Mixed Models based on Likelihood Boosting

Florian Reithinger

München 2006

Mixed Models based on Likelihood Boosting

Florian Reithinger

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Florian Reithinger
aus München

München, den 28. September 2006

Erstgutachter: Prof. Dr. Gerhard Tutz
Zweitgutachter: PD. Dr. Christian Heumann
Tag der mündlichen Prüfung: 20. Dezember 2006

Contents

List of Tables

LIST OF TABLES

List of Figures

Notation

Mixed model notation

$$
\begin{aligned}
y_{it} &= x_{it}^T\beta + z_{it}^T b_i + \epsilon_{it} = x_{it}^T\beta + \textstyle\sum_{j=1}^{c} z_{itj} b_i^{(j)} + \epsilon_{it}, \qquad i \in \{1,\dots,n\} \qquad t \in \{1,\dots,T_i\} \\
y_i &= X_i\beta + Z_i b_i + \epsilon_i, \qquad i \in \{1,\dots,n\} \\
y &= X\beta + \mathbb{Z}b + \epsilon = X\beta + \textstyle\sum_{j=1}^{c} Z_{.(j)} b^{(j)} + \epsilon \\
y_{(i)} &= x_{(i)}^T\beta + z_{(i)}^T b + \epsilon_{(i)}, i \in \{1,\dots,N\}
\end{aligned}
$$

y_{it}	Response of cluster i at observation t
$h(.)$	Response function, inverse of link-function g
x_{it}	Design vector for fixed effects (cluster i, measurement t)
$z_{it}^T = [z_{it1},\dots,z_{itc}]$	design vector (cluster i, measurement t)
X_i	Design matrix for fixed effects (cluster i)
Z_i	Design matrix for random effects (cluster i)
$Z = (Z_1^T,\dots,Z_n^T)^T$	Design matrix for random effects (stacked version)
n	Clusters in total
$X^T = (X_1^T,\dots,X_n^T)$	Design matrix for fixed effects (complete dataset)
$\mathbb{Z}$	Design matrix for random effects (usually block-diagonal version, complete dataset)
	$\mathbb{Z} = \text{bdiag}(Z_1,\dots,Z_n)$ in longitudinal settings
$x_{(i)}^T$	i-th row of X
$z_{(i)}^T$	i-th row of $\mathbb{Z}$
$y_{(i)}^T$	i-th element of y
$\beta^T = (\beta_1,\dots,\beta_p)$	Parameter vector for fixed effects
b_i	Vector of random effects for cluster i
$b = (b_1^T,\dots,b_n^T)^T$	Vector of random effects
$N = \sum_{i=1}^{n} T_i$	Observations in total

Notation

$p(b;\rho)$	Mixing density
$\tilde{p}(a;\rho)$	Standardized mixing density with standardized random variable a
ρ	Parameter vector for the covariance structure of random effects
θ	ρ and nuisance parameters
$Q(\rho)$	Covariance matrix for the random effect b_i
$\mathbb{Q}(\rho)$	Covariance matrix for the random effect b
$V_i := V_i(\theta)$	Marginal covariance of the cluster i
$V := V(\theta)$	Marginal covariance over all clusters
ϱ	Correlation between two covariates
$f(.)$	Density or conditional density of y or y given b
$q = \sum_{j=1}^{c} q_j$	Dimension of b, dimension of $b^{(j)}$
$\mathbb{E}(.)$	Expectation
σ_ϵ^2	Error term, $e_{it} \sim N(0, \sigma_\epsilon^2)$
θ	d-dimensional vector of parameters for variance components
trace()	Trace of a matrix
$f(y,b) = f(y\|b)p(b;\rho)$	Joint density of y and b
rows(A,i,j)	Submatrix of matrix A from row i to row j
elem(y,i,j)	Subvector of vector y from element i to element j
$vec(A)$	Symmetric direct operator on symmetric matrix A.
	Vector from the lower triangular entries of matrix A
$vech(A)$	Vech operator on symmetric matrix A
	Vector from rows of matrix A.
c	Random design matrix has c components of the
$Z_{.(j)}$	Partitioned random effect design matrix associated with component
$b^{(j)}$	Partitioned random effect associated with component j

Additve mixed model notation

$$
\begin{array}{lll}
y_{it} & = x_{it}^T\beta + \sum_{j=1}^m \phi_{itj}^T\alpha_j + z_{it}^T b_i + \epsilon_{it} & , \qquad i \in 1,\dots,n \qquad t \in \{1,\dots,T_i\} \\
y_i & = X_i\beta + \sum_{j=1}^m \Phi_{ij}\alpha_j + Z_i b_i + \epsilon_i & = X_i\beta + \Phi_{i.}\alpha + Z_i b_i + \epsilon_i, \qquad i\{\in 1,\dots,n\} \\
Y & = X\beta + \Phi\alpha + \mathbb{Z}b + \epsilon & = X\beta + \sum_{j=1}^m \Phi_{.j}\alpha_j + \mathbb{Z}b + \epsilon \\
y_{(i)} & = x_{(i)}^T\beta + \Phi_{(i)}^T\alpha + z_{(i)}^T b + \epsilon_{(i)} & = x_{(i)}^T\beta + \sum_{j=1}^m (\phi^{(j)}(u_{(i)j}))^T\alpha_j + z_{(i)}^T b + \epsilon_{(i)}, i \in \{1,\dots,N\} \\
 & & = x_{(i)}^T\beta + \sum_{j=1}^m \phi_{(i)j}^T\alpha_j + z_{(i)}^T b + \epsilon_{(i)}, i \in \{1,\dots,N\}
\end{array}
$$

$\alpha_{(j)}(.)$	Unspecified j-th function
u_{itj}	Measured covariate for the j-th unspecified function in cluster i at measurement t
$\alpha_{(j)}(u_{itj})$	Function evaluation of the measured covariate for j-th function $\alpha_{(j)}(.)$ in cluster i at measurement t
M	Dimension of the spline basis
m	Number of unspecified functions
$\phi_s^{(j)}(.)$	s-th basis function for variable j
$\phi^{(j)}(.)^T = (\phi_1^{(j)}(.),\dots,\phi_M^{(j)}(.))$	Basis functions for variable j (M-dimensional, vector)
$\phi_{itj} = \phi^{(j)}(u_{itj})$	Function evaluation of covariate u_{itj} (vector)
$u_{i.j}^T = (u_{i1j},\dots,u_{iT_ij})$	Vector of covariates needed for function j in cluster i
$\Phi_{ij} = \Phi_{i.j} = (\phi_{i1j},\dots,\phi_{iT_ij})^T$	Matrix for elementwise basis Function evaluations for the j-th function of covariates $u_{i.j}$
$u_{..j}^T = (u_{1.j}^T,\dots,u_{n.j}^T)$	Vector of covariates needed for function j (complete dataset)
$\Phi_{.j} := \Phi_{..j} = (\Phi_{1.j}^T,\dots,\Phi_{n.j}^T)^T$	Matrix for elementwise basis function evaluations for the j-th function of covariates $u_{..j}$
$\Phi_{i.} := \Phi_{i..} = (\Phi_{i.1},\dots,\Phi_{i.m})$	Matrix for basis function evaluation for Covariates $u_{i.1},\dots,u_{i.m}$ in cluster i
$\Phi := \Phi_{...} = (\Phi_{1..}^T,\dots,\Phi_{n..}^T)^T = (\Phi_{..1},\dots,\Phi_{..m})$	Matrix for basis function evaluation of all covariates
α_j	M-dimensional vector of basis coefficients needed for approximation $\alpha_{(j)}(u) = (\phi^{(j)})^T\alpha_j$
$\alpha^T = (\alpha_1^T,\dots,\alpha_m^T)$	Vector of all basis coefficients
$\Phi_{(i)}$	i-th row of matrix $\Phi_{...}$
K_α	Penalty matrix for all components including fixed effects
$u_{(i)j}$	i-th entry of vector $u_{..j}$
λ	Smoothing parameter
$X_{\Phi i} = [X_i, \Phi_{i1},\dots,\Phi_{im}]$	Generalized design matrix for fixed and smooth effects
$\phi^{(i,j)}(.) = \phi^{(i)}(.) \odot \phi^{(j)}(.)$	Elementwise Kronecker product of $\phi^{(i)}(.)$ and $\phi^{(j)}(.)$
$\alpha_s^{(j)}$	Coefficient s for the j-th smooth component
$\phi_{(i)j}$	i-th row of $\Phi_{.j}$

Boosted additive mixed model notation

$X_{i(r)} = [X_i, \Phi_{i.r}]$	Designmatrix for the r-th component including fixed effects
K_r	Penalty matrix for the r-th component including fixed effects
$\hat{\delta}_r$	Weak learner for the r-th component including fixed effects in a boosting step
$\hat{\beta}^{(l)}, \hat{\alpha}^{(l)}, \delta^{(l)}, \eta^{(l)}$	Ensemble estimates in the l-th boosting step
$\eta_{i(r)}^{(l)}$	Predictor using the r-th component in boosting step l in cluster i
$M_r^{(l)}$	Projection matrix for residuals on component r of l-th boosting step to the weak learner $\hat{\delta}_r$, given the selection before
$H_r^{(l)}$	Local hat matrix for projection for residuals on component r of l-th boosting step to the predictor $\hat{\eta}_r^{(l)}$, given the selection before
$G_r^{(l)}$	Global hat matrix for projection for y on component r in the l-th boosting step to the predictor $\hat{\eta}_r^{(l)}$, given the selection before
j_l	Selected component in the l-th boosting step
$S_r^{(l)}$	Selection criterion in the l-th boosting step using the trace of $G_r^{(l)}$
$M^{(l)} := M_{j_l}^{(l)}$	Short notation, if j was selected in the l-th boosting step
$H^{(l)} := H_{j_l}^{(l)}$	Short notation, if j was selected in the l-th boosting step
$G^{(l)} := G_{j_l}^{(l)}$	Short notation, if j was selected in the l-th boosting step
k	Number of flexible splines
$R_i := [\Phi_{i1}\alpha_1, \ldots, \Phi_{ik}\alpha_k]$	Random design matrix for flexible splines
$R_i^{(l)} := [\Phi_{i1}\alpha_1^{(l)}, \ldots, \Phi_{ik}\alpha_k^{(l)}]$	Random design matrix for flexible splines in the l-th boosting step
	Design matrix for unspecified functions
$\tilde{Z}_i^{(l)} = [Z_i, R_i^{(l)}]$	Random design matrix for cluster i for parametric and smooth covariates
$X_{.(r)} = [X_{1(r)}^T, \ldots, X_{n(r)}^T]^T$	Design matrix for component r (complete dataset)
$\eta_{.(r)}$	Predictor with component r (complete dataset)

Preface

This thesis has been written during my activities as research assistant at the Department of Statistics of, Ludwigs-Maximilians-University, Munich. The financial support from *Sonderforschungsbereich 386* is greatfully acknowledged.

First I would like to express my thank to my doctoral advisor Gehard Tutz, who gave the important impulse and and advised this thesis. I would also like to thank all professors, all other colleagues, other research assistants and everybody, who was accessible for questions or problems.

I thank Prof. Brian Marx to agree to be the external referee.

In addition I want to thank Dr. Christian Heumann for being my second referee. The discussions around mixed models were important for the ingredients of this thesis. Especially Stefan Pilz was a big help as a personal mentor.

An useful aid was the Leibniz-Rechenzentrum in Munich, which allowed me to do my computations on the high-performance-cluster. The computations around the simulation studies did approximately take ten-thousand hours, where only the time on the cpu unit was measured. Around three-thousand datasets of different length were analyzed here. The number crunshers of the department did not deliver enough power to compute most of the settings which were checked in this thesis.

Special thanks to my parents who supported my interests in mathematics and to my wife Daniela, who spent patiently a lot of time at home with me on statistical papers, work and managed with our son *Vinzenz*.

Chapter 1

Introduction

1.1 Mixed Models and Boosting

The methodology of linear mixed model in combination with penalized splines has become popular in the past few years.

The idea of mixed models was originally developed in 1953 by Henderson (1953). He derived the mixed models to analyze longitudinal data by assuming a latent, unobserved structure in the data. The response was continuous and the structure was assumed to be linear. An example for longitudinal data may be patients with many repeated measures on each patient. The latent structure in this case may be on the individual level of the patient. So individuality is getting important in this context. For the statistical analysis, the observed covariates are considered to be conditionally independent given the patient and the patients are themselves are assumed to be independent. The latent structure may be only at the individual level of the patients (random intercept) or individual level and slope of these patient (random intercept and slope). A nice overview on mixed models is given by Verbeke & Molenberghs (2001). Another way of modeling longitudinal data with weak assumptions is using the generalized estimation equations (GEE), see, Liang & Zeger (1986). In this thesis, only mixed models are investigated.

The influence of covariates is often reflected insufficiently because the assumed parametrization for continuous covariates is usually very restrictive. For models with continuous covariates, a huge repertoire of nonparametric methods have been developed within the past few years. The effect of a continuous covariate on the response is specified by a smooth function. A smooth function is meant to be sufficient differentiable. Representatives of the nonparametric methods are kernel regression estimators, see, Gasser & Müller

(1984), Staniswalis (1989), for regression splines is Eubank (1988), Friedman & Silverman (1989), for local polynomial fitting is Hastie & Loader (1993), Fan & Gijbels (1996) and for smoothing splines is Silverman (1984), Green (1987). A nice overview on nonparametric methods may be found in Simonoff (1996).

The used representation for smooth effects in this thesis combines the different approaches for spline regression. When polynomial splines are used one has to specify the degree of its polynomial pieces as well as the decomposition of the range by a finite number of knots. The decomposition of polynomial splines can be expressed by a vector space. There exists a basis representation for every element of this space. That is why the approximated smooth effects can be parameterized. The regression spline can be reduced to a strict parametric structure, which is a great benefit of this approach.

The goodness of the approximation by polynomial splines is determined by the decomposition of the range. A large number of knots increase the fit of the data but for data with huge noise, the estimated curves are very wiggly. One way to control the variability of the estimated function is the adaptive selection of knots and positions, see Friedman (1991) and Stone, Hansen, Kooperberg & Truong (1997). An alternative approach is to use penalization techniques. In the latter case the penalty terms are focused on the basis coefficients of the polynomial spline representation. Two concepts for penalization have been established in recent years. One concept encompasses the truncated power series as suggested in Ruppert, Wand & Carroll (2003). In this case, one uses the ridge penalty. The other concept is maintained by Eilers & Marx (1996). They use the B-spline basis together with a penalization of neighboured basis coefficients which is called P-splines. Both concepts have the advantage that the estimation of parameters can be obtained by the maximizing of a penalized likelihood function.

The crucial part of a penalized likelihood is that the smoothing parameter λ, which controls the variability of the estimated functions, has to be optimized . One idea suggested by Eilers and Marx (Eilers & Marx (1996)) is to optimize the AIC criterion which measures the likelihood of the model given the fixed smoothing parameter λ. The likelihood is penalized by the effective degrees of freedom in the model, see Hastie & Tibshirani (1990). Another idea is to use the cross-validation criterion which is a computational burden in many data situation. Another driven criterion is the generalized cross validation criterion established by Craven & Wahba (1979). Recent investigations on this criterion are documented in Wood (2004). Another strategy to optimize this tuning parameter is based on mixed models.

The reason for the popularity of mixed models in the 90's is the comment of Terry Speed (Speed (1991)) on Robinson's article (Robinson (1991)) on BLUP equations. Terry Speed states that the maximization of a penalized likelihood is equivalent to the solutions of the

BLUP equations in a linear mixed model. These statements were picked up by Wand (2000), Parise, Wand, Ruppert & Ryan (2001), Ruppert, Wand & Carroll (2003) and Lin & Zhang (1999). So nowadays smoothing is often connected to penalized splines or it is seen as a suitable method to find reliable estimates for the smoothing parameter.

Boosting originates in the machine learning community where it has been proposed as a technique to improve classification procedures by combining estimates with reweighted observations. Recently it has been shown that boosting is also a way of fitting an additive expansion in basis functions when the single basis functions represent the results of one iteration of the boosting procedure. The procedure is quite similar to the method of gradient descent by the use of specific loss functions, see Breiman (1999) and Friedman, Hastie & Tibshirani (2000). Since it has been shown that reweighting corresponds to minimizing a loss function iteratively (Breiman (1999), Friedman (2001)), boosting has been extended to regression problems in a L_2-estimation framework by Bühlmann & Yu (2003). Tutz & Binder (2006) introduced the likelihood-based boosting concept for all kinds of link functions and distributions.

The aim of this thesis it to combine the mixed model methodology with boosting approaches. Especially the concept of componentwise boosting is introduced where in each iteration step, only one variable is allowed to be updated. This is a useful strategy if one tries to optimize a huge number of continuous variables. It is a very robust method in terms of algorithmic optimization. Part of the algorithmic structure is that one can do variable selection since among all covariates, only one is selected to be optimized which is a nice add-on in the boosting methodology.

Often the application of an additive mixed model is too restrictive because each cluster may have its own specific function. So one idea is to compute a joint smooth function of the continuous covariates and a random, cluster-specific smooth function as suggested by Ruppert, Wand & Carroll (2003), Wu & Liang (2004) or Verbyla, Cullis, Kenward & Welham (1999). If a joint additive structure with a random intercept is not sufficient to capture the variation of subjects, then one may extend the model by cluster specific modifications on the joint spline function, which is realized by a random effect. This kind of model is simply structured and needs only two additional parameters, the variance of the slope and the covariance between slope and intercept. It is therefore very parsimonious and allows simple interpretation. By using few additional parameters it has a distinct advantage over methods that allows subjects to have their own function, yielding as many functions as subjects (see for example Verbyla, Cullis, Kenward & Welham (1999) and Ruppert, Wand & Carroll (2003)).

An adequate formulation, investigation and interpretation of regression models needs an explicit consideration of the feature space of response variables. So in some areas the

assumption of a normal distribution, which is part of the classical regression models, has to be generalized to regression models for discrete responses. The statistical concept for these regression models was build by Nelder & Wedderburn (1972). They introduced the *Generalized Linear Models.* Many publications based on these models were published by Kay & Little (1986), Armstrong & Sloan (1989) in the medical research, Amemiya (1981), Maddala (1983) in economics and many other areas.

Heavy numerical problems arise if one tries to do inference in generalized linear models for longitudinal data. And the problems are not only restricted to the generalized linear mixed models. In the generalized estimation equations, the optimization is not a trivial thing, see Liang & Zeger (1986). These problems originate in the fact that the marginal distribution is not analytically accessible for generalized linear mixed models. Therefor complicate integrals have to be solved. In the mixed model one can do analytical integration by using some nice properties of gaussian random variables. But in generalized linear mixed models numerical integration has to be done. This can be either done by using the Laplace approximation (Breslow & Clayton (1993), Schall (1991), Wolfinger (1994)) using a normal approximation or either using integration points based methods like Gauss-Hermite quadrature (Hedeker & Gibbons (1996), Pinheiro & Bates (1995)), or Monte-Carlo integration (McCulloch (1997), McCulloch (1994), Booth & Hobert (1999)). One may use direct methods or the EM algorithm to get parameter estimates.

In the context of categorial the adequacy is not only restricted to the consideration of a discrete response structure. Properties of the variables are often reflected in a bad way using linear assumptions on the covariates, see Lin & Zhang (1999). Again, aim of this thesis is to extend generalized linear mixed models by nonparametric effects. The two remaining strategies for optimization is on the one side the approach discussed by Ruppert, Wand & Carroll (2003) and on the other side boosted generalized semi-parametric mixed models pictured in this thesis.

1.2 Guideline trough This Thesis

Chapter two gives a short introduction of linear mixed models. The different strategies of optimizing a linear mixed model as well as a robust variant of a linear mixed model are proposed.

The semi-parametric mixed models are part of the third chapter. The nonparametric models are sketched briefly as well as how nonparametric approaches are handled. It is mentioned that which of the problems arise if nonparametric modeling is used.

In the fourth chapter, the semi-parametric mixed models are extended to the flexible mixed models where all clusters have a common development of the covariate and each cluster has its distinct modifications in the sense that the effect of a covariate are strengthened or attenuated individually.

The generalized linear mixed models are the topic of the fifth chapter. An overview of the most popular methods are given here since there is no canonic way to estimate a generalized linear mixed model.

The sixth chapter deals with generalized semi-parametric mixed models. The Laplacian approximation is used to implement the boosting idea into the generalized linear mixed model framework. In simulation studies the results are compared to the optimized model based on the mixed model approach for additive models (see Ruppert, Wand & Carroll (2003)).

A short summary on the results, the given problems as well as an outlook on further development and questions are given which have been accumulated in the course of this thesis in the last.

Chapter 2

Linear Mixed Models

2.1 Motivation: CD4 Data

The data was collected within the Multicenter AIDS Cohort Study (MACS), which followed nearly 5000 gay or bisexual men from Baltimore, Pittsburgh, Chicago and Los Angeles since 1984 (see Kaslow, Ostrow, Detels, Phair, Polk & Rinaldo (1987), Zeger & Diggle (1994)). The study includes 1809 men who were infected with HIV at study entry and another 371 men who were seronegative at entry and seroconverted during the follow-up. In the study 369 seroconverters ($n = 369$) with 2376 measurements in total ($N = 2376$) were used and two subjects were dropped since covariate information was not available. The interesting response variable is the number or percent of CD4 cells (CD4) by which progression of disease may be assessed. Covariates include years since seroconversion (time), packs of cigarettes a day (cigarettes), recreational drug use (drugs) with expression yes or no, number of sexual partners (partners), age (age) and a mental illness score (cesd).

In this study, we have a repeated measurement design because every seroconverter has several measurement of covariates at different time points. For the i-th seroconverters, T_i repeated measurement were observed. For example the first seroconverter in the dataset has three repeated measurements, so in this case is $T_i = 3$. The described observations for the i-th seroconverter at repeated measurement t can then be addressed by $CD4_{it}$ for the response and for the corresponding covariates age_{it}, $partners_{it}$, $drugs_{it}$, $cesd_{it}$ and $time_{it}$.

If one has only one observation on each seroconverter, then $T_i = 1$ for all seroconverters ($i \in 1, \ldots, n$). So one can use standard cross sectional methods, because the measurement error for each seroconverter can be assumed to be independent. In the case of repeated

measurements ($T_i \neq 1$) one has clusterd observations with error term ($\epsilon_{i1}, \ldots, \epsilon_{iT_i}$) for the i-th person.

One approach to modeling data of this type is based on mixed models. Here the observations ($CD4_{it}, t = 1, \ldots, T_i$) for the i-th seroconverter are assumed to be conditional independent. In other words, given the level of the i-th seroconverter, the errors for repeated measurements of this person may assumed to be independent. The unknown level for the i-th seroconverter is expressed in mixed models by the so called random intercept b_i. A common assumption on random intercepts is that they are Gausssian distributed with $b_i \sim N(0, \sigma_b^2)$. σ_b^2 is the random intercept variance.

A mixed model with linear parameters age, partners, drugs, time and cesd the form is given by

$$CD4_{it} = \beta_0 + \beta_1 age_{it} + \beta_2 drugs_{it} + \beta_3 time_{it} + b_i + \epsilon_{it}$$

for $i = 1, \ldots, n$ and $t = 1, \ldots, T_i$, where ϵ_{it} is the error term. This can also be rewritten in vector notation for the i-th seroconverter as

$$CD4_i = \beta_0 + \beta_1 age_i + \beta_2 partners_i + \beta_3 drugs_i + \beta_4 time_i + 1_{T_i} b_i + \epsilon_i$$

where $CD4_i^T = (CD4_{i1}, \ldots, CD4_{iT_i})$, $age_i^T = (age_{i1}, \ldots, age_{iT_i})$, $drugs_i^T = (drugs_{i1}, \ldots, drugs_{iT_i})$, $time_i^T = (time_{i1}, \ldots, time_{iT_i})$ and $\epsilon_i^T = (\epsilon_{i1}, \ldots, \epsilon_{iT_i})$. 1_{T_i} is a vector of the length T_i with ones. The assumption on the model may be

$$\begin{pmatrix} \epsilon_i \\ b_i \end{pmatrix} = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 I_{T_i} & 0 \\ 0 & \sigma_b^2 \end{pmatrix} \right),$$

where the errors and random intercepts of the i-th person are not correlated with those of the j-th person ($j \neq i$). If the vector notation without an index is preferred, one can also write

$$CD4 = \beta_0 + \beta_1 age + \beta_2 partners + \beta_3 drugs + \beta_4 time + \mathbb{Z}b + \epsilon,$$

where $CD4^T = (CD4_1^T, \ldots, CD4_n^T)$, $age^T = (age_1^T, \ldots, age_n^T)$, $drugs^T = (drugs_1^T, \ldots, drugs_n^T)$, $time^T = (time_1^T, \ldots, time_n)^T$, $\epsilon^T = (\epsilon_1^T, \ldots, \epsilon_n^T)$ and $b^T = (b_1, \ldots, b_n)$ The matrix $\mathbb{Z}$ is then a blockdiagonal matrix of $1_{T_1}, \ldots, 1_{T_n}$. The assumption on the mixed model can than reduced to

$$\begin{pmatrix} \epsilon \\ b \end{pmatrix} = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 I_N & 0 \\ 0 & \sigma_b^2 I_n \end{pmatrix} \right).$$

Since one may use a short notation without addressing the variable names $CD4$, age, $drugs$, $cesd$, $time$, then one set generally response to $y_{it} := CD4_{it}$. The

variables that are responsible for the fixed effects are put into the vector $x_{it}^T := (1, age_{it}, drugs_{it}, time_{it})$.. The variables associated with the random effect are stacked in blockdiagonal entries in the matrix $\mathbb{Z}$. The short term notation is with $X_i^T = (x_{i1}, \ldots, x_{iT})$, $X^T = (X_1^T, \ldots, X_n^T)$, $y_i^T = (y_{i1}, \ldots, y_{iT})$, $y^T = (y_1^T, \ldots, y_n^T)$ and $\beta^T = (\beta_0, \ldots, \beta_4)$

$$y = X\beta + \mathbb{Z}b + \epsilon.$$

For example, one might extend the mixed model to a mixed model with random slopes as

$$CD4_{it} = \beta_0 + \beta_1 age_{it} + \beta_2 drugs_{it} + \beta_3 time_{it} + b_i^{(1)} + time_{it} b_i^{(2)} + \epsilon_{it},$$

allowing a random variation in the slope for the linear time effect. Using the vector $z_{it}^T := (1, time_{it})$, $Z_i^T := (z_{i1}, \ldots, z_{iT_i})$, $\mathbb{Z} = bdiag(Z_1, \ldots, Z_n)$ and $b_i^T = (b_i^{(1)}, b_i^{(2)})$, where $b_i^{(1)}$ is the random intercept and $b_i^{(2)}$ is the random slope, one can write

$$y = X\beta + \mathbb{Z}b + \epsilon.$$

The assumption on the random effects of the i-th seroconverter may be $b_i \sim N(0, Q)$, where Q is a 2×2-covariance matrix. This covariance matrix may be assumed to be the same for all persons. The random effects of the persons are not correlated within each other. This may be denoted by

$$\begin{pmatrix} \epsilon \\ b \end{pmatrix} = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 I_N & 0 \\ 0 & \mathbb{Q} \end{pmatrix} \right)$$

with $\mathbb{Q}$ being the n-times blockdiagonal matrix of Q.

