothing

6.1.1 Structure

For univariate penalized spline smoothing (4.2), we considered three classes of non-linear functions:

1. $f_1(x) = -2.5 + x + 5d(0.3 - x)^2$
2. $f_2(x) = 1 + x + d(log(0.1 + 5x) - x)$
3. $f_3(x) = 1 + 2x + 1.5d(cos(\frac{1}{2}\pi + 2\pi x) - 2x)$.

Each class depends on the parameter d controlling the degree of non-linearity of the functions. For increasing d, the non-linearity of f_1, f_2, and f_3 is increased. This corresponds to a higher signal-to-noise ratio ${}^{\tau^2}\!/_{\sigma^2}$. On the other hand, when d equals zero, the three functions reduce to linear functions in x. Setting $d = 0$ yields

1. $f_1(x) = -2.5 + x$
2. $f_2(x) = 1 + x$
3. $f_3(x) = 1 + 2x$.

The following seven values were considered for d:

$$d \in \{0, 0.1, 0.2, 0.4, 0.8, 1.2, 1.6\}$$

The courses of f_1, f_2, and f_3 for varying values of the non-linearity parameter d are shown in Figure 6.1. Furthermore, we chose the sequence of sample sizes as

$$nseq = 30, 50, 100, 200.$$

For each of the 168 settings[1], we generated $nrep = 250$ data sets (containing x and y) as follows:

1. x of length $n \in nseq$ was chosen equidistantly from the interval $[0, 1]$.
2. The response variable y was generated as

$$y = f_k(x) + \varepsilon, \text{ with } k \in \{1, 2, 3\}, \ \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

   with the respective non-linearity parameter d. In analogy to Greven and Kneib (2010) the error variance was set to σ^2=1.

[1] 2(estimation types)× 7($dseq$) × 4($nseq$) × 3(functions).

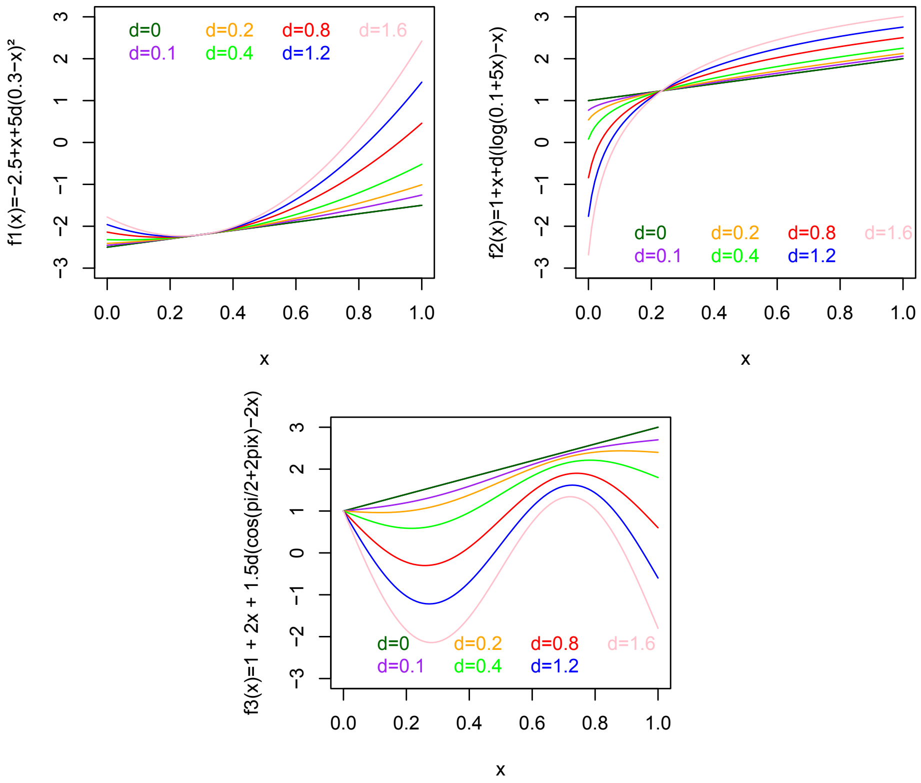

Figure 6.1: *Functions estimated non-parametrically in the simulation study using penalized spline smoothing for varying d.*

Since model comparison via the marginal AIC using restricted maximum likelihood estimation requires equal fixed effects (compare Section 5.1.1), a re-parametrization of the original data was carried out at the end of the data generation step. That is, a non-linear model was estimated to the original data followed by the extraction of the matching design matrix $\boldsymbol{X}$. This matrix is composed of an intercept column (consisting of one's) and a second column of which the entries are transformations of the original x. These transformations were then used for the estimation of the linear model, such that the linear model as well as the non-linear model used the same design matrix $\boldsymbol{X}$. Note that this was conducted at the end of the generation step and that the calculation of functions f_1, f_2, and f_3 was still carried out with the original data x.

Because Greven and Kneib (2010) showed that there is a close agreement between the consideration of Φ_1 and $\Phi_0 + 1$ (compare Section 5.1.2), we focused in the simulation studies of this work on the case with known σ^2. Thus, primary the case $\Phi_0 + 1$ was considered. Note that this step simplifies the calculations, as especially Φ_1 of the approximate cAIC (5.18) is numerically very expensive and possibly instable. Obviously, the asymptotic version of the conventional cAIC (5.10) is not affected by this step. For the marginal AIC, the error variance is accounted for by adding one in any case.
For the covariance based measures, the consideration of unknown error variance does not involve additional expenses (no additional bootstrap replications are needed). For this

reason, Efron's measures with unknown σ^2 (5.48) were also included in the simulations.[2]

In the main simulation step, for each

- f in $fseq = f_1,\ f_2,\ f_3$
- n in $nseq = 30, 50, 100, 200$
- d in $dseq = 0, 0.1, 0.2, 0.4, 0.8, 1.2, 1.6$
- and both ML and REML estimation,

the two models m_1 and m_2 were fitted to the corresponding data, followed by the extraction of all relevant model components. In analogy to the simulation studies of Greven and Kneib (2010), cubic B-Splines with ten inner knots and a second order difference penalty were used to specify the non-parametric effects. The mixed model representation from Section 4.3 yields a mixed model with a fixed linear effect in x, and random effects accounting for the deviation from this linear effect.

For a more detailed depiction of the functions and their structure, see Appendix C and the attached `R − Code` (on the accompanying disc).

6.1.2 Components

The following model components were extracted for the linear model m_1 [3]

- the design matrix $\boldsymbol{X}$,
- the estimated predictors $\boldsymbol{X}\hat{\beta}_1$,
- the maximized log-likelihood $log\left(f(y|\hat{\beta}_1)\right)$,
- and the estimated error variance $\hat{\sigma}_1^2$.

For the more complex model m_2, we extracted

- the design matrix $\boldsymbol{Z}$ of the representation as a mixed model,
- the estimated fixed effects vector $\hat{\beta}_2$,
- the estimated predictors $\boldsymbol{X}\hat{\beta}_2 + \boldsymbol{Z}\hat{b}$,

[2] More precise, the implementation with re-estimated error variance in each bootstrap sample (compare Section 5.1.2).

[3] In the following, indices 1 and 2 denote whether the quantities belong to model m_1 and m_2.

- the estimated fixed part of the predictor $\boldsymbol{X}\hat{\beta}_2$,
- the maximized conditional log-likelihood $log\left(f(y|\hat{\beta}_2, \hat{b}, \hat{\tau}^2)\right)$,
- the maximized marginal log-likelihood (under ML and REML estimation),
- the estimated random effects variance $\hat{\tau}^2$,
- the estimated error variance $\hat{\sigma}_2^2$,
- and the estimated covariance of the response vector y, $\widehat{Cov}(y) = \hat{\boldsymbol{V}}$.

Based on these quantities, the degrees of freedom and the AICs to be compared were computed, comprising

- the degrees of freedom and the AIC for the linear model (m_1), denoted as `df_m1`, `AIC_m1` (cf. equation (5.1)),
- the conventional degrees of freedom (`dfconvent_m2`) and the conventional conditional AIC (`AICconvent_m2`) for model m_2 (cf. equation (5.10)),
- the analytic degrees of freedom (`dfanalyt_m2`) and the corresponding conditional AIC (`AICanalyt_m2`) for model m_2 (cf. equation (5.27)),
- the approximate degrees of freedom (`dfapprox_m2`) and the associated conditional Akaike information criterion (`AICapprox_m2`) for model m_2 (cf. equation (5.14)),
- the conditional and the joint version (with and without an estimation of the error variance in each bootstrap replication) of the covariance based degrees of freedom and the corresponding conditional AIC for varying numbers of bootstrap replications for model m_2 (cf. equation (5.46) and equation (5.48))[4],
- the degrees of freedom based on Yu and Yau (2011) in its three representations (`dfyuyau_m2` (5.70), `dfyuyau_tausq_in_num_m2` (5.74) and in the representation depending on the conventional measure `dfyuyau_rho_tausq_in_num_m2` (5.79)) as well as the associated conditional Akaike information criteria for m_2 (`AICyuyau_m2`, `AICyuyau_tausq_in_num_m2` and `AICyuyau_rho_tausq_in_num_m2`)[5] (cf. equation (5.67)),
- the degrees of freedom returned by function `logLik {mgcv}` (`dfmgcv_m2`)and the corresponding AIC for the complex model (`AICmgcv_m2`),
- and the marginal degrees of freedom (`mdf_m2`) and the marginal AIC (`mAIC_m2`) for the non-linear model (cf. equation ((5.5) and (5.6)).

An overview of all measures including their titles is given in Table C.1 in Appendix C.

[4] For the exact names see Table C.1 in Appendix C.

[5] Where the index `rho` denotes the representation as function of the conventional degrees of freedom (see (5.70) and (5.79))

6.1.3 Technical Details

All calculations were performed with the statistical software program `R` (R Development Core Team, 2011). The model m_1 was estimated using the function `lm` of the `basic` package, and for the non-linear model m_2 the function `gamm` {`mgcv`} was employed. The latter fits the specified model to the data by a call to the function `lme` {`nlme`} (see Appendix E.1.2) in the case of normal errors and identity link.

Note that since spring 2011 Wood (2011) facilitated the estimation of penalized splines represented as mixed models by use of another function: `gam` in package `mgcv`. This function is commonly used to fit generalized additive models with integrated smoothness estimation. Wood demonstrated in a simulation study that `gam` is numerically more stable and works faster than the estimation by `gamm`.[6] Moreover, for the generalized (non-Gaussian) case, REML estimation is only possible by the use of the function `gam`, as for `gamm` one can only specify REML estimation in the case of Gaussianity. For this reason,we also tried to use `gam` for the estimations.

However, several difficulties arose from the fact that the approach to use `gam` for estimations based on the mixed model representation has not been frequently used so far, which prevented further application of this function as part of this work. First, the reparametrization used was not traceable as the function `gam` does not work internally with independent and identically distributed random effects as it was considered in our simulation studies. Thus, the extraction of the design matrix of the fixed effects, $\boldsymbol{X}$, turned out to be rather complicated under maximum likelihood estimation. Second, as the literature on the algorithms used for the estimation (in the generalized case) is sparse, it remained uncertain in what way exactly the effects and variance components are estimated using `gam`. And third, for the function `logLik.gam`, which is used to extract the maximized log-likelihood and the degrees of freedom which are automatically returned by the package `mgcv` for the use of `gam`-models, there is no possibility to request the use of the REML likelihood (Wood, personal communication). Hence, an entire comparison including the automatically returned measures by the use of `gam` was not feasible.

Except for the described difficulties, one significant advantage of using the function `gamm` is that it has also been used in the simulation studies of Greven and Kneib (2010) who compared the marginal degrees of freedom with the conventional, the approximative, and the analytic degrees of freedom in the linear mixed model. Thus, using `gamm` allowed to compare the current results to the results of Greven and Kneib (2010) and made an extension of their analysis to the degrees of freedom, i.e. the covariance based degrees of freedom and the degrees of freedom based on Yu and Yau (2011), possible. For a description of the use of the function `gamm` {`mgcv`}, see Appendix E.1.2.

The estimation algorithm (using `gamm`) did not always converge. For the cases of convergence failure all parameters were set to '`NA`', such that the number of models which did not converge is available (see the results in Subsection 6.1.4). Furthermore, convergence errors in the computation of the covariance based degrees of freedom were intercepted, counted, and the generation of the respective bootstrap sample was repeated.

[6] At least for the data used in Wood (2011).

Apart from the parameters $nrep$, $dseq$, $nseq$, $fseq$, x, and σ^2, some more input variables had to be specified in order to compute the approximate degrees of freedom based on Liang et al. (2008) and the covariance based degrees of freedom of Efron (2004). First, a value for the disturbance h in (5.19) had to be assigned for the computation of the approximate degrees of freedom. Second, the sequences of numbers of replications (for both versions) for the bootstrap approximations of the covariance based measures had to be specified.

In this simulation study, we chose the small value h to be $h = 0.0001$ as in the simulations of Greven and Kneib (2010). Note that we compared a sequence of numbers in a sub-simulation, but as there was no noticeable change in the resulting degrees of freedom, no other values were considered in the main simulation study due to the high computational costs.

Concerning the number of bootstrap replications, a distinction between the conditional and the joint version was made.

For the conditional version of the covariance approximation, 200 bootstrap replications were used. This number is the result of a detailed analysis on changes of the frequency of selecting the more complex model by varying the number of bootstrap replications. As hardly any changes could be observed between 200 and more replications, one can assume that this number is sufficiently large, at least for a similar setting, i.e for one unknown variance component of $Cov(b) = \boldsymbol{G}$ and a maximal sample size of $n = 200$.

For the joint version, 200 bootstrap replications turned out to be insufficient as additional variability is introduced stemming from the estimation of the random effects variance τ^2 in each bootstrap replication. The analysis with a constant sequence of numbers of replications (`Bootseq`) showed that the performance of the joint covariance based cAIC became worse for increasing sample size. For this reason, we used sequences (`Bootseq`) varying with the sample size n. Based on several tests on adequate sizes, the numbers of replications were chosen as follows. Note that in addition to the total number of bootstrap replications (varying with n), also 80% of it was considered in order to check whether changes in the performance can be detected between both replication numbers or if the lower number would already be sufficient.

1. For $n = 30$: $800, 1000$ bootstrap replications were used.
2. For $n = 50$: $1200, 1500$ bootstrap replications were used.
3. For $n = 100$: $1600, 2000$ bootstrap replications were used.
4. For $n = 200$: $2000, 2500$ bootstrap replications were used.

Consequently, the replication numbers for increasing sample size become comparatively large which implies high computational costs. However, it should be noted that the disadvantages for larger sample sizes do not necessarily devalue the measure itself as one main idea of bootstrap methods is to present an alternative whenever asymptotics do not apply due to small sample sizes. Moreover, in contrast to the approximate cAIC which needs n model fits, the covariance based measure is generalizable to the non-Gaussian case (compare (5.81)).

As in the simulation studies of Greven and Kneib (2010), we introduced a check for zero variance of the form

$$\left| log\left(f(y|\hat{\beta}_1)\right) - log\left(f(y|\hat{\beta}_2, \hat{b}, \hat{\tau}^2)\right)\right| > 5 \times 10^{-03} \tag{6.1}$$

in the implementation for most of the measures. This step was carried out because the variance is not exactly estimated to zero due to numerical imprecision. For those cases where the absolute difference was greater than 5×10^{-03}, the penalty terms were set to the penalty term of the simpler model m_1. In this simulation study the degrees of freedom for model m_1 were equal to three.[7]

The absolute difference of the maximized log-likelihood of model m_1 and model m_2 was used instead of the estimated parameter $\hat{\tau}^2$ itself, e.g. $\hat{\tau}^2 > \epsilon$ ($\epsilon > 0$), as the scaling of variances complicates the search of a suitable threshold value. The threshold 5×10^{-03} is based on tests conducted for the simulation studies of Greven and Kneib (2010).

In the following, the binary variable, indicating whether the estimated variance is considered to be zero or not (based on the check for zero variance (6.1)), will be denoted as `var_null`, with

$$\texttt{var\_null} = \begin{cases} 0, & \text{if the absolute difference is greater than } 5 \times 10^{-03} \\ 1, & \text{else.} \end{cases}$$

The check for zero variance (6.1) was included in the implementation of the following measures:

- For the conventional degrees of freedom *var_null* was considered as it is proved that the degrees of freedom simplify to those of the linear model for zero random effects variance. One can therefore avoid computations by introducing the check for zero variance.

- For the analytic degrees of freedom the check for zero variance was used for the same reasons and because in parameter s in Theorem 1 a check for variance components which are estimated to zero is implicitly included. This is not the case for its approximate version of Liang et al. (2008) for which the derivatives are used. Therefore a check is not necessary for the approximate degrees of freedom.

- For the covariance based measures two variants were considered (in the final version). In the first, the check for zero variance was only introduced such that for the joint version the random effects were drawn from a $\mathcal{N}(0, 0)$ distribution (i.e. set to zero) instead of from $\mathcal{N}(0, \hat{\tau}^2)$ distribution for an absolute difference of the maximized log-likelihoods greater than the threshold. The corresponding cAIC will be further denoted as `AICcov_m2_joint` and the conditional analogue (for which no check was included) as `AICcov_m2_cond`[8]. The check for zero variance was introduced here as the results of the analysis without a check indicated numerical problems in the

[7] $2 + 1$ as $\Phi_0 + 1$ was considered in order to account for the error variance.

[8] Note that the corresponding number of bootstrap replications is added in the way: e.g. `AICcov_m2_cond_Boot200`.

joint case for small values of d.[9] Note that the joint version contains more sources of variability as the random effects are as well drawn from a distribution. This makes it more sensitive to numerical imprecisions and instabilities in the estimation.
The second variant contains the check for zero variance for either bootstrap version, the joint and the conditional. The degrees of freedom were set to the degrees of freedom of the linear model whenever

$$\left| log \left(f(y|\hat{\beta}_1) \right) - log \left(f(y|\hat{\beta}_2, \hat{b}, \hat{\tau}^2) \right) \right| \leq 5 \times 10^{-03}.$$

This step was conducted as – especially for large sample sizes – both measures still suffered from numerical imprecisions in the range of small d. It also enabled a better comparison to the other measures. The cAIC with a check like this are denoted as `AICcov_m2_joint_check` and `AICcov_m2_cond_check`[10].

- The check for zero variance was also inserted in the computation of the degrees of freedom of Yu and Yau (2011) as numerous numerical difficulties (such as cancellation) arose in the computation for small estimates of the random effects variance $\hat{\tau}^2$, leading to negative and very large values for the degrees of freedom. Note that all three representations of the degrees of freedom of Yu and Yau (2011) suffered from this problems and differed (although shown to be theoretically equivalent) very much without the introduction of the check for zero variance.

Note that whenever a matrix was inverted of which it was not sure that it was invertible, it was checked whether the inversion was successful or not. For failure, the respective measure was set to '`NA`'.

In the main simulation step, the function `foreach` {`foreach`} was applied in order to compute the $nrep = 250$ ML and REML estimations. Note that it is only possible on Unix systems to conjoin the packages `foreach` and `doMC` in order to execute `foreach` loops in parallel by using the binary operator %`dopar`% instead of %`do`% which evaluates the expression sequentially. The number of worker processes, that should be used to parallelize the tasks, has to be specified as otherwise the tasks are executed sequentially. The simulations studies of this work were run on a Unix system using all 24 processors available.[11]

[9] Small values of d are associated with a large number of estimations of the random effects variance equal to zero.

[10] And the corresponding number of bootstrap replications is included in the name.

[11] This can be specified with the command: `registerDoMC(cores = 24)`.

6.1.4 Results

In this subsection, first the results of the selection frequency of the non-linear model will be presented, followed by the analysis of the various degrees of freedom and their relationships, visualized by scatter plots. Moreover, some technical details concerning the implementation and the numeric will be given.

Selection Frequency of the Non-Linear Model

Corresponding to the theoretical findings of Greven and Kneib (2010), the conventional cAIC (5.10) led to the largest proportion of decisions for the complex model (m_2) in all settings. The marginal AIC ((5.5) and (5.6)) in contrast showed by far the lowest selection frequency of model m_2 – thus favored the linear model – as expected from the theory and the simulations studies of Greven and Kneib (2010).
The curves of the model choice performance of the approximate cAIC (5.14), the analytic cAIC (5.23) and the cAIC of Yu and Yau (2011) (5.67) lay in between the curves of the conventional cAIC and the marginal AIC. This result applied to either ML or REML estimation, to all sample sizes ($n \in nrep$) and to all three functions f_k ($k \in \{1, 2, 3\}$). Note that all three representations of the degrees of freedom of Yu and Yau always coincided with the check for zero variance (6.1) and only one representation was included in the final simulation study. Results for the function f_1 and for the sample sizes $n = 30$ and $n = 200$ (under ML and REML estimation) are shown in Figure 6.2. Note that an 'optimal curve' would be zero for true linearity ($d = 0$) and would grow rapidly up to one for higher values of d.[12] Complete results can be found in Appendix C.

The results indicated moreover that the function `logLik.gamm{mgcv}` automatically returns the marginal AIC, as not only the selection frequencies but also the degrees of freedom (see Figure 6.11) and therefore the AICs of the two measures coincided exactly in each of the settings. `AICmgcv_m2` was therefore excluded from the further analysis and the figures.
In a comparison of the different implementations of the covariance based cAICs, one could see that the joint version was more affected by both

- the introduction of the check for zero variance (6.1) which sets the degrees of freedom to those of the linear model and
- the re-estimation of the error variance in each bootstrap sample.

[12] Comparable to an optimal ROC-curve.

