Model choice
and
variable selection
in
mixed & semiparametric models

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
'Doctor rerum naturalium'
der Georg-August-Universität Göttingen

im Promotionsprogramm PhD School of Mathematical Sciences (SMS)
der Georg-August University School of Science (GAUSS)

vorgelegt von
Benjamin Säfken
aus Oldenburg

Göttingen, 2015

Betreuungsausschuss

Prof. Dr. Thomas Kneib, Lehrstühle für Statistik und Ökonometrie, Georg-August-Universität Göttingen

Prof. Dr. Tatyana Krivobokova, Institut für Mathematische Stochastik, Georg-August-Universität Göttingen

Mitglieder der Prüfungskommision

Referent: Prof. Dr. Thomas Kneib, Lehrstühle für Statistik und Ökonometrie, Georg-August-Universität Göttingen

Koreferentin: Prof. Dr. Tatyana Krivobokova, Institut für Mathematische Stochastik, Georg-August-Universität Göttingen

Weitere Mitglieder der Prüfungskommision

Prof. Dr. Simon Wood, Department of Mathematical Sciences, University of Bath

Prof. Dr. Andrea Krajina, Institut für Mathematische Stochastik, Georg-August-Universität Göttingen

Prof. Dr. Felix Krahmer, Institut für Numerische und Angewandte Mathematik, Georg-August-Universität Göttingen

Prof. Dr. Dominic Schuhmacher, Institut für Mathematische Stochastik, Georg-August-Universität Göttingen

Tag der mündlichen Prüfung: 27. März 2015

Acknowledgements

The work on this thesis has been a both stunning and painful process. I could not have completed my work without the support, encouragement and empathy of a select circle of people. I want to thank . . .

. . . Thomas for persuading me to come to Göttingen, for being optimistic about my work when I wasn't, for always offering clever and friendly advice and for the "Gelehrtenrepublik".

. . . my friend and colleague Holger for three years of laughter, fun and shared suffering in the office known as the all-time "office-of-the-month".

. . . Elisabeth for the warm welcome in Göttingen, shared breakfast and for her constant efforts to convince me to go climbing.

. . . Hauke, Alex and all the other colleagues who made the time in the blue tower invaluable.

. . . all the people that made my research visit to Bath possible. Especially Simon for the fruitful discussions and for joining my thesis committee.

. . . my parents Jean & Hermann for 30 years of incredible support and for believing in me when no one else did.

But most of all I want to thank my daughter Frida and my wife Katha for making me happy, cheering me up and making all the effort worthwhile.

Benjamin Säfken

Zusammenfassung

Mithilfe semiparametrischer und gemischter Modelle kann eine Vielzahl verschiedenartiger Datentypen und -strukturen in Regressionsmodellen berücksichtigt werden. Räumliche und zeitliche Strukturen diskreter und stetiger Daten können ebenso flexibel behandelt werden wie z. B. Daten mit funktionaler Struktur. Diese steigende Flexibilität verlangt von einem Statistiker, sich in zunehmendem Maße zwischen konkurrierenden Modellen zu entscheiden.

In der Modellwahl spielen die Freiheitsgrade, als Maß für die Komplexität von Modellen, eine zentrale Rolle. In dieser Arbeit werden drei Ansätze, jeweils für verschiedene Verteilungen der (konditionalen) Zielgröße, zur Schätzung der Freiheitsgrade in gemischten und semiparametrischen Modellen entwickelt. Da sich semiparametrische Modelle als gemischte Modelle darstellen lassen, können die gleichen Modellwahlverfahren für beide Modellklassen verwendet werden.

Durch die Verwendung Steinscher Methoden können die Freiheitsgrade für eine Gruppe von Verteilungen aus der Exponentialfamilie bestimmt werden. Die entwickelten Methoden zur Bestimmung der Freiheitsgrade werden anhand eines Datenbeispieles zum Baumwachstum veranschaulicht.

Freiheitsgrade für eine größere Gruppe von Verteilungen lassen sich durch Kreuzvalidierung und Bootstrap-Verfahren bestimmen. Auch lässt sich eine approximative Steinsche Methode für weitere Verteilung geeignet anpassen.

Basierend auf dem Satz über implizite Funktionen lassen sich in Modellen mit normalverteilter Zielgröße die Freiheitsgrade von Varianz- und Glättungsparametern analytisch berechnen. Berücksichtigt man diese Freiheitsgrade nicht, kann dies zu verzehrter Modellwahl führen. Neben der methodischen Entwicklung werden auch geometrische Eigenschaften der Freiheitsgrade von Varianz- und Glättungsparameter analysiert. Des Weiteren werden numerische Probleme bei der Berechnung der Freiheitsgrade behandelt.

Abstract

Semiparametric and mixed models allow different kinds of data structures and data types to be considered in regression models. Spatial and temporal structures of discrete or spatial data can be treated as flexibly as, for instance, functional data. This growing flexibility increasingly requires a statistician to make choices between competing models.

In model selection the degrees of freedom play an important role as a measure of model complexity. In this thesis three approaches for the estimation of the degrees of freedom in mixed and semiparametric models are developed, each for different distributions of the (conditional) responses. The interpretation of semiparametric models as mixed models justifies using the same model selection techniques for both model classes.

By using Steinian methods, the degrees of freedom can be determined for a group of distributions belonging to the exponential family. The developed methods for determining the degrees of freedom are illustrated by an example of tree growth data.

For a larger class of distributions the degrees of freedom can be determined by cross-validation and bootstrap methods. Additionally, an approximate Steinian method can be adapted for further distributions.

Based on the implicit function theorem the degrees of freedom of a variance or smoothing parameter can de derived analytically if the response is normally distributed. Failure to take these degrees of freedom into account can lead to biased model selection. In addition to the methodological derivation, the geometrical properties of the degrees of freedom of the variance and smoothing parameters are analysed. Furthermore, numerical problems in the computation of the degrees of freedom are considered.

Contents

Introduction

1 Semiparametric regression & mixed models

Regression models allow for the quantification of the influence of information (the covariates), given on numerous different scales, on the distribution of the outcomes. The information can influence different properties of the distribution of the outcome. This thesis is concerned with mean regression. Hence, the influence of the information on the distribution is only modelled through the mean of the distribution.

In this setting the distributions of the outcomes $y_1, \dots, y_n$, also called responses, are represented by a statistical model

$$f(y_i|\mu_i, \phi), \ i = 1, \dots, n, \tag{1}$$

with observation dependent on mean parameter μ_i and further distributional parameter ϕ, that do not need to be univariate. The means depend on predictors $\eta_i = g(\mu_i)$ mapped to each other by a link function $g(\cdot)$ that ensures the means to lie on a certain support, such that $f(\cdot|\mu_i)$ defines the envisaged probability distribution. The predictor contains the the covariates influencing the outcomes, including the parametric structure of the influence of the information:

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \tag{2}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are design matrices containing the covariates, $\boldsymbol{\beta}$ are fixed parameters and $\boldsymbol{u}$ are random or penalized parameters. The distinction between the two types of parameters allows for a joint modelling of mixed models and semiparametric models in one framework. The theoretical difference between both approaches lies in the assumptions about the random or penalized parameters $\boldsymbol{u}$.

In the mixed model framework these are random parameters following a normal distribution:

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

with covariance matrix $\boldsymbol{D}$ depending on a vector of variance parameters.
In the framework of semiparametric regression the parameters $\boldsymbol{u}$ are penalized by a quadratic penalty

$$-\frac{1}{2}\boldsymbol{u}^t\boldsymbol{D}^{-1}\boldsymbol{u},$$

with the inverse covariance matrix as penalty matrix. In this formulation the inverse variance parameters are proportional to smoothing parameters controlling the extent of penalization. The connection between the semiparametric and the mixed model framework has been known for a long time (Green, 1984; Wahba, 1985) and very general methods for the representation of semiparametric models as mixed models are available (Kneib, 2005). In either framework the estimated (or predicted) model coefficients are equivalent to the penalized maximum likelihood estimates.
Historically, the estimation of the smoothing or variance parameters in the semiparametric and the mixed model framework rely on distinct paradigms. In the semiparametric framework, an aspect also reflected in the spline smoothing literature, the smoothing parameter was chosen based on optimality criteria that approximate the average mean squared error. The most prominent criterion is generalized cross-validation (GCV, Craven and Wahba, 1978).
In the mixed model framework the variance parameters of the random effects are estimated by maximizing the (restricted) marginal likelihood (REML), where the fixed and random effects are profiled and integrated out (Patterson and Thompson, 1971; Laird and Ware, 1982). In case (1) is not normal, a Laplace approximation can be used (Wood, 2011; Wood, Pya, and Säfken, 2014). This is equivalent to an empirical Bayes approach where a noninformative prior is assigned to the fixed parameters. Recent work by Reiss and Ogden (2009) and Krivobokova (2013) shows that the finite sample performance of marginal likelihood based smoothing and variance parameter selection is superior under certain settings.

The framework presented above is applicable to many types of models. For instance, if (1) is an exponential family distribution and the random parameters in (2) are omitted it becomes the well-known generalized linear model (GLM). However, the response distributions are not limited to the exponential family. Further non-exponential family distributions for the responses are possible (for example scaled t, negative binomial or beta distributions).

Although not considered here, regression models can go far beyond the mean regression (Kneib, 2013) and thus also the influence of the covariates on further distributional parameters could be considered, leading to models for location, scale and shape (Rigby and Stasinopoulos, 2005).

The information that is influencing the mean is measured on multiple scales. For instance the covariates containing the information may be continuous, discrete, spatial or even functional. A vast variety of this information can be represented in the mixed model formulation (2).

Commonly mixed models are used for longitudinal data analysis (Laird and Ware, 1982) and cluster analysis. Thus the data consists of repeated measurements of a subject or cluster, and subject- or cluster-specific random effects are assigned. But also more complex hierarchical grouping structures are possible.

In the semiparametric framework low rank smoothing splines are used to model the influence of a smooth function of a covariate. The most prominent classes are thin plate regression splines (Wood, 2003) and P-splines (Eilers and Marx, 1996). It is also possible to model smooth interactions and continuous spatial effects with tensor product splines or radial basis functions.

Another prominent class are Gaussian Markov random fields (Rue and Held, 2005) for discrete spatial data inheriting a neighbourhood structure. Moreover, with Gaussian Markov random fields, in general all latent Gaussian models as presented in the Bayesian framework in Rue, Martino, and Chopin (2009) and Lindgren, Rue, and Lindström (2011) are available.

Even functional data can be analysed with the help of representation (2), i.e. signal regression (Ramsay and Silverman, 1997, 2002).

In statistical applications many different sources of information often influence the mean of a distribution. In this case, more than one of the preceding classes is incorporated into the predictor (2). This leads to structured additive or generalized additive models (GAM, Hastie and Tibshirani, 1990; Wood, 2006). These models are very flexible and have become very popular in recent years.

A few examples of the preceding models for the (covariate) information are given in Figure (1).

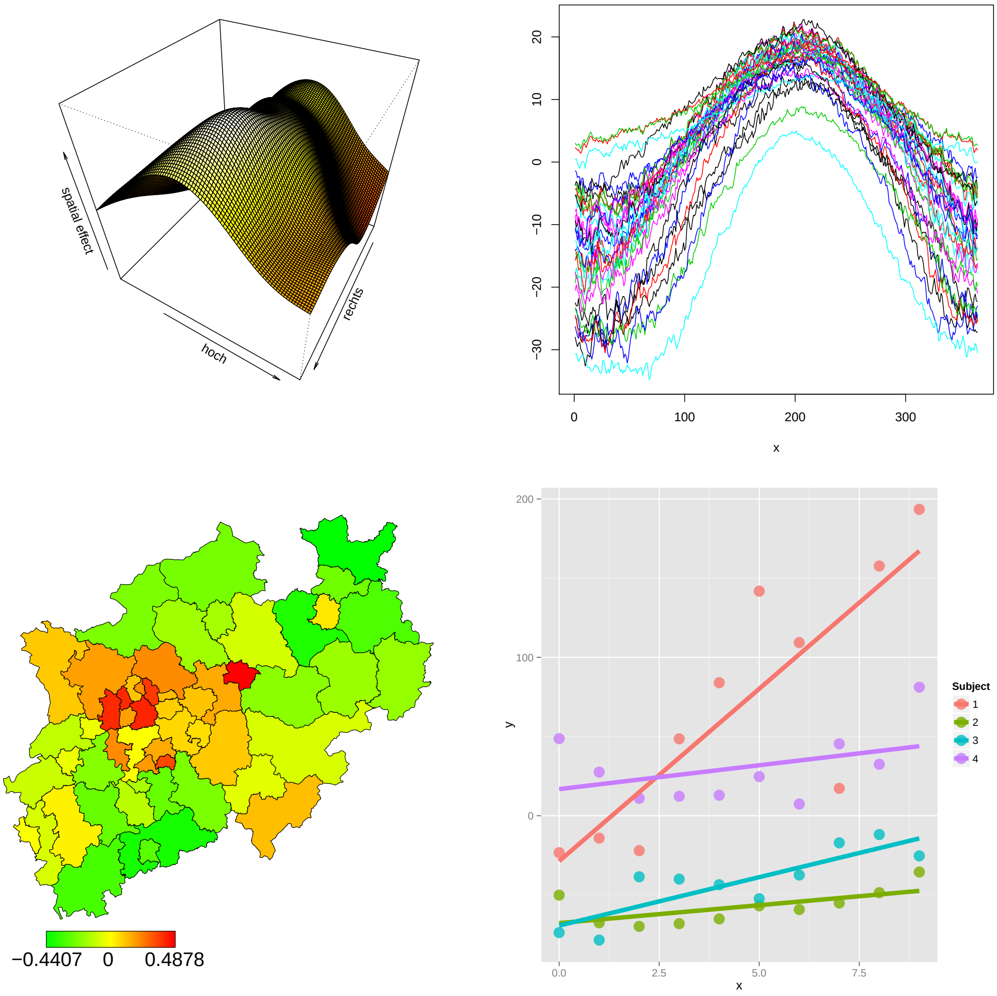

Figure 1: Examples of covariate information that can be modelled with (2). Upper left: two-dimensional spline for continuous spatial data; upper right: functional data observed on a grid; lower left: Gaussian Markov random field for discrete spatial data; lower right: longitudinal data with subject specific random intercept and random slope.

2 Model choice & variable selection

There is increasing need for model choice and variable selection in semiparametric and mixed models. There are different reasons for this. Firstly, the increased flexibility and complexity of the previously discussed models can lead to severe overfitting. This is indicated by the high dimensionality of semiparametric and mixed models. Secondly, with accelerating computation power and storage capacities the ease with which data are collected has increased and hence more data are available. These large amounts of data need to be analysed and appropriate models have to be determined. This development has recently been discussed under the term 'Big Data'.

The increasing interest in model choice and variable selection not only in semiparametric and mixed models is reflected by the number of publications in this field. Detailed and comprehensive overviews can be found in Claeskens and Hjort (2010) and Burnham and Anderson (2010). Model selection, as such, is becoming more and more an integral part of the process of statistical modelling. However, model choice cannot be simply reduced to finding the one true model that generates the data. Hence, there are several aspects that need to be considered in the process of model choice and variable selection.

There might not even be a true underlying data-generating mechanism or at least none that can be described through the limited human power of thought. It is nevertheless the objective of the model selection process to find an appropriate model that describes data, although it might be far from the truth. The so-called *principle of parsimony* reflects this. This heuristic principal, often known as Ockham's razor, can be found in many fields of science. For instance, the principle is expressed in the following quotation:

> The basic concepts and laws which are not logically further reducible constitute the indispensable and not rationally deducible part of the theory. It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.
>
> – Albert Einstein, *On the Method of Theoretical Physics*

This principle induces a trade-off between complexity and model fit. In statistics this trade-off is often referred to as the bias-variance trade-off and employed in the estimation process in order to avoid over- and underfitting.

Furthermore, the model selection process needs to account for the objective of the underlying scientific analysis and the research question to be explored with the help of the model and the data. Thus, pure mean prediction might not be reasonable when the research question aims at other properties of a distribution.

Model selection based on the AIC (Akaike information criterion) is probably the most common strategy when choosing between competing models (Akaike, 1973). This may be due to the simplicity and broad applicability of the criterion. It measures the fit of a model by the log likelihood and penalises it with the model complexity, given by the degrees of freedom:

$$AIC = -2 \cdot \text{log-likelihood} + 2 \cdot \text{degrees of freedom}.$$

In simple regression settings the degrees of freedom are the number of parameters. However, the AIC is not only justified by heuristically balancing between complexity and fit. It can rather be derived as an estimator of the relative Kullback-Leibler distance (Kullback and Leibler, 1951). The Kullback-Leibler distance, known from information theory, measures the distance between two probability distributions. Hence, the AIC has the information theoretic interpretation of estimating the relative information that is lost by the model when used as an approximation to a true underlying data-generating process.

An instructive overview about how closely the AIC is linked to other model selection techniques such as bootstrap and cross-validation is given by Efron (2004). It highlights the connection between the degrees of freedom and covariance penalties, effected by the estimation of the prediction error of a regression model.

There have been several extensions to the AIC. For example, in the case of small sample size or highly overparameterized models Hurvich and Tsai (1989) proposed a corrected criterion called AIC_C. In linear mixed models parts of the parameters are random variables, hence a natural choice would be to base the AIC on the marginal model, i.e. the model with the random effects integrated out. This leads to a biased criterion as stated in Greven and Kneib (2010). An AIC based on the conditional likelihood, with the random effects treated as penalized parameters, was introduced by Vaida and Blanchard (2005). However, the degrees of freedom were derived assuming the variance parameters of the random effects to be known. Plugging in estimated variance-covariance matrices induces a bias that leads to a preference for larger models with more random effects (Greven and Kneib, 2010). A correction to avoid that bias was proposed by Liang, Wu, and Zou (2008) using an identity known from Stein (1972). Greven and Kneib (2010) show that there is even an analytical representation of the correction. Similar results for the Schwarz criterion or Bayesian information criterion (BIC) are derived by Delattre, Lavielle, and Poursat (2014).

Extensions of the conditional AIC to exponential family distributions have recently been proposed by different authors. One approach that suffers from the same flaws as the criterion proposed by Vaida and Blanchard (2005) is presented by Donohue, Overholser, Xu, and Vaida (2011). Another asymptotic AIC proposed by Yu and Yau (2012) needs strict regulatory conditions in terms of the estimation technique used for estimating the random effects covariance parameters. An AIC that overcomes these weaknesses is proposed by Wood et al. (2014).

3 Outline

This thesis looks at the problem of the estimation of the prediction error and the degrees of freedom from different perspectives. Each perspective is connected to a research question or a specific aim, which are treated in the corresponding chapter. In the following the research questions and aims of each chapter are stated. They offer a good guideline through the thesis:

The **first** major research question that is discussed in this thesis is:

How can the conditional AIC proposed in Greven and Kneib (2010) be extended to further exponential family distributions?

A unified framework for the estimation of the conditional AIC in generalized linear mixed models is developed in the first chapter. A direct extension of the findings of Greven and Kneib (2010) is possible for some distributions. For these distributions extensive simulations and an application are presented. The approach, however, does not apply to some important exponential family distributions such as the binomial. The reasoning of these limitations are then presented. This chapter is based on an extended and modified version of the paper:

> Saefken, Kneib, van Waveren and Greven (2014). A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models. *Electronic Journal of Statistics* **8**, 201-225.

The **second** research question this thesis is concerned with is:

How can the conditional prediction error be measured and what possibilities are there for estimating this?

Different prediction error measures are defined in the second chapter. Estimation of the prediction error is directly linked to covariance penalties introduced in Efron (2004). A conditional version of these is presented and estimation techniques are discussed and their behaviour is analysed in a simulation study.

In semiparametric regression and hierarchical modelling the parameters of interest can be split up into primary (regression) parameters and secondary (variance or smoothing) parameters.

The **third** research question examined in this thesis refers to following:

In what way do the secondary parameters effect the degrees of freedom of a model?

The development of the degrees of freedom of the secondary parameters and their analysis can, from a methodological perspective, be seen as the most challenging part of this thesis. The proposed approach is applicable to very general estimation methods for the secondary parameters and has interesting geometrical properties. An investigation of these geometrical properties is possible with the help of methods from differential geometry. The findings and the importance of considering the degrees of freedom of the secondary parameters are demonstrated in a simulation study.

The aim of the **fourth** chapter is as follows:

A broad introduction into the computational aspects of the estimation of the conditional AIC for different response distributions.

Implementations presented in the fourth chapter are available as R-package `cAIC4` on CRAN. The methods are explained based on the R implementations and accompanied by real data examples. The following paper has evolved on basis of this chapter:

> Saefken, Ruegamer, Greven and Kneib (2015). Conditional model selection in mixed-effect models with cAIC4. *Working paper.*

Chapter 1

Kullback-Leibler distance in exponential families

1.1 Bias correction for the conditional AIC in GLMMs

1.1.1 Generalized linear mixed models

Consider a Generalized linear mixed model (GLMM) with predictor

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}$$

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

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

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

$$\log\left(f(y_i|\boldsymbol{\beta}, \boldsymbol{u})\right) = \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \tag{1.1.1}$$

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

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

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

1.1.2 Akaike information criterion

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

$$AIC = -2\log\left(f(\boldsymbol{y}|\hat{\gamma}(\boldsymbol{y}))\right) + 2\nu \tag{1.1.2}$$

with log likelihood $(\log f(\cdot|\hat{\gamma}(\boldsymbol{y})))$ and $\nu = dim(\boldsymbol{\gamma})$ is an asymptotically unbiased estimator for the Akaike information. A direct extension of the AIC to GLMMs based on the marginal model would be the marginal AIC,

$$mAIC = -2\log\left(f(\boldsymbol{y}|\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\vartheta}})\right) + 2(p+q) \tag{1.1.3}$$

where $f(\boldsymbol{y}|\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\vartheta}})$ is the maximized marginal likelihood. If the dispersion parameter ϕ is estimated, the bias correction in (1.1.3) changes to $2(p+q+1)$. Using the marginal model implies that the focus is on the fixed effects and that new data $\boldsymbol{z}$ does not share the random effects of $\boldsymbol{y}$. However, the marginal AIC may be inappropriate for variable selection in linear mixed effect models if the focus is on clusters rather than on the population, as stated in Vaida and Blanchard (2005). Even under the marginal model it is not an (asymptotically) unbiased estimator of the Akaike information as shown for the Gaussian case by Greven and Kneib (2010).

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

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

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

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

where

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

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

$$2\Phi_0(\boldsymbol{y}) = 2\sum_{i=1}^{n}\frac{\partial \hat{y}_i}{\partial y_i} = 2\text{tr}\left(\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{y}}\right). \tag{1.1.4}$$

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

1.1.3 Bias correction

For GLMMs with responses following an exponential family distribution as in (1.1.1) and unknown random effects variance parameters $\boldsymbol{\vartheta}$, we derive the following bias correction:

Proposition 1.1.1. *In GLMMs with responses following an exponential family distribution and unknown* $\boldsymbol{\vartheta}$*, the bias correction for* $-2\log f(\boldsymbol{y}|\hat{\boldsymbol{\beta}},\hat{\boldsymbol{u}})$ *as an estimator of cAI is*

$$\begin{aligned} 2\Psi &= 2\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{y},\boldsymbol{u}}\left[\frac{y_i-\mu_i}{\phi}\hat{\theta}_i(\boldsymbol{y})\right] \\ &= \frac{2}{\phi}\sum_{i=1}^{n}\left\{\mathbb{E}_{\boldsymbol{y},\boldsymbol{u}}\left[y_i\hat{\theta}_i(\boldsymbol{y})\right]-\mu_i\mathbb{E}_{\boldsymbol{y},\boldsymbol{u}}\left[\hat{\theta}_i(\boldsymbol{y})\right]\right\}. \qquad (1.1.5)\end{aligned}$$

If ϕ is estimated, ϕ in the first expression is replaced by $\hat{\phi}$.

Proof of Proposition 1.1.1. The conditional log-likelihood is then given as

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

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

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

Now the bias correction can be calculated by

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

1.2 Stein's method for exponential families

1.2.1 Continuous distributions

The proposed bias correction in (1.1.5) suffers from the use of the true but unknown mean μ and therefore cannot be applied directly. Liang et al. (2008) solved this problem by using a formula known from Stein (1972) which turns (1.1.5) into (1.1.4). The following result extends the idea of Stein to continuous exponential family distributions and is a slight modification of Hudson (1978).

Theorem 1.2.1. *Let y be continuous and have density given by (1.1.1). For a differentiable function $m : \mathbb{R} \to \mathbb{R}$ that vanishes on the limits of the support of y if the limits of the support are finite and $\mathbb{E}[|m'(y)|] < \infty$ if the limits are infinite, it holds that*

$$\mathbb{E}\Big[m'(y)\Big] = \mathbb{E}\left[-\left(\frac{\theta}{\phi} + \frac{\partial}{\partial y} c(y,\phi)\right) m(y)\right]. \tag{1.2.1}$$

Proof. Notice first that for a density from an exponential family like (1.1.1)

$$f'(x) = f(x) \cdot (\frac{\partial}{\partial x} c(x,\phi) + \frac{\theta}{\phi}) \tag{1.2.2}$$

holds. Since $\lim_{x\to a} m(x) = \lim_{x \to b} m(x) = 0$, $\lim_{x\to a} f(x) < \infty$ and $\lim_{x\to b} f(x) < \infty$ we have

$$\begin{aligned}
\mathbb{E}\Big[m'(X)\Big] &= \int_a^b m'(x) f(x) dx \\
&= \int_a^b m'(x) \exp\left(\frac{\theta x - b(\theta)}{\phi} + c(x,\phi)\right) dx \\
&= [m(x)f(x)]_a^b - \int_a^b f'(x) m(x) dx \\
&= -\int_a^b f(x) \cdot (\frac{\partial}{\partial x} c(x,\phi) + \frac{\theta}{\phi}) m(x) dx \\
&= \mathbb{E}\left[\left(-\frac{\partial}{\partial x} c(x,\phi) - \frac{\theta}{\phi}\right) m(X)\right].
\end{aligned}$$

The third equality is obtained on integration by parts. Since $m(x)f(x)$ vanishes when $x \to a, b$, the fourth equality holds. □

If y is Gaussian, formula (1.2.1) simplifies to

$$\mathbb{E}\left[m'(y)\right] = \mathbb{E}\left(\frac{y-\mu}{\sigma^2}m(y)\right), \tag{1.2.3}$$

the formula known from Stein. Applied to the bias correction (1.1.5) this yields the bias correction $2\Phi_0$ known from Liang et al. (2008). The theorem can also be applied to obtain bias corrections for other exponential family distributions as stated in the following. For y exponentially distributed with mean μ, $y \sim \mathcal{E}(\frac{1}{\mu})$, and letting $m(y) = \int_0^y g(x)\ dx$, equation (1.2.1) becomes

$$\mu\mathbb{E}\left[g(y)\right] = E\left[\int_0^y g(x)dx\right]. \tag{1.2.4}$$

We use this equation to derive an analytically accessible version of (1.1.5).

Corollary 1.2.1. *Let $y_i|\boldsymbol{u} \sim \mathcal{E}(\frac{1}{\mu_i})$. Then an unbiased estimator of the cAI is*

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

with

$$\Psi = \sum_{i=1}^{n} y_i\hat{\theta}_i(\boldsymbol{y}) - \int_0^{y_i} \hat{\theta}_i(\boldsymbol{y}_{-i}, x)dx \tag{1.2.5}$$

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

Proof of Corollary 1.2.1. Let $y_i|\boldsymbol{u} \sim \mathcal{E}(\frac{1}{\mu_i})$, then we can rewrite the bias correction

(1.1.5) with the help of equation (1.2.4):

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

where $\boldsymbol{y}_{-i}$ is the vector of observed responses without the $i-$th observation. □

1.2.2 Discrete distributions

A similar identity to Theorem 1.2.1 also holds for discrete random variables from an exponential family distribution. The following theorem is due to Hudson (1978).

Theorem 1.2.2. *Let y be a discrete random variable taking values in $\mathbb{N}_0=\{0,1,2,\ldots\}$ and let y have probability function given by (1.1.1). For $m:\mathbb{N}\to\mathbb{R}$ with $\mathbb{E}[|m(y)|]<\infty$ it holds that*

$$\exp(\theta)\mathbb{E}\left(m(y)\right)=\mathbb{E}\left[t(y)m(y-1)\right] \qquad (1.2.6)$$

where

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

Proof. Let $h:\mathbb{N}\to\mathbb{R}$ satisfy $\mathbb{E}\left[|h(X)|\right]<\infty$. Then for $\zeta=\exp(\theta)$ we can write the probability function of X as

$$\mathbb{P}(X=n)=p(n)=\zeta^n\exp(-b(\log(\zeta)))\exp(c(n)).$$

The calculation is now straitforward

$$\begin{aligned}
\zeta\mathbb{E}\left[h(X)\right] &= \zeta\sum_{n=0}^{\infty} h(n)p(n) \\
&= \zeta\sum_{n=0}^{\infty} h(n)\zeta^{n}\exp(b(\log(\zeta))\exp(c(n)) \\
&= \sum_{n=0}^{\infty} h(n)\zeta^{n+1}\exp(b(\log(\zeta))\exp(c(n)) \\
&= \sum_{m=1}^{\infty} h(m-1)\zeta^{m}\exp(b(\log(\zeta)))\exp(c(m-1)) \\
&= \sum_{m=1}^{\infty} h(m-1)\zeta^{m}\exp(b(\log(\zeta)))t(m)\exp(c(m)) \\
&= \sum_{m=0}^{\infty} h(m-1)\zeta^{m}\exp(b(\log(\zeta)))t(m)\exp(c(m)) \\
&= \mathbb{E}\left[t(X)h(X-1)\right]
\end{aligned}$$

The last but one equation follows because $t(0) = 0$. □

For y Poisson distributed with parameter λ, $y \sim \mathcal{P}(\lambda)$, equation (1.2.6) simplifies to

$$\lambda\mathbb{E}\left[m(y)\right] = \mathbb{E}\left[ym(y-1)\right], \tag{1.2.7}$$

with $ym(y-1) = 0$ if $y = 0$ by convention. This is an identity due to Chen (1975). With the help of this identity the bias correction (1.1.5) can be made analytically accessible.

Corollary 1.2.2. *Let $y_i|\boldsymbol{u} \sim \mathcal{P}(\lambda_i)$. Then an unbiased estimator of the cAI is*

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

with

$$\Psi = \sum_{i=1}^{n} y_i\left(\hat{\theta}_i(\boldsymbol{y}) - \hat{\theta}_i(\boldsymbol{y}_{-i}, y_i - 1)\right), \tag{1.2.8}$$

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

Corollary 1.2.2 gives an alternative derivation of the result in Lian (2012), which highlights the connection to the normal case.

Proof of Corollary 1.2.2. If $y_i|\boldsymbol{u} \sim \mathcal{P}(\lambda_i)$ then equation (1.1.5) becomes with the help of equation (1.2.7):

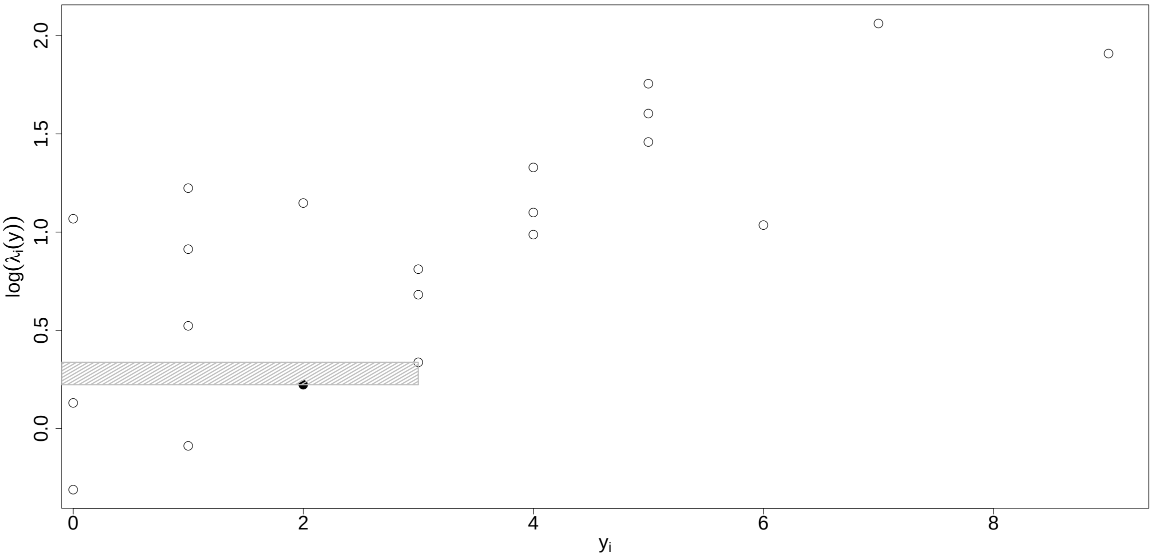

Figure 1.1: The plot shows how the bias correction in (1.2.8) is obtained.

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

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

1.3 Limits of the approach

Theorem 1.2.1 and Theorem 1.2.2 can be extended to further distributions. For instance the generalized SURE formula (Lemma 2) in Shen and Huang (2006) is a generalisation of Theorem 1.2.1 and Theorem 1.2.2 to distributions not necessarily from the exponential family. Although the formula has been obtained in a different context, it is closely related to the findings in Section 1.2 and gives further insight on how identities for further distributions could potentially be derived. On the other hand, formulas as in Theorems 1.2.1 and 1.2.2 do not necessarily lead to bias correction terms computable from observable quantities for all distributions, as discussed below.

1.3.1 Continuous distributions

For example, if y follows a gamma distribution with mean μ and scale parameter ν, i.e. $y \sim \mathcal{G}(\mu,\nu)$, identity (1.2.1) is

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

This can be rewritten in terms of μ

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

In contrast to the $\nu = 1$ case, this identity cannot be used to remove the true but unknown parameter μ_i in the bias correction term (1.1.5) unless we could rewrite the estimator of the canonical parameter in (1.1.5) by

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

for a function $m(\cdot)$ fulfilling the requirements in Theorem 1.2.1. Since this does not seem possible, Theorem 1.2.1 cannot be used to rewrite the bias correction term (1.1.5) for a gamma distribution with $\nu \neq 1$.

1.3.2 Discrete distributions

Similarly, applying Theorem 1.2.2 to the negative binomial distribution where y has the probability function

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

identity (1.2.1) becomes

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

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

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

The second part of the bias correction (1.1.5), i.e. $\mu_i\mathbb{E}_{y,\boldsymbol{u}}(\hat{\theta}_i(y))$, could therefore only be replaced if the estimator for the canonical parameter $\hat{\theta}_i(\cdot)$ can be written as

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

for some arbitrary function $m(\cdot)$ as in Theorem 1.2.2. This is not possible.

Theorem 1.2.2 cannot be applied to the binomial distribution $\mathcal{B}(n,p)$ since a binomially distributed random variable only takes values in $\{0,1,\ldots,n\} \subset \mathbb{N}_0$. Extending the distribution by defining $P(y=n+k)=0\ \forall k \in \mathbb{N}$ does not yield an identity which could be applied to the bias correction (1.1.5), for the same reason as in the case of the negative binomial distribution.

1.4 Simulation study

In the first part of this simulation study, we concentrate on random intercept models. The bias corrections (1.2.5) and (1.2.8) are analysed in two different ways. First, we compare the precision and the variability of different bias corrections as estimators of the correction term obtained by estimating the relative Kullback-Leibler distance with the log-likelihood. In a second step, the model choice behaviour of the bias correction for exponential responses (1.2.5) and Poisson distributed responses (1.2.8) is assessed.

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

1.4.1 Random intercept model

Exponential distribution

First, we will focus on the precision and the variability of the proposed bias correction (1.2.5). We therefore consider a model with an exponentially distributed response y_{ij} and a random intercept u_i with

$$\mu_{ij} = \exp(\beta_0 + \beta_1 x_j + u_i);\ i = 1, \ldots, m;\ j = 1, \ldots, n_i, \tag{1.4.1}$$

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

We compare the proposed estimator for the bias correction Ψ obtained from refitting the model for each i with the true bias BC defined by (1.1.5), the asymptotically unbiased estimator $\hat{\rho}_{ml}$ proposed by Yu and Yau (2012) and the estimator $\hat{\rho}_{Don}$ of Donohue et al. (2011). The true bias correction BC is derived by averaging 30,000 samples of (1.1.5) based on model (1.4.1). This criterion used as a benchmark is not available in practice since for its calculation the true mean μ has to be known.

For the proposed bias correction Ψ as in (1.2.5), an integral needs to be evaluated. Since this can not be done analytically, it is approximated by adaptive quadrature. The resulting bias correction is used to obtain the proposed cAIC.

The cAIC suggested by Yu and Yau (2012) is included to assess the performance of an asymptotically unbiased estimator of the cAI in finite sample settings. Similarly to the cAIC suggested by Vaida and Blanchard (2005) for Gaussian responses, the cAIC

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

Table 1.1: Mean estimated values of four different estimators of the bias correction (1.1.5) and the corresponding standard deviations (indicated by σ with corresponding index) of model (1.4.1) for different cluster sizes (n_i), number of clusters (m) and variances of random effects (τ^2). The true bias correction BC is derived by (1.1.5), the estimator Ψ is directly calculated by (1.2.5), $\hat{\rho}_{ml}$ is the estimator proposed by Yu and Yau (2012) and $\hat{\rho}_{Don}$ is the estimator proposed by Donohue et al. (2011)

proposed by Donohue et al. (2011) requires known random effects variance parameters. For known random effects variance parameters, the criterion is consistent. In our simulated random intercept model, τ^2 would need to be known. Since in many applications this will not be the case, we use the proposed bias correction of Donohue et al. (2011) with the estimated variance parameter $\hat{\tau}^2$ assumed true.

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

Table 1.1 shows the results. They suggest that, that the proposed estimator performs well although numerical integration was used. The estimator $\hat{\rho}_{ml}$ has the tendency to underestimate the true bias correction for positive true τ^2 and to overestimate it for true $\tau^2 = 0$. This may be due to the fact that a non-canonical link function was used, while the authors derive their estimator only for canonical links. Furthermore the authors do not use PQL as fitting method, see 1.4.3 for a short remark.