2.2 The Model

The linear mixed model for longitudinal data was introduced by Henderson (1953). Let the data be given by $(y_{it}, x_{it}), i = 1, \ldots, n, t = 1, \ldots, T_i$ with y_{it} connected to observation t in cluster i and x_{it} denoting a vector of covariates which may vary across the observations within one cluster. $N = \sum_{i=1}^n T_i$ is the number of observations in total.

For the simplicity of presentation, let the number of observations within one cluster T do not depend on the cluster. Let x_{it} and z_{it} are design vectors composed from given covariates. We set $X_i^T = (x_{i1}, \ldots, x_{iT}), Z_i^T = (z_{i1}, \ldots, z_{iT}), y_i^T = (y_{i1}, \ldots, y_{iT}), X^T = (X_1^T, \ldots, X_n^T), y^T = (y_1^T, \ldots, y_n^T), , \mathbb{Z} =$ bdiag$(Z_1, \ldots, Z_n)$. The basic idea of the random effect models it to model the joint distribution of the observed covariate y and an unobservable random effect b.

The assumption on the distribution of y given the random effect b is

$$y|b \sim N(X\beta + \mathbb{Z}b, \sigma_\epsilon^2 I_N).$$

Here the (conditional) distribution $f(y|b)$ follows a Gaussian normal distribution with mean $X\beta + \mathbb{Z}b$ and covariance $\sigma_\epsilon^2 I_N$.

The assumption on the random effects b and the error component ϵ is

$$\begin{pmatrix} \epsilon \\ b \end{pmatrix} = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 I_N & 0 \\ 0 & \mathbb{Q}(\rho) \end{pmatrix} \right).$$

Here the distribution $p(b;\rho)$ of the random effects is Gaussian distribution with mean zero and covariance matrix $\mathbb{Q}(\rho)$. In this case the structural parameter ρ specifies the covariance matrix $\mathbb{Q}(\rho) = \text{cov}(b)$. Since the mean of $p(b;\rho)$ is assumed to be zero, so $p(b)$ is fully specified up to the unknown covariance matrix $\mathbb{Q}(\rho)$. An overview on parameterized covariance matrices and its derivatives is given in the appendix.

If one assumes that the observations within and between clusters given the random effects are independent, the joint density of y, $f(y \mid b)$=$\prod_{i=1}^{n} f(y_i \mid b_i)$ with $f(y_i \mid b_i) = \prod_{t=1}^{T_i} f(y_{it} \mid b_i)$ reduces to product densities, which can be handled easily. Here $f(y_i \mid b_i)$ and $f(y_{it} \mid b_i)$ are also Gaussian densities with mean $X_i\beta + Z_i b_i$ and $x_{it}^T\beta + z_{it}^T b_i$ and covariance $\sigma_\epsilon^2 I_{T_i}$ and σ_ϵ^2.

One reason for the popularity of this assumption is that it is easy to understand and its usefulness in the context of longitudinal data. More complex structures are possible, i.e., clusters are correlated or the observations within the clusters are correlated in a distinct way. Since correlation problems can easy be expressed in the Gaussian framework we start with the linear mixed model specified by Gaussian mixing densities and conditional outcomes that are normally distributed.

One gets the marginal densities as

$$f(y) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \int f(y_i|b_i) p(b_i;\rho) db_i,$$

where $p(b_i;\rho)$ is the density of $N(0, Q(\rho))$. In this case $\mathbb{Q}(\rho) = \text{bdiag}(Q(\rho), \ldots, Q(\rho))$. In other word, each cluster has the same structure $Q(\rho)$. So only $Q(\rho)$ has to be estimated to get $\mathbb{Q}(\rho)$.

Let θ be the vector of variance parameters $\theta^T = (\sigma, \rho)^T$. The result of this consideration is the marginal form of a Gaussian random effects model

$$y_i \sim N(X_i\beta, V_i(\theta)),$$

where

$$V_i(\theta) = \sigma_\epsilon^2 I_T + Z_i Q(\rho) Z_i^T.$$

In matrix notation

$$y \sim N(X\beta, \mathbb{V}(\theta))$$

with $\mathbb{V}(\theta) = \text{bdiag}(V_1(\theta), \ldots, V_n(\theta))$. The joint density of y and b is reduced to a marginal density by integrating out the random effect b.

As already mentioned, the advantage of this restriction is that this parametrization is easy to handle regarding numerical aspects. The operations on huge covariance matrices and design matrices can be reduced to operations on block matrices of the block diagonal matrix. This is a well conditioned problem in the numerical sense. Here, $Q(\rho)$ is the covariance matrix of the random effects within one cluster which is assumed to be the same in each cluster.

Gaussian random effect models have an advantage, because on the basis of

$$f(y) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \int f(y_i|b_i) p(b_i; \rho) db_i,$$

one can easily switch between the marginal and conditional views. The n integrals can be solved analytically by using the marginal distribution. For arbitrary mixtures of conditional and random distribution, it is not possible, in general. The log-likelihood for β and θ is given by

$$l(\beta, \theta|y) = \sum_{i=1}^{n} \log(f(y_i)) = -\frac{1}{2} \sum_{i=1}^{n} \log(|V_i(\theta)|) + \sum_{i=1}^{n} (y_i - X_i\beta)^T V_i(\theta)^{-1} (y_i - X_i\beta).$$

So the estimator $\hat{\beta}$ is obtained by solving the following equation, which is derived from the log-likelihood by differentiating with respect to β

$$\left(\sum_{i=1}^{n} (X_i^T V_i^{-1} X_i) \right) \beta = \left(\sum_{i=1}^{n} X_i^T V_i^{-1} y_i \right). \tag{2.1}$$

As shown in Harville (1976) and described in Harville (1977) b_i can be estimated by

$$\hat{b}_i = Q(\rho) Z_i^T V_i(\theta)^{-1} (y_i - X_i\hat{\beta}). \tag{2.2}$$

Harville (1976) shows, that the solutions of equation 2.1 and 2.2 are equivalent to the solution of the BLUP-equation

$$\begin{bmatrix} X^T W X & X W \mathbb{Z} \\ \mathbb{Z}^T W X & \mathbb{Z}^T W \mathbb{Z} + \mathbb{Q}(\rho)^{-1} \end{bmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{bmatrix} X^T W y \\ \mathbb{Z}^T W y \end{bmatrix} \tag{2.3}$$

with $W = \frac{1}{\sigma_\epsilon^2} I$.

The estimator $\hat{b}$ is called BLUP (best linear unbiased predictor) which minimizes $\mathbb{E}((\hat{b} - b)^T(\hat{b} - b))$, see Harville (1976). Additional effort is necessary if $Q(\rho)$ or the structural parameters ρ are not known. A usual way to solve these problems are often based on the restricted log-likelihood. Therefore profile likelihood concepts are used, which alternatingly plug in the estimate for the variance components and the estimate for the fixed effects.

A detailed introduction in linear mixed models is given by Robinson (1991), McCulloch & Searle (2001). Especially on longitudinal mixed models, information can be found in Verbeke & Molenberghs (2001) and Harville (1976) and Harville (1977).

2.3 The Restricted Log-Likelihood

The restricted log-likelihood is based on Patterson & Thompson (1971) . It was reviewed by Harville (1974), Harville (1977) and by Verbeke & Molenberghs (2001). It is given by

$$l_r(\beta, \theta) = -\frac{1}{2}\sum_{i=1}^{n} \log(|V_i(\theta)|) + \sum_{i=1}^{n}(y_i - X_i\beta)^T V_i(\theta)^{-1}(y_i - X_i\beta) - \frac{1}{2}\sum_{i=1}^{n} \log(|X_i^T V_i(\theta) X_i|).$$

The restricted log-likelihood differs from the log-likelihood by an additional component, since

$$l_r(\beta, \theta) = l(\beta, \theta) - \frac{1}{2}\sum_{i=1}^{n} \log(|X_i^T V_i(\theta) X_i|). \tag{2.4}$$

Differentiating $l_r(\beta, \theta)$ with respect to β results in the same equation as differentiating $l(\beta, \theta)$ with respect to β. An important question is now why $l_r(\beta, \theta)$ should be used for the further computation. By plugging in the estimates alternatingly, degrees of freedom for the estimate of the variance components θ are lost. The loss of degrees is compensated by the additional component in the restricted log likelihood. Details can be found in Harville (1977)

Since $l_r(\beta, \theta)$ is nonlinear in θ , $l_r(\beta, \theta)$ has to be maximized by a Fisher-Scoring algorithm.

The estimation of the variance components is based on the profile log-likelihood that is obtained by plugging in the estimates $\hat{\beta}$ in the marginal log-likelihood formula 2.4.

Differentiation with respect to $\theta^T = (\sigma_\varepsilon, \rho^T) = (\theta_1, \ldots, \theta_d)$ yields

$$s(\hat{\beta}, \theta) = \frac{\partial l(\hat{\beta}, \theta)}{\theta} = (s(\hat{\beta}, \theta)_i)_{i=1,\ldots,d}$$

and

$$F(\hat{\beta}, \theta) = -E(\frac{\partial^2 l(\hat{\beta}, \theta)}{\partial\theta\partial\theta^T}) = (F(\hat{\beta}, \theta)_{i,j})_{i,j=1,\ldots,d}$$

with

$$s(\hat{\beta}, \theta)_i = \frac{\partial l(\hat{\beta}, \theta)}{\partial\theta_i} = \quad -\tfrac{1}{2}\sum_{k=1}^n \text{trace}\left((P_k(\theta))^{-1}\frac{\partial V_k(\theta)}{\partial\theta_i}\right)$$

$$+\tfrac{1}{2}\sum_{k=1}^n (y_k - \eta_k)^T V_k(\theta)^{-1}\frac{\partial V_k(\theta)}{\partial\theta_i}V_k(\theta)^{-1}(y_k - \eta_k).$$

P_k is defined in Harville (1977).

$$P_k(\theta) = V_k(\theta)^{-1} - V_k(\theta)^{-1}X_k\left(\sum_{k=1}^n X_k^T V_k(\theta)^{-1}X_k\right)^{-1}X_k^T V_k(\theta)^{-1}$$

and

$$F(\hat{\beta}, \theta)_{i,j} = \frac{1}{2}\sum_{k=1}^n \text{trace}\left((P_k(\theta))^{-1}\frac{\partial V_k(\theta)}{\partial\theta_i}(P_k(\theta))^{-1}\frac{\partial V_k(\theta)}{\partial\theta_j}\right),$$

where

$$\frac{\partial V_k(\theta)}{\partial\theta_i} = \begin{cases} 2\sigma_\epsilon I_{T_k} & \text{if } i = 1 \\ Z_k\frac{\partial Q(\rho)}{\partial\theta_j}Z_k^T & \text{if } j = i, i \neq 1. \end{cases}$$

The estimator $\hat{\theta}$ can now be obtained by running a common Fisher scoring algorithm with

$$\hat{\theta}^{(s+1)} = \hat{\theta}^{(s)} + F(\hat{\beta}, \theta^{(s)})^{-1}s(\hat{\beta}, \hat{\theta}^{(s)}).$$

where s denotes the iteration index of the Fisher scoring algorithm. If Fisher scoring has converged, the resulting $\hat{\theta}$ represents the estimates of variances for the considered step.

2.3.1 The Maximum Likelihood Method

In special cases, it is necessary to use the ML instead of the REML, because the Fisher-Scoring in the REML-methods may be affected by numerical problems. Especially when

there are many covariates, which have no effect on the response, the REML estimator then do not converge.

On the other hand it is criticized that the maximum likelihood estimator for σ^2 does not take into account the loss of degrees of freedom when plugging in $\hat{\beta}$.

The estimation of the variance components is based on the profile log-likelihood that is obtained by plugging in the estimates $\hat{\beta}$ in the marginal log-likelihood

$$l(\hat{\beta};\theta) \quad = -\tfrac{1}{2}\sum_{i=1}^{n}\log(|V_i(\theta)|) + \sum_{i=1}^{n}(y_i-\hat{\eta})^T V_i(\theta)^{-1}(y_i-\hat{\eta}).$$

Differentiation with respect to $\theta^T = (\sigma_\varepsilon, \rho^T) = (\theta_1, \ldots, \theta_d)$ yields

$$s(\hat{\beta},\theta) = \frac{\partial l(\hat{\beta},\theta)}{\partial\theta} = (s(\hat{\beta},\theta)_i)_{i=1,\ldots,d}$$

and

$$F(\hat{\beta},\theta) = -E(\frac{\partial^2 l(\hat{\beta},\theta)}{\partial\theta\partial\theta^T}) = (F(\hat{\beta},\theta)_{i,j})_{i,j=1,\ldots,d}$$

with

$$s(\hat{\beta},\theta)_i = \tfrac{\partial l(\hat{\beta},\theta)}{\theta_i} = \quad -\tfrac{1}{2}\sum_{k=1}^{n}\text{trace}\left((V_k(\theta))^{-1}\,\tfrac{\partial V_k(\theta)}{\theta_i}\right)$$

$$+\tfrac{1}{2}\sum_{k=1}^{n}(y_k-\eta_k)^T V_k(\theta)^{-1}\tfrac{\partial V_k(\theta)}{\theta_i}V_k(\theta)^{-1}(y_k-\eta_k)$$

and

$$F(\hat{\beta},\theta)_{i,j} = \frac{1}{2}\sum_{k=1}^{n}\text{trace}\left((V_k(\theta))^{-1}\,\frac{\partial V_k(\theta)}{\partial\theta_i}\,(V_k(\theta))^{-1}\,\frac{\partial V_k(\theta)}{\partial\theta_j}\right),$$

where

$$\frac{\partial V_k(\theta)}{\partial\theta_i} = \begin{cases} 2\sigma_\epsilon I_{T_k} & \text{if } i=1 \\ Z_k\frac{\partial Q(\rho)}{\partial\theta_j}Z_k^T & \text{if } j=i, i\neq 1. \end{cases}$$

2.4 Estimation with Iteratively Weighted Least Squares

The estimation algorithm can be described as following:

Compute good start values $\hat{\beta}_0$ and $\hat{\theta}_0$. The value of $\hat{\beta}_0$ can be the estimator from a linear model. The elements of θ_0 are set to be small values, i.e. 0.1.

1. set $k = 0$

2. compute $\hat{\beta}^{(k+1)}$ by solving the equation $l(\beta, \hat{\theta}^{(k)})$ above with plugged in $\hat{\theta}^{(k)}$

3. compute $\hat{\theta}^{(k+1)}$ in $l(\hat{\beta}, \theta)$ by running a Fisher scoring algorithm with plugged in $\hat{\beta}^{(k+1)}$.

4. stop, if all stopping criteria are reached, else start in 1 with $k = k + 1$.

This algorithm corresponds to the iteratively weighted least squares algorithms. Alternatively, the variance parameters can be obtained by using the EM-algorithm originally described in Laird & Ware (1982). Later, Lindstrom & Bates (1990) suggested that the Newton-Raphson-algorithm should be preferred over the EM-algorithm.

2.5 Estimation with EM-Algorithm - The Laird-Ware Method

The idea of this maximization method is based on Laird & Ware (1982). Indirect maximation of the marginal density starts from the joint log-density of the observed data $y = (y_1, \ldots, y_n)$ and the unobservable effects $\delta = (\beta, b_1, \ldots, b_n)$. The joint log-likelihood is

$$\log f(y, \delta|\theta) = \log f(y|\delta; \sigma_\epsilon^2) + \log p(b_1, \ldots, b_n; \rho)$$

From the model assumptions one obtains, up to constants,

$$\begin{aligned} S_1(\sigma_\epsilon^2) \quad &\propto -\tfrac{1}{2} N \log \sigma_\epsilon^2 - \tfrac{1}{2\sigma_\epsilon^2} \textstyle\sum_{i=1}^n \epsilon_i^T \epsilon_i, \\ S_2(Q(\rho)) \quad &\propto -\tfrac{n}{2} \log \det(Q(\rho)) - \tfrac{1}{2} \textstyle\sum_{i=1}^n b_i^T Q(\rho)^{-1} b_i \\ &= -\tfrac{n}{2} \log \det(Q(\rho)) - \tfrac{1}{2} \textstyle\sum_{i=1}^n tr(Q(\rho)^{-1} b_i^T b_i) \end{aligned}$$

Next we start in the EM-framework with building the conditional expections with respect to $\theta^{(p)}$

$$M(\theta|\theta^p) = \mathbb{E}\{S_1(\sigma_\epsilon^2)|y; \theta^{(p)}\} + \mathbb{E}\{S_2(Q(\rho))|y; \theta^{(p)}\} \tag{2.5}$$

which is called the E-step. In detail are the E-equations

$$\begin{aligned}
\mathbb{E}\{S_1(\sigma_\epsilon^2)|y;\theta^{(p)}\} &= -\tfrac{1}{2}\textstyle\sum_{i=1}^n T_i \log(\sigma_\epsilon^2) \\
&\quad -\tfrac{1}{2\sigma_\epsilon^2}\textstyle\sum_{i=1}^n [(\epsilon_i^{(p)})^T\epsilon_i^{(p)} + tr\text{cov}(\epsilon_i|y_i;\theta^{(p)}], \\
\mathbb{E}\{S_2(Q(\rho))|y;\theta^{(p)}\} &= -\tfrac{n}{2}\log\det(Q(\rho^{(p)})) \\
&\quad -\tfrac{1}{2}\textstyle\sum_{i=1}^n [tr[Q(\rho^{(p)})^{-1}b_i^{(p)}(b_i^{(p)})^T] + \text{cov}(b_i|y_i;\theta^{(p)})
\end{aligned} \tag{2.6}$$

with current residuals $\epsilon_i^{(p)} = y_i - X_i\hat{\beta}^{(p)} - Z_i\hat{b}_i^{(p)}$. Differentiation of (2.6) with respect to σ_ϵ and $Q(\rho)$ yields the M-equations (Equations that maximize the E-equations)

$$\begin{aligned}
\sigma_\epsilon^{2(p+1)} &= -\tfrac{1}{N}\textstyle\sum_{i=1}^n (\epsilon_i^{(p)})^T\epsilon_i^{(p)} + tr\text{cov}(\epsilon_i|y_i;\theta^{(p)}] \\
Q(\rho^{(p+1)}) &= \tfrac{1}{n}\textstyle\sum_{i=1}^n [b_i^{(p)}(b_i^{(p)})^T + \text{cov}(b_i|y_i;\theta^{(p)})].
\end{aligned} \tag{2.7}$$

We use projection matrices according Laird, Lange & Stram (1987)

$$P(\theta^{(p)}) = \left(V_i(\theta^{(p)}\right)^{-1} \tag{2.8}$$

in ML-Estimation and

$$P(\theta^{(p)}) = \left(V_i(\theta^{(p)}\right)^{-1} - \left(V_i(\theta^{(p)}\right)^{-1} X_i(X_i^T\left(V_i(\theta^{(p)}\right)^{-1}X_i)^{-1}X_i^T\left(V_i(\theta^{(p)}\right)^{-1} \tag{2.9}$$

in REML-Estimation with $V(\rho^p) = \sigma_\epsilon^{2(p+1)} + Z_iQ(\rho^{(p+1)})Z_i^T$.

Therefore we can denote with with projection matrix (2.8) or (2.9) as described in Laird, Lange & Stram (1987)

$$\begin{aligned}
\sigma_\epsilon^{2(p+1)} &= -\tfrac{1}{N}\textstyle\sum_{i=1}^n (\epsilon_i^{(p)})^T\epsilon_i^{(p)} + \sigma^{2(p)}tr(I - \sigma^{2(p)}P_i(\theta^{(p)})] \\
Q(\rho^{(p+1)}) &= \tfrac{1}{n}\textstyle\sum_{i=1}^n [b_i^{(p)}(b_i^{(p)})^T + Q(\rho^{(p)})(I - Z_i^TP_i(\theta^{(p)})Z_i)Q(\rho^{(p)})]
\end{aligned} \tag{2.10}$$

The estimates $\hat{\beta}^{(p)}$ and $\hat{b}_i^{(p)}$ are obtained in the usual way

$$\begin{aligned}
\hat{\beta}^{(p)} &= \left(\textstyle\sum_{i=1}^n X_i(V_i(\theta^{(p)}))^{-1}X_i\right)^{-1}\sum_{i=1}^n X_i^T(V_i(\theta^{(p)}))^{-1}y_i \\
\hat{b}_i^{(p)} &= Q(\rho^{(p)})Z_i^T(V_i(\theta^{(p)}))^{-1}(y - X\hat{\beta}^{(p)}).
\end{aligned} \tag{2.11}$$

The EM-Algorithm is now

1. Calculate start value $\theta^{(0)} = ({\sigma_\epsilon}^{(0)}, \rho^{(0)})$.

2. For $p = 1, 2, \ldots$ compute $\hat{\delta}^{(p)} = (\hat{\delta}^{(p)}, \hat{b}_1^{(p)}, \ldots, \hat{b}_n^{(p)})$ with variance-covariance components replaced by their current estimates $\theta^{(p)} = ({\sigma_\epsilon}^{(p)}, \rho^{(p)})$, together with current residuals $\epsilon_i^{(p)} = y_i - X_i\hat{\beta}^{(p)} - Z_i\hat{b}_i^{(p)}$ and posterior covariance matrices $\text{cov}(\epsilon_i|y_i; \theta^{(p)})$ and $\text{cov}(b_i|y_i; \theta^{(p)})$. This step may be seen as the E-step.

3. Do the M-step to compute updates with 2.10.

4. If the condition
$$\frac{||\theta^{(p+1)} - \theta^{(p)}||}{||\theta^{(p)}||}$$
is accomplished, convergence of the EM-algorithm is achieved. If not start in step 2 with $\theta^{(p+1)}$ as update for $\theta^{(p)}$.

More information on the estimation of mixed models via EM-algorithm can be found in Laird & Ware (1982). Later, Lindstrom & Bates (1988) compare the Newton-Raphson method to EM-estimation. Especially fast algorithm reparametrization can be found here. Laird, Lange & Stram (1987) gave detailed information on EM-estimation and algorithmic acceleration. Alternatively, the gradient algorithm as described in Lange (1995), can also be used which is closely related to the EM-Algorithm.