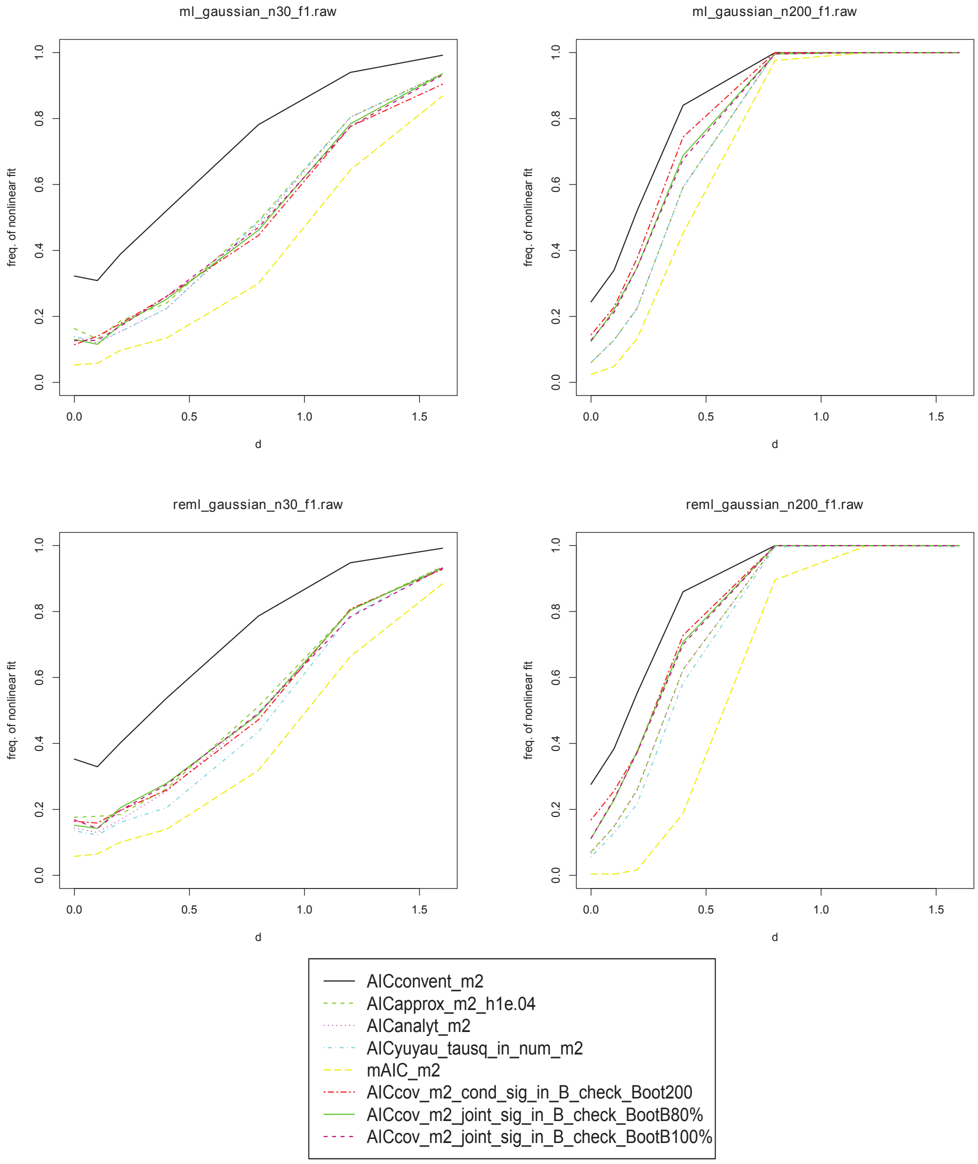

Figure 6.2: *Proportion of simulation replications where the non-linear model m_2 is favored by the respective AIC for function f_1 and sample sizes $n = 30$ and $n = 200$. Above: ML estimation, Below: REML estimation.*

One can exemplarily see in Figure 6.3 that the performance with the check for zero variance and with re-estimated error variances (`sig_in_B`) was (almost in all settings) superior[13] to the other implementations for both the conditional as well as the joint version. The further presentation of the results will therefore be restricted to `AICcov_m2_cond_sig_in_B_check_B200` and the joint analogues, which greatly enhances the clarity of the figures. The associated selection frequency curves lay – as for the other corrected cAICs – between that of the conventional cAIC and that of the marginal AIC (see the green, the dot-dashed red and the dashed purple curves in Figure 6.2).

[13] In the sense of being closer to the curve of the analytic cAIC.

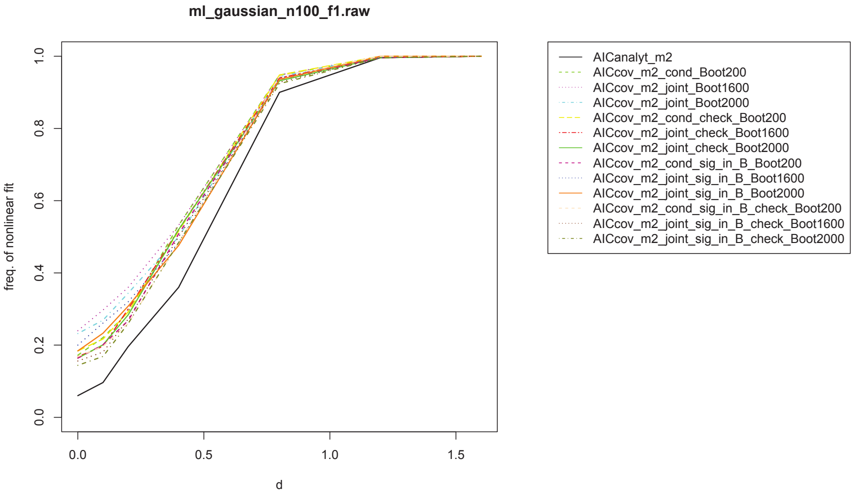

Figure 6.3: *Proportion of simulation replications where the non-linear model m_2 is favored by the respective covariance based AIC and the analytic cAIC for function f_1, ML estimation and a sample size of $n = 100$.*

The comparison of the approximate cAIC and its analytic version showed that the selection frequency of model m_2 was – especially for small values of the non-linearity parameter d – larger for the approximate measure for the case of small sample size n. For $n = 100$ and above the two curves coincided under maximum likelihood as well as restricted maximum likelihood estimation (for all settings). This result can be seen in Figure 6.2 (dashed green curve and dotted purple curve). The observed differences can be traced back to failures of the numerical computation. In many settings, the curve of the approximate cAIC lay above that of the analytic cAIC due to an underestimation of the approximate penalty term. The observed difference between the analytic and the approximate curves might be reduced in future simulations by introducing the check for zero variance into the computation of the approximate degrees of freedom. This would additionally speed up the computations (compare Chapter 8). As an aside, we found that the two measures differed even more when the function `gam` instead of `gamm` was used (see technical details above (Subsection 6.1.3)).

The cAIC of Yu and Yau and the analytic cAIC led (almost generally) to the same decisions in case of maximum likelihood estimation. However, as the former has not been constructed under restricted maximum likelihood estimation, a considerable difference could be observed under REML estimation (see Figure 6.2, dotted purple curve and dot-dashed blue curve). Here, the curve of the cAIC of Yu and Yau lay below that of the analytic cAIC (for all settings) resulting in a greater number of decisions in favor of the linear model.

Regarding the covariance based cAICs (with the check for zero variance and re-estimated error variances), a slight tendency in favor of the joint version could be observed. In most of the cases when the results showed a clear difference between the selection frequency of the conditional cAIC and its joint counterparts, the curves corresponding to the joint measures lay (slightly) closer to that of the analytic cAIC (see for example Figure 6.4, dotted purple curve and green curve and dashed purple curve). Note however that this finding could not be observed throughout all settings and did not apply to all values of the non-linearity parameter d (see Figure 6.5). Moreover, one could see that the selection frequency of the joint cAIC with 80% of the bootstrap replications was very similar to that with 100% of the replications used (see the green curve and the dashed purple curve in the right graphics in Figure 6.2), indicating that the number of bootstrap replications was sufficiently large. For large sample sizes, the two curves were almost indistinguishable.

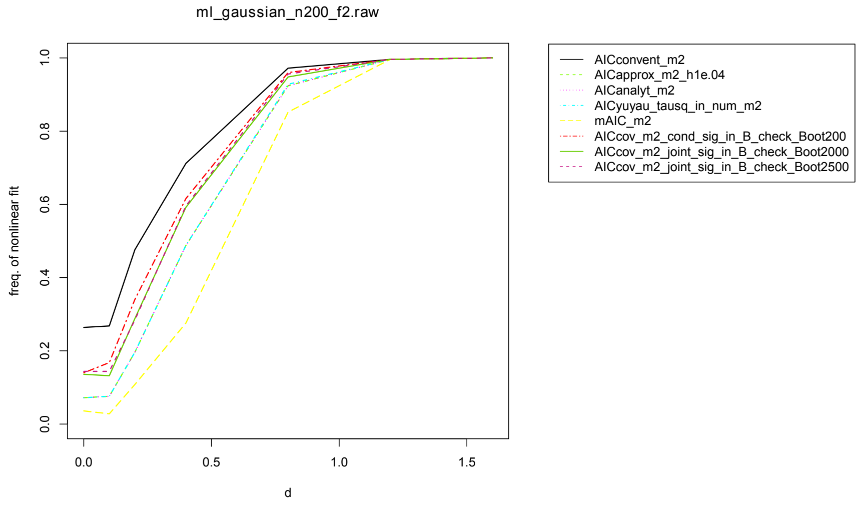

Figure 6.4: *Proportion of simulation replications where the non-linear model m_2 is favored by the respective AIC for function f_2, ML estimation and a sample size of $n = 200$. Here, the curve of the joint cAIC lies considerably closer to the analytic curve than its conditional counterpart.*

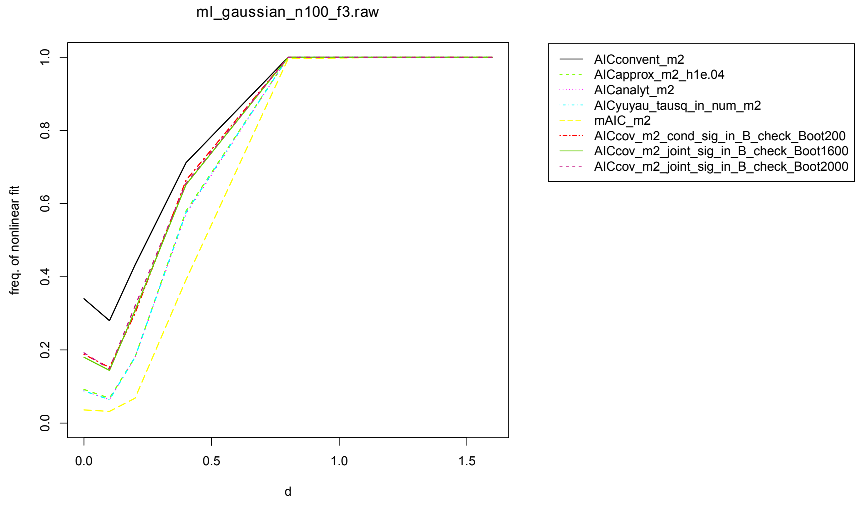

Figure 6.5: *Proportion of simulation replications where the non-linear model m_2 is favored by the respective AIC for function f_3, ML estimation and a sample size of $n = 100$. Here, no noticeable difference in the selection frequency of the joint and the conditional cAICs can be observed.*

For most of the settings, the three curves of the covariance based cAICs (`AICcov_m2_cond_sig_in_B_check_Boot200`, `AICcov_m2_joint_sig_in_B_check_BootB80%` and `AICcov_m2_joint_sig_in_B_check_BootB100%`)[14] were closer to the analytic curve for smaller sample sizes. For large n, the three curves were shifted upwards in direction of the conventional curve (see Figure 6.2). No noticeable differences between ML and REML estimation could be observed. For the three underlying functions (f_1, f_2 and f_3) one could see some differences concerning the closeness of the three curves to each other and to that of the analytic measure. Furthermore, the closeness of the conditional to the joint curves could not be traced back to a systematic effect depending on the sample size, nor depending on the non-linearity parameter d.
One could see that, especially for small values of d, the curves of the covariance based cAICs sometimes tended to be unsteady, to have unexpected kinks and to differ from the behavior for greater values of the non-linearity parameter (see for example Figure 6.6). This occurred much more frequent without the check for zero variance, but sometimes even when the check was included. This suggests that the check did not remedy all numerical problems. It should moreover be noted that the computation of the covariance based cAICs was not (completely) stable, i.e. a repeated run of the simulations (based on the same data) led to different decisions (at least without the check for zero variance), especially in the range of small ds. It was therefore difficult to attain a clear preference for either the conditional or the joint version. Yet, as will become clear in the next section, the results of the second simulation study support the – here slightly indicated – preference for the joint measure.

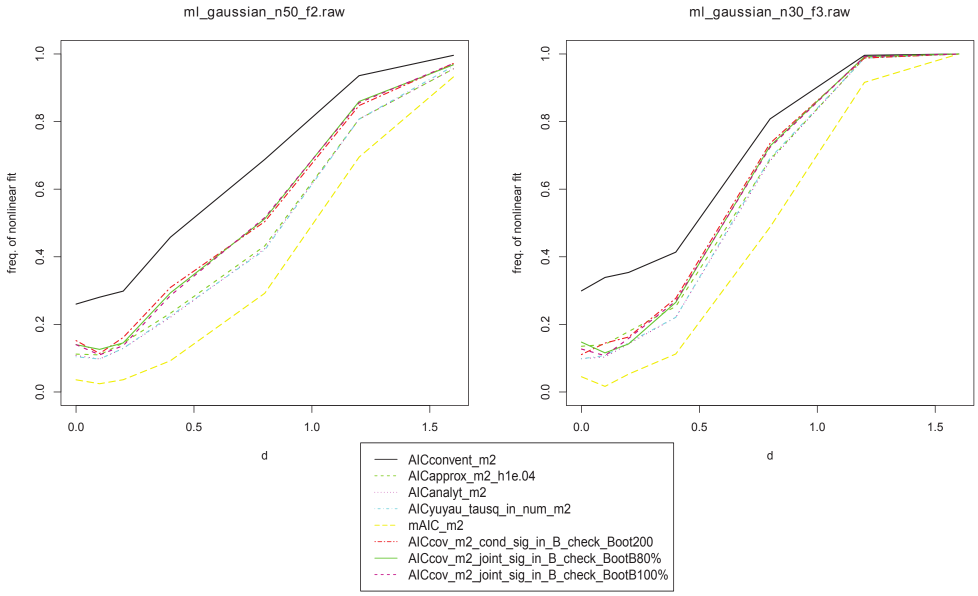

Figure 6.6: *Proportion of simulation replications where the non-linear model m_2 is favored by the respective AIC under ML estimation. On the left for function f_2 and a sample size of $n = 50$. On the right for function f_3 and a sample size of $n = 30$. In the range of small values of non-linearity parameter d one can observe kinks in the curves of the covariance based.*

[14] With B denoting the number of bootstrap replications used.

Convergence failure (compare the technical details above (Subsection 6.1.3)) in the estimation of the models in each bootstrap sample occurred much rarer with the introduction of the check for zero variance.[15] The model estimation failed the most often under ML estimation and for small values of n. For the joint measure many more failures could be observed than for the conditional. This was probably due to the fact that considerably more replications were used and thus more models had to be estimated. Note that as the greatest number of estimation failures lay below 1% of the estimations performed[16], these numerical issues presumably did not affect the interpretation of the resulting curves and are only mentioned here for reasons of integrity.

Degrees of Freedom

A more precise insight in the connection between the degrees of freedom could be obtained by analyzing the respective scatter plots. The left scatter plot in Figure 6.7 exemplarily shows that the approximate degrees of freedom did not exactly coincide with the analytic degrees in case of small sample sizes and small values of d (red ellipse). As can be seen in the right scatter plot the differences disappeared for larger values of n (for the same d). Moreover, one can observe jumps of the analytic degrees of freedom in this figure.[17] The degrees were either equal to three (for $\left|log\left(f(y|\hat{\beta}_1)\right) - log\left(f(y|\hat{\beta}_2, \hat{b}, \hat{\tau}^2)\right)\right| \leq 5 \times 10^{-03}$) or greater than four, but no values arose in between. For the approximate degrees of freedom, this effect could (with some numerical deviations) also be observed, as well as for the degrees of freedom of Yu and Yau under ML estimation. The source of these jumps has not been identified so far.

[15] In total almost 3 times less failures occurred for the conditional version and for the joint version it was more than 2.5 times less.

[16] And below 2% for the implementations without the check for zero variance.

[17] Note that the jumps occurred also in the simulations of Greven and Kneib (2010).

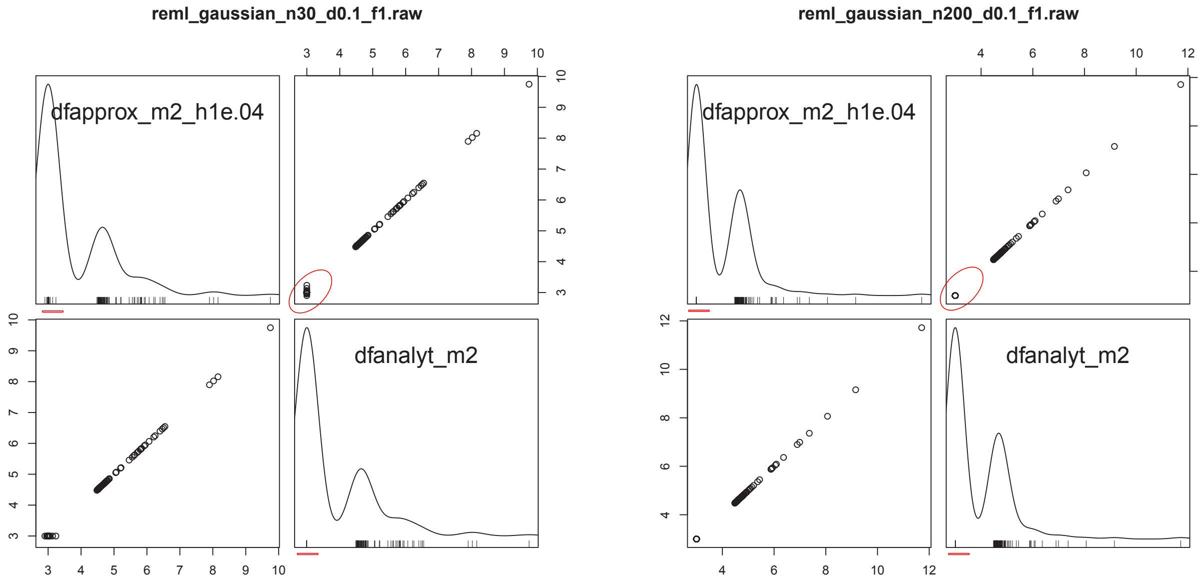

Figure 6.7: *Scatter plot matrix of the analytic degrees of freedom and the approximate degrees of freedom for function f_1, REML estimation and $d = 0.1$. In the left plot one can see the results for $n = 30$ and in the right for a sample size of $n = 200$. The red ellipses and lines highlight the differences for small and large sample size in the behavior of the approximate degrees of freedom.*

As indicated by the selection frequency plots, the cAIC of Yu and Yau and the analytic cAIC were very similar under maximum likelihood estimation. This correspondence was also observable (especially for large sample sizes) in the scatter plots of the associated degrees of freedom (see for example the right plot in Figure 6.8). However, for small sample sizes there were still some differences, as can be seen in the left plot in Figure 6.8. Under REML estimation the degrees of freedom of Yu and Yau differed from the analytic degrees (see Figure 6.9). Extremely large and even negative values appeared for `dfyuyau_tausq_in_num_m2` (see Figure 6.10).

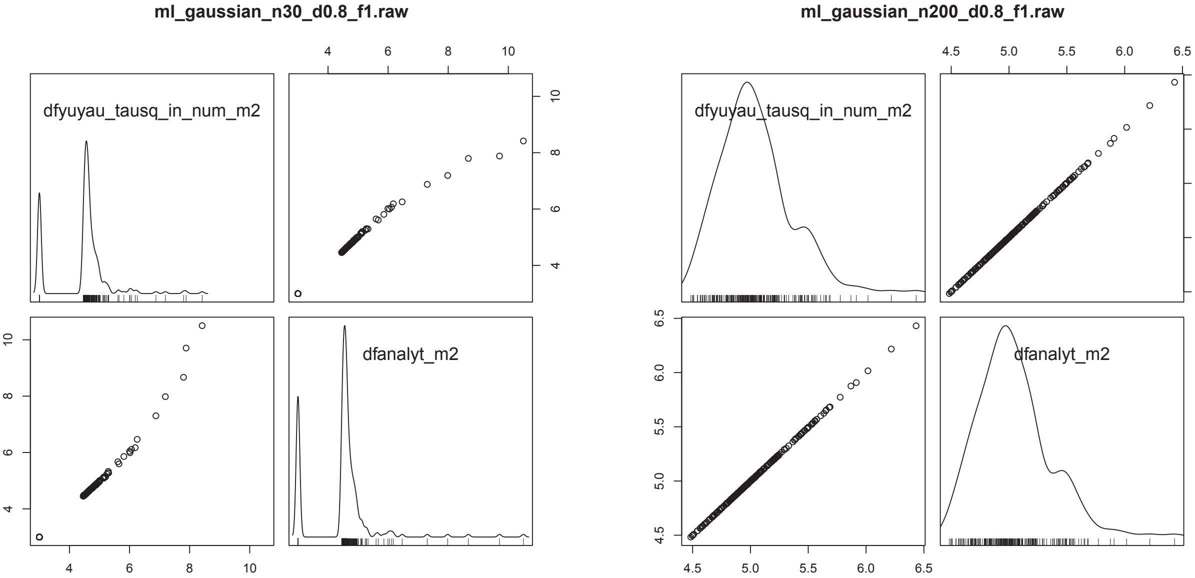

Figure 6.8: *Scatter plot matrix of the analytic degrees of freedom and the degrees of freedom of Yu and Yau for function f_1 under ML estimation for $d = 0.8$. On the left, the sample size is $n = 30$ on the right it is $n = 200$.*

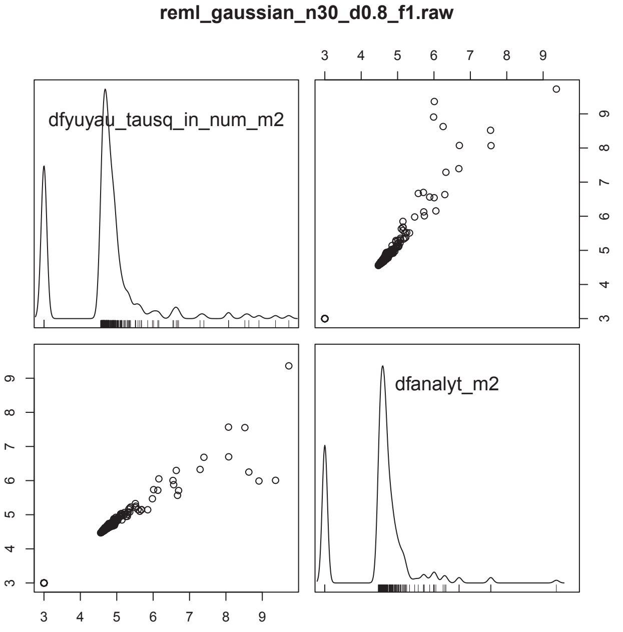

Figure 6.9: *Scatter plot matrix of the analytic degrees of freedom and the degrees of freedom of Yu and Yau for function f_1 under REML estimation for sample size $n = 30$ and for $d = 0.8$.*

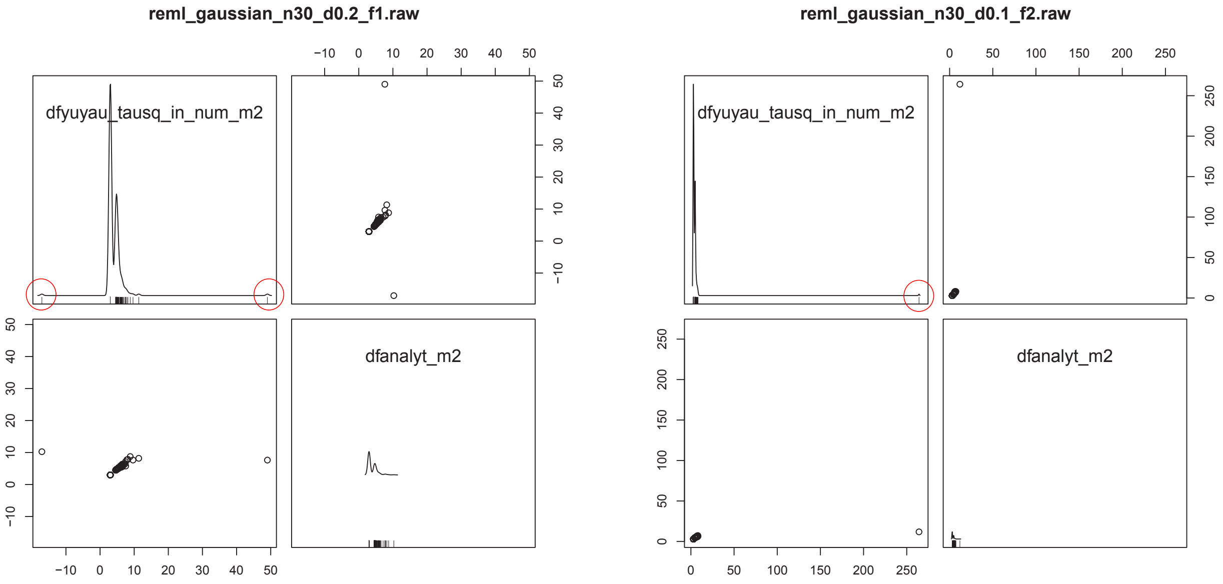

Figure 6.10: *Scatter plot matrix of the analytic degrees of freedom and the degrees of freedom of Yu and Yau for function f_1, REML estimation and sample size $n = 30$. On the left, the non-linearity parameter $d = 0.2$ on the right it is $d = 0.1$. Negative and very large values of the degrees of freedom of Yu and Yau are highlighted by red circles.*

In Figure 6.11 one can see that the function `logLik.gamm{mgcv}` automatically returns the marginal degrees of freedom. This has already been indicated by the selection frequency plots. There could, however, have been a minimal difference of the two measures – as only the proportion was shown to be identical in the selection frequency plots – which could be ruled out by the analysis of the scatter plots (and further analysis of the results).