The estimator $\hat{\rho}_{Don}$ consistently underestimates BC, as it ignores variability due to

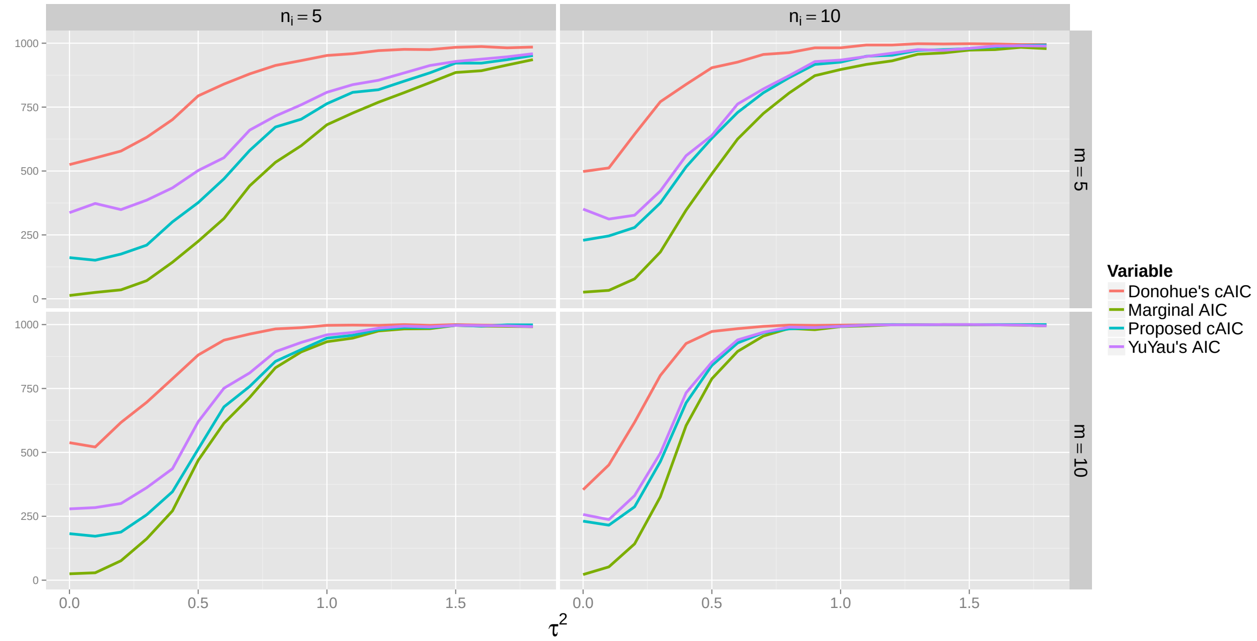

Figure 1.2: Results for the random intercept model with exponentially distributed responses. The y-axis shows the number of simulation replications out of 1000 where the more complex model was favoured by the different AICs.

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

We now consider the behaviour of the proposed bias correction (1.2.5) when selecting random effects. Therefore, consider the same settings as in model (1.4.1) but with the random effect variances as $\tau^2 = 0,\ 0.1,\ 0.2, \dots,\ 1.8$, respectively. For each of the settings, 1,000 data sets are generated and one model containing a random intercept ($\tau^2 \geq 0$) and another (generalized linear) model without random effects are fitted to each data set. The random effects model is fitted by PQL, see Breslow and Clayton (1993) and Wood (2006).

We compute the frequency of selecting the model including the random intercept ($\tau^2 > 0$), which is chosen whenever the proposed AIC is smaller than an AIC derived from the model without a random intercept ($\tau^2 = 0$). As reference AICs for the model without random intercept we use (1.1.2) for the marginal AIC, Donohue's cAIC and Yu & Yau's cAIC. For the proposed cAIC we use formula (1.2.5) with a generalized linear model as reference. Thus, for each AIC we use as a reference the AIC it reduces to in the null model without intercept.

The marginal AIC as defined in (1.1.3) requires the marginal log-likelihood, which is obtained by Laplace approximation. The results for different settings and AICs are displayed in Figure 1.2.

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

Poisson distribution

Investigating the precision and variability of the bias correction (1.2.8), we consider a random intercept model with Poisson distributed responses and subject specific random intercept, $y_{ij}|u_i \sim \mathcal{P}(\lambda_{ij})$. A logarithmic link function is used

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 x_j + u_i;\ i = 1, \dots, m;\ j = 1, \dots, n_i, \tag{1.4.2}$$

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

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

Table 1.2: Mean estimated values of four different estimators of the bias correction (1.1.5) and the corresponding standard deviations (indicated by σ with corresponding index) of model (1.4.2) for different cluster sizes (n_i), number of clusters (m) and variances of random effects (τ^2). The true bias correction BC is derived by (1.1.5), the estimator Ψ is directly calculated by (1.2.8), $\hat{\rho}_{ml}$ is the estimator proposed by Yu and Yau (2012) and $\hat{\rho}_{Don}$ is the estimator proposed by Donohue et al. (2011)

The results are shown in Table 1.4.1. The proposed estimator Ψ combines high precision with low variance. Compared to the estimates with exponentially distributed responses, $\hat{\rho}_{ml}$ performs well although it shows a tendency towards overestimation and has high variances especially for a larger number of small clusters. The estimator $\hat{\rho}_{Don}$ underestimates the true bias correction as it did in the previous setting.

As for the simulation study with exponentially distributed responses, we also assess the model choice behaviour of bias correction (1.2.8). The settings are the same as in model (1.4.2) except the random effects variance, that is $\tau^2 = 0,\ldots,0.8$. Then 1,000 data sets for each setting are generated. Two models are applied to the data, one model containing a random intercept ($\tau^2 \geq 0$) and another model without random effects ($\tau^2 = 0$). The frequency of selecting the more complex model, including the random effects is computed for different AICs. Just as for the exponential responses, PQL was used as model fitting method. The different proposed AICs are the same as in the exponential model (1.4.1): 1) the proposed bias correction Ψ as in (1.2.8); 2) the cAIC suggested by Yu and Yau (2012); 3) the cAIC proposed by Donohue et al. (2011), with the estimated variance parameter $\hat{\tau}^2$ plugged in as true τ^2; 4) the marginal AIC

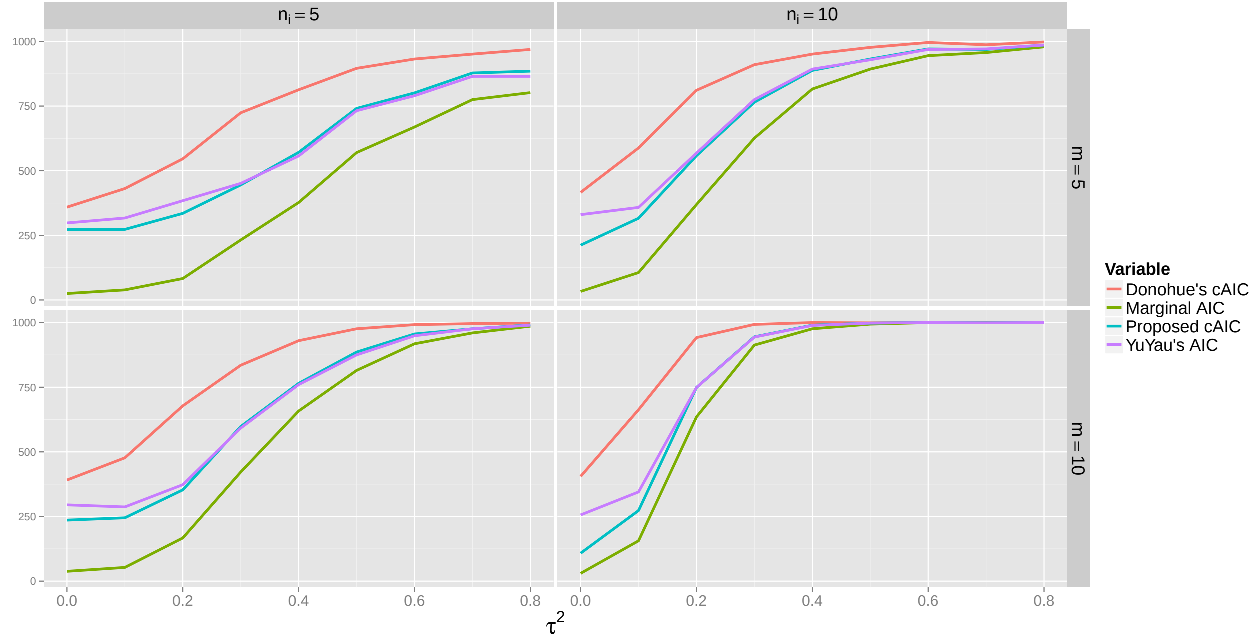

Figure 1.3: Results for the random intercept model with Poisson distributed responses. The y-axis shows the number of simulation replications out of 1,000 where the more complex model was favoured by the different AICs.

as defined in (1.1.3), which is obtained by Laplace approximation.

The results are displayed in Figure 1.3. They are similar to the results observed for exponential responses. The marginal AIC chooses the model including the random effects only very rarely even for random effects variances larger than zero. On the other hand, the AIC proposed by Donohue et al. (2011) chooses to include random effects very often, even if the model was simulated without random effects. The proposed criterion and Yu and Yau's asymptotic critierion behave similar, with a stronger preference for the larger model when the variance is zero or small for Yu and Yau's AIC. The asymptotically unbiased criterion proposed by Yu and Yau (2012) behaves as expected. For larger cluster sizes and increasing number of clusters the model choice behaviour gets better.

1.4.2 Penalized spline smoothing

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

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

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

In this setting, we choose the smooth function to be

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

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

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

The smooth function is estimated by a penalized spline

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

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

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

Table 1.3: Mean estimated values of four different estimators of the bias correction (1.1.5) and the corresponding standard deviations (indicated by σ with corresponding index) of model (1.4.3) for different sample sizes n and different degrees of nonlinearity d. The estimator Ψ is directly calculated by (1.2.5), BC is derived by (1.1.5), $\hat{\rho}_{ml}$ is the estimator proposed by Yu and Yau (2012) and $\hat{\rho}_{Don}$ is the estimator proposed by Donohue et al. (2011)

Exponential distribution

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

$$\log(\mu_i) = 1 + x_i + d\left(\frac{1}{3} - x_i\right)^2, \ i = 1, \ldots, n. \tag{1.4.3}$$

For nonlinearity parameters $d = 0, 0.5, 1$ the averaged estimated bias corrections and corresponding standard deviations are derived from 1,000 data sets simulated from model (1.4.3).

The results in Table 1.4.2 indicate that the bias correction Ψ in (1.2.5) and BC in (1.1.5) have the same expected value, as was shown analytically in Corollary (1.2.1). The high variance $\sigma_{\hat{\rho}_{ml}}$ of the estimator proposed by Yu and Yau (2012) is due to outliers that occur, caused by numerical instability. The estimator $\hat{\rho}_{Don}$ does not change a lot for differing levels of nonlinearity and underestimates the bias correction term.

The model choice behaviour of the same criteria as in (1.4.1) is assessed in the same way as in the random intercept model. For each setting and each value of nonlinearity $d = 0,\ 0.25,\ 0.5,\ 0.75,\ 1,\ 1.5,\ 2,\ 2.5,\ 3,\ 3.5$ and 4, 1,000 data sets are generated, and a linear and a nonlinear model are fitted to the data. The frequency of selecting the more complex, nonlinear model for each criterion is computed.

Figure 1.4 shows the results. The marginal AIC behaves as expected and chooses the nonlinear model only very rarely. The proposed cAIC based on the bias correction (1.2.5) shows similar behaviour to the other settings. For increasing sample size, Yu and Yau (2012) show an unexpected behaviour. The cAIC by Yu and Yau (2012) selects the nonlinear model with a proportion increasing with sample size, even for zero or small variances τ^2, and for the largest sample size more often than the cAIC proposed by Donohue et al. (2011). Since this behaviour seems to contradict the findings of Yu and Yau (2012), a short discussion is given in 1.4.3.

For Poisson distributed responses $y_i \sim \mathcal{P}(\mu_i)$, model (1.4.3) stays the same but, due to a different signal-to-noise ratio, we choose a different sequence of nonlinearity parameters. In order to compare the precision and variability of the different bias corrections, we choose the nonlinearity parameter $d = 0$, 0.8, 1.6. For each level of nonlinearity and for the sample sizes $n = 25$, 50, 75, 100 the estimated bias corrections are listed in Table 1.4.2. The results show that the proposed estimator Ψ is close to the bias correction BC

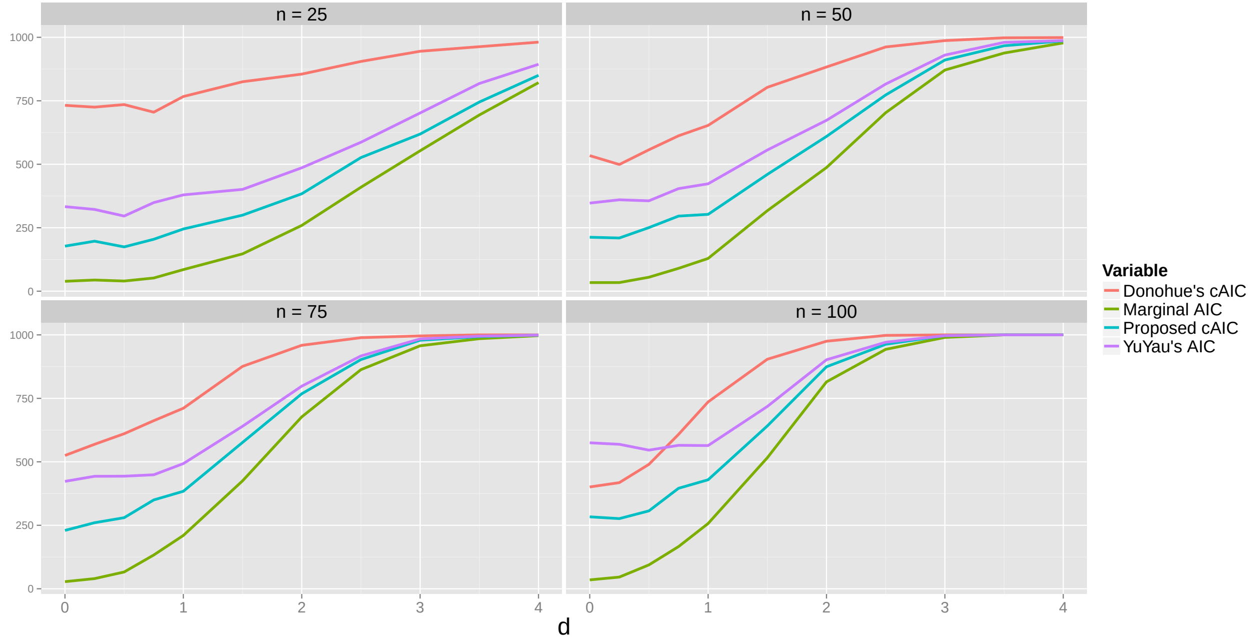

Figure 1.4: Results for the spline smoothing model with exponentially distributed responses. The y-axis shows the number of simulation replications where out of 1,000 the more complex model was favoured by the different AICs.

Poisson distribution

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

Table 1.4: Mean estimated values of four different estimators of the bias correction (1.1.5) and the corresponding standard deviations (indicated by σ with corresponding index) of model (1.4.3) with Poisson distributed responses for different sample sizes n and different degrees of nonlinearity d. The estimator Ψ is directly calculated by (1.2.8), BC is derived by (1.1.5), $\hat{\rho}_{ml}$ is the estimator proposed by Yu and Yau (2012) and $\hat{\rho}_{Don}$ is the estimator proposed by Donohue et al. (2011)

derived by 30,000 times reestimating model (1.4.3) with Poisson distributed responses and calculating 1.1.5. The BC bias correction is not applicable in practice since the true unknown mean μ has to be known for its calculation. The high variance of the estimator $\hat{\rho}_{ml}$, proposed by Yu and Yau (2012) indicates some very large values which seem to be due to numerical instabilities.

The selection frequencies are derived for nonlinearity levels $d =$ 0, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6. They are shown in Figure 1.5. They behave similar to the ones observed for the smoothing spline model with exponentially distributed responses. The unexpected behaviour of the cAICs proposed by Yu and Yau (2012) and Donohue et al. (2011) are not as pronounced as for exponentially distributed responses. Nevertheless, the bias correction of Yu and Yau (2012) is occasionally smaller than the one proposed by Donohue et al. (2011) in this setting as well.

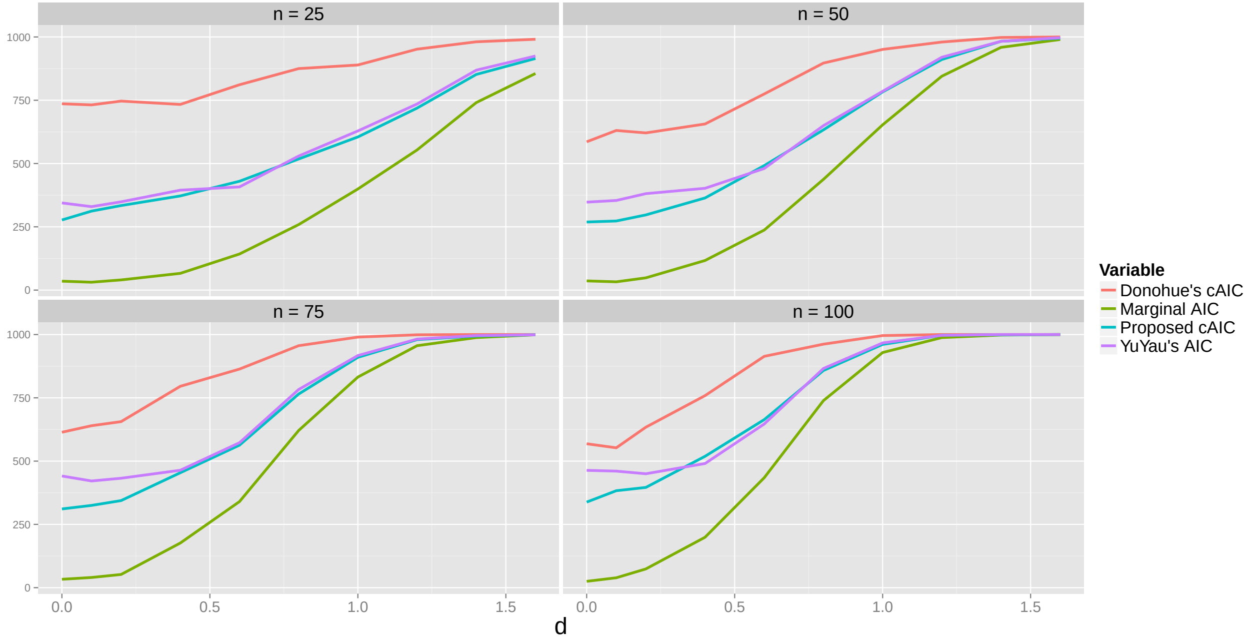

Figure 1.5: Results for the spline smoothing model with Poisson distributed responses. The y-axis shows the number of simulation replications out of 1000 where the more complex model was favoured by the different AICs.

1.4.3 General remarks

The cAIC proposed by Yu and Yau (2012) is used here as an ad-hoc criterion since it is one of the few available benchmarks for model selection in generalized linear mixed models. The criterion was derived for ML estimation of the variance parameters based on McGilchrist (1994), while our models were fitted using the REML based PQL method proposed by Breslow and Clayton (1993). Despite the difference between REML and ML, the two approaches are similar to each other in maximizing the joint likelihood of y and $\boldsymbol{u}$ as mentioned by McGilchrist (1994). However, the main objection to the application of the cAIC proposed by Yu and Yau (2012) may be that the models (1.4.3) and (1.4.1) have a non-canonical link although the criterion of Yu and Yau (2012) requires canonical link functions.

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

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

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

A disadvantage regarding the proposed estimator (1.2.5) when using numeric integration is that for each datum the integral needs to be calculated. Therefore, if for one i in (1.2.5) the integral can not be calculated the bias correction can not be obtained. This may be a problem particularly in large data sets and for instance, if there is collinearity in the data.

The implementation of the proposed method to derive (1.2.5) based on numerical integration takes 330 s for model (1.4.1) with random-effects variance 1 and five clusters with five observations each on a 2.80-GHz personal computer. The computational cost depends on how precise the numerical integration is and on the size of the data set.

For data from model (1.4.2) with random effects variance 1 and five clusters with five observations each it takes about 3 s to compute (1.2.8) on a 2.80-GHz personal computer. This leave-one-out implementation is increasingly time-consuming for larger data sets and less time-consuming if there are many zeros in the observed responses.

1.5 Example: Modelling tree growth with water availability

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

In this case study, we show how the proposed estimator of the Kullback-Leibler distance for exponential responses influences the selection of models for tree growth. The study is based on a sub-sample of 2655 trees, from a 28.5 ha large area that is located in the core zone of the Hainich National Park, Thuringia, Germany. A map of the study area including the tree types is given in Figure 1.6. The National Park is part of Germany's largest continuous broad-leaved forest. To estimate tree growth, in 1999 and 2007 for each tree within the study area the Diameter at Breast Height (DBH), i.e. at about 1.30 m, was mapped, see Butler Manning (2007). The difference in diameter is the dependent variable growth. We only consider beech, which accounted for 90 % of the recorded trees. We included only trees with 10–30 cm DBH because they can be reasonably assumed to have completed the phase of highest mortality due to competition (self-thinning), without yet reproducing themselves. Furthermore, we excluded trees for which no positive growth was recorded as these measurements seem to be erroneous.

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

Water availability is a good proxy for abiotic resource availability on rich soils, because water availability, apart from light, mainly limits tree-growth influencing the predominance of beech. To estimate spatial variation in water availability due to soil properties, we use the soil depth (sd_i) as covariate. The soil depth was measured on a $50\text{m} \times 50\text{m}$ grid on the study area. From these observations a smooth plane was estimated and predictions for each tree derived. A contour plot of the soil depth is given in Figure 1.7 and corresponding the smooth surface is plotted in Figure 1.8. A second available covariate, the Topographic Wetness Index (twi_i), is calculated from a Digital

Elevation Model and measures water availability determined by the topography, see Böhner, McCloy, and Strobl (2006). The tree-specific predictions of the Topographic Wetness Index were also derived by an two-dimensional thin plate regression spline, see Wood (2006).

Our aim is to find a model that best describes the tree growth with the help of the given covariates. Hence, we choose the model with the lowest estimated Kullback-Leibler distance from a set of candidate models. We concentrate on the selection of linear versus nonlinear modelling of the continuous covariates. This corresponds to the selection of random effects in the mixed model framework. We model the DBH difference using an exponential distribution, as using a gamma distribution resulted in a dispersion parameter estimate of 0.98 for model 1.5.1 that is very close to one.

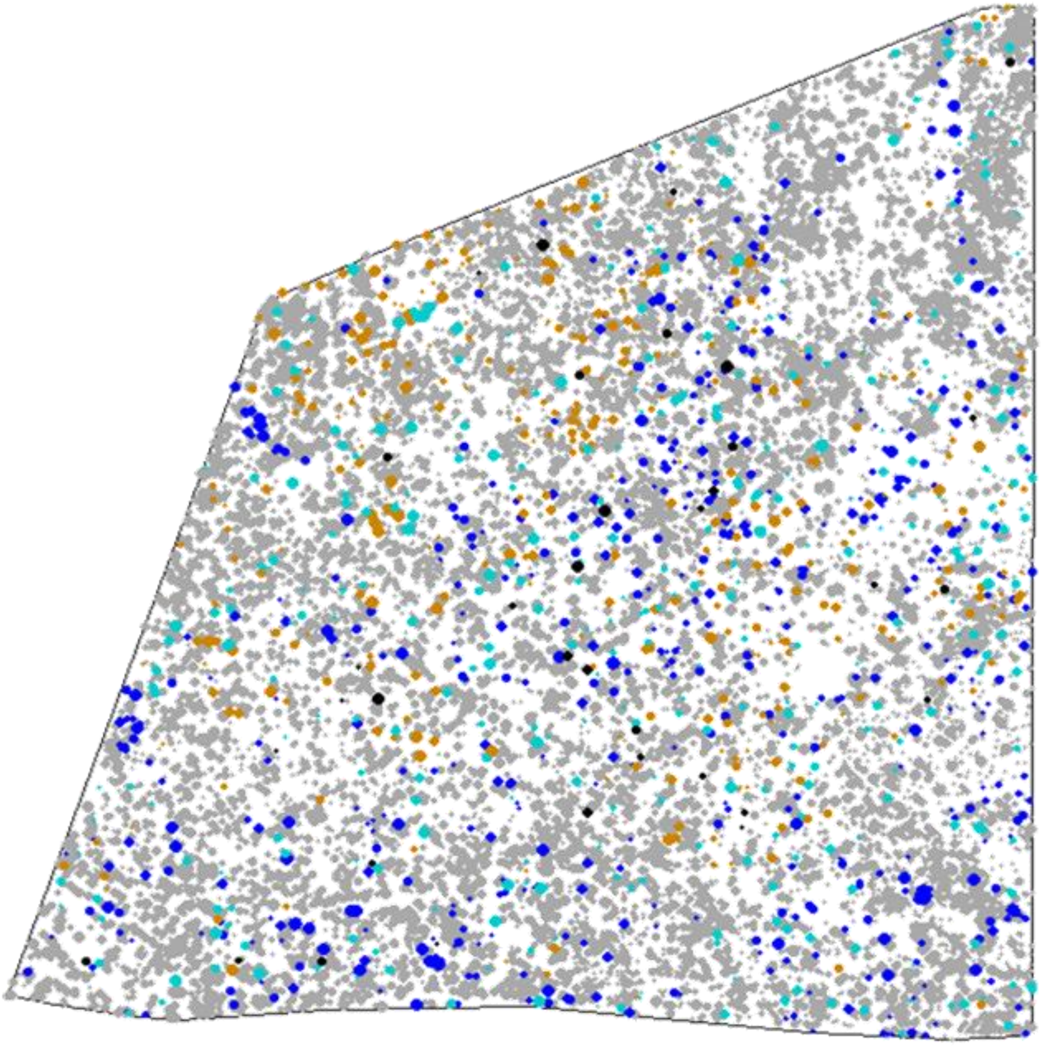

Figure 1.6: The study area with european beech (grey) and ash (blue) and further tree types. Further tree types are hornbeam (brown) and norway maple (light blue).

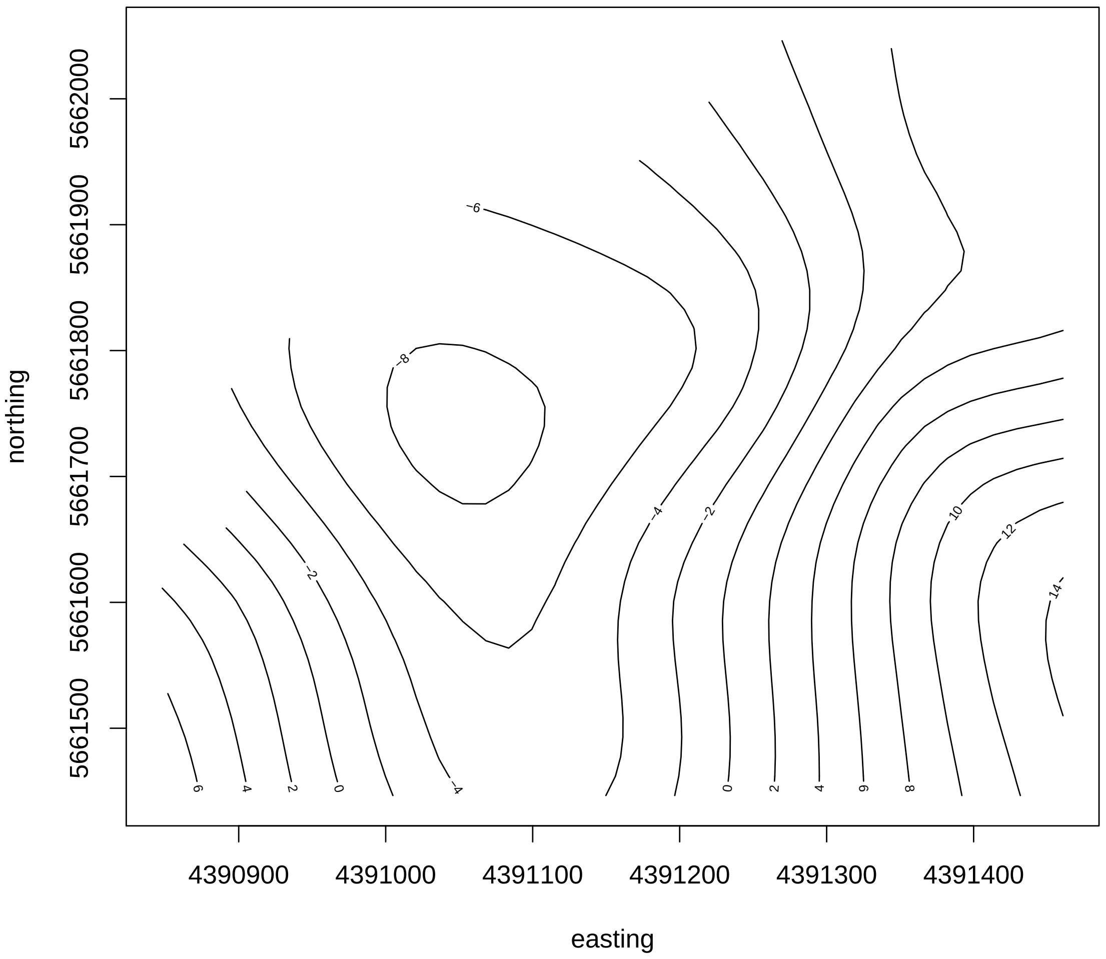

Figure 1.7: A contour plot of the smooth effect of the soil depth within the study area modelled with a thin plate regression spline. Used for the prediction of the soil depth at each tree.

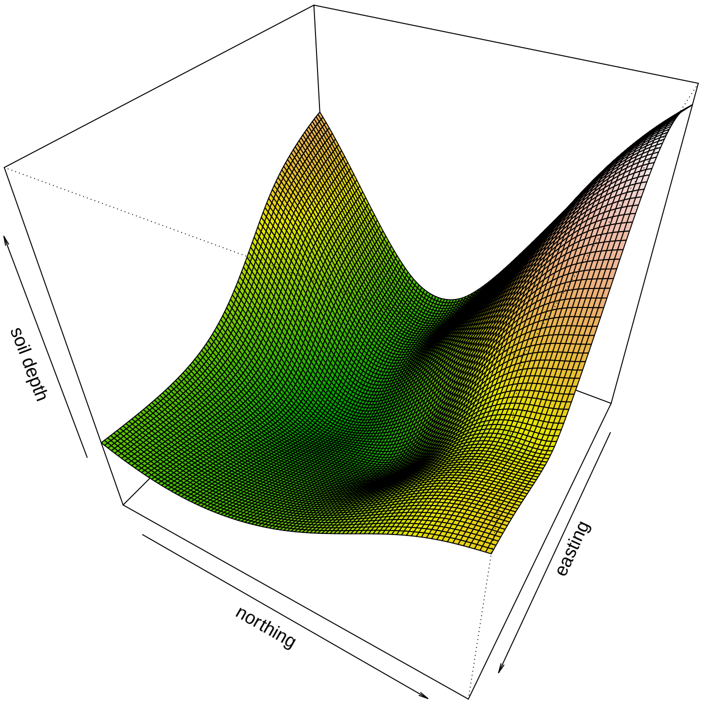

Figure 1.8: A perspective plot of the smooth soil depth within the study area.

1.5.1 Univariate smooth function

In order to investigate the model choice behaviour of the mAIC and the proposed AIC with bias correction (1.2.5) in a simple model, we consider a univariate smoothing example, based on the tree growth data. We estimate the effect of soil depth on the tree growth and include the KRAFT-class to account for differing growth potentials due to light availability. For the mean of the tree growth μ, we obtain the following model:

$$\log(\mu_i) = \beta_0 + \beta_1 k_i + m(sd_i), \; i = 1, \ldots, 2655, \tag{1.5.1}$$

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

We distinguish between a linear model (M1) in which the function $m(\cdot)$ is a linear function and a semiparametric model (M2) with nonlinear function $m(\cdot)$. The semiparametric model is fit by PQL. Both estimated functions are plotted in Figure 1.9. The mAIC for the linear model (M1) is 6,258 and for the semiparametric model (M2) the mAIC is 6,276. The conditional AIC based on (1.2.5) for the linear model (M1) is 6,257 and for the semiparametric model (M2) it is 6,235. Therefore, the mAIC chooses the model (M1) with $m(\cdot)$ as linear function and the proposed conditional AIC chooses the model (M2) with $m(\cdot)$ as nonlinear function.

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

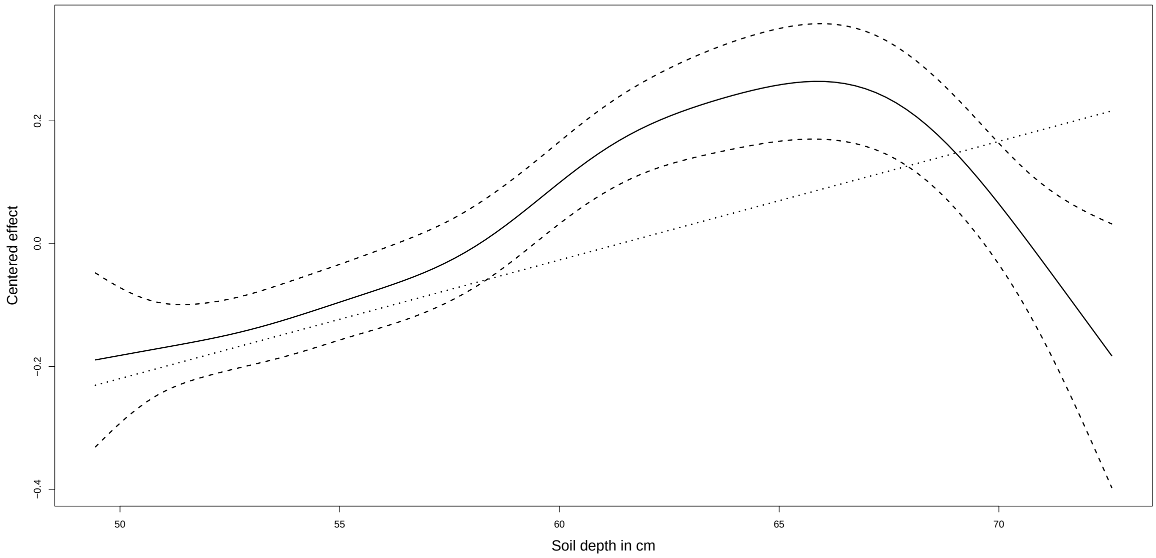

Figure 1.9: The linear effect (dotted line) and the nonlinear effect (solid line) with confidence interval (dashed lines) of the soil depth on the tree growth. The pointwise confidence intervals are defined using twice the standard deviation of the estimator.

1.5.2 Generalized additive model

In a more sophisticated approach, we consider a model incorporating possibly nonlinear effects of three covariates and one linear effect of the KRAFT-class k. Accordingly, we extend model (1.5.1) to a generalized additive model, see Wood (2006):

$$\log(\mu_i) = \beta_0 + \beta_1 k_i + m_1(sd_i) + m_2(twi_i) + m_3(nnd1_i) \tag{1.5.2}$$

or

$$\log(\mu_i) = \beta_0 + \beta_1 k_i + m_1(sd_i) + m_2(twi_i) + m_3(nnd2_i), \tag{1.5.3}$$

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

The functions $m_1, \ldots, m_3$ may either be linear or nonlinear functions. In the model selection process, we choose between the two possibilities for each of the three functions in the two models. In consequence we can choose from a set of 16 candidate models. We expect all covariates to have an effect on growth and therefore do not include models into the model selection process that completely omit one of the covariates, except $nnd1$ and $nnd2$ respectively. All possible models and the corresponding AIC values can be found in Table 1.5.2.

The model selection process is based on two criteria, the proposed conditional AIC with associated bias correction (1.2.5) and the conditional AIC proposed by Donohue et al. (2011). The marginal AIC is omitted because we cannot extract the design matrices $Z_i,\ i = 1, 2, 3$ corresponding to the random effects parametrization of the smoothing splines that are needed to derive the Laplace approximation. This problem does not occur in univariate smoothing models since there is no need to split up the design matrix Z corresponding to the random effect. The conditional AIC proposed by Yu and Yau (2012) could not be calculated due to the need for inverting matrices that are singular.

The two criteria both choose the model including the second nearest-neighbour distance with all three effects modelled as nonlinear functions. When comparing each specific model in order to assess whether to include the second or the first nearest-neighbour

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

Table 1.5: Estimated Kullback-Leibler distance for 16 models fitted to the tree growth data. The first four columns indicate if the effects of the covariates are modelled by linear ($-$) or nonlinear ($\sim$) functions, corresponding to the absence and presence of random effects. Two different estimators of the Kullback-Leibler distance are listed in the table: The AIC based on the bias correction 1.2.5 ($cAIC$) and the AIC proposed by Donohue et al. (2011) ($dAIC$)

distance, both criteria in each case favour the model with the second nearest-neighbour distance. For all models, except the two models only containing linear effects, the proposed conditional AIC is larger and therefore penalizes more than the conditional AIC proposed by Donohue et al. (2011). This confirms the behaviour observed in the simulation study, i.e. that the criterion by Donohue et al. (2011) underestimates the bias correction.

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

$$\log(\mu_i) = \beta_0 + \beta_1 k_i + m_1(sd_i) + \beta_2 twi_i + m_3(nnd1_i) \tag{1.5.4}$$

and

$$\log(\mu_i) = \beta_0 + \beta_1 k_i + \beta_2 sd_i + m_2(twi_i) + m_3(nnd2_i), \tag{1.5.5}$$

our proposed estimator chooses the first model (1.5.4), while the estimator proposed by Donohue et al. (2011) chooses the latter (1.5.5).

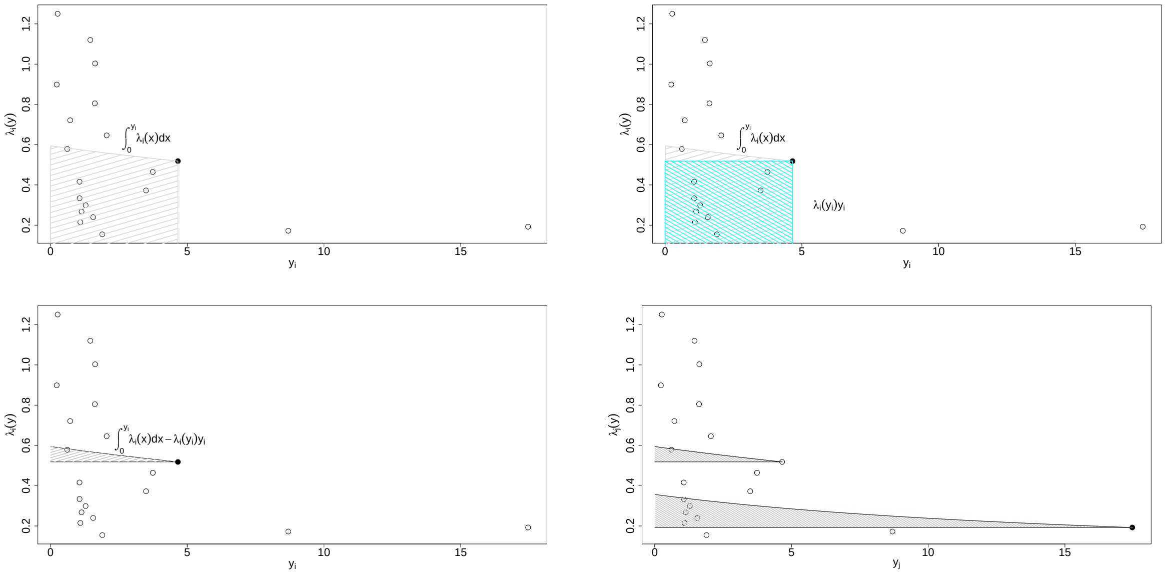

Figure 1.10: The four plots show how the bias correction in (1.2.5) is obtained.