2.6 Robust Linear Mixed Models

The marginal distribution of a linear mixed models is

$$y_i \sim N(X_i\beta, V_i(\theta)). \tag{2.12}$$

This assumption on the distribution is now replaced by the robust variant as suggested in Lange, Roderick, Little & Taylor (1989)

$$y_i \sim t_T(X_i\beta, \Psi_i(\theta), \nu), \tag{2.13}$$

where $t_k(\mu, \Psi, \nu)$ denotes the k-variate t-distribution as given in Cornish (1954), Ψ is the scaling matrix which has the function of Σ in the mixed model concept. The additional parameter ν must be positive and can be noninteger. It has the funtion of a robustness

parameter, since it downweights outlying cases. We set now $\mu = X\beta$. The marginal density is

$$f(y|\mu,\Psi,\nu) = \frac{|\Psi|^{-1/2}\Gamma((\nu+k)/2)}{(\Gamma(1/2)))^k\Gamma(\nu/2)\nu^{k/2}} \left(1+\frac{(y-\mu)^T\Psi^{-1}(y-\mu)}{\nu}\right)^{-(\nu+k)/2}. \tag{2.14}$$

Important properties for $\nu \geq 2$ are:

- $y \sim t_k(\mu,\Psi,\nu)$
- $\mathbb{E}(y) = \mu$ and $\text{Cov}(y) = \Sigma = \frac{\nu\Psi}{\nu-2}$ for $(\nu > 2)$
- $b|y \sim \frac{\chi^2_{\nu+k}}{\nu+\delta^2}$ with $\delta^2 = (y-\mu)^T\Psi^{-1}(y-\mu)$,
  $\chi^2_k(.)$ denotes the Chi-Square distribution with k degrees of freedom
- $\frac{\delta^2}{k} \sim F_{k,\nu}$

According to these properties, the model can be derived from multivariate normal-distribution with scaling variable b_i

$$y_i|b_i \sim N_T(\mu_i, \frac{\Psi(\theta)}{b_i}) \text{ where } b_i \sim \frac{\chi^2_T}{T}. \tag{2.15}$$

The log-likelihood for model (2.13) ignoring constants is

$$\begin{aligned} l(\beta,\rho,\nu) &\approx \textstyle\sum_{i=1}^n l_i(\beta,\rho,\nu) \\ &\text{with} \\ l_i(\beta,\rho,\nu) &= \tfrac{1}{2}\log|\Psi_i(\rho)| - \tfrac{1}{2}(\nu+T)\log\left(1+\tfrac{\delta_i^2(\beta,\rho)}{\nu}\right) \\ &\quad -\tfrac{1}{2}T\log(\nu) + \log\left[\Gamma\left(\tfrac{\nu+T}{2}\right)\right] - \log\Gamma(\tfrac{\nu}{2}). \end{aligned} \tag{2.16}$$

The likelihood-equations regarding β are closely related to the likelihood-equations of a linear mixed model. Setting the first derivative of (2.16) regarding β zero yields

$$\sum_{i=1}^n w_i X_i^T \Psi(\rho)^{-1}(y_i-\mu_i) = 0 \tag{2.17}$$

with the weight $w_i = \frac{\nu+T}{\nu+\delta_i^2}$.

The log-likelihood (2.16) can be maximized via a Fisher-Scoring-Algorithm. First we collect all necessary parameters in $\gamma = (\beta, \rho, \nu)$.

One has to rewrite (2.14) into

$$f(y|\gamma) = |\Psi(\rho)|^{-1/2} g((y-\mu)^T \Psi(\rho)^{-1}(y-\mu), \nu) \tag{2.18}$$

with

$$g(s,\nu) = \frac{\Gamma((\nu+k)/2)}{\Gamma(1/2)^k \Gamma(\nu/2)\nu^{k/2}} \left(1 + \frac{s}{\nu}\right)^{-(\nu+k)/2}. \tag{2.19}$$

The first derivatives are

$$\begin{aligned}
&\tfrac{l(\gamma)}{\partial\beta} = -2\tfrac{\partial g(\sigma^2,\nu)}{\partial\sigma^2}\tfrac{1}{g(\sigma^2,\nu)}\tfrac{\partial\mu}{\beta}\Psi(\rho)^{-1}(y-\mu),\\
&\tfrac{l(\gamma)}{\partial\rho_i} = -\tfrac{1}{2}\text{tr}\left(\Psi(\rho)^{-1}\tfrac{\Psi(\rho)}{\partial\rho_i}\right) - \tfrac{\partial g(\sigma^2,\nu)}{\partial\sigma^2}\tfrac{1}{g(\sigma^2,\nu)}(y-\mu)^T\Psi(\rho)^{-1})\tfrac{\partial\Psi(\rho)}{\partial\rho_i}\Psi(\rho)^{-1}(y-\mu),\\
&\tfrac{l(\gamma)}{\partial\nu} = \tfrac{\partial g(\sigma^2,\nu)}{\partial\nu}\tfrac{1}{g(\sigma^2,\nu)}.
\end{aligned} \tag{2.20}$$

So one can write

$$s(\gamma) = \begin{bmatrix} \frac{l(\gamma)}{\partial\beta} \\ \frac{l(\gamma)}{\partial\rho} \\ \frac{l(\gamma)}{\partial\nu} \end{bmatrix} \tag{2.21}$$

with $\frac{l(\gamma)}{\partial\rho}^T = (\frac{l(\gamma)}{\partial\rho_1}, \ldots, \frac{l(\gamma)}{\partial\rho_d})$.

The elements of the expected Fisher-matrix are

$$
\begin{aligned}
F_{\beta\beta} &= -\mathbb{E}\left(\tfrac{l(\gamma)}{\partial\beta^T\partial\beta}\right) &&= \tfrac{\nu+k}{\nu+k+2}\tfrac{\partial\mu^T}{\partial\beta^T}\Psi(\rho)^{-1}\tfrac{\partial\mu}{\partial\beta}, \\
F_{\rho_i\rho_j} &= -\mathbb{E}\left(\tfrac{l(\gamma)}{\partial\rho_i\partial\rho_j}\right) &&= \tfrac{\nu+k}{\nu+k+2}\tfrac{1}{2}\mathrm{tr}\left(\Psi(\rho)^{-1}\tfrac{\Psi(\rho)}{\partial\rho_i}\Psi(\rho)^{-1}\tfrac{\Psi(\rho)}{\partial\rho_j}\right), \\
& && -\tfrac{1}{2(\nu+k+2)}\left(\Psi(\rho)^{-1}\tfrac{\Psi(\rho)}{\partial\rho_i}\right)\left(\Psi(\rho)^{-1}\tfrac{\Psi(\rho)}{\partial\rho_j}\right), \\
F_{\rho_i\nu} &= -\mathbb{E}\left(\tfrac{l(\gamma)}{\partial\rho_i\partial\nu}\right) &&= \tfrac{1}{(\nu+k+2)(\nu+k)}\mathrm{tr}\left(\Psi(\rho)^{-1}\tfrac{\Psi(\rho)}{\partial\rho_i}\right), \\
F_{\nu\nu} &= -\mathbb{E}\left(\tfrac{l(\gamma)}{\partial\nu\partial\nu}\right) &&= -\tfrac{1}{2}[\tfrac{1}{2}TG\left(\tfrac{\nu+k}{2}\right) - \tfrac{1}{2}TG\left(\tfrac{\nu}{2}\right) \\
& && +\tfrac{k}{\nu(\nu+k)} - \tfrac{1}{\nu+k} + \tfrac{\nu+2}{\nu(\nu+k+2)}], \\
F_{\beta\rho_i} &= -\mathbb{E}\left(\tfrac{l(\gamma)}{\partial\beta\partial\rho_i}\right) &&= \mathbb{E}\left(\tfrac{l(\gamma)}{\partial\beta\partial\nu}\right) = 0,
\end{aligned}
\tag{2.22}
$$

where $\mathrm{TG}(x) = \frac{d^2}{d^2x}\log(\Gamma(x))$ is the trigamma function. The partitioned Fisher matrix is

$$
F(\gamma) = \begin{bmatrix} F_{\beta\beta} & 0 & 0 \\ 0 & F_{\rho\rho} & F_{\rho\nu} \\ 0 & F_{\nu\rho} & F_{\nu\nu} \end{bmatrix}.
$$

The log-likelihood function (2.14) can be maximized using Fisher-Scoring-Algorithm. Lange, Roderick, Little & Taylor (1989) compare Fisher-Scoring to EM-estimation. Algorithmic details and proofs for Score and Fisher matrix are given in this paper as well as alternative assumptions on robust linear mixed models.

In the literature, one can find the extension of linear random effect models to semiparametric linear mixed models, see Ishwaran & Takahara (2002) and Zhang & Davidian (2001). The semiparametric term refers to the unknown density of the random effects density or a random measure with unknowns random effects. This terminology is often misleading because semiparametric modeling also refers to additive modeling of continuous covariates, which underlying structure is deterministic.

For more flexibility in the random effects structure, mixture models got very popular in the past. A mixture model is obtained by finitely mixing the conditional distribution. For non-Bayesian semiparametric approaches to linear mixed models, see Verbeke & Lessafre (1996) and Aitkin (1999) who have used a finite mixture approach with implementation by the EM algorithm.

A more Bayesian based framework is founded on the Dirichlet process as discussed in Ferguson (1973). It is applied on random effect models by the idea of random partition

structures and Chinese Restaurant processes (CR process). Later, Brunner, Chan, James & A.Y.Lo (2001) extended these ideas to the weighted Chinese Restaurant Processes which they applied to Bayesian Mixture models. Brunner, Chan, James & A.Y.Lo (1996) provided the general methodology related to i.i.d weighted Chinese Restaurant algorithms (WCR-Algorithms). Ishwaran & Takahara (2002) combined the iid WCR algorithm with REML estimates for inference in Laird-Ware random effect models. Naskar, Das & Ibrahim (2005) used WCR and EM algorithm for survival data.

Chapter 3

Semi-Parametric Mixed Models

There is an extensive amount of literature on the linear mixed model, starting from Henderson (1953), Laird & Ware (1982) and Harville (1977). Nice overviews including more recent work is described in Verbeke & Molenberghs (2001), McCulloch & Searle (2001). Generally, the influence of covariates is restricted to a strictly parametric form in linear mixed models. While in regression models, much work has been done to extend the strict parametric form to more flexible forms of semi- and nonparametric regression, but much less has been done to develop flexible mixed model. For overviews on semiparametric regression models, see Hastie & Tibshirani (1990), Green & Silverman (1994) and Schimek (2000).

A first step to more flexible mixed models is the generalization to additive mixed models where a random intercept is included. With response y_{it} for observation t on individual/cluster i and covariates $u_{i1}, \ldots u_{im}$, the basic form is

$$y_{it} = \beta_0 + \alpha_{(1)}(u_{i1}) + \cdots + \alpha_{(m)}(u_{im}) + b_{i0} + \varepsilon_{it}, \tag{3.1}$$

where $\alpha_{(1)}(.), \ldots, \alpha_{(m)}(.)$ are unspecified functions of covariates $u_{i1}, \ldots, u_{im}, b_{i0}$ is a subject-specific random intercept with $b_{i0} \sim N(0, \sigma_b^2)$ and ε_{it} is an additional noise variable. Estimation for this model may be based on the observation that regression models with smooth components may be fitted by mixed model methodology. Speed (1991) indicated that the fitted cubic smoothing splines is a best linear unbiased predictor. Subsequently the approach has been used in several papers to fit mixed models, see e.g. Verbyla, Cullis, Kenward & Welham (1999), Parise, Wand, Ruppert & Ryan (2001), Lin & Zhang (1999), Brumback & Rice (1998), Zhang, Lin, Raz & Sowers (1998),Wand (2003). Bayesian approaches have also been considered, see e.g., by Fahrmeir & Lang (2001), Lang, Adebayo, Fahrmeir & Steiner (2003), Fahrmeir, Kneib & Lang (2004) and Kneib & Fahrmeir (2006).

3.1 Short Review on Splines in Semi-Parametric Mixed Models

In linear mixed models the strict linear and parametric terms can be extended by incorporating a continuous covariate u which has an additive functional influence in the sense of

$$y_{it} = x_{it}^T \beta + \alpha(u_{it}) + z_{it}^T b_i + \epsilon_{it}. \tag{3.2}$$

In the following, we consider several ways to approximated an unknown function $\alpha(.)$.

3.1.1 Motivation: The Interpolation Problem

For simplicity, we first consider the approximation of a function $\alpha(.)$ with known values at measurement points (knots) u_{it}. Then we start with observations $(u_{it}, \alpha(u_{it})), i = 1, \ldots, n$.

The interpolation problem is given by finding a function $s(.)$ with property

$$s(u_{it}) = \alpha(u_{it}), i = 1, \ldots, n \quad t = 1, \ldots, T_i.$$

Spline theory is a common way to solve this problem. The function $\alpha(.)$ may be approximated by a Spline function $s(.)$. A spline is based on a set of knots $K = \{k_1, \ldots, k_{\tilde{M}}\}$ in the range $[k_1, k_{\tilde{M}}]$. K is the set of ordered observations $K = \{k_j | k_j \leq k_{j+1}\}$. It has elements k_j which are ordered. The smallest value is k_1

The spline is of degree d, $d \in \mathbb{N}_0$ on K, if s(.) is $d-1$ times continuously differentiable and for every $u \in [k_j, k_{j+1})$ $s(u)$ is a polynomial of degree d, $j = 1, \ldots, \tilde{M} - 1$. A spline on $[k_j, k_{j+1})$ of degree d or order d+1 may be represented by

$$s(u) = a_j^{[d]} u^d + a_j^{[d-1]} u^{d-1} + \cdots + a_j^{[1]} u^1 + a_j^{[0]}.$$

The vector space for splines with degree d for the given knots K is denoted by $S_d(K)$.

Splines interpolate given data points u_{it} and their known function evaluations $\alpha(u_{it})$ by using piecewise polynomials to connect these data points. Other interpolation strategies are also possible, i.e., trigonometric interpolation or classical polynomial interpolation, but these methods often have disadvantages in the numeric sense. In Figure 3.1, such an interpolation problem is given for known pairs $(u_{it}, s(u_{it}))$. Since the data points are the given knots, $s(.)$ is a polynomial between successive u-values. In the case of cubic splines

first and second derivative are continuous at the observation points. The interpolation spline is given in Figure 3.2.

The set of all splines of degree d on the knots K is a $\tilde{M} + d - 1 = M$ subspace of the vector space, which contains the $d-1$ times differentiable functions. That is why a spline can be expressed by a set of M linear independent basis functions $\phi_j(.), j \in 1, M$. So $S_d(K)$ can be described by the spline basis $\mathcal{B} = \{\phi_1(u), \ldots, \phi_M(u)\}$.

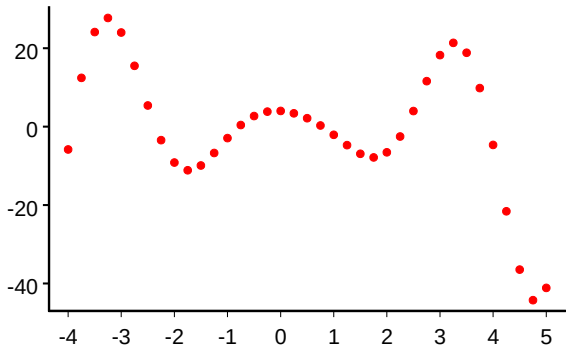


Figure 3.1: Interpolation problem - The values u_{it} are on the x-axis and the corresponding $s(u_{it})$ are on the y-axis. These observed points should be part of continous functin $s(.)$

Function $\alpha(u)$ with $u \in [a, b]$ may be approximated by a spline $s(u)$ using basis functions so that

$$\alpha(u) \approx s(u) = \sum_{j=1}^{M} \phi_j(u)\alpha_j = \phi(u)^T\alpha.$$

The problem in (3.2) for known function evaluations $\alpha(u_{it})$ can be written as

$$y_{it} = x_{it}^T\beta + \phi(u_{it})^T\alpha + z_{it}^T b_i + \epsilon_{it}, \tag{3.3}$$

since $s(u_{it}) = \phi(u_{it})^T\alpha$ is the spline interpolation of $\alpha(u_{it})$.

3.1.2 Popular Basis Functions

Truncated Power Series One basis for $S_d(K)$ for a given set of knots K with degree d is

$$\begin{aligned} &\phi_1(u) = 1, \phi_1(u) = u, \ldots, \phi_{d+1}(u) = u^d, \\ &\phi_{d+2}(u) = (u - k_2)_+^d, \ldots, \phi_{d+\tilde{M}-1}(u) = (u - k_{\tilde{M}-1})_+^d, \end{aligned} \tag{3.4}$$

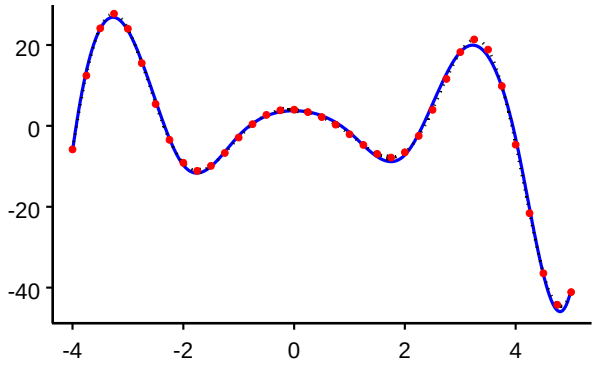


Figure 3.2: Spline solution of the interpolation problem. (The solid points are the given points, the blue line is the spline function (interpolation spline)

where

$$(u-k_j)_+^d = \begin{cases} (u-k_j)^d & \text{, if } u \geq k_j \\ 0 & \text{, else} \end{cases}.$$

The basis is

$$\mathcal{B} = \{\phi_1(u), \ldots, \phi_M(u)\}.$$

B-splines B-splines for degree d with $\tilde{M}$ inner knots in the range $[a, b]$ can be recursively defined by

"De Boor" recursion for B-splines

$$\phi_j^0(u) = \chi_{[k_j,k_{j+1}]}(u) = \begin{cases} 1 & \text{, if } \quad k_j \leq u < k_{j+1} \\ 0 & \text{, else} \end{cases}, \tag{3.5}$$

$$\phi_j^d(u) = \tfrac{k_{j+d-1}-u}{k_{j+d-1}-k_{j+1}}\phi_{j+1}^{d-1}(x) + \tfrac{u-k_j}{k_{j+d}-k_j}\phi_j^{d-1}(u).$$

For the construction of the B-spline basis with the recursion (3.5), outer knots are necessary in the form of $k_1 <= \cdots <= k_{d-1} <= a_1$ and $a_2 <= k_{\tilde{M}+d} <= \cdots <= k_{\tilde{M}+2*d-1}$, which are usually based on equidistant knots. Then the B-splines basis for $S_d(K)$ is

$$\mathcal{B} = \{\phi_1(u), \ldots, \phi_M(u)\}.$$

3.1.3 Motivation: Splines and the Concept of Penalization - Smoothing Splines

A main problem in statistics is that the function evaluation $\alpha(u_{it})$ is not observable from the data. Instead, only the response y_{it} is observable, which is a sum of the unknown values $\alpha(u_{it})$ and ϵ_{it}. These values have to be estimated from the data.

In this subsection, all observation are taken as knots ($\tilde{M} == N$). K is the set of ordered observations $K = \{k_j | k_j \leq k_{j+1}\}$. It has elements u_{it}, which are ordered. It is also possible that a (equidistant) grid of knots is given. This is a useful condition, especially in the regression spline context . That is why $\tilde{M}$ was used instead of N.

We change the cluster representation of the simple semi-parametric model (3.2)

$$y_{it} = x_{it}^T \beta + \phi(u_{it})^T \alpha + z_{it}^T b_i + \epsilon_{it},$$

which is in matrix form

$$Y = X\beta + \Phi\alpha + \mathbb{Z}b + \epsilon, \tag{3.6}$$

to the elementwise measurement representation

$$y_{(i)} = x_{(i)}^T \beta + \phi(u_{(i)})^T \alpha + z_{(i)}^T b + \epsilon_{(i)},$$

where $._{(i)}$ indicates i-th row vector for matrix Z,X or $._{(i)}$ indicates the i-the element of vector Y,u,ϵ and $\phi(u_{(i)})^T = (\phi_1(u_{(i)}), \ldots, \phi_M(u_{(i)}))$ is the basis function evaluation of $u_{(i)}$.

The job of spline smoothing is primary to find a good estimation $s(u_{it}) = \phi^T(u_{it})\hat{\alpha}$ for the unknown function evaluations $\alpha(u_{it}) \approx \hat{s}(u_{it})$. The difficulty of equation (3.6) is that it is not identifiable since $dim(y) = N$ and $dim(\Phi) = N \times (N + d - 1)$. To solve this problems further restrictions to the estimation concepts have to be made. Since equation (3.6) is the equation of a mixed model with assumption $\epsilon \sim N(0, \sigma_\epsilon^2 I)$ and $b \sim N(0, \mathbb{Q}(\rho))$ the estimates for fixed effects β, α and structural parameters $\theta^T = (\sigma_\epsilon^2, \rho)$ would normally be obtained by ML or REML for cases $dim(y) > dim((\beta^T, \alpha^T)^T)$.

That is why the idea of penalizing the roughness of estimated curves was born. The roughness of curves $s(.)$ is controlled by a penalty term and a smoothing parameter λ. Consider the minimization problem for β, α, θ

$$\sum_{i=1}^{N} l_i(\beta, \alpha, \theta) - \frac{1}{2}\lambda \int (\alpha''(u))^2 du \rightarrow \min \tag{3.7}$$

where $l_{(i)}(\beta, \alpha, \theta)$ is the likelihood contribution of observation $y_{(i)}$. It is assumed that $\alpha(.)$ has continuous first and second derivatives $\alpha'(.)$ and $\alpha''(.)$ with $\alpha''(.)$ is quadratically integrable. That is in detail the function class described by the Sobolev space (see Alt (1985) and Adams (1975)). In other words, $\alpha(.)$ can be approximated by a spline functions $s(.)$ that is based on an smoothing parameter λ. To show the dependence of λ let $s_\lambda(u)$ the spline function, which is the result of the minimization problem in formula 3.7 for given λ. $l(\beta, \alpha, \theta) = \sum_{i=1}^{N} l_i(\beta, \alpha, \theta)$ is the marginal likelihood for model (3.6).

The bias$(\hat{s}(u), \alpha(u))$ of $\alpha(u)$ and $\hat{s}(u)$ is increasing for big λ's. The principal trade-off between the bias and the variance of the estimated is reflected by the mean squared error

$$\begin{aligned} \mathbb{E}(s_\lambda(u) - \alpha(u))^2 &= \text{var}(\hat{s}_\lambda(u)) + [\mathbb{E}(\hat{s}_\lambda(u)) - \alpha(u))]^2 \\ &= \text{var}(\hat{s}_\lambda(u)) + (\text{bias}(\hat{s}_\lambda(u), \alpha(u)))^2. \end{aligned}$$

In other words large values of λ lead to underfitting, to small values to overfitting. Getting an optimal λ and an optimal spline s is a difficult statistical problem to solve.

The maximization problem (3.7) may be solved by a natural cubic smoothing spline as described in Reinsch (1967) without the need of a basis function representation of $s(.)$. The concept of Reinsch (1967) and De Boor (1978) can be transferred to the mixed model methology where minimizing (3.7) for given ρ is equivalent to

$$\sum_{i=1}^{N} l_i(\beta, \alpha, \theta)^2 - \frac{1}{2}\lambda s' \tilde{K} s \rightarrow \min, \tag{3.8}$$

where $\hat{y}_{(i)} = x_{(i)}^T \hat{\beta} + s_\lambda(u_{(i)}) + z_{(i)}^T \hat{b}$, $s^T = (s(k_1), \ldots, s(k_M))$. For details on the cubic splines $s(.)$ and penalty matrix $\tilde{K}$ see Fahrmeir & Tutz (2001).

3.1.4 Motivation: The Concept of Regression Splines

The basic idea of regression splines is to work with only a small number of knots $(\tilde{M} << N)$.

One has to find suitable knots K in that sense that the placement and also the number of knots are responsible for the roughness of the curve. This concept my be understood as adaptive selection of knots and their placement. Here the number and position of knots strongly determine the degree of smoothing. The position of knots may be chosen uniformly over the data, at appropriate quantiles or by more complex data-driven schemes.

For a detailed discussion of these issues see Friedman & Silverman (1989),Wand (2000) or Stone, Hansen, Kooperberg & Truong (1997)

Another idea is to take a (equidistant) fixed grid of knots. That is the main difference to smoothing splines, since the knots are chosen individually (how many knots, range of the interval where the knots coming from). A spline function for a fixed grid without penalization is visualized in Figure 3.4. For this Figure and Figure 3.3, a random effects model with only one smooth covariate ($\alpha(u) = sin(u)$) was used. Therefore forty clusters with five repeated measurements each were simulated. The random effect was assumed to be $N(0, \sigma_b^2)$, $\sigma_b^2 = 2$ and the error term was assumed to be $N(0, \sigma_\epsilon^2)$, $\sigma_\epsilon^2 = 2$. In Figure 3.3 and 3.3, the concept of regression splines was used. In Figure 3.3, the smoothing parameter λ was set to zero and in Figure 3.4, it was set to sixty. It is obvious that the roughness of the curve has to be penalized. On the other hand a spline function is desired that is independent of the placement and the number of knots.