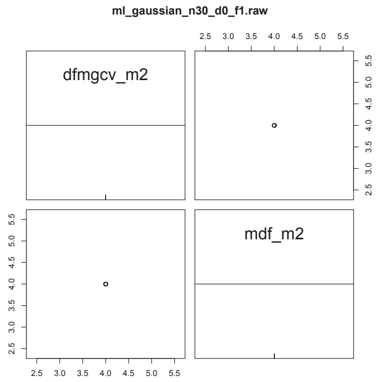

Figure 6.11: *Scatter plot matrix of the marginal degrees of freedom and the degrees of freedom automatically returned by package* `logLik.gamm{mgcv}` *for function f_1, ML estimation, $n = 30$ and $d = 0$. One can exemplarily see here that the two degrees of freedom are equal. They were always equal to four as we considered the case of one random effect and without any covariates. Recall that the marginal degrees are given as $2(p + q + 1)$ in the ML case and as $2(q + 1)$ in the REML case (cf. (5.5) and (5.6)).*

Implementation and Numerical Issues

Overall, the simulation time amounted to almost ten days (including the implementations with and without the check for zero variance (6.1) of the covariance based cAICs). In the estimation of the more complex model m_2 154 convergence failures (amounting to less than 1% of all 32,000 simulations) occurred. For these cases all measures were set to '`NA`'. No non-invertible matrices appeared in the estimation of the various measures.

Some major numerical problems occurred in the computation of the degrees of freedom of Yu and Yau, which is why the check for zero variance was introduced in the implementation. It should be noted that without the check for zero variance the selection frequency curves did not – also not under ML estimation – resemble the curves of the analytic cAIC. Without the check for zero variance, highly negative and very large values appeared for the degrees of freedom of Yu and Yau (see Figure 6.10) and the three representations (`dfyuyau_m2`, `dfyuyau_tausq_in_num_m2` and in the representation depending on the conventional measure, `dfyuyau_rho_tausq_in_num_m2`) did not correspond. These problems could be traced back to numerical cancellation for small values of $\hat{\tau}^2$. For the representation in which it is divided by the estimated random effects variance, it seems very natural that problems arise. Yet, the representations in which $\hat{\tau}^2$ appears only in the numerator were also problematic, probably due to the fact that terms which include the (estimated) random effects variance have to be inverted. A detailed analysis of the components of the computation of `dfyuyau_tausq_in_num_m2` moreover showed that matrix $\boldsymbol{U}$ in equation (5.74) was responsible for at least parts of the numerical difficulties. Although it theoretically is a symmetric matrix, some eigenvalues of $\boldsymbol{U}$ turned out to be complex numbers. To prevent these computational inaccuracies, the matrix was artificially made symmetric by using the function `forceSymmetric` of the `Matrix`-package.

6.2 Random Intercept Model

The main structure of the simulations for random intercept models remained the same as in the simulations of penalized spline smoothing. However, as the structure of the simulated data was rather different and another function was used for the estimations, the second simulation study will also be quickly described in the following. Furthermore, a summary of the results will be given and the findings will be compared to the results of the first simulation study (see Section 6.3).

6.2.1 Structure

For the analysis of the random intercept models (compare Definition 6), N clusters of each $J_i = J$, $\forall i$, observations were considered, whereby the number of groups was chosen as

$$N = 10$$

and the cluster sizes were specified as

$$J \in \{3, 6, 9, 12\}.$$

The random effects b_{0i} in equation (3.45) were drawn independently from a $\mathcal{N}(0, d)$ distribution, such that the random effects variance $\tau^2 = d$ again is a measure of the signal-to-noise ratio ${}^{\tau^2}\!/_{\sigma^2}$ as in Section 6.1.[18]
As in the simulation study using penalized spline smoothing, only the case of known error variance was considered and again σ^2 is set to one. Note that no intercept was used in the generation of the data, i.e. $\beta_0 = 0$. For the random effects variance d the same seven values as in Section 6.1 were used, thus

$$d \in \{0, 0.1, 0.2, 0.4, 0.8, 1.2, 1.6\}$$

was considered. Obviously, the sample size n can be determined as

$$n = N \times J.$$

Consequently, there were 56 settings[19] for which $nrep = 250$ data sets (containing y and id, a variable specifying the cluster structure) were generated as follows:

1. The response variable y was generated as the sum of a random intercept $b_{0i} \sim \mathcal{N}(0, d)$ for each cluster and an error term $\varepsilon \sim \mathcal{N}(0, 1)$.
2. A factor variable id with values $1 : N(= 10)$, specifying to which cluster the respective observation belongs, was added.

[18] Compare Greven and Kneib (2010).
[19] 2(estimation types)× 7($dseq$) × 4(values of J).

The two models m_1 (linear model) and m_2 (the random intercept model), which were fitted in the following, had the form

$$
\begin{aligned}
m_1 &: y = \beta_0 + \varepsilon_i, \\
m_2 &: y = \beta_0 + b_{0i} + \varepsilon_i,
\end{aligned}
$$

for $i = 1, \ldots, N$.

Note that for the random intercept model, no additional re-parameterizations to ensure comparability of model m_1 and m_2 had to be taken into consideration as the fixed effects design matrix only comprised a column of ones, and thus corresponds to the global intercept of which the simpler model m_1 consisted (except for the error term). Thus, the fixed effects design matrix $\boldsymbol{X}$ was the same for both models.

The loops in the main simulation step cycled through

- the cluster sizes $J \in \{3, 6, 9, 12\}$ and
- the non-linearity parameter $d \in \{0, 0.1, 0.2, 0.4, 0.8, 1.2, 1.6\}$.

Again, for each d and J, the models m_1 and m_2 were fitted to the corresponding data under each estimation method, i.e. by ML estimation and by REML estimation. The following extraction of the required components could be carried out straightforward, in contrast to the extraction in the previous simulation, as no additional functions had to be used.
For further information on the implementation see the attached R-code (on disc).

6.2.2 Components

The same model components were extracted for the models m_1 and m_2 as for penalized spline smoothing. The thereupon computed degrees of freedom and AICs are denoted in analogy to the previous simulation with the difference that a different function was used for the estimation and therefore the degrees of freedom and the maximized log-likelihood automatically returned by the program does not correspond to that of Section 6.1. Instead of `dfmgcv_m2` and `AICmgcv_m2`, the associated measures are denoted as `dfnlme_m2` and `AICnlme_m2` in accordance with the package used (see below).

6.2.3 Technical Details

As before, `R` was used for the simulation. More precisely, the function `lm {basic}` was used for the estimation of the simpler model m_1 and the fit of the random intercept model was performed with the use of the function `lme` of the package `nlme` (compare 3.1.7 and

Appendix E.1.1). Note that the same function was used for the simulation study using random intercept models in Greven and Kneib (2010) and that the results are thus comparable.

To facilitate comparison, convergence failures would have been treated as in the penalized smoothing simulation, i.e. set to "`NA`" (no "`NA`"s occurred (see Subsection 6.2.4)).

The disturbance h in the definition of the approximate degrees of freedom (5.19), was again set to $h = 0.0001$ and the number of bootstrap replications was adjusted to the sample size. The following numbers were considered:

1. For $J \times N = 30$: $800, 1000$ bootstrap replications were used.
2. For $J \times N = 60$: $1200, 1500$ bootstrap replications were used.
3. For $J \times N = 90$: $1600, 2000$ bootstrap replications were used.
4. For $J \times N = 120$: $1600, 2000$ bootstrap replications were used.

Note that the check for zero variance (6.1) from the simulation study using penalized spline smoothing was also introduced in this simulation study for the computation of Vaida and Blanchard's conventional cAIC, the analytic cAIC of Greven and Kneib (2010), the bootstrap based measures based on Efron (2004), and the conditional Akaike information criterion proposed by Yu and Yau (2011) (in its three representations), with no changes to Section 6.1.

The parallelization of the main simulation step was done as in the first simulation study (compare the technical details in Subsection 6.1.3).

6.2.4 Results

The results which will be given for the simulation using random intercept models include – as in the previous section – the selection frequencies of the more complex model (m_2), an analysis of the degrees of freedom themselves and finally some technical details on the implementation. Note that the presentation of the results will be followed by a comparison of the results of the two simulation studies in the next section (Section 6.3).

Selection Frequency of the Non-linear Model

The selection frequency curves clearly correspond to the theoretical findings of Greven and Kneib (2010). Similar to the first simulation study, the conventional cAIC (5.10) showed the highest selection frequency of the non-linear model (m_2) throughout all settings, whereas the marginal AIC ((5.5) and (5.6)) led to the lowest number of decisions in favor of model m_2. The curves of the corrected cAICs were all placed in between these two extremes. Results for group sizes $J = 3$ and $J = 12$ under either ML and REML estimation are shown in Figure 6.12. Complete results can be found in Appendix C.

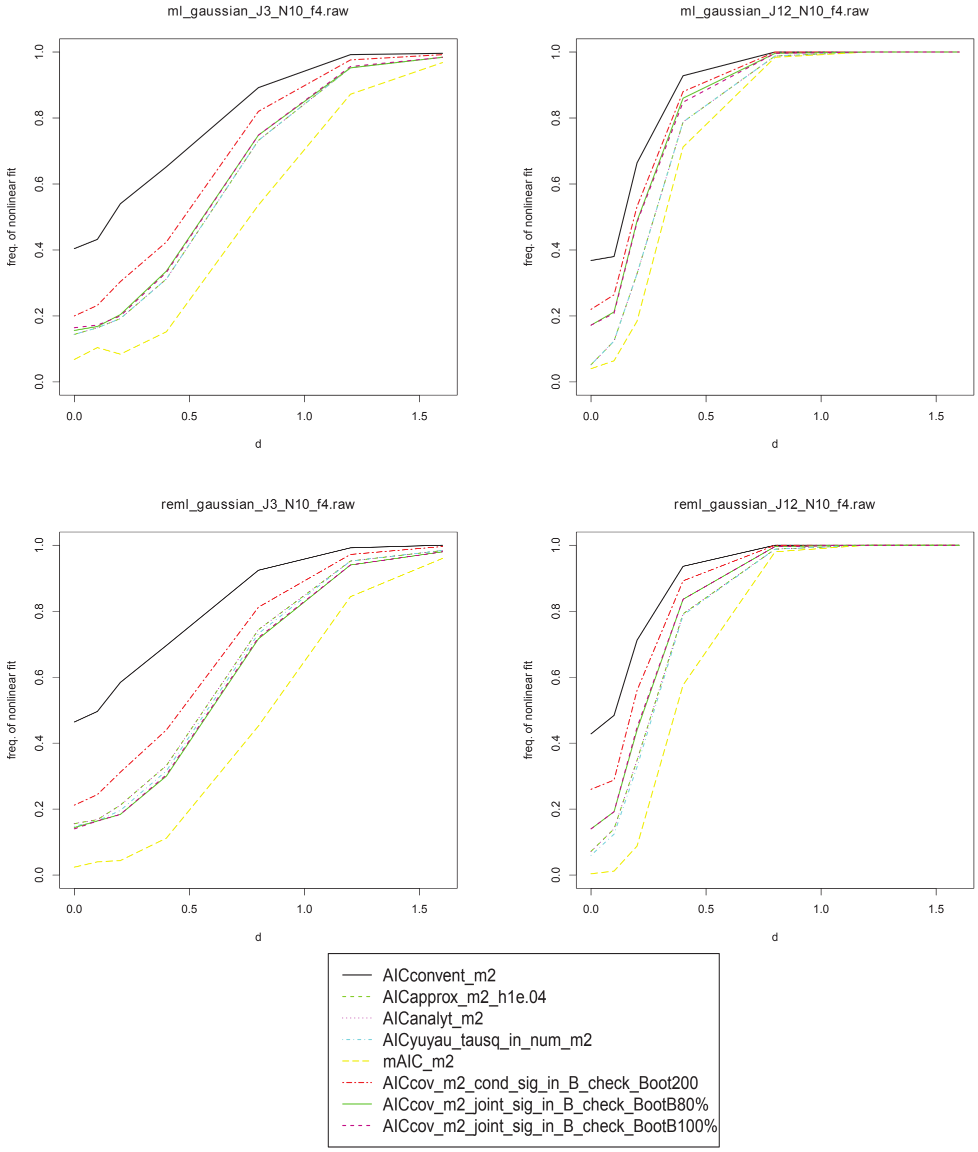

Figure 6.12: *Proportion of simulation replications where the non-linear model m_2 is favored by the respective AIC for group sizes $J = 3$ and $J = 12$. Above: ML estimation, Below: REML estimation.*

Note that again all three representation of the degrees of freedom of Yu and Yau (2011) coincided (when the check for zero variance (6.1) was included). Hence, only `dfyuyau_tausq_in_num_m2` (5.74) was further used in the simulation study. Moreover, it turned out that the automatically returned degrees of freedom of the function `logLik.lme{nlme}` are equal to the marginal degrees of freedom (as it is the case for the corresponding function in package `mgcv`).[20] Due to this equality, only the marginal AIC was included in the further analysis.

In contrast to the simulation using penalized spline smoothing, the conditional as well as the joint versions (80% and 100% of the bootstrap replications) of the covariance based selection frequency curves remained almost unaffected by the introduction of the check for zero variance. However, both – and especially the joint versions – were highly affected by the re-estimation of the error variance (5.48). In analogy to the first simulation study, it turned out that the covariance based measures with re-estimated error variance (and with the check for zero variance included in the implementation) were superior to the other variants, as one can exemplarily see in Figure 6.13. The presentation of the results will therefore (and for reasons of comparability to the first simulation study) be restricted to `AICcov_m2_cond_sig_in_B_check_Boot200` and its joint counterparts.

It can be seen (e.g. in Figure 6.12) that the curves of the joint version that used only 80% of the bootstrap replications (green curve) almost coincided with those for which all bootstrap replications were taken into account (dashed purple curve). This (again) indicates that the selection of the number of bootstrap replications was sufficiently large (compare the results of the first simulation study in Subsection 6.1.4).
In contrast to the first simulation study, the results for random intercept models showed a clear preference for the joint version over the conditional version as the corresponding curves lay much closer to the analytic curve. This applied to all settings and can be seen e.g. in Figure 6.12. One explanation for the superiority of the joint version is that it accounts for more variability since the random effects were redrawn for each bootstrap sample.
As in the first simulation study, one could see that the covariance based selection frequency curves (`AICcov_m2_cond_sig_in_B_check_B200`, `AICcov_m2_joint_sig_in_B_check_B80%`, and `AICcov_m2_joint_sig_in_B_check_B100%`) departed from the analytic curve for larger sample sizes. For great values of $n = J \times N$ one could observe an upward shift in direction of the conventional curve (see Figure 6.12). Again, no visible difference could be found between the results of ML and REML estimation.

A Comparison of the approximate cAIC of Liang et al. (5.14) and its analytic version (5.23) showed that the associated curves exactly corresponded to each other with the exception of one setting. For group size $J = 9$ a minimal discrepancy could be observed under REML estimation in the range of small values of the non-linearity parameter d (see the dashed green curve and the dotted purple curve in Figure 6.14). Details on the actual values of the degrees of freedom will be given in the following passage.

[20] Function `gamm{mgcv}` calls function `lme{nlme}` in the case of normal errors and identical link. It is therefore obvious that both functions lead to the same automatically returned degrees of freedom.

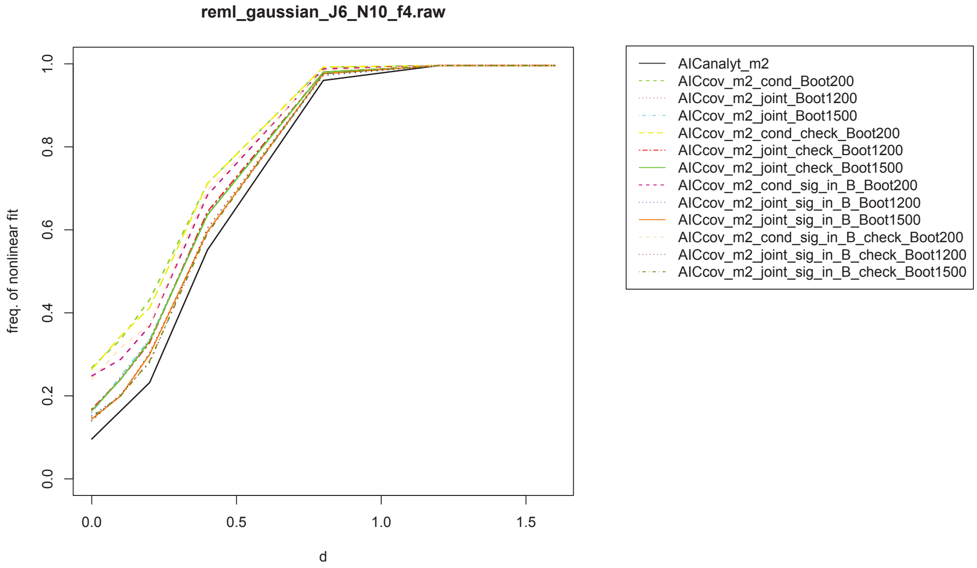

Figure 6.13: *Proportion of simulation replications where the non-linear model m_2 is favored by the respective covariance based AIC and the analytic cAIC under REML estimation and for a group size of $J = 6$.*

Similar to the simulation using penalized spline smoothing, the selection frequency curves of the analytic cAIC and of the cAIC of Yu and Yau coincided under maximum likelihood estimation. For the random intercept models, the two curves were even identical throughout all ML settings. Under REML estimation, the curves associated to `AICyuyau_tausq_in_num_m2` lay again below the analytic curves. The measure of Yu and Yau thus led more often to decisions in favor of the simpler model m_1 than it was the case for the analytic cAIC. It should be noted however that for large group sizes the two curves were almost identical.

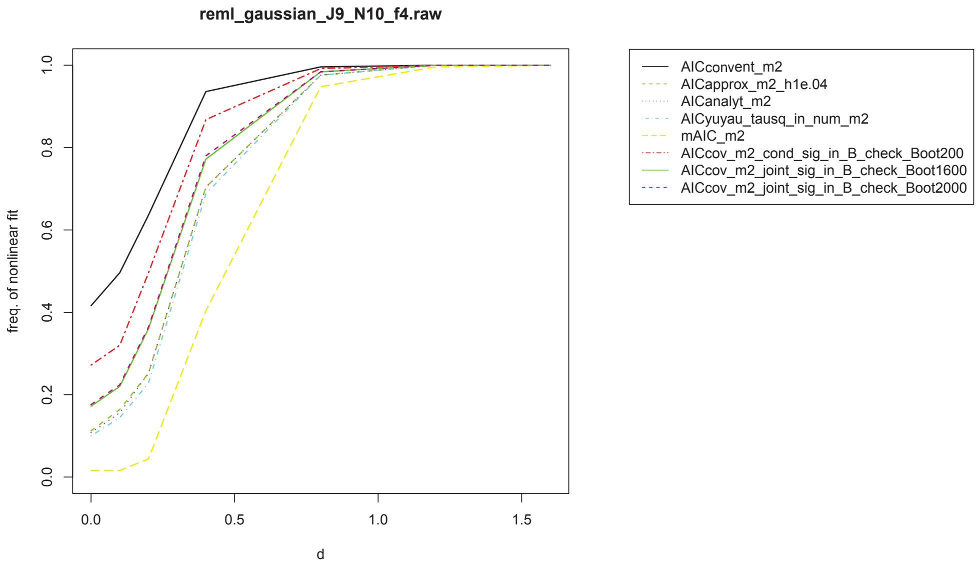

Figure 6.14: *Proportion of simulation replications where the non-linear model m_2 is favored by the respective AICs under REML estimation and for a group size of $J = 9$. For small values of d one can see a slight difference of the approximate and the analytic curve.*

Degrees of Freedom

The scatter plot matrices in Figure 6.15 exemplarily show that the approximate and the analytic degrees of freedom were equal except for minor deviations. These outliers could be mostly found for small values of d. In the left plot one can see the results for group size $J = 12$ under ML estimation for true linearity ($d = 0$). The red circles show the values which do not correspond between the approximate and the analytic degrees of freedom. The right scatter plot matrix displays the results for $J = 9$ under REML estimation and true linearity. Here, one outlier could be detected for the approximate measure (red circle). Recall, that in the selection frequency curves a slight deviation could be observed for the same setting (REML, $J = 9$ and small values of d). However, the other deviations in the degrees of freedom did not affect the selection frequencies of the non-linear model. It was therefore essential to additionally analyze the scatter plots in order to investigate the behavior of the measures.