1.6 A fast approximation of the exponential bias correction

The bias correction for exponentially distributed responses (1.2.5) is computationally inefficient. Its calculation involves numerically computing as many integrals as there are data points. Therefore, a fast approximation is desirable. Such an approximation is suggested in the following. First the bias correction (1.2.5) can be written in terms of the rate parameter λ of an exponential distribution as

$$\Psi = \sum_{i=1}^{n} \int_0^{y_i} \hat{\lambda}_i(y_{-i}, x)dx - y_i\hat{\lambda}_i(y). \tag{1.6.1}$$

The main idea is now to approximate the function $\hat{\lambda}_i(y_{-i}, x)$ in order to find a efficient way to compute the integral in (1.6.1). Since the expectation of the conditional distribution of $y|b$ is $\frac{1}{\lambda}$ and in case $\hat{\lambda}_i(y_{-i}, x)$ is a consistent estimator, we try to find

$$\hat{\lambda}_i(y_{-i}, x) \approx \frac{1}{x+a} + b \tag{1.6.2}$$

for some parameters a and b and for each data point $i = 1, \ldots, n$. Choosing this kind of functional form is also supported in figure (1.6).

Solving the integral in (1.6.1) with this approximations gives

$$\begin{aligned}\int_0^{y_i} \hat{\lambda}_i(y_{-i},x)dx &\approx \int_0^{y_i} \frac{1}{x+a}+bdx \\ &= \log|y_i+a|-\log|a|+by_i \end{aligned} \tag{1.6.3}$$

The value of the estimator function at point y_i is already known from the fitting process, i.e.

$$\hat{\lambda}_i(y_{-i},y_i)=\hat{\lambda}_i^1.$$

Furthermore, we can evaluate the value of the estimator function for some small value x_0 near 0, i.e.

$$\hat{\lambda}_i(y_{-i},x_0)\approx\hat{\lambda}_i(y_{-i},0)=\hat{\lambda}_i^0,$$

where x_0 should not be exactly zero since the estimator function will often not be able to handle zero values.

It is now possible to find unique values a,b such that the approximate function in (1.6.2) coincides with the true estimator function in the two points x_0 and y_i. In this case the parameters must fulfil the equations

$$\hat{\lambda}_i^0=\frac{1}{a}+b$$

and

$$\hat{\lambda}_i^1=\frac{1}{y_i+a}+b.$$

Solving these two equations w.r.t a and b gives

$$a=-\frac{y_i}{2}+\sqrt{\frac{y_i}{\hat{\lambda}_i^0-\hat{\lambda}_i^1}+\left(\frac{y_i}{2}\right)}$$

and

$$b=\hat{\lambda}_i^0-\frac{1}{a}.$$

Putting things together we obtain

$$\begin{aligned}\int_0^{y_i} \hat{\lambda}_i(y_{-i}, x)dx &= \log\left|\frac{y_i}{2} + \sqrt{\frac{y_i}{\hat{\lambda}_i^0 - \hat{\lambda}_i^1} + \left(\frac{y_i}{2}\right)^2}\right| \\ &- \log\left|-\frac{y_i}{2} + \sqrt{\frac{y_i}{\hat{\lambda}_i^0 - \hat{\lambda}_i^1} + \left(\frac{y_i}{2}\right)^2}\right| \\ &+ \left(\hat{\lambda}_i^0 - \frac{1}{-\frac{y_i}{2} + \sqrt{\frac{y_i}{\hat{\lambda}_i^0 - \hat{\lambda}_i^1} + \left(\frac{y_i}{2}\right)^2}}\right) y_i.\end{aligned}$$

Since for each data point only one estimator function must be evaluated, this substantially decreases the computational cost and is about the same as for an n-fold cross-validation.

Chapter 2

Generalized Steinian covariance penalties

In this chapter, we discuss general methods for estimating the conditional prediction error in mixed models. The framework of mixed models, as any model with quadratic penalties, need not be limited to exponential family distributions, but can be extended to distributions beyond the exponential family for example, the beta or scaled-t distribution. For a discussion on a general framework for inference in such models see Wood et al. (2014). Conditionality here refers to the perspective of prediction, meaning that the prediction is conditioned on the random effects, i.e. future data are assumed to share the same random effects as the observed data. Different measures to assess prediction error, the so-called q-class of prediction errors, are presented and their representation as conditional covariance penalties with the help of the so-called optimism theorem (Efron, 2004) are discussed. A conditional version of the optimism theorem for mixed models is then presented.

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

A special focus in the simulation study is on the model choice behaviour of these estimation techniques and the different error functions, especially when comparing a complex model incorporating random effects with simpler models, that exclude the random effects, as in the simulation study in section 1.4.

2.1 Conditional prediction error in mixed models

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

$$f(y_i|\mu_i, \phi).$$

The mean is linked to a predictor by a componentwise function, i.e.

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

and the scale parameter ϕ is constant for all y_i, $i = 1, \ldots, n$. The predictor is split up into fixed parameters $\boldsymbol{\beta}$ and random parameters $\boldsymbol{u}$. For the random effects, we assume normality (although most results do not depend on the random effects distribution):

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

with positive semi-definite covariance matrix $\boldsymbol{D}$. The covariance matrix depends on a parameter, $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{\tau}^2)$. The parameter may be multivariate. In the simulation study we focus on $\boldsymbol{D} = \tau^2\boldsymbol{I}$.

2.1.1 q-class of error measures

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

$$Q(y, \hat{\mu}) = q(\hat{\mu}) + q'(\hat{\mu})(y - \hat{\mu}) - q(y).$$

This q-class of error measures was introduced by Efron (1986). In the following, we give some examples for common choices of error measures. A natural error measure that belongs to the q-class of error measures is, for instance, the squared error.

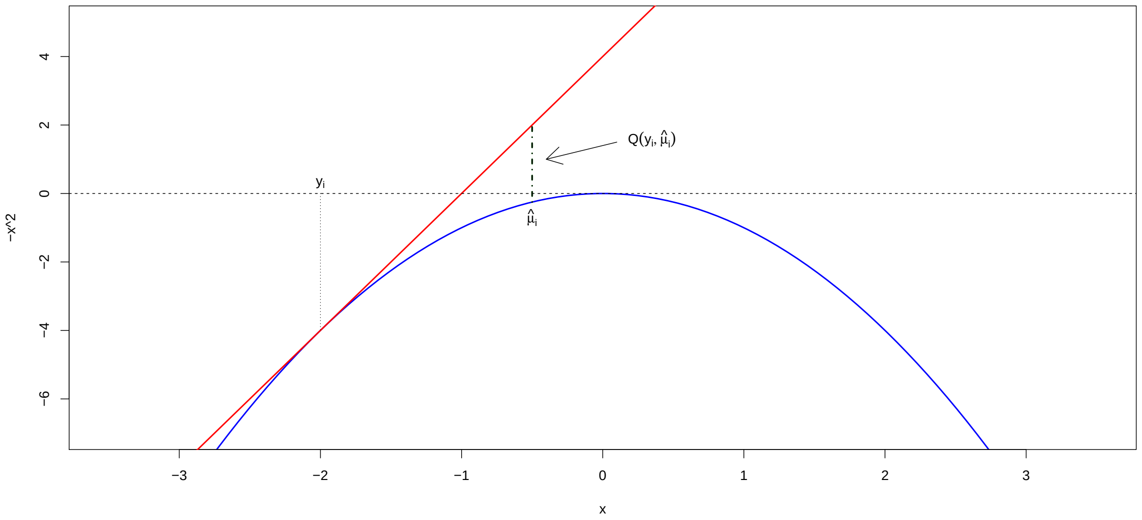

Figure 2.1: Illustration of the construction of the squared error in Example I. The tangency (red) of the concave function $q(y) = -y^2$ (blue) at y_i is evaluated at $\hat{\mu}_i$. The error $Q(y_i, \hat{\mu}_i)$ is than given by the difference between the tangency and $q(\hat{\mu}_i)$.

Example I

Choosing as a concave function

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

results in the squared error measure, i.e.

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

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

Example II

The triangular function on the unit interval

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

leading to the counting error

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

Another choice that is applicable for a large class of probability distributions is the deviance function. For exponential family distributions with natural parameter ϑ, mean $\mu = b'(\vartheta)$, scale parameter ϕ and with the function $b(\cdot)$ the same as in in 1.1.1, the deviance error is defined by

Example III

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

and thus

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

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

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

$$\theta = -\frac{q'(\hat{\mu})}{2}.$$

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

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

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

For data $\boldsymbol{y} = (y_1, \ldots, y_n)$ with predictions $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$, the total error is defined as the sum of the component errors, i.e.

$$Q(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} Q(y_i, \hat{\mu}_i) \tag{2.1.1}$$

2.1.2 Conditional covariance penalties

In order to assess the true prediction error, the quantity (2.1.1) is too optimistic since the predicted mean $\hat{\boldsymbol{\mu}}$ depends on the observed data $\boldsymbol{y}$. The obvious interest is how well the model will fit future data from the same underlying data generating process. Hence, the quantity of interest is the expected prediction error w.r.t. future data, say $\boldsymbol{z}$:

$$\mathbb{E}_{\boldsymbol{z}}(Q(\boldsymbol{z}, \hat{\boldsymbol{\mu}})).$$

If, however, the regression model under consideration contains more than one source of randomness, such as random effects, the type of prediction is not unique. In mixed models, future values may not share the same random effects as the ones that were used for fitting the model. The prediction should then be based on

$$\hat{\boldsymbol{\mu}}_m = \mathbb{E}(\boldsymbol{y}).$$

This is known as the marginal mean corresponding to the mean of the marginal distribution of the data $\boldsymbol{y}$. On the other hand, the future values at which the prediction is targeted can hold the same random effects as the observed data. Thus only one source of randomness is considered for the prediction. In this case the appropriate mean is

$$\hat{\boldsymbol{\mu}}_c = \mathbb{E}(\boldsymbol{y}|\boldsymbol{u}),$$

which is known as the conditional mean and corresponds to the mean of the conditional distribution of the data $\boldsymbol{y}|\boldsymbol{u}$. For instance, in a Gaussian model the conditional and marginal means correspond to the predictors with or without the predicted random effects,

$$\hat{\boldsymbol{\mu}}_m = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

and

$$\hat{\boldsymbol{\mu}}_c = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{u}}.$$

For other distributions, the distinction is not as obvious. The densities of the marginal distributions are often not analytically accessible. Thus in a mixed model the appropriate mean, that needs to be plugged into the error function, depends on the focus of the prediction. If the prediction focus lies on the population and future values may have any even unobserved random effects, the marginal mean is suitable. If, on the other hand the prediction focus lies on a cluster or individual associated with a random effect, the mean of choice should be the conditional mean. While for mixed models the appropriate mean depends on the prediction focus, in many applications of the mixed model in which the mixed model framework is a vehicle for estimation, such as penalized regression, the prediction is always assumed to share the same random effects.

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

$$\mathbb{E}_{\boldsymbol{z}|\boldsymbol{u}}\left(Q\left(\boldsymbol{z}, \hat{\boldsymbol{\mu}}\right)\right).$$

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

$$\mathbb{E}_{y_i,\boldsymbol{u}}\left(\mathbb{E}_{z|\boldsymbol{u}}\left(Q\left(z, \hat{\mu}_i\right)\right)\right) = \mathbb{E}_{y_i,\boldsymbol{u}}\left(Q\left(y_i, \hat{\mu}_i\right) + 2\mathrm{cov}\left(\hat{\theta}_i, y_i\right)\right), \tag{2.1.2}$$

Proof. For the i-th conditionally expected error component we have

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

and the observed error is

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

Thus the difference between observed and expected error is

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

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

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

□

Hence, the total conditional prediction error is

$$\mathbb{E}_{\boldsymbol{y},\boldsymbol{u}}\left(\mathbb{E}_{\boldsymbol{z}|\boldsymbol{u}}\left(Q\left(\boldsymbol{z},\hat{\boldsymbol{\mu}}\right)\right)\right) = \mathbb{E}_{\boldsymbol{y},\boldsymbol{u}}\left(Q\left(\boldsymbol{y},\hat{\boldsymbol{\mu}}\right) + 2\sum_{i=1}^{n}\mathrm{cov}\left(\theta_i,y_i\right)\right), \tag{2.1.3}$$

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

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

indicates the data vector in which the i-th component is excluded, the prediction error of the i-th component is also conditioned on $\boldsymbol{y}_{-i}$. Hence, the data conditioned or "fixed data" version of (2.1.2) is

$$\mathbb{E}_{y_i,\boldsymbol{u}|\boldsymbol{y}_{-i}}\left(\mathbb{E}_{z|\boldsymbol{u}}\left(Q\left(z,\hat{\mu}_i\right)\right)\right) = \mathbb{E}_{y_i,\boldsymbol{u}|\boldsymbol{y}_{-i}}\left(Q\left(y_i,\hat{\mu}_i\right) + 2\mathrm{cov}\left(\hat{\theta}_i,y_i\right)\right). \tag{2.1.4}$$

2.2 Estimating conditional covariance penalties

The covariance penalties are in general not observable and therefore need to be estimated. There are several possible approaches that will be discussed here. Very general methods that can be applied to any distributional and parametric setting are the bootstrap and cross-validation that are presented in Section 2.2.2 and 2.2.3. The latter can even be applied to nonparametric settings. Nevertheless, for certain distributions it is possible to derive estimators that are preferable. These can be seen as generalizations of the *Steinian* type estimators that were presented in Chapter 1 and are thus refereed to as the *Steinian* in Efron (2004). Unlike the estimators in Chapter 1, those presented in this section are not unbiased but are reasonable approximations, as shown in the subsequent simulation study. Such estimators are proposed for the gamma- and the Bernoulli-distribution. With the help of Theorem 2.2.1 it is furthermore possible to gain insight into the circumstances under which these estimator are available. Furthermore, there is an interesting connection between the *Steinian* type estimators that are presented here and the conditional "fixed data" bootstrap, thereby the *Steinian* type estimators appear to be large sample approximations of the conditional bootstrap estimators that keep all responses except the i-th fixed.

2.2.1 Plug-in conditional covariance penalty

As stated above, the formulas (1.2.1) and (1.2.6) only hold for a limited number of parametric distribution families, i.e. the Gaussian, Poisson and exponential families. There are attempts to generalize such kind of *Steinian* formulas to further distributions, although there is up to date no generalization that leads to unbiased estimates of the conditional covariance penalties for all distributions. One such generalized *Steinian* formula for a large class of distributions is given in Shen and Huang (2006):

Theorem 2.2.1. *Let y be a continuous random variable with probability density function $f(\cdot)$ with support (a,b) where $a \in \mathbb{R} \cup \{-\infty\}$ and $b \in \mathbb{R} \cup \{\infty\}$ and $\theta = \theta(y)$ a differentiable function such that $\mathbb{E}\left(\theta(y)(y-\mu)\right) < \infty$ with $\mu = \mathbb{E}(y)$, then*

$$cov(\theta(y), y) = \mathbb{E}\left(\theta'(y)V(y,\mu)\right) \tag{2.2.1}$$

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

The function $V(y,\mu) = \frac{1}{f(y)} \int_a^y (\mu - t) f(t) dt$ is counterintuitive, but its expectation is the variance, i.e. $\mathbb{E}\left(V(y,\mu)\right) = \mathrm{Var}(y)$.
A similar identity also holds for discrete random variables and is given here although it is not used in the subsequent discussion:

Theorem 2.2.2. *Let y be a discrete random variable with probability function $p(\cdot)$ on the support $\mathcal{T} \subset \mathbb{Z}$ and $\theta = \theta(y)$ a function such that $\mathbb{E}\left(\theta(y)(y-\mu)\right) < \infty$ with $\mu = \mathbb{E}(y)$, then*

$$cov(\theta(y), y) = \mathbb{E}\left(\Delta\theta(y)V(y,\mu)\right) \tag{2.2.2}$$

with $V(y,\mu) = \frac{1}{p(y)} \sum_{t \in \mathcal{T}}^{y} (\mu - t) p(t)$ and $\Delta\theta(y) = \theta(y^+) - \theta(y)$, where y^+ the smallest number in $\mathcal{T}$ that is larger then y and $y^+ = y$ if it is the largest value in its support $\mathcal{T}$.

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

Example I

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

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

Example II

For the Poisson distribution

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

Example III

For the Beta distribution on (α, β)

$$V(y,\mu) = \frac{1}{\alpha+\beta} y(1-y).$$

Example VI

For the Gamma distribution with mean μ and scale parameter ν

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

The formula (2.2.1), and the same holds for its discrete analogue in (2.2.2), does not automatically lead to an observable quantity because in general we cannot plug in an estimate of $V(y,\mu)$, since

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

We therefore need to distinguish between two cases: Either the "variance" function is independent of y, i.e. $V(y,\mu) = V(\mu)$, or the "variance" function somehow depends on y. The first case is unproblematic since we have $\mathbb{E}\left(\theta'(y)V(y,\mu)\right) = \mathbb{E}\left(\theta'(y)\right)\mathrm{Var}(y)$ and we can plug in an estimate of $\mathrm{Var}(y)$ as is done in the bias correction for the normal distribution, see (1.2.3). However, we can make some progress in the second case for the gamma distribution or more generally for all distributions, for which

$$V(y,\mu) = y \cdot \tilde{V}(\mu),$$

with $\tilde{V}(\mu)$ only depending on μ not on y. In this case, for y and $\theta(y)$ as in (2.2.1) and $\theta(y)$ additionally being s-times continuously differentiable with $s \in \mathbb{N}$ and $\theta^{(s+1)}(y) = 0$, the conditional covariance penalty is

$$\operatorname{cov}(\theta, y) = \operatorname{Var}(y) \sum_{i=1}^{s} \mathbb{E}(\theta^{(i)}) \cdot \tilde{V}(\mu)^{i-1}. \tag{2.2.3}$$

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

$$\mathbb{E}(h(y)y) = \mathbb{E}(h(y))\,\mathbb{E}(y) + \operatorname{cov}(h(y), y),$$

in combination with Theorem 2.2.1. Thus the covariance can be rewritten:

$$\begin{aligned}
\operatorname{cov}(\theta(y), y) &= \mathbb{E}\left(\theta'(y)V(y,\mu)\right) \\
&= \mathbb{E}\left(\theta'(y) \cdot y\right)\tilde{V}(\mu) \\
&= \left[\mathbb{E}\left(\theta'(y)\right)\mathbb{E}(y) + \underbrace{\operatorname{cov}(\theta'(y), y)}_{=\mathbb{E}\left(\theta''(y)V(y,\mu)\right)}\right]\tilde{V}(\mu) \\
&= \mathbb{E}\left(\theta'(y)\right)\underbrace{\mathbb{E}(y)\tilde{V}(\mu)}_{=\operatorname{Var}(y)} + \mathbb{E}\left(\theta''(y)\underbrace{V(y,\mu)}_{=y\tilde{V}(\mu)}\right)\tilde{V}(\mu) \\
&= \mathbb{E}\left(\theta'(y)\right)\operatorname{Var}(y) + \mathbb{E}\left(\theta''(y)y\right)\tilde{V}(\mu)^2 \\
&= \mathbb{E}\left(\theta'(y)\right)\operatorname{Var}(y) + \mathbb{E}\left(\theta''(y)\right)\mathbb{E}(y)\tilde{V}(\mu)^2 + \operatorname{cov}(\theta''(y), y)\tilde{V}(\mu)^2 \\
&= \operatorname{Var}(y)\mathbb{E}\left(\theta'(y)\right) + \operatorname{Var}(y)\mathbb{E}\left(\theta''(y)\right)\tilde{V}(\mu) + \operatorname{cov}(\theta''(y), y)\tilde{V}(\mu)^2 \\
&= \ldots \\
&= \operatorname{Var}(y)\sum_{i=1}^{s}\mathbb{E}\left(\theta^{(i)}(y)\right)\tilde{V}(\mu)^{i-1}
\end{aligned}$$

□

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

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

the covariance penalty can be approximated by

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

For the mean μ in $\tilde{\theta}(y)$, a plug-in estimator can be used. Either the estimated mean $\hat{\mu}$ or the estimator of the saturated model, i.e. the observed values y, are applicable. Thus using a first order Taylor expansion and the observed values as estimators of the mean and substituting an estimator for the variance $\widehat{\operatorname{Var}(y)} = V(\hat{\mu})$, formula 2.2.3 can be estimated by:

$$\operatorname{cov}(\theta(y), y) \approx V(\hat{\mu})\mathbb{E}\left(\theta'(y)\right). \tag{2.2.4}$$

Note that the subsequent result in equation (2.2.3) can be generalized by allowing for linear translations of the random variable to be separated from the "variance" function. Under the same conditions that need to hold for the formula (2.2.3) and, additionally, with $a, b \in \mathbb{R}$ and $V(y, \mu) = (a + by) \cdot \tilde{V}(\mu)$, the covariance penalty is

$$\operatorname{cov}(\theta, y) = \operatorname{Var}(y) \sum_{i=1}^{s} \mathbb{E}(\theta^{(i)}) \cdot b^{i-1} \tilde{V}(\mu)^{i-1}. \tag{2.2.5}$$

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

$$\operatorname{cov}_i = \mu_i(1-\mu_i)\left(\hat{\theta}_i(1) - \hat{\theta}_i(0)\right). \tag{2.2.6}$$

since

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

Thus, substituting an estimator for the variance $\mu_i(1-\mu_i)$ leads to the estimate

$$\sum_{i=1}^{n} \widehat{\text{cov}}_i = \sum_{i=1}^{n} \hat{\mu}_i(1-\hat{\mu}_i)\left(\hat{\theta}_i(1)-\hat{\theta}_i(0)\right). \tag{2.2.7}$$

In Section 2.2.2 we will show the close connection between this estimator and the bootstrap estimator.

2.2.2 Conditional parametric bootstrap

A direct way of estimating the conditional covariance penalty is the parametric bootstrap with conditional random effects. This means that every bootstrap sample is taken from the conditional distribution with conditional mean $\hat{\boldsymbol{\mu}}_c$. For a large number of bootstrap simulations B from the originally fitted model $\hat{\boldsymbol{\theta}}|\boldsymbol{u} = \hat{\boldsymbol{\theta}}_c$ the covariance of the parameters of main interest $\hat{\boldsymbol{\theta}}_c$ and the bootstrap samples $\boldsymbol{z}$ is calculated by

$$\sum_{i=1}^{n} \widehat{\text{cov}}_i = \sum_{i=1}^{n} \frac{1}{B-1} \sum_{j=1}^{B} \hat{\theta}_{ij}(z_{ij})\left(z_{ij} - \overline{\boldsymbol{z}}_{i\cdot}\right), \tag{2.2.8}$$

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

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

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

with

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

The bootstrap here needs B model refits. Another possibility for a bootstrap estimate would be a "fixed data" bootstrap, in which for each bootstrap sample $n-1$ data points are fixed and only the i-th datum is resampled. This corresponds to the estimate in (2.1.4). However, this approach is computationally burdensome since for each observed response a whole set of bootstrap samples and model fits must be computed and thus the total error estimation requires $n \cdot B$ model fits. Nonetheless, there is an approximation of the bootstrap estimate that makes it less computationally expensive and that coincides with the approximation of formula (2.2.4). Therefore, consider the bootstrap estimator for the i-th covariance penalty with all cases but the i-th fixed. Instead of evaluating the main parameter of interest we use a Taylor approximation around the estimated mean $\hat{\mu}_i$ yielding

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

Accordingly, the bootstrap estimator of the i-th conditional covariance penalty can be approximated by

$$\begin{aligned}
\widehat{\mathrm{cov}}_i &= \frac{1}{B-1}\sum_{j=1}^{B} \hat{\theta}_i(z_j)\,(z_j - \overline{\boldsymbol{z}}_i) \\
&\approx \frac{1}{B-1}\sum_{j=1}^{B} \left(\hat{\theta}_i(\hat{\mu}_i) + \left.\frac{\partial \hat{\theta}_i}{\partial z}\right|_{\hat{\mu}_i} (z_j - \hat{\mu}_i) \right) (z_j - \overline{\boldsymbol{z}}_i) \\
&\approx \left.\frac{\partial \hat{\theta}_i}{\partial z}\right|_{\hat{\mu}_i} \frac{1}{B-1}\sum_{j=1}^{B} (z_j - \hat{\mu}_i)\,(z_j - \overline{\boldsymbol{z}}_i) \\
&\approx \left.\frac{\partial \hat{\theta}_i}{\partial z}\right|_{\hat{\mu}_i} \widehat{\mathrm{Var}(y_i)}, \qquad (2.2.9)
\end{aligned}$$

with the last approximation holding if the number of bootstrap samples tends to infinity, $B \to \infty$. A similar link between the bootstrap and the plug-in estimator for Bernoulli conditional covariance penalties is derived in 2.3.2.

2.2.3 Conditional cross-validation

The probably most popular method for prediction error estimation is cross-validation. Compared to conditional parametric bootstrap and the plug-in estimates, the conditional cross-validation has the advantage that it is not model-dependent. On the other hand, just like the plug-in estimates, the conditional cross-validation is a local method in the sense that, for estimation of the covariance penalty, it only changes the i-th data point. Let $\hat{\mu}_{-i}$ be the estimated mean with the i-th observation deleted, i.e. the estimator based on the reduced data set $\boldsymbol{y}_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$. Then an obvious estimator of the conditional covariance penalty is

$$\sum_{i=1}^{n} \widehat{\text{cov}}_i = \frac{1}{2} \sum_{i=1}^{n} \left[Q(y_i, \hat{\mu}_{-i}) - Q(y_i, \hat{\mu}_i) \right]. \tag{2.2.10}$$

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

$$\sum_{i=1}^{n} \widehat{\text{cov}}_i = \sum_{i=1}^{n} b(\hat{\vartheta}_{-i}) - b(\hat{\vartheta}_i) + y(\hat{\vartheta}_i - \hat{\vartheta}_{-i}), \tag{2.2.11}$$

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

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

2.3 Simulations

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

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

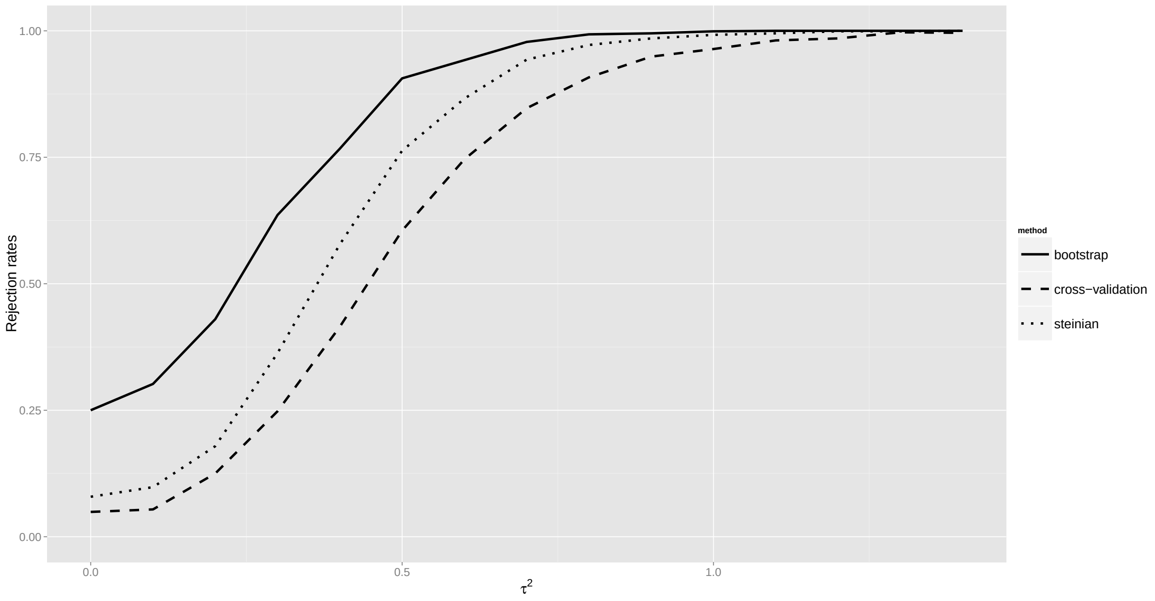

Figure 2.2: Frequency of choosing the complex model (2.3.1) against a simple model only including an intercept. The conditional covariance penalties are estimated by bootstrap (2.2.8), cross- validation (2.2.11) and the *Steinian* estimator (2.3.3).

2.3.1 Gamma distribution

Deviance error

In order to corroborate the theoretical results that evolved in the previous sections, we deploy a random intercept model with conditionally gamma distributed responses. Thus the conditional density of the data generating process is given by

$$f(y_{ij}|\mu_{ij},\nu) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu_{ij}}\right)^{\nu} y^{\nu-1}\exp\left(-\frac{\nu y}{\mu_{ij}}\right) \tag{2.3.1}$$

for $i = 1,\dots,n$ and $j = 1,\dots,m$, with

$$\mu_{ij} = \exp(\beta_0 + u_j) \quad \text{and} \quad \boldsymbol{u} \sim \mathcal{N}\left(0, \tau^2 \boldsymbol{I}\right).$$

The number of individuals is set to 12 and the number of observations is 6. The variance parameter of the random intercept τ^2 varies between 0 and 1.6. For each setting 1,000 data sets of model (2.3.1) are generated. The scale parameter $\phi = \frac{1}{\nu} = 1$ is constant for all observations.

The error for the gamma distributed observations is assessed with the deviance error in Example III, hence the total expected prediction error that is ought to be minimized is

$$\mathbb{E}_{\boldsymbol{z}|\boldsymbol{u}}\left(Q\left(\boldsymbol{z},\hat{\boldsymbol{\mu}}\right)\right)=Q\left(\boldsymbol{y},\hat{\boldsymbol{\mu}}\right)+2\sum_{i=1}^{n}\sum_{j=1}^{m}\operatorname{cov}\left(\hat{\theta}_{ij},y_{ij}\right), \tag{2.3.2}$$

with the natural parameter

$$\hat{\theta}_{ij}=\hat{\vartheta}_{ij}=\left(-\frac{1}{\hat{\mu}_{ij}}\right).$$

Thus, based on formula (2.2.3) we approximate the conditional covariance penalty for gamma distributed responses of the i-th data point by

$$\widehat{\operatorname{cov}}_{ij}\approx\widehat{V}_{ij}\frac{\partial\hat{\theta}_{ij}}{\partial y_{ij}}. \tag{2.3.3}$$

Notice that for the gamma distribution, as for any exponential family distribution, the *Steinian* type estimator of the conditional covariance (2.3.3) can be simplified as follows:

$$\widehat{\operatorname{cov}}_{ij}\approx\widehat{V}_{ij}\frac{\partial\hat{\theta}_{ij}}{\partial y_{ij}}=\hat{\phi}\frac{\partial\hat{\mu}_{ij}}{\partial\hat{\theta}_{ij}}\frac{\partial\hat{\theta}_{ij}}{\partial y_{ij}}=\hat{\phi}\frac{\partial\hat{\mu}_{ij}}{\partial y_{ij}}. \tag{2.3.4}$$

Next to this estimate, the conditional covariance penalties are estimated by conditional parametric bootstrap and conditional cross-validation, see equations (2.2.8) and (2.2.10). The estimation of the random effects variance parameter is done by REML estimation based on the R-package `mgcv` version `1.8-2`, see Wood (2011). The bootstrap needs 500 model fits, while the cross-validation only needs $n\cdot m=72$ model fits. The derivatives in (2.3.4) are calculated on the basis of the algorithm in Gilbert and Varadhan (2012).

A null model rejection rate plot similar to those introduced in Greven and Kneib (2010) is displayed in Figure 2.2. This plot shows the frequency of selecting the more complex model (2.3.1), incorporating random effects to the simpler models with only one intercept. For each data set the total expected prediction error (2.3.2) of the simple and complex model is estimated and the model with less prediction error is chosen. The cross-validation estimate here behaves similar to the marginal AIC in Greven and Kneib (2010). For a random effects variance of 0.5, the cross-validation only chooses to include random effects in roughly six out of ten cases, whereas the *Steinian* estimator does so in more than three-quarter of the cases. The bootstrap shows a good behaviour

τ^2	cov_b	cov_{cv}	cov_s	$\mathsf{sd}(\mathsf{cov}_b)$	$\mathsf{sd}(\mathsf{cov}_{cv})$	$\mathsf{sd}(\mathsf{cov}_s)$
0	3.54	3.26	2.95	0.97	3.38	2.62
0.1	3.66	3.66	3.29	1.13	3.75	2.78
0.2	3.92	4.57	4.10	1.33	3.79	2.97
0.3	4.70	6.23	5.54	1.86	4.17	4.04
0.4	5.88	7.90	11.93	2.40	4.26	160.65
0.5	7.18	9.68	8.12	2.52	4.12	2.38
0.6	8.32	10.74	8.87	2.54	4.71	2.09
0.7	9.50	11.88	9.61	2.37	4.38	1.57
0.8	10.53	12.87	10.15	2.09	4.43	1.16
0.9	11.17	13.39	10.44	1.89	4.70	1.02
1	11.75	13.93	10.72	1.56	4.80	0.78
1.1	12.16	14.18	10.92	1.41	4.62	0.64
1.2	12.54	14.60	11.07	1.11	4.94	0.52
1.3	12.77	14.51	11.19	1.02	4.66	0.47
1.4	13.04	15.34	11.32	0.83	5.01	0.36

Table 2.1: Mean and standard error of the conditional covariance penalty samples estimated by bootstrap (cov_b), cross-validation (cov_{cv}) and the Steinian estimate 2.2.7 (cov_s) with 1,000 samples from a model with 12 individuals and 6 observations per individual for the conditional gamma distribution 2.3.1.

although it incorporates random effects in a quarter of the cases although there are none in the underlying data-generating mechanism.

The estimated conditional covariance penalties and the corresponding standard deviations are listed in Table 2.1. The table shows that all three methods give similar estimates for the conditional covariance penalties. However, the standard deviation of the cross-validation and even more of the bootstrap method increases with rising random effects variance τ^2. The *Steinian* type estimator (2.3.4) shows the exact opposite behaviour. The value of the *Steinian* estimator for random effects variance $\tau^2 = 0.4$ is salient. This is due to the fact that the estimates have not appropriately been corrected for numerical anomalies that arise from the use of numerical derivation. However, the standard deviation already indicates that there is an outlier. The model choice behaviour displayed in Figure 2.2 showing no unexpected conduct for $\tau^2 = 0.4$ assures this.

The six boxplots in Figure 2.3 confirm how close the three estimates of the conditional covariance are to each other and how the variability of the estimates changes with increasing τ^2.

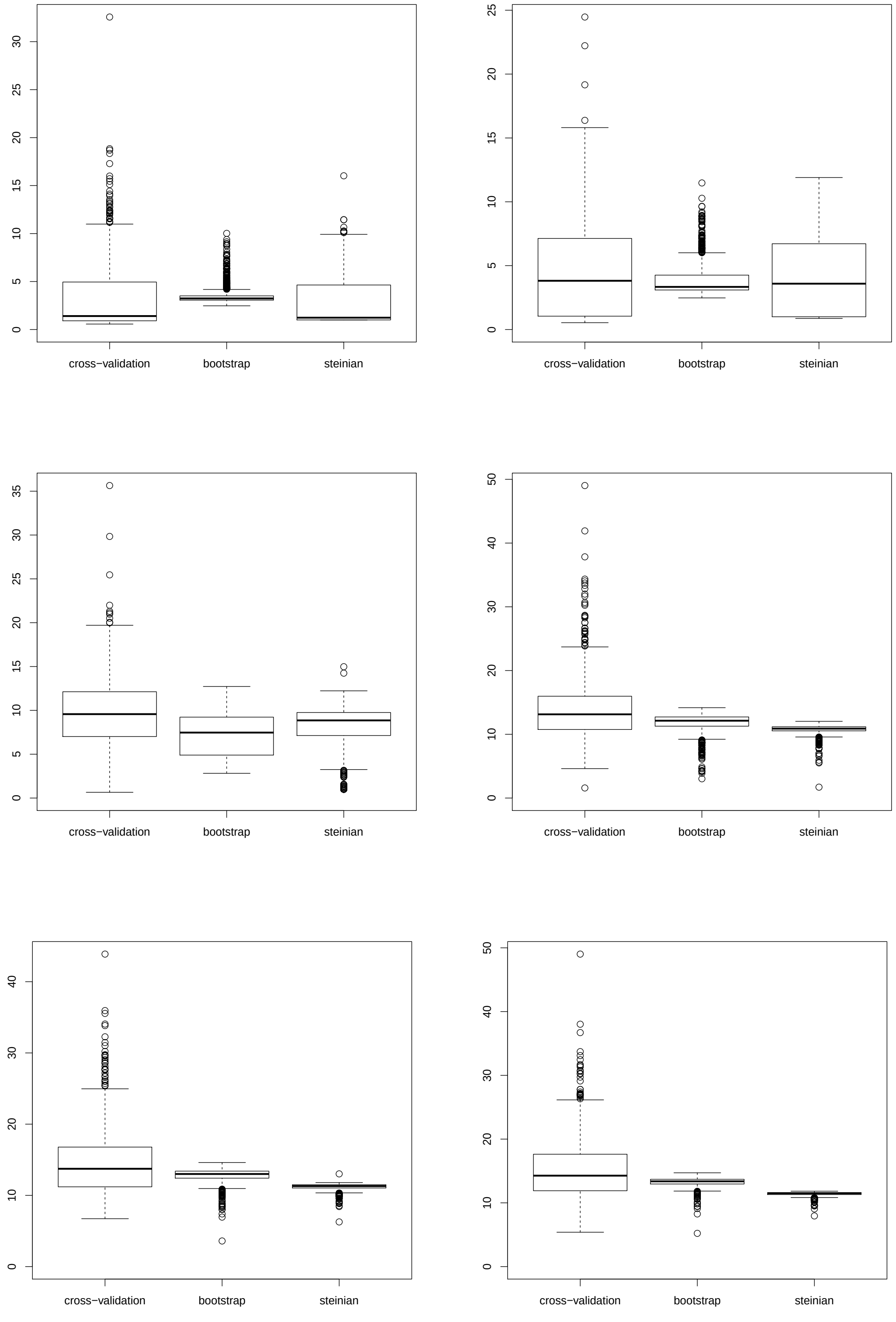

Figure 2.3: Boxplots of the estimated covariance penalties for 13 individuals, 7 observations per individual and variance of random effects equal to 0, .2, .5, 1, 1.3, 1.5 respectively.