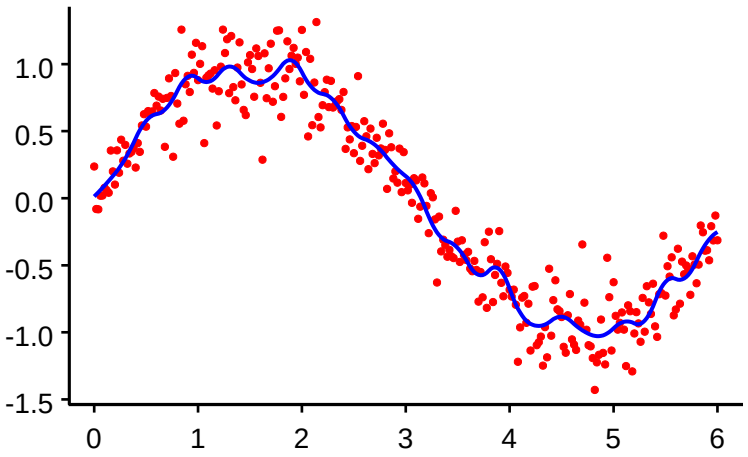


Figure 3.3: The red points describe the given data, the blue line in the figure is a spline computed with 40 knots and B-splines of degree 3. (no penalization)

Again the penalization problem for given ρ is

$$\sum_{i=1}^{N} l_i(\beta, \alpha, \theta) - \lambda P(s(.)) \rightarrow \min, \tag{3.9}$$

where $s(u)$ has the functional form $s(u) = \sum_{j=1}^{M} \phi_j(u)\alpha_j$ and $P(s(.))$ is a roughness functional. A widely used roughness functional is the integrated squared second derivative

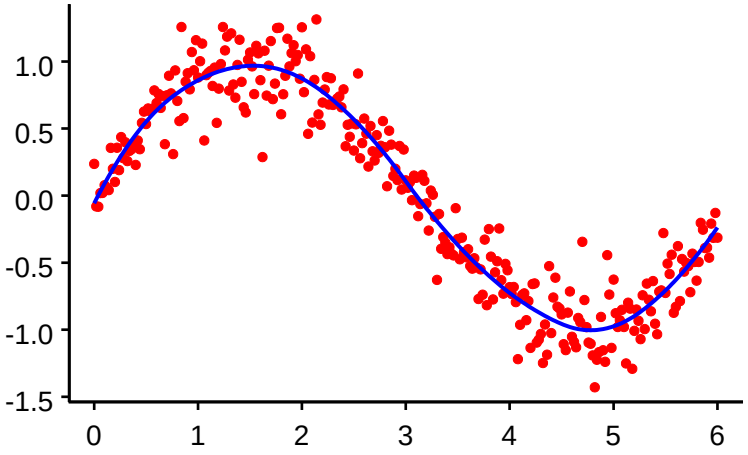


Figure 3.4: The spline in the figure is the optimal spline with respect to the penalty and λ. Using the penalty term reduces the influence of knots. Here 40 knots and B-splines of degree 3 were used.

$P(s(.)) = \int (s''(u))du$ as an measure for the curvature or the roughness of the function $s(.)$. If the basis function representation of a spline is used the penalty term has the form

$$P(s(.)) = \alpha^T K \alpha \tag{3.10}$$

where K is a matrix with entries $K_{ij} = \int \phi_i''(u)\phi_j''(u)du$.

Eilers & Marx (1996) introduced a penalty term where adjacent B-spline coefficients are connected to each other in a very distinct way. The penalty term is based on $\tilde{K} = (D^l)^T D^l$, where D^l is a contrast matrix of order l which contrasts polynomials of the order l. Using B-splines with penalization K one penalizes the difference between adjacent categories in the form $\lambda\alpha^T K\alpha = \lambda \sum_j \{\triangle^l \alpha_j\}^2$. $\triangle$ is the difference operator with $\triangle\alpha_j = \alpha_{j+1} - \alpha_j$, $\triangle^2\alpha_j = \triangle(\triangle\alpha_j)$ etc., for details see Eilers & Marx (1996). Usually the order of the penalized differences is the same as the order of the spline (B-Spline).

In Figure 3.4, one can see that penalization reduces the influence of knots, which has an effect on the roughness of the curve. So penalization reduces also the influence of the number and placement of the knots. Another number of knots with different placements would deliver a quite similar spline function solution.

The difference matrix D is needed to compute the difference matrix of the l-th order

D^l corresponding to B-Spline penalization (see Eilers & Marx (1996)) in an recursive scheme. With D being the $(M-1)\times(M)$ contrast matrix

$$D = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}$$

one obtains higher order differences by the recursion $D^l = DD^{l-1}$ which is a $(M-l)\times M$ matrix. This can be used for a more simple and intuitive definition of the penalty than equation 3.10.

A similar argumentation is used for the truncated Power Series basis where the penalty matrix is simply set to

$$K = \text{bdiag}(0_{(d)\times(d)}, I_{M-d}).$$

where $0_{(d)\times(d)}$ is a d-dimensional quadratic zero matrix, and $I_{(M-d)}$ is the identity matrix of dimension $(M-d)$.

3.1.5 Identification Problems: The Need of a Semi-Parametric Representation

Problems in additive modeling based on splines arise, if intercepts or splines for other covariates are used. If no further restriction of the splines is made, the resulting splines are not clearly identifiable. This is illustrated in the following example

Example 3.1 : Rewriting an additive term to a semi-parametric term
One can write the additive term without parametric terms

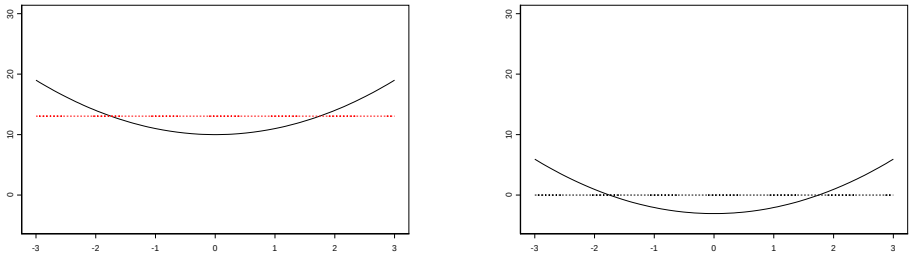
$$\alpha(u) = 10 + u^2, \text{ for } u \in [-3,3]$$

to a semi-parametrical representation

$$\alpha(u) = \beta_0 + \tilde{\alpha}(u) = 10 + u^2, \text{ for } u \in [-3,3]$$

with $\beta_0 = 10$ and $\tilde{\alpha}(u) = u^2$. But also $\beta_0 = 5$ and $\tilde{\alpha}(u) = 5 + u^2$ is a valid semiparametric parametrization for the additive term $\alpha(u)$.

The interest is often in the population mean level and in the absolute deviations from this mean as a function of a continuous covariates. It is a natural idea to center the continuous covariates



(a) (b)

Figure 3.5: (a) may be seen as pure additive spline smoothing, (b) may be seen that $\alpha(u)$ describes the absolute deviation from zero. In this case the level is $\tilde{\tilde{\beta}}_0 = \frac{\int_a^b \alpha(u)du}{b-a}$. The desired property of the additive term $\tilde{\tilde{\alpha}}(u)$ is $\int_a^b \tilde{\tilde{\alpha}}(u) = 0$

around zero. Figure (3.5) shows the difference in the interpretation. Nice benefit of this restriction is that the semi-parametrical representation is identifiable. □

Also two or more additive terms have to be rewritten to semi-parametric terms, since additive terms should be identifiable. This is illustrated in the

Example 3.2 : Rewriting additive terms to semi-parametric terms
One can write additive terms

$$\begin{aligned} \alpha_{(1)}(u) &= 10 + u^2, \text{ for } u \in [-3,3] = [a_{(1)}, b_{(1)}] \\ \alpha_{(2)}(v) &= -5 + v^3, \text{ for } v \in [-3,3] = [a_{(2)}, b_{(2)}] \end{aligned}$$

to the additive predictor

$$\mu^{add}(u,v) = \alpha_{(1)}(u) + \alpha_{(2)}(v).$$

Again $\tilde{\alpha}_{(1)}(u) = 5 + u^2$ *and* $\tilde{\alpha}_{(2)}(v) = v^3$ *corresponds to the same additive predictor since* $\mu^{add}(u,v) = \tilde{\alpha}_{(1)}(u) + \tilde{\alpha}_{(2)}(v)$*. Using the same idea described in example 3.1, one gets identifiable additive terms by the reparametrization of the additive predictor to semi-parametric terms*

$$\mu^{add}(u,v) = \beta_0 + \tilde{\tilde{\alpha}}_{(1)}(u) + \tilde{\tilde{\alpha}}_{(2)}(v)$$

with properties $\int_{a_{(1)}}^{b_{(1)}} \tilde{\tilde{\alpha}}_{(1)}(u) = 0$, $\int_{a_{(2)}}^{b_{(2)}} \tilde{\tilde{\alpha}}_{(2)}(v) = 0$ *and* $\tilde{\tilde{\beta}} = \frac{\int_{a_{(1)}}^{b_{(1)}} \alpha(u)du}{b_{(1)}-a_{(1)}} + \frac{\int_{a_{(2)}}^{b_{(2)}} \alpha(v)dv}{b_{(2)}-a_{(2)}}$. □

So the basic idea for the identifiable additive terms $\tilde{\alpha}(u)$ is to rewrite additive terms $\alpha(u)$ to a semi-parametric consideration with $\int_a^b \tilde{\alpha}(u) = 0$. So $\tilde{\alpha}(u)$ has the form

$$\tilde{\alpha}(u) = \alpha(u) - \int_a^b \frac{\alpha(u)}{b-a} du. \tag{3.11}$$

Using simple analysis, one can show that equation (3.11) holds $\int_a^b \tilde{\alpha}(u) = 0$.

Since $\alpha(u)$ is often approximated by a spline function that is composed of basis functions $\alpha(u) \approx \phi^T(u)\alpha$, where $\phi^T(u) = (\phi_1(u), \ldots, \phi_M(u))$, the discrete version using the coefficients of basis functions can also used to get regularized $\tilde{\alpha}^T = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_M)$ with restriction $\sum_{j=1}^M \tilde{\alpha}_j = 0$. A regularized version may be obtained by

$$\tilde{\alpha} = \alpha - \frac{\sum_{j=1}^M \alpha_j}{M}, \qquad \tilde{\alpha}_m = -\sum_{j=1}^{M-1} \tilde{\alpha}_j.$$

There term $\frac{\sum_{j=1}^M \alpha_j}{M}$ is often understood as shift in the level of the function $\alpha(u)$. For detailed information on these restrictions, see the Appendix A.1.

3.1.6 Singularities: The Need of a Regularization of Basis Functions

Let a semi-parametrization be given by

$$y_{(i)} = x_{(i)}^T\beta + \sum_{j=1}^M \phi_j(u_{(i)})\alpha_j + z_{(i)}^T b + \epsilon_{(i)} = \begin{bmatrix} 1 & \tilde{x}_{(i)} & \phi(u_{(i)})^T \end{bmatrix} \begin{bmatrix} \beta_0 \\ \tilde{\beta} \\ \alpha \end{bmatrix} + z_{(i)}^T b + \epsilon_{(i)}$$

where $x_{(i)}^T = (1, \tilde{x}_{(i)})$ and $\beta^T = (\beta_0, \tilde{\beta}^T)$. For a truncated power series bases, the column corresponding to the intercept is absolute linear dependent on $\phi_1(u) = 1$. The rows for the design matrix for fixed effects can be written as $\begin{bmatrix} 1 & \tilde{x}_{(i)} & 1 & \phi_2(u_{(i)}) & \ldots & \phi_M(u_{(i)}) \end{bmatrix}$. It is obvious that the design matrix has not full rank. The same problem affects also B-splines. A B-spline basis is a specific decomposition of 1. A main property of B-splines is that one has $\sum_{j=1}^M \phi_j(u) = 1$. There exists the linear combination of the design matrix $\begin{bmatrix} 1 & \tilde{x}_{(i)} & \sum_{j=1}^M \phi_j(u_{(i)}) \end{bmatrix}$, which shows once again problems in the rank of the design matrix.

The same problem is arising when more than one additive functions are used. For truncated power series one has columns $\phi_1^{(k)} = 1$ and $\phi_1^{(l)} = 1$ in the design matrix for additive components k and l. For B-splines the corresponding sums $\sum_{j=1}^M \phi_j^{(k)} = 1$ and $\sum_{j=1}^M \phi_j^{(l)} = 1$ are absolute linear dependent.

To solve these singularities specific transformations T must be applied to $\Phi(u)$ with

$$\tilde{\Phi}(u) = T\Phi(u),$$

where $T\Phi(u)$ has full rank. Generally these transformations also affects the penalty matrix K that occur in a penalized likelihood function. See the Appendix A.1 for a detailed discussion on these transformations.

3.2 The Model

Let the data be given by $(y_{it}, x_{it}, u_{it}, z_{it}), \quad i = 1, \ldots, n, \quad t = 1, \ldots, T_i$, where y_{it} is the response for observation t within cluster i and $x_{it}^T = (x_{it1}, \ldots, x_{itp}), \quad u_{it}^T = (u_{it1}, \ldots, u_{itm}), \quad z_{it}^T = (z_{it1}, \ldots, z_{its})$ are vectors of covariates, which may vary across clusters and observations. The semi-parametric mixed model that is considered in the following has the general form

$$\begin{aligned} y_{it} &= x_{it}^T\beta + \sum_{j=1}^{m} \alpha_{(j)}(u_{itj}) + z_{it}^T b_i + \epsilon_{it} \\ &= \mu_{it}^{par} + \mu_{it}^{add} + \mu_{it}^{rand} + \epsilon_{it} \end{aligned} \tag{3.12}$$

where

$\mu_{it}^{par} = x_{it}^T\beta$ is a linear parametric term,

$\mu_{it}^{add} = \sum_{j=1}^{m} \alpha_{(j)}(u_{itj})$ is an additive term with unspecified influence functions $\alpha_{(1)}, \ldots, \alpha_{(m)}$,

$\mu_{it}^{rand} = z_{it}^T b_i$ contains the cluster-specific random effect $b_i, \quad b_i \sim N(0, Q(\rho))$, where $Q(\rho)$ is a parameterized covariance matrix and

ϵ_{it} is the noise variable, $\epsilon_{it} \sim N(0, \sigma^2 I), \quad \epsilon_{it}, b_i$ independent.

In spline methodology, the unknown functions $\alpha_{(j)}$ are approximated by basis functions. A simple basis is known as the truncated power series basis of degree d, yielding

$$\alpha_{(j)}(u) = \gamma_0^{(j)} + \gamma_1^{(j)} u + \ldots \gamma_d^{(j)} u^d + \sum_{s=1}^{\tilde{M}} \alpha_s^{(j)} (u - k_s^{(j)})_+^d \ ,$$

where $k_1^{(j)} < \ldots < k_{\tilde{M}}^{(j)}$ are distinct knots. More generally, one uses

$$\alpha_{(j)}(u) = \sum_{s=1}^{M} \alpha_s^{(j)} \phi_s^{(j)}(u) = \alpha_j^T \phi^{(j)}(u), \tag{3.13}$$

where $\phi_s^{(j)}$ denotes the s-th basis function for variable $j, \alpha_j^T = (\alpha_1^{(j)}, \ldots, \alpha_M^{(j)})$ are unknown parameters and $\phi^{(j)}(u)^T = (\phi_1^{(j)}(u), \ldots, \phi_M^{(j)}(u))$ represent the vector-valued evaluations of the basis functions.
For semi- and nonparametric regression models, Eilers & Marx (1996), Marx & Eilers

(1998) proposed the numerically stable B-splines which have also been used by Wood (2004). For further investigation of basis functions, see also Wand (2000), Ruppert & Carroll (1999).

By collecting observations within one cluster the model has the form

$$\begin{aligned} y_i &= X_i\beta + \Phi_{i1}\alpha_1 + \ldots + \Phi_{im}\alpha_m + Z_i b_i + \epsilon_i, \\ & \begin{bmatrix}\epsilon_i \\ b_i\end{bmatrix} \sim N\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\sigma_\varepsilon^2 I & \\ & Q(\rho)\end{pmatrix}\right), \end{aligned} \tag{3.14}$$

where $X_i\beta$ contains the linear term, $\Phi_{ij}\alpha_j$ represents the one additive term and $Z_i\beta$ the random term. Vectors and matrices are given by $y_i^T = (y_{i1}, \ldots, y_{iT_i})$, $X_i^T = (x_{i1}, \ldots, x_{iT_i}), \Phi_{ij}^T = (\phi^{(j)}(u_{i1j}), \ldots, \phi^{(j)}(u_{iT_ij}))$, $Z_i^T = (z_{i1}, \ldots, z_{iT_i})$, $\epsilon_i^T = (\epsilon_{i1}, \ldots, \epsilon_{iT_i})$. In the case of the truncated power series the "fixed" term $\gamma_0^{(j)} + \gamma_1^{(j)}u + \ldots + \gamma_d^{(j)}u^d$ is taken into the linear term $X_i\beta$ without specifying X_i and β explicitly.
In matrix form one obtains

$$y = X\beta + \Phi_{.1}\alpha_1 + \ldots + \Phi_{.m}\alpha_m + Zb + \epsilon$$

where $y^T = (y_1^T, \ldots, y_n^T)$, $b^T = (b_1^T, \ldots, b_n^T)$, $\epsilon^T = (\epsilon_1^T, \ldots, \epsilon_n^T)$,
$X^T = (X_1^T, \ldots, X_n^T)$, $\Phi_{.j}^T = (\Phi_{1j}^T, \ldots, \Phi_{nj}^T)$, $Z^T = (Z_1^T, \ldots, Z_n^T)$.
Parameters to be estimated are the fixed effects, which are collected in $\delta^T = (\beta^T, \alpha_1^T, \ldots, \alpha_m^T)$ and the variance specific parameters $\theta^T = (\sigma_\varepsilon, \rho^T)$ which determine the covariances $cov(\epsilon_{it}) = \sigma_\varepsilon^2 I_{T_i}$ and $cov(b_i) = Q(\rho)$. In addition one wants to estimate the random effects b_i. Since b_i is a random variable, the latter is often called prediction rather than estimation. We set $X_{\Phi i.} = [X_i, \Phi_{i1}, \ldots, \Phi_{im}]$.

3.3 Penalized Maximum Likelihood Approach

Starting from the marginal version of the model

$$\begin{aligned} & y_i = X_i\beta + \Phi_{i1}\alpha_1 + \ldots + \Phi_{im}\alpha_m + \epsilon_i^* \\ & \text{or} \\ & y_i = X_{\Phi i}\delta + \epsilon_i^* \qquad\qquad , \\ \epsilon_i^* \sim N(0, V_i(\theta)), & \quad V_i(\theta) = \sigma_\epsilon^2 I_{T_i} + Z_i Q(\rho) Z_i^T, \end{aligned} \tag{3.15}$$

estimates for δ may be based on the *penalized log-likelihood*

$$l_p(\delta;\theta) = -\frac{1}{2}\sum_{i=1}^n \log(|V_i(\theta)|) - \sum_{i=1}^n \frac{1}{2}(y_i - X_{\Phi i.}\delta)^T (V_i(\theta))^{-1}(y - X_{\Phi i.}\delta) - \frac{1}{2}\delta^T K \delta, \tag{3.16}$$

where $\delta^T K\delta$ is a penalty term which penalized the coefficients $\alpha_1, \ldots, \alpha_n$. For the truncated power series an appropriate penalty is given by

$$K = Diag(0, \lambda_1 I, \ldots, \lambda_m I),$$

where I denotes the identity matrix and λ_j steers the smoothness of the function $\alpha_{(j)}$. For $\lambda_j \to \infty$ a polynomial of degree d is fitted. P-splines ((Eilers & Marx, 1996)) use $K = D^T D$ where D is a matrix that builds the difference between adjacent parameters yielding the penalty $\delta^T K\delta = \Sigma_j \lambda_j \Sigma_s (\alpha_{s+1}^{(j)} - \alpha_s^{(j)})^2$ or higher differences.

From the derivative of $l_p(\delta, \theta)$, one obtains the estimation equation $\partial l_p(\delta, \phi)/\partial\delta = 0$ which yields

$$\sum_{i=1}^{n} (X_{\Phi i.}^T (V_i(\theta))^{-1} y_i) = (\sum_{i=1}^{n} (X_{\Phi i.}^T (V_i(\theta))^{-1} X_{\Phi i.} + K)^{-1})\hat{\delta}$$

and therefore

$$\hat{\delta} = (\sum_{i=1}^{n} (X_{\Phi i.}^T (V_i(\theta))^{-1} X_{\Phi i.} + K))^{-1} \sum_{i=1}^{n} X_{\Phi i.}^T (V_i(\theta))^{-1} y_i$$

which depends on the variance parameters θ. It is well known that maximization of the log-likelihood with respect to θ yields biased estimates since maximum likelihood does not take into account that fixed parameters have been estimated (see Patterson & Thompson (1974)). The same holds for the penalized log-likelihood (3.16). Therefore for the estimation of variance parameters often restricted maximum likelihood estimates (REML) are preferred which are based on the log-likelihood

$$\begin{aligned} l_r(\delta, \theta) &= -\frac{1}{2} \sum_{i=1}^{n} \log(|V_i(\theta)|) - \frac{1}{2} \sum_{i=1}^{n} (y_i - X_{\Phi i.}\beta)^T V_i(\theta)^{-1} (y_i - X_{\Phi i.}\beta) \\ &\quad -\frac{1}{2} \sum_{i=1}^{n} \log(|X_{\Phi i.}^T V_i(\theta) X_{\Phi i.}|), \end{aligned}$$

see, Harville (1974), Harville (1977) and Verbeke & Molenberghs (2001).
The restricted log-likelihood differs from the log-likelihood by an additional component. One has

$$l_r(\delta, \theta) = l(\delta, \theta) - \frac{1}{2} \sum_{i=1}^{n} \log(|X_{\Phi i.}^T V_i(\theta) X_{\Phi i.}|).$$

It should be noted that for the estimation of θ, the penalization term $\delta^T K\delta$ may be omitted since it has no effect. Details on REML is given in the Appendix.

BLUP Estimates

Usually one also wants estimates of the random effects. Best linear unbiased prediction (BLUP) is a framework to obtain estimates for δ and $b_1, \ldots, b_n$ for given variance components. There are several ways to motivate BLUP (see Robinson (1991)). One way is to consider the joint density of y and b which is normal and maximize with respect to δ and b. By adding the penalty term $\delta^T K \delta$ one has to minimize

$$\sum_{i=1}^{n} \frac{1}{\sigma^2}(y_i - X_{\Phi i.}\delta - Z_i b_i)^T (y_i - X_{\Phi i.}\delta - Z_i b_i) + b_i^T Q(\rho)^{-1} b_i + \delta^T K \delta, \qquad (3.17)$$

where $X_{\Phi i.} = [X_i, \Phi_{i1}, \ldots, \Phi_{im}]$, $\mathbb{Q}(\rho) = Diag(Q(\rho) \ldots Q(\rho))$.
With $X_\Phi^T = (X_{\Phi.1}^T \ldots X_{\Phi.m}^T)$ the criterion (3.17) may be rewritten as

$$\frac{1}{\sigma^2}(y - X_\Phi \delta - \mathbb{Z}b)^T (y - X_\Phi \delta - \mathbb{Z}b) + b^T \mathbb{Q}(\rho)^{-1} b^T + \delta^T K \delta$$

which yields the "ridge regression" solution

$$\begin{bmatrix} \hat{\delta} \\ \hat{b} \end{bmatrix} = \left(C^T \tfrac{1}{\sigma_\varepsilon^2} I C + B \right)^{-1} C^T \frac{1}{\sigma_\varepsilon^2} I y$$

with $C = (X_\Phi, Z)$ and

$$B = \begin{pmatrix} K & 0 \\ 0 & \mathbb{Q}(\rho)^{-1} \end{pmatrix}.$$

Some matrix derivation shows that $\hat{\delta}$ has the form

$$\hat{\delta} = (X_\Phi^T V(\theta)^{-1} X_\Phi + K)^{-1} X^T V(\theta)^{-1} y,$$

where $V(\theta) = Diag(V_1(\theta) \ldots V_n(\theta))$, and for the vector of random coefficients $b^T = (b_1^T, \ldots, b_n^T)$ one obtains

$$\hat{b} = \mathbb{Q}(\rho)\mathbb{Z}^T V(\theta)^{-1}(y - X_\Phi \hat{\delta}).$$

In simpler form BLUP estimates are given by

$$\hat{\delta} = (\sum_{i=1}^{n} (X_{\Phi i.}^T (V_i(\theta))^{-1} X_{\Phi i.} + K))^{-1} \sum_{i=1}^{n} X_{\Phi i.}^T (V_i(\theta))^{-1} y_i,$$
$$\hat{b}_i = Q Z_i^T V_i(\theta)^{-1}(y_i - X_{\Phi i.}\hat{\delta}).$$

3.4 Mixed Model Approach to Smoothing

It is necessary to specify the smoothing parameters $\lambda_1, \ldots, \lambda_m$ for the computation of $\hat{\delta}$. Using cross-validation techniques, problems arise, if the number of smooth covariates is high. An approach that works for moderate number of smooth covariates uses the ML or REML estimates of variance components. The basic concept ist to reformulate the estimation as a more general mixed model. Let us consider again the criterion for BLUP estimates (3.17) which has the form

$$\begin{aligned} & \frac{1}{\sigma^2}(y - X_\Phi\delta - \mathbb{Z}b)^T(y - X_\Phi\delta - \mathbb{Z}b) + \alpha^T K_\alpha \alpha + b^T \mathbb{Q}(\rho)^{-1} b \\ = \; & \frac{1}{\sigma^2}(y - X\beta - \Phi\alpha - \mathbb{Z}b)^T(y - X\beta - \Phi\alpha - \mathbb{Z}b) + (\alpha^T b^T) \begin{pmatrix} K_\alpha & 0 \\ 0 & \mathbb{Q}(\rho)^{-1} \end{pmatrix} \begin{pmatrix} \alpha \\ b \end{pmatrix} \end{aligned} \tag{3.18}$$

where $\Phi = [\Phi_{.1} \ldots \Phi_{.m}]$ and K_α for the truncated power series has the form $K_\alpha = Diag(\lambda_1 I, \ldots, \lambda_m I)$.
Thus (3.18) corresponds to the BLUP criterion of the mixed model

$$y = X\beta + \begin{bmatrix} \Phi & \mathbb{Z} \end{bmatrix} \begin{pmatrix} \alpha \\ b \end{pmatrix} + \epsilon$$

$$\text{with} \qquad \begin{pmatrix} \alpha \\ b \\ \epsilon \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K_\alpha^{-1} & 0 & 0 \\ 0 & \mathbb{Q}(\rho) & 0 \\ 0 & 0 & \sigma_\varepsilon^2 I \end{pmatrix} \right).$$

Since $K_\alpha = Diag(\lambda_1 I, \ldots, \lambda_m I)$ the smoothing parameters $\lambda_1, \ldots, \lambda_m$ correspond to the variance of the random effects $\alpha_1, \ldots, \alpha_m$ for which $cov(\alpha_j) = \lambda_j I$ is assumed. Thus $\alpha_1, \ldots, \alpha_m$ are treated as random effects for the purpose of estimation. REML estimates yield $\hat{\lambda}_1, \ldots, \hat{\lambda}_m$. For alternative basis function like B-splines some reformulation is necessary to obtain the simple independence structure of the random effects (see Appendix A.1).