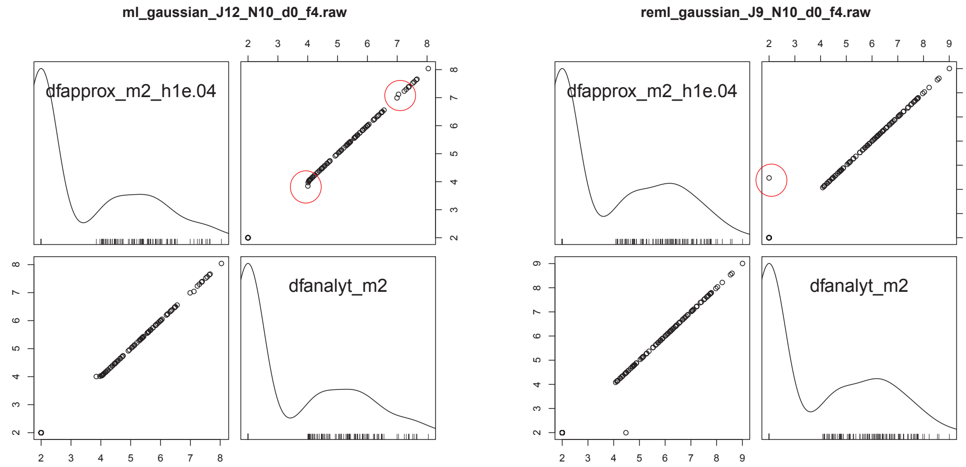

Figure 6.15: *Scatter plot matrix of the analytic degrees of freedom and the approximate degrees of freedom for true linearity $d = 0$. On the left: Results for group size $J = 12$ under ML estimation. On the right: Results for group size $J = 9$ under REML estimation. The red circle highlights the deviations of the approximate degrees of freedom to the analytic degrees of freedom.*

No negative nor very large values occured for the degrees of freedom of Yu and Yau in this simulation study. It can be seen in Figure 6.16 that under ML estimation (left) the degrees of freedom exactly corresponded to the analytic degrees. The right plot shows that under REML estimation there was a shift in accordance with the findings of the analysis of the selection frequency.
The jumps observed in the analysis of the first simulation study also appeared in the random intercept simulation (see Figure 6.15). For the analytic degrees of freedom (and due to the equality also for the degrees of Yu and Yau under ML estimation) the jumps could be detected throughout all settings. For the approximate degrees of freedom they appeared for most settings.

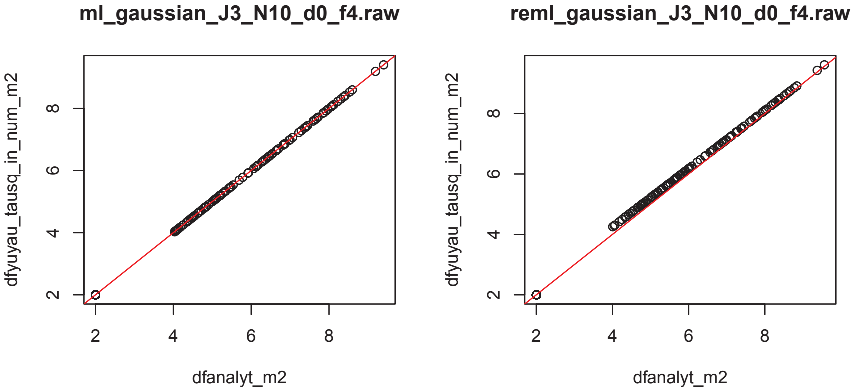

Figure 6.16: *Scatter plots of the analytic degrees of freedom and the degrees of freedom of Yu and Yau for true linearity $d = 0$ and a group size of $J = 3$. On the left: Results under ML estimation. On the right: Results under REML estimation. The angle bisector is marked as a red line.*

Implementation and Numerical Issues

The simulation using random intercept models run approximately one day. This is considerably shorter than the running time of the first simulation study. Note that the computation time differences arose for several reasons. First, the number of settings was substantially smaller in the second simulation study.[21] Moreover, the maximum sample size for the random effect models was $n = 120$ whereas it was $n = 200$ in the first simulation. Computational failure due to divergence in the estimation of the non-linear model (m_2) and in the computation of the covariance based measures was another cause for longer estimation times. No such convergence failures occurred in the simulations for random intercept models, neither in the estimation of model m_2, nor in the computation of the bootstrap based measures.
As in the first simulation no non-invertible matrices occurred.

It should moreover be noted that due to the numerical challenges which arose in the computation of the degrees of freedom of Yu and Yau without the check for zero variance (6.1), we directly implemented the cAIC of Yu and yau here with the check for zero variance included.

[21] Keep in mind that three functions were considered in the first simulation study.

6.3 Comparison of the Two Simulation Studies

In summary, we found that the main results of the two simulation studies largely agreed. The curves of all corrected measures lay in between those of the conventional cAIC and the marginal AIC for either simulation study. Moreover, the closeness of the approximate cAIC to the analytic cAIC could be observed for both studies (with numerical deviations). Furthermore, the results showed that the measure of Yu and Yau differed from the analytic measure under REML estimation, although it was almost identical to the analytic cAIC under ML estimation. For most settings, it turned out that the curves of the covariance based measures lay in between the analytic and the conventional curve.

However, it could be seen that the preference for the joint over the conditional version of the covariance based cAIC was much clearer in the simulation using random intercept models. We furthermore found that the computations of the first simulation were much more susceptible to numerical imprecision and that convergence errors occurred in contrast to the second simulation. This might have been due to the clearly more complex structure (e.g. the correlation structure between the responses) of the simulations on penalized spline smoothing compared to that of the simulations on random intercept models.

It should be kept in mind that approximations was performed in the first simulation study, as the underlying functions f_1, f_2 and f_3 were approximated by polynomial splines (see Chapter 4). One therefore did not only have to deal with an **estimation error**, but also with an **approximation error**. Besides, it should be considered that the normal distribution, from which the random effects were drawn, was an **auxiliary construction**. This was due to the fact that the generation of the data was based on one of the three functions (f_1, f_2, f_3) and the assumptions of the LMM were therefore not (exactly) satisfied. This could be the source of

- the observed differences in the behavior of the cAICs between the three functions. It is possible that the reflection of the underlying functions was of varying quality. The drawing of the random effects might have been unequally representative for f_1, f_2 and f_3.

- the poorer performance of the joint version in the first simulation study (compared to the clear preference of the joint over the conditional version in the second simulation study) as the random effects were re-drawn for each bootstrap sample.

Chapter 7

Case study

In addition to the analysis of the behavior of the AICs in the two simulation studies described in the previous chapter, we conducted the following application on a real data set on childhood malnutrition in Zambia in order to illustrate the practical relevance of the selection of random effects via AICs.
First, the background and the relevance of the data will be quickly elucidated (based on Kandala et al. (2001) and Greven and Kneib (2010)), followed by a brief explanation of the data set and the analysis of some descriptive properties. Then, two univariate smoothing models will be presented for which it was to decide whether non-linear modeling was required or not. This was done by representing the models as mixed models and computing the various AICs which were then compared to the AIC of the corresponding linear model.

7.1 Background and Relevance

Malnutrition – especially among children – is considered to be one of the most urgent and challenging health problems in developing countries such as Zambia and is therefore of great political relevance. It is considered to be one of the main indicators for deprivation and is associated with high mortality rates and poor labor productivity. According to Kandala et al. (2001), no less than 42 percent of Zambian children under the age of five are classified to be stunted, i.e. chronically malnourished (compare the operationalization of stunting in the following) and 18 percent as severely stunted.

In order to investigate the development of acute and chronic malnutrition, regular surveys are produced by demographic and health organizations. The data set on chronic malnutrition of children in the African state Zambia used in this work is the result of the 1992 Demographic and Health Survey (DHS) conducted by Macro International and the Zambian statistical agency.

A representative sample of 6299 women of reproductive age was drawn through stratified clustered sampling. The women were asked to answer questions on themselves and on their children that were born within the five previous years, comprising maternal and child health, education, family planning and other information.

Childhood malnutrition is usually assessed by the anthropometric status of the child, such as weight and height, relative to a reference standard which accounts for the age of the child. Generally, three types of malnutrition are distinguished: Acute undernutrition (measured as insufficient weight for height), chronic undernutrition or stunting (measured as insufficient height for age) and underweight (measured as insufficient weight for age) which can be a result of the first two types of malnutrition. As in the case study of Greven and Kneib (2010), the focus in this work lies on chronic undernutrition, quantified by the Z-score

$$zscore_i = (cheight_i - m)/s, \text{ for child } i, \tag{7.1}$$

where $cheight_i$ denotes the individual height of the child, m refers to the median height of children of the same age from a reference population and s is the corresponding standard deviation of the reference population. A Z-score less than minus two classifies the respective child as stunted and a value less than minus three indicates severe chronic undernutrition.

7.2 Data Description

The data set on childhood malnutrition consists of 4421 observations[1], each with information on the dependent variable (in the following regression models) in form of the Z-score (7.1), and data on the situation of the child (gender, duration of breastfeeding and age) as well as on the mother's age, height, body mass index (BMI)[2], educational status and work. Moreover, the residential district of the family is available. As Kandala et al. (2001) have shown, some of these determinants have a non-linear influence on the chronic undernutrition of children. An overview of the explanatory variables and their coding can be found in the supplementary material in Appendix D.

For the investigation of the behavior of the AICs from Section 5.1, a subsample of 1600 observations was randomly chosen from the data set.[3]
In the subsample, 764 of the children were male and 836 female, with an average age of 27.29 months. The mean age of the mothers at birth was 26.50 years. For a total of 385 children, the duration of breastfeeding was less than a month (of which 11 children were of age less than a month). The average duration of breastfeeding was 11.03 months. Less than half (901) of the mothers stated to be employed and most of the mothers (1002) went to primary school but not to elementary school or higher.

[1] The entire data set is larger (6299 obs.), here only complete cases are taken into account.

[2] The body mass index is based on an individual's height and the weight and calculated as the weight in kg divided by the square of height in meters.

[3] Note that this is the same subsample as in Greven and Kneib (2010).

7.3 Univariate Smoothing Models

Generally, the aim is to determine a regression model that – with the covariates available – best approximates the true underlying data generation mechanism. Here, the analysis was restricted to univariate modeling as it sufficed to investigate the behavior of the Akaike information criteria and enabled to take the computational expensive measures of Efron (2004) and Liang et al. (2008) into account.

Two univariate smoothing models were analyzed, the first regarding the influence of the age of the child in months (`cage`) on the Z-score (7.1) and the other that of the determinant `mage` (age of the mother at birth in years). The models were estimated based on the representation as linear mixed models followed by the computation of the respective marginal AIC and the conditional AICs as in Subsection 5.1.2.

We aimed to answer the question whether the respective explanatory variable had a non-linear effect on the dependent variable (the Z-score) or not – corresponding to the selection of random effects. This was assessed by comparing the AICs of the univariate smoothing models to the AICs of the respective linear models, similar to the simulation studies in Chapter 6.

The non-linear models were estimated by using the function `gamm` of the `R`-package `mgcv` (see Appendix E.1.2) and the linear models with the function `lm` of the `basic` package. In analogy to the first simulation study in 6.1, we used cubic B-splines with ten inner knots and a second order difference penalty – penalizing the deviations from the linear model – to specify the non-parametric effects.
Note that for the further analysis, the Z-score (7.1) was centered and standardized. Moreover, prior to the model estimations, an auxiliary linear mixed model was fitted to the data in order to obtain the fixed effects after re-parametrization. For the extraction of the fixed and random effects, the function `extract.lmeDesign` was again used.[4]

The explicit choice of the two covariates `cage` and `mage` was made in order to illustrate two different situations. One where the influence was clearly non-linear (`cage`), and the other where not all criteria led to the same decision (`mage`) as will be shown in the following.
The estimated linear and non-linear effects obtained by ML and REML estimation for the two covariates are shown in Figures 7.1 and 7.2. One can see that under ML as well as under REML estimation, a clearly non-linear curve was estimated for the covariate `cage`, whereas for the variable `mage` the curves – especially in the maximum likelihood case – were much closer to the linear estimation.

In order to answer the question on the need for non-linear modeling for this data, we used the same Akaike information criteria as in the simulation studies in Chapter 6. For the conditional version of the covariance based penalty term, 200 bootstrap replications were used. As we found in the simulation studies of the previous chapter that for the joint version the number of bootstrap replications needed to be increased with sample

[4] Compare the simulation study in Section 6.1.

size, the calculations were based on 2000 bootstrap replications for the joint measure. The disturbance in the computation of the approximate AIC by Liang et al. (2008) was – in analogy to the simulation studies – chosen as $h = 0.0001$. All AICs of the non-linear model were then compared to the Akaike information criterion of the simpler (linear) model. The calculations run approximately 2.2 hours.

The results in Table 7.1 and in Table D.2 in Appendix D show that under ML as well as under REML estimation, all criteria for the complex model ($m2$) indicated that the age of the child (`cage`) had a non-linear effect on the Z-score because they were all smaller than the associated AIC of the linear model $m1$. Under either estimation method, the smallest Akaike information criterion was given by the conditional covariance based measure with a constant error variance based on Efron (2004). In accordance with the theoretical findings of Vaida and Blanchard (2005) and Greven and Kneib (2010), the criterion which was closest to the AIC of the linear model, under both estimation methods, was the marginal AIC which tended to make a choice in favor of the simpler model. As in the simulation studies, one could see that the function `logLik.gamm` of the package `mgcv` automatically returns the marginal AIC. The results also showed that the AIC of Yu and Yau (2011) was equivalent to the analytic AIC in the case of maximum likelihood estimation, but – as it has been constructed only under ML estimation – it had a greater value than the analytic measure under REML. For the approximate cAIC, the same values were obtained as for its analytic version.

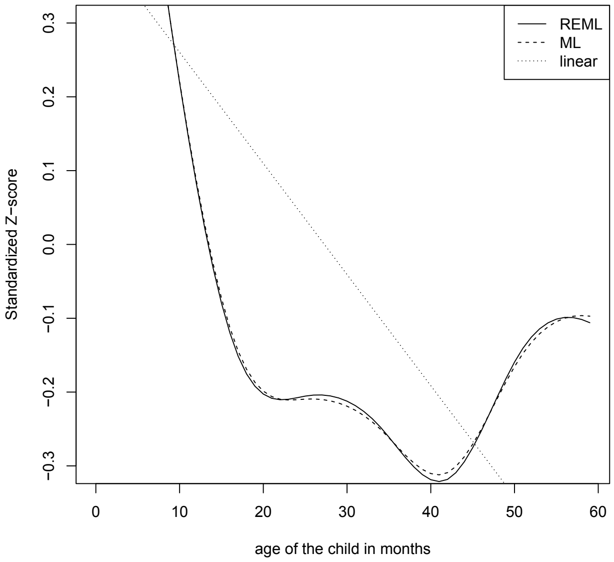

Figure 7.1: *Estimated linear and non-linear effects obtained by ML and REML for covariate* `cage`

name of AIC	ML estimation	REML estimation
`AIC_m1`	4434.04	4434.04
`AICconvent_m2`	4315.16	4314.77
`AICapprox_m2_h1e − 04`	4316.39	4316.10
`AICanalyt_m2`	4316.39	4316.10
`AICcov_m2_cond_Boot200`	**4315.15**	**4313.95**
`AICcov_m2_cond_sig_in_B_Boot200`	4315.21	4313.99
`AICcov_m2_joint_Boot2000`	4316.44	4314.80
`AICcov_m2_joint_sig_in_B_Boot2000`	4316.44	4314.81
`AICyuyau_tausq_in_num_m2`	4316.39	4316.55
`AICmgcv_m2`	4327.29	4333.59
`mAIC_m2`	4327.29	4333.59

Table 7.1: *cAICs and mAIC for linear (m_1) and non-linear (m_2) modeling of univariate continuous covariate effects of covariate* `cage`*. For both ML and REML, the smallest AIC is marked in bold.*

For the variable `mage`, the situation was rather different and not all criteria led to the same decision (see Table 7.2 and Table D.3 in Appendix D). Under both estimation methods, the conventional cAIC was the smallest and lay below the AIC of the linear model. This corresponds to the theoretical findings of Greven and Kneib (2010) who showed that ignoring the uncertainty in the random effects variance (as is the case for the conventional cAIC) leads to the selection of the more complex model, unless $\hat{\tau}^2 = 0$ (compare 5.1.2). In addition, the two variants of the joint covariance based cAICs led to the selection of the complex model under ML, whereas under REML the two variants of the conditional analogue were smaller than `AIC_m1`.
It should be remarked that a greater number of replications for the covariance based measure might have been necessary as the sample size was comparatively large (compared to the maximum sample size of $n = 200$ in the simulations studies in Chapter 6). There was evidence that a replication number of $B = 1000$ was not sufficiently large for the joint measure as this led to a different decision as the actual choice for covariate *mage*.
All other criteria (marked with a (*) in Table 7.2) decided in favor of the linear model under either estimation method. Under REML estimation, the degrees of freedom of Yu and Yau (2011) were again greater than the corresponding analytic degrees.
Note that, according to the check for zero variance based on the maximized log-likelihood difference (6.1), the random effects variance was not estimated to be zero under either method. Thus, the consideration of the additional implementation including the check for zero variance of the covariance based degrees of freedom would have given no additional insight.

No convergence errors occurred in the computations, neither in the initial calculation of the non-linear models for the influence of `cage` or `mage`, nor within the computation of the cAICs. Furthermore, no non-invertible matrices appeared for which the associated measure would have been set to ‘`NA`’.

Finally, it should be is pointed out that, as expected, the random effects variance for either covariate was estimated to be larger under REML estimation than under ML estimation. Also, the maximized log-likelihoods were greater under REML.

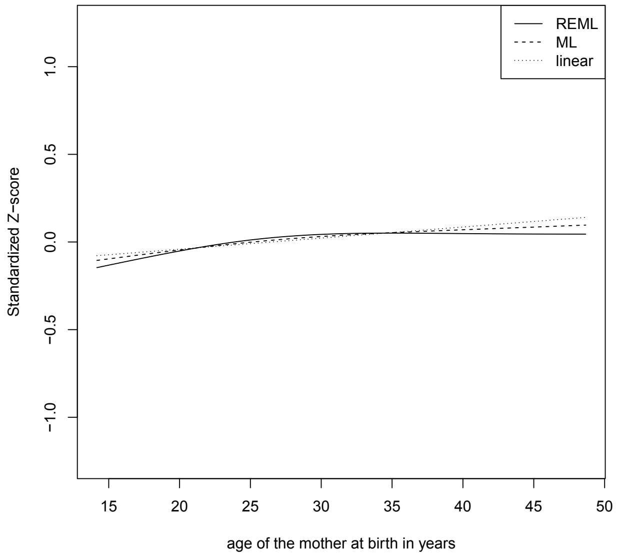

Figure 7.2: *Estimated linear and non-linear effects obtained by ML and REML for covariate* `mage`

name of AIC	ML estimation	REML estimation
AIC_m1	4542.58	4542.58
AICconvent_m2	**4541.96**	**4541.69**
AICapprox_m2_h1e − 04	4546.85*	4543.30*
AICanalyt_m2	4546.85*	4543.30*
AICcov_m2_cond_Boot200	4542.72*	4542.30
AICcov_m2_cond_sig_in_B_Boot200	4542.73*	4542.34
AICcov_m2_joint_Boot2000	4542.53	4542.66*
AICcov_m2_joint_sig_in_B_Boot2000	4542.55	4542.68*
AICyuyau_tausq_in_num_m2	4546.85*	4547.11*
AICmgcv_m2	4544.54*	4551.19*
mAIC_m2	4544.54*	4551.19*

Table 7.2: *cAICs and mAIC for linear (m_1) and non-linear (m_2) modeling of univariate continuous covariate effects of covariate* `mage`*. Under both ML and REML, the smallest AIC is marked in bold and those which are greater than the AIC of the linear model are emphasized with a star (*).*

It should be mentioned that although some of the children in the data set had the same mother, no additional random effects for the mothers were considered for several reasons. First, this would have become computationally very expensive as more than a thousand person-specific random effects would have to be included and it could have led to computation problems. Second, the number of mothers with several children in the study is relatively small and third, the results should be comparable to the results of Greven and Kneib (2010) who proceeded in the same way.

Chapter 8

Further Considerations

In the following, some considerations on extensions of our simulation studies (in Chapter 6) as well as theoretical aspects will be presented, ranging from general extensions to enhancements of specific cAICs. In particular, different modifications for the covariance based cAIC will be given.

A very interesting and crucial next step would be to conduct a similar simulation study for the generalized case, i.e. for GLMMs, where distributions beyond the Gaussian one are considered. This would permit to evaluate the behavior of the different criteria in this more flexible and more complex situation. It seems possible that the analysis in GLMMs would actually lead to changes in the results, especially concerning the cAIC of Yu and Yau (2011) (5.67). In our simulation studies we found that the criterion of Yu and Yau was almost equal to the analytic cAIC under maximum likelihood estimation. This might change in the generalized case if the asymptotic does not behave like it does for the case of LMMs.

So far, two cAICs allow the selection of random effects in GLMMs: The cAIC based on the covariance penalty of Efron (2004) ((5.46) and (5.48)) and the cAIC of Yu and Yau. In order to compare more measures in the generalized case than these two, a next step could be to apply those without generalized forms to the working model. A long term objective is clearly to find an analytical formulation for the generalized case.

Note that for most distributions of the exponential family, such as a Bernoulli or a Poisson distribution, the distinction between a known and an unknown dispersion parameter ceases as ϕ is a constant, i.e. $\phi = 1$ (see Table 3.1 in Subsection 3.2.1). Nevertheless, simulation studies for GLMMs are (technically) more demanding, since the marginal distribution is inaccessible, which is why approximations have to be used. Note that the results depend on the type of approximation. As the function `gamm` of the `R`-package `mgcv` does not permit to specify REML estimation in the generalized case (see Subsection 6.1.3), it would be advisable to use the function `gam` {`mgcv`} for the estimation of the penalized spline models. Some functions which can be used for the estimation in generalized random intercept models have been described in Subsection 3.2.6. The associated simpler models would then be GLMs instead of LMs and could be estimated by using the function `glm` of the `basic` package in `R`.