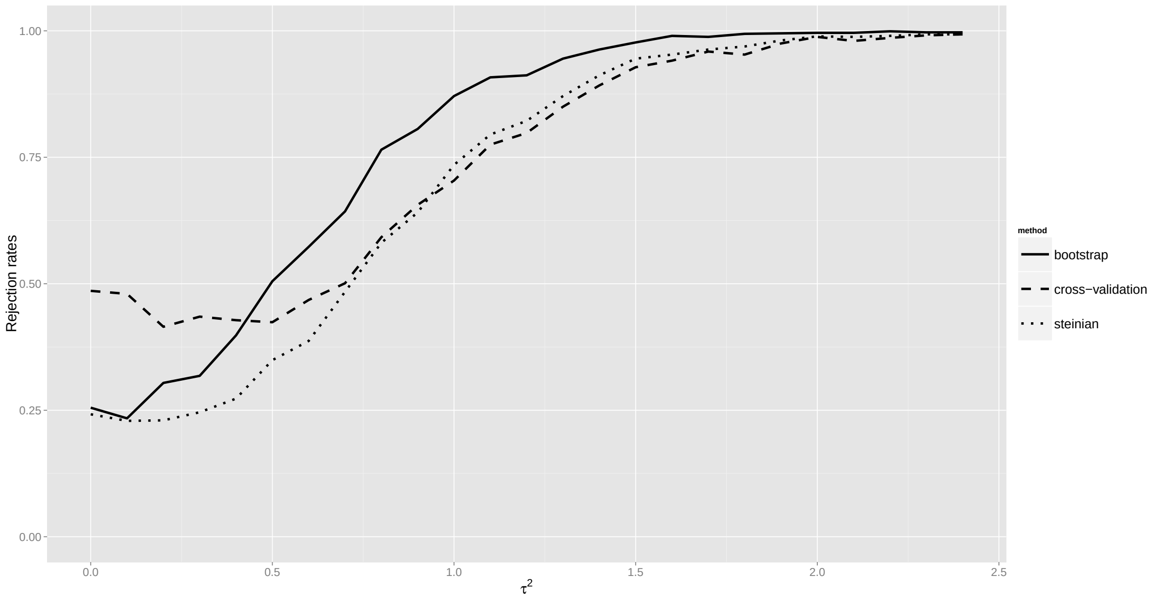

Figure 2.4: Frequency of choosing the complex model (2.3.5) against a simple model only including an intercept. The conditional covariance penalties are estimated by bootstrap (2.2.8), cross-validation (2.2.11) and the *Steinian* estimator (2.2.7).

2.3.2 Bernoulli distribution

Binomial deviance error

For the Bernoulli distribution, we conduct a simulation study that analyses the behaviour of the estimator of the expected conditional deviance error and the corresponding conditional covariance penalty (2.2.7) and compares it to the bootstrap and to cross-validation. Additionally, we can also compute the "true" conditional covariance penalties from formula (2.2.6), since the true mean of the data generating mechanism is known. As in the simulation study on the gamma distribution, rejection rates of a simple model without random effects in comparison to a complex model including random effects are given for different degrees of variance of the random effects.

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

$$f(y_{ij}|\mu_{ij}) = (1-\mu_{ij})^{1-y_{ij}}\mu_{ij}^{y_{ij}} \quad \text{for} \quad i=1,\dots,n \quad \text{and} \quad j=1,\dots,m \tag{2.3.5}$$

with

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + u_j \quad \text{and} \quad \boldsymbol{u} \sim \mathcal{N}\left(0, \tau^2 \boldsymbol{I}\right).$$

The number of individuals is set to 13 and 28 and the number of observations is 7 and 13 respectively. The variance parameter of the random intercept τ^2 varies between 0 and 2.4. For each setting, 1,000 data sets of model (2.3.5) are generated and the covariance penalties of the model are estimated by the different estimation techniques proposed in the preceding sections. The bootstrap estimate is based on 800 bootstrap samples.

The models are fitted with the REML method implemented in the R-package `lme4`, see Bates, Mächler, Bolker, and Walker (2014b) and Bates, Mächler, Bolker, and Walker (2014a). The conditional bootstrap and the *Steinian* are estimated with the R-package `cAIC4`, see Saefken, Ruegamer, Greven, and Kneib (2014), that is presented in Chapter 4. The bootstrap here needs 800 model fits, while the cross-validation and the *Steinian* only need $n \cdot m$ model fits. The fits needed for the *Steinian* are faster than for cross-validation, since the data set remains unchanged except for one response value in each computation, which leads to rapid convergence. Thus, from the computational perspective the *Steinian* performs best.

Figure 2.4 displays the frequencies of how often the complex model (2.3.5) is favoured against a simple model with only an intercept. This means we choose the model that minimizes the expected conditional prediction error

$$\mathbb{E}_{\boldsymbol{z}|\boldsymbol{u}}\left(Q\left(\boldsymbol{z}, \hat{\boldsymbol{\mu}}\right)\right).$$

The error function Q is the deviance error from Example III and the total conditional prediction error is thus assessed by

$$\mathbb{E}_{\boldsymbol{z}|\boldsymbol{u}}\left(Q\left(\boldsymbol{z}, \hat{\boldsymbol{\mu}}\right)\right) = Q\left(\boldsymbol{y}, \hat{\boldsymbol{\mu}}\right) + 2\sum_{i=1}^{n}\sum_{j=1}^{m} \operatorname{cov}\left(\hat{\theta}_{ij}, y_{ij}\right), \tag{2.3.6}$$

with the logit parameter

$$\hat{\theta}_{ij} = \hat{\vartheta}_{ij} = \log\left(\frac{\hat{\mu}_{ij}}{1-\hat{\mu}_{ij}}\right).$$

It is, however, not clear how the covariance penalties of the simple model can be estimated. In order to stay consistent with the estimation of the conditional covariance penalties of the complex models, the covariance penalties of the simple models are estimated with bootstrap, cross-validation and the *Steinian* applied to the generalized linear model. The bootstrap and the *Steinian* in one quarter of the cases choose the complex model although the true underlying model does not incorporate random effects. Notice that the behaviour of choosing too many parameters is rather common for AIC-like criteria. For instance, the significance level of the AIC in standard settings is approximately 0.157, see Greven and Kneib (2010). The bootstrap chooses the true complex model slightly more often than the *Steinian* for increasing random effects variance. The cross-validation, on the other hand, performs worse as the chance of selecting the false complex model is almost 0.5.

For the 1,000 estimated conditional covariance penalties, boxplots for different sizes of the random effects variance parameters are given in Figure 2.5. The performance of cross-validation, bootstrap and the *Steinian* highly depend on the random effects variance parameter. This makes a comparison difficult. The variability of the *Steinian* with the true mean plugged in is larger than the variability when the estimated mean is used.

The empirical means and the standard errors of the conditional covariance samples are listed in Table 2.2.
Notice that the bootstrap estimate here is unconditional in the sense that all cases of the data set are varied in each set of covariance penalty estimates. The bootstrap that is conditioned on $\boldsymbol{y}_{-i}$, i.e. in which for the estimation of the i-th covariance penalty only the i-th case is resampled for Bernoulli responses is approximately equal to the *Steinian* (2.2.7), which can be seen as follows:

$$\begin{aligned}\widehat{\text{cov}}_i &= \frac{1}{B-1}\sum_{j=1}^{B}\hat{\theta}_i(z_j)\,(z_{ij}-\overline{\boldsymbol{z}}_{i\cdot})\\ &= \frac{B_1}{B-1}\hat{\theta}_i(1)\,(1-\overline{\boldsymbol{z}}_{i\cdot})+\frac{B_0}{B-1}\hat{\theta}_i(0)\,(-\overline{\boldsymbol{z}}_{i\cdot})\,,\end{aligned}$$

where B_0 is the number of zeros in the bootstrap sample and B_1 is the number of ones in the bootstrap sample, $\hat{\theta}_i(1)=\log\left(\frac{\hat{\mu}_i(1)}{1-\hat{\mu}_i(1)}\right)$ is the estimated logit (the natural parameter) with $y_i=1$ and $\hat{\theta}_i(0)=\log\left(\frac{\hat{\mu}_i(1)}{1-\hat{\mu}_i(1)}\right)$ the same, but $y_i=0$ is assumed. The mean of the bootstrap sample $\overline{\boldsymbol{z}}_{i\cdot}=\frac{1}{B}\sum_{j=1}^{B}z_{ij}$ is approximately equal to $\hat{\mu}_i$ as $B\to\infty$. Thus we have

τ^2	cov_b	cov_{cv}	cov_s	$\mathsf{sd}(\mathsf{cov}_b)$	$\mathsf{sd}(\mathsf{cov}_{cv})$	$\mathsf{sd}(\mathsf{cov}_s)$
0.0	2.94	2.69	2.55	0.68	2.31	2.07
0.1	2.95	2.67	2.54	0.67	2.33	2.09
0.2	3.00	3.03	2.86	0.68	2.44	2.17
0.3	3.07	3.20	3.02	0.80	2.55	2.29
0.4	3.16	3.60	3.39	0.91	2.67	2.41
0.5	3.40	4.32	4.04	1.11	2.83	2.58
0.6	3.64	4.80	4.50	1.35	2.90	2.67
0.7	3.92	5.42	5.08	1.49	2.92	2.71
0.8	4.40	6.20	5.82	1.73	2.86	2.68
0.9	4.81	6.75	6.34	1.92	2.77	2.61
1.0	5.27	7.37	6.96	2.04	2.45	2.34
1.1	5.65	7.78	7.36	2.00	2.24	2.17
1.2	6.06	8.06	7.61	2.11	2.29	2.18
1.3	6.50	8.45	7.95	2.13	2.02	1.91
1.4	6.85	8.76	8.26	2.00	1.80	1.72
1.5	7.25	9.03	8.47	1.90	1.66	1.52
1.6	7.54	9.26	8.68	1.83	1.47	1.30
1.7	7.90	9.44	8.85	1.67	1.37	1.24
1.8	7.93	9.47	8.81	1.73	1.31	1.16
1.9	8.16	9.57	8.89	1.57	1.33	1.09
2.0	8.51	9.78	8.97	1.33	1.28	1.00
2.1	8.48	9.71	8.94	1.41	1.33	1.06
2.2	8.68	9.86	8.92	1.30	1.41	1.09
2.3	8.79	9.88	8.93	1.19	1.44	1.07
2.4	8.77	9.95	8.85	1.24	1.55	1.12

Table 2.2: Mean and standard error of the conditional covariance penalty samples estimated by bootstrap (cov_b), cross-validation (cov_{cv}) and the Steinian estimate 2.2.7 (cov_s) with 1,000 samples from a model with 13 individuals and 7 observations per individual for the conditional Bernoulli distribution 2.3.5.

$$
\begin{aligned}
\widehat{\text{cov}}_i &= \frac{B_1}{B-1}\hat{\theta}_i(1)\,(1-\overline{\boldsymbol{z}}_{i\cdot}) - \frac{B_0}{B-1}\hat{\theta}_i(0)\,(\overline{\boldsymbol{z}}_{i\cdot}) \\
&\approx \frac{B_1}{B-1}\hat{\theta}_i(1)\,(1-\hat{\mu}_i) - \frac{B_0}{B-1}\hat{\theta}_i(0)\,(\hat{\mu}_i) \\
&\approx \hat{\mu}_i\hat{\theta}_i(1)\,(1-\hat{\mu}_i) - (1-\hat{\mu}_i)\,\hat{\theta}_i(0)\,(\hat{\mu}_i) \\
&= \hat{\mu}_i\,(1-\hat{\mu}_i)\left(\hat{\theta}_i(1)-\hat{\theta}_i(0)\right).
\end{aligned}
$$

Since the bootstrap estimates are optimal if the number of bootstrap resamples B tends to infinity, the *Steinian* estimate can be seen as the optimal conditional (in the sense of $z|\boldsymbol{y}_{-i}$) bootstrap estimate. The unconditioned bootstrap, however, in view of the simulation results seems to be biased. This may be due to the correlation structure that is imposed by the random effects. Hence, the responses are only conditionally independent (in the sense of random effects).

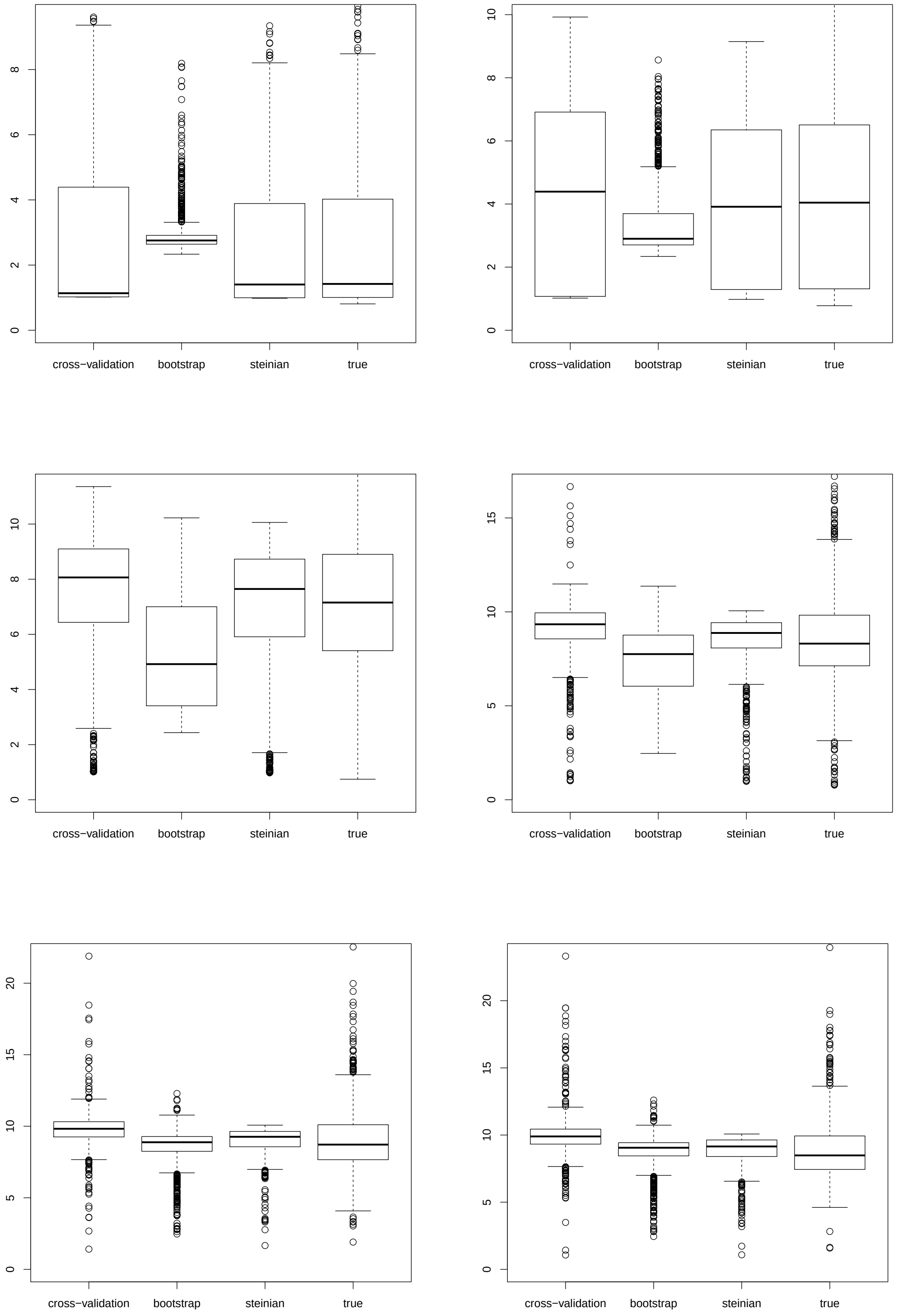

Figure 2.5: Boxplots of the estimated covariance penalties for 13 individuals, 7 observations per individual and variance of random effects equal to 0, 0.5, 1, 1.5, 2, 2.4 respectively.

2.3.3 Scaled t distribution

Squared error

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

$$f(y_{ij}|\mu_{ij},\nu,\sigma) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}\sigma}\left(1+\frac{1}{\nu}\left(\frac{y_{ij}-\mu_{ij}}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}} \tag{2.3.7}$$

for $i = 1,\dots,n$ and $j = 1,\dots,m$, with

$$\mu_{ij} = \exp(\beta_0 + u_j) \quad \text{and} \quad \boldsymbol{u} \sim \mathcal{N}\left(0,\tau^2\boldsymbol{I}\right).$$

The number of individuals is set to 17 and the number of observations is 7 respectively. The variance parameter of the random intercept τ^2 varies between 0 and 1.6. For each setting 1,000 data sets of model (2.3.7) are generated. For ease of computation we expect the remaining parameters to be fixed and known, i.e. $\nu = 7$ and $\sigma = 1$.

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

$$\mathbb{E}_{\boldsymbol{z}|\boldsymbol{u}}\sum_{i=1}^{n}\sum_{j=1}^{m}(z_{ij}-\hat{\mu}_{ij})^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\hat{\mu}_{ij})^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{m}\operatorname{cov}(\hat{\mu}_{ij},y_{ij}). \tag{2.3.8}$$

Thus, based on formula (2.2.9), we approximate the conditional covariance penalty for scaled t distributed responses of the ij-th data point by

$$\widehat{\operatorname{cov}}_{ij} \approx \widehat{\operatorname{Var}(y_{ij})}\left.\frac{\partial\hat{\mu}_{ij}}{\partial z}\right|_{\hat{\mu}_{ij}}. \tag{2.3.9}$$

Next to this approximate estimate, the conditional covariance penalties are estimated by conditional parametric bootstrap and conditional cross-validation, see equations (2.2.8) and (2.2.10). The bootstrap estimate is based on 800 bootstrap samples.

τ^2	cov_b	cov_{cv}	cov_s	$\mathsf{sd}(\mathsf{cov}_b)$	$\mathsf{sd}(\mathsf{cov}_{cv})$	$\mathsf{sd}(\mathsf{cov}_s)$
0	3.98	3.53	3.58	0.82	3.25	3.13
0.1	4.17	4.25	4.32	1.10	3.58	3.45
0.2	4.58	5.72	5.67	1.46	4.00	3.72
0.3	5.78	8.29	8.13	2.22	4.12	3.73
0.4	7.48	10.67	10.34	2.92	3.98	3.37
0.5	9.59	12.71	12.29	3.16	3.53	2.80
0.6	11.57	14.44	13.82	2.84	3.05	2.15
0.7	13.15	15.61	14.89	2.60	2.90	1.77
0.8	14.35	16.43	15.67	2.30	2.80	1.53
0.9	15.18	17.18	16.24	1.97	2.92	1.30
1.0	15.98	17.61	16.80	1.53	2.80	1.02
1.1	16.52	18.20	17.17	1.25	2.90	0.86
1.2	16.91	18.38	17.47	0.99	2.77	0.71
1.3	17.27	18.89	17.75	0.87	2.90	0.63
1.4	17.46	19.01	17.89	0.74	2.87	0.56
1.5	17.69	19.15	18.08	0.62	2.89	0.47
1.6	17.86	19.43	18.23	0.55	2.76	0.41
1.7	17.97	19.60	18.32	0.50	3.05	0.37
1.8	18.07	19.51	18.42	0.45	2.84	0.34

Table 2.3: Mean and standard error of the conditional covariance penalty samples estimated by bootstrap (cov_b), cross-validation (cov_{cv}) and the Steinian estimate 2.2.7 (cov_s) with 1,000 samples from a model with 17 individuals and 7 observations per individual for the conditional scaled t distribution.

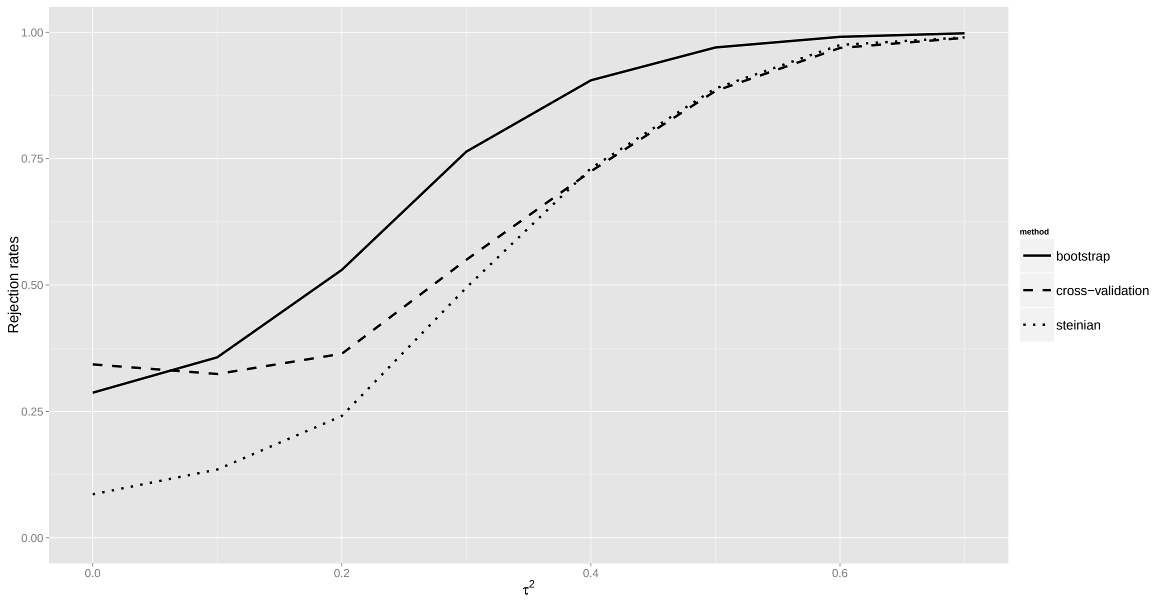

Figure 2.6: Frequency of choosing the complex model (2.3.7) against a simple model only including an intercept. The conditional covariance penalties are estimated by bootstrap (2.2.8), cross-validation (2.2.11) and the *Steinian* estimator (2.2.9).

The models are fitted with the R-package `mgcv` version `1.8-2`, see Wood et al. (2014), this package uses a REML criterion to find the optimal random effects variance parameter τ^2. The bootstrap here needs 800 model fits, while the cross-validation and the *Steinian* only need $n \cdot m = 119$ model fits. The derivatives in (2.3.9) are numerically approximated based on the algorithm in Gilbert and Varadhan (2012).

The estimated conditional covariance penalties and the corresponding standard deviations are listed in Table 2.3. The means of all three estimation techniques are similar for all random effects variances. In many cases the *Steinian* lies between the cross-validation and the bootstrap estimate. In combination with the results in the selection frequency plot 2.6 this gives evidence to the superior behaviour of the *Steinian*. Although the covariance penalty for $\tau^2 = 0$ is smaller for the *Steinian* than for the bootstrap, the *Steinian* selects the null model more often than the bootstrap. So the *Steinian* seems to penalize in the "right" situations.

The convergence to a selection rate (of the more complex model) of one with rising signal-to-noise ratio τ^2 is fast, as can be seen in Figure 2.6. The distribution under consideration is close to the Gaussian for which the convergence rate is also high. However, the squared error function is also a possible influencing factor. Moreover, the *Steinian* has lower variance than the cross-validation and bootstrap in all settings. The reduced variability can also be observed in the boxplots in Figure 2.7.

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

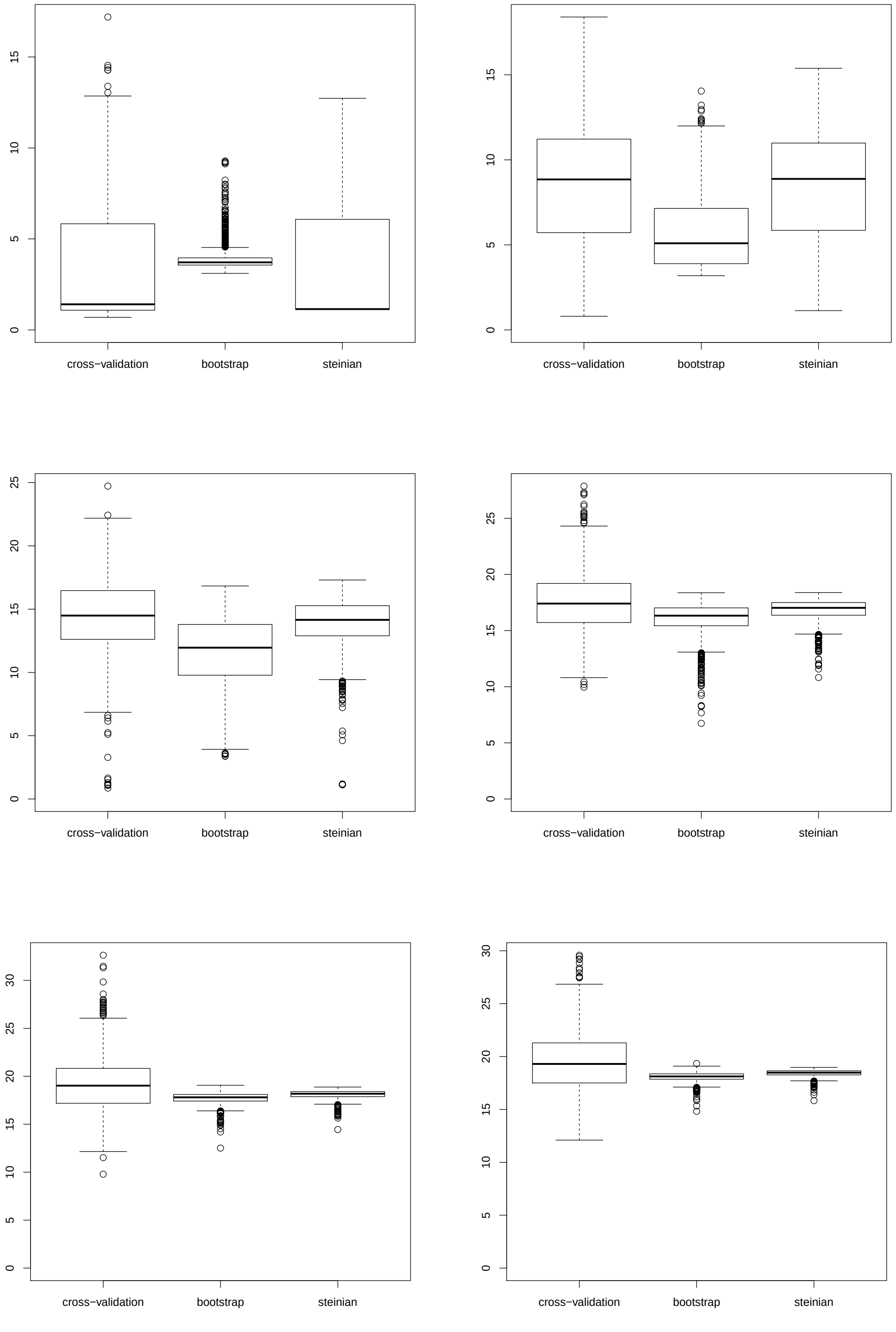

Figure 2.7: Boxplots of the estimated covariance penalties for 8 individuals, 6 observations per individual and variance of random effects equal to 0, 0.3, 0.6, 1, 1.5, 1.8 respectively.

Chapter 3

Degrees of freedom of the smoothing parameter

This chapter is concerned with developing the degrees of freedom for semiparametric regression models, taking into account the uncertainty induced by estimation of the smoothing parameter. The new representation of the degrees of freedom of semiparametric regression models will allow for differentiation between the degrees of freedom of the regression parameters and the degrees of freedom associated with the smoothing parameter.

These degrees of freedom of the smoothing parameter largely depend on the method that is used for smoothing parameter selection. Common methods are generalized cross-validation (GCV) and restricted maximum likelihood (REML) that are introduced in the first part of this chapter. The notation of Reiss and Ogden (2009) is convenient for the derivation of the degrees of freedom of a univariate smoothing parameter and will therefore be adopted in the first part. A spectral representation of the degrees of freedom of the smoothing parameter gives further insight into its genesis. Simulations show that the degrees of freedom associated with the smoothing parameter are larger than zero. Suggestions on how this could be proven theoretically are given.

In the second part, a more general notation with multiple smoothing parameters and a wide variety of possible penalty terms and smoothing parameter selection criteria are introduced. The geometry of the smoothing process is analysed and an interesting extension that permits an ostensive insight into the degrees of freedom of the smoothing parameter is given with the help of results from differential geometry.

The degrees of freedom of the smoothing parameters are explicitly calculated in the

general notation for four different smoothing parameter selection criteria. Based on these derivations, the degrees of freedom of the smoothing parameters are implemented, so that they can be used with the R-package `mgcv`, see Wood (2011).
In the simulation study this implementation is used to show how the degrees of freedom of the smoothing parameters impact model selection in semiparametric regression models. This is closely linked to the findings of Greven and Kneib (2010).

3.1 The degrees of freedom for penalized splines in two frameworks

Consider n observed values from the data-generating process

$$y_i = f(x_i) + \epsilon_i$$

for some smooth function $f(\cdot) : \mathbb{R} \longrightarrow \mathbb{R}$, $x_i \in \mathbb{R}$ and $\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$ for all $i = 1, \ldots, n$. In order to estimate (predict) the unknown function $f(\cdot)$ low rank spline smoothers are applied. This allows for a parametric representation of $f(\cdot)$ and the data can therefore be written in compact matrix notation

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon},$$

with $\boldsymbol{y} \in \mathbb{R}^n$ and full rank $(n \times p)$ and $(n \times q)$ matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$. The vector $\boldsymbol{\epsilon}$ consists of independent and identically distributed errors ϵ_i. We distinguish between the unpenalized parameters $\boldsymbol{\beta}$ and the penalized parameters $\boldsymbol{u}$.
The following results can be derived in a substantially simpler framework by assuming the subsequent assumptions to hold:

1. The design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are full rank $(n \times p)$- and $(n \times q)$-matrices respectively.

2. $\boldsymbol{X}$ and $\boldsymbol{Z}$ are orthogonal, i.e. $\boldsymbol{X}^t\boldsymbol{Z} = \boldsymbol{0}$.

3. $\boldsymbol{y}$ is not in the column space of $\boldsymbol{X}$.

For a large part of semiparametric regression models these assumptions do not hold. In the second part of this chapter, the results will be derived without the most prohibitive assumption (2) having to hold.

3.1.1 Estimation based on cross-validation

A penalty term $\lambda\boldsymbol{S}$, with smoothing parameter $\lambda \in (0,\infty)$ and positive definite penalty matrix $\boldsymbol{S}$, is associated to the penalized parameters. The regression parameters are estimated via minimization of the penalized least squares criterion

$$(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu})^t(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu})+\lambda\boldsymbol{u}^t\boldsymbol{S}\boldsymbol{u}. \tag{3.1.1}$$

In this setting the so-called hat-matrix

$$\boldsymbol{H}=\boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t-\boldsymbol{Z}\left(\boldsymbol{Z}^t\boldsymbol{Z}+\lambda\boldsymbol{S}\right)^{-1}\boldsymbol{Z}^t$$

and the fixed smoothing parameter residual matrix

$$\boldsymbol{P}=\boldsymbol{I}-\boldsymbol{H}=\boldsymbol{I}-\boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t+\boldsymbol{Z}\left(\boldsymbol{Z}^t\boldsymbol{Z}+\lambda\boldsymbol{S}\right)^{-1}\boldsymbol{Z}^t$$

play an important role for the choice of λ, see Reiss and Ogden (2009). The fixed smoothing parameter residual matrix $\boldsymbol{P}$ also substantially eases the derivation of the degrees of freedom and therefore will be used hereafter.

In order to find the right balance between fidelity to the data and smoothness of the function $f(\cdot)$ the smoothing parameter λ is chosen by the generalized cross-validation criterion proposed by Craven and Wahba (1978). The criterion in this setting is given by

$$GCV(\lambda)=\frac{(\boldsymbol{y}-\hat{\boldsymbol{y}})^t(\boldsymbol{y}-\hat{\boldsymbol{y}})/n}{(1-\mathrm{tr}(\boldsymbol{H})/n)^2}=n\frac{\boldsymbol{y}\boldsymbol{P}^2\boldsymbol{y}}{\mathrm{tr}(\boldsymbol{P})^2}.$$

It is an approximation of the leave-one-out cross-validation criterion. Minimizing this criterion we obtain the equation that the optimal smoothing parameter λ chosen by GCV has to fulfil, i.e.

$$\frac{d}{d\lambda}GCV(\lambda)=\mathrm{tr}(\boldsymbol{P}^2)\boldsymbol{y}^t\boldsymbol{P}^2\boldsymbol{y}-\mathrm{tr}(\boldsymbol{P})\boldsymbol{y}^t\boldsymbol{P}^3\boldsymbol{y}=0. \tag{3.1.2}$$

3.1.2 Empirical Bayes estimation

In the Bayesian framework, a noninformative prior is assigned to the unpenalized parameters, i.e. the density is proportional to a constant

$$p(\boldsymbol{\beta}) \propto c$$

and a Gaussian prior is assigned to the penalized parameters

$$\boldsymbol{u} \sim \mathcal{N}\left(\boldsymbol{0}, (\lambda \boldsymbol{S})^{-1}\right),$$

with positive definite covariance matrix $(\lambda \boldsymbol{S})^{-1}$. Maximizing the joint posterior distribution for known covariance matrix $(\lambda \boldsymbol{S})^{-1}$ corresponds to minimizing the penalized least squares criterion (3.1.1). In consequence, for a predefined secondary parameter λ the estimators of the regression coefficients coincide in both frameworks. The covariance parameter λ is assumed unknown but fixed in the empirical Bayes approach. In this case the covariance matrix of the data is

$$\operatorname{cov}(\boldsymbol{y}) = \boldsymbol{V} = \sigma^2 \boldsymbol{I} + \frac{1}{\lambda} \boldsymbol{Z} \boldsymbol{S}^{-1} \boldsymbol{Z}.$$

The covariance parameter λ can therefore be estimated by maximizing the restricted marginal log-likelihood

$$\begin{aligned}\log(p(\boldsymbol{y}|\lambda)) &\propto \log\left(\int p(y|\boldsymbol{\beta}, \boldsymbol{u}, \lambda) p(\boldsymbol{u}|\lambda) d\boldsymbol{u}\right) \\ &= -\frac{1}{2}\left(\log\left|\hat{\sigma}^2 \boldsymbol{V}\right| + \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^t \left(\hat{\sigma}^{-2} \boldsymbol{V}^{-1}\right) \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) + \log\left|\hat{\sigma}^{-2} \boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{X}\right|\right).\end{aligned}$$

In mixed model methodology this is known as the profile restricted maximum likelihood. The restricted maximum likelihood criterion that needs to be optimized in order to find the optimal λ can be written with the expressions for the estimated parameters plugged in as

$$REML(\lambda) = -\frac{1}{2}\left(\log|\boldsymbol{V}| + (n-p)\log\left(\boldsymbol{y}^t \boldsymbol{P} \boldsymbol{y}\right) + \log\left|\boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{X}\right|\right).$$

Thus the parameter λ has to fulfil

$$\frac{d}{d\lambda} REML(\lambda) = (n-p)\boldsymbol{y}^t \boldsymbol{P}^2 \boldsymbol{y} - \operatorname{tr}(\boldsymbol{P})\boldsymbol{y}^t \boldsymbol{P} \boldsymbol{y} = 0. \tag{3.1.3}$$

For the remainder we assume that the smoothing parameter fulfilling either equation (3.1.3) or equation (3.1.2) does not lie on the boundary of the parameter space, i.e. $\lambda \in (0,\infty)$.

3.1.3 General concept of degrees of freedom

The key to determine the correct degrees of freedom in semiparametric regression is to consider the dependence of the hat matrix on the smoothing parameter. The degrees of freedom are defined as the trace of the Jacobian matrix of the estimated values treated as functions of the data. Since the smoothing parameter depends on the data, the Jacobian is not equal to the hat matrix, but rather needs a correction term for the uncertainty induced by smoothing parameter selection. In the following the difference between λ and $\hat{\lambda}$ will be suppressed in order to simplify the notation. The correction term can be obtained with the help of the chain rule as follows:

$$\begin{aligned}
\operatorname{tr}\left(\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{y}}\right) &= \operatorname{tr}\left(\frac{\partial \boldsymbol{H}\boldsymbol{y}}{\partial \boldsymbol{y}}\right) \\
&= \operatorname{tr}(\boldsymbol{H}) + \operatorname{tr}\left(\frac{\partial \boldsymbol{H}}{\partial \lambda} \cdot \boldsymbol{y} \cdot \frac{\partial \lambda}{\partial \boldsymbol{y}^t}\right) \\
&= \operatorname{tr}(\boldsymbol{H}) + \frac{\partial \lambda}{\partial \boldsymbol{y}^t} \cdot \frac{\partial \boldsymbol{H}}{\partial \lambda} \cdot \boldsymbol{y}
\end{aligned}$$

The second part of the degrees of freedom term can be viewed as the contribution of uncertainty of the smoothing parameter towards the degrees of freedom. The most challenging part of the correction terms is the derivative of the smoothing parameter w.r.t. the data, i.e.

$$\gamma_{\boldsymbol{y}} = \frac{\partial \lambda}{\partial \boldsymbol{y}^t}.$$

This is due to the fact that the function $\lambda(\boldsymbol{y})$ is only implicitly defined by the criterion that is used to find the optimal smoothing parameter. Nevertheless, under the regularity conditions the derivative is analytically accessible with the help of the implicit function theorem: In general may $\mathcal{C}(\boldsymbol{y},\lambda)$ be the criterion the optimal smoothing parameter needs to fulfil. So $\mathcal{C}(\boldsymbol{y},\lambda)$ is either equal to (3.1.3) or (3.1.2). By the multidimensional chain rule it follows that

$$\frac{\partial \mathcal{C}(\boldsymbol{y},\lambda)}{\partial \boldsymbol{y}^t} + \frac{\partial \mathcal{C}(\boldsymbol{y},\lambda)}{\partial \lambda} \cdot \frac{\partial \lambda}{\partial \boldsymbol{y}^t} = 0$$

and hence

$$\frac{\partial \lambda}{\partial \boldsymbol{y}^t} = -\frac{\partial \mathcal{C}(\boldsymbol{y},\lambda)}{\partial \lambda}^{-1} \frac{\partial \mathcal{C}(\boldsymbol{y},\lambda)}{\partial \boldsymbol{y}^t}.$$

For the explicit calculations of the degrees of freedom of the smoothing parameter the gradients of $\mathcal{C}(\boldsymbol{y},\lambda)$ with respect to the data $\boldsymbol{y}$ and the smoothing parameter λ are

$$b = \frac{\partial \mathcal{C}(\boldsymbol{y},\lambda)}{\partial \lambda} \text{ and } \boldsymbol{g} = \frac{\partial \mathcal{C}(\boldsymbol{y},\lambda)}{\partial \boldsymbol{y}^t}.$$

Thus the implicit derivative in this case can be obtained via

$$\boldsymbol{\gamma_y} = -b^{-1}\boldsymbol{g}. \tag{3.1.4}$$

Noting that $\frac{\partial \boldsymbol{H}}{\partial \lambda} = -\frac{\partial \boldsymbol{P}}{\partial \lambda}$ and using the result from Reiss and Ogden (2009) Appendix A, i. e. $\frac{\partial \boldsymbol{P}}{\partial \lambda} = \lambda^{-1}\left(\boldsymbol{P} - \boldsymbol{P}^2\right)$, the degrees of freedom of the smoothing parameters are

$$df(\lambda) = \boldsymbol{\gamma_y} \cdot \frac{\partial \boldsymbol{H}}{\partial \lambda} \cdot \boldsymbol{y} = \boldsymbol{\gamma_y} \cdot \lambda^{-1}\left(\boldsymbol{P}^2 - \boldsymbol{P}\right) \cdot \boldsymbol{y}.$$

In case the optimizing criterion is GCV, the Hessian is given by

$$b = \lambda^{-1}\boldsymbol{y}^t\left[\left(3\boldsymbol{P}^4 - 4\boldsymbol{P}^3\right)\mathrm{tr}(\boldsymbol{P}) + \left(4\boldsymbol{P}^2 - \boldsymbol{P}^3\right)\mathrm{tr}(\boldsymbol{P}^2) - 2\boldsymbol{P}^2 tr(\boldsymbol{P}^3)\right]\boldsymbol{y}$$

and the derivative of the criterion w.r.t. the data $\boldsymbol{y}$ is given by

$$\boldsymbol{g} = 2\boldsymbol{y}^t\left(\mathrm{tr}(\boldsymbol{P}^2)\boldsymbol{P}^2 - \mathrm{tr}(\boldsymbol{P})\boldsymbol{P}^3\right)$$

In case the optimizing criterion is REML, the Hessian is given by

$$b = 2\lambda^{-1}\boldsymbol{y}^t\left[\left(\mathrm{tr}(\boldsymbol{P}^2) - 2\mathrm{tr}(\boldsymbol{P})\right)\boldsymbol{P} + (n - p + \mathrm{tr}(\boldsymbol{P}))\boldsymbol{P}^2 - (n-p)\boldsymbol{P}^3\right]\boldsymbol{y}$$

and the derivative of the criterion w.r.t. the data $\boldsymbol{y}$ is given by

$$\boldsymbol{g} = 2\boldsymbol{y}^t\left((n-p)\boldsymbol{P}^2 - \mathrm{tr}(\boldsymbol{P})\boldsymbol{P}\right).$$

3.2 Spectral representation of the degrees of freedom of the smoothing parameter

The spectral decomposition of the residual matrix yields $\boldsymbol{P} = \sum_{i=1}^{n} \nu_i \boldsymbol{v}_i \boldsymbol{v}_i^t$, with eigenvalues $\nu_1,\dots,\nu_n$ and eigenvectors $\boldsymbol{v}_1,\dots,\boldsymbol{v}_n$. As known from Reiss and Ogden (2009) the eigenvalues and corresponding eigenvectors can be split up into three groups:

i) The q positive eigenvalues $1 > \nu_1 = \frac{\lambda}{\lambda+\xi_1} \geq \ldots \geq \nu_q = \frac{\lambda}{\lambda+\xi_q}$, with the q eigenvalues of $\boldsymbol{Z}\boldsymbol{S}^{-1}\boldsymbol{Z}^t$ denoted by $\xi_1,\dots,\xi_q$.

ii) The $n-p-q$ eigenvalues $\nu_{q+1} = \ldots = \nu_{n-p} = 1$.

iii) The p eigenvalues $\nu_{n-p+1} = \ldots = \nu_n = 0$.