3.5 Boosting Approach to Additive Mixed Models

Boosting originates in the machine learning community where it has been proposed as a technique to improve classification procedures by combining estimates with reweighted

observations. Since it has been shown that reweighting corresponds to minimizing iteratively a loss function (Breiman (1999), Friedman (2001)) boosting has been extended to regression problems in a L_2-estimation framework by Bühlmann & Yu (2003). In the following boosting is used to obtain estimates for the semiparametric mixed model. Instead of using REML estimates for the choice of smoothing parameters, the estimates of the smooth components are obtained by using "weak learners" iteratively. The weak learner is the estimate of δ based on a fixed, very large smoothing parameter λ, which is used for all components. By iterative fitting of the residual and selection of components (see algorithm) the procedure adapts automatically to the possibly varying smoothness of components.

3.5.1 Short Review of Likelihood Boosting

Basic idea was to improve the misclassification rates, see Schapire (1990). The basic concept is to use a classifier iteratively with differing weights on the observations and to combine the results in a committee. It has been shown the misclassification error can be reduced dramatically. Recently it has been shown that boosting is a way of fitting an additive expansion in basis functions when the single basis functions represent the results of one iteration of the boosting procedure. The procedure is based on gradient descent by the use of specific loss functions, see Breiman (1999) and Friedman, Hastie & Tibshirani (2000).

Example 3.3 : Functional Gradient Descend

The objective is to minimize $E[L(y, f(x))]$ for general loss function $L(y, f(x))$ in a simple regression context where y is the response and $f(x)$ is a function of the predictor x.

1. *Initialization*
2. *Fit $\hat{f}_0(x) = B(x, \gamma)$ to data (y_i, x_i), where B is a (parameterized) regressor function (learner) Set $m = 0$.*
3. *Negative gradient*
   Determine the negative gradient $r_i = -\partial L(y_i, f^{(m-1)})/\partial f$ and fit $B(x, \gamma)$ to data (r_i, x_i)
4. *Determine the step size ν by minimizing*
   $\sum_i L(y_i, f^{(m-1)}(x_i) + \nu B(x_i, \hat{\gamma}))$
5. *Increase m by one and repeat steps 2 and 3*

□

From this view it is no longer restricted to classification problems. Friedman, Hastie & Tibshirani (2000) replace the exponential loss function, which underlies the classical Adaboost, by the binomial log likelihood yielding LogitBoost. Bühlmann & Yu (2003) investigate L_2 loss, which yields the L_2-Boost algorithm. Tutz & Binder (2006) introduced the likelihood-based boosting concept for all kinds of link functions and exponential family distributions. As in generalized linear models let $y_i|x_i$ have a distribution from a simple exponential family $f(y_i|x_i) = \exp\{(y_i\theta_i - b(\theta_i))/\phi + c(y_i, \phi)\}$, where θ_i is the canonical parameter and ϕ is a dispersion parameter. Instead of assuming a linear predictor $\eta_i = x_i^T\beta$ in each boosting step the fitting of a simple learner $\eta_i = \eta(x_i, \gamma)$ is assumed, where γ is a finite ore infinite-dimensional parameter. If the learner is a regression spline, γ describes the coefficients of the spline functions. The likelihood to be maximized is given by

$$l(\gamma) = \sum_{i=1}^{n} l(y_i, \eta_i) = \sum_{i=1}^{n} (y_i\theta_i - b(\theta_i))/\phi + c(y_i, \phi),$$

where the canonical parameter θ_i is a function of $\eta_i = \eta(x_i, \gamma)$.

Example 3.4 : Likelihood-Based Boosting for Regression Models

1. *Initialization: For given data $(y_i, x_i), i = 1, \ldots, n$, fit the intercept model $\mu^{(0)} = h(\eta^{(0)})$ by maximizing the likelihood $l(\gamma)$ yielding $\hat{\eta}^{(0)} = \eta^{(0)}, \hat{\mu}^{(0)} = h(\hat{\eta}^{(0)})$.*

   For $l = 0, 1, \ldots,$

2. *Fit the model*

$$\mu_i = h(\hat{\eta}^{(l)}(x_i) + \eta(x_i, \gamma)) \tag{3.19}$$

   to data $(y_i, x_i), i = 1, \ldots, n$, where $\hat{\eta}^{(l)}(x_i)$ is treated as an offset and $\eta(x_i)$ is estimated by the learner $\eta(x_i, \hat{\gamma}^{(l)})$. Set $\hat{\eta}^{(l+1)}(x_i) = \hat{\eta}^{(l)}(x_i) + \hat{\eta}(x_i, \hat{\gamma}^{(l)})$.

3. *Stop, if the chosen information criterion could not be improved in the following step*

The estimate $\eta(x_i, \hat{\gamma})$ may represent a spline or some other learner determined by γ. □

The structure of the algorithm in Example 3.4 is used to incorporate variable selection by componentwise learners. Bühlmann & Yu (2005) proposed the concept of sparse boosting. In each iteration, only the contribution of a single variable is determined. A simple learner of this type, which has often been used in boosting, is a tree with only two terminal nodes (stumps). With stumps the selection of the variable to be updated is done implicitly by tree methodology. When using regression splines, model fitting within the algorithm

contains a selection step in which one variable is selected and only the corresponding function is updated. The componentwise update has the advantage that the selection of variables is performed by the fitting of simple models, which contain only one variable.

Example 3.5 : Componentwise Boosting

Step 2 (Model fit)

1. *Fit the model*
$$\mu_{is} = h(\hat{\eta}^{(l)}(x_i) + \eta(x_{i(s)}, \gamma_s)) = h(\eta_{is}) \tag{3.20}$$
to data $(y_i, x_i), i = 1, \ldots, n$*, where* $\hat{\eta}^{(l)}(x_i)$ *is treated as an offset and* $\eta(x_{i(s)})$ *is estimated by the learner* $\eta(x_{i(s)}, \hat{\gamma}_s^{(l)})$*.* $x_{i(s)}$ *stands for the s-th covariate, and* γ_s *for the corresponding coefficient.*

2. *Selection: Select from* $s \in \{1, \ldots, p\}$ *the variable* j *that leads to the smallest* $IC_s^{(l)}$*. The chosen information criterion* IC *in the* l*-th boosting step for variables* s*,* $IC_s^{(l)}$ *is computed commonly by using the log-likelihood* $\sum_{i=1}^{n} l(y_i, \eta_{is})$ *and by using a suitable measure for the effective degrees of freedom. A suitable measure is the trace of the projection matrix, which is responsible for the projection of* y *to* $\hat{\eta}_{is}$*. Common information criteria, based on the trace of the projection matrix, are* AIC *or* BIC*.*

3. *Update:*
$$\hat{\eta}^{(l+1)}(x_i) = \hat{\eta}^{(l)}(x_i) + \eta(x_{i(s)}, \hat{\gamma}_s)$$

□

The estimation step in example 3.5 is similar to the generic functional gradient descend in example3.3. For details see Bühlmann & Yu (2003), but it is not an example in the strict sense. Here ν is set to an constant $\nu = 1$. Functional gradient descend uses the negative gradient of a global loss function evaluated at observations as response in the next iteration. However, the negative derivative of the likelihood yields values that may be considered as responses only in special cases, for example, if the response is unrestricted and continuous. Therefore in the general case the algorithm is a one step improvement of the given estimate represented by the offset that uses the derivative of a penalized likelihood.

3.5.2 The Boosting Algorithm for Mixed Models

The following algorithm uses componentwise boosting. Componentwise boosting means that only one component of the predictor, in our case one smooth term $\Phi_{ij}\alpha_j$, is refitted at a time. That means that a model containing the linear term and only one smooth component is fitted in one iteration step. For simplicity we will use the notation

$$X_{i(r)} = [X_i \; \Phi_{ir}] \quad , \quad \delta_r^T = (\beta^T, \alpha_r^T)$$

for the design matrix with predictor $X_{i(r)} = X_i\beta + \Phi_{ir}\alpha_r$.
The corresponding penalty matrix is denoted by K_r, which for the truncated power series has the form

$$K_r = Diag(0, \; \lambda I).$$

One wants to optimize model (3.12) in the following.
BoostMixed

1. *Initialization*
   Compute starting values $\hat{\beta}^{(0)}, \hat{\alpha}_1^{(0)}, \ldots \hat{\alpha}_m^{(0)}$ and set $\eta_i^{(0)} = X_i\hat{\beta}^{(0)} + \Phi_{i1}\hat{\alpha}_1^{(0)} + \ldots + \Phi_{im}\hat{\alpha}_m^{(0)}$.

2. *Iteration*
   For l=1,2,...

   (a) *Refitting of residuals*

   i. Computation of parameters
   For $r \in \{1, \ldots, m\}$ the model for residuals

   $$y_i - \eta_i^{(l-1)} \sim N(\eta_{i(r)}, V_i(\theta))$$

   with

   $$\eta_{i(r)} = X_{i(r)}\delta_r = X_i\beta + \Phi_{ir}\alpha_r$$

   is fitted, yielding

   $$\hat{\delta}_r = (\sum_{i=1}^{n}(X_{i(r)}^T(V_i(\theta^{(l-1)}))^{-1}X_{i(r)} + K_r))^{-1}\sum_{i=1}^{n}X_{i(r)}^T(V_i(\theta^{(l-1)}))^{-1}(y_i - \eta_i^{(l-1)}).$$

   ii. Selection step
   Select from $r \in \{1, \ldots, m\}$ the component j that leads to the smallest $AIC_r^{(l)}$ or $BIC_r^{(l)}$ as given in Section 3.5.3.

iii. Update

Set $$\hat{\beta}^{(l)} = \hat{\beta}^{(l-1)} + \hat{\beta},$$

and

$$\hat{\alpha}_r^{(l)} = \begin{cases} \hat{\alpha}_r^{(l-1)} & \text{if } r \neq j \\ \hat{\alpha}_r^{(l-1)} + \hat{\alpha}_r & \text{if } r = j, \end{cases}$$

$$\hat{\delta}^{(l)} = ((\hat{\beta}^{(l)})^T,\ (\hat{\alpha}_1^{(l)})^T,\ \dots (\hat{\alpha}_m^{(l)})^T)^T.$$

Update for $i = 1, \dots, n$

$$\eta_i^{(l)} = \eta_i^{(l-1)} + X_{i(j)}\hat{\delta}_j.$$

(b) *Computation of Variance Components*

The computation is based on the penalized log-likelihood

$$l_p(\theta|\eta^{(l)}; \delta_l) \quad = -\tfrac{1}{2}\sum_{i=1}^n \log(|V_i(\theta)|) + \sum_{i=1}^n (y_i - \eta_i^{(l)})^T V_i(\theta)^{-1}(y_i - \eta_i^{(l)})$$

$$-\tfrac{1}{2}(\hat{\delta}^{(l)})^T K_r \hat{\delta}^{(l)}.$$

Maximization yields $\hat{\theta}^{(l)}$.

This algorithm was inspired by the concept of an boosted Information-Criterion as developed in Bühlmann & Yu (2005) which they call sparse boosting. The objective of the selection of starting values is to select the most relevant variables in order to avoid huge variances for the error term in the beginning of the iteration. The computation of the starting values is very similar to the boosting algorithm itself. It starts with $\hat{\beta}^{(0)} = \hat{\alpha}_1^{(0)} = \cdots = \alpha_m^{(0)} = \eta_i^{(0)} = 0$ but the iterations for $l = 1, 2, \dots,$ in 2. are slightly modified. The first modification is that in (a) the covariance $V_i(\theta)$ is replaced by the simpler covariance matrix $\sigma^2 I$. Therefore, step (b) is replaced by the variance estimate $(\hat{\sigma}^2)^{(l)} = \frac{1}{N}\sum_{i=1}^n (y_i - \eta_i^{(l)})^T(y_i - \eta_i^{(l)})$. The iteration stops if $|(\hat{\sigma}^2)^{(l)} - (\hat{\sigma}^2)^{(l-1)}| < 10$. The variables that have been selected until this crude stopping criterion is met form a subset $\{s_1, \dots, s_{\tilde{m}}\}$. The initial estimates then are set to $(\hat{\beta}^{(0)}, \hat{\alpha}_{s_1}^{(0)}, \dots, \hat{\alpha}_{s_{\tilde{m}}}^{(0)})^T = (\sum_{i=1}^n (\tilde{X}_i^T \tilde{X}_i + \tilde{K}))^{-1} \sum_{i=1}^n \tilde{X}_i^T y_i$ with $\tilde{X}_i = [X_i, \Phi_{is_1}, \dots, \Phi_{is_{\tilde{m}}}]$ and $\tilde{K} = diag(0, \lambda K_{s_1}, \dots, \lambda K_{s_{\tilde{m}}})$. The other components are set to zero.

We chose componentwise boosting techniques because they turn out to be very stable in the high dimensional case where many potential predictors are under consideration. In this case, the procedure automatically selects the relevant variables and may be seen as a tool for variable selection with respect to unspecified smooth functions. In the case of few predictors, one may also use boosting techniques without the selection step by refitting the residuals for the full model with design matrix $[X_i \Phi_{i1} \dots \Phi_{im}]$.

3.5.3 Stopping Criteria and Selection in BoostMixed

In boosting, often cross-validation is used to find the appropriate complexity of the fitted model (e.g. Dettling & Bühlmann (2003)). In the present setting cross-validation turns out to be too time consuming to be recommended. An alternative is to use the effective degrees of freedom which are given by the trace of the hat matrix (compare Hastie & Tibshirani (1990)). In the following the hat matrix is derived.
For the derivation of the hat matrix the matrix representation of the mixed model is preferred (see (3.15))

$$y = X\beta + \Phi_1\alpha_1 + \ldots + \Phi_m\alpha_m + \epsilon^*,$$

$$\text{where} \quad \epsilon^* \sim N(0, V), \quad V(\theta) = Diag(V_1(\theta), \ldots, V_n(\theta)).$$

Since in one step only one component is refitted one has to consider the model for the residual refit of the rth component

$$residual = X_{.(r)}\delta_r,$$

$$\text{where} \quad X_{.(r)}^T = (X_{1(r)}^T \ldots X_{n(r)}^T), \quad X_{i(r)} = [X_i\ \Phi_{ir}], \quad \delta_r^T = (\beta^T, \alpha_r^T).$$

The refit in the lth step is given by

$$\begin{aligned} \hat{\delta}_r &= \left(X_{.(r)}^T V(\theta^{(l-1)})^{-1} X_{.(r)} + \lambda K_r\right)^{-1} X_{.(r)}^T V^{-1}(\theta^{(l-1)})(y - \eta^{(l-1)}) \quad (3.21) \\ &= M_r^{(l)}(y - \eta^{(l-1)}), \end{aligned}$$

where

$$M_r^{(l)} = \left(X_{.(r)}^T V(\theta^{(l-1)})^{-1} X_{.(r)} + \lambda K_r\right)^{-1} X_r^T V^{-1}(\theta^{(l-1)})$$

refers to the rth component in the lth refitting step. Then the corresponding fit has the form

$$\hat{\eta}_r^{(l)} = X_r\hat{\delta}_r = X_r M_r^{(l)}(y - \hat{\eta}^{(l-1)}) = H_r^{(l)}(y - \eta^{(l-1)}),$$

where

$$H_r^{(l)} = X_{.(r)} M_r^{(l)}.$$

Let now j_l denote the index of the variable that is selected in the lth boosting step and $H^{(l)} = H_{je}^{(l)}$ denote the resulting "hat" matrix of the refit. One obtains with starting matrix $H^{(0)}$

$$\eta^{(1)} = H^{(0)}y + H^{(1)}(y - \hat{\eta}^{(0)}) = (H^{(0)} + H^{(1)}(I - M^{(0)}))y$$

and more general

$$\hat{\eta}^{(l)} = G^{(l)}y,$$

where

$$G^{(l)} = \sum_{s=0}^{l} H^{(s)} \prod_{k=0}^{s-1} (I - H^{(k)})$$

is the global hat matrix after the lth step. It is sometimes useful to rewrite G as

$$G^{(l)} = I - \prod_{k=0}^{l} (I - H^{(k)})$$

(compare Bühlmann & Yu (2003)).
For the selection step one evaluates the hat matrices for candidates which for the rth component in the lth step have the form

$$G_r^{(l)} = G^{(l-1)} + H_r^{(l)} \prod_{k=0}^{l-1} (I - H^{(k)}).$$

Given the hat matrix $G_r^{(l)}$, complexity of the model may be determined by information criteria. When considering in the lth step the rth component one uses the criteria

$$\begin{aligned} AIC_r^{(l)} &= -2 \left\{ -\frac{1}{2} \sum_{i=1}^{n} log(V(\hat{\theta}^{(l-1)})) + \sum_{i=1}^{n} (y_i - \hat{\eta}^{(l-1)})^T V_i (\hat{\theta}^{(l-1)})^{-1} (y_i - \hat{\eta}^{(l-1)}) \right\} \\ &\quad + 2\, trace\, (G_r^{(l)}), \\ BIC_r^{(l)} &= -2 \left\{ -\frac{1}{2} \sum_{i=1}^{n} log(V(\hat{\theta}^{(l-1)})) + \sum_{i=1}^{n} (y_i - \hat{\eta}_i^{(l-1)}) (V_k(\theta)^{(l-1)})^{-1} (y_i - \eta_i^{(l-1)}) \right\} \\ &\quad + 2\, trace(\, G_r^{(l)}\,) log(n). \end{aligned}$$

In the rth step, one selects from $r \in \{1, \ldots, m\}$ the component that minimizes $AIC_r^{(l)}$(or $BIC_r^{(l)}$) and obtains $AIC^{(l)} = AIC_{j_l}^{(l)}$ if j_l is selected in the rth step. If $AIC_r^{(l)}$(or $BIC_r^{(l)}$) is larger than the previous criterion $AIC^{(l-1)}$ iteration stops. It should be noted that in contrast to common boosting procedures, the selection step reflects the complexity of the refitted model. In common componentwise boosting procedures (e.g. Bühlmann & Yu (2003)) one selects the component that maximally improves the fit and then evaluates if the fit including complexity of the model deteriorates. The proposed procedure selects the component in a way that the new lack-of-fit, including the augmented complexity, is minimized. In our simulations the suggested approach showed superior performance.

In the following, the initialization of the boosting algorithm is shortly sketched. The basic concept is to select few relevant variables in order to obtain stable estimates of variance components. Therefore for large λ (in our application $\lambda = 1000$), the total model is fitted

using the full design matrix $X_\Phi = [X, \Phi_1, \ldots, \Phi_m]$ and covariance matrix $V_i(\theta) = I$. Then in a stepwise way the variables are selected (usually up to 5) that yield the best fit. These yield the initial estimates $\hat{\beta}^{(0)}, \alpha_1^{(0)}, \ldots, \alpha_m^{(0)}$ and the initial hat matrix $G^{(0)}$.

3.5.4 Standard Errors

Approximate standard errors for the parameter β and the functions $\alpha_{(j)}(u) = \Phi_{(j)}(u)^T \alpha_j$ may be derived by considering the iterative refitting scheme. For the estimated parameter in the $l-th$ step $\delta^{(l)}$ one obtains

$$\hat{\delta}^{(l)} = \hat{\delta}^{(l-1)} + M^{(l)}(y - \hat{\eta}^{(l-1)})$$

where $M^{(l)}$ is a matrix that selects the components β and α_{j_l} which are updated in the $l-th$ step. It is given by

$$(M^{(l)})^T = \left(\ (M_{j_l,1}^{(l)})^T, 0, \ldots, (M_{j_l,2}^{(l)})^T, \ldots, 0 \ \right),$$

where $M_{j_l,1}, M_{j_l,2}$ denote the partitoning of $M_{j_l}^{(l)}$ into components that refer to β and α_{j_l} respectively, i.e.

$$M_{j_l}^{(l)} = \begin{pmatrix} M_{j_l,1} \\ M_{j_l,2} \end{pmatrix}.$$

One obtains for the refitting of δ with starting matrix $M^{(0)}$

$$\begin{aligned} \hat{\delta}^{(1)} &= M^{(0)}y + M^{(1)}(y - \hat{\eta}^{(0)}) \\ &= M^{(0)}y + M^{(1)}(I - H^{(0)})y, \end{aligned}$$

and more general

$$\hat{\delta}^{(l)} = D^{(l)}y,$$

where

$$D^{(l)} = \sum_{s=0}^{l} M^{(s)} \prod_{k=0}^{s-1} (I - H^{(k)}).$$

With L denoting the last refit one obtains with $\hat{\delta} = \hat{\delta}^{(L)}, D = D^{(L)}$, for the covariance of $\hat{\delta}$

$$\begin{aligned} cov(\hat{\delta}) &= D \ cov(y) D^T \\ &= D \ V(\theta) D^T. \end{aligned}$$

Approximate variances follow by using $\hat{\theta} = \hat{\theta}^{(L)}$ to approximate $V(\theta)$. Standard errors for β and $\alpha_{(j)}(u) = \Phi_{ij}^T \alpha_j$ are then easily derived since $\hat{\delta}^T = (\hat{\beta}^T, \hat{\alpha}_j^T)$.

In boosting, the crucial tuning parameter is the number of iterations. The smoothing parameter that is used in the algorithm should be chosen large to obtain a weak learner. The number of iterations increases for large λ. In order to limit the number of iterations we modified the algorithm slightly. If more than 1000 iterations are needed until the stopping criterion is met, then the algorithm is restarted with $\lambda/2$; the halving procedure is repeated if $\lambda/2$ also needs more than 1000 iterations.

3.5.5 Visualizing Variable Selection in Penalty Based Approaches

In spline methodology, the unknown functions $\alpha_{(j)}$ are often approximated by basis functions. A simple basis is known as the truncated power series basis of degree d, yielding

$$\alpha_{(j)}(u) \approx \gamma_0^{(j)} + \gamma_1^{(j)} u + \ldots \gamma_d^{(j)} u^d + \sum_{s=1}^{\tilde{M}} \alpha_s^{(j)} (u - k_s^{(j)})_+^d \ ,$$

where $k_1^{(j)} < \ldots < k_{\tilde{M}}^{(j)}$ are distinct knots.

If one uses

$$\alpha_{(j)}(u) = u\alpha_j$$

the underlying function is approximated by an linear term. So in this case the basis function approach encompasses only one knot and $\Phi_1^{(j)}(u) = u$ is the identity function. In the shrinkage theory the likelihood is penalized by $-(\alpha_j)^2$. So the corresponding penalty matrix for the boosted linear effects is an $m \times m$ identity matrix.

One can study the coefficients build-up in a similar way as in LASSO (see Tibshirani (1996)) since the linear model is a special case with one knot and identity as basis function. Using ridge penalty, the result is a linear mixed model with parametric main effects. So many variables may be included, but only few contain information on the response. Smooth effects can be compared by considering $\tilde{\alpha}_j = \int_0^1 |\alpha_{(j)}(u_j)| du_j, j = 1, \ldots, m$ in build-up graphics. For parametric effects, the variables are transformed to the interval $[0, 1]$ and centering the effects around zero yields $\tilde{\beta}_k = \frac{\beta_k}{2} * 0.25, k = 1, \ldots, p$. The integral corresponds to the area of the centered functions, see Figure 3.6.

Example 3.6 : Generalization of build-up graphics

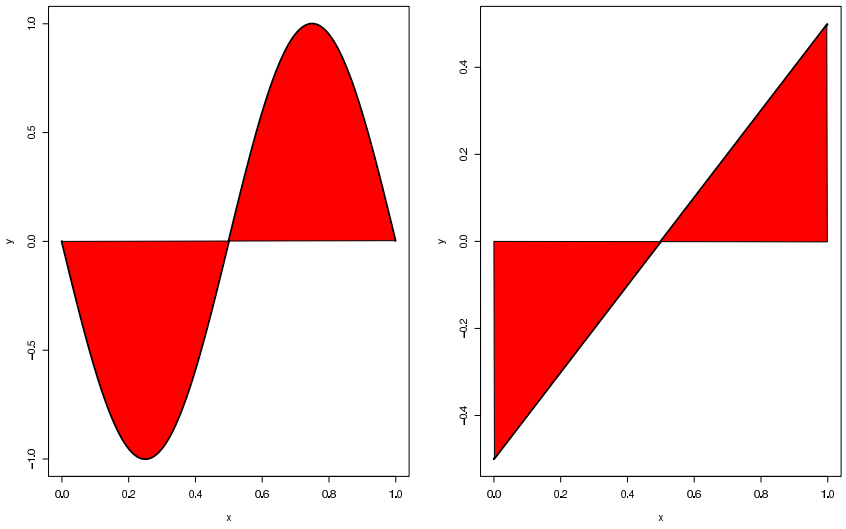


Figure 3.6: The red area is the integral for $|sin(u)|$ on the left side and for u on the right side. It measures the strength of influence on the response

To demonstrate the generalization of build-up graphics we used the underlying random intercept model

$$y_{it} = b_i + \sum_{j=1}^{19} \alpha_{(j)}(u_{it}) + \epsilon_{it}, i = 1, \ldots, 80, t = 1, \ldots, 5$$

with the smooth components given by

$$\begin{aligned} \alpha_{(i)}(u) &= \tfrac{5}{i} * sin(u) \quad && u \in [-3, 3], i = 1, \ldots, 5 \\ \alpha_{(i)}(u) &= 0 && u \in [-3, 3], i = 6, \ldots, 19. \end{aligned} \tag{3.22}$$

The variances for error term and random intercepts were taken to be $\sigma_\epsilon^2 = \sigma_b^2 = 2$. *Figure 3.7 show the build-up graphic for smooth effects. Figure 3.8 shows the true underlying functions* $(\alpha_{(1)}, \ldots, \alpha_{(6)})$ *and their corresponding estimates. For this study the smooth functions* $\alpha_{(i)}, i = 1, \ldots, 19$ *wer specified in the model. What is getting obvious is that the strength of functions are reflected in the build up graphics. The area under the curves may be interpreted as that part of the response which could be explained by these curves. It is also a measure for the importance of curves according the order of the estimated values* $\tilde{\alpha}_j, j = 1, \ldots, 19$.