Except for simulations for GLMMs, another future objective could be the extension of the cAIC of Yu and Yau (2011) to restricted maximum likelihood estimation as well as to others than the canonical link function.[1]

For the Gaussian case, one could think moreover of an extension to more general covariance matrices $\boldsymbol{R}$ of the error terms, which were considered here $\boldsymbol{R} = \sigma^2 \boldsymbol{I}_n$.
Due to occasional failure of the numerical computation of the approximate degrees of freedom of Liang et al. (2008) (5.14), it is worth thinking about including the check for zero variance (6.1) also for this measure, which would additionally speed up the computation.

In this work we concentrated on the selection of one random effect. It could be interesting to extend this analysis to more than one random effect. The inspection of the effect of the presence of random effects on the selection of fixed effects could also be subject of interest for future analyses (cf. Greven and Kneib (2010)).

Finally, it could also be interesting to consider the topic in the Bayesian framework.

In the following, some modifications for the covariance based measure ((5.46) and (5.48)) will be considered.
First, Greven (2011b) showed that the second term of the bias correction (5.17) (underlined in equation (8.2)) in the case of **unknown error variance $\boldsymbol{\sigma^2}$**,

$$BC = cAI - E_{g(y,b)}\left[-2\ log\left(f(y|\hat{\beta}(y), \hat{b}(y), \hat{\sigma}^2(y))\right)\right] \tag{8.1}$$

$$= 2\ E_{g(y,b)}\left[\sum_{i=1}^{n}(y_i - \mu_i)\frac{\hat{\mu}_i}{\hat{\sigma}^2}\right] + \underline{2\ E_{g(y,b)}\left[\sum_{i=1}^{n} c(y_i, \hat{\sigma}^2) - E_{g(y^*|b)}\left[c(y_i, \hat{\sigma}^2)\right]\right]}, \tag{8.2}$$

with y^* distributed as y, does not cancel out. Hence, this term needs to be taken into account. In the case of Gaussianity and the canonical link function, one obtains[2]

$$c(y_i, \sigma^2) = -\frac{y_i^2}{2\sigma^2} - \frac{1}{2}\ log\left(2\pi\sigma^2\right). \tag{8.3}$$

Applying the computational formula for the variance (Steiner (1796 - 1863)), one obtains for the second moment of $y_i^* \sim \mathcal{N}(\mu, \sigma^2)$

$$E(y_i^{*2}) = Var(y_i^*) + [E(y_i^*)]^2 = \sigma^2 + \mu^2.$$

Thus, the bias correction becomes

$$BC = 2\ E_{g(y,b)}\left[\sum_{i=1}^{n}(y_i - \mu_i)\frac{\hat{\mu}_i}{\hat{\sigma}^2}\right] + \sum_{i=1}^{n} E_{g(y,b)}\left[\frac{\sigma^2 + \mu^2 - y_i^2}{\hat{\sigma}^2}\right]. \tag{8.4}$$

Since $\hat{\sigma}^2$ depends on the response y_i, it cannot be pulled out of the expectation with respect to $g(y, b)$ and the term is not exactly zero. Greven suggested approximating this

[1] This would become important e.g. in the case of an exponential distribution where the canonical link function is inadequate because it does not guarantee that the mean is non-negative (see Tutz (1011)).

[2] Greven (2011b)

expectation – in analogy to the previous proceeding – by using a bootstrap. To this end, the error variance σ^2 and the mean μ_i are fixed at the estimated quantities and $\hat{\sigma}^2$ is re-estimated in each bootstrap sample. One obtains the following approximation of the second term in (8.4):

$$\frac{1}{B}\sum_{\xi=1}^{B}\frac{n\hat{\sigma}^2 + n\hat{\mu}^2 - \left(y^{\xi}\right)^T y^{\xi}}{\left(\hat{\sigma}^2\right)^{\xi}}, \tag{8.5}$$

with $\left(\hat{\sigma}^2\right)^{\xi}$ denoting the estimated error variance in bootstrap sample ξ ($\xi = 1, \ldots, B$).

A second modification in the computation of the joint covariance based measure should be studied more closely. Note that this alternative proceeding is computationally very expensive, which is why it has not been treated in detail within the scope of this work. The analysis of this modification seems very interesting, especially as – unexpectedly – the re-estimation of the error variances (instead of using the constant variance) highly affected the results (see Subsections 6.1.4 and 6.2.4). One can therefore expect a similar impact on the outcome, which is why the modification should be considered in future simulations. The outline of this approach will be given in the following.

As discussed in Subsection 5.1.2, the difference $y^{*\xi} - y^{*\cdot}$ ($\xi = 1, \ldots, B$) does not estimate $\boldsymbol{X}\beta + \boldsymbol{Z}b$ in the joint case. Thus, Greven (2011b) suggested to replace the difference $(y^{*\xi} - y^{*\cdot})$ by $\varepsilon^{*\xi} = y^{*\xi} - \boldsymbol{X}\hat{\beta} - \boldsymbol{Z}b^{*\xi}$. The alternative idea[3] is to overcome this problem by drawing a number ($B1$) of random effects $b^{*\xi}$ as

$$b_i^{*\xi} \sim \mathcal{N}(0, \hat{\tau}^2),\ i = 1, \ldots, n,\ \xi = 1, \ldots, B1, \tag{8.6}$$

and for each of the random effects a number ($B2$) of error terms

$$\varepsilon_i^{*\xi k} \sim \mathcal{N}(0, \hat{\sigma}^2),\ i = 1, \ldots, n,\ \xi = 1, \ldots, B1,\ k = 1, \ldots, B2. \tag{8.7}$$

Then, for each error term, the associated response $y_i^{*\xi k}$ is determined as

$$y_i^{*\xi k} = \boldsymbol{X}_i\hat{\beta} + \boldsymbol{Z}_i b_i^{*\xi} + \varepsilon_i^{*\xi k},\ i = 1, \ldots, n, \xi = 1, \ldots, B1,\ k = 1, \ldots, B2. \tag{8.8}$$

For each response variable, the (non-linear) model is fitted, yielding an estimator for the linear predictor and the error variance. Note that as for the other variants, one can either use the constant error variance or the specific variances of each bootstrap replication.[4] In a next step, the random effects specific means are determined as

$$\begin{aligned} y^{*\xi\cdot} &= \frac{1}{B2}\sum_{k=1}^{B2} y^{*\xi k} && (8.9)\\ &= \frac{1}{B2}\sum_{k=1}^{B2} \boldsymbol{X}\hat{\beta} + \boldsymbol{Z}b^{*\xi} + \varepsilon^{*\xi k} && (8.10)\\ &= \boldsymbol{X}\hat{\beta} + \boldsymbol{Z}b^{*\xi} + \underbrace{\frac{1}{B2}\sum_{k=1}^{B2} \varepsilon^{*\xi k}}_{\xrightarrow{B2\to\infty} 0}. && (8.11) \end{aligned}$$

[3] Greven (2011b)

[4] As mentioned above, the second term of the BC should be included additionally when assuming unknown error variance.

Hence, for a large number of errors drawn per random effect, $B2$, the $b^{*\xi}$ specific means average to $\boldsymbol{X}\hat{\beta} + \boldsymbol{Z}b^{*\xi}$. The random effects specific means are then used for the construction of the estimator instead of $y_i^{*\cdot}$ as before. Finally, the approximation of the first term of the BC becomes

$$\sum_{i=1}^{n}\sum_{\xi=1}^{B1}\frac{1}{B2-1}\sum_{k=1}^{B2}(y_i^{*\xi k} - y_i^{*\xi\cdot})\frac{\hat{\eta}_i^{*\xi k}}{(\hat{\sigma}^2)^{*\xi k}}\ . \tag{8.12}$$

For the algorithm see Appendix B. Note that for an unknown error variance, the second term of the bias correction (see (8.1)) should be additionally taken into account, as described above, as the following modification only effects the first term of the BC.

The comparison of the results of the two simulation studies (6.1 and 6.2) showed that the covariance based AIC did not perform as well for the smoothing splines as for the random intercept models. One explanation is that in the former additional inaccuracy was introduced by drawing from a Gaussian distribution which is only an auxiliary construction (see Section 6.3). Therefore, another possibility to modify the computation of the covariance based degrees of freedom would be to refrain from assuming Gaussian distribution by using non-parametric bootstrap methods. Asymptotically, the two bootstrap methods should be equivalent, but they can differ for finite sample size. Note that non-parametric bootstrap could be inappropriate for small sample sizes.

Furthermore, it would be recommendable to also estimate the linear model (m_1) with the bootstrap methods used for the computation of the non-linear model (m_2) as this would allow to better compare the models due to more similar variability. It would furthermore make possible to better understand the behavior of the criteria.

In summary, we presented various extensions to our simulations. The most important next step would be to try out various modifications for the covariance based cAIC and to apply the same bootstrap methods to the linear model m_1 for a better comparison. The resultant criteria could then be applied in a simulation study for generalized linear mixed models in which they would be compared to the cAIC of Yu and Yau and (possibly) to the other criteria which can be applied to the working model.

Chapter 9

Conclusion

In this thesis, we considered model selection via Akaike information criteria in mixed models. The focus lay in particular on the selection of random effects. We concentrated on estimators of the conditional Akaike information (cAI), which take the estimation uncertainty in the random effects into account. So far, the behavior of an approximate corrected conditional Akaike information criterion (5.14) and its analytic analogue (5.23) have been studied in simulation studies for linear mixed models by Greven and Kneib (2010).

The objective of this thesis was to investigate the behavior of two additional corrected conditional Akaike information criteria (cAIC) for which a generalization beyond the Gaussian distribution is available: The cAIC of Yu and Yau (2011) (5.67) and the cAIC based on a covariance penalty of Efron (2004) ((5.46) and (5.48)). Using simulations, we draw a comparison between these two measures and the approximate, the analytic and the uncorrected cAIC (5.10) in order to determine whether the covariance based cAIC and the recently suggested cAIC of Yu and Yau are appropriate alternatives to the analytic cAIC in the special case of LMMs. Applying their generalized forms would then be a way to perform model selection in GLMMs as long as no analytic version has been derived.

Furthermore, we demonstrated two methods to compute the covariance based cAIC, and we examined which method is more adequate for the selection of random effects in mixed models. In this context, we also studied the influence of the error variance on the results. In addition to the performance of the various cAICs, numerical and implementational aspects were included in the decision which of the newly considered cAICs is most promising to serve as an adequate model selection criterion in generalized linear mixed models.

We conducted two simulation studies to examine the behavior of the measures in two different situations. In the first, the linear mixed model served as an inferential tool in the estimation for penalized spline smoothing. The second simulation study used random intercept models.

The results of both simulation studies mainly agreed. However, we discovered that the results of the simulation based on penalized splines smoothing were more sensitive to numerical imprecisions and that the preference for either the joint or the conditional version of the covariance based cAIC was here not as distinct as for the simulation based on random intercept models. This can be ascribed to the more complex correlation structure for penalized splines compared to random intercept models. Another reason is that approximations were made for penalized spline smoothing and that the mixed model was only an inferential tool, but did not reflect the true underlying structure.

The simulations showed that the cAIC of Yu and Yau is almost identical (in our settings) to the analytic cAIC under ML estimation. However, under REML estimation the cAIC of Yu and Yau turned out to favor the simpler model. In addition, extremely large and even negative degrees of freedom arose under REML estimation. Moreover, we had to deal with several numerical problems in the implementation of this measure. The computational costs for the cAIC of Yu and Yau, however, were comparably low (compared to the approximate and the covariance based cAIC). It should be noted that it might possibly perform worse in the case of GLMMs, if the asymptotic does not behave like it did for LMMs.
Finally, we found that the version of the covariance based cAIC with redrawn random effects and re-estimated error variance for each bootstrap sample performed better than all other alternatives which were considered. In many settings, the measure showed a behavior relatively similar to that of the analytic cAIC. For large sample sizes, however, it turned out to favor the more complex model and to differ from the analytic measure. Further modifications are needed for the case of re-estimated error variances (see for details Chapter 8). Computationally, the covariance based measure was very expensive, as it turned out that many bootstrap replications were needed to obtain a reliable estimator. For practical use, it is thus essential to review our implementation.

In summary, we showed that the cAIC of Yu and Yau and the covariance based cAIC are both promising approaches for the selection of random effects in generalized linear mixed models, although further considerations are needed for both criteria. Compared to the marginal and the uncorrected conditional AIC, which clearly favor the simpler or the more complex model, respectively, the cAIC of Yu and Yau and the covariance based cAIC are bias corrected AICs which led in many situations to the same decisions as the corrected analytic cAIC.

Appendix A

Proofs and Derivations

Proof 1. Minimization of $E_y\left[KLD(g,\hat{f})\right]$ is equivalent to maximization of $\{constant - T\}$[1]:

$$\begin{aligned}
E_y\left[KLD(g,\hat{f}(z))\right] &= \int_{\mathbb{R}} KLD(g,\hat{f}(z))g(y)\ dy \\
&= \int_{\mathbb{R}}\left[\int_{\mathbb{R}} log\left\{\frac{g(z)}{\hat{f}(z)}\right\} g(z)\ dz\right] g(y)\ dy \\
&= \int_{\mathbb{R}}\left[\int_{\mathbb{R}} log\left(g(z)\right)\ g(z)\ dz - \int_{\mathbb{R}} log\left(\hat{f}(z)\right) g(z)dz\right] g(y)\ dy \\
&= \int_{\mathbb{R}} log\left(g(z)\right) g(z)\ dz - \int_{\mathbb{R}}\left[\int_{\mathbb{R}} log\left(\hat{f}(z)\right) g(z)\ dz\right] g(y)\ dy \\
&= constant - E_y\left[\int_{\mathbb{R}} log\left(\hat{f}(z)\right) g(z)\ dz\right] \\
&= constant - E_y\left[E_z\left[log\left(\hat{f}(z)\right)\right]\right], \\
&= constant - T,
\end{aligned}$$

where $\hat{f}(z)$ denotes $f(z|\hat{\psi}(y))$. Thus, minimizing $E_y\left[KLD(g,\hat{f}(z))\right]$ is equivalent to maximizing $\{constant - T\}$. □

[1] Heumann et al. (2010)

Proof 2. Conversion of the conditional LMM into the marginal LMM[2]:

$$
\begin{aligned}
f(y) &= \int f(y|b)f(b)db = \int f(y,b)db \\
&\propto \int exp\left\{-\frac{1}{2}\,(y-\boldsymbol{X}\beta-\boldsymbol{Z}b)^T\boldsymbol{R}^{-1}(y-\boldsymbol{X}\beta-\boldsymbol{Z}b)-\frac{1}{2}\,b^T\boldsymbol{G}^{-1}b\right\}db \\
&= \int \exp\left\{-\frac{1}{2}\left[(y-\boldsymbol{X}\beta)^T\boldsymbol{R}^{-1}(y-\boldsymbol{X}\beta)-2\,(y-\boldsymbol{X}\beta)^T\boldsymbol{R}^{-1}\boldsymbol{Z}b+b^T\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}b+b^T\boldsymbol{G}^{-1}b\right]\right\}db \\
&= \int exp\left\{-\frac{1}{2}\left[\begin{pmatrix} y-\boldsymbol{X}\beta \\ b\end{pmatrix}^T\begin{pmatrix}\boldsymbol{R}^{-1} & -\boldsymbol{R}^{-1}\boldsymbol{Z} \\ -\boldsymbol{Z}\boldsymbol{R}^{-1} & \boldsymbol{G}^{-1}+\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\end{pmatrix}\begin{pmatrix} y-\boldsymbol{X}\beta \\ b\end{pmatrix}\right]\right\}db \\
&\overset{\text{Schur compl.}}{=} \int exp\left\{-\frac{1}{2}\left[\begin{pmatrix} y-\boldsymbol{X}\beta \\ b\end{pmatrix}^T\begin{pmatrix}\boldsymbol{V} & \boldsymbol{Z}\boldsymbol{G} \\ \boldsymbol{G}\boldsymbol{Z}^T & \boldsymbol{G}\end{pmatrix}^{-1}\begin{pmatrix} y-\boldsymbol{X}\beta \\ b\end{pmatrix}\right]\right\}db, \\
&\text{with } \boldsymbol{V}=\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T+\boldsymbol{R}.
\end{aligned}
$$

Thus, the density belongs to the Gaussian distribution

$$
\begin{pmatrix} y \\ b\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{X}\beta \\ 0\end{pmatrix},\begin{pmatrix}\boldsymbol{V} & \boldsymbol{Z}\boldsymbol{G} \\ \boldsymbol{G}\boldsymbol{Z}^T & \boldsymbol{G}\end{pmatrix}\right).
$$

□

Derivation 1. Derivation of Henderson's mixed model equations[3]:

Consider the penalized generalized least-squares criterion (3.23). It can be re-formulated as

$$
\begin{aligned}
GLS_{pen}(\beta,b) &= (y-\boldsymbol{X}\beta-\boldsymbol{Z}b)^T\boldsymbol{R}^{-1}(y-\boldsymbol{X}\beta-\boldsymbol{Z}b)+b^T\boldsymbol{G}^{-1}b \\
&= (y-\boldsymbol{X}\beta)^T\boldsymbol{R}^{-1}(y-\boldsymbol{X}\beta)-2b^T\boldsymbol{Z}^T\boldsymbol{R}^{-1}(y-\boldsymbol{X}\beta)+b^T\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}b+b^T\boldsymbol{G}^{-1}b \\
&= y^T\boldsymbol{R}^{-1}y-2\beta^T\boldsymbol{X}^T\boldsymbol{R}^{-1}y+\beta^T\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}\beta-2b^T\boldsymbol{Z}^T\boldsymbol{R}^{-1}y+2b^T\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}\beta \\
&\quad + b^T\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}b+b^T\boldsymbol{G}^{-1}b
\end{aligned}
$$

The first derivative yields

$$
\begin{aligned}
\frac{\partial}{\partial\beta}\,GLS_{pen}(\beta,b) &= -2\boldsymbol{X}^T\boldsymbol{R}^{-1}y+2\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}\beta+2b^T\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} \\
\frac{\partial}{\partial b}\,GLS_{pen}(\beta,b) &= -2\boldsymbol{Z}^T\boldsymbol{R}^{-1}y+2\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}\beta+2\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}b+2\boldsymbol{G}^{-1}b.
\end{aligned}
$$

[2] Konrath (2009)

[3] Konrath (2009)

The result is set to zero resulting in

$$\begin{aligned}
0 &\overset{!}{=} -2\boldsymbol{X}^T\boldsymbol{R}^{-1}y + 2\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}\beta + 2b^T\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}\\
&\Leftrightarrow \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}\hat{\beta} + \hat{b}^T\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{R}^{-1}y\\
&\Leftrightarrow \left(\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}, \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\right)\begin{pmatrix}\hat{\beta}\\ \hat{b}\end{pmatrix} = \boldsymbol{X}^T\boldsymbol{R}^{-1}y
\end{aligned}$$

and for the random effects vector

$$\begin{aligned}
0 &\overset{!}{=} -2\boldsymbol{Z}^T\boldsymbol{R}^{-1}y + 2\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}\beta + 2\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}b + 2\boldsymbol{G}^{-1}b\\
&\Leftrightarrow \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}\hat{\beta} + (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})\hat{b} = \boldsymbol{Z}^T\boldsymbol{R}^{-1}y\\
&\Leftrightarrow (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}, \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})\begin{pmatrix}\hat{\beta}\\ \hat{b}\end{pmatrix} = \boldsymbol{Z}^T\boldsymbol{R}^{-1}y
\end{aligned}$$

Altogether, one obtains Henderson's mixed model equations

$$\begin{pmatrix}\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}\end{pmatrix}\begin{pmatrix}\hat{\beta}\\ \hat{b}\end{pmatrix} = \begin{pmatrix}\boldsymbol{X}^T\boldsymbol{R}^{-1}y\\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}y.\end{pmatrix}$$

□

Derivation 2. Derivation of the hat matrix $\boldsymbol{H}_1$ in the LMM:

Consider the LMM (3.1.3) with $\boldsymbol{R} = \sigma^2\boldsymbol{I}_n$. Alternatively, it can be displayed in the form

$$y = \boldsymbol{B}\delta + \varepsilon,$$

where

$$\delta = (\beta^T, b^T)^T \text{ and } \boldsymbol{M} = [\boldsymbol{X}, \boldsymbol{Z}],$$

$$f(\delta) \propto exp\left\{-\frac{1}{2\tau^2}\delta^T\boldsymbol{K}\delta\right\}$$

with

$$\boldsymbol{K} = \begin{pmatrix}0 & & 1 & & \\ & \ddots & & \ddots & \\ & & 0 & & 1\end{pmatrix},$$

the number of zeros corresponding to the dimension of β and the number of ones to the dimension of b. The estimation therefore yields

$$\hat{\delta} = (\boldsymbol{M}^T\boldsymbol{M} + \lambda^{-1}\boldsymbol{K})^{-1}\boldsymbol{M}^Ty.$$

For y, it follows that

$$\hat{y} = \boldsymbol{M}(\boldsymbol{M}^T\boldsymbol{M} + \lambda^{-1}\boldsymbol{K})^{-1}\boldsymbol{M}^T y.$$

Thus the matrix that maps the observed data vector y into the fitted vector $\hat{y}$, is

$$\boldsymbol{H}_1 = \boldsymbol{M}(\boldsymbol{M}^T\boldsymbol{M} + \lambda^{-1}\boldsymbol{K})^{-1}\boldsymbol{M}^T.$$

For the derivation in the more general setting and further information, see Vaida and Blanchard (2005) and Hodges and Sargent (2001).