Denote by $\boldsymbol{w} = (w_1,\dots,w_n) = \left(\boldsymbol{y}^t\boldsymbol{v}_1,\dots,\boldsymbol{y}^t\boldsymbol{v}_n\right)$ the response data with respect to the basis formed by the eigendecomposition of $\boldsymbol{P}$. Define

$$Q_k = \boldsymbol{y}^t\boldsymbol{P}^k\boldsymbol{y} = \sum_{i=1}^{q} w_i^2 \left(\frac{\lambda}{\lambda+\xi_i}\right)^k + \sum_{i=q+1}^{n-p} w_i^2$$

and

$$t_k = tr\left(\boldsymbol{P}^k\right) = n-p-q+\sum_{i=1}^{q}\left(\frac{\lambda}{\lambda+\xi_i}\right)^k.$$

Thus we can rewrite the degrees of freedom of the smoothing parameter in terms of the spectral representation. Notice that the degrees of freedom of the smoothing parameter is essentially a fraction of two numbers. Hence, the numerator for the degrees of freedom of the smoothing parameter estimated by REML is

$$\begin{aligned}
\frac{\lambda}{2}\cdot\boldsymbol{g}\cdot\left(\boldsymbol{P}^2-\boldsymbol{P}\right)\boldsymbol{y} &= \boldsymbol{y}^t\left((n-p)\boldsymbol{P}^2 - \mathrm{tr}(\boldsymbol{P})\boldsymbol{P}\right)\cdot\left(\boldsymbol{P}^2-\boldsymbol{P}\right)\boldsymbol{y}\\
&= \boldsymbol{y}^t\left((n-p)\boldsymbol{P}^4 - \mathrm{tr}(\boldsymbol{P})\boldsymbol{P}^3\right)\boldsymbol{y} - \boldsymbol{y}^t\left((n-p)\boldsymbol{P}^3 - \mathrm{tr}(\boldsymbol{P})\boldsymbol{P}^4\right)\boldsymbol{y}\\
&= (n-p)\left(\boldsymbol{y}^t\boldsymbol{P}^4\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{P}^3\boldsymbol{y}\right) + \mathrm{tr}(\boldsymbol{P})\left(\boldsymbol{y}^t\boldsymbol{P}^2\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{P}^3\boldsymbol{y}\right)\\
&= t_0\left(Q_4-Q_3\right) + t_1\left(Q_2-Q_3\right),
\end{aligned}$$

in the preceding notation of the spectral representation. For the denominator of the degrees of freedom of the smoothing parameter estimated by REML we have

$$\begin{aligned}\frac{\lambda}{2}\cdot b &= \boldsymbol{y}^t\left[\left(\mathrm{tr}(\boldsymbol{P}^2)-2\mathrm{tr}(\boldsymbol{P})\right)\boldsymbol{P}+(n-p+\mathrm{tr}(\boldsymbol{P}))\,\boldsymbol{P}^2-(n-p)\boldsymbol{P}^3\right]\boldsymbol{y}\\ &= (n-p)\left(\boldsymbol{y}^t\boldsymbol{P}^2\boldsymbol{y}-\boldsymbol{y}^t\boldsymbol{P}^3\boldsymbol{y}\right)+\mathrm{tr}(\boldsymbol{P})\left(\boldsymbol{y}^t\boldsymbol{P}^2\boldsymbol{y}-\boldsymbol{y}^t\boldsymbol{P}\boldsymbol{y}\right)+\boldsymbol{y}^t\boldsymbol{P}\boldsymbol{y}\left(\mathrm{tr}(\boldsymbol{P}^2)-\mathrm{tr}(\boldsymbol{P})\right)\\ &= t_0\,(Q_2-Q_3)+t_1\,(Q_2-Q_1)+Q_1\,(t_2-t_1)\,.\end{aligned}$$

Combining both results, the degrees of freedom of the smoothing parameter based on REML estimation in compact spectral representation are

$$df(\lambda)=\frac{\boldsymbol{g}\cdot\left(\boldsymbol{P}^2-\boldsymbol{P}\right)\boldsymbol{y}}{b}=\frac{t_0\,(Q_4-Q_3)+t_1\,(Q_2-Q_3)}{t_0\,(Q_2-Q_3)+t_1\,(Q_2-Q_1)+Q_1\,(t_2-t_1)}.$$

Along the same lines the degrees of freedom of the smoothing parameter estimated with GCV can be derived in the compact spectral notation. Doing so the numerator is

$$\begin{aligned}\frac{\lambda}{2}\cdot\boldsymbol{g}\cdot\left(\boldsymbol{P}^2-\boldsymbol{P}\right)\boldsymbol{y} &= \boldsymbol{y}^t\left(\mathrm{tr}(\boldsymbol{P}^2)\boldsymbol{P}^2-\mathrm{tr}(\boldsymbol{P})\boldsymbol{P}^3\right)\cdot\left(\boldsymbol{P}^2-\boldsymbol{P}\right)\boldsymbol{y}\\ &= \mathrm{tr}(\boldsymbol{P}^2)\boldsymbol{y}^t\boldsymbol{P}^4\boldsymbol{y}-\mathrm{tr}(\boldsymbol{P})\boldsymbol{y}^t\boldsymbol{P}^5\boldsymbol{y}-\mathrm{tr}(\boldsymbol{P}^2)\boldsymbol{y}^t\boldsymbol{P}^3\boldsymbol{y}+\mathrm{tr}(\boldsymbol{P})\boldsymbol{y}^t\boldsymbol{P}^4\boldsymbol{y}\\ &= t_1\,(Q_4-Q_5)+t_2\,(Q_4-Q_3)\,.\end{aligned}$$

The denominator, if smoothing parameter estimation is done with GCV, is

$$\begin{aligned}\lambda\cdot b &= \boldsymbol{y}^t\left[\left(3\boldsymbol{P}^4-4\boldsymbol{P}^3\right)\mathrm{tr}(\boldsymbol{P})+\left(4\boldsymbol{P}^2-\boldsymbol{P}^3\right)\mathrm{tr}(\boldsymbol{P}^2)-2\boldsymbol{P}^2tr(\boldsymbol{P}^3)\right]\boldsymbol{y}\\ &= (3Q_4-4Q_3)\,t_1+(4Q_2-Q_3)\,t_2-2Q_2t_3.\end{aligned}$$

Thus the degrees of freedom of the smoothing parameter in spectral representation with GCV-based smoothing parameter estimation is

$$df(\lambda)=\frac{\boldsymbol{g}\cdot\left(\boldsymbol{P}^2-\boldsymbol{P}\right)\boldsymbol{y}}{b}=2\frac{t_1\,(Q_4-Q_5)+t_2\,(Q_4-Q_3)}{(3Q_4-4Q_3)\,t_1+(4Q_2-Q_3)\,t_2-2Q_2t_3}.$$

3.2.1 Is $df(\boldsymbol{\lambda})$ always larger than zero?

An apparent question is whether the degrees of freedom associated with the smoothing parameter are always positive. The degrees of freedom that are associated with the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are always positive, as can be seen from the spectral representation, since

$$df(\boldsymbol{\beta},\boldsymbol{u}) = \operatorname{tr}(\boldsymbol{H}) = p+q-\sum_{i=1}^{q}\left(\frac{\lambda}{\lambda+\xi_i}\right) \geq 0.$$

For the degrees of freedom of the smoothing parameter this is not obvious. Extensive simulation based on the model in Gu and Wahba (1991) shows that this is probable. More explicitly, the simulations suggest that

$$\boldsymbol{g}\cdot\left(\boldsymbol{P}^2-\boldsymbol{P}\right)\boldsymbol{y} \geq 0 \quad\text{and}\quad b \leq 0.$$

A proof of these results is not available at the moment. A possible approach might be the following: Suppose the smoothing parameter was estimated by optimization of the REML criterion. Then, using the derivative of the REML criterion set to zero (3.1.3) it follows that in the spectral representation the equation

$$\frac{t_0}{t_1} = \frac{Q_1}{Q_2} \tag{3.2.1}$$

is fulfilled. The first step to prove that $df(\lambda) \geq 0$ could be to show that

$$\begin{aligned} \boldsymbol{g}\cdot\left(\boldsymbol{P}^2-\boldsymbol{P}\right)\boldsymbol{y} &\geq 0 \\ \Leftrightarrow\quad \underbrace{\frac{t_0}{t_0+t_1}}_{>\frac{1}{2}}Q_4+\underbrace{\frac{t_1}{t_0+t_1}}_{<\frac{1}{2}}Q_2 &\geq Q_3. \end{aligned}$$

However, since $Q_{k-1} \geq Q_k \geq Q_{k+1}$, it is not clear why this is the case. Another possibility is to use (3.2.1)

$$\begin{aligned} \boldsymbol{g} \cdot \left(\boldsymbol{P}^2 - \boldsymbol{P}\right)\boldsymbol{y} &\geq 0 \\ \Leftrightarrow \frac{Q_3 - Q_2}{Q_4 - Q_3} &\geq \frac{t_0}{t_1} \\ \Leftrightarrow \frac{Q_3 - Q_2}{Q_4 - Q_3} &\geq \frac{Q_1}{Q_2}, \end{aligned}$$

considering that the differences of the quadratic forms can be adequately expressed by

$$Q_k - Q_{k+1} = \sum_{i=1}^{q} w_i^2 \left(\frac{\lambda}{\lambda + \xi_i}\right)^k \left(\frac{\xi_i}{\lambda + \xi_i}\right).$$

Despite these results it seems that some link between $\lambda \in (0, \infty)$, q, p and the ξ_is is needed in order to prove the assumption.

3.3 Degrees of freedom in case of multivariate λ

To generalize the previous results we consider data $\boldsymbol{y} \in \mathbb{R}^n$ that is generated by a data generating process

$$\boldsymbol{y} = \sum_{k=1}^{s} \mathcal{I}_k(f) + \boldsymbol{\epsilon} \tag{3.3.1}$$

with f being r-times continuously differentiable functions of m covariates and $\mathcal{I}_k$ are bounded linear functionals. The error term is normally distributed with mean zero and covariance matrix $\sigma^2\boldsymbol{R}$, i.e. $\boldsymbol{\epsilon} \sim N\left(\boldsymbol{0}, \sigma^2\boldsymbol{R}\right)$. The bounded linear functionals may be rather different. Possible examples are evaluation functionals on $\mathbb{R}^2$, i.e.

$$\mathcal{I}_{ki}(f) = f_k(z_{1i}, z_{2i}),\ i = 1, \ldots, n,$$

or, for instance, linear functionals from the vector space $C^2[a,b]$ of two times continuously differentiable functions on the interval $[a,b]$ to the real numbers, i.e.

$$\mathcal{I}_{ki}(f) = \int_a^b f_k(z) g_i(z) dz\ , i = 1, \ldots, n, \tag{3.3.2}$$

for known functions $g_i \in C^2[a,b]$.

To estimate models such as (3.3.1) in practice, low ranked roughness penalized basis expansion methods are applied. This leads to the semiparametric linear model that most can generally be expressed as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.3.3}$$

where $\boldsymbol{X}$ is a $(n \times (p+q))$ design matrix and $\boldsymbol{\beta}$ is an unknown parameter vector of dimension $p+q$. Both $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are combined matrices and parameters of those parameters and matrices associated with the s terms in (3.3.1). To each of the terms we associate a penalty term. The combined penalty matrix $\boldsymbol{S} = \sum_{k=1}^{s} \lambda_k \boldsymbol{S}_k$ is of rank q and its null space has dimension p. Thus model (3.3.3) is fitted by the optimization problem

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \sum_{k=1}^{s} \lambda_k \boldsymbol{\beta}^t \boldsymbol{S}_k \boldsymbol{\beta}. \tag{3.3.4}$$

This formulation is general, as it does not distinguish between the penalized and unpenalized parts of the parameter vector $\boldsymbol{\beta}$.

For a multivariate smoothing parameter, i.e. multiple penalty terms associated with the model, the correct degrees can be specified by

$$\operatorname{tr}\left(\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{y}}\right) = \operatorname{tr}(\boldsymbol{H}) + \sum_{k=1}^{s} \frac{\partial \lambda_k}{\partial \boldsymbol{y}^t} \cdot \frac{\partial \boldsymbol{H}}{\partial \lambda_k} \cdot \boldsymbol{y}.$$

Let h_{ij} indicate the element of the hat matrix $\boldsymbol{H}$ in row i and column j. The degrees of freedom are deduced from

$$\begin{aligned}
\frac{\partial \hat{y}_i}{\partial y_i} &= \frac{\partial \sum_{j=1}^{n} h_{ij} y_j}{\partial y_i} \\
&= \sum_{j=1}^{n} \frac{\partial h_{ij} y_j}{\partial y_i} \\
&= \sum_{j=1}^{n} \left(\frac{\partial h_{ij}}{\partial y_i} y_j + h_{ij} \frac{\partial y_j}{\partial y_i} \right) \\
&= \sum_{j=1}^{n} \frac{\partial h_{ij}}{\partial y_i} y_j + h_{ii}.
\end{aligned}$$

Thus

$$\begin{aligned}
\operatorname{tr}\left(\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{y}}\right) &= \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial h_{ij}}{\partial y_i} y_j + h_{ii} \\
&= \sum_{i=1}^{n} h_{ii} + \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial h_{ij}}{\partial y_i} y_j \\
&= \operatorname{tr}(\boldsymbol{H}) + \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial h_{ij}}{\partial \lambda}\frac{\partial \lambda}{\partial y_i} y_j \\
&= \operatorname{tr}(\boldsymbol{H}) + \sum_{i=1}^{n} \frac{\partial \lambda}{\partial y_i}\frac{\partial \boldsymbol{h}_{j\cdot}}{\partial \lambda} \boldsymbol{y} \\
&= \operatorname{tr}(\boldsymbol{H}) + \frac{\partial \lambda}{\partial \boldsymbol{y}^t}\frac{\partial \boldsymbol{H}}{\partial \lambda} \boldsymbol{y}.
\end{aligned}$$

Since the smoothing parameters λ_j need to be positive, the optimization for finding the optimal smoothing parameter is often done with respect to the log smoothing parameter $\rho_j = \log(\lambda_j)$ for reasons of simplicity, see Wood (2008).
For a multivariate log smoothing parameter $\boldsymbol{\rho} \in \mathbb{R}^s$ the degrees of freedom are

$$\begin{aligned}
\operatorname{tr}\left(\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{y}}\right) &= \operatorname{tr}(\boldsymbol{H}) + \sum_{k=1}^{s}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial h_{ij}}{\partial \lambda_k}\frac{\partial \lambda_k}{\partial y_i} y_j \\
&= \operatorname{tr}(\boldsymbol{H}) + \sum_{k=1}^{s}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial h_{ij}}{\partial \lambda_k}\frac{\partial \lambda_k}{\partial \rho_k}\frac{\partial \rho_k}{\partial y_i} y_j \\
&= \operatorname{tr}(\boldsymbol{H}) + \sum_{k=1}^{s} \lambda_k \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial h_{ij}}{\partial \lambda_k}\frac{\partial \rho_k}{\partial y_i} y_j \\
&= \operatorname{tr}(\boldsymbol{H}) + \sum_{k=1}^{s} \lambda_k \frac{\partial \rho_k}{\partial \boldsymbol{y}^t}\frac{\partial \boldsymbol{H}}{\partial \lambda_k} \boldsymbol{y}.
\end{aligned}$$

The degrees of freedom cannot only be written as the sensitivity of the estimated values with respect to the observed data, but can also be expressed in terms of the regression parameters, i.e.

$$\operatorname{tr}\left(\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{y}^t}\boldsymbol{X}\right) + \sum_{k=1}^{s} \frac{\partial \lambda_k}{\partial \boldsymbol{y}^t}\frac{\partial \boldsymbol{X}\hat{\boldsymbol{\beta}}}{\partial \lambda_k}. \tag{3.3.5}$$

This formulation highlights the fact that the first part of the degrees of freedom term

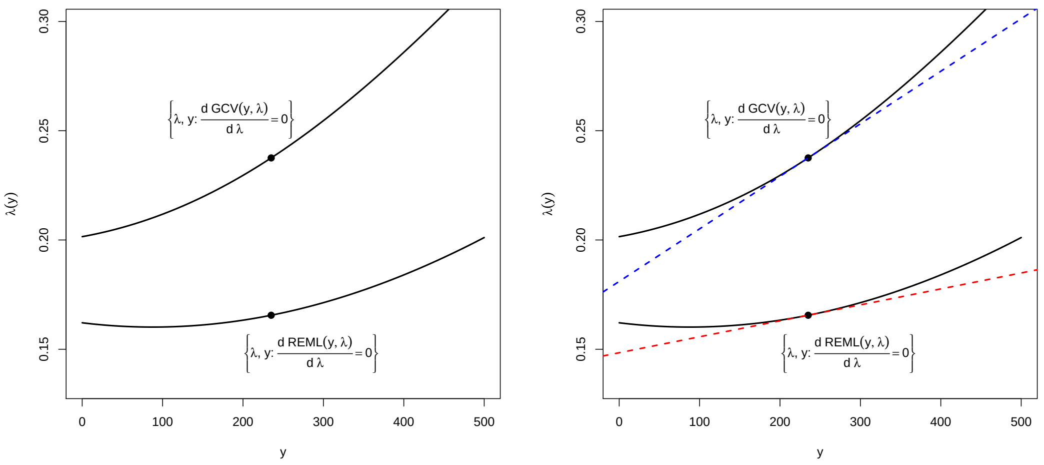

Figure 3.1: Interpretations of the smoothing parameter selection: 1. The implicitly defined function $\lambda(\boldsymbol{y})$ for one data point for $REML$ and GCV criterion. 2. The set of all values of the smoothing parameter and one data point that fulfil the $REML$ or GCV criterion. On the right hand side additionally the derivatives of the implicit functions at the point y_i.

contributes towards the regression coefficients and it is known that $df(\beta_j) \leq 1$ with penalization leading to a reduction of the degrees of freedom of the regression parameters. On the other hand, the second part corresponds to the degrees of freedom of the smoothing parameter λ and consists of two parts. The first one $\frac{\partial \lambda}{\partial \boldsymbol{y}^t}$ is a measure of the sensitivity of the smoothing parameter towards the data. The second part $\frac{\partial \boldsymbol{X}\hat{\boldsymbol{\beta}}}{\partial \lambda}$ accounts for the effect of the smoothing parameter on the regression parameters. Thus, the degrees of freedom that correspond to the k-th smoothing parameter are defined by

$$df(\lambda_k) = \lambda_k \frac{\partial \rho_k}{\partial \boldsymbol{y}^t} \frac{\partial \boldsymbol{H}}{\partial \rho_k} \boldsymbol{y}.$$

The most challenging part in the preceding formula in terms of computation is the derivative of the log smoothing parameter with respect to the data, i.e. $\frac{\partial \rho_k}{\partial \boldsymbol{y}^t}$. The function $\rho_k(\boldsymbol{y})$ is only implicitly defined, see Figure 3.1. Thus we need the implicit function theorem. Let

$$\boldsymbol{\Gamma}_{\boldsymbol{y}} = \frac{\partial \boldsymbol{\rho}}{\partial \boldsymbol{y}^t}$$

be the $(s \times n)$-matrix containing the derivatives of the log smoothing parameters with respect to the data. Then

$$\boldsymbol{B}\boldsymbol{\Gamma}_{\boldsymbol{y}} = -\boldsymbol{G}, \tag{3.3.6}$$

where $\boldsymbol{B}$ is the Hessian of the log smoothing parameters and $\boldsymbol{G}$ is the matrix containing the derivatives of the optimization criterion (for instance, REML$(\boldsymbol{y},\boldsymbol{\rho})$ or GCV$(\boldsymbol{y},\boldsymbol{\rho})$). More generally, let $\mathcal{C}(\boldsymbol{y},\boldsymbol{\rho})$ be the derivatives of the optimization criterion with respect to the log smoothing parameters and the data, then

$$\boldsymbol{G} = \frac{\partial \mathcal{C}(\boldsymbol{y},\boldsymbol{\rho})}{\partial \boldsymbol{y}^t} \quad \text{and} \quad \boldsymbol{B} = \frac{\partial \mathcal{C}(\boldsymbol{y},\boldsymbol{\rho})}{\partial \boldsymbol{\rho}^t} \tag{3.3.7}$$

with

$$\mathcal{C}(\boldsymbol{y},\boldsymbol{\rho}) = \frac{\partial REML(\boldsymbol{y},\boldsymbol{\rho})}{\partial \boldsymbol{\rho}} \quad \text{or} \quad \mathcal{C}(\boldsymbol{y},\boldsymbol{\rho}) = \frac{\partial GCV(\boldsymbol{y},\boldsymbol{\rho})}{\partial \boldsymbol{\rho}}.$$

Hence, if the criterion has a maximum at $\hat{\boldsymbol{\rho}}$, the Hessian $\boldsymbol{B}$ with respect to the smoothing parameter will be negative definite. This is the case if the criterion $REML(\boldsymbol{\rho},\boldsymbol{y})$ is maximized. If, on the other hand, the criterion has a minimum at $\hat{\boldsymbol{\rho}}$, as for GCV, the Hessian $\boldsymbol{B}$ with respect to the smoothing parameter will be positive definite. However, in the following we will assume that $\boldsymbol{B}$ is positive definite since otherwise we can use the implicit function theorem with $-\boldsymbol{B}$ and $-\boldsymbol{G}$ with the same result $\boldsymbol{\Gamma}_{\boldsymbol{y}}$. Since $\boldsymbol{B}$ is therefore symmetric and positive definite, the Cholesky-decomposition can be applied, i.e.

$$\boldsymbol{B} = \boldsymbol{L}\boldsymbol{L}^t. \tag{3.3.8}$$

The solution can thus be obtained by a series of forward and backward solves. Thus, first obtain

$$\boldsymbol{A} = -\boldsymbol{L}^{-1}\boldsymbol{G}$$

and then solve the equations

$$\boldsymbol{L}^t\boldsymbol{\Gamma}_{\boldsymbol{y}} = -\boldsymbol{A},$$

in order to efficiently obtain the derivative matrix of the log smoothing parameters w.r.t. the data.

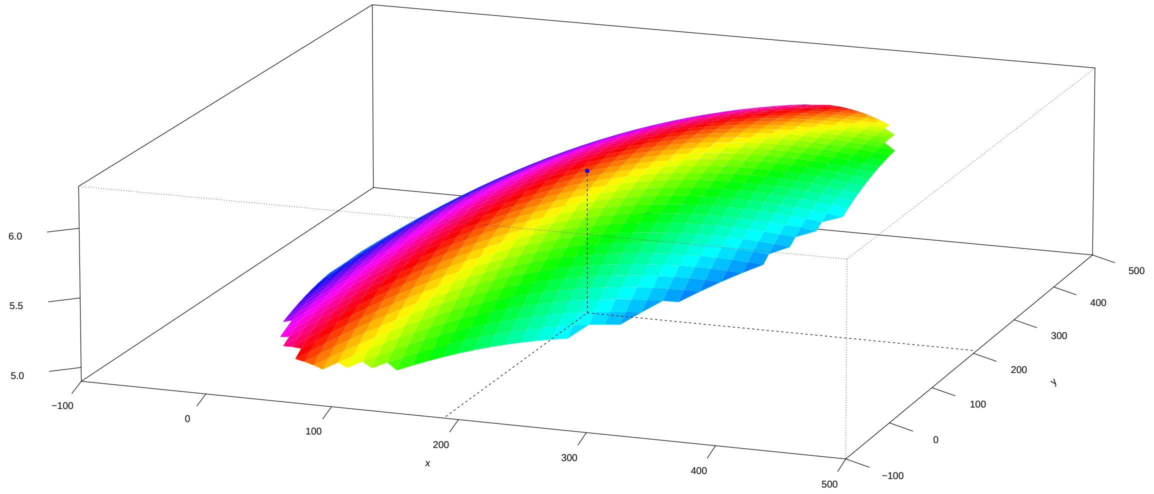

Figure 3.2: The 2-dimensional submanifold $\mathcal{M}$ of the Euclidean $\mathbb{R}^3$ defined by the criterion $\mathcal{C}$. Here all values of $\boldsymbol{y}$ are fixed except for two. The x-axis corresponds to y_1, the y-axis to y_2 and the z-axis corresponds to the smoothing parameter.

3.3.1 Geometry of the degrees of freedom of the smoothing parameter

Smoothing parameter selection has an interesting geometrical interpretation. Roughly speaking, if the response data is varied in the neighbourhood of the observed response, then the data and the s corresponding optimal (log) smoothing parameters define a (differentiable) manifold of dimension n in the $(n+s)$-dimensional space. This can be adequately visualised in particular for $s = 1$, varying only one or two of the response data points while keeping the remaining data fixed. For dimension one the manifold is a curve, that is given in Figure 3.1. The curves are plotted for REML and GCV and can either be interpreted as (smoothing parameter) functions of the varied response data point or as a set of all values in the two-dimensional space that fulfil the REML or GCV criterion. A two-dimensional manifold, in which two response data points are varied and the remaining data points are fixed, is given in Figure 3.2. In this case the manifold is a two-dimensional surface in the three-dimensional space.

This geometric interpretation of the smoothing parameter selection process also gives a geometric understanding of the degrees of freedom of the (log) smoothing parameter, or more precisely of the sensitivity of the (log) smoothing parameter towards the observed response variables. By calculating the derivatives of the (log) smoothing

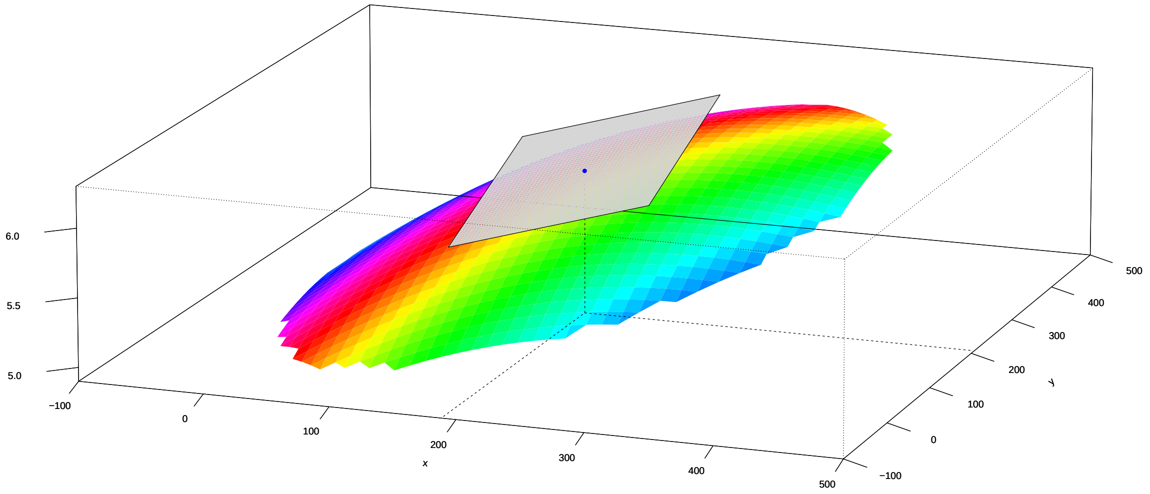

Figure 3.3: The same manifold $\mathcal{M}$ as in Figure 3.2 and the affine tangent space $\boldsymbol{a} + T_{\boldsymbol{a}}\mathcal{M}$ at point $\boldsymbol{a}$ defined by the criterion $\mathcal{C}$.

parameters with respect to the data the tangent space of the manifold at the observed response data point and the corresponding (log) smoothing parameters is obtained. For the one dimensional manifold in the two-dimensional plane the tangent space is a line, as can be seen in the right panel in Figure 3.1. For the two-dimensional manifold in the three dimensional space the tangent space corresponds to a plane through the observed response data points and the corresponding (log) smoothing parameter, see Figure 3.3.1. This interpretation links the smoothing parameter selection process and the degrees of freedom of the smoothing parameter to the field of differential geometry.

A more formal description of the geometric structure of the smoothing process is given in the following:
Suppose we observed data $\boldsymbol{y} = (y_1, \ldots, y_n)$ and let $U_\epsilon = \{\boldsymbol{x} | \, \|\boldsymbol{x} - \boldsymbol{y}\| < \epsilon\}$ be an open ball around $\boldsymbol{y}$ and assume that for all $\boldsymbol{x} \in U_\epsilon$ the differentiable criterion $\mathcal{C}\,(\boldsymbol{y}, \boldsymbol{\rho})$ has a unique and real optimum. Define $\mathcal{M} := \{(\boldsymbol{x}, \boldsymbol{\rho}) \in U_\epsilon \subset \mathbb{R}^{n+s} : \mathcal{C}\,(\boldsymbol{x}, \boldsymbol{\rho}) = 0\}$. Then $\mathcal{M}$ is a n-dimensional differentiable manifold of $\mathbb{R}^{n+s}$.

Proof. Since the j-th component of the criterion $\mathcal{C}_j\,(\boldsymbol{y}, \boldsymbol{\rho}) : U \longrightarrow \mathbb{R}, j = 1 \ldots, s$ is differentiable by definition and for all $\boldsymbol{x} \in \mathcal{M}$ the $s \times s$ Hessian at $\boldsymbol{x} \in \mathcal{M}$

$$\boldsymbol{B}_{\boldsymbol{x}} = \frac{\partial \mathcal{C}\,(\boldsymbol{x}, \boldsymbol{\rho})}{\partial \boldsymbol{\rho}}$$

is positive definite ($\mathcal{C}$ has a relative maximum) and hence has rank s, $\mathcal{M}$ is a differentiable manifold by the definition of a manifold in Forster (1999), page 128. $\square$

The assumptions can even be made more precise. The condition that needs to be fulfilled for $\mathcal{M}$ to be a differentiable manifold is that the combined matrices $(\boldsymbol{G_x} : \boldsymbol{B_x})$ must have full rank s for all $\boldsymbol{x} \in \mathcal{M}$:

$$\operatorname{rank}(\boldsymbol{G_x} : \boldsymbol{B_x}) = \operatorname{rank}\begin{pmatrix} \frac{\partial \mathcal{C}_1(\boldsymbol{x},\boldsymbol{\rho})}{\partial x_1} & \cdots & \frac{\partial \mathcal{C}_1(\boldsymbol{x},\boldsymbol{\rho})}{\partial x_n} & \frac{\partial \mathcal{C}_1(\boldsymbol{x},\boldsymbol{\rho})}{\partial \rho_1} & \cdots & \frac{\partial \mathcal{C}_1(\boldsymbol{x},\boldsymbol{\rho})}{\partial \rho_s} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{C}_s(\boldsymbol{x},\boldsymbol{\rho})}{\partial x_1} & \cdots & \frac{\partial \mathcal{C}_s(\boldsymbol{x},\boldsymbol{\rho})}{\partial x_n} & \frac{\partial \mathcal{C}_s(\boldsymbol{x},\boldsymbol{\rho})}{\partial \rho_1} & \cdots & \frac{\partial \mathcal{C}_s(\boldsymbol{x},\boldsymbol{\rho})}{\partial \rho_s} \end{pmatrix} = s.$$

or equivalent that the gradients of $\mathcal{C}$,

$$\begin{pmatrix} \frac{\partial \mathcal{C}_1(\boldsymbol{x},\boldsymbol{\rho})}{\partial x_1} \\ \vdots \\ \frac{\partial \mathcal{C}_1(\boldsymbol{x},\boldsymbol{\rho})}{\partial x_n} \\ \frac{\partial \mathcal{C}_1(\boldsymbol{x},\boldsymbol{\rho})}{\partial \rho_1} \\ \vdots \\ \frac{\partial \mathcal{C}_1(\boldsymbol{x},\boldsymbol{\rho})}{\partial \rho_s} \end{pmatrix}, \ldots, \begin{pmatrix} \frac{\partial \mathcal{C}_s(\boldsymbol{x},\boldsymbol{\rho})}{\partial x_1} \\ \vdots \\ \frac{\partial \mathcal{C}_s(\boldsymbol{x},\boldsymbol{\rho})}{\partial x_n} \\ \frac{\partial \mathcal{C}_s(\boldsymbol{x},\boldsymbol{\rho})}{\partial \rho_1} \\ \vdots \\ \frac{\partial \mathcal{C}_s(\boldsymbol{x},\boldsymbol{\rho})}{\partial \rho_s} \end{pmatrix},$$

are linearly independent for every $\boldsymbol{x} \in \mathcal{M}$.
The tangent space of a manifold at point $(\boldsymbol{y}, \hat{\boldsymbol{\rho}}) = \boldsymbol{a} \in \mathcal{M} \subset \mathbb{R}^{n+s}$ can be defined by, see Forster (1999), page 148:

$$T_{\boldsymbol{a}}\mathcal{M} := \{\boldsymbol{v} \in \mathbb{R}^{n+s} : \left\langle \boldsymbol{v}, \frac{\partial \mathcal{C}_j(\boldsymbol{y},\boldsymbol{\rho})}{\partial \boldsymbol{a}} \right\rangle = 0 \text{ for } j = 1, \ldots, s\},$$

with

$$\left\langle \boldsymbol{v}, \frac{\partial \mathcal{C}_j(\boldsymbol{y},\boldsymbol{\rho})}{\partial \boldsymbol{a}} \right\rangle = v_1 \cdot \frac{\partial \mathcal{C}_j(\boldsymbol{y},\boldsymbol{\rho})}{\partial a_1} + \ldots + v_{n+s} \cdot \frac{\partial \mathcal{C}_j(\boldsymbol{y},\boldsymbol{\rho})}{\partial a_{n+s}} =$$
$$v_1 \cdot \frac{\partial \mathcal{C}_j(\boldsymbol{y},\boldsymbol{\rho})}{\partial y_1} + \ldots + v_n \cdot \frac{\partial \mathcal{C}_j(\boldsymbol{y},\boldsymbol{\rho})}{\partial y_n} + v_{n+1} \cdot \frac{\partial \mathcal{C}_j(\boldsymbol{y},\boldsymbol{\rho})}{\partial \hat{\rho}_1} + \ldots + v_{n+s} \cdot \frac{\partial \mathcal{C}_j(\boldsymbol{y},\boldsymbol{\rho})}{\partial \hat{\rho}_s}.$$

Suppose now $\dim(\boldsymbol{\rho}) = 1$, then an element of the tangent space can be derived by setting all v_k to zero except the i-th and the $(n+1)$-th. If $v_{n+1} = 1$ for the i-th component v_i it has to follow that

$$v_i = -\frac{\frac{\partial \mathcal{C}(\boldsymbol{y},\rho)}{\partial \rho}}{\frac{\partial \mathcal{C}(\boldsymbol{y},\rho)}{\partial y_i}} = \left(\frac{\partial \rho}{\partial y_i}\right)^{-1} = \boldsymbol{\gamma}_{y_i}^{-1},$$

with $\boldsymbol{\gamma}_{y_i}$ the i-th element of the derivative vector $\boldsymbol{\gamma}_{\boldsymbol{y}}$, see (3.1.4). Hence, a basis of the n-dimensional tangent space are thus given by

$$\begin{pmatrix} \boldsymbol{\gamma}_{y_1}^{-1} \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ \boldsymbol{\gamma}_{y_2}^{-1} \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \dots, \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \boldsymbol{\gamma}_{y_n}^{-1} \\ 1 \end{pmatrix}.$$

Thus, by calculating the degrees of freedom of the smoothing parameter a basis of the tangent space $T_a\mathcal{M}$ is derived.