□

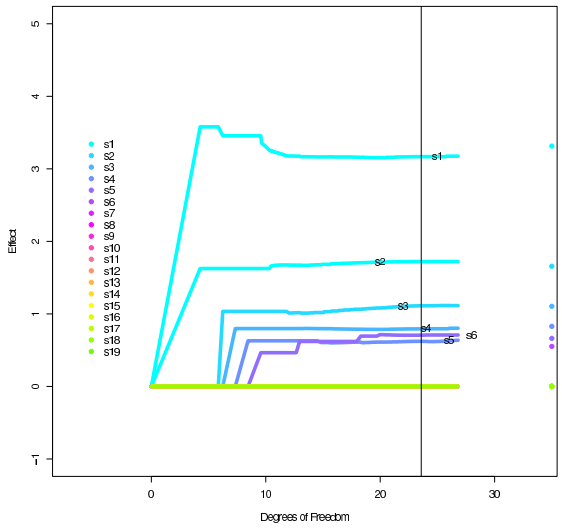


Figure 3.7: Build-up graphic for the smooth effects. On the right side one can see the true values $\int |\alpha_{(i)}(u)|du$ for the coefficients build ups.

3.6 Simulation Studies

3.6.1 BoostMixed vs. Mixed Model Approach

Study 1 and 2

We present part of a simulation study in which the performance of BoostMixed models is compared to alternative approaches. The underlying model is the random intercept model

$$y_{it} = b_i + \sum_{j=1}^{40} c * \alpha_{(j)}(u_{it}) + \epsilon_{it}, i = 1, \ldots, 80, t = 1, \ldots, 5$$

with the smooth components given by

$$\begin{aligned} \alpha_{(1)}(u) &= sin(u) \quad u \in [-3, 3], \\ \alpha_{(2)}(u) &= cos(u) \quad u \in [-2, 8], \\ \alpha_{(3)}(u) &= u^2 \qquad u \in [-3, 3], \\ \alpha(u) &= 0 \qquad\quad u \in [-3, 3], j = 4, \ldots, 40. \end{aligned} \tag{3.23}$$

The vectors $u_{it}^T = (u_{it1}, \ldots, u_{it40})$ have been drawn independently with components following a uniform distribution within the specified interval. For the covariates constant

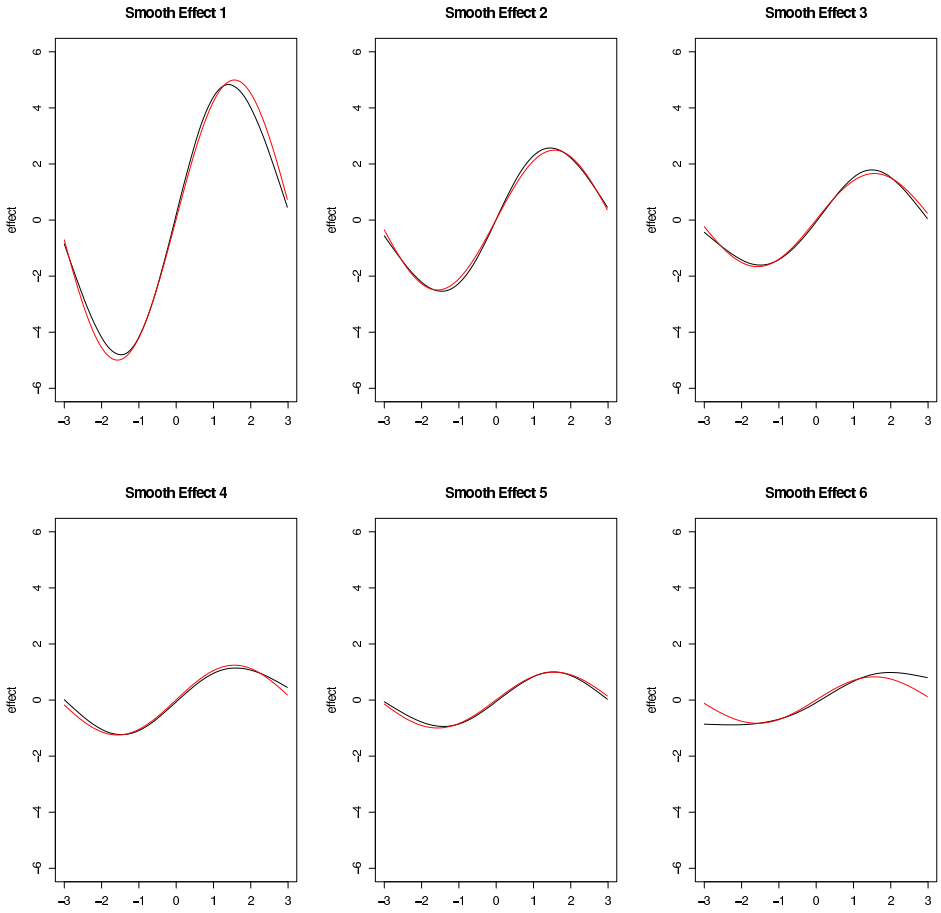


Figure 3.8: The red line are the estimated effects, the black lines are the true underlying functions.

correlation is assumed, i.e. $corr(u_{itr}, u_{its}) = \rho$. In study 1, ρ is set to $\rho = 0.1$. For study 2, ρ was chosen to be $\rho = 0.5$. The constant c determines the signal strength of the covariates. The random effect and the noise variable have been specified by $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 2$ and $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 2$. In the part of the study which is presented the number of observations has been chosen by $n = 80, T = 5$.

The fit of the model is based on B-splines of degree 3 with 15 equidistant knots. The performance of estimators is evaluated separately for the structural components and the variance. By averaging across 100 datasets we consider mean squared errors for $\eta, \sigma_b^2, \sigma_\varepsilon^2$ given by

$$\text{mse}_\eta = \sum_{i=1}^n \sum_{t=1}^T (\eta_{it} - \hat{\eta}_{it})^2, \hat{\eta}_{it} = x_{it}^T \hat{\beta}, \qquad \text{mse}_\beta = ||\beta - \hat{\beta}||^2,$$

$$\text{mse}_{\sigma_b^2} = ||\sigma_b^2 - \hat{\sigma}_b^2||^2, \qquad \text{mse}_{\sigma_\epsilon^2} = ||\sigma_\epsilon^2 - \hat{\sigma}_\epsilon^2||^2.$$

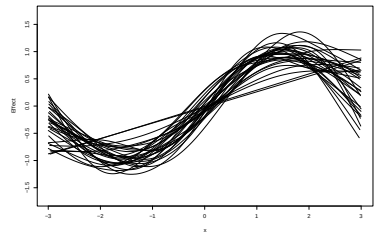
as well as the mean squared error for the smooth components

$$\text{mse}_\alpha = \sum_{i=1}^n \sum_{t=1}^{T_i} \sum_{j=1}^p (\alpha_{(j)}(u_{itj}) - \hat{\alpha}_{(j)}(u_{itj}))^2,$$

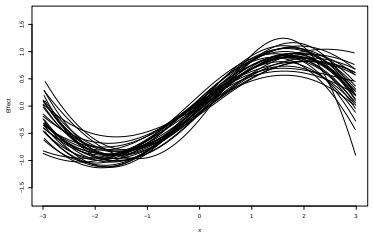
which corresponds to the estimation of parameters in linear mixed models.

For illustration, in Figure 3.9, the Mixed Model approach to smooth components (MM) from study 1 is compared with BoostMixed for 30 datasets. It is seen that both methods detect the underlying smooth functions fairly well. However, it is seen that the mixed model approach has higher variability. For example for some datasets the component $\alpha_{(1)}$ has been strongly oversmoothed yielding straight lines (rather than the sin function).

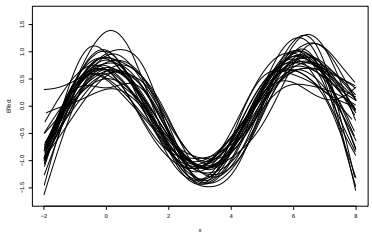
In Tables 3.1 and 3.2 the resulting mean squared errors are given for the low correlation case $(\rho = 0.1)$ (study 1) and the medium correlation case $(\rho = 0.5)$ (study 2). It is seen that for all components mean squared errors are smaller when BoostMixed is used. The difference is rather large for high dimensional predictors which include noisy covariates. But it should be noted that also in the case, where only the variables are included which carry information, the mean squared errors are still smaller when BoostMixed is used. For higher number of predictors (p>20), the Mixed Model fit did not work and therefore, no values are shown in Table 3.1 and 3.2. The strongest reduction in terms of mean squared error is found for the estimation of mse_η the effect becomes stronger with increasing signal c and parameters p, see for example $mse_\eta = 41.946$ for BoostMixed and $mse_\eta = 50.448$ for the additive model with $c = 1, p = 3$. In Figure 3.10 and 3.11 the mean squared errors are given for the pure information case (p=3) and the case that includes several noise variables (p=15).



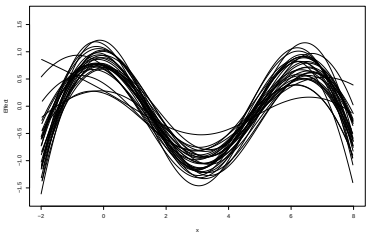
(a) $\hat{\alpha}_{(1)}$



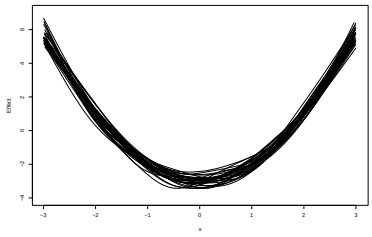
(b) $\hat{\alpha}_{(1)}$



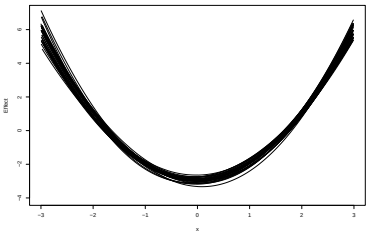
(c) $\hat{\alpha}_{(2)}$



(d) $\hat{\alpha}_{(2)}$



(e) $\hat{\alpha}_{(3)}$



(f) $\hat{\alpha}_{(3)}$

Figure 3.9: Study 1: Thirty functions computed with mixed model methods(left panels) and boosting (right panels) $(c = 1, p = 3)$

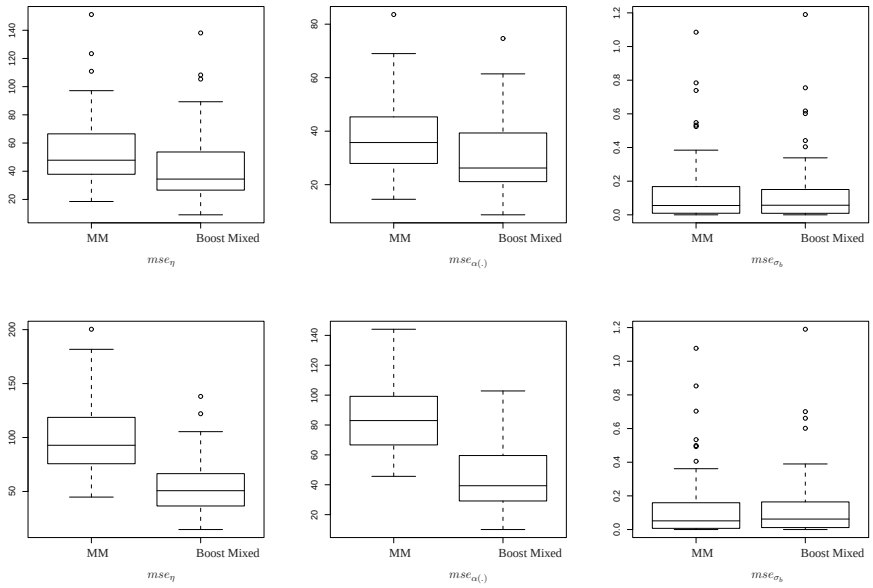


Figure 3.10: Boxplot for mse_η, mse_f and mse_{σ_b} additive simulation study with $c = 1$ and $p = 3$ (top) and $c = 1$ and $p = 15$ (bottom)

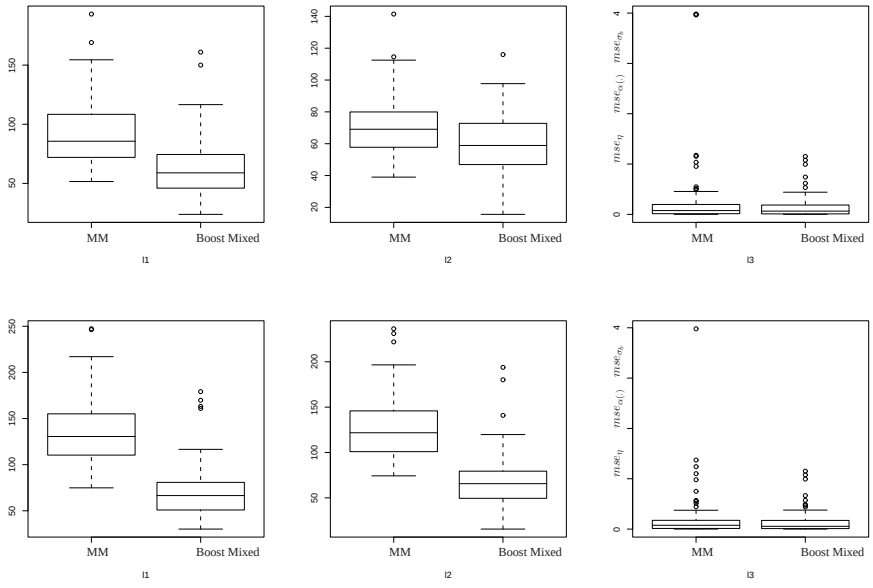


Figure 3.11: Boxplot for mse_η, mse_f and mse_{σ_b} additive simulation study with $c = 10$ and $p = 3$ (top) and $c = 1$ and $p = 15$ (bottom)

		MM						BoostMixed									
c	p	mse_η	mse_α	mse_{σ_b}	$\text{mse}_{\sigma_\epsilon}$	Steps	Time	mse_η	mse_α	mse_{σ_b}	$\text{mse}_{\sigma_\epsilon}$	Steps	Time	FalsePos	FalseNeg	Initial	Selected
0.5	3	45.791	37.701	0.026	0.117	13	0.09	42.079	36.867	0.026	0.115	9.9	0.0	0.0	0.1	2.0	2.9
0.5	6	55.721	48.399	0.030	0.117	18	0.41	48.666	45.112	0.028	0.114	10.2	0.0	0.4	0.1	2.0	3.3
0.5	15	88.005	85.470	0.031	0.129	25	7.03	62.501	62.270	0.029	0.114	9.7	0.1	0.9	0.2	2.0	3.7
0.5	25							73.134	74.790	0.030	0.116	9.8	0.1	1.2	0.3	2.0	3.9
1.0	3	50.448	37.422	0.024	0.126	8	0.06	41.946	31.226	0.026	0.119	19.7	0.0	0.0	0.0	2.0	3.0
1.0	6	60.520	48.547	0.024	0.120	15	0.33	42.773	32.237	0.026	0.120	19.7	0.1	0.0	0.0	2.0	3.0
1.0	15	92.705	85.021	0.028	0.120	21	6.05	46.662	36.725	0.029	0.120	20.0	0.2	0.2	0.0	2.0	3.2
1.0	25							50.440	41.102	0.028	0.118	20.2	0.3	0.3	0.0	2.0	3.3
5.0	3	71.243	60.651	0.032	0.187	12	0.08	53.399	47.592	0.031	0.181	144.6	0.4	0.0	0.0	1.9	3.0
5.0	6	82.051	72.296	0.031	0.185	14	0.32	55.396	49.947	0.031	0.182	146.9	0.4	0.1	0.0	1.9	3.1
5.0	15	116.472	113.781	0.036	0.190	20	5.87	57.510	52.545	0.032	0.182	145.2	2.3	0.2	0.0	1.9	3.2
5.0	25							58.533	53.910	0.034	0.182	145.5	3.4	0.2	0.0	1.9	3.2
10.0	3	88.045	71.694	0.027	0.264	14	0.10	62.981	59.701	0.029	0.139	495.6	1.1	0.0	0.0	3.0	3.0
10.0	6	98.669	84.396	0.026	0.226	17	0.40	62.981	59.701	0.029	0.139	495.6	2.6	0.0	0.0	3.0	3.0
10.0	15	132.549	125.730	0.033	0.239	24	7.11	65.726	62.807	0.033	0.139	492.1	6.7	0.1	0.0	3.0	3.1
10.0	25							66.588	63.895	0.033	0.139	490.9	12.0	0.1	0.0	3.0	3.1

Table 3.1: Study 1: Comparison between additive mixed model fit and BoostMixed ($\rho = 0.1$).

		MM						BoostMixed									
c	p	mse_η	mse_α	mse_{σ_b}	$\text{mse}_{\sigma_\epsilon}$	Steps	Time	mse_η	mse_α	mse_{σ_b}	$\text{mse}_{\sigma_\epsilon}$	Steps	Time	FalsePos	FalseNeg	Initial	Selected
0.5	3	46.503	34.107	0.022	0.133	13	0.09	45.416	36.576	0.026	0.136	9.9	0.0	0.0	0.1	2.0	2.9
0.5	6	57.421	48.626	0.024	0.133	18	0.42	50.530	43.280	0.028	0.139	10.3	0.0	0.3	0.1	2.0	3.2
0.5	15	90.615	92.066	0.029	0.135	28	8.30	64.707	61.314	0.032	0.140	11.0	0.1	0.8	0.2	2.0	3.7
0.5	25							72.285	70.857	0.035	0.141	11.5	0.2	1.1	0.2	2.0	3.9
1.0	3	49.449	40.515	0.033	0.146	9	0.06	40.716	34.440	0.035	0.145	17.4	0.0	0.0	0.0	2.0	3.0
1.0	6	60.771	54.728	0.037	0.148	16	0.37	42.105	36.107	0.037	0.143	17.6	0.1	0.1	0.0	2.0	3.0
1.0	15	93.651	97.541	0.038	0.151	21	6.41	43.327	37.663	0.037	0.144	17.7	0.2	0.1	0.0	2.0	3.1
1.0	25							46.404	41.527	0.036	0.145	17.9	0.4	0.2	0.0	2.0	3.2
5.0	3	72.155	62.797	0.023	0.153	12	0.09	53.174	49.862	0.025	0.153	109.6	0.3	0.0	0.0	3.0	3.0
5.0	6	82.856	77.115	0.025	0.157	14	0.33	53.663	50.515	0.026	0.154	109.5	0.6	0.0	0.0	3.0	3.0
5.0	15	114.390	118.645	0.028	0.156	18	5.25	54.918	51.990	0.026	0.154	109.4	1.5	0.1	0.0	3.0	3.1
5.0	25							56.471	53.814	0.027	0.154	109.1	2.6	0.1	0.0	3.0	3.1
10.0	3	93.000	77.369	0.029	0.230	14	0.09	68.369	63.423	0.030	0.184	430.2	1.1	0.0	0.0	3.0	3.0
10.0	6	103.896	92.147	0.028	0.225	15	0.34	69.027	64.432	0.030	0.184	430.0	2.2	0.0	0.0	3.0	3.0
10.0	15							70.142	65.935	0.031	0.180	428.9	5.7	0.1	0.0	3.0	3.1
10.0	25							73.504	70.497	0.031	0.181	427.1	7.9	0.2	0.0	3.0	3.2

Table 3.2: Study 2: Comparison between additive mixed model fit and BoostMixed ($\rho = 0.5$).

For a more extensive analysis of BoostMixed five further simulation studies with same setting (3.23) (except study 7), but different values for T, n, ρ were made. In all studies 100 datasets were generated

Study 3 - medium cluster
The random effect and the noise variable have been specified by $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 3$ and $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 3$. In the part of the study which is presented the number of observations has been chosen by $n = 40, T = 5$. Pairwise correlation was taken to be $corr(u_{itr}, u_{its}) = 0.1$. Details can be found in Table C.3 and Figure C.3.

Study 4 - big clusters
The random effect and the noise variable have been specified by $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 3$ and $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 6$. In the part of the study which is presented the number of observations has been chosen by $n = 20, T = 10$. Pairwise correlation was taken to be $corr(u_{itr}, u_{its}) = 0.1$. Details can be found in Table C.4 and Figure C.4.

Study 5 - small clusters
The random effect and the noise variable have been specified by $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 3$ and $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 6$. In the part of the study which is presented the number of observations has been chosen by $n = 100, T = 2$. Pairwise correlation was taken to be $corr(u_{itr}, u_{its}) = 0.1$. Details can be found in Table C.5 and Figure C.5.

Study 6 - big dataset
The random effect and the noise variable have been specified by $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 9$ and $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 12$. In the part of the study which is presented the number of observations has been chosen by $n = 250, T = 20$. Pairwise correlation was taken to be $corr(u_{itr}, u_{its}) = 0.1$. Details can be found in Table C.6 and Figure C.6.

Study 7 - many additive covariates
The random effect and the noise variable have been specified by $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 2$ and $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 2$. In the part of the study which is presented the number of observations has been chosen by $n = 40, T = 5$. Pairwise correlation was taken to be $corr(u_{itr}, u_{its}) = 0.1$. The additive term have functions $\alpha_{(1)}(u) = \frac{c}{s} in(u), u \in [-3, 3]$, $\alpha_{(2)}(u) = \frac{c}{2} cos(u), u \in [-2, 8]$, $\alpha_{(3)}(u) = \frac{c}{3} u^2, u \in [-3, 3]$,$\alpha_{(4)}(u) = \frac{c}{4} sin(u), u \in [-3, 3]$, $\alpha_{(5)}(u) = \frac{c}{5} cos(u), u \in [-2, 8]$, $\alpha_{(6)}(u) = \frac{c}{6} u^2, u \in [-3, 3]$. The other functions are set to $\alpha_{(j)}(u) = 0, u \in [-3, 3], j = 7, \ldots, 40$. Details can be found in Table C.7.

If one wants to summarize the results of the studies BoostMixed seems to be a powerful competitor to the mixed model approach. In studies 3 to 5 which are simulation studies with only 200 observations in total BoostMixed delivers comparable, sometimes worse MSE_η than the mixed model approach for small signals (c=0.5). Especially in these

cases the selection aspect in BoostMixed is important since some relevant variables were not selected which downgrades the MSE_η. For large signals BoostMixed is in most cases superior to the mixed model approach. The mixed model approach seems to be more sensitive to higher signals for the influence of covariates. One may see a difference of the performance in datasets with small and huge clusters. Therefore in study 3 one can see worse selection for small signals and good results for huge signals. Instead, in study 4 the differences in the MSE_η are not as noticeable as in study 3, the MSE_{σ_ϵ} show much better results for BoostMixed. If one switches now to studies 1 and 2 which have 400 observations in total, the efficiency of selecting relevant variables is improved for small signals which is reflected in the comparable MSE_η for $c = 0.5$. Also for correlated data the results did not change. The difference of both methods disappear using large datasets as in study 6. In this case all relevant variables were selected. BoostMixed shows only slight better results for huge signals. Study 7 is a little bit different from the other studies, since this study has more relevant covariates and this study encompasses a forward selection procedure. In this sense, BoostMixed is compared to the mixed model approach with all covariates (MM) and to mixed model approach with an integrated forward selection (forward). It is quite similar to the BoostMixed algorithm since one starts with the intercept model. In every step all remaining covariates are fitted separately. The covariate characterized by the best improvement of the BIC-Criterion is taken into the model and seen as relevant. The selection is stopped if the complexity criterion can not improved any more. Compared to the forward selection procedure BoostMixed selects more relevant variables. On the other side BoostMixed delivers slightly bad results in the MSE_η compared to the mixed model approach. The time and computation complexity is getting tremendous if putting many covariates (p>20) in the forward selection procedure. For small covariates BoostMixed is a very fast selection strategy compared to the forward selection procedure.

Simulation studies for linear effects with a short discussion (Study 9 - Study 14) can be found in Appendix C.3. The results of the common linear models are compared to the boosted versions. Details on the underlying structure and the results are in Appendix C.3.