Proof 3. Optimism Theorem of Efron[4]:

Recall the definitions of Section 5.1.2. The true predictive error can be written as

$$Err_i = err_i + O_i,$$

i.e. as a sum of the apparent error and the optimism O_i. This directly gives equation (5.42). By definition of $Q(y, \hat{\mu})$, one can calculate

$$\begin{aligned} Err_i &= q(\hat{\mu}_i) + \dot{q}(\hat{\mu}_i)(\mu_i - \hat{\mu}_i) - E\left\{q(y_i^0)\right\} \text{ and} \\ err_i &= q(\hat{\mu}_i) + \dot{q}(\hat{\mu}_i)(y_i - \hat{\mu}_i) - q(y_i). \end{aligned}$$

This results in

$$\begin{aligned} O_i &= Err_i - err_i \\ &= \dot{q}(\hat{\mu}_i)(\mu_i - y_i) - E\left\{q(y_i^0)\right\} + q(y_i) \\ &= 2\hat{\lambda}_i(y_u - \mu_i) - E\left\{q(y_i^0)\right\} + q(y_i). \end{aligned} \tag{A.1}$$

Due to the fact that y^0 is independently drawn from the same mechanism as y, taking expectations in (A.1) yields

$$\begin{aligned} E(O_i) = \Omega_i &= E\left[2\hat{\lambda}_i(y_u - \mu_i) - E\left[q(y_i^0)\right] + q(y_i)\right] \\ &= E\left[2\hat{\lambda}_i(y_i - \mu_i)\right] - E\left[E\left[q(y_i^0)\right]\right] + E\left[q(y_i)\right] \end{aligned}$$

which is equal to 2 $Cov(\hat{\lambda}_i, y_i)$. □

[4] Efron (2004)

Derivation 3. Derivation of the matrix $\boldsymbol{H}_{\tau^2\tau^2}$ for the cAIC of Yu and Yau:

$$\begin{aligned}\boldsymbol{H}_{\tau^2\tau^2} &= -\frac{\partial^2 h_a}{\partial\tau^2\partial\tau^2}\\ &= -\frac{\partial^2}{\partial\tau^2\partial\tau^2}\left\{-\frac{1}{2}(log\left\{det(H_{22})\right)\} - \frac{\nu}{2}log\left(\tau^2\right) - \frac{1}{2\tau^2}b^Tb\right\}\end{aligned}$$

with the rule for derivation of $log(det)$

$$= -\frac{\partial}{\partial\tau^2}\left\{-\frac{1}{2}tr\left\{(\frac{1}{\sigma^2}\boldsymbol{Z}^T\boldsymbol{Z}+\frac{1}{\tau^2}\boldsymbol{I}_\nu)^{-1}(-\frac{1}{\tau^4})\right\} - \frac{\nu}{2\tau^2} + \frac{1}{2\tau^4}b^Tb\right\}$$

switching trace and derivation yields

$$= \frac{1}{2}tr\left\{\frac{\partial^2}{\partial\tau^2}\left[(\frac{1}{\sigma^2}\boldsymbol{Z}^T\boldsymbol{Z}+\frac{1}{\tau^2}\boldsymbol{I}_\nu)^{-1}(-\frac{1}{\tau^4})\right]\right\} - \frac{\nu}{2\tau^4} + \frac{1}{\tau^6}b^Tb$$

applying the product and the chain rule of derivative gives

$$\begin{aligned}&= \frac{1}{2}tr\left\{-\frac{\sigma^4}{\tau^8}(\boldsymbol{Z}^T\boldsymbol{Z}+\frac{\sigma^2}{\tau^2}\boldsymbol{I}_\nu)^{-2} + 2\frac{\sigma^2}{\tau^6}(\boldsymbol{Z}^T\boldsymbol{Z}+\frac{\sigma^2}{\tau^2}\boldsymbol{I}_\nu)^{-1}\right\}\\ &\quad - \frac{\nu}{2\tau^4} + \frac{1}{\tau^6}b^Tb\end{aligned}$$

with $\boldsymbol{Z}^T\boldsymbol{Z} = \frac{\sigma^2}{\tau^2}(\boldsymbol{I}_\nu + \frac{\tau^2}{\sigma^2}\boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{I}_\nu)$ and $tr\left\{\boldsymbol{I}_\nu\right\} = \nu$ it follows

$$= \frac{1}{\tau^6}b^Tb - \frac{1}{2\sigma^4}tr\left\{\left[(\boldsymbol{I}_\nu + \frac{\tau^2}{\sigma^2}\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{Z}\right]^2\right\}.$$

Proof 4. Formulation of the penalty of Yu and Yau in dependence of the conventional penalty term:

$$\hat{\rho}_{ml} = tr\left\{(\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}} - \boldsymbol{H}_{\tilde{\theta}\tilde{\theta}}\boldsymbol{H}_{\tau^2\tau^2}^{-1}\boldsymbol{H}_{\tau^2\tilde{\theta}})^{-1}\boldsymbol{H}^*\right\}|_{\hat{b},\hat{\tau}^2}$$

with the Woodbury formula yields

$$\begin{aligned}
&= tr\left\{\left[\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}}^{-1} + \boldsymbol{H}_{\tilde{\theta}\tilde{\theta}}^{-1}\boldsymbol{H}_{\tilde{\theta}\tau^2}(\boldsymbol{H}_{\tau^2\tau^2} - \boldsymbol{H}_{\tau^2\tilde{\theta}}\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}}^{-1}\boldsymbol{H}_{\tilde{\theta}\tau^2})^{-1}\boldsymbol{H}_{\tau^2\tilde{\theta}}\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}}\right]\boldsymbol{H}^*\right\}|_{\hat{b},\hat{\tau}^2} \\
&= \hat{\rho} + tr\left\{\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}}^{-1}\boldsymbol{H}_{\tilde{\theta}\tau^2}(\boldsymbol{H}_{\tau^2\tau^2} - \boldsymbol{H}_{\tau^2\tilde{\theta}}\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}}^{-1}\boldsymbol{H}_{\tilde{\theta}\tau^2})^{-1}\boldsymbol{H}_{\tau^2\tilde{\theta}}\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}}^{-1}\boldsymbol{H}^*\right\}|_{\hat{b},\hat{\tau}^2}
\end{aligned}$$

as τ^2 is scalar this is equal to

$$= \hat{\rho} + \frac{\boldsymbol{H}_{\tau^2\tilde{\theta}}\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}}^{-1}\boldsymbol{H}^*\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}}^{-1}\boldsymbol{H}_{\tilde{\theta}\tau^2}}{\boldsymbol{H}_{\tau^2\tau^2} - \boldsymbol{H}_{\tau^2\tilde{\theta}}\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}}^{-1}\boldsymbol{H}_{\tilde{\theta}\tau^2}}|_{\hat{\tilde{\theta}},\hat{\tau^2}}.$$

□

Derivation 4. Derivation of the formulation of the penalty of Yu and Yau with τ^2 only in the numerator[5]:

The derivation of the penalty term which for which the random effects variance does not appear in the denominator is based on equation (5.68).
Applying the BLUP

$$\begin{aligned}
\hat{b} &= \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(y - \boldsymbol{X}\hat{\beta}) \\
&= \boldsymbol{G}_*\boldsymbol{Z}\boldsymbol{V}_*^{-1}(y - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}y) \\
&= \frac{\tau^2}{\sigma^2}\boldsymbol{Z}^T\boldsymbol{A}_*y,
\end{aligned}$$

with $\boldsymbol{A}_* = \boldsymbol{V}_*^{-1} - \boldsymbol{V}_*^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_*^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}_*^{-1}$, one obtains

$$\begin{aligned}
\boldsymbol{H}_{\tilde{\theta}\tilde{\theta}} - \boldsymbol{H}_{\tilde{\theta}\tau^2}\boldsymbol{H}_{\tau^2\tau^2}^{-1}\boldsymbol{H}_{\tau^2\tilde{\theta}}\ |_{\hat{b}} &= \frac{1}{\sigma^2}\begin{pmatrix}\boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{Z} + \frac{\sigma^2}{\tau^2}\boldsymbol{I}_\nu\end{pmatrix} - \frac{1}{\tau^4}\begin{pmatrix}0 \\ \frac{\tau^2}{\sigma^2}\boldsymbol{Z}^T\boldsymbol{A}_*y\end{pmatrix} \\
&\times \left(\frac{1}{\tau^2\sigma^4}y^T\boldsymbol{A}_*\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{A}_*y - \frac{1}{2\sigma^4}tr\left\{\left[(\boldsymbol{I}_\nu + \frac{\tau^2}{\sigma^2}\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{Z}\right]^2\right\}\right)^{-1} \\
&\times \frac{1}{\tau^4}\begin{pmatrix}0 & \frac{\tau^2}{\sigma^2}y^T\boldsymbol{A}_*\boldsymbol{Z}\end{pmatrix} \\
&= \frac{1}{\sigma^2}\begin{pmatrix}\boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{Z} + \frac{1}{\tau^2}\boldsymbol{U}\end{pmatrix},
\end{aligned}$$

[5] Greven (2011b)

with

$$\boldsymbol{U} = \sigma^2 \boldsymbol{I}_\nu - \frac{\sigma^2 \boldsymbol{Z}^T \boldsymbol{A}_* y y^T \boldsymbol{A}_* \boldsymbol{Z}}{y^T \boldsymbol{A}_* \boldsymbol{Z} \boldsymbol{Z}^T \boldsymbol{A}_* y - \frac{\tau^2}{2} tr \left\{ \left[(\boldsymbol{I}_\nu + \frac{\tau^2}{\sigma^2} \boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \boldsymbol{Z} \right]^2 \right\}}.$$

Applying the inversion formula for block-matrices (with the use of the Schur complement of $\boldsymbol{Z}^T\boldsymbol{Z} + \frac{1}{\tau^2}\boldsymbol{U}$) leads to

$$\left(\frac{1}{\sigma^2} \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{Z} + \frac{1}{\tau^2}\boldsymbol{U} \end{pmatrix} \right)^{-1} =$$

$$\sigma^2 \begin{pmatrix} (\boldsymbol{X}^T\boldsymbol{X} - \tau^2\boldsymbol{T})^{-1} & -\tau^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z}(\tau^2\boldsymbol{Z}^T\boldsymbol{P}_0\boldsymbol{Z} + \boldsymbol{U})^{-1} \\ -\tau^2(\boldsymbol{U}^T\tau^2\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} - \tau^2\boldsymbol{T})^{-1} & \tau^2(\tau^2\boldsymbol{Z}^T\boldsymbol{P}_0\boldsymbol{Z} + \boldsymbol{U})^{-1} \end{pmatrix},$$

with

$$\begin{aligned} \boldsymbol{P}_0 &= \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T, \\ \boldsymbol{T} &= \boldsymbol{X}^T\boldsymbol{Z}(\tau^2\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}. \end{aligned}$$

Denoting

$$\begin{aligned} A_3 &= \boldsymbol{X}^T\boldsymbol{X} - \tau^2\boldsymbol{T} \text{ and} \\ A_4 &= (\tau^2\boldsymbol{Z}^T\boldsymbol{P}_0\boldsymbol{Z} + \boldsymbol{U}). \end{aligned}$$

results in the formula (5.74).

Appendix B

Algorithms and Bootstrap estimation

Algorithm 1. (Penalized Iteratively Reweighted Least-Squares algorithm (PIRLS))

The penalized Iteratively Reweighted least-squares algorithm is an extension of the Iteratively Reweighted least-squares algorithm used for the estimation in generalized linear models. The latter leads the estimation problem in the GLM back to an iteratively weighted least-squares problem. The parameter is estimated as a linear approximation of the (in general non-linear) score equations $S(\beta) = 0$ (compare (3.53)).
Starting from an initial value $\hat{\beta}^{(0)}$, a tangent to the score-function in $\hat{\beta}^{(0)}$ is constructed by using a first order Taylor expansion of $S(\beta)$ around $\hat{\beta}^{(0)}$

$$\begin{aligned} S(\beta) &\approx S(\hat{\beta}^{(0)}) + S'(\hat{\beta}^{(0)})(\beta - \hat{\beta}^{(0)}) && \text{(B.1)} \\ &= S(\hat{\beta}^{(0)}) - I'(\hat{\beta}^{(0)})(\beta - \hat{\beta}^{(0)}), && \text{(B.2)} \end{aligned}$$

where $I(\beta)$ denotes the Fisher information. An improved solution $\hat{\beta}^{(1)}$ is obtained as the zero of the tangent

$$\hat{\beta}^{(1)} = \hat{\beta}^{(0)} + I(\hat{\beta}^{(0)})^{-1} S(\hat{\beta}^{(0)}). \tag{B.3}$$

A further improvement, $\hat{\beta}^{(2)}$, is achieved via a linearization on the basis of $\hat{\beta}^{(1)}$. The described procedure is iteratively repeated until the solutions do not differ anymore or until a stop criterion is reached, e.g.

$$\frac{\left\| \hat{\beta}^{(k)} - \hat{\beta}^{(k+1)} \right\|}{\left\| \hat{\beta}^{(k)} \right\|} < \varepsilon \text{ (with } \varepsilon > 0), \tag{B.4}$$

where $\|\cdot\|$ denotes the Euclidean norm and ε is a given threshold (Fahrmeir et al., 2007; Scheipl, 2009).

For GLMMs, a penalized version of this method is used. Here, the aim is to predict the random effects b for given β, θ_*, and ϕ (compare (3.2.5)). First, the score-function and the Fisher information have to be specified.
The score function is given by

$$S(b) = \frac{\partial}{\partial b} log \left\{ \mathcal{L}(\beta, \theta_*, \phi, b) \right\} = \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{\Delta} (y - \mu) - \boldsymbol{G}(\theta_*)^{-1} b, \tag{B.5}$$

with the weight matrix

$$\boldsymbol{W} = diag\left(\frac{1}{\phi b''(\vartheta_i)}\left(\frac{\partial \mu_i}{\partial \eta}\right)^2\right)_{i=1,\ldots,n} \tag{B.6}$$

and

$$\boldsymbol{\Delta} = diag\left(\frac{\partial \eta_i}{\partial \mu}\right)_{i=1,\ldots,n}. \tag{B.7}$$

In matrix notation, the Fisher matrix of the random effects vector in matrix notation is

$$I(b) = E_b\left[-\frac{\partial^2}{\partial b \partial b^T} log\left\{\mathcal{L}(\beta, \theta_*, \phi, b)\right\}\right] = \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{G}(\theta_*)^{-1}, \tag{B.8}$$

with again $\boldsymbol{W}$ denoting the weight matrix from above (compare Scheipl (2009)).

Let $\boldsymbol{W}^{(0)}$ denote $\boldsymbol{W}(b^{(0)})$, $\boldsymbol{\Delta}^{(0)} = \boldsymbol{\Delta}(b^{(0)})$, and $\mu^{(0)} = \mu(b^{(0)})$. For given $\boldsymbol{G}(\theta_*)$ and $\boldsymbol{W}^{(0)-1}$ the model can be rewritten with the help of pseudo-observations $\tilde{y}$ as a linear mixed model of the form

$$\tilde{y}|b \sim \mathcal{N}(\boldsymbol{Z}b, \boldsymbol{W}^{(0)-1}) \tag{B.9}$$

$$b \sim \mathcal{N}(0, \boldsymbol{G}(\theta_*)), \tag{B.10}$$

with pseudo-data or alternatively *working response*

$$\tilde{y} = \boldsymbol{Z}b^{(0)} + \boldsymbol{\Delta}^{(0)}(y - \mu^{(0)}). \tag{B.11}$$

The expression "iteratively reweighted" is used to emphasize the fact that the parameter estimates $\hat{b}^{(k)}$ are determined for a fixed weight matrix $\boldsymbol{W}$ and then the weights are updated to the current estimates. Thus, the complete PIRLS algorithm (for given β, θ_* and ϕ) is as follows:

Step 1 An initial value $\hat{b}^{(0)}$ and a stop criterion are chosen and k is put to 0.

Step 2 The working response $\tilde{y}^{(k)}$ and the weights function $\boldsymbol{W}^{(k)}$ are computed.

Step 3 The resulting weighted least-squares problem yielding an estimator for b are solved.

Step 2 and 3 are iterated until the stop criterion is fulfilled.

Algorithm 2. (Laplace Approximation)[1]

The idea of the Laplace approximation is to approximate a k-dimensional integral of the form $\int_{\mathbb{R}^k} exp(f(\theta))d\theta$ by a Gaussian distribution. It has been constructed for known functions $f(\theta)$ which are twice differentiable, unimodal and bounded. The solution is as follows:

1. Determination of the maximum of the function f, yielding $\theta_{max} = argmax\ f(\theta)$
2. Approximation of $f(\theta)$ by a second-order Taylor expansion around θ_{max}

$$f(\theta) \approx f(\theta_{max}) + \frac{1}{2}(\theta - \theta_{max})^T \underbrace{\left(\frac{\partial^2}{\partial\theta\partial\theta} f(\theta_{max})\right)}_{-\boldsymbol{P}} (\theta - \theta_{max}) \tag{B.12}$$

3. Approximation of the integrand by inserting the result of the quadratic approximation of f yields

$$\int_{\mathbb{R}^k} exp(f(\theta))d\theta \approx \int_{\mathbb{R}^k} exp(f(\theta_{max}) \underbrace{-\frac{1}{2}(\theta - \theta_{max})^T \boldsymbol{P}^{-1}(\theta - \theta_{max})}_{\text{Kernel of } \mathtt{N}(\theta_{max}, \boldsymbol{P}^{-1})})d\theta \tag{B.13}$$

Thus, the integral $\int_{\mathbb{R}^k} exp(f(\theta))d\theta$ can be approximated by

$$\int_{\mathbb{R}^k} exp(f(\theta))d\theta \approx exp(f(\theta_{max}))\sqrt{\frac{(2\pi)^k}{|\boldsymbol{P}|}}. \tag{B.14}$$

This method can be used for the numerical estimation of the components of GLMMs. The Laplace approximation then is applied to the marginal log-likelihood

$$\begin{aligned} log\{\mathcal{L}(\beta, \theta_*, \phi)\} &= log\,(f(y|\beta, \theta_*, \phi)) = log\left\{\int f(y|b, \beta, \phi) f(b|\theta_*) db\right\} \qquad \text{(B.15)} \\ &= log\left\{\int exp\left\{\frac{y^T\vartheta - b(\vartheta)}{\phi} - c(y, \phi)\right\} \frac{1}{\sqrt{|\boldsymbol{G}(\theta_*)|}} exp\left\{-\frac{1}{2} b^T \boldsymbol{G}(\theta_*)^{-1} b\right\} db\right\}, \end{aligned}$$

yielding as approximation

$$\begin{aligned} log\{\mathcal{L}(\beta, \theta_*, \phi)\} &\approx log\left\{\mathcal{L}(\beta, \hat{b}, \phi)\right\} \frac{1}{2} log|\boldsymbol{G}(\theta_*)| - \frac{1}{2}\hat{b}^T \boldsymbol{G}(\theta_*)\hat{b} \qquad \text{(B.16)} \\ &+ log\left\{\int exp\left(-\frac{1}{2}(b - \hat{b})^T \boldsymbol{I}(\hat{b})(b - \hat{b})\right) db\right\} \\ &\propto log\left\{\mathcal{L}(\beta, \hat{b}, \phi)\right\} - \frac{1}{2} log|\boldsymbol{G}(\theta_*)| - \frac{1}{2}\hat{b}^T \boldsymbol{G}(\theta_*)\hat{b} - \frac{1}{2} log|\boldsymbol{I}(\hat{b})|, \qquad \text{(B.17)} \end{aligned}$$

[1] Greven (2009), Scheipl (2009)

with $\boldsymbol{I}(b)$ denoting the Fisher information, i.e. the expectation of the negative second derivative of the log-likelihood with respect to the random effects vector

$$\boldsymbol{I}(b) = -E_b\left[\frac{\partial^2}{\partial b \partial b^T} log\left\{\mathcal{L}(\beta, \theta_*, \phi, b)\right\}\right] \tag{B.18}$$

$$= \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}(\theta_*)^{-1}, \tag{B.19}$$

where $\boldsymbol{W}$ is the weight matrix of the form

$$\boldsymbol{W} = diag\left(\frac{1}{\phi b''(\vartheta_i)}\left(\frac{\partial \mu_i}{\partial \eta}^2\right)\right)_{i=1,\ldots,n}. \tag{B.20}$$

Algorithm 3. (Bootstrap estimation for the covariance penalty term in the LMM)

In the following, the algorithm for the bootstrap estimation of the covariance penalty term in the case of normal errors will be described. Note that the modifications regarding the check for zero variance (6.1) are not included in the outline. In addition to the bootstrap algorithm described in this paragraph, the description of the alternative, computationally more complex variant of the joint measure (8.12) is given in the next paragraph.

The idea of this bootstrap algorithm is to estimate the covariance based penalty term (for known error variance (5.46) and for unknown error variance (5.48)) in its two versions:

- The **conditional** version, where the random effects are kept constant and
- the **joint** version, in which the random effects are also drawn from a distribution

In general, for *parametric* bootstrap, the bootstrap replications are constructed from the estimated (assumed) distribution

$$\hat{f} \rightarrow y^*$$

and the parameters, here denoted as μ, are then estimated in each bootstrap sample

$$y^* \rightarrow \hat{\mu}^* = m(y^*).$$

In this work, the bootstrap estimation is based on model components resulting from the estimation of the models which are compared via cAIC. Given these quantities, the following steps are executed.

Conditional

Step 1 A sufficiently large number of bootstrap replications (B) is chosen.[2]

Step 2 For each bootstrap replication $\xi = 1, \ldots, B$, new observations are generated as

$$y_i^{*\xi} = \boldsymbol{X}_i \hat{\beta} + \boldsymbol{Z}_i \hat{b}_i + \varepsilon_i^{*\xi}, \; i = 1, \ldots, n, \tag{B.21}$$

with $\hat{\beta}$ and $\hat{b}$ the BLUP-estimators for the linear mixed model, $\boldsymbol{X}$ and $\boldsymbol{Z}$ the associated design matrices and

$$\varepsilon_i^{*\xi} \sim \mathcal{N}(0, \hat{\sigma}^2), \; i = 1, \ldots, n, \tag{B.22}$$

where $\hat{\sigma}^2$ denotes the estimated error variance from the LMM.

Step 2 In each bootstrap sample, a model is fitted to the new data $(y_1^{*\xi}, \ldots, y_n^{*\xi})$, $\xi = 1, \ldots, B$, yielding an estimator for the linear predictor $\eta^{*\xi}$ – in the case of normal errors and identity link equal to the expectation $\mu^{*\xi}$ – and for the error variance σ^2.

Step 3 Next, for each i ($i = 1, \ldots, n$) the mean of the observations across all bootstrap samples is calculated

$$y_i^{*\cdot} = \frac{1}{B} \sum_{\xi=1}^{B} y_i^{*\xi}. \tag{B.23}$$

Step 4 The contribution to the estimator of the covariance of y_i and $\hat{\mu}_i$ of each bootstrap sample is calculated:

$$(y_i^{*\xi} - y_i^{*\cdot})\hat{\mu}_i^{*\xi}, \; \xi = 1, \ldots, B \; i = 1, \ldots, n \tag{B.24}$$

and is divided by either

(a) the estimated error variance from the initial LMM, $\hat{\sigma}^2$, yielding

$$(y_i^{*\xi} - y_i^{*\cdot})\frac{\hat{\mu}_i^{*\xi}}{\hat{\sigma}^2}, \; i = 1, \ldots, n, \tag{B.25}$$

or by

(b) the estimated error variances specific to each bootstrap replication, $(\hat{\sigma}^2)^{*\xi}$, for $\xi = 1 \ldots, B$, resulting in

$$(y_i^{*\xi} - y_i^{*\cdot})\frac{\hat{\mu}_i^{*\xi}}{(\hat{\sigma}^2)^{*\xi}}, \; i = 1, \ldots, n. \tag{B.26}$$

[2] What an adequate number is, can be learned from simulations (compare Chapter 6).