For multivariate log smoothing parameter the basis of the tangent space can also be given in terms of the matrix $\boldsymbol{\Gamma}_{\boldsymbol{y}}$ containing the derivatives of the log smoothing parameters with respect to the data. Notice that the tangent space $T_{\boldsymbol{a}}\mathcal{M}$ can be seen as the kernel of the combined matrices $\boldsymbol{G}$ and $\boldsymbol{B}$ with dimensions $s \times n$ and $s \times s$ from equation (3.3.7):

$$T_{\boldsymbol{a}}\mathcal{M} = \ker(\boldsymbol{G} : \boldsymbol{B}) = \{\boldsymbol{v} \in \mathbb{R}^{n+s} : (\boldsymbol{G} : \boldsymbol{B})\,\boldsymbol{v} = \boldsymbol{0}\}.$$

By equation (3.3.6), this is equivalent to

$$\ker(-\boldsymbol{B}\boldsymbol{\Gamma}_{\boldsymbol{y}} : \boldsymbol{B}).$$

Setting $v_i = 1$ and $v_k = 0$ for $k = 1,\dots,i-1,i+1,\dots,n$, the remaining $v_{n+1}\dots v_{n+s}$ are implicitly given by

$$\sum_{j=1}^{s}(v_{n+j} - \frac{\partial \rho_j}{\partial y_i})\frac{\partial \mathcal{C}_l(\boldsymbol{y},\boldsymbol{\rho})}{\partial \rho_j} = 0,\ l \ = \ 1,\dots,s.$$

Hence by setting $v_{n+j} = \frac{\partial \rho_j}{\partial y_i} = \boldsymbol{\Gamma}_{\boldsymbol{y}}^{i,j}$, where $\boldsymbol{\Gamma}_{\boldsymbol{y}}^{i,j}$ denotes the i,j-th element of the matrix $\boldsymbol{\Gamma}_{\boldsymbol{y}}$, an n-dimensional basis of $T_{\boldsymbol{a}}\mathcal{M}$ is

$$\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ \mathbf{\Gamma}_{\boldsymbol{y}}^{1,1} \\ \vdots \\ \mathbf{\Gamma}_{\boldsymbol{y}}^{1,s} \end{pmatrix}, \dots, \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \mathbf{\Gamma}_{\boldsymbol{y}}^{n,1} \\ \vdots \\ \mathbf{\Gamma}_{\boldsymbol{y}}^{n,s} \end{pmatrix}.$$

3.4 Four explicit calculations of the degrees of freedom

The derivation of the degrees of freedom of the smoothing parameters under rather general settings needs longer derivations than in those in the first part of this chapter. A number of smoothing parameter selection criteria are very common in practice. For these we derive the degrees of freedom in a general setting, i.e. multiple smoothing parameters and assumption (2) do not need to hold. In order to make these derivations broadly applicable we follow the notations in Wood (2008) and Wood (2011). Some of the derivatives are already stated in Wood (2008) and Wood (2011). Nevertheless, we explicitly give them for the special case of Gaussian data.
Based on these calculations an implementation in R for `gamObjects` from the R-package `mgcv`, version `1.8-2` was developed.

3.4.1 Restricted Maximum Likelihood

The negative restricted maximum log likelihood criterion that ought to be minimized, as is given in Wood (2011) formula (4), is

$$-l_r(\boldsymbol{\rho}, \sigma^2) = \frac{D_p(\hat{\boldsymbol{\beta}})}{2\sigma^2} + K + \frac{n-p}{2} \log(2\pi\sigma^2).$$

with

$$2K = \log\left|\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{S}\right| - \log|\boldsymbol{S}|_+$$

where $|\boldsymbol{S}|_+$ denotes the product of non-zero eigenvalues of $\boldsymbol{S}$ and d the dimension of the null space of $\boldsymbol{S}$. The penalized deviance is

$$D_p(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^t \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) + \hat{\boldsymbol{\beta}}^t\boldsymbol{S}\hat{\boldsymbol{\beta}}.$$

To obtain the profile restricted maximum log likelihood criterion σ^2 needs to be profiled out by its estimator

$$\hat{\sigma}^2 = \frac{D_p(\hat{\boldsymbol{\beta}})}{n-p}.$$

Hence, the negative profile restricted maximum log likelihood criterion up to a constant becomes

$$-REML(\boldsymbol{\rho}) = \frac{(n-p)}{2}\log\left(D_p(\hat{\boldsymbol{\beta}})\right) + K.$$

We first need to derive

$$\boldsymbol{G} = -\frac{\partial^2 REML(\boldsymbol{\rho})}{\partial\boldsymbol{\rho}\partial\boldsymbol{y}^t} \quad \text{and} \quad \boldsymbol{B} = -\frac{\partial^2 REML(\boldsymbol{\rho})}{\partial\boldsymbol{\rho}\partial\boldsymbol{\rho}^t}.$$

Therefore we obtain

$$-\frac{\partial REML(\boldsymbol{\rho})}{\partial\rho_j} = \frac{n-p}{2}\frac{1}{D_p(\hat{\boldsymbol{\beta}})}\frac{\partial D_p(\hat{\boldsymbol{\beta}})}{\partial\rho_j} + \frac{\partial K}{\partial\rho_j}.$$

Thus, for the j-th row of $\boldsymbol{G}$ we obtain

$$-\frac{\partial^2 REML(\boldsymbol{\rho})}{\partial\rho_j\partial\boldsymbol{y}^t} = \frac{n-p}{2D_p(\hat{\boldsymbol{\beta}})}\left(\frac{\partial^2 D_p(\hat{\boldsymbol{\beta}})}{\partial\rho_j\partial\boldsymbol{y}^t} - \frac{1}{D_p(\hat{\boldsymbol{\beta}})}\frac{\partial D_p(\hat{\boldsymbol{\beta}})}{\partial\rho_j}\frac{\partial D_p(\hat{\boldsymbol{\beta}})}{\partial\boldsymbol{y}^t}\right)$$

and for the i-th and j-th element of $\boldsymbol{B}$ we get

$$-\frac{\partial^2 REML(\boldsymbol{\rho})}{\partial\rho_j\partial\rho_i} = \frac{n-p}{2D_p(\hat{\boldsymbol{\beta}})}\left(\frac{\partial^2 D_p(\hat{\boldsymbol{\beta}})}{\partial\rho_j\partial\rho_i} - \frac{1}{D_p(\hat{\boldsymbol{\beta}})}\frac{\partial D_p(\hat{\boldsymbol{\beta}})}{\partial\rho_j}\frac{\partial D_p(\hat{\boldsymbol{\beta}})}{\partial\rho_i}\right) + \frac{\partial^2 K}{\partial\rho_j\partial\rho_i}.$$

The derivatives of the penalized deviance are

$$\frac{\partial D_p(\hat{\boldsymbol{\beta}})}{\partial\rho_j} = 2\left(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}\right)^t \boldsymbol{X}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_j} + 2\frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\rho_j}\boldsymbol{S}\hat{\boldsymbol{\beta}} + \exp(\rho_j)\hat{\boldsymbol{\beta}}^t\boldsymbol{S}_j\hat{\boldsymbol{\beta}},$$

$$\frac{\partial D_p(\hat{\boldsymbol{\beta}})}{\partial\boldsymbol{y}^t} = 2\left(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}\right)^t\left(\boldsymbol{X}\frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\boldsymbol{y}^t} - \boldsymbol{I}\right) + 2\frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\boldsymbol{y}^t}\boldsymbol{S}\hat{\boldsymbol{\beta}},$$

$$
\begin{aligned}
\frac{\partial^2 D_p(\hat{\boldsymbol{\beta}})}{\partial\rho_j\partial\boldsymbol{y}^t} = 2\Bigg(&\frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\rho_j}\boldsymbol{X}^t(\boldsymbol{X}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}^t}-\boldsymbol{I})+\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{y}\right)^t\boldsymbol{X}\frac{\partial^2\hat{\boldsymbol{\beta}}}{\partial\rho_j\partial\boldsymbol{y}^t}\\
&+\hat{\boldsymbol{\beta}}^t\boldsymbol{S}\frac{\partial^2\hat{\boldsymbol{\beta}}}{\partial\rho_j\partial\boldsymbol{y}^t}+\frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\rho_j}\boldsymbol{S}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}^t}+\exp(\rho_j)\hat{\boldsymbol{\beta}}^t\boldsymbol{S}_j\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}^t}\Bigg)
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial^2 D_p(\hat{\boldsymbol{\beta}})}{\partial\rho_j\partial\rho_i} =& 2\left(\frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\rho_j}\boldsymbol{X}^t\boldsymbol{X}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_j}+\hat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{X}\frac{\partial^2\hat{\boldsymbol{\beta}}}{\partial\rho_j\partial\rho_i}-\boldsymbol{y}^t\boldsymbol{X}\frac{\partial^2\hat{\boldsymbol{\beta}}}{\partial\rho_j\partial\rho_i}\right)\\
&+2\left(\frac{\partial^2\hat{\boldsymbol{\beta}}^t}{\partial\rho_j\partial\rho_i}\boldsymbol{S}\hat{\boldsymbol{\beta}}+\frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\rho_j}\boldsymbol{S}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_i}+\exp(\rho_j)\hat{\boldsymbol{\beta}}^t\boldsymbol{S}_j\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_i}+\exp(\rho_i)\hat{\boldsymbol{\beta}}^t\boldsymbol{S}_i\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_j}\right)
\end{aligned}
$$

for $j \neq i$ and otherwise

$$
\begin{aligned}
\frac{\partial^2 D_p(\hat{\boldsymbol{\beta}})}{\partial\rho_i^2} =& 2\left(\frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\rho_i}\boldsymbol{X}^t\boldsymbol{X}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_i}+\hat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{X}\frac{\partial^2\hat{\boldsymbol{\beta}}}{\partial\rho_i^2}-\boldsymbol{y}^t\boldsymbol{X}\frac{\partial^2\hat{\boldsymbol{\beta}}}{\partial\rho_i^2}\right)\\
&+2\left(\frac{\partial^2\hat{\boldsymbol{\beta}}^t}{\partial\rho_i^2}\boldsymbol{S}\hat{\boldsymbol{\beta}}+\frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\rho_i}\boldsymbol{S}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_i}+2\exp(\rho_i)\hat{\boldsymbol{\beta}}^t\boldsymbol{S}_i\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_i}\right)+\exp(\rho_i)\hat{\boldsymbol{\beta}}^t\boldsymbol{S}_i\hat{\boldsymbol{\beta}}.
\end{aligned}
$$

The derivatives of the regression parameters with respect to the smoothing parameters and the data are

$$
\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}^t} = \left(\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{S}\right)^{-1}\boldsymbol{X}^t,
$$

and

$$
\begin{aligned}
\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_j} &= -\exp(\rho_j)\left(\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{S}\right)^{-1}\boldsymbol{S}_j\left(\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{S}\right)^{-1}\boldsymbol{X}^t\boldsymbol{y}\\
&= -\exp(\rho_j)\left(\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{S}\right)^{-1}\boldsymbol{S}_j\hat{\boldsymbol{\beta}}.
\end{aligned}
$$

The second derivatives are

$$
\frac{\partial^2\hat{\boldsymbol{\beta}}}{\partial\rho_j\partial\boldsymbol{y}^t} = -\exp(\rho_j)\left(\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{S}\right)^{-1}\boldsymbol{S}_j\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}^t}
$$

and

$$\frac{\partial^2 \hat{\boldsymbol{\beta}}}{\partial \rho_j \partial \rho_i} = -\left(\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{S}\right)^{-1}\left(\exp(\rho_j)\boldsymbol{S}_j\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_i}+\exp(\rho_i)\boldsymbol{S}_i\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_j}\right)$$

if $j \neq i$ and otherwise, i.e. $j = i$

$$\frac{\partial^2 \hat{\boldsymbol{\beta}}}{\partial \rho_i^2} = \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_i} - 2\left(\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{S}\right)^{-1}\left(\exp(\rho_i)\boldsymbol{S}_i\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_i}\right).$$

The derivatives of K with respect to the log smoothing parameters are still left:

$$\begin{aligned}\frac{\partial K}{\partial \rho_j} &= \frac{1}{2}\frac{\partial}{\partial \rho_j}\left(\log\left|\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{S}\right| - \log|\boldsymbol{S}|_+\right)\\ &= \frac{1}{2}\exp(\rho_j)\left(\operatorname{tr}\left(\left(\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{S}\right)^{-1}\boldsymbol{S}_j\right)+\operatorname{tr}\left(\boldsymbol{S}^{-1}\boldsymbol{S}_j\right)\right),\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 K}{\partial \rho_j \partial \rho_i} &= \frac{1}{2}\frac{\partial^2}{\partial \rho_j \partial \rho_i}\left(\log\left|\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{S}\right| - \log|\boldsymbol{S}|_+\right)\\ &= \frac{1}{2}\exp(\rho_j+\rho_i)\left(\operatorname{tr}\left(\boldsymbol{S}^{-1}\boldsymbol{S}_j\boldsymbol{S}^{-1}\boldsymbol{S}_i\right) - \left(\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{S}\right)^{-1}\boldsymbol{S}_j\left(\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{S}\right)^{-1}\boldsymbol{S}_i\right)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2 K}{\partial \rho_j^2} &= \frac{1}{2}\frac{\partial^2}{\partial \rho_j^2}\left(\log\left|\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{S}\right| - \log|\boldsymbol{S}|_+\right)\\ &= \frac{\partial K}{\partial \rho_j} + \frac{1}{2}\exp(2\rho_j)\left(\operatorname{tr}\left(\boldsymbol{S}^{-1}\boldsymbol{S}_j\boldsymbol{S}^{-1}\boldsymbol{S}_j\right) - 2\left(\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{S}\right)^{-1}\boldsymbol{S}_j\right).\end{aligned}$$

Stable calculation of these derivatives and especially for the $\log|\boldsymbol{S}|_+$ part needs some transformations, see Wood (2011) Appendix B for details.

3.4.2 Marginal Likelihood

If smoothing parameter estimation is done via maximum marginal likelihood rather than with restricted maximum likelihood, the estimator of the variance changes to

$$\hat{\sigma}^2 = \frac{D_p(\hat{\boldsymbol{\beta}})}{n}.$$

Furthermore, the regression parameters need to be separated into penalized and unpenalized components for the part $\log\left|\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{S}\right|$ in K, see Wood (2011) section 2.1 for details. Let $\boldsymbol{U}$ be the orthogonal matrix corresponding to the reparametrization, then the profile marginal likelihood criterion is

$$-ML(\boldsymbol{\rho}) = \frac{n}{2}\log\left(D_p(\hat{\boldsymbol{\beta}})\right) + \frac{1}{2}\log\left|\boldsymbol{U}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{U}+\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}\right| - \frac{1}{2}\log|\boldsymbol{S}|_+ .$$

Thus, all derivatives from the restricted maximum likelihood stay the same up to the factor $\frac{n}{n-p}$, except that in the derivatives of $\log\left|\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{S}\right|$ the matrices $\boldsymbol{X}$, $\boldsymbol{S}$ and $\boldsymbol{S}_j$ are replaced by $\boldsymbol{U}\boldsymbol{X}$, $\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}$ and $\boldsymbol{U}\boldsymbol{S}_j\boldsymbol{U}$ respectively.

3.4.3 Akaike Information Criterion, unbiased risk estimator & Mallows Cp

Following Wood (2008) the smoothing parameter for known scale parameter σ^2 can be estimated by minimizing the generalized AIC or the unbiased risk estimator ($UBRE$) that is given in Wood (2008) section 2.1 by

$$UBRE(\boldsymbol{\rho}) = D(\hat{\boldsymbol{\beta}}) + 2\gamma\sigma^2\tau,$$

whereby $D(\hat{\boldsymbol{\beta}}) = 2\sigma^2\left(l_s(\sigma^2) - l(\hat{\boldsymbol{\beta}})\right)$ is the deviance of the model with the saturated likelihood $l_s(\sigma^2)$ and γ is an ad-hoc tuning parameter (Wood, 2008) and

$$\tau = \operatorname{tr}(\boldsymbol{H}) = \operatorname{tr}\left(\boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{S}\right)^{-1}\boldsymbol{X}^t\right)$$

is the trace of the hat matrix. Thus, for the first derivative with respect to the smoothing parameter we have

$$\frac{\partial UBRE(\boldsymbol{\rho})}{\partial\rho_j} = \frac{\partial D(\hat{\boldsymbol{\beta}})}{\partial\rho_j} + 2\gamma\sigma^2\frac{\partial\tau}{\partial\rho_j}$$

with

$$\frac{\partial\tau}{\partial\rho_j} = -\exp(\rho_j)\operatorname{tr}\left(\frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\boldsymbol{y}}\boldsymbol{S}_j\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}^t}\right)$$

and

$$\frac{\partial D(\hat{\boldsymbol{\beta}})}{\partial\rho_j} = 2\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{y}\right)^t\boldsymbol{X}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_j}. \tag{3.4.1}$$

The j-th row of $\boldsymbol{G}$ is therefore

$$\frac{\partial^2 UBRE(\boldsymbol{\rho})}{\partial\rho_j\partial\boldsymbol{y}^t} = \frac{\partial^2 D(\hat{\boldsymbol{\beta}})}{\partial\rho_j\partial\boldsymbol{y}^t} = 2\frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\rho_j}\boldsymbol{X}^t\left(\boldsymbol{X}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}^t} - \boldsymbol{I}\right) + 2\left(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}\right)^t\boldsymbol{X}\frac{\partial^2\hat{\boldsymbol{\beta}}}{\partial\rho_j\partial\boldsymbol{y}^t}.$$

The second derivatives w.r.t. the log smoothing parameters can hence be obtained by

$$\frac{\partial^2 UBRE(\boldsymbol{\rho})}{\partial\rho_j\partial\rho_i} = \frac{\partial D(\hat{\boldsymbol{\beta}})}{\partial\rho_j\partial\rho_i} + 2\gamma\sigma^2\frac{\partial^2\tau}{\partial\rho_j\partial\rho_i}$$

with

$$\begin{aligned}\frac{\partial D(\hat{\boldsymbol{\beta}})}{\partial\rho_j\partial\rho_i} &= 2\frac{\partial}{\partial\rho_i}\left[\left(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}\right)^t\boldsymbol{X}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_j}\right]\\ &= 2\frac{\partial}{\partial\rho_i}\left[\hat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{X}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_j} - \boldsymbol{y}^t\boldsymbol{X}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_j}\right]\\ &= 2\left(\frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\rho_i}\boldsymbol{X}^t\boldsymbol{X}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_j} + \hat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{X}\frac{\partial^2\hat{\boldsymbol{\beta}}}{\partial\rho_j\partial\rho_i} - \boldsymbol{y}^t\boldsymbol{X}\frac{\partial^2\hat{\boldsymbol{\beta}}}{\partial\rho_j\partial\rho_i}\right)\\ &= 2\left(\frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\rho_i}\boldsymbol{X}^t\boldsymbol{X}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_j} + \left(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}\right)^t\boldsymbol{X}\frac{\partial^2\hat{\boldsymbol{\beta}}}{\partial\rho_j\partial\rho_i}\right). \qquad (3.4.2)\end{aligned}$$

The second derivatives of the trace of the hat matrix with respect to the log smoothing parameters are

$$\begin{aligned}\frac{\partial^2\tau}{\partial\rho_j\partial\rho_i} &= -\exp(\rho_j)\mathrm{tr}\left(\frac{\partial^2\hat{\boldsymbol{\beta}}^t}{\partial\boldsymbol{y}\partial\rho_i}\boldsymbol{S}_j\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}^t} + \frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\boldsymbol{y}}\boldsymbol{S}_j\frac{\partial^2\hat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}^t\partial\rho_i}\right)\\ &= -\exp(\rho_j)\mathrm{tr}\left(-\exp(\rho_i)\frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\boldsymbol{y}}\boldsymbol{S}_i\left(\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{S}\right)^{-1}\boldsymbol{S}_j\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}^t}\right.\\ &\quad\left. - \frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\boldsymbol{y}}\boldsymbol{S}_j\exp(\rho_i)\left(\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{S}\right)^{-1}\boldsymbol{S}_i\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}^t}\right)\\ &= 2\ \exp(\rho_j + \rho_i)\mathrm{tr}\left(\frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\boldsymbol{y}}\boldsymbol{S}_i\left(\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{S}\right)^{-1}\boldsymbol{S}_j\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}^t}\right) \qquad (3.4.3)\end{aligned}$$

and

$$\frac{\partial^2\tau}{\partial\rho_j^2} = \frac{\partial\tau}{\partial\rho_j} + 2\ \exp(2\rho_j)\mathrm{tr}\left(\frac{\partial\hat{\boldsymbol{\beta}}^t}{\partial\boldsymbol{y}}\boldsymbol{S}_j\left(\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{S}\right)^{-1}\boldsymbol{S}_j\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}^t}\right). \qquad (3.4.4)$$

3.4.4 Generalized cross-validation

The generalized cross-validation criterion (GCV) as in Wood (2008) is

$$GCV(\boldsymbol{\rho}) = \frac{nD(\hat{\boldsymbol{\beta}})}{(n-\gamma\tau)^2},$$

whereby $D(\hat{\boldsymbol{\beta}}) = 2\sigma^2\left(l_s - l(\hat{\boldsymbol{\beta}})\right)$ is the deviance of the model and γ again is an ad-hoc tuning parameter (Wood, 2008).

The derivative of the GCV w.r.t. the j-th log smoothing parameter is

$$\frac{\partial GCV(\boldsymbol{\rho})}{\partial \rho_j} = \frac{n\frac{\partial}{\partial \rho_j}D(\hat{\boldsymbol{\beta}})}{(n-\gamma\tau)^2} + \frac{2nD(\hat{\boldsymbol{\beta}})}{(n-\gamma\tau)^3}\gamma\frac{\partial \tau}{\partial \rho_j},$$

where $\frac{\partial}{\partial \rho_j}D(\hat{\boldsymbol{\beta}})$ is defined in 3.4.1.

Accordingly, the second derivatives w.r.t. the log smoothing parameters and the data $\boldsymbol{y}^t$ and as such also the j-th row of the matrix $\boldsymbol{G}$ is

$$\begin{aligned}\frac{\partial^2 GCV(\boldsymbol{\rho})}{\partial \rho_j \partial \boldsymbol{y}^t} &= \frac{n}{(n-\gamma\tau)^2}\left[\frac{\partial^2}{\partial \rho_j \partial \boldsymbol{y}^t}D(\hat{\boldsymbol{\beta}})\right] + \frac{2n\gamma}{(n-\gamma\tau)^3}\frac{\partial \tau}{\partial \rho_j}\frac{\partial}{\partial \boldsymbol{y}^t}\left[D(\hat{\boldsymbol{\beta}})\right] \\ &= \frac{2n}{(n-\gamma\tau)^2}\left[\frac{\partial \hat{\boldsymbol{\beta}}^t}{\partial \rho_j}\boldsymbol{X}^t\left(\boldsymbol{X}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{y}^t} - \boldsymbol{I}\right) + \left(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}\right)^t \boldsymbol{X}\frac{\partial^2 \hat{\boldsymbol{\beta}}^t}{\partial \rho_j \partial \boldsymbol{y}^t}\right] \\ &+ \frac{4n\gamma}{(n-\gamma\tau)^3}\frac{\partial \tau}{\partial \rho_j}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^t\left(\boldsymbol{I} - \boldsymbol{X}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{y}^t}\right).\end{aligned}$$

The second derivatives w.r.t. the i-th and j-th log smoothing parameters and therefore the i-th and j-th entry of the matrix $\boldsymbol{B}$ is

$$\begin{aligned}\frac{\partial^2 GCV(\boldsymbol{\rho})}{\partial \rho_j \partial \rho_i} &= \frac{\partial}{\partial \rho_i}\left[\frac{n\frac{\partial}{\partial \rho_j}D(\hat{\boldsymbol{\beta}})}{(n-\gamma\tau)^2} + 2\gamma\frac{nD(\hat{\boldsymbol{\beta}})}{(n-\gamma\tau)^3}\frac{\partial \tau}{\partial \rho_j}\right] \\ &= \frac{2\gamma n}{(n-\gamma\tau)^3}\left(\frac{(n-\gamma\tau)}{2\gamma}\cdot\frac{\partial^2 D(\hat{\boldsymbol{\beta}})}{\partial \rho_j \partial \rho_i} + \frac{\partial D(\hat{\boldsymbol{\beta}})}{\partial \rho_j}\frac{\partial \tau}{\partial \rho_i}\right. \\ &\left. + \frac{\partial D(\hat{\boldsymbol{\beta}})}{\partial \rho_i}\frac{\partial \tau}{\partial \rho_j} + \frac{3\gamma}{(n-\gamma\tau)}D(\hat{\boldsymbol{\beta}})\frac{\partial \tau}{\partial \rho_i}\frac{\partial \tau}{\partial \rho_j} + D(\hat{\boldsymbol{\beta}})\frac{\partial^2 \tau}{\partial \rho_j \partial \rho_i}\right),\end{aligned}$$

with $\frac{\partial^2 \tau}{\partial \rho_j \partial \rho_i}$ the same as in 3.4.3 and 3.4.4 and $\frac{\partial^2}{\partial \rho_j \partial \rho_i}D(\hat{\boldsymbol{\beta}})$ is the same as in 3.4.2.

3.5 Simulations

In this simulation study we want to investigate different aspects of the behaviour of the corrected degrees of freedom in several settings. Correcting the degrees of freedom for the uncertainty that is induced by the estimation of the smoothing parameter has been examined for REML and ML in the context of the conditional AIC in Greven and Kneib (2010). The corrected conditional AIC in this setting is

$$\text{cAIC} = -\frac{\text{cRSS}}{\sigma^2} + 2df\,(\boldsymbol{\beta}, \boldsymbol{u}) + 2df\,(\boldsymbol{\lambda}), \tag{3.5.1}$$

with $df\,(\boldsymbol{\beta}, \boldsymbol{u})$ and especially $df(\boldsymbol{\lambda})$ defined in (3.3.5) and conditional residual sum of squares cRSS. Conditionality here refers to the penalized parameters that are conditioned when treated as random in an empirical Bayes framework (see Section 3.1.2). The conventional AIC ignores the degrees of freedom of the smoothing parameter $df(\boldsymbol{\lambda})$. Notice that the AIC here is used for smoothing parameter selection, but is used for model selection after the smoothing parameter selection process.

Greven and Kneib (2010) show that neglecting this dependence leads to biased model selection. Similar to the simulation study in Section 1.4 and 2.3, this bias becomes apparent for the choice between $\lambda = \infty$ vs $\lambda \in (0, \infty)$. While this behaviour is known for REML and ML, it is unknown for GCV and UBRE. This simulation study shows that this effect is particularly conspicuous for GCV and UBRE.
Another question is how the degrees of freedom of the smoothing parameter $df(\lambda)$ interact with the degrees of freedom of the penalized parameters $df(\beta)$ or the basis dimension of the smooth term associated with the smoothing parameter.

3.5.1 Random intercept model

A simulation setting that is not comprised by model (3.3.1), but nevertheless can be estimated by minimizing (3.3.3) is a mixed model, and here we focus on the special case of a random intercept model. It is not common to estimate mixed models with the help of criteria like GCV or UBRE, but it gives an easy to interpret simulation model, in which the signal-to-noise ratio is given by the variance of the random intercepts denoted by $\tau^2 = \frac{\sigma^2}{\lambda}$, with smoothing parameter $\lambda = \exp(\rho)$ and noise component σ^2. Thus 1,000 data sets are generated from the model

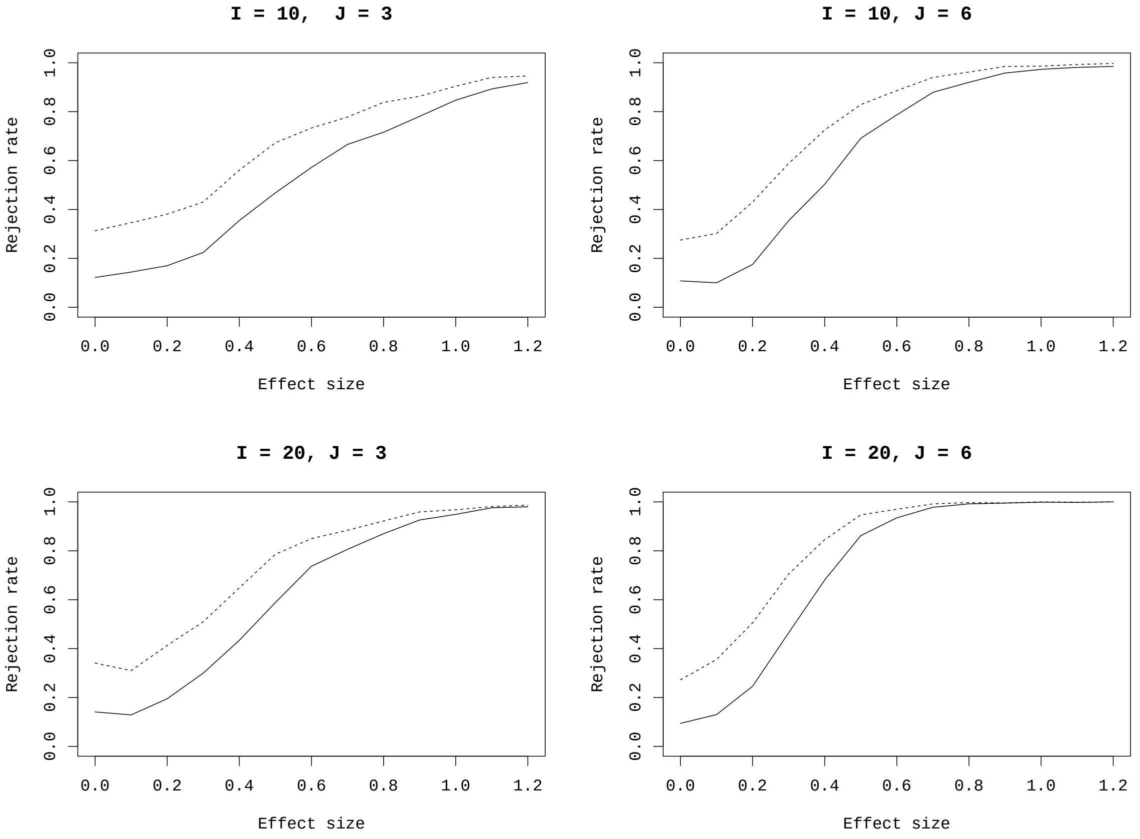

Figure 3.4: Results for the random intercept model with the smoothing parameter chosen by GCV. The y-axis shows the number of simulation replications out of 1,000 where the more complex model was favoured by the different AICs. The full line (——) corresponds to the selection frequency of the corrected AIC (3.6.2) with the degrees of freedom of the smoothing parameter. The dashed line (- - -) corresponds to the conventional AIC ignoring the uncertainty induced by the smoothing parameter.

$$y_{ij} = \beta_0 + u_j + \epsilon_{ij} \quad \text{for} \quad i = 1, \dots, n \quad \text{and} \quad j = 1, \dots, m,$$

with

$$\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \tau^2 \boldsymbol{I}_m & \mathbf{0} \\ \mathbf{0} & \sigma^2 \boldsymbol{I}_n \end{pmatrix} \right).$$

This is a balanced random intercept model since the cluster sizes m do not depend on the clusters. In this simulation we set $\sigma^2 = 1$, $\beta_0 = 0$, $\tau^2 = 0, 0.1, \dots, 1.2$ and the number of clusters and cluster sizes to $10, 20$ and $3, 6$ respectively.

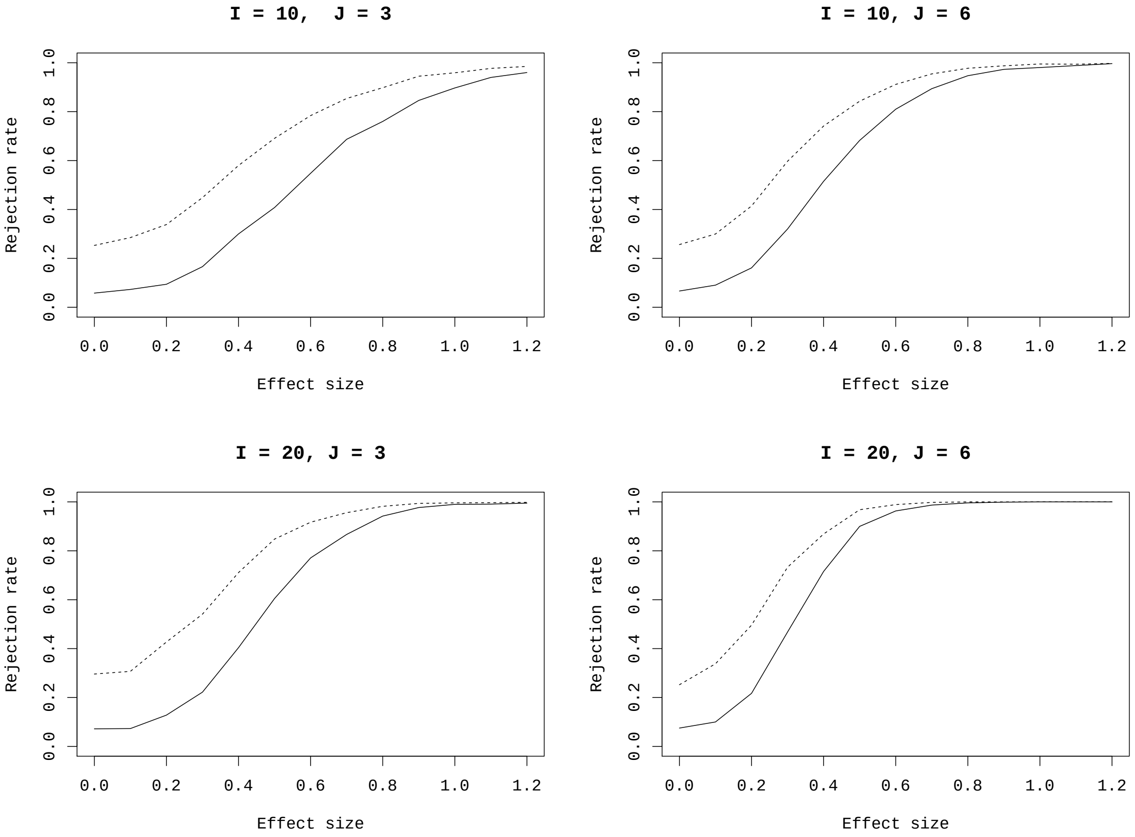

Figure 3.5: Results for the random intercept model with the smoothing parameter chosen by UBRE. The y-axis shows the number of simulation replications out of 1,000 where the more complex model was favoured by the different AICs. The full line (——) corresponds to the selection frequency of the corrected AIC (3.6.2) with the degrees of freedom of the smoothing parameter. The dashed line (- - -) corresponds to the conventional AIC ignoring the uncertainty induced by the smoothing parameter.

For each setting and data set, a random intercept model and a linear model with just one intercept are fitted. This corresponds to a complex model with $\lambda \in (0, \infty)$ and a model with $\lambda = \infty$. For each of the smoothing parameter selection criteria the frequency of choosing the random intercept model is computed. The frequencies are plotted for each criterion and each number of clusters and cluster sizes against the increasing signal-to-noise ratio, i.e. the inverse smoothing parameter.
The models are fitted with the R-package mgcv version 1.8-2 (Wood, 2011). The degrees of freedom of the smoothing parameter (or log-smoothing parameters) are implemented following the results in Section 3.4. The implementation is fast and computes the degrees of freedom instantaneously, with many parts of the calculation being already available from the model fit. No numerical problems occurred. The Cholesky-

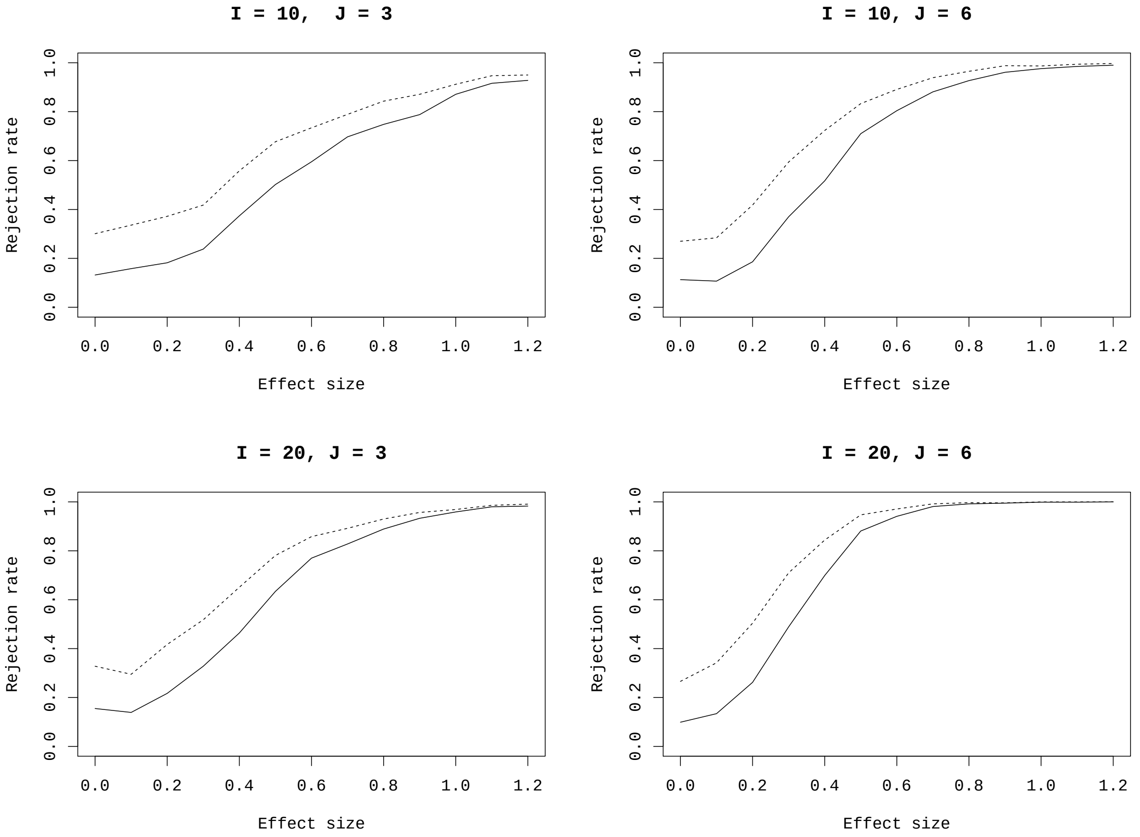

Figure 3.6: Results for the random intercept model with the smoothing parameter chosen by REML. The y-axis shows the number of simulation replications out of 1,000 where the more complex model was favoured by the different AICs. The full line (——) corresponds to the selection frequency of the corrected AIC (3.6.2) with the degrees of freedom of the smoothing parameter. The dashed line (- - -) corresponds to the conventional AIC ignoring the uncertainty induced by the smoothing parameter.

decomposition (4.3.1) was computable in all cases. Thus the Hessian $\boldsymbol{B}$ was positive definite in all models.