3.6.2 Pointwise Confidence Band for BoostMixed

In the following the focus is to get reliable confidence bands for smooth components. In this section estimated confidence bands are compared to the empirical confidence bands, that were computed from 250 datasets.

The underlying model is the random intercept model

$$y_{it} = b_i + \sum_{j=1}^{40} c * \alpha_{(j)}(u_{it}) + \epsilon_{it}, i = 1, \ldots, 80, t = 1, \ldots, 5$$

with a setting of covariates as described in (3.23).

The vectors $u_{it}^T = (u_{it1}, \ldots, u_{it3})$ have been drawn independently with components following a uniform distribution within the specified interval. For the covariates constant correlation is assumed, i.e. $corr(u_{itr}, u_{its}) = 0.1$. The constant c determines the signal strength of the covariates. The random effect and the noise variable have been specified by $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 2$ and $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 2$. On the presented study the number of observations has been chosen by $n = 80, T = 5$ and $n = 40, T = 5$. The smoothing parameter was fixed to $\lambda = 1000$.

The pointwise confidence intervals were computed using the covariance $\hat{cov}(\hat{\delta})$. Since $\delta^T = (\beta, \alpha_1^T, \ldots, \alpha_m^T)$ one can obtain $\hat{cov}(\hat{\delta}_j)$ by a decomposition of $\hat{cov}(\hat{\delta})$. So easily the pointwise confidence intervals for component j can be computed by $cov(X_j\hat{\delta}_j)$. Taking the diagonal elements and multiplying the square root together with the 0.975 percent and 0.025 percent quantile of the normal distribution on the estimates $\Phi_j\hat{\alpha}$ delivers the upper and lower 0.95 pointwise estimated confidence bands.

The empirical 0.95 percent pointwise confidence intervals were computed by getting the empirical 0.975 and 0.025 percent quantiles of all estimated functions. As one can see in Figures 3.12 and 3.6.2 the estimated pointwise confidence intervals are an good approximation to the empirical confidence bounds.

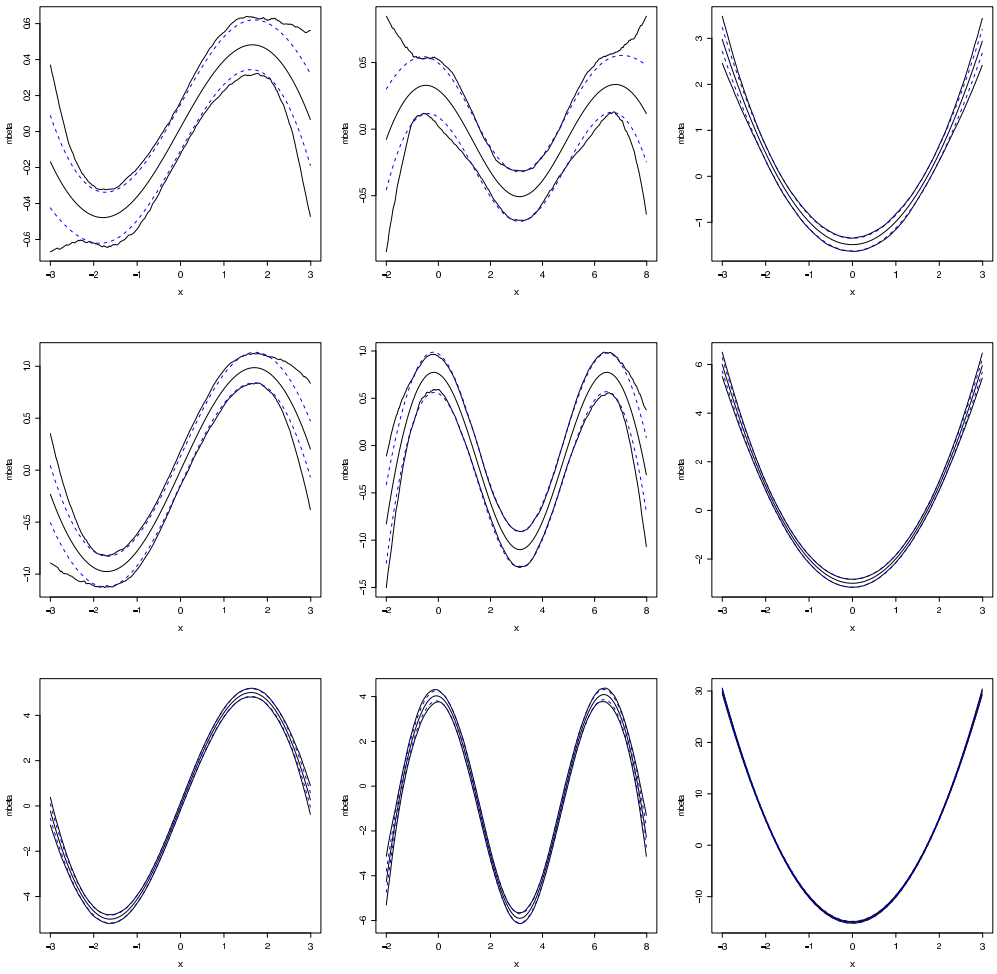


Figure 3.12: 0.95 percent pointwise confidence bands. The blue one are the averaged estimated confidence bands. The solid in the middle is the averaged smooth component. Upper and lower solid lines are the empirical pointwise confidence bands. The upper tree components are for signal $c = 0.5$, the middle components are for $c = 1.0$ and the bottom functions are for signal $c = 5$. There are $n = 80$ clusters.

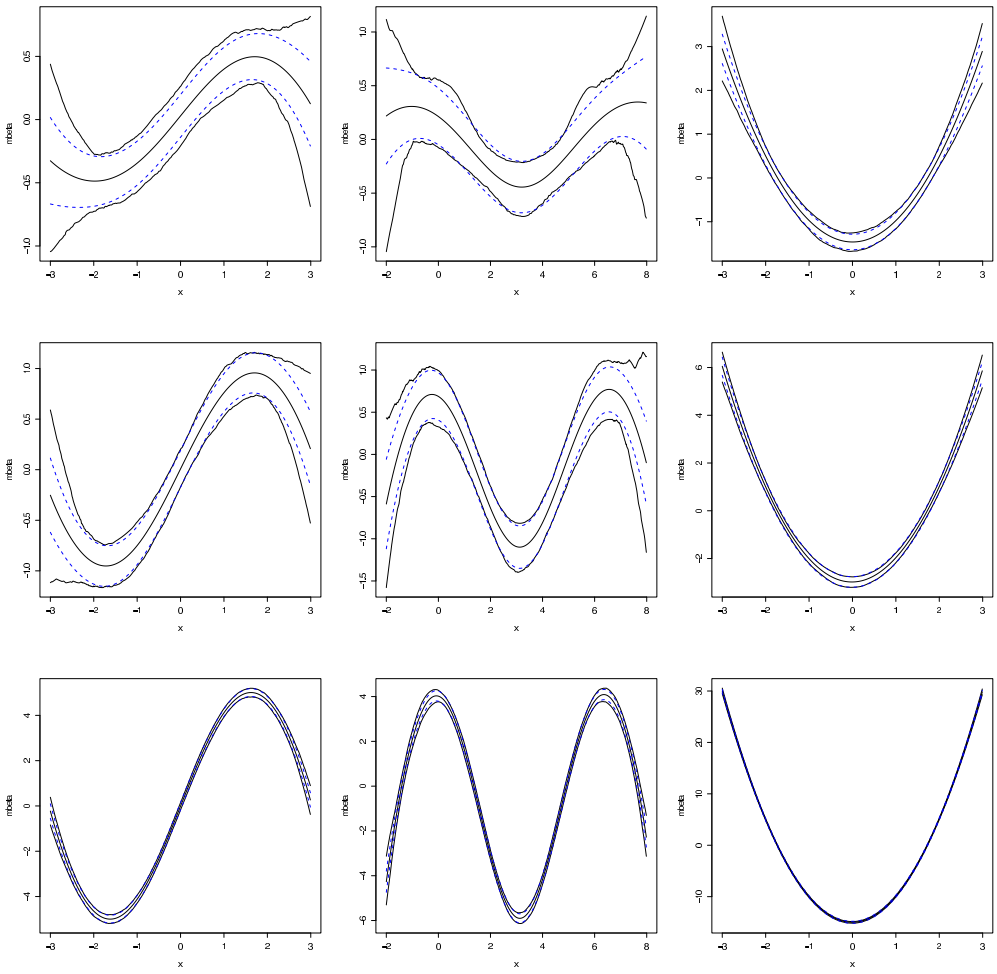


Figure 3.13: 0.95 percent pointwise confidence bands. The blue one are the averaged estimated confidence bands. The solid in the middle is the averaged smooth component. Upper and lower solid lines are the empirical pointwise confidence bands. The upper tree components are for signal $c = 0.5$, the middle components are for $c = 1.0$ and the bottom functions are for signal $c = 5$. There are $n = 40$ clusters.

3.6.3 Choosing an Appropriate Smoothing Parameter and an Appropriate Selection Criterion

The focus is in the following on getting reliable confidence bands for smooth components. In this section estimated confidence bands are compared to the empirical confidence bands, that were computed from 250 datasets.

Study 8 The underlying model is the random intercept model

$$y_{it} = b_i + \sum_{j=1}^{40} c * \alpha_{(j)}(u_{it}) + \epsilon_{it}, i = 1, \ldots, 80, t = 1, \ldots, 5$$

with a setting of covariates as described in (3.23).

The vectors $u_{it}^T = (u_{it1}, \ldots, u_{it3})$ have been drawn independently with components following a uniform distribution within the specified interval. For the covariates constant correlation is assumed, i.e. $corr(u_{itr}, u_{its}) = 0.1$. The constant c determines the signal strength of the covariates. The random effect and the noise variable have been specified by $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 2$ and $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 2$. In the part of the study which is presented the number of observations has been chosen by $n = 80, T = 5$ and $n = 40, T = 5$. c was set to $c = 0.5$.

The smoothing parameters were chosen on a grid from $[0, 2400]$ with steps of 50 for 3,5,15, and 25 smooth covariates. Then the distributions of the mean squared errors is considered and compared to the distribution of the mixed model approach. Selection and stopping criterion were chosen to be BIC or AIC.

As Figure 3.14 demonstrates the influence of taking λ different from 1000 is marginal. One has only to choose a lambda that is appropriate large. We found 1000 to be a good choice. Figure 3.15 includes also the selection aspect.

For detailed graphics for different signal strengths $c = 0.5, 1$ and $c = 5$, see Figures C.7, C.8 and C.9 for BIC, Figures C.10, C.11 and C.12 for AIC.

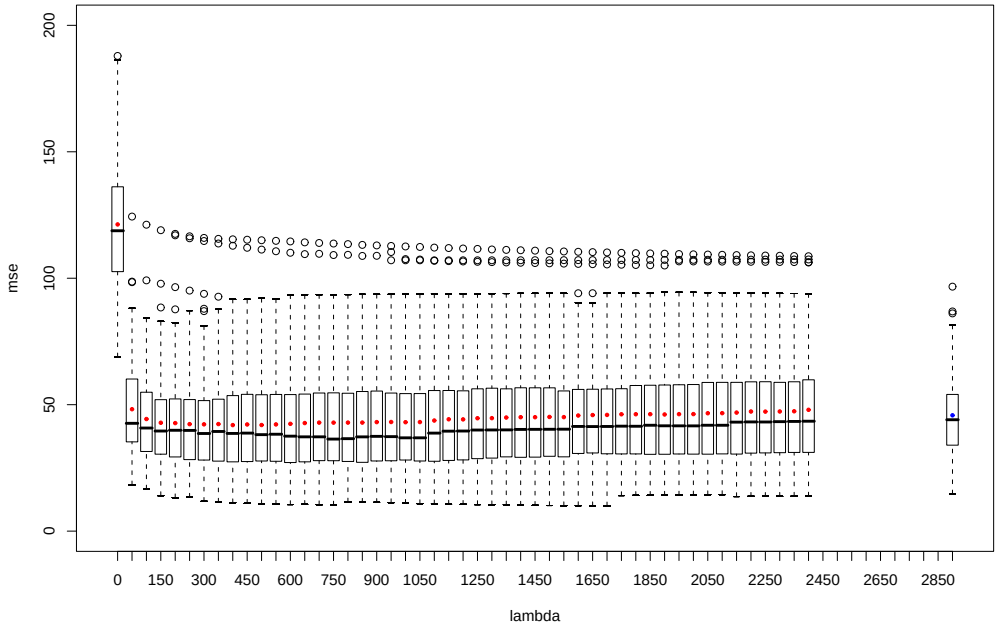
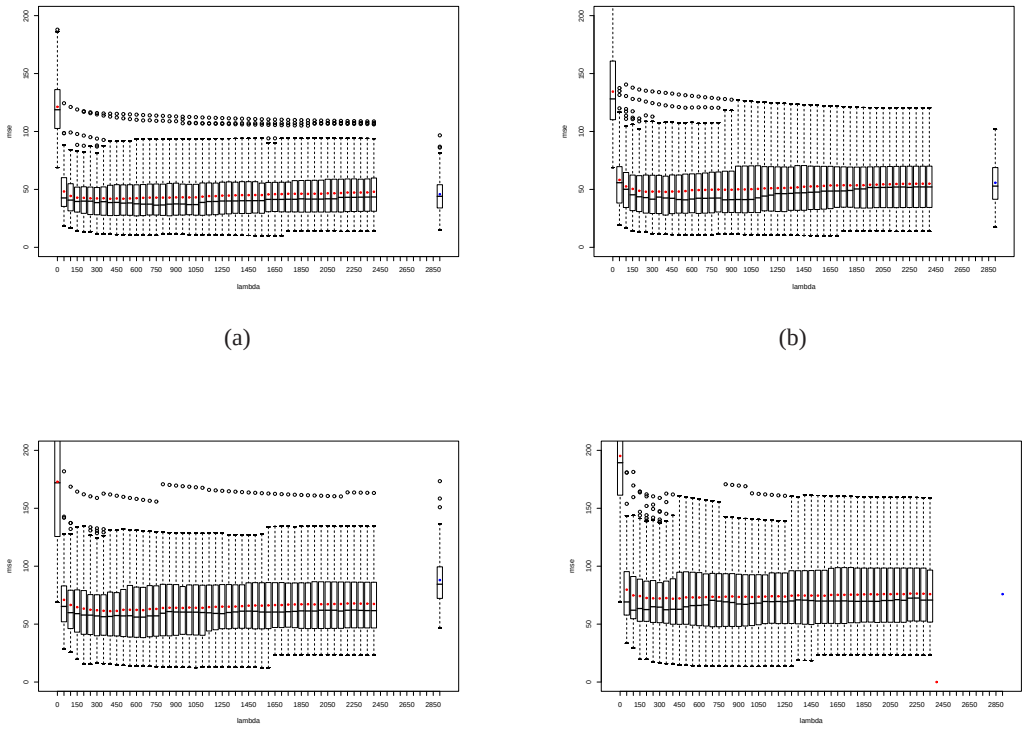


Figure 3.14: The errors for 3 smooth effects in the model are presented by boxplots. BIC was used as selection and stopping criterion. The red points are the means for the MSEs depending on different lambdas. On the right side the distribution of the MSEs of the mixed model approch is plotted. The blue point is the mean of the MSEs of the mixed model approach. c was chosen to be $c = 0.5$



(a) (b)

(c) (d)

Figure 3.15: The distributions of the mean squared errors for different counts of smooth effects in the model are presented by boxplots. BIC was used as selection and stopping criterion. The red points are the means for the MSEs depending on different λ. On the right side the distribution of the MSEs of the mixed model approach is plotted. The blue point is the mean of the MSEs of the mixed model approach. (a) 3 smooth effects used (b) 5 smooth effects used (c) 15 smooth effects used and (d) 25 smooth effects used. c was chosen to be $c = 0.5$

3.6.4 Surface-Smoothing

As described in Hämmerlin & Hoffmann (1992) and Dierckx (1993) the concept of one-dimensional splines can be extended to d-dimensional splines by using tensor products on the spline basis $\mathcal{B}_1, \ldots, \mathcal{B}_d$.

In this context only the case $d = 2$ is considered. Therefore a spline basis is needed which can be derived with the elementwise Kronecker product $\odot$ which is defined for a $N \times M_1$ matrix A and a $N \times M_2$ matrix B.

$c_{(i)}$ is a vector of the i-th row of C, $a_{(i)}$ is the i-th row of A and $b_{(i)}$ is the i-th row of B.

The elementwise Kronecker product can be described

$$C := A \odot B \text{ with } c_{(i)} = a_{(i)}^T \otimes b_{(i)}^T.$$

So one can write

$$\phi_{(1,2)}(u_{(i)1}, u_{(i)2}) = \phi_{(1)} \odot \phi_{(2)} := \phi_{(1)}(u_{(i)1}) \odot \phi_{(2)}(u_{(i)2})$$

with $\phi_{(1)}(u_{(i)1})$ being the basis functions for the covariate $u_{(i)1}$ (first covariate, i-th measurement) and $\phi_{(2)}(u_{(i)1})$. The resulting spline basis is

$$\mathcal{B} := \{\phi_{(1,2)}^{(1)}, \ldots, \phi_{(1,2)}^{(M_1 * M_2)}\}.$$

Instead of using the $\odot$ operator, the Kronecker product $\otimes$ delivers the same result. The $\odot$ product is especially useful for matrices $\Phi_{(1)}$ and $\Phi_{(2)}$, where the products are computed row-wise.

So any interaction $\alpha(u_{(i)1}, u_{(i)2})$ between two covariates $u_{(i)1}$ and $u_{(i)2}$ can be approximated by splines

$$\alpha(u_{(i)1}, u_{(i)2}) \approx \phi_{(1,2)}(u_{(i)1}, u_{(i)2})^T \alpha$$

with ϕ consisting of columns $\phi_{(1,2)}^{(1)}, \ldots, \phi_{(1,2)}^{(M_1 * M_2)}$ and α as a vector of coefficients with length $M_1 * M_2$.

The only difficulty is to get an corresponding penalty matrix that penalizes the differences of adjacent basis functions. In 2-dimensional settings the definition of adjacent basis functions is not unique. The adjacent basis functions may be seen as the one on the main-axis (4 neighbors) or all surrounding basis functions (8 neighbors). For details on the construction of tensor splines see Marx & Eilers (2005) and Eilers, Currie & Durban (2006)

Four neighbors For the first case the penalty matrix is easily obtained for $M_1 = M_2$ by

$$D^{(1)} = D_{(M_1-d)\times M_1} \otimes I_{(M_1)} \text{ and } D^{(2)} = I_{(M_1)} \otimes D_{(M_1-d)\times M_1}.$$

The penalty term is then

$$-\lambda_{(1)}\alpha^T K^{(1)}\alpha - \lambda_{(2)}\alpha^T K^{(2)}\alpha = -\alpha^T \begin{bmatrix} \lambda_{(1)}K^{(1)} & 0 \\ 0 & \lambda_{(1)}K^{(1)} \end{bmatrix} \alpha = -\alpha^T K \alpha$$

with $K^{(1)} = (D^{(1)})^T D^{(1)}$ and $K^{(2)} = (D^{(2)})^T D^{(2)}$.

Eight neighbors Penalizing with more than four neighbors is difficult to derive. Still $M_1 = M_2$ is assumed. Therefore a location matrix L ($M_1 \times M_1$ matrix) is needed with $L_{i,j} = i * M_1 + j$. Next necessary item is vector $p^T = \triangle^d I_{(d+1)}$, where $\triangle^d$ is the d-dimensional difference operator.

So one can penalize the diagonal differences with $((M_1 - d) * M_1) \times (M_1 * M_1)$ matrices $D^{(3)}$ and $D^{(4)}$ by recursion in $k = 1, \ldots, d+1$, $i = 1, \ldots, M_1 - d$ and $j = 1, \ldots, M_1 - d$

$$D^{(3)} = D^{(3)}(i,j,k) = D_{L_{i,j},L_{i+k,j+k}} = p_k$$

and

$$D^{(4)} = D^{(4)}(i,j,k) = D_{L_{i,j},L_{i+d+2-k,j+k}} = p_k,$$

where p_k is the k-th element of p^T

The penalty term is then

$$\begin{aligned} &-\lambda_{(1)}\alpha^T K^{(1)}\alpha - \lambda_{(2)}\alpha^T K^{(2)}\alpha - \lambda_{(3)}\alpha^T K^{(3)}\alpha - \lambda_{(4)}\alpha^T K^{(2)}\alpha \\ &= -\alpha^T K \alpha \end{aligned}$$

with $K^{(3)} = (D^{(3)})^T D^{(3)}$ and $K^{(4)} = (D^{(4)})^T D^{(4)}$.

In the simulation the underlying model is an random intercept model with

$$y_{it} = \sum_{j=1}^{5} \sum_{l=j+1}^{6} \alpha_{(j,l)}(u_{itj}, u_{itl}) + b_{i0} + \epsilon_{it} \tag{3.24}$$

where ϵ_{it} is independent $N(0, \sigma^2)$ with $\sigma^2 = 0.5$, b_{i0} is independent $N(0, \sigma_b^2)$ with $\sigma_b^2 = 1.0$. $T = 5$ and $i = 4000$.

$\alpha_{(1,2)}(u_{it1}, u_{it2})$ is the density function for the two-dimensional normal distribution with correlation $\rho = 0.5$. $\alpha_{(3,4)}(u_{it3}, u_{it4}) = sin(u_{it3}) * sin(u_{it3})$ and $\alpha_{(5,6)}(u_{it5}, u_{it6}) =$

$exp(u_{it5}) * u_{it6}$. All covariates were drawn uniformly from the interval $[-3, 3]$. All other interactions of covariates have influence zero on the response. 100 datasets of the model (3.24) were generated. Comparisons to the R-function *gamm* from the R-Package *mgcv* (Version 1.3.12) were tried. Since computations did not lead to stable estimators, only the relevant effects $\alpha_{(1,2)}(u_{it1}, u_{it2})$, $\alpha_{(3,4)}(u_{it3}, u_{it4})$, $\alpha_{(5,6)}(u_{it5}, u_{it6})$ was specified in estimation. For the simulation study only the main axes were penalized (four neighbors). Estimates for one dataset is given in Figure 3.16 and 3.17.

The result of the study is given in Table 3.3.

	MM				BoostMixed				
c	MSE_η	MSE_{σ_b}	MSE_{σ_ϵ}	Steps	MSE_η	MSE_{σ_b}	MSE_{σ_ϵ}	Selected	Steps
0.5	50.605	0.133	0.023	16.1	44.400	0.133	0.025	2.9	27.9
1	53.324	0.147	0.034	11.3	39.049	0.147	0.034	3.0	55.2
5	76.088	0.155	0.024	12.8	52.205	0.155	0.025	3.0	385.0

Table 3.3: Comparison between additive mixed model fit and BoostMixed.

One can see in Table 3.3 that the results are quite comparable. BoostMixed does perform better than the mixed model approach in the MSE_η. The mean squared errors for the random effects variance are nearly the same. The mean squared error for the error component is sometimes larger using BoostMixed.