Step 5 The contributions are added up and divided by $(B-1)$ yielding

(a) for constant error variance

$$\frac{1}{B-1}\sum_{\xi=1}^{B}(y_i^{*\xi}-y_i^{*\cdot})\frac{\hat{\mu}_i^{*\xi}}{\hat{\sigma}^2},\ i=1,\dots,n, \tag{B.27}$$

and

(b) for sample specific error variances

$$\frac{1}{B-1}\sum_{\xi=1}^{B}(y_i^{*\xi}-y_i^{*\cdot})\frac{\hat{\mu}_i^{*\xi}}{(\hat{\sigma}^2)^{*\xi}},\ i=1,\dots,n. \tag{B.28}$$

Step 6 The sum of all individual estimations is taken, resulting in

(a)

$$\begin{aligned} gdf &= \sum_{i=1}^{n}\frac{1}{B-1}\sum_{\xi=1}^{B}(y_i^{*\xi}-y_i^{*\cdot})\frac{\hat{\mu}_i^{*\xi}}{\hat{\sigma}^2} && \text{(B.29)}\\ &= \frac{1}{\hat{\sigma}^2}\ \sum_{i=1}^{n}\frac{1}{B-1}\sum_{\xi=1}^{B}(y_i^{*\xi}-y_i^{*\cdot})\hat{\mu}_i^{*\xi}, && \text{(B.30)}\end{aligned}$$

or for specific error variances

(b)

$$gdf=\sum_{i=1}^{n}\frac{1}{B-1}\sum_{\xi=1}^{B}(y_i^{*\xi}-y_i^{*\cdot})\frac{\hat{\mu}_i^{*\xi}}{\hat{\sigma}^{2*\xi}}\ . \tag{B.31}$$

Joint

Step 1 A sufficiently large number of bootstrap replications (B) is chosen.[3]

Step 2 For each bootstrap replication $\xi=1,\dots,B$, new observations are generated as

$$y_i^{*\xi}=\boldsymbol{X}_i\hat{\beta}+\boldsymbol{Z}_ib^{*\xi}+\varepsilon_i^{*\xi},\ i=1,\dots,n, \tag{B.32}$$

with $\hat{\beta}$ the BLUP-estimator for the linear mixed model, $\boldsymbol{X}$ and $\boldsymbol{Z}$ the associated design matrices as in the conditional case and (for $i=1,\dots,n$)

$$b_i^{*\xi}\sim\mathcal{N}(0,\hat{\tau}^2) \tag{B.33}$$

$$\varepsilon_i^{*\xi}\sim\mathcal{N}(0,\hat{\sigma}^2), \tag{B.34}$$

where $\hat{\sigma}^2$ denotes the estimated error variance from the LMM (as in the conditional case) and $\hat{\tau}^2$ is the estimated random effects variance from the linear mixed model.

[3] What an adequate number is, can be learned from simulations (compare Chapter 6).

Step 3 In each bootstrap sample, a linear mixed model is fitted to the new data $(y_1^{*\xi}, \ldots, y_n^{*\xi})$, $\xi = 1, \ldots, B$, yielding an estimator for the linear predictor $\eta^{*\xi}$ – in the LMM equal to the expectation $\mu^{*\xi}$ – and for the error variance σ^2.

Step 4 Next, the contributions to the covariance of y_i and $\hat{\mu}_i$ are estimated as

$$\varepsilon_i^{*\xi}\hat{\mu}_i^{*\xi},\ \xi = 1, \ldots, B,\ i = 1, \ldots, n \tag{B.35}$$

end are divided by either

(a) the estimated error variance from the initial LMM, $\hat{\sigma}^2$, yielding

$$\varepsilon_i^{*\xi}\frac{\hat{\mu}_i^{*\xi}}{\hat{\sigma}^2},\ i = 1, \ldots, n \tag{B.36}$$

or by

(b) the estimated error variances specific to each bootstrap replication, $(\hat{\sigma}^2)^{*\xi}$, for $\xi = 1, \ldots, B$, resulting in

$$\varepsilon_i^{*\xi}\frac{\hat{\mu}_i^{*\xi}}{(\hat{\sigma}^2)^{*\xi}},\ i = 1, \ldots, n. \tag{B.37}$$

Step 5 The contributions are added up and divided by B[4], yielding

(a) for constant error variance

$$\frac{1}{B}\sum_{\xi=1}^{B}\varepsilon_i^{*\xi}\frac{\hat{\mu}_i^{*\xi}}{\hat{\sigma}^2},\ i = 1, \ldots, n \tag{B.38}$$

and

(b) for sample specific error variances

$$\frac{1}{B}\sum_{\xi=1}^{B}\varepsilon_i^{*\xi}\frac{\hat{\mu}_i^{*\xi}}{(\hat{\sigma}^2)^{*\xi}},\ i = 1, \ldots, n. \tag{B.39}$$

Step 6 The sum of all individual estimators is taken, resulting in

(a)

$$gdf = \sum_{i=1}^{n}\frac{1}{B}\sum_{\xi=1}^{B}\varepsilon_i^{*\xi}\frac{\hat{\mu}_i^{*\xi}}{\hat{\sigma}^2} \tag{B.40}$$

$$= \frac{1}{\hat{\sigma}^2}\ \sum_{i=1}^{n}\frac{1}{B}\sum_{\xi=1}^{B}\varepsilon_i^{*\xi}\hat{\mu}_i^{*\xi}, \tag{B.41}$$

[4] In this variant one does not have to account for an estimated mean and thus it is divided by B rather than $B - 1$.

or for specific error variances

(b)

$$gdf = \sum_{i=1}^{n} \frac{1}{B} \sum_{\xi=1}^{B} \varepsilon_i^{*\xi} \frac{\hat{\mu}_i^{*\xi}}{(\hat{\sigma}^2)^{*\xi}} \ . \tag{B.42}$$

Algorithm 4. (Alternative Bootstrap Estimation for the Joint Covariance Penalty Term in the LMM)

In this paragraph, the alternative for the computation of the joint covariance based measure will be outlined. Note that the computational cost is rather high.[5]

The measure is based on the idea to replace the average of the responses of the conditional computation ($y_i^{*\cdot}$) by a random effects specific average, such that the mean becomes $\boldsymbol{X}_i\hat{\beta} + \boldsymbol{Z}_i\hat{b}_i^{*\xi}$ instead of $\boldsymbol{X}_i\hat{\beta}$ (see Chapter 8). Note that again, it can be distinguished between the computation with constant error variance and the approach with re-estimated error variance in each sample. As the second variant turned out to be more adequate in the simulation studies in Chapter 6, the following will be restricted to non-constant error variances.

The proceeding is as follows:

Step 1 Sufficiently large numbers $B1$ (number of random effects) and $B2$ (number of error terms drawn for each random effect) are chosen.[6] Note that the computational expense rises rather rapidly with increasing numbers $B1$ and $B2$ as it indicates the number of models to be estimated.[7]

Step 2 $B1$ random effects are drawn from a $\mathcal{N}(0, \hat{\tau}^2)$ distribution, yielding

$$b_i^{*1}, \dots b_i^{*B1}, \text{ for } i = 1, \dots, n, \tag{B.43}$$

where $\hat{\tau}^2$ is the estimator of the random effects variance from the LMM.

Step 3 For each of the $B1$ random effects, $B2$ errors are drawn as

$$\varepsilon_i^{*\xi k} \sim \mathcal{N}(0, \hat{\sigma}^2), \ i = 1, \dots, n, \ \xi = 1, \dots, B1, \ k = 1, \dots, B2 \tag{B.44}$$

and $\hat{\sigma}^2$ denoting the estimated error variance.

Step 4 Based hereon, the associated responses $y_i^{*\xi k}$ are computed as

$$y_i^{*\xi k} = \boldsymbol{X}_i\hat{\beta} + \boldsymbol{Z}_i b_i^{*\xi} + \varepsilon_i^{*\xi k}, \ i = 1, \dots, n, \ \xi = 1, \dots, B1, \ k = 1, \dots, B2. \tag{B.45}$$

[5] Depending on the choices of the two replication numbers.

[6] Again, what numbers are sufficiently large can be learned from simulation studies.

[7] $B1 \times B2$ models have to be estimated in total.

Step 5 In a next step, to each of the responses $y^{*\xi k}$ a model is fitted, each yielding an estimator for the linear predictor $\eta^{\xi k}$, which is – in the case of normal errors and identity link – equal to the expectation $\mu^{\xi k}$. Moreover, an estimation of the error variance is obtained: $(\hat{\sigma}^2)^{*\xi k}$, $\xi = 1, \dots, B1$ and $k = 1, \dots, B2$. Note that for model failure the errors are re-drawn for the respective random effects and new responses are generated.

Step 6 Next, the mean of the responses is calculated for each random effect (across k)

$$y_i^{*\xi\cdot} = \sum_{k=1}^{B2} y_i^{*\xi k}, \; i = 1, \dots, n. \tag{B.46}$$

Step 7 The contributions to the estimator of the covariance are then determined by using the random effects specific mean of the responses and the sample specific error variances, yielding

$$\sum_{k=1}^{B2} (y_i^{*\xi k} - y_i^{*\xi\cdot}) \frac{\hat{\mu}_i^{*\xi k}}{(\hat{\sigma}^2)^{*\xi k}}, \; i = 1, \dots, n, \; \xi = 1, \dots, B1. \tag{B.47}$$

Step 8 This quantity is divided by $(B2-1)$[8] and he sum is taken with respect to the random effects $\xi = 1, \dots, B1$, yielding

$$\sum_{\xi=1}^{B1} \frac{1}{B2-1} \sum_{k=1}^{B2} (y_i^{*\xi k} - y_i^{*\xi\cdot}) \frac{\hat{\mu}_i^{*\xi k}}{(\hat{\sigma}^2)^{*\xi k}} \; . \tag{B.48}$$

Step 9 The individual estimators are then added, resulting in

$$\sum_{i=1}^{n} \sum_{\xi=1}^{B1} \frac{1}{B2-1} \sum_{k=1}^{B2} (y_i^{*\xi k} - y_i^{*\xi\cdot}) \frac{\hat{\mu}_i^{*\xi k}}{(\hat{\sigma}^2)^{*\xi k}} \; . \tag{B.49}$$

[8] The subtraction of 1 shall account for the estimated mean.

Appendix C

Supplement to the Simulation Studies

In the following, the complete results of the two simulation studies will be presented. This includes the plots of the selection frequencies for function f_1, f_2 and f_3 of the simulation study using penalized splines in Section 6.1 and those of the random intercept simulation in Section 6.2. The plots cover all settings, i.e. ML as well as REML estimation and all sample sizes. Note that for reasons of space, the scatter plot matrices of the degrees of freedom will not be listed here.

name of AIC	description
AIC_m1	AIC of the linear model
AICconvent_m2	conventional df (5.10)
AICapprox_m2_h1e.04	approximate cAIC(5.14) with $h = 0.0001$
AICanalyt_m2	analytic cAIC (5.23)
AICcov_m2_cond_Boot200	covariance based cAIC (5.46) (conditional version) with constant σ^2 and 200 bootstrap replications
AICcov_m2_cond_sig_in_B_Boot200	covariance based cAIC (5.48) (conditional version) with re-estimated σ^2 and 200 bootstrap replications
AICcov_m2_cond_check_Boot200	covariance based cAIC (5.46) with the check for zero variance (conditional version) with constant σ^2 and 200 bootstrap replications
AICcov_m2_cond_sig_in_B_check_Boot200	covariance based cAIC (5.48) with the check for zero variance (conditional version) with re-estimated σ^2 and 200 bootstrap replications
AICcov_m2_joint_BootB	covariance based cAIC (5.46) (joint version) with constant σ^2 and B bootstrap replications
AICcov_m2_joint_sig_in_B_BootB	covariance based cAIC (5.48) (joint version) with re-estimated σ^2 and B bootstrap replications
AICcov_m2_joint_check_BootB	covariance based cAIC (5.46) with the check for zero variance (joint version) with constant σ^2 and B bootstrap replications
AICcov_m2_joint_sig_in_B_check_BootB	covariance based cAIC (5.48) with the check for zero variance (joint version) with re-estimated σ^2 and B bootstrap replications
AICyuyau_tausq_in_num_m2	cAIC of Yu and Yau (5.67) in the representation where $\hat{\tau}^2$ appears only in the numerator; not expressed depending on the conventional measure
AICmgcv_m2	AIC automatically returned by function logLik.gamm {mgcv}
AICnlme_m2	AIC automatically returned by function logLik.lme {nlme}
mAIC	marginal AIC ((5.5) and (5.6))

Table C.1: *Names of the AICs in the simulation studies in Chapter 6. The associated degrees of freedom are named in the same way. The term* AIC *is simply replaced with* df*, e.g.* dfanalyt_m2*.*

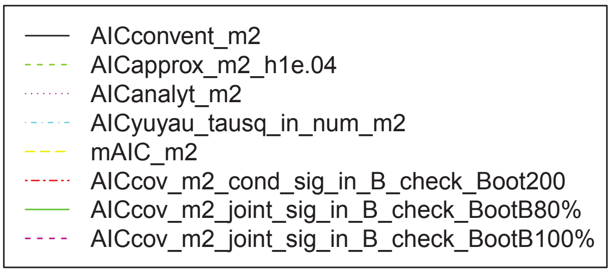

Figure C.1: *Legend for the selection frequency curves in figures C.2, C.3, C.4 and C.5.*

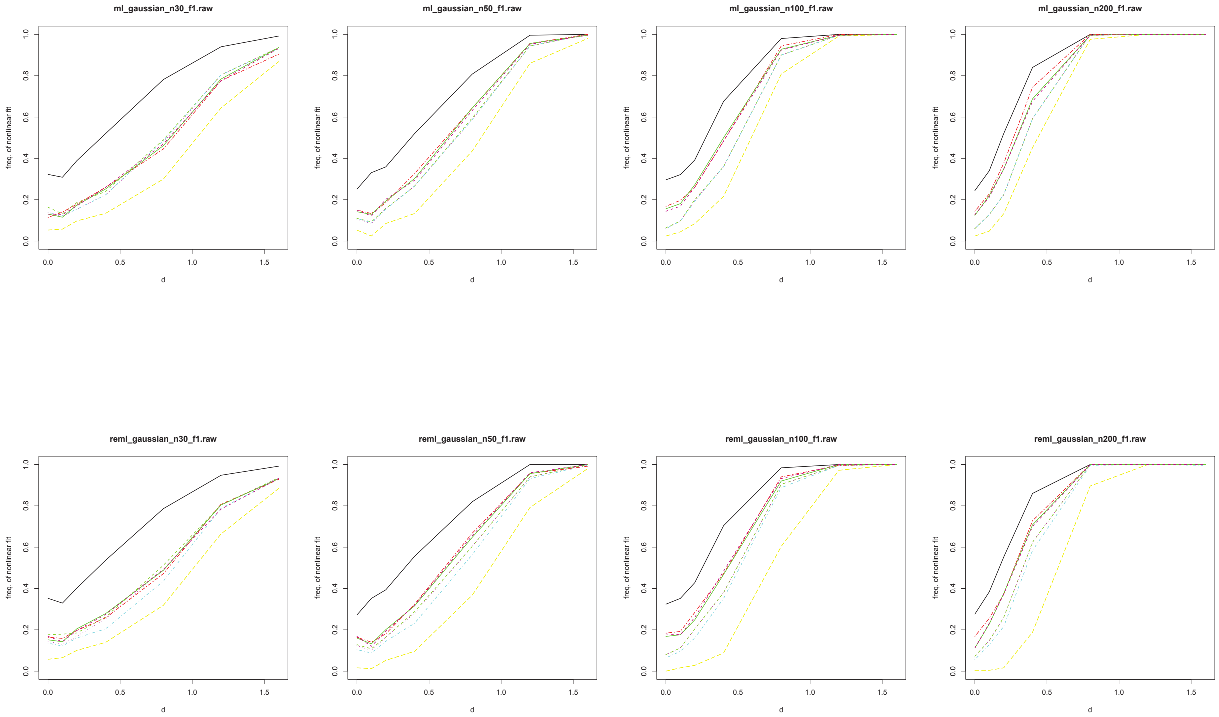

Figure C.2: *Complete results for function f_1 of the first simulation study (Section 6.1): Proportion of simulation replications where the non-linear model m_2 is favored by the respective AIC.*

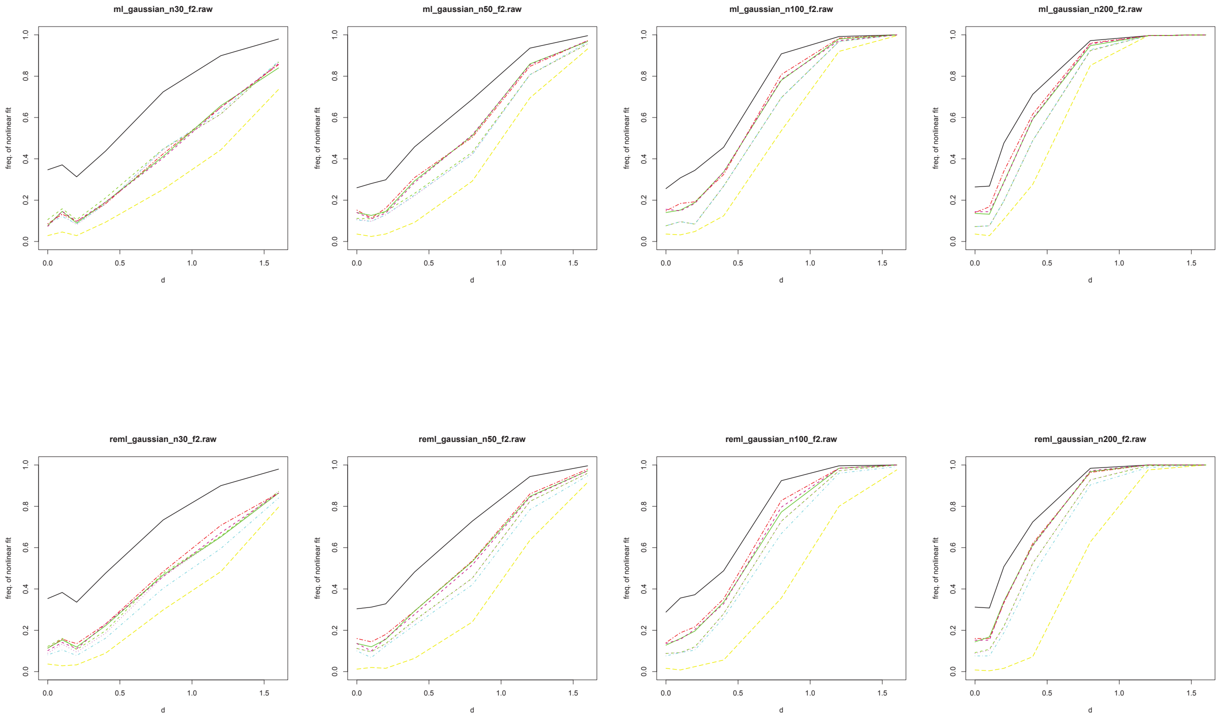

Figure C.3: *Complete results for function f_2 of the first simulation study (Section 6.1): Proportion of simulation replications where the non-linear model m_2 is favored by the respective AIC.*

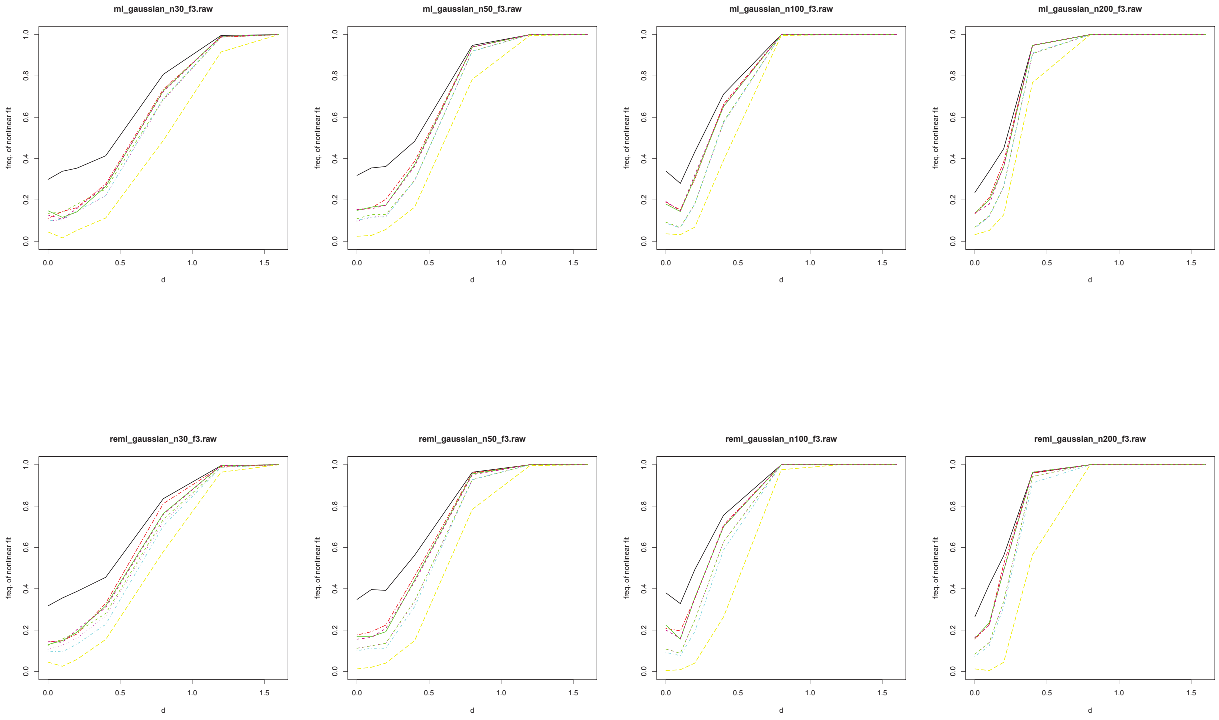

Figure C.4: *Complete results for function f_3 of the first simulation study (Section 6.1): Proportion of simulation replications where the non-linear model m_2 is favored by the respective AIC.*

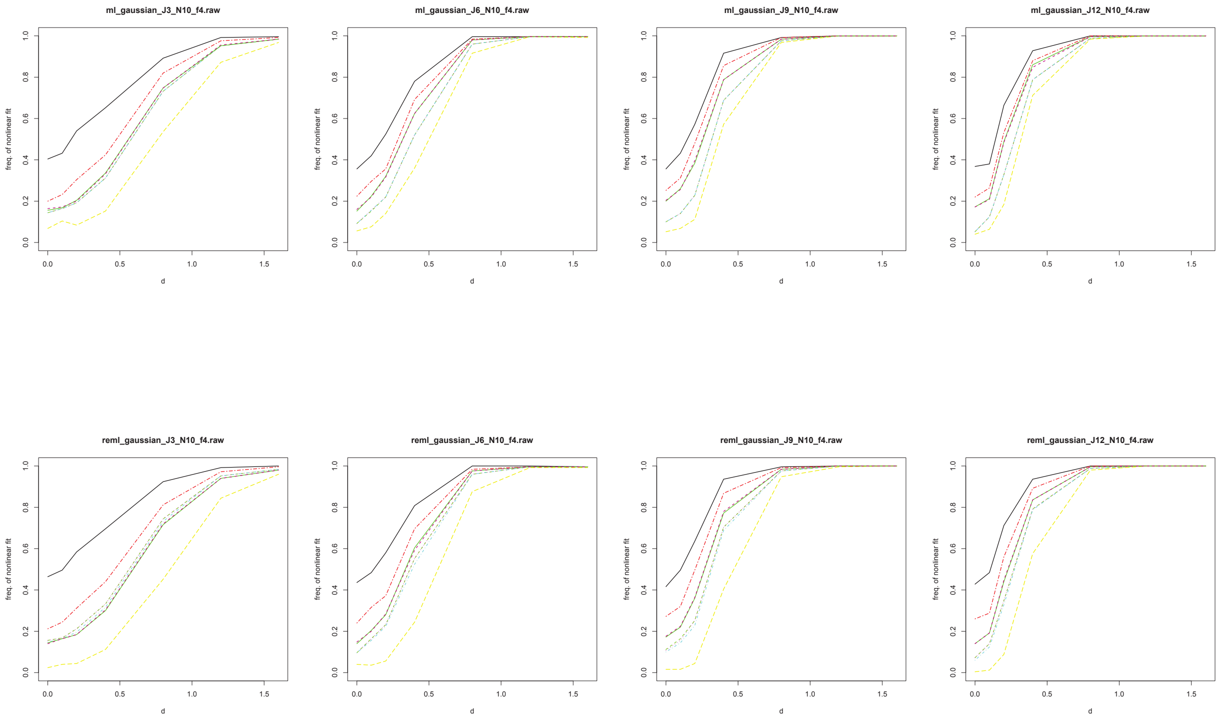

Figure C.5: *Complete results for the second simulation study (Section 6.2): Proportion of simulation replications where the non-linear model* m_2 *is favored by the respective AIC.*

Appendix D

Supplement to the Case Study

In the following, the variable description and the complete results of the case study on childhood malnutrition in Zambia will be presented.