The results are plotted for the different number of clusters and cluster sizes in Figure 3.4 for GCV, Figure 3.5 for UBRE, Figure 3.6 for REML and Figure 3.7 for ML. The plots for REML and ML confirm the results in Greven and Kneib (2010). For smoothing parameter selection criteria GCV and UBRE the behaviour of biased model selection persists. Especially for the UBRE criterion, which assumes the model error σ^2 is known, the bias in the model selection is even stronger than the bias for REML or ML. Hence, the rejection rates of the simple linear model differ more between the AIC with and without the degrees of freedom of the smoothing parameter. This emphasizes

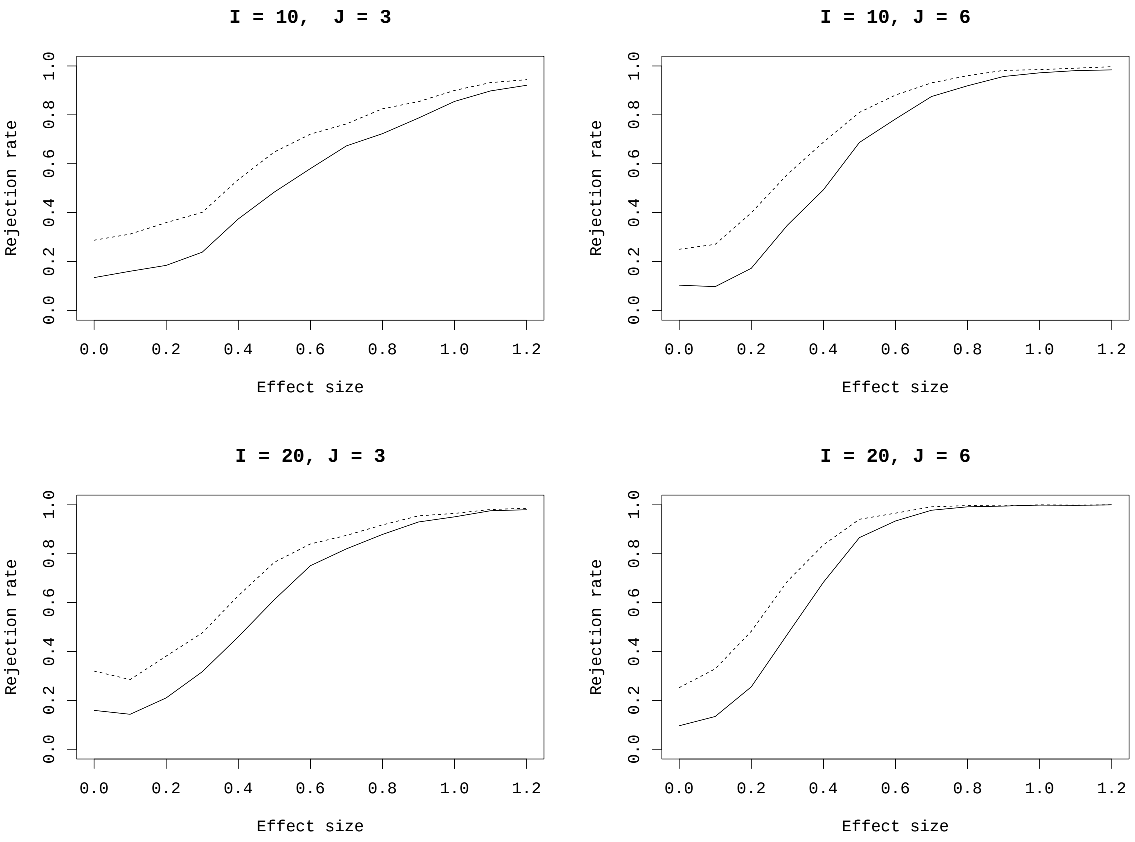

Figure 3.7: Results for the random intercept model with the smoothing parameter chosen by ML. The y-axis shows the number of simulation replications out of 1,000 where the more complex model was favoured by the different AICs. The full line (——) corresponds to the selection frequency of the corrected AIC (3.6.2) with the degrees of freedom of the smoothing parameter. The dashed line (- - -) corresponds to the conventional AIC ignoring the uncertainty induced by the smoothing parameter.

that the degrees of freedom of the smoothing parameter need to be considered for model selection no matter by which criterion the smoothing parameters are implicitly specified.

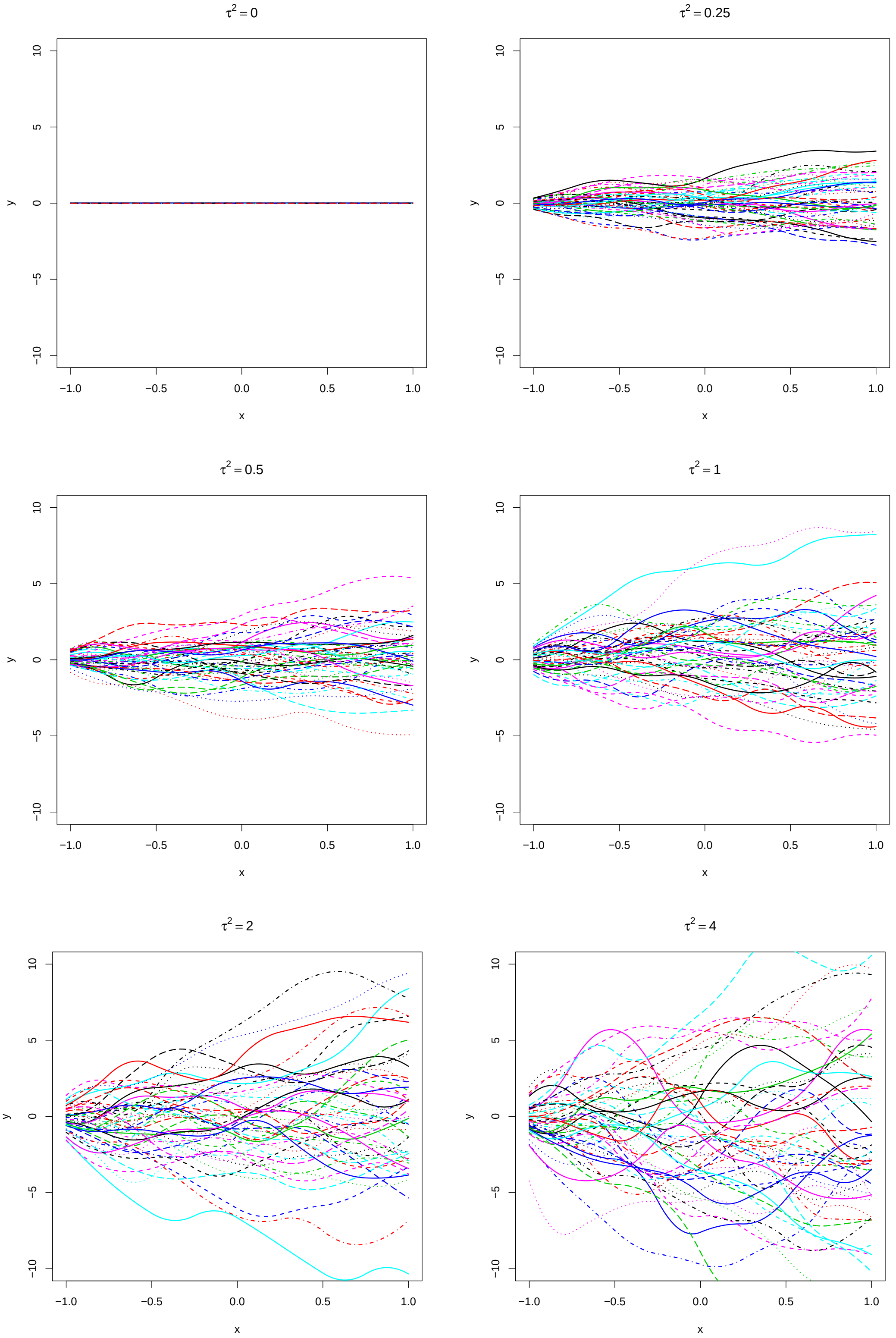

Figure 3.8: The random functions evolving for B-spline parameters β generated from a random walk with variance equal to 0, 0.25, 0.5, 1, 2 and 4.

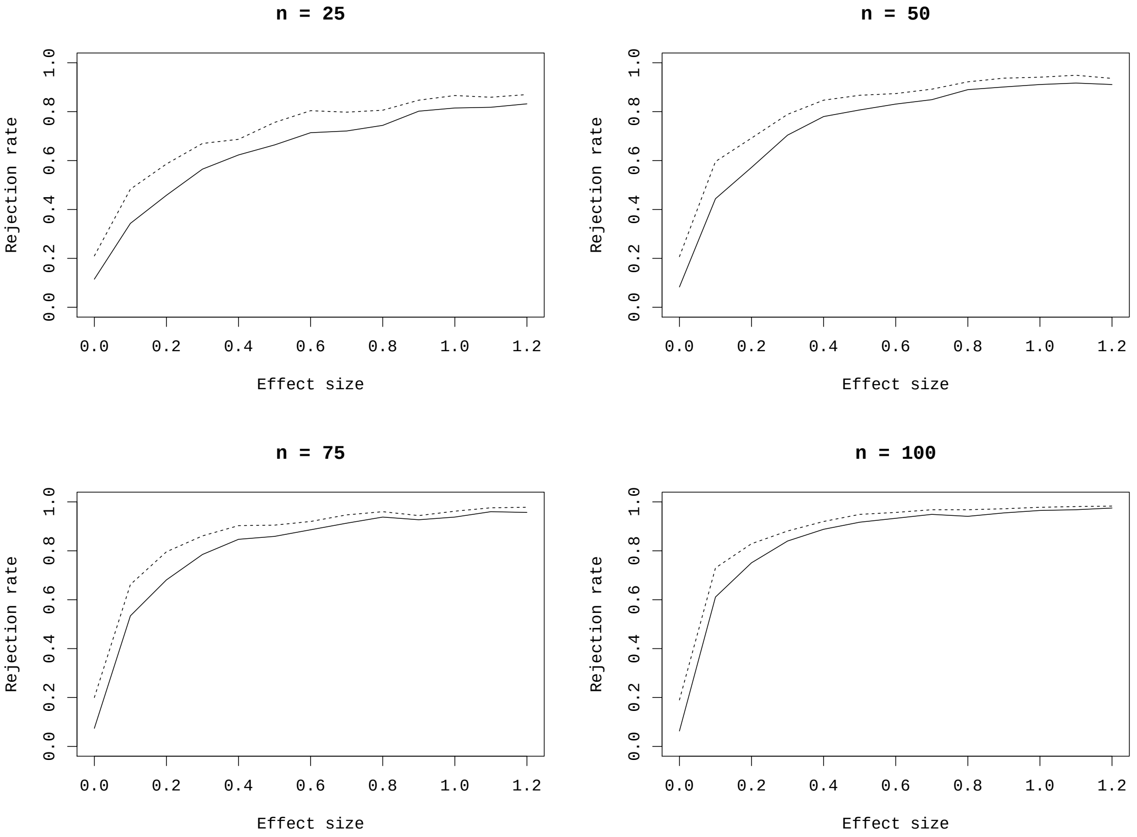

Figure 3.9: Results for the smoothing model with the smoothing parameter chosen by REML. The y-axis shows the number of simulation replications out of 1,000 where the more complex model was favoured by the different AICs. The full line (——) corresponds to the selection frequency of the corrected AIC (3.6.2) with the degrees of freedom of the smoothing parameter. The dashed line (- - -) corresponds to the conventional AIC ignoring the uncertainty induced by the smoothing parameter.

3.5.2 Penalized B-spline model

A common example of model (3.3.3) is a penalized spline model. This part of the simulation study adapts the findings from the random intercept model simulation to penalized spline smoothing. We therefore consider an empirical Bayes setting, meaning that we do not think of a true underlying function, but rather assume the coefficients that define a function to be random. This makes the results from the random intercept model directly comparable to the results in this section. Hence, we assume that the data is generated by

$$y_i = f(x_i) + \epsilon_i \quad \text{for} \quad i = 1, \dots, n, \tag{3.5.2}$$

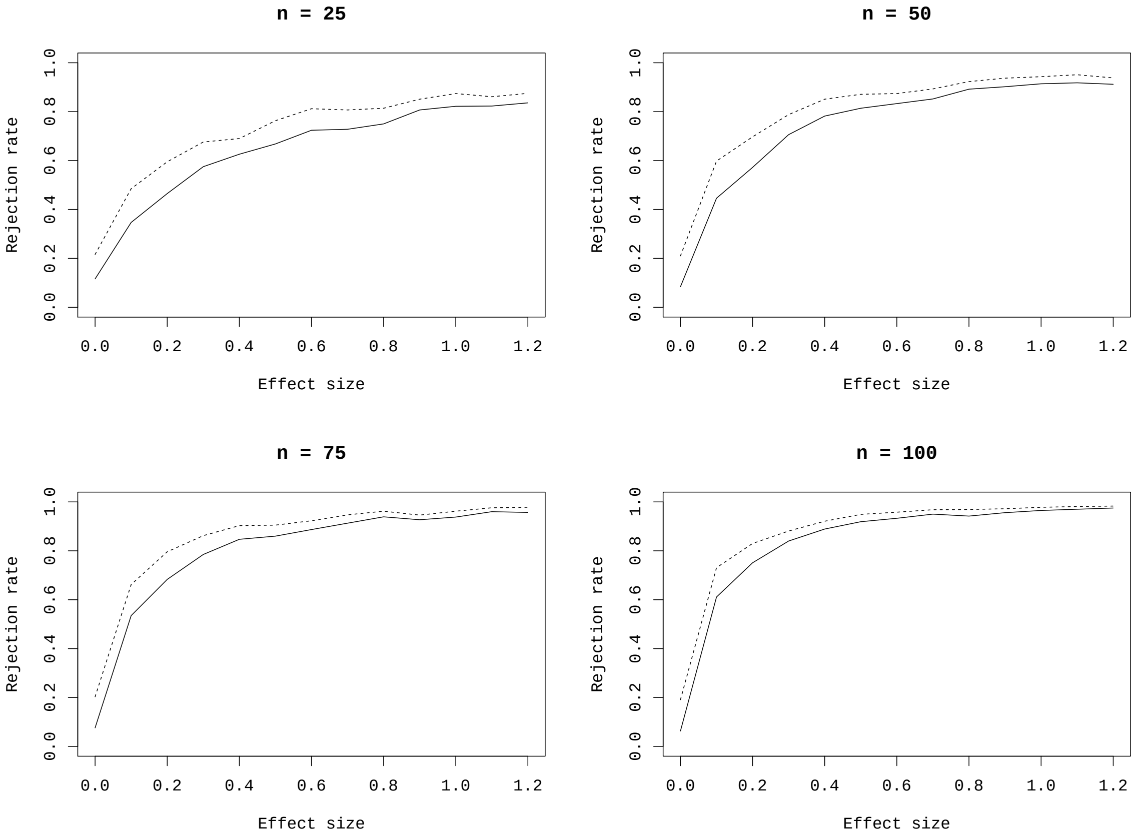

Figure 3.10: Results for the smoothing model with the smoothing parameter chosen by ML. The y-axis shows the number of simulation replications out of 1,000 where the more complex model was favoured by the different AICs. The full line (——) corresponds to the selection frequency of the corrected AIC (3.6.2) with the degrees of freedom of the smoothing parameter. The dashed line (- - -) corresponds to the conventional AIC ignoring the uncertainty induced by the smoothing parameter.

with the random function

$$f(x) = \sum_{i=1}^{d} \beta_j B_j(x) \quad \text{on} \quad x \in [-1, 1], \tag{3.5.3}$$

defined by the sum of scaled B-spline basis functions $B_i(\cdot)$, see Eilers and Marx (1996). The coefficients β_i are generated by a random walk of order one with variance parameter τ^2:

$$\beta_k | \beta_{k-1} \sim \mathcal{N}\left(0, \tau^2\right) \tag{3.5.4}$$

or

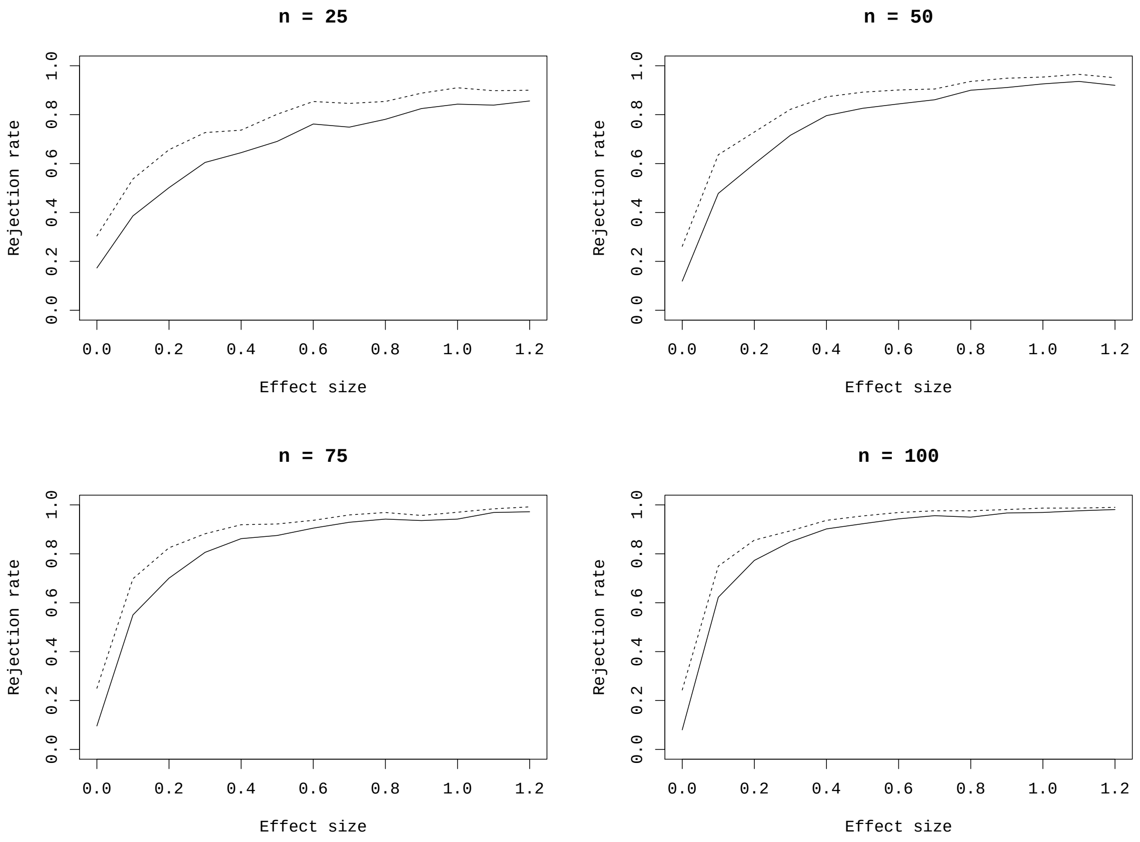

Figure 3.11: Results for the smoothing model with the smoothing parameter chosen by GCV. The y-axis shows the number of simulation replications out of 1,000 where the more complex model was favoured by the different AICs. The full line (——) corresponds to the selection frequency of the corrected AIC (3.6.2) with the degrees of freedom of the smoothing parameter. The dashed line (- - -) corresponds to the conventional AIC ignoring the uncertainty induced by the smoothing parameter.

$$\beta_k = \beta_{k-1} + u, \quad \text{with} \quad u \sim \mathcal{N}\left(0, \tau^2\right).$$

Some examples of random functions generated by this mechanism with 10 basis functions are plotted in Figure 3.8 for different random effects parameters. The examples indicate that the variance parameter τ^2 impacts the roughness of the function. The lower the variance parameter the closer the functions are to a horizontal line through zero, with equality in $\tau^2 = 0$ in the upper left panel.

For sample sizes $n = 25, 50, 100$ and 400, there are $1,000$ data sets generated of model (3.5.2), with varying variance parameter between $\tau^2 = 0$ and $\tau^2 = 4$ and fixed intercept

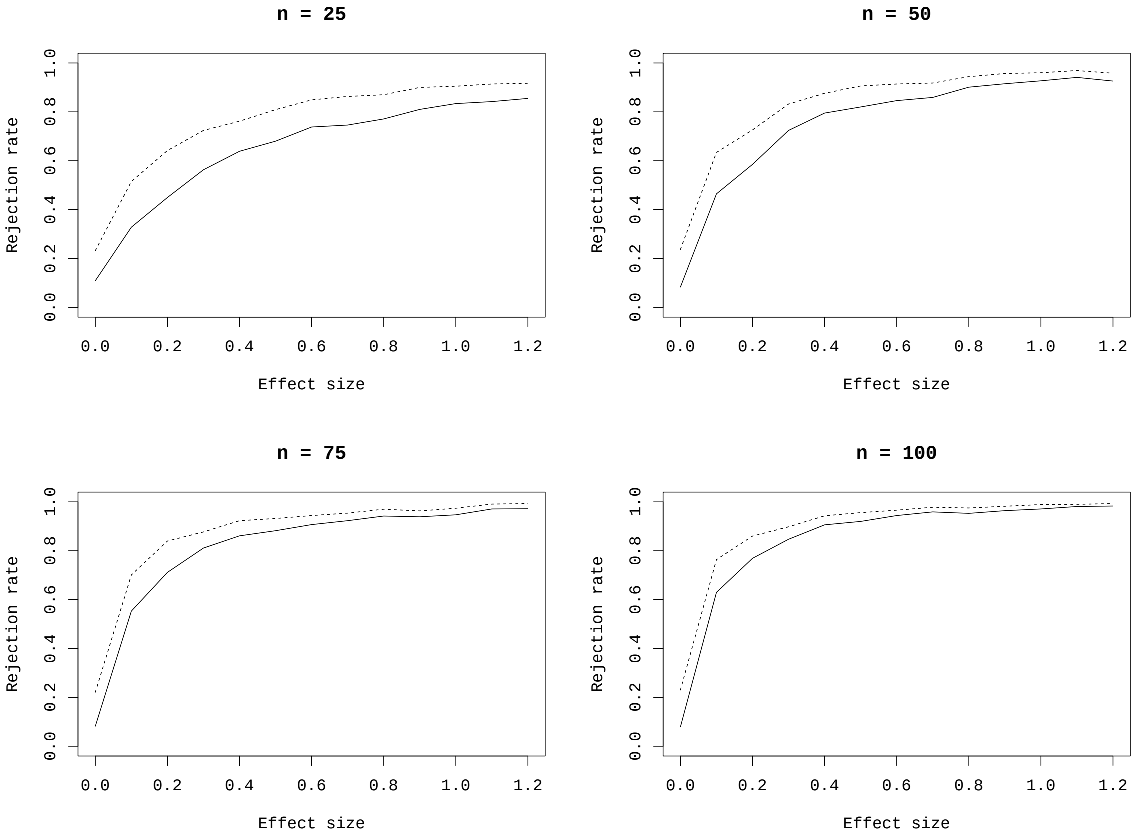

Figure 3.12: Results for the smoothing model with the smoothing parameter chosen by UBRE. The y-axis shows the number of simulation replications out of 1,000 where the more complex model was favoured by the different AICs. The full line (——) corresponds to the selection frequency of the corrected AIC (3.6.2) with the degrees of freedom of the smoothing parameter. The dashed line (- - -) corresponds to the conventional AIC ignoring the uncertainty induced by the smoothing parameter.

$\beta_0 = 0$. In each sample, the coefficients of the random functions $\beta_1, \ldots, \beta_{10}$ (β_0 is fixed) are resampled and the response data are generated from the random functions. This corresponds to the resampling of the random intercept in previous simulation settings. A B-spline model like the data-generating model and a linear model with only one intercept is fitted to the data. As in the random intercept simulation study, the frequency of choosing the complex model by AIC is computed. The AIC considers the degrees of freedom not only from the regression coefficients but also from the smoothing parameter (see equation 3.3.5).

The models are again fitted with the `R`-package `mgcv` version `1.8-2` (Wood, 2011) and the degrees of freedom of the smoothing parameter (or log-smoothing parameters) are calculated from the implementation following the results in Section 3.4. Numerical

problems arose from not positive definite matrix $\boldsymbol{B}$. Thus the Cholesky-decomposition (4.3.1) was not doable in all cases. This was not the case for REML, one case for ML, 10 cases for UBRE and 25 cases for GCV.
The results for REML and ML as smoothing parameter selection criterion are plotted in Figures 3.9 and 3.10. The rise in the selection frequency of the complex spline model is sharper than for the random intercept model in Figures 3.6 and 3.7. The results for GCV and UBRE as smoothing parameter selection criteria are plotted in Figures 3.11 and 3.12. They also rise sharper than the corresponding selection frequency plots for the random intercept. The difference between the frequency curves with and without the degrees of freedom of the smoothing parameter $df(\lambda)$ is larger for GCV and UBRE based smoothing parameter estimation than for REML and ML.

3.6 Example: Canadian weather data

The degrees of freedom of the smoothing parameter can be obtained in different semiparametric regression settings. As an application the degrees of freedom of the smoothing parameter are derived from a functional linear model applied to Canadian weather data.

3.6.1 Canadian weather data

The Canadian weather data are a popular example in functional data analysis (Ramsay and Silverman, 1997, 2002). The data set contains the average daily temperature and precipitation at 35 different locations in Canada over the years 1960 to 1994. The average temperature as functions of the days of the year are given in Figure 3.13.

The data set also contains the annual precipitation of the 35 locations. It makes sense to assume that the temperature has some kind of effect on the precipitation. The influence of average annual temperature on the log annual precipitation can be analysed with the help of a functional linear model.

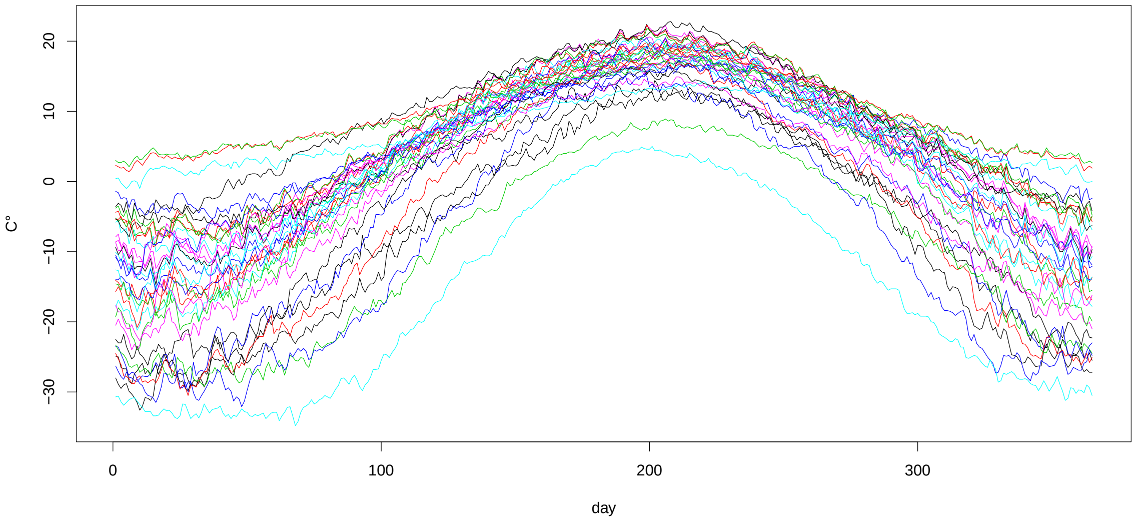

Figure 3.13: For 35 different locations in Canada the average temperatures of the years 1960 to 1994 as functions the day of the year.

3.6.2 Functional predictor regression

For the analysis of the influence of the average annual temperature on the log annual precipitation model (3.3.1) with the linear functional (3.3.2) is assumed as data-generating mechanism. Thus for the log annual precipitation of the i-th location the model is

$$y_i = \beta_0 + \int f(z) g_i(z) dz + \epsilon_i \ , i = 1, \dots, 35, \tag{3.6.1}$$

where $g_i(\cdot)$ is the average annual temperature, called signal or functional predictor, for the i-th location measured at a grid of points $1, \dots, 365$ and ϵ_i is a normal error with mean zero and variance σ^2. The function $f(\cdot)$ is represented by a penalized B-spline (3.5.3) with a random walk assumption on the parameters (3.5.4) and 10 basis functions. The data is analysed with a first and a second order random walk. Thus the penalty matrix is $\boldsymbol{S} = \boldsymbol{\Delta}^t \boldsymbol{\Delta}$, where Δ is a dth-order differencing matrix with $d = 1, 2$. Since the functions f and g_i are evaluated at discrete points the integral (3.6.1) is approximated by a sum:

$$\int f(z) g_i(z) dz = \frac{1}{h} \sum_{k=1}^{365} f(x_k) g_i(x_k)$$

with $f(x_k) = \sum_{j=1}^{10} \beta_j B_j(x_k)$. Hence, if a first order random walk is used and the smoothing parameter lies on the boundary of the parameter space, i.e. $\lambda \to \infty$, the function f is constant and the model (3.6.1) becomes a linear model

$$y_i = \beta_0 + \beta_1 \bar{z}_i + \epsilon_i, \tag{3.6.2}$$

with the mean of the average annual temperature over the year as covariate, i.e. $\bar{z}_i \propto \frac{1}{365}\sum_{k=1}^{365} g_i(x_k)$.

If the parameters follow a second order random walk and the smoothing parameter tends to infinity $f(x_k) = \beta_1 + \beta_2 x_k$ becomes linear and accordingly model (3.6.1) reduces to

$$y_i = \beta_0 + \beta_1 \bar{z}_i + \beta_2 \sum_{k=1}^{365} g_i(x_k) x_k + \epsilon_i. \tag{3.6.3}$$

The models are fitted with the R-package `mgcv` version `1.8-2` (Wood, 2011). The smoothing parameters are estimated with GCV, REML and ML and the degrees of freedom of the regression parameters and the smoothing parameters are calculated based on the derivations of Section 3.4.

	$df(\lambda)$	$df(\boldsymbol{\beta})$	cRSS
GCV	0.55	5.75	22.68
REML	0.43	4.31	19.84
ML	0.21	4.09	19.15
linear	0	2	6.56

Table 3.1: The estimated degrees of freedom of the regression parameters and the smoothing parameter from model (3.6.1), with $d = 1$, and the corresponding conditional residual sum of squares. In the last row the degrees of freedom, i.e. the number of parameters, and the residual sum of squares of model 3.6.2 are listed.

The results for first order random walk, $d = 1$, are listed in Table 3.1 and for the second order random walk, $d = 2$, the results are listed in Table 3.2. Based on definition the conditional AIC can be derived with these quantities. The AIC of the linear model (3.6.2) is -7.12. The model (3.6.1), with first order random walk assumption, gives lower conditional AIC for all three smoothing parameter selection criteria. Thus a functional modelling is preferable. With the second order random walk proposal, the AIC of the linear model (3.6.3) is -14.86. The conditional AICs of the semiparametric models again are smaller for all three smoothing parameter selection methods. Therefore a semiparametric model is also preferable in case of a second order random walk assumption. Hence, the effect of the temperature on the annual precipitation should not be modelled by a linear function of the mean of the annual temperature.

	$df(\lambda)$	$df(\boldsymbol{\beta})$	cRSS
GCV	0.58	5.47	22.85
REML	0.12	4.00	19.07
ML	0.26	3.89	18.81
linear	0	3	11.34

Table 3.2: The estimated degrees of freedom of the regression parameters and the smoothing parameter from model (3.6.1), with $d = 2$, and the corresponding conditional residual sum of squares. In the last row the degrees of freedom, i.e. the number of parameters, and the residual sum of squares of model 3.6.3 are listed.

The coefficient functions that were estimated for the first order random walk are given in Figure 3.14 for the three smoothing parameter selection criteria. The functions estimated on the bases of REML and ML are nearly indistinguishable in contrast to the function that was estimated based on GCV, that is less smooth.
The coefficient functions that were estimated for the second order random walk are given in Figure 3.15 for the three smoothing parameter selection criteria. The estimated functions are similar but with smaller confidence intervals.

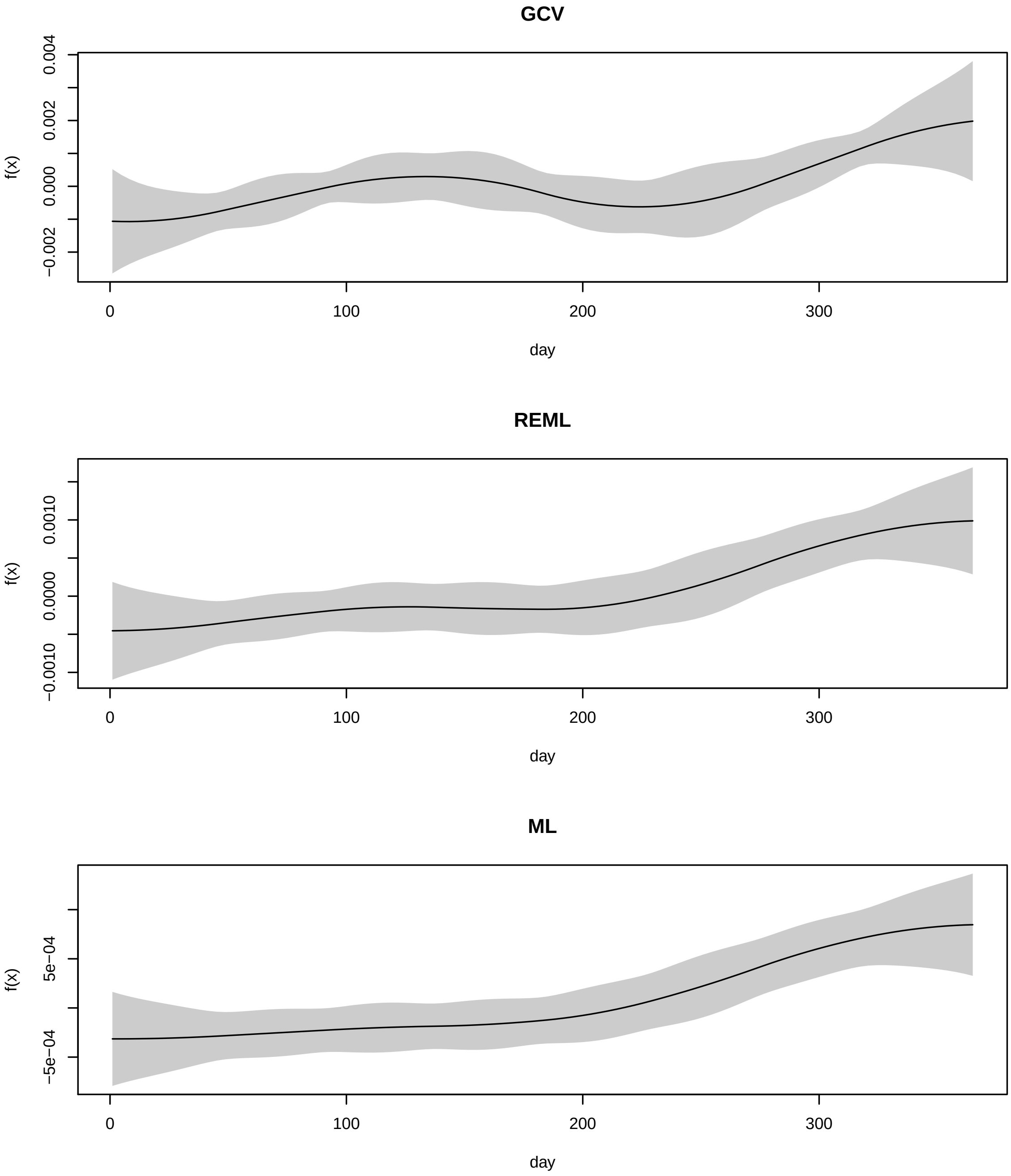

Figure 3.14: The coefficient function $f(\cdot)$ from model (3.6.1), with a first order random walk assumption for the regression parameters. The smoothing parameter was estimated with GCV, REML and ML.

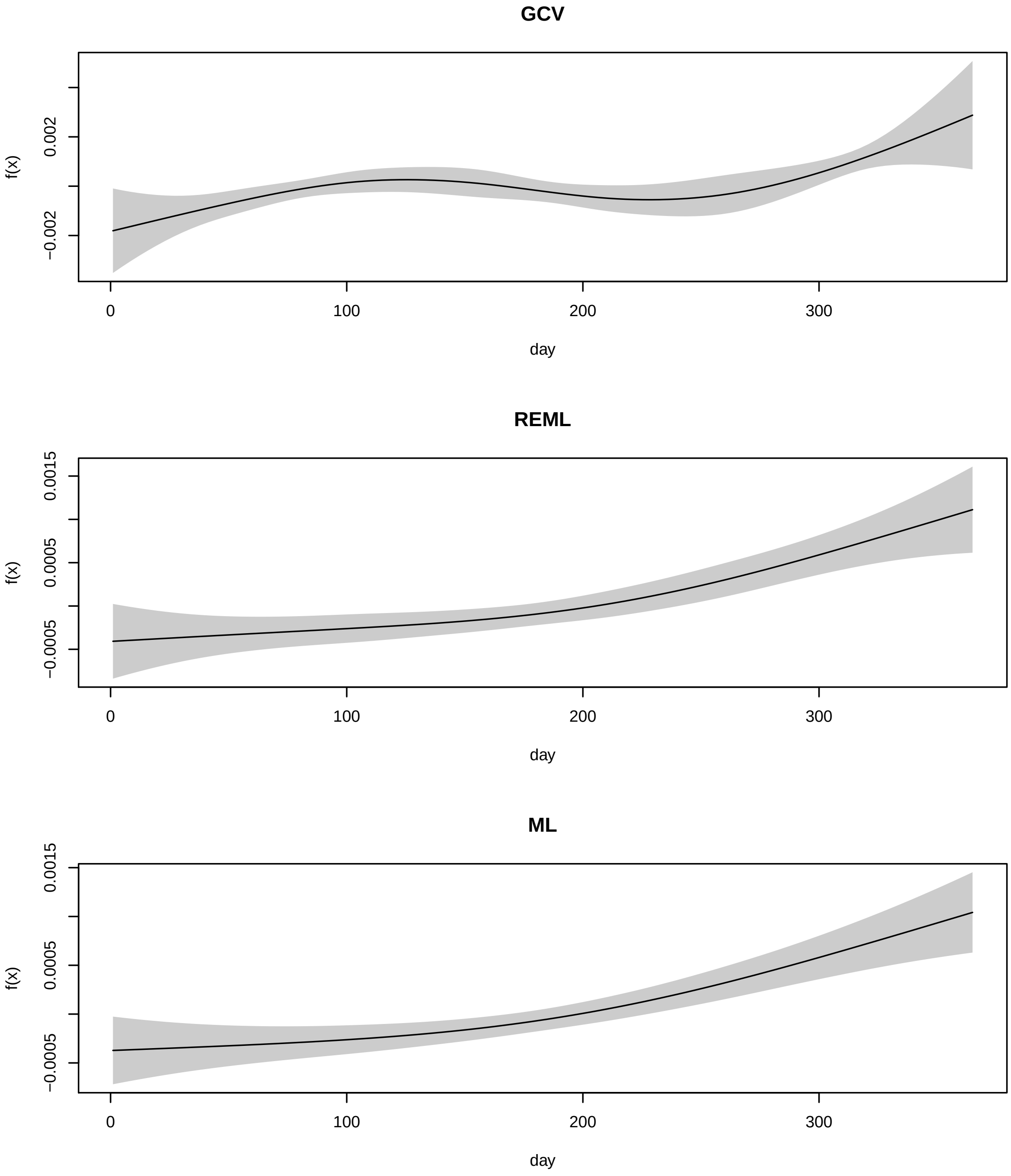

Figure 3.15: The coefficient function $f(\cdot)$ from model (3.6.1), with a second order random walk assumption for the regression parameters. The smoothing parameter was estimated with GCV, REML and ML.

Chapter 4

Fast and stable computation of the conditional AIC

The R-package lme4 is popular when it comes to inference in mixed models. This is due to the exceptionally fast and generic implementation. Nevertheless, up to date conditional prediction and model selection are not supported in the package. The conditional AIC proposed by Greven and Kneib (2010) offers an unbiased AIC based on the conditional distribution of the model. However, the computation of this criterion is not as simple as it is for other AIC criteria. This chapter describes techniques for fast and stable computation of the conditional AIC in mixed models estimated with lme4, as they are implemented in the R-package cAIC4. In addition to translating the findings of Greven and Kneib (2010) to the model formulations used in Bates et al. (2014b), this chapter deals with the implementation of conditional AICs proposed for non-Gaussian settings in Chapter 2 and Chapter 1. The methods are accompanied by examples, mainly taken from lme4 (Bates et al., 2014b).