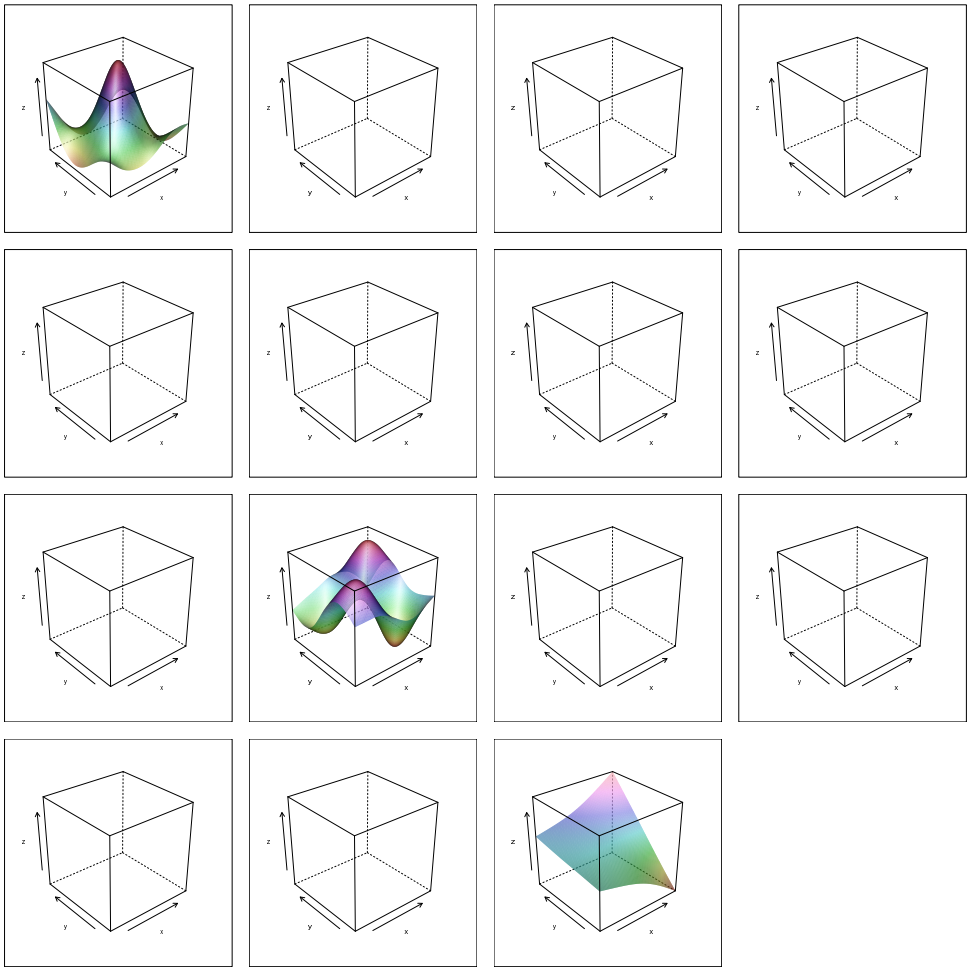


Figure 3.16: Surfaceplot for smoothed Interactions for 6 covariates for one selected dataset

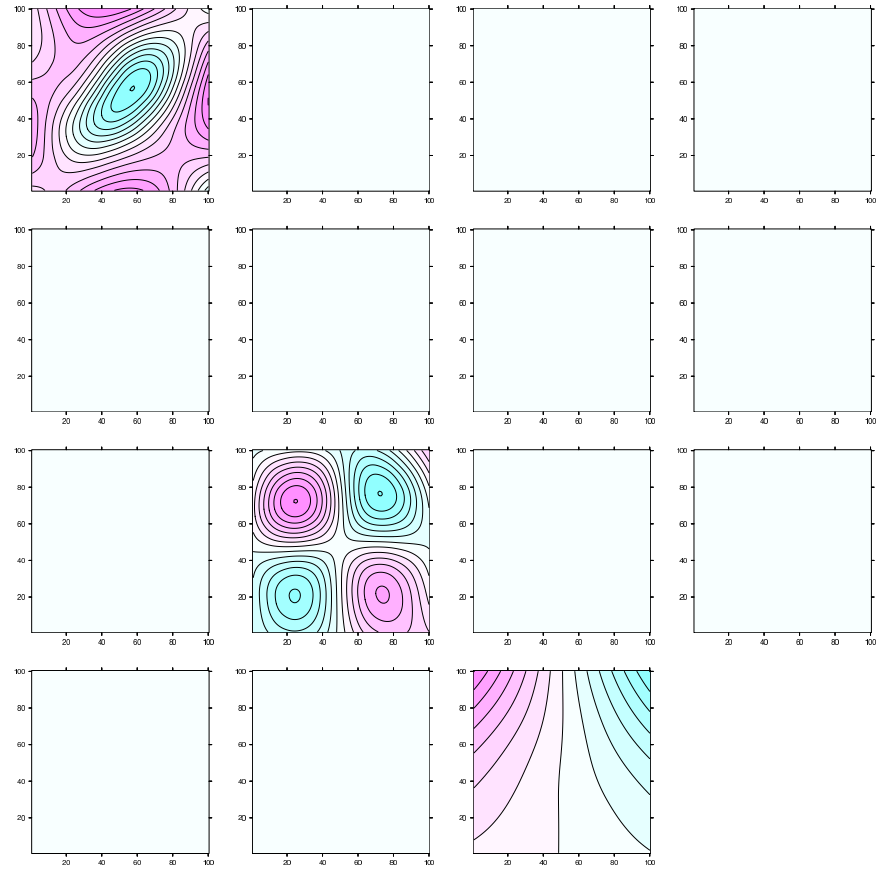


Figure 3.17: Levelplot for smoothed interactions for 6 covariates for one selected dataset

3.7 Application

3.7.1 CD4 data

Zeger & Diggle (1994) motivate extensively the interest in the typical time course of CD4 cell decay and the variability across subjects. Since the forms of the effects is not known, time since seroconversion, age and the mental illness score may be considered as unspecified additive effects. Figure 3.18 shows the smooth effect of time on CD4 cell decay for a random intercept model together with the data, Figure 3.19 shows the observations for three men with differing number of observed time points (dashed lines) and the fitted curves for individual time decay.

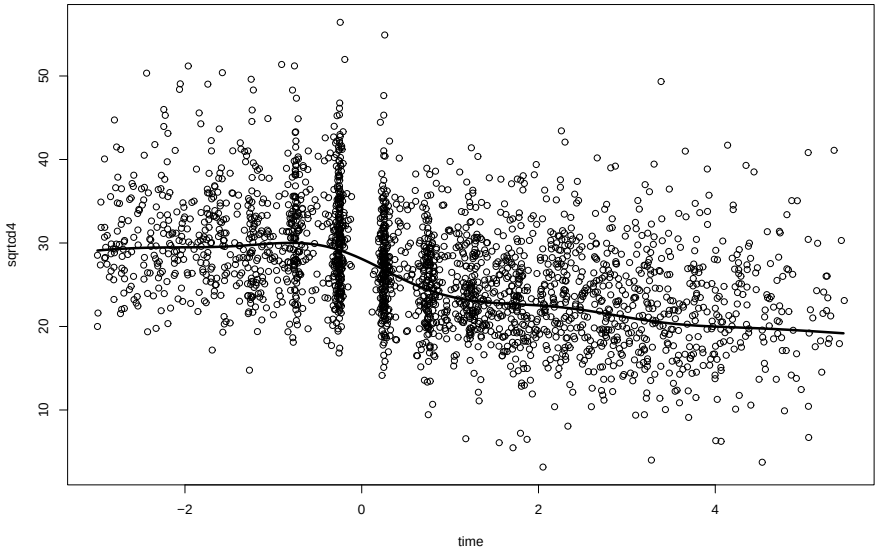


Figure 3.18: Smoothed time effect on the CD4 cell from Multicenter AIDS Cohort Study (MACS)

For the AIDS Cohort Study MACS we considered the semi-parametric mixed model from Section 1

$$y_{it} = \mu_{it}^{par} + \mu_{it}^{add} + b_{it} + \epsilon_{it},$$

where y_{it} denotes the square root CD4 counts of cells for subject i on measurement t (taken at irregular time intervals). The parametric and nonparametric term are given by

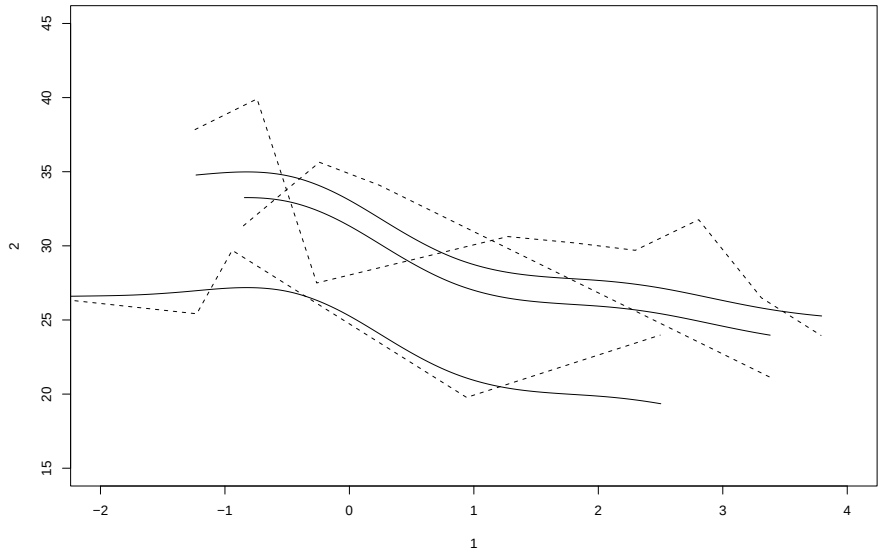


Figure 3.19: Smoothed time effect on the CD4 cell from Multicenter AIDS Cohort Study (MACS) and the decay of CD4 cells of 3 members of the study over time

$$\begin{aligned} \mu_i^{\text{par}} &= \beta_0 + drugs_i\beta_D + partners_i\beta_P, \\ \mu_{it}^{\text{add}} &= \alpha_T(time) + \alpha_A(age_i) + \alpha_C(cesd). \end{aligned}$$

where $cesd$ is a mental illness score. The square root transformation has been used since the original CD4 cell number varies over a wide range. The estimated effect of time was modelled smoothly with the resulting curve given in Figure 3.18. This smooth curve can be compared to the results of Zeger & Diggle (1994) who applied generalized estimation equations. In Figure 3.20 the smooth effects of age, the mental illness score and time are given. It is seen that there is a slight increase in CD4 cells for increasing age and a decease with higher values of the mental illness score. Table 3.4 shows the estimates for the parameters. Comparison between BoostMixed and the mixed model approach shows that the estimates are well comparable.

	BoostMixed		Mixed Model	
Intercept	24.6121	(0.294)	24.8233	(0.286)
Drugs	0.5211	(0.279)	0.5473	(0.292)
partners	0.0633	(0.049)	0.0595	(0.034)
σ_ϵ	4.2531	-	4.26138	-
σ_b	4.3870	-	4.43180	-

Table 3.4: Estimates for the AIDS Cohort Study MACS with BoostMixed and mixed model approach (standard deviations given in brackets)

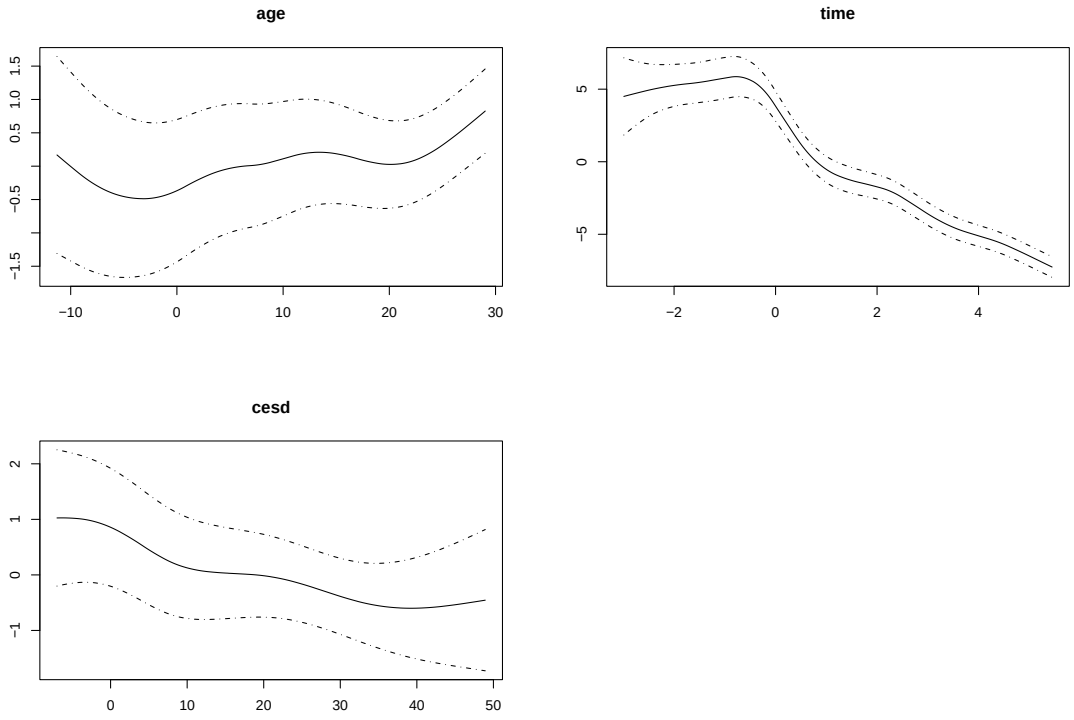


Figure 3.20: Estimated effect of age, the illness score cesd and time based on BoostMixed

Chapter 4

Extending Semi-Structured Mixed Models to incorporate Cluster-Specific Splines

The semiparametric additive model (3.12) allows for additive effects of covariates, including multivariate random effects. For example random slopes for linear terms are already included. Setting $z_{it} = x_{it}$ model (3.12) is a random slope model

$$y_{it} = \sum_{j=1}^{m} \alpha_{(j)}(u_{itj}) + x_{it}^T \beta + z_{it}^T b_i + \varepsilon_{it},$$

where b_i represents random slopes on the variables x_{it}. Quite a different challenge is the incorporation of random effects in additive terms. For simplicity of presentation we restrict consideration to one smooth effect. Let the smooth random intercept model

$$y_{it} = \beta_0 + \alpha(u_i) + b_{i0} + \varepsilon_{it},\ b_{i0} \sim N(0, \sigma^2),$$

be extended to

$$y_{it} = \beta_0 + \alpha(u_i) + \alpha(u_i) b_{i1} + b_{i0} + \varepsilon_{it}, \tag{4.1}$$

with $\Big(b_{i0}, b_{i1}\Big) \sim N(0, Q(\rho))$.
As usual the smooth component has to be centered for reasons of identifiability of effects, in our applications $\sum_i \alpha(u_i) = 0$ has been used. That means the "random slope" b_{i1} in model (4.1) is a parameter that, quite similar to random slopes in linear mixed models, lets the strength of the variable vary across subjects. The dependence on variable u_i becomes

$$\alpha(u_i) + \alpha(u_i) b_{i1} = \alpha(u_i)(1 + b_{i1})$$

showing that $\alpha(u_i)$ represents the basic effect of variable u_i but this effect can be stronger for individuals if $b_{i1} > 0$ and weaker if $b_{i1} < 0$. Thus b_{i1} strengthens or attenuates the effect of the variable u_i. If the variance of b_{i1} is very large it may even occur that $b_{i1} < 1$ meaning that the effect of u_i is "inverted" for some individuals. If $\alpha(u_i)$ is linear with $\alpha(u_i) = \beta u_i$, the influence term is given by $\alpha(u_i)(1 + b_{i1}) = u_i(\beta + \tilde{b}_{i1})$ where $\tilde{b}_{i1} = \beta b_{i1}$ represents the usual term in linear mixed models with random slopes. Thus comparison with the linear mixed model should be based on the rescaled random effect $\tilde{\beta}_{i1}$ with $E(\tilde{\beta}_{i1}) = 0, \text{Var}(\tilde{\beta}_{i1}) = \beta^2 \text{Var}(\beta_{i1})$.

The main problem in model (4.1) is the estimation of the random effects. If $\alpha(u)$ is expanded in basis functions by $\alpha(u) = \sum_s \alpha_s \phi_s(u)$ one obtains

$$\alpha(u_i)b_i = \sum_s \alpha_s b_i \phi_s(u),$$

which is a multiplicative model since α_s *and* b_i are unknown and cannot be observed. However, boosting methodology may be used to obtain estimates for the model. The basic concept in boosting is that in one step the refitting of $\alpha(u_i)$ is done by using a weak learner which in our case corresponds to large λ in the penalization term.

Thus in one step the change from iteration $\alpha^{(l)}$ to $\alpha^{(l+1)}$ is small. Consider the model in vector form with predictor $\eta_i^T = (\eta_{i1}, \ldots, \eta_{iT_i})$ with

$$\eta_i = \mathbf{1}\beta_0 + \Phi_{i.}\alpha + (\mathbf{1}\, \Phi_{i.}\alpha) \begin{pmatrix} b_i \\ b_{i1} \end{pmatrix},$$

where $1^T = (1, \ldots, 1)$ is a vector of 1s, $\Phi_{i.}$ is the corresponding matrix containing evaluations of basis functions and $\alpha^T = (\alpha_1, \ldots \alpha_M)$ denotes the corresponding weights. Then the refitting of residuals in the iteration step is modified in the following way.

Let $\eta_i^{(l-1)}$ denote the estimate from the previous step. Then the refitting of residuals (without selection) is done by fitting the model

$$y_i - \eta_i^{(l-1)} \sim N(\eta_i,\, V_i(\theta))$$

with

$$\eta_i = \mathbf{1}\beta_0 + \Phi_{i.}\alpha + (1,\, \Phi_{i.}\hat{\alpha}^{(l-1)}) \begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix}, \tag{4.2}$$

where β_0, α are the parameters to be estimated and the estimate from the previous step $\hat{\alpha}^{(l-1)}$ is considered as known parameter. With resulting estimates $\hat{\beta}_0$, $\hat{\alpha}$ the corresponding update step takes the form

$$\hat{\alpha}^{(l)} = \hat{\alpha}^{(l-1)} + \hat{\alpha} \quad , \quad \hat{\beta}_0^{(l)} = \hat{\beta}_0^{(l-1)} + \hat{\beta}_0.$$

The basic idea behind the refitting is that forward iterative fitting procedures like boosting are weak learners. Thus the previous estimate is considered as known in the last term of (4.2). Only the additive term $\Phi_{i\cdot}\alpha$ is refitted within one iteration step. Of course after the refit the variance components corresponding to (b_{i0}, b_{i1}) have to be estimated.

4.1 General Model with Cluster-Specific Splines

Let the data be given by $(y_{it}, x_{it}, u_{it}, z_{it}), \quad i = 1, \ldots, n, \quad t = 1, \ldots, T_i$, where y_{it} is the response for observation t within cluster i and $x_{it}^T = (x_{it1}, \ldots, x_{itp}), \quad u_{it}^T = (u_{it1}, \ldots, u_{itm}), \quad z_{it}^T = (z_{it1}, \ldots, z_{itq_i})$ are vectors of covariates, which may vary across clusters and observations. The semi-parametric mixed model with cluster-specific splines that is considered in the following has the form

$$\begin{aligned} y_{it} &= x_{it}\beta + \textstyle\sum_{j=1}^m \alpha_{(j)}(u_{itj}) + z_{it}^T b_i^{(1)} + \sum_{j=1}^k \alpha_{(j)}(u_{itj}) b_{i(j)}^{(2)} + \epsilon_{it} \\ &= \mu_{it}^{par} + \mu_{it}^{add} + \mu_{it}^{rand} + \mu_{it}^{cl} + \epsilon_{it} \end{aligned}$$

where $b_i = [b_i^{(1)}, (b_i^{(2)})^T]^T \sim N(0, Q(\rho))$ is a partitioned random effect and $Q(\rho)$ is a parameterized covariance matrix and

$\mu_{it}^{par} = x_{it}^T\beta$ is a linear parametric term,

$\mu_{it}^{add} = \sum_{j=1}^m \alpha_{(j)}(u_{it_j})$ is an additive term with unspecified influence functions $\alpha_{(1)}, \ldots, \alpha_{(m)}$,

$\mu_{it}^{rand} = z_{it}^T b_i^{(1)}$ contains the cluster-specific random effect $b_i^{(1)}$,

$\mu_{it}^{cl} = \sum_{j=1}^k \alpha_{(j)}(u_{itj}) b_{i(j)}^{(2)}$ is a modification of additive terms $\alpha_{(1)}, \ldots, \alpha_{(k)}$ by cluster specific linear random effects $b_{i(j)}^{(2)}$ with $(b_i^{(2)})^T = (b_{i(1)}^{(2)}, \ldots, b_{i(k)}^{(k)})$, and

ϵ_{it} is the noise variable, $\epsilon_{it} \sim N(0, \sigma_\epsilon^2 I), \quad \epsilon_{it}, b_i$ independent.

To approximate the nonlinear functions one uses

$$\alpha_{(j)}(u) = \sum_{s=1}^M \alpha_s^{(j)} \phi_s^{(j)}(u) = \alpha_j^T \phi^{(j)}(u) \tag{4.3}$$

where $\phi_s^{(j)}$ denotes the s-th basis function for variable $j, \alpha_j^T = (\alpha_1^{(j)}, \ldots, \alpha_M^{(j)})$ are unknown parameters and $\phi_j(u)^T = (\phi_1^{(j)}(u), \ldots, \phi_M^{(j)}(u))$ represent the vector-valued

evaluations of the basis functions.

By collecting observations within one cluster the model has the form

$$\begin{aligned} y_i &= X_i\beta + \Phi_{i1}\alpha_1(1 + b^{(2)}_{i(1)}) + \ldots + \Phi_{ik}\alpha_k(1 + b^{(2)}_{i(k)}) \qquad (4.4)\\ &\quad +\Phi_{i,k+1}\alpha_{k+1} + \ldots + \Phi_{im}\alpha_m + Z_i b_i^{(1)} + \epsilon_i, \\ &\quad \begin{bmatrix}\epsilon_i \\ b_i\end{bmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma^2_\varepsilon I & \\ & Q(\rho)\end{pmatrix}\right), \qquad (4.5)\end{aligned}$$

where $X_i\beta$ contains the linear term, $\Phi_{ij}\alpha_j$ represents the additive term, $Z_i\beta$ the random term and $b^T = ((b^{(1)})^T, (b^{(2)})^T)$. Vectors and matrices are given by $y_i^T = (y_{i1}, \ldots, y_{iT_i})$, $X_i^T = (x_{i1}, \ldots, x_{iT_i}), \Phi_{ij}^T = (\phi^{(j)}(u_{i1j}), \ldots, \phi^{(j)}(u_{iT_ij}))$, $Z_i^T = (z_{i1}, \ldots, z_{iT_i})$, $\epsilon_i^T = (\epsilon_{i1}, \ldots, \epsilon_{iT_i})$. In the case of the truncated power series the "fixed" term $\gamma_0^{(j)} + \gamma_1^{(j)}u + \ldots + \gamma_d^{(j)}u^d$ is taken into the linear term $X_i\beta$ without specifying X_i and β explicitly.
In matrix form one obtains

$$y = X\beta + \Phi_1\alpha_1 + \ldots + \Phi_m\alpha_m + \mathbb{Z}b^{(1)} + \mathbb{R}b^{(2)} + \epsilon,$$

$$y = X\beta + \Phi_1\alpha_1 + \ldots + \Phi_m\alpha_m + \tilde{\mathbb{Z}}b + \epsilon,$$

where $y^T = (y_1^T, \ldots, y_n^T)$, $b^T = (b_1^T, \ldots, b_n^T)$, $b^T = ((b^{(1)})^T, (b^{(2)})^T)$, $\epsilon^T = (\epsilon_1^T, \ldots, \epsilon_n^T)$,
$X^T = (X_1^T, \ldots, X_n^T)$, $\Phi_j^T = (\Phi_{1j}^T, \ldots, \Phi_{nj}^T)$, $\mathbb{Z}^T = diag(Z_1^T, \ldots, Z_n^T)$, $R_{i.} := R_{i.}(\alpha_1, \ldots, \alpha_k) = [\Phi_{i1}\alpha_1, \ldots, \Phi_{ik}\alpha_k]$, $R = diag(R_{1.}, \ldots, R_{n.})$ and $\tilde{\mathbb{Z}} = [\mathbb{Z}, \mathbb{R}]$. Parameters to be estimated are the fixed effects, collected in $\delta^T = (\beta^T, \alpha_1^T, \ldots, \alpha_m^T)$ and the variance specific parameters $\theta^T = (\sigma_\varepsilon, \rho^T)$ which determine the covariances $cov(\epsilon_{it}) = \sigma^2_\varepsilon I_{T_i}$ and $cov(b_i) = Q(\rho)$.

4.1.1 The Boosting Algorithm for Models with Cluster-Specific Splines

The following algorithm uses componentwise boosting. Componentwise boosting means that only one component of the predictor, in our case one smooth term $\Phi_{ij}\alpha_j$, is refitted at a time. That means that a model containing the linear term and only one smooth component is fitted in one iteration step. For simplicity we will use the notation

$$X_{i(r)} = [X_i\ \Phi_{ir}] \quad , \quad \delta_r^T = (\beta^T, \alpha_r^T)$$

for the design matrix with predictor $X_{i(r)} = X_i\beta + \Phi_{ir}\alpha_r$.
The corresponding penalty matrix is denoted by K_r, which for the truncated power series has the form

$$K_r = Diag(0,\ \lambda I).$$

BoostMixed

1. *Initialization*
   Compute starting values $\hat{\beta}^{(0)}, \hat{\alpha}_1^{(0)}, \ldots \hat{\alpha}_m^{(0)}$ and set $\eta_i^{(0)} = X_i\hat{\beta}^{(0)} + \Phi_{i1}\hat{\alpha}_1^{(0)} + \ldots + \Phi_{ik}\hat{\alpha}_k^{(0)}$ and set $R_{i.}^{(0)} := R_{i.}(\hat{\alpha}_1^{(0)}, \ldots, \hat{\alpha}_k^{(0)}) = (\Phi_{i1}\hat{\alpha}_1^{(0)}, \ldots, \Phi_{ik}\hat{\alpha}_k^{(0)})$, $V_i^{(0)} = (\sigma_\epsilon^{(0)})^2 I + \tilde{Z}_i^{(0)} Q(\rho^{(0)}(\tilde{Z}_i^{(0)})^T$, where $\tilde{Z}_i^{(0)} = (Z_i, R_{i.}^{(0)})$.

2. *Iteration*
   For l=1,2,...

   (a) *Refitting of residuals*

   i. Computation of parameters
   For $r \in \{1, \ldots, m\}$ the model for residuals

   $$y_i - \eta_i^{(l-1)} \sim N(\eta_{i(r)}, V_i^{(l-1)}(\theta^{(l-1)}))$$

   with

   $$\eta_{i(r)} = X_{i(r)}\delta_r = X_i\beta + \Phi_{ir}\alpha_r$$

   is fitted, yielding

   $$\hat{\delta}_r = (\sum_{i=1}^{n}(X_{i(r)}^T(V_i^{(l-1)}(\theta^{(l-1)}))^{-1}X_{i(r)} + K_r))^{-1}\sum_{i=1}^{n} X_{i(r)}^T(V_i^{(l-1)}(\theta^{(l-1)}))^{-1}(y_i - \eta_i^{(l-1)}).$$

   ii. Selection step
   Select from $r \in \{1, \ldots, m\}$ the component j that leads to the smallest $AIC_r^{(l)}$ or $BIC_