Variable	**Description**
`csex`	gender of the child (1 = male, 0 = female)
`cfeed`	duration of breastfeeding (in months)
`cage`	**age of the child (in months)**
`mage`	**age of the mother (at birth, in years)**
`mheight`	height of the mother (in cm)
`mbmi`	body mass index of the mother
`medu`	education of the mother (1 = no education, 2 = primary school, 3 = elementary school, 4 = higher)
`mwork`	employment status of the mother (1 = employed, 0 = unemployed)
`district`	residential district (54 districts altogether)

Table D.1: *Explanatory variables in the Zambia data set. Source: Greven and Kneib (2010).*

name of measure	ML estimation	REML estimation
tausq2	1.81	2.37
ll1	-2214.02	-2214.02
ll2	-2150.72	-2150.31
var_null	0.00	0.00
df_m1	3.00	3.00
AIC_m1	4434.04	4434.04
dfconvent_m2	6.86	7.08
AICconvent_m2	4315.16	4314.77
dfapprox_m2_h1e − 04	7.47	7.74
AICapprox_m2_h1e − 04	4316.39	4316.10
dfanalyt_m2	7.47	7.74
AICanalyt_m2	4316.39	4316.10
dfcov_m2_cond_sig_in_B_Boot200	6.88	6.69
conv_error_m2_cond	0.00	0.00
AICcov_m2_cond_Boot200	4315.15	4313.95
AICcov_m2_cond_sig_in_B_Boot200	4315.21	4313.99
dfcov_m2_joint_Boot2000	7.50	7.09
dfcov_m2_joint_sig_in_B_Boot2000	7.50	7.10
conv_error_m2_joint	0.00	0.00
AICcov_m2_joint_Boot2000	4316.44	4314.80
AICcov_m2_joint_sig_in_B_Boot2000	4316.44	4314.81
Loglik_mgcv_m2	-2159.64	-2162.79
dfmgcv_m2	4.00	4.00
AICmgcv_m2	4327.29	4333.59
dfyuyau_tausq_in_num_m2	7.47	7.97
AICyuyau_tausq_in_num_m2	4316.39	4316.55
mll2	-2159.64	-2162.79
mdf_m2	4.00	4.00
mAIC_m2	4327.29	4333.59

Table D.2: *Complete table of measures for covariate* cage

name of measure	ML estimation	REML estimation
tausq2	0.01	0.04
ll1	-2268.29	-2268.29
ll2	-2267.69	-2267.07
var_null	0.00	0.00
df_m1	3.00	3.00
AIC_m1	4542.58	4542.58
dfconvent_m2	3.29	3.77
AICconvent_m2	4541.96	4541.69
dfapprox_m2_h1e − 04	5.74	4.58
AICapprox_m2_h1e − 04	4546.85	4543.30
dfanalyt_m2	5.74	4.58
AICanalyt_m2	4546.85	4543.30
dfcov_m2_cond_sig_in_B_Boot200	3.67	4.09
conv_error_m2_cond	0.00	0.00
AICcov_m2_cond_Boot200	4542.72	4542.30
AICcov_m2_cond_sig_in_B_Boot200	4542.73	4542.34
dfcov_m2_joint_Boot2000	3.58	4.26
dfcov_m2_joint_sig_in_B_Boot2000	3.58	4.26
conv_error_m2_joint	0.00	0.00
AICcov_m2_joint_Boot2000	4542.53	4542.66
AICcov_m2_joint_sig_in_B_Boot2000	4542.55	4542.68
Loglik_mgcv_m2	-2268.27	-2271.60
dfmgcv_m2	4.00	4.00
AICmgcv_m2	4544.54	4551.19
dfyuyau_tausq_in_num_m2	5.73	6.48
AICyuyau_tausq_in_num_m2	4546.85	4547.11
mll2	-2268.27	-2271.60
mdf_m2	4.00	4.00
mAIC_m2	4544.54	4551.19

Table D.3: *Complete table of measures for covariate* mage

Appendix E

R-code

E.1 LMM implementation in R

E.1.1 lme{nlme}

This function is suitable for the estimation of linear mixed models as in Section 3.1 and is called by function gamm {mgcv} used in the simulation study using penalized spline smoothing (6.1). Moreover, it was used in the second simulation study (6.2) for the estimation of the random intercept models.

Function lme{nlme} is used as follows[1]

lme(fixed, data, random, correlation, weights, subset, method, control, ...),

with the arguments

- object: An object inheriting from class lme, representing a fitted linear mixed model
- fixed: Specification of the fixed effects part of the model. A two-sided linear formula object with the response variable on the left of a ∼ operator and the terms separated by + operators on the right,
  e.g. response ∼ time (with time being a fixed effect).
- data: An optional data frame containing the variables named in fixed, random, correlation, weights, and subset. By default the variables are taken from the environment from which lme is called.
- random: Specification of the random effects part of the model.
  e.g. random = 1|subject: Random intercepts for every subject,
  or random = 1 + time|subject: Random intercepts and slopes for every subject.
  Moreover, multilevel models containing several random effects can be specified. In order to divide the data into groups, function groupedData() can be applied.

[1] R Development Core Team (2011)

- `correlation`: An optional `corStruct` object describing the within-group correlation structure. See the documentation of `corClasses` for a description of the available `corStruct` classes.

- `weights`: An optional `varFunc` object or one-sided formula describing the within-group heteroscedasticity structure. If given as a formula, it is used as the argument to `varFixed`, corresponding to fixed variance weights. Defaults to NULL, corresponding to homoscedastic within-group errors.

- `subset`: An optional expression indicating the subset of the rows of data that should be used in the fit. Default: all observations included.

- `method`: Specification if the estimation approach: either `"REML"` or `"ML"`. Default: `"REML"`.

- `control`: A list of control values for the estimation algorithm to replace the default values returned by the function `lmeControl`. Defaults to an empty list.

The extraction of the model components and predictions can be straightforwardly done by the commands

- `predict(level = 0)`: Extraction of the prediction on population level.

- `predict(level = j)`: Extraction of the prediction on level j,
  e.g. `level = 1` corresponds to the cluster level in the second simulation study (6.2).

- `fixed.effects`: Extraction of the fixed effects.

- `random.effects`: Extraction of the random effects.

- `getVarCov`: Random effects covariance matrix,
  e.g. $\hat{\tau}^2$ in the simulation study.

For a more detailed explanation (and more arguments and functions) see Pinheiro and Bates (2000).

E.1.2 gamm {mgcv}

Function `gamm` is used for the computation of generalized additive mixed models – models which include unknown smooth functions as well as random effects. In this work, it was utilized in the first simulation study (6.1) for the estimation of the non-linear model m_2. Technically, the function performs the re-parameterizations needed for the representation as mixed models as in Section 4.3 and calls function `lme` {`nlme`} (see above) in the case of Gaussianity with identical link and function `gammPQL` of package `mgcv` otherwise to actually estimate the model and then "unscrambles" the returned object such that it has the form of a `gam` object.[2] According to Wood (2006), the function is "basically a wrapper function for `lme`, or the GLMM fitting routine `glmmPQL`(...)". He also points out that it occurs often that numerical problems occur in the estimation, or failure of the PQL iterations in the generalized case.

Function `gamm` {`mgcv`} is used as follows.[3]

$$\texttt{gamm}(\texttt{formula}, \texttt{random}, \texttt{correlation}, \texttt{family}, \texttt{data}, \texttt{subset}, \texttt{niterPQL}, \texttt{method}, ...),$$

with the arguments

- `formula`: A formula like in a `GLM` with the difference that smooth terms can added to the right side of the formula,
  e.g. `response` $\sim$ `s(time)`.
  Note that models must contain at least one random effect: either a smooth with non-zero smoothing parameter, or a random effect specified in argument random.
  A smooth term

  $$\texttt{s}(\texttt{x}, \texttt{bs} = \text{`}\texttt{ps}\text{'}, \texttt{m} = \texttt{c}(2, 2))$$

  in the `formula` argument, specifies a cubic B-spline basis and a second order difference penalty on the coefficients[4], whereby the input `ps` stands for P-splines and in option $\texttt{m} = \texttt{c}(2, 2)$ the first entry specifies the order of the spline and the second gives the order of the difference penalty.
- `random`: Optional random effects structure, specified as in a call to function `lme`.
- `correlation`: An optional correlation structure object as used to define correlation structures in `lme`.
- `family`: In contrast to function `lme`, which is only capable to treat the case of normal errors, the `family` command allows to chose a distribution of the one-parametric exponential family and a link function. The default is set to `gaussian` with identity link.

[2] Wood (2006)

[3] R Development Core Team (2011)

[4] By default, ten inner knots are used.

- `data`: A data frame or list containing the model response variable and covariates required by the formula. By default the variables are taken from `environment(formula)`, typically the environment from which `gamm` is called.
- `subset`: An optional vector specifying a subset of observations to be used in the fitting process.
- `niterPQL`: Maximum number of PQL iterations (if any).
- `method`: Estimation method, either maximum likelihood estimation, specified by '`ML`' or restricted maximum likelihood estimation ('`REML`'). Note that this specification is ignored in the generalized case. Thus it is only possible to use both methods in the case of normal error terms and identity link, when function *lme* is called directly.

The outcome is a list of two items, a `gam` part and a `lme` part. An overview of the model fit is obtained by

- `summary(model$lme)`: For details on the underlying `lme` fit and by
- `summary(model$gam)`: For a summary of the style of function `gam {mgcv}`.

The extraction of the model components can by done by

- `predict(model$gam)` or `predict(model$lme)`: Extraction of the prediction
- `coef(model$lme)`[1:ncol($\boldsymbol{X}$)][5]: Extraction of the fixed effects vector, where $\boldsymbol{X}$ denotes the design matrix of the fixed effects and ncol denotes the number of columns.

The extraction of the design matrices, $\boldsymbol{X}$ and $\boldsymbol{Z}$, as well as the extraction of the estimated error variance and the smoothing parameter was performed by the use of function `extract.lmeDesign`, which is based on function `extract.lmeDesign` of the package `RLRsim` and was already used for the simulation studies of Greven and Kneib (2010). For more details, please see the attached `R`-code on disc.

[5] Already costumized to the simulation using penalized spline smoothing in 6.1.

E.2 Attached R-Code on Disc

Please note that the R-code of the simulation studies and of the case study is attached on a disc. The files can be divided into three categories. The first comprises the R-code of the simulation study using penalized splines smoothing (with the ending gamm). The second includes the R-code of the simulation study using random intercept models (with the ending RI). Note that some files are used in both simulations and have thus no specific ending. The third category covers the R-code of the case study on malnutrition in Zambia. As we used penalized spline smoothing for the estimations in the case study, the files of the first simulation study are additionally used. The following packages have to be installed to conduct the simulations studies:

- mgcv
- nlme
- foreach
- [optional] doMC (only for Unix systems)
- quantreg
- car
- Matrix.

The code is fully commented. Note that many parts are based on/taken from the simulation studies of Greven and Kneib (2010).

The structure of the R-code of the first simulation study will be briefly described in the following (it can be directly transfered to the second simulations study):

1. The data (gaussian.Rdata) is generated by using the file gendata.R (which calls the file fcts_corrected.r which in turn calls Biometrika_paper_Psplines.r).
2. The main simulation step is performed in the sim_gaussian_selbst_gamm which uses the data (gaussian.Rdata) and calls
   - Gesamt_AIC_Spline_Sim_neu_gamm.r In Gesamt_AIC_Spline_Sim_neu_gamm.r all degrees of freedom and cAICs are computed, it calls:
     - fcts_corrected.r
     - Biometrika_paper_Psplines.r
     - dfnaive.r
     - dfanalyt.r
     - dfliang_gamm.r
     - dfefron_gamm_schranke.r
     - dfyuyau_tausq_in_numerator.r

  - dfmarginal.r

- Biometrika_paper_Psplines_gamm.r.

The results are returned in a folder called results_gamm. The selection frequency plots for all settings are obtained by the file summary_selbst_gamm.r which calls plotAIC_corrected_gamm_na_exclude.r and Farbskala.r. The resulting pdf-file is called results_gamm_na_exclude.r

Note that some additional files are included, such as the implementations of all representations of the cAIC of Yu and Yau (2011) and alternative implementations of the covariance based cAIC of Efron (2004).

Appendix F

Abbreviations and Symbols

AI	Akaike information
AIC	Akaike information criterion
cAIC	Conditional Akaike information criterion
mAIC	Marginal Akaike information criterion
(G)LM	(Generalized) linear model
(G)LMM	(Generalized) linear mixed model
KLD	Kullback-Leibler distance
BC	Bias correction
ML	Maximum likelihood
REML	Restricted maximum likelihood
(g)df	(Generalized) degrees of freedom
pmf	Probability mass function
pdf	Probability density function
i.i.d.	Independent and identically distributed
TP-Basis	Truncated powers basis
BLUE	Best linear unbiased estimator
(G)LS	(Generalized or weighted) least-squares
(E)BLUP	(Empirical) best linear unbiased predictor
(P)IRLS	(Penalized) Iteratively Reweighted least-squares
LA	Laplace approximation
PQL	Penalized Quasi-Likelihood
(A)GQ	(Adaptive) Gaussian quadrature
pen	Penalized
NA	Not available

Table F.1: *Abbreviations used in this thesis.*

$\mathbb{R}$	Real numbers
$\forall$	For all
$\Leftrightarrow$	If and only if
$exp(\cdot)$	Exponential function
$log(\cdot)$	Natural logarithm function
$tr(\cdot)$	Trace function
$det(\cdot)$	Determinant of a matrix
$\|\boldsymbol{V}\|$	Determinant of matrix $\boldsymbol{V}$
$diag(\cdot)$	Diagonal matrix
$id()$	Identity function
$\boldsymbol{I}_n$	$n \times n$ Identity matrix
x^T	x transposed
$\boldsymbol{V}^{1/2}$	(E.g.) Cholesky square root of matrix $\boldsymbol{V}$
$\frac{\partial f(y)}{\partial y}$	First partial derivative of $f(y)$ with respect to y
$\frac{\partial^2 f(y)}{\partial y^2}$	Second partial derivative of $f(y)$ with respect to y
$f'(\cdot)$	First derivative of function f
$f''(\cdot)$	Second derivative of function f
$\hat{\theta}$	Estimation of θ
$\approx$	Approximate
$\propto$	Proportional to
$\sim$	Distributed
$\mathcal{N}(\mu, \boldsymbol{\Sigma})$	Normal distribution with mean μ and covariance matrix $\boldsymbol{\Sigma}$
$E_g(\boldsymbol{X})$	Expectation of $\boldsymbol{X}$ with respect to g
$E_g(\boldsymbol{X}\|b)$	Conditional (to b) expectation of $\boldsymbol{X}$ with respect to g
$Var_g(\boldsymbol{X})$	Variance of $\boldsymbol{X}$ with respect to g
$E_g(\boldsymbol{X})$	Covariance of $\boldsymbol{X}$ with respect to g
$g(y\|b)$	Conditional distribution of y given b
$g(y, b)$	Joint distribution of y and b
$\mathcal{L}(\cdot)$	Likelihood
$l(\cdot)$	Log-likelihood
β_0	Intercept

Table F.2: *Symbols used in this thesis.*

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, volume 1, pages 267–281. Springer Verlag.

Akaike, H. (1983). Information measures and model selection. *Bulletin of the International Statistical Institute*, 50(1):277–291.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52:345–370.

Burnham, K. and Anderson, D. (2002). *Model selection and multimodel inference: A practical information-theoretic approach.* Springer-Verlag.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory.* John Wiley & Sons, Inc., New York.

Crainiceanu, C. M. and 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.

DeLeeuw, J. (1988). *Model selection in multinomial experiments*, pages 118–138. Dijkstra, T.K., New York.

DeLeeuw, J. (1992). Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle. *Breakthroughs in statistics*, 1:599–609.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, series B*, 39(1):1–38.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7(1).

Efron, B. (2004). The Estimation of Prediction Error: Covariance Penalties and Cross Validation. *Journal of the American Statistical Association*, 99(467):619–632.

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-Splines and penalties. *Statistical Science*, 11(2):89–121.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, 14(3):715–745.

Fahrmeir, L., Kneib, T., and Lang, S. (2007). *Regression: Modelle, Methoden und Anwendungen*. Springer-Verlag, Berlin, Heidelberg, 2nd edition.

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of The Royal Society A: Mathematical, Physical and Engineering Sciences*, 222:309–368.

Greven, S. (2008). *Non-Standard Problems in Inference for Additive and Linear Mixed Models*. PhD thesis, Ludwig-Maximilians-Universität München.

Greven, S. (2009). Advanced Longitudinal Data Analysis. Lecture notes.

Greven, S. (2011a). The cAIC of Yu and Yau in Linear Mixed Models. Ideas on the conditional AIC.

Greven, S. (2011b). The joint covariance-based cAIC. Ideas on the conditional AIC.

Greven, S. and Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, 97(4):773–789.

Hämmerlin, G. and Hoffmann, K.-H. (1994). *Numerische Mathematik*. Springer, Berlin, 4th edition.

Heumann, C., Fahrmeir, L., Dargatz, C., and Bayerstadler, A. (2010). Testen und Schätzen II. Lecture notes.

Hodges, J. and Sargent, D. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, 88(2):367.

Kandala, N., Lang, S., Klasen, S., and Fahrmeir, L. (2001). Semi-parametric analysis of the socio-demographic and spatial determinants of undernutrition in two African countries. *Research in Official Statistics*, 4(1):81–99.

Kneib, T. (2003). Bayes-Inferenz in generalisierten geoadditiven gemischten Modellen (Korrigierte Version). Master's thesis, Ludwig-Maximilians.Universität München.

Konrath, S. (2009). Gemischte Modelle. Lecture notes.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathtematical Statistics*, 22(1):79–86.

Kurtz, A. K. (1948). A Research Test of the Rorschach Test. *Personnel Psychology*, 1:41–51.

Liang, H., Wu, H., and Zou, G. (2008). A Note on Conditional AIC for Linear Mixed-Effects Models. *Biometrika*, 95(3):773–778.

Mallows, C. L. (1973). Some Comments on C P. *Technometrics*, 15(4):661–675.

Mansmann, U. (2009). Analyse longitudinaler Daten. Lecture notes.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, Boca Raton, Florida, 2nd edition.

Pinheiro, J. and Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer, New York.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Mixed Models*, volume 1, chapter 4, pages 91–111. Cambridge University Press, New York.

Scheipl, F. (2009). Gemischte Modelle. Lecture notes.

Self, S. G. and 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):pp. 605–610.

Shannon, C. (1948). The mathematical theory of communication. 1963. *M.D. computing: computers in medical practice*, 14(4):306–17.

Stein, C. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9(6):1135–1151.

Stone, C. (1982). Local asymptotic admissibility of a generalization of Akaike's model selection rule. *Annals of the Institute of Statistical Mathematics*, 34:123–133.

Takeuchi, K. (1976). Distribution of informal statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences), in Japanese*, (153):12–18.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.

Tutz, G. (2010/11). Generalisierte lineare Modelle. Lecture notes.

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

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Statistics and Computing Series. Springer-Verlag, New York, NY.

Walker, S. (1996). An EM Algorithm for Nonlinear Random Effects Models. *Biometrics*, 52(3):934–944.

Wood, S. (2006). *Generalized additive models: an introduction with R*. Texts in statistical science. Chapman & Hall/CRC.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Society*, pages 3–36.

Ye, J. (1998). On Measuring and Correction the Effects of Data Mining and Model Selection. *Journal of the American Statistical Association*, 93:120–131.

Yu, D. and Yau, K. K. W. (2011). Conditional Akaike information criterion for generalized linear mixed models. Scientific report.

List of Figures

List of Tables

Declaration

I hereby declare that I have written this diploma thesis on my own

and have used no other than the stated sources and aids.

Hiermit versichere ich, dass ich die von mir vorgelegte Arbeit

selbstständig verfasst habe

und andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt wurden.

München, 11. August 2011