4.1 The model representation

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

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

where $\boldsymbol{y}$ is n-dimensional, $\boldsymbol{\beta}$ is p-dimensional and $\boldsymbol{u}$ is q-dimensional. Hence, the

matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are $(n \times p)$ and $(n \times q)$ design matrices.
The unconditional distribution of the random effects $\boldsymbol{u}$ follows a multivariate Gaussian with mean $\mathbf{0}$ and positive semidefinite $(q \times q)$ covariance matrix, $\boldsymbol{D}_{\boldsymbol{\theta}}$,

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

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

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

with the lower triangular covariance factor $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$ and the variance parameter σ^2 of the error term of the conditional responses. Notice that the notation of $\boldsymbol{\theta}$ changes to previous chapters. This notation is adapted from Bates et al. (2014b) in order to increase comparability with standard notation in `lme4`. The R-package `lme4` owes part of its rapid computation to exploiting the structure of linear models. This structure is omitted in the above formulation and is mainly due to the sparse structure of $\boldsymbol{Z}$ and $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$.

4.2 Dealing with the boundary issues

A major issue in obtaining the conditional AIC is to account for potential parameters of $\boldsymbol{\theta}$ on the boundary of the parameter space. This needs to be done in order to ensure positive semidefiniteness of the covariance matrix $\boldsymbol{D}_{\boldsymbol{\theta}}$.
The restructuring of the model in order to obtain the cAIC is done automatically by `cAIC4`. To gain insight into the restructuring, an understanding of the mixed model formulas used in `lme4` is essential. For an in depth explanation of how the formula module of `lme4` works, see (Bates et al., 2014b, Section 2.1).
Suppose we want to fit a mixed model with two grouping factors `g1` and `g2`. Within the first grouping factor `g1` there are several continuous variables `v1`, `v2` and `v3` and within the second grouping factor there is only one variable `x`. Thus there are not only random intercepts but also random slopes that are possibly correlated within the groups. Such a model with response `y` would be called in `lme4` by

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

In mixed models fitted with `lme4`, the random effects covariance matrices $\boldsymbol{D}_{\boldsymbol{\theta}}$ always have block-diagonal structure. For instance, in the example from above the Cholesky factorized blocks of $\boldsymbol{D}_{\boldsymbol{\theta}}$ associated with each random effects term are

```
getME(m, "ST")

$g2
            [,1] [,2]
[1,]  1.18830317  NaN
[2,] -0.01488366    0

$g1
              [,1]        [,2] [,3] [,4]
[1,]  1.0184624863  0.00000000  NaN  NaN
[2,] -0.1438760320  0.05495795  NaN  NaN
[3,] -0.0007342374  0.19904408    0  NaN
[4,] -0.0883652260 -1.36463451  Inf    0
```

If any of the diagonal elements of the blocks are zero, the corresponding random effects terms are deleted from the formula. In `lme4` this is done conveniently by the component names list

```
m@cnms

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

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

Thus a new model formula can be obtained by designing a new list of components names:

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

$g2
[1] "(Intercept)"

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

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

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

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

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

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

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

The conditional AIC is computed with the new model. If there are no random effect terms left in the formula, a linear model and the conventional AIC are returned. The deleteZeroComponents functions additionally account for several special cases that may occur.

4.3 Computational matters

The corrected conditional AIC proposed in Greven and Kneib (2010) accounts for the uncertainty induced by the estimation of the random effects covariance parameters $\boldsymbol{\theta}$. In order to adapt the findings of Greven and Kneib (2010), a number of quantities from the `lmer` model fit need to be extracted and transformed. These computations are presented in the following. They are designed to minimize the computational burden and maximize the numerical stability. Parts of the calculations needed, for instance the Hessian of the ML/REML criterion, can also be found in Bates and DebRoy (2004). Notice, however, that `lme4` does not explicitly calculate these quantities but uses derivative free optimizers for the profile likelihoods. Some of the results used in this chapter are closely linked to those developed in Chapter 3.
A core ingredient of mixed models is the covariance matrix of the marginal responses $\boldsymbol{y}$. The inverse of the scaled covariance matrix $\boldsymbol{V}_0$ will be used in the following calculations:

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

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

$$\boldsymbol{L}_{\boldsymbol{\theta}}\boldsymbol{L}_{\boldsymbol{\theta}}^t = \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^t \boldsymbol{Z}^t \boldsymbol{Z} \boldsymbol{\Lambda}_{\boldsymbol{\theta}} + \boldsymbol{I}_q. \tag{4.3.1}$$

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

$$\begin{aligned} \boldsymbol{\Lambda}_{\boldsymbol{\theta}} \left(\boldsymbol{L}_{\boldsymbol{\theta}}^t\right)^{-1} \boldsymbol{L}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^t &= \left(\boldsymbol{Z}^t\boldsymbol{Z} + \left(\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^t\right)^{-1} \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{-1} \right)^{-1} \\ \Leftrightarrow \qquad \boldsymbol{I} - \boldsymbol{V}_0^{-1} &= \left(\boldsymbol{L}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^t \boldsymbol{Z}^t \right)^t \left(\boldsymbol{L}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^t \boldsymbol{Z}^t \right). \end{aligned}$$

Hence, the inverse of the scaled variance matrix $\boldsymbol{V}_0^{-1}$ can be efficiently computed with the help of the R-package `Matrix` that provides methods specifically for sparse matrices:

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

Notice that

```
solve(getME(m, "L"), Lambdat, system = "P")
```

accounts for a fill-reducing permutation matrix $\boldsymbol{P}$ associated (and stored) with $\boldsymbol{L}_{\boldsymbol{\theta}}$ (Bates et al., 2014b) and is thus equivalent to

```
P %*% Lambdat
```

Another quantity needed for the calculation of the corrected degrees of freedom in the conditional AIC are the derivatives of the scaled covariance matrix of the responses $\boldsymbol{V}_0$ with respect to the j-th element of the parameter vector $\boldsymbol{\theta}$:

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

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

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

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

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

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

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

Increased numerical stability can be derived by scaling the derivative matrices with their Frobenius-norm

```
  Wlist[[j]] <- Wlist[[j]]/norm(Wlist[[j]], type = "F")
```

Furthermore, the fixed random effects residual matrix introduced in Chapter 3 (denoted by $\boldsymbol{P}$ there) is essential to derive the corrected AIC of Theorem 3 in Greven and Kneib (2010). Adapting their notation, the matrix is

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

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

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

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

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

Accordingly, the computation in R can be done as follows:

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

With these components, the Hessian matrix

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

and the matrix

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

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

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

The inverse of $\boldsymbol{B}$ does not need to be calculated instead, if $\boldsymbol{B}$ is positive definite, a Cholesky decomposition and two backward solves are sufficient (cf. Chapter 3).

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

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

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

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

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

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

4.4 Example: sleepstudy

An example that is often used in connection with the R-package lme4 is the sleepstudy data from a study of the daytime performance changes of the reaction time during chronic sleep restriction (Belenky, Wesensten, Thorne, Thomas, Sing, Redmond, Russo, and Balkin, 2003). Eighteen volunteers were only allowed to spend three hours of their daily time in bed for one week. The speed (mean and fastest 10% of responses) and lapses (reaction times greater than 500 ms) on a psychomotor vigilance task were measured several times. The averages of the reaction times are saved as response variable Reaction in the data set. Each volunteer has an identifier Subject. Additionally the number of days of sleep restriction for each measurement is listed in the covariate Days. Thus the structure of the data set that will be analysed is

```
str(sleepstudy)

'data.frame': 180 obs. of  3 variables:
 $ Reaction: num  250 259 251 321 357 ...
 $ Days    : num  0 1 2 3 4 5 6 7 8 9 ...
 $ Subject : Factor w/ 18 levels "308","309","310",..: 1 1 1 1 1 1 ...
```

An example of what the sleepstudy data look like, can be derived by a sample of the 180 measurements:

```
sleepstudy[sample(180, 10),]

    Reaction Days Subject
124 280.5891    3     351
81  241.6083    0     335
39  305.3495    8     330
122 300.0576    1     351
169 350.7807    8     371
46  293.3187    5     331
82  273.9472    1     335
56  309.9976    5     332
29  261.0125    8     310
123 269.8939    2     351
```

Further insight into the data can be gained by a lattice plot from the `lattice`-package. The average reaction times of each volunteer are plotted against the days of sleep restriction with the corresponding linear regression line. Such a plot can be found in Figure 4.1:

```
require(lattice)
xyplot(Reaction ~ Days | Subject, sleepstudy, type = c("g","p","r"),
index = function(x,y) coef(lm(y ~ x))[1],
xlab = "Days of sleep restriction",
ylab = "Average reaction time (ms)", aspect = "xy")
```

The conditional AIC can be used to find the model that best predicts future observations, assuming that future observations share the same random effects as the ones used for the model fitting. In the case of this data set, using the cAIC for model choice corresponds to finding the model that best predicts future reaction times of the volunteers who took part in the study.
After looking at the lattice plot, a first model that could be applied is one with a random intercept and a random slope for `Days` within each volunteer (`Subject`):

$$y_{ij} = \beta_0 + \beta_1 \cdot \text{day}_{ij} + u_{j0} + u_{j1} \cdot \text{day}_{ij} + \epsilon_{ij} \tag{4.4.1}$$

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

$$\begin{pmatrix} u_{j0} \\ u_{j1} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \tau_{12}^2 \\ \tau_{12}^2 & \tau_2^2 \end{pmatrix}\right).$$

In the preceding notation $\tau_1^2 = \theta_1$, $\tau_2^2 = \theta_2$ and $\tau_{12}^2 = \theta_3$. That τ_{12}^2 is not necessarily zero indicates that the random intercept and the random slope are allowed to be correlated.

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

Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ 1 + Days + (1 + Days | Subject)
   Data: sleepstudy
REML criterion at convergence: 1743.628
Random effects:
 Groups   Name        Std.Dev. Corr
 Subject  (Intercept) 24.740
```

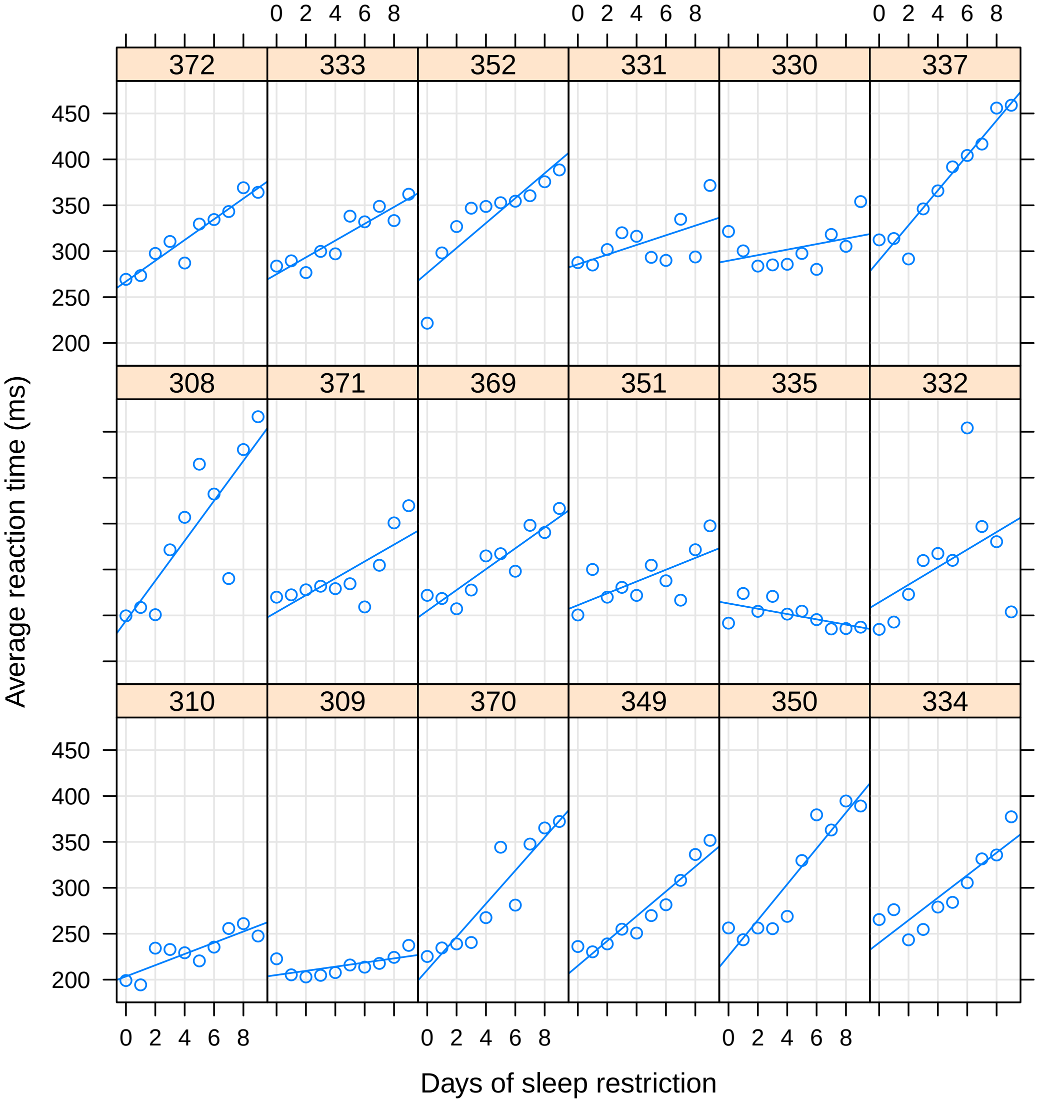

Figure 4.1: Lattice plot of the sleepstudy data. For each volunteer there is one panel. The identification number of each volunteer is in the heading of the panels. In the panels the days of sleep restriction is plotted against the reaction time and a regression line is added for each volunteer/panel.

```
          Days          5.922   0.07
 Residual              25.592
Number of obs: 180, groups:  Subject, 18
Fixed Effects:
(Intercept)         Days
     251.41        10.47
```

The output shows that the within-subject correlation between the random intercepts u_{j0} and the random slopes u_{j1} is low, being estimated as 0.07. Hence, there seems to be no evidence that the initial reaction time of the volunteers has systematic impact on the pace of ascending reaction time following the sleep restriction.

Consequently, a suitable model might be one in which the correlation structure between both is omitted beforehand. The model for the response therefore stays the same as in (4.4.1), but the random effects covariance structure is predefined as

$$\begin{pmatrix} u_{j0} \\ u_{j1} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & 0 \\ 0 & \tau_2^2 \end{pmatrix}\right).$$

Such a model without within-subject correlation is called by

```
(m2 <- lmer(Reaction ~ 1 + Days + (1|Subject) + (0 + Days|Subject),
sleepstudy))

Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ 1 + Days + (1 | Subject) + (0 + Days | Subject)
   Data: sleepstudy
REML criterion at convergence: 1743.669
Random effects:
 Groups    Name        Std.Dev.
 Subject   (Intercept) 25.051
 Subject.1 Days         5.988
 Residual              25.565
Number of obs: 180, groups:  Subject, 18
Fixed Effects:
(Intercept)         Days
     251.41        10.47
```

Notice that the estimates of standard deviations of the random effects do not differ much between the first and the second model. To decide which model is more appropriate in terms of subject specific prediction, the conditional AIC can be used. Calling the `caic`-function from the `cAIC4`-package gives the output:

```
cAIC(m1)

$loglikelihood
[1] -824.507

$df
[1] 31.30192

$reducedModel
NULL

$new
[1] FALSE

$caic
[1] 1711.618
```

The conditional log likelihood and the corrected degrees of freedom are the first two elements of the resulting list. The third element is called `reducedModel` and is the model without the random effects covariance parameters that were estimated to lie on the boundary of the parameter space (cf. Section 4.2) and `NULL` if there were none on the boundary. The fourth element says if such a new model was fit because of the boundary issue, which was not the case here. The last element is the conditional AIC as proposed in Greven and Kneib (2010).
The cAIC of the second model `m2` is:

```
cAIC(m2)$caic

[1] 1710.426
```

From a conditional perspective, the second model is thus preferred to the first one. This confirms the assertion that the within-subject correlation can be omitted in the model.

There are a number of further possible models for this data. For instance, the random slope could be excluded from the model. In this model the pace of ascending reaction time does not systematically vary between the volunteers. This model is estimated by

```
(m3 <- lmer(Reaction ~ 1 + Days + (1|Subject), sleepstudy))

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

The conditional AIC of this model is

```
cAIC(m3)$caic

[1] 1767.118
```

This is far larger than the cAIC for the two preceding models. The lattice plot in Figure (4.1) already indicated that there is strong evidence of subject-specific (random) slops. This is also reflected by the cAIC.

The conditional AIC is appropriate for choosing between a simple null model without any random effects and a complex model incorporating random effects, as noted by Greven and Kneib (2010). Thus it is reasonable to compare the cAIC of the three previous mixed models with the standard AIC for a linear model.

```
AIC(lm(Reaction ~ 1 + Days, sleepstudy))

[1] 1906.293
```

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

4.5 Conditional AIC for Poisson and binary

The cAIC4-package additionally offers a conditional AIC for conditionally Poisson distributed responses and an approximate conditional AIC for binary data. The Poisson cAIC is given in Corollary (1.2.2), see also Saefken et al. (2014) and the degrees of freedom term for the binary data are proposed in Chapter 2, equation (2.2.7) (Efron, 2004).

Making use of the fast refit() function of the lme4-package, both cAICs can be computed moderately fast, since $n-d$ and n model refits are required, with n being the number of observations and d the number of responses that are zero for the Poisson responses. The cAIC for Poisson response is in the following computed with the grouseticks data set from the lme4-package as illustration. The methods for deriving the degrees of freedom for binary responses are somewhat similar and are therefore omitted.

The grouseticks data set was originally published in Elston, Moss, Boulinier, Arrowsmith, and Lambin (2001). It contains information about the aggregation of parasites, so-called sheep ticks on red grouse chicks. The covariates of the data set are given in Table 4.1.

Covariate	Discription
INDEX	identifier of the chick
TICKS	the number of ticks sampled
BROOD	the brood number
HEIGHT	height above sea level in meters
YEAR	the year as 95, 96 or 97
LOCATION	the geographic location code

Table 4.1: The covariates of the grouseticks data set.

The number of ticks is the response variable. In a first model, the expected number of ticks λ_{ijk} is modelled in dependence of the year and the height as fixed effects, and for each grouping variables BROOD, INDEX and LOCATION a random intercept is incorporated. The full model is

$$\log\left(\mathbb{E}\left(\texttt{TICKS}_{ijk}\right)\right) = \log\left(\lambda_{ijk}\right) = \beta_0 + \beta_1 \cdot \texttt{YEAR}_{ijk} + \beta_2 \cdot \texttt{HEIGHT}_{ijk} + u_i + u_j + u_k \tag{4.5.1}$$

with random effects distribution

$$\begin{pmatrix} u_i \\ u_j \\ u_k \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & 0 & 0 \\ 0 & \tau_2^2 & 0 \\ 0 & 0 & \tau_3^2 \end{pmatrix}\right).$$

Before fitting the model the covariates `HEIGHT` and `YEAR` are centred and stored in the data set `grouseticks_cen`.

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

Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: TICKS ~ YEAR + HEIGHT + (1|BROOD) + (1|INDEX) + (1|LOCATION)
   Data: gs
     AIC      BIC   logLik deviance df.resid
1845.482 1869.476 -916.741 1833.482      397
Random effects:
 Groups   Name        Std.Dev.
 INDEX    (Intercept) 5.458e-01
 BROOD    (Intercept) 1.211e+00
 LOCATION (Intercept) 2.326e-05
Number of obs: 403, groups:  INDEX, 403; BROOD, 118; LOCATION, 63
Fixed Effects:
(Intercept)         YEAR       HEIGHT
    0.47235     -0.48026     -0.02571
```

The degrees of freedom for a conditional AIC with conditionally Poisson distributed responses as in (4.5.1) is

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l} y_{ijk}\left(\log\left(\hat{\lambda}_{ijk}\right) - \log\left(\hat{\lambda}_{ijk}^{-}\right)\right),$$

where $\hat{\lambda}_{ijk}^{-}$ is the predicted number of ticks estimated with the observation $y_{ijk} - 1$ instead of y_{ijk}. Notice that for $y_{ijk} = 0$ no new model fit is required and the new estimator is set to the old $\log\left(\hat{\lambda}_{ijk}\right) = \log\left(\hat{\lambda}_{ijk}^{-}\right)$.

Hence, for the computation of the degrees of freedom the non-zero responses need to be identified and a matrix with the responses in each column needs to be created.

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

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

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

Now the `refit()` function can be applied to the columns of the matrix in order to obtain the estimates $\log\left(\hat{\lambda}_{ijk}^{-}\right)$ from the reduced data.

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

The computation of the degrees of freedom is straightforward:

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

[1] 205.6579
```

The conditional AIC based on these degrees of freedom is obtained by call of the `cAIC`-function

```
cAIC(p1)

$loglikelihood
[1] -572.0136
```

```
$df
[1] 205.6579

$reducedModel
NULL

$new
[1] FALSE

$caic
[1] 1555.343
```

The output is the same as for Gaussian linear mixed models. It becomes apparent that there is a substantial difference between the conditional and the marginal AIC: In the output of the model the marginal AIC is reported to be 1845.48. Notice that the marginal AIC is biased (Greven and Kneib, 2010).

With regard to the full model, the standard deviations of the random effects are rather low. It may thus be possible to exclude one of the grouping from the model, only maintaining two random effects. There are three possible models with one of the random effects terms excluded.

If the random intercept associated with LOCATION is excluded, the model is

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

[1] 1555.344
```

The conditional AIC is almost the same as for the full model. It may thus make sense to choose the reduced model and for the prediction of the number of ticks not to make use of the random intercept associated with the LOCATION grouping.

Another possible model can be obtained by omitting the random intercepts for the INDEX grouping structure instead of those associated with LOCATION. This would make the model considerably easier, since each chick has an INDEX and hence a random intercept is estimated for each observation.

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

[1] 1842.189
```

In comparison with the two preceding models the large cAIC documents that the subject-specific random intercept for each observation should be included.

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

```
formel <- TICKS ~ YEAR + HEIGHT + (1|INDEX) + (1|LOCATION)
p4  <- glmer(formel, family = "poisson", data = grouseticks_cen)
cAIC(p4)$caic

[1] 1594.411
```

The cAIC is higher than the cAICs for the full model and the model without the LOCATION grouping structure. Consequently, either the full model or the model without the LOCATION grouping structure is favoured by the cAIC. Keeping in mind that a model should be as simple as possible the author favours the latter.

Conclusion & Outlook

1 Summary

The basic underlying idea of model selection is to choose the model with the lowest estimated prediction error. Different measures of the prediction error are possible, as suggested in Chapter 2. The difference between the apparent and the expected prediction error is the covariance penalty. If the error is measured with the deviance error, the covariance penalties are the degrees of freedom of the model. Estimation of the covariance penalties can be done by bootstrap, cross-validation or the Steinian method. For Gaussian models the Steinian is particularly suitable. The method from Stein (1972) makes the degrees of freedom an observable quantity. Thus comparing different models by the estimated deviance prediction error and estimating the degrees of freedom using the Steinian method is equivalent to AIC based model selection.

The framework presented in Chapter 1 generalizes this method as far as possible. As a result, methods for Possion and exponentially distributed responses are available. In contrast to the Gaussian, these methods involve refitting of the model and hence have higher computational cost. The fact that the theoretical results are useful when analysing data is demonstrated in an application to a data set on tree growth.

Leaving the limited framework of the first chapter, far more general approaches are possible. These are presented in Chapter 2. Other prediction error measures and many more distributions, even beyond the exponential family, are available in this framework. It pays for this universal applicability by substantially less rigorous derivations. For the estimation of the covariance penalties and thus also for the prediction errors, three different schemes are presented: cross-validation, bootstrap and an approximate Steinian method. The approximate Steinian method for continuous distributions is a generalization of the one for the Gaussian. The Steinian covariance penalty is the scaled sum of the derivatives of the estimated mean with respect to the observed data at the fitted values. Hence, it measures the sensitivity of the estimator with respect to the data. This emphasizes the connection to data perturbation methods (Shen and Huang, 2006).

The popularity of the AIC is based, at least to a certain extent, on the fact that the degrees of freedom for linear models are the number of parameters. This makes the AIC ease to compute. In semiparametric models the degrees of freedom are slightly more involved. The degrees of freedom are commonly estimated by the trace of the hat matrix. This ignores the uncertainty induced by the estimation of the smoothing parameter. For REML based estimation this was first recognized by Greven and Kneib (2010). Chapter 3 provides a much wider framework by defining the degrees of freedom of the smoothing parameter for general smoothing parameter estimation methods, for instance GCV or Mallows C_p. Furthermore, the degrees of freedom of the smoothing parameter have interesting geometrical properties that affirm how appealing this approach is from a theoretical point of view.

If results and methods from statistics persist is becoming increasingly dependent on the availability of a stable and fast implementation in a statistical software. Chapter 4 presents the R-package `cAIC4` that accounts for this dependency. The `cAIC4` package implements many of the methods that this thesis deals with, especially the conditional AIC for Gaussian mixed models by Greven and Kneib (2010). The challenge of this chapter, in contrast to the preceding chapters, is not methodological but technical. The `cAIC4` implementation comprises several ideas of how the computation of the conditional AIC can be made faster and more reliable. This is achieved, for instance, by avoiding explicit computation of inverse matrices and exploiting sparse structures.

To sum up: the first three chapters of this thesis provide general frameworks for the estimation of the conditional AIC in mixed models, the estimation of the prediction error in mixed models, and the estimation of the degrees of freedom in semiparametric regression respectively. The fourth chapter gives insight into computational aspects of the three preceding chapters.

2 Future research

Model choice and variable selection in mixed and semiparametric regression remain an active field of research. This thesis offers several starting points for future research in this field. Research questions arise in methodological, computational and applied fields of statistics, making this topic both challenging and interesting. A few examples of future directions of research and possible extensions of this thesis are given below:

An approximate conditional AIC that overcomes many of the problems denoted in Chapter 1 is proposed in Wood et al. (2014). However, the behaviour of this AIC needs to be analysed in different situations, for example the small sample performance especially in comparison with the unbiased conditional AIC as proposed in Chapter 3. Furthermore, the approximate conditional AIC is only used in connection with REML based smoothing parameter estimation, although smoothing parameter uncertainty is also apparent for smoothing parameter choice with prediction error criteria, as stated in Chapter 3.

Another substantial field of future research is model choice and variable selection for regression models that go beyond the mean. For instance, selecting models with structured additive predictors for location, scale and shape or with multivariate responses offer interesting perspectives. Particularly fast available information criteria for such high-dimensional and computational costly models are more advantageous than other methods like cross-validation or bootstrap. Additionally, regression models that do not, or at least do not merely, focus on the mean of the distribution might require other prediction error measures than the prominent deviance error and the closely-linked Kullback-Leibler distance. Thus a focussed information criterion (Claeskens and Hjort, 2003) adapted to structured additive predictors for location, scale and shape, accounting for the smoothing parameter uncertainty, might be preferable.

From a methodological point of view, the findings in Chapter 3 suggest that results are attainable. Further insight into the behaviour of GCV and REML from the perspective of the degrees of freedom would be especially desirable. This could build upon the results of Reiss and Ogden (2009). Furthermore, Chapter 3 relates the degrees of freedom of the smoothing parameter to the field of differential geometry. This link has not been fully exploited yet.

The conditional AIC estimates the conditional relative deviance error of a semiparametric or mixed model, taking into account the uncertainty induced by the smoothing parameter. It is therefore applicable for choosing among models with distinct semiparametric structure. In this thesis it was mostly used for the choice between presence and absence of a semiparametric component. There are several other choices to make in semiparametric (and mixed) modelling, for instance, the basis dimension, the number of knots for a smoothing spline or the order of the random walk for P-splines and Gaussian Markov random fields. All this has not been investigated in this thesis and could possibly enhance the use of the cAIC.

Since in many situations it is not reasonable to assume there exists one true model, a prediction should be based on more than one model. Model averaging (Hjort and Claeskens, 2003) comprises this approach. For mixed models this was suggested by Zhang, Zou, and Liang (2014). An adaptation to semiparametric models and an implementable version in R have yet to be carried out. Furthermore, the process of model selection is an integral part of model fitting. Thus standard errors and confidence intervals after model selection could increase accuracy of the final model. This has recently been investigated by Efron (2013) and a connection to the cAIC seems feasible.

The marginal AIC is frequently used for model selection in mixed models. As shown by Greven and Kneib (2010), the marginal AIC is biased. A criterion that overcomes this problem, for instance based on bootstrap methods, would be worthwhile. Moreover, there are challenging prediction problems from applied sciences where both prediction focuses (marginal and conditional) are needed. For instance, in forestry the above ground biomass of a pixel (section of land) is predicted. The pixels are grouped by tree stands. Thus, a mixed model seems sensible. If the biomass is predicted for a pixel from a stand with predicted random effect, a conditional AIC is appropriate. If, on the other hand, the biomass is predicted for a pixel from a new stand, a marginal AIC is appropriate. Thus, for one data set two distinct models may be needed.

Bibliography

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

Bates, D. M. and S. DebRoy (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis 91*(1), 1–17.

Bates, D. M., M. Mächler, B. Bolker, and S. Walker (2014a). Fitting Linear Mixed-Effects Models using lme4.

Bates, D. M., M. Mächler, B. Bolker, and S. Walker (2014b). lme4: Linear mixed-effects models using Eigen and S4.

Belenky, G., N. J. Wesensten, D. R. Thorne, M. L. Thomas, H. C. Sing, D. P. Redmond, M. B. Russo, and T. J. Balkin (2003). Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. *Journal of sleep research 12*(1), 1–12.

Böhner, J., K. R. McCloy, and J. Strobl (Eds.) (2006). *SAGA - analysis and modelling applications*, Volume 115 of *Göttinger geographische Abhandlungen*. Göttingen: Goltze.

Breslow, N. E. and D. G. Clayton (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association 88*(421), 9.

Burnham, K. P. and D. R. Anderson (2010). *Model selection and multimodel inference: A practical imformation-theoretic approach* (2. ed., [Nachdr.] ed.). New York, NY [u.a.]: Springer.

Butler Manning, D. (2007). *Stand structure, gap dynamics and regeneration of a semi-natural beech forest on limestone in central Europe: A case study: Univ.,* *Diss.–Freiburg (Breisgau), 2007*, Volume 38 of *Schriftenreihe Freiburger forstliche Forschung*. Freiburg: Waldbau-Institut.

Chen, L. H. Y. (1975). Poisson Approximation for Dependent Trials. *The Annals of Probability 3*(3), 534–545.

Claeskens, G. and N. L. Hjort (2003). The Focused Information Criterion. *Journal of the American Statistical Association 98*(464), 900–916.

Claeskens, G. and N. L. Hjort (2010). *Model selection and model averaging*, Volume 2010: 1 of *Cambridge series in statistical and probabilistic mathematics*. Cambridge: Cambridge University Press.

Crainiceanu, C. M. and D. Ruppert (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(1), 165–185.

Crainiceanu, C. M., D. Ruppert, G. Claeskens, and M. P. Wand (2005). Exact Likelihood Ratio Tests for Penalised Splines. *Biometrika* (Vol. 92, No. 1), 91–103.

Craven, P. and G. Wahba (1978). Smoothing noisy data with spline functions. *Numerische Mathematik 31*(4), 377–403.

Davison, A. C. (2008). *Statistical models*, Volume 11 of *Cambridge series in statistical and probabilistic mathematics*. Cambridge: Cambridge University Press.

Delattre, M., M. Lavielle, and M.-A. Poursat (2014). A note on BIC in mixed-effects models. *Electron. J. Statist.*, 456–475.

Donohue, M. C., R. Overholser, R. Xu, and F. Vaida (2011). Conditional Akaike information under generalized linear and proportional hazards mixed models. *Biometrika 98*(3), 685–700.

Efron, B. (1986). How Biased is the Apparent Error Rate of a Prediction Rule? *Journal of the American Statistical Association 81*(394), 461–470.

Efron, B. (2004). The Estimation of Prediction Error. *Journal of the American Statistical Association 99*(467), 619–632.

Efron, B. (2013). Estimation and Accuracy After Model Selection. *Journal of the American Statistical Association 109*(507), 991–1007.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B -splines and penalties. *Statistical Science 11*(2), 89–121.

Einstein, A. (1934). On the Method of Theoretical Physics. *Philosophy of Science 1*(2), 163–169.

Elston, D. A., R. Moss, T. Boulinier, C. Arrowsmith, and X. Lambin (2001). Analysis of aggregation, a worked example: numbers of ticks on red grouse chicks. *Parasitology 122*(05), 563–569.

Fahrmeir, L., T. Kneib, S. Lang, and B. D. Marx (2013). *Regression Models, Methods and Applications* (2 ed.). Springer.

Forster, O. (1999). *Analysis 3, Integralrechnung im IRn und Anwendungen* (3., überarb. Aufl. ed.), Volume / Otto Forster ; 3 of *Studium Aufbaukurs Mathematik*. Wiesbaden: vieweg.

Gilbert, P. and R. Varadhan (2012). numDeriv: Accurate Numerical Derivatives.

Golub, G. H. and Van Loan, Charles F (1996). *Matrix computations* (3rd ed. ed.). Johns Hopkins studies in the mathematical sciences. Baltimore: Johns Hopkins University Press.

Green, P. J. (1984). Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)* (46), 149–192.

Greven, S. and T. Kneib (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika 97*(4), 773–789.

Gu, C. and G. Wahba (1991). Minimizing GCV/GML Scores with Multiple Smoothing Parameters via the Newton Method. *SIAM Journal on Scientific and Statistical Computing 12*(2), 383–398.

Harville, D. A. (1974). Bayesian Inference for Variance Components Using Only Error Contrasts. *Biometrika 61*(2), 383.

Hastie, T. and R. Tibshirani (1990). *Generalized additive models* (1st ed ed.), Volume 43 of *Monographs on statistics and applied probability*. London and New York: Chapman and Hall.

Hjort, N. L. and G. Claeskens (2003). Frequentist Model Average Estimators. *Journal of the American Statistical Association 98*(464), 879–899.

Hodges, J. S. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika 88*(2), 367–379.

Hudson, H. M. (1978). A Natural Identity for Exponential Families with Applications in Multiparameter Estimation. *The Annals of Statistics 6*(3), 473–484.

Hurvich, C. M. and C.-L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika 76*(2), 297–307.

Kass, R. E. and D. Steffey (1989). Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models). *Journal of the American Statistical Association 84*(407), 717.

Kauermann, G., T. Krivobokova, and L. Fahrmeir (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society. Series B (Methodological)* (71).

Kneib, T. (2005). *Mixed model based inference in structured additive regression.* Dr. Hut - Verlag.

Kneib, T. (2013). Beyond mean regression. *Statistical Modelling 13*(4), 275–303.

Krivobokova, T. (2013). Smoothing parameter selection in two frameworks for penalized splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(4), 725–741.

Kullback, S. and R. A. Leibler (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

Laird, N. M. and J. H. Ware (1982). Random-Effects Models for Longitudinal Data. *Biometrics 38*(4), 963–974.

Lian, H. (2012). A note on conditional Akaike information for Poisson regression with random effects. *Electronic Journal of Statistics 6*(0), 1–9.

Liang, H., H. Wu, and G. Zou (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika 95*(3), 773–778.

Lindgren, F., H. Rue, and J. Lindström (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(4), 423–498.

McGilchrist, C. (1994). Estimation in Generalized Mixed Models. *Journal of the Royal Statistical Society. Series B (Methodological) 56*(1), 61–69.

McLean, M. W., G. Hooker, A.-M. Staicu, F. Scheipl, and D. Ruppert (2014). Functional Generalized Additive Models. *Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America 23*(1), 249–269.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General) 135*(3), 370.

Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*(3), 545–554.

Ramsay, J. O. and B. W. Silverman (1997). *Functional data analysis*. Springer series in statistics. New York: Springer.

Ramsay, J. O. and B. W. Silverman (2002). *Applied functional data analysis: Methods and case studies*. Springer series in statistics. New York: Springer.

Reiss, P. T. and R. T. Ogden (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society. Series B (Methodological)* (71), 505–523.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*(3), 507–554.

Rue, H. and L. Held (2005). *Gaussian Markov random fields: Theory and applications*, Volume 104 of *Monographs on statistics and applied probability*. Boca Raton: Chapman & Hall/CRC.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 319–392.

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

Saefken, B., T. Kneib, C.-S. van Waveren, and S. Greven (2014). A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models. *Electronic Journal of Statistics 8*(0), 201–225.

Saefken, B., D. Ruegamer, S. Greven, and T. Kneib (2014). cAIC4: Conditional Akaike information criterion for lme4.

Self, S. G. and K.-Y. Liang (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *Journal of the American Statistical Association* (82), 605–610.

Shen, X. and H.-C. Huang (2006). Optimal Model Assessment, Selection, and Combination. *Journal of the American Statistical Association 101*(474), 554–568.

Shen, X., H.-C. Huang, and J. Ye (2004). Adaptive Model Selection and Assessment for Exponential Family Distributions. *Technometrics 46*(3), 306–317.

Stein, C. (Ed.) (1972). *A bound for the error in the normal approximation to the distribution of a sum of dependent random variables*, Sixth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, Calif. University of California Press.

Stram, D. O. and Jae Won Lee (1994). Variance Components Testing in the Longitudinal Mixed Effects Model. *Biometrics 50*(4), 1171–1177.

Vaida, F. and S. Blanchard (2005). Conditional Akaike information for mixed-effects models. *Biometrika 92*(2), 351–370.

Wahba, G. (1985). A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem. *Ann. Statist.*, 1378–1402.

Wahba, G. (1990). *Spline models for observational data*, Volume 59 of *CBMS-NSF Regional Conference series in applied mathematics.* Philadelphia, Pa.: Society for Industrial and Applied Mathematics.

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference.* Springer texts in statistics. New York: Springer.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(1), 95–114.

Wood, S. N. (2006). *Generalized Additive Models: An introduction with R.*

Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(3), 495–518.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (73), 3–36.

Wood, S. N. (2013a). A simple test for random effects in regression models. *Biometrika*.

Wood, S. N. (2013b). On p-values for smooth components of an extended generalized additive model. *Biometrika 100*(1), 221–228.

Wood, S. N., N. Pya, and B. Säfken (2014). On doubly generalized additive models. *Working Paper*.

Ye, J. (1998). On Measuring and Correcting the Effects of Data Mining and Model Selection. *Journal of the American Statistical Association 93*(441), 120–131.

Yu, D. and K. K. W. Yau (2012). Conditional Akaike information criterion for generalized linear mixed models. *Computational Statistics & Data Analysis 56*(3), 629–644.

Zhang, X., G. Zou, and H. Liang (2014). Model averaging and weight choice in linear mixed-effects models. *Biometrika 101*(1), 205–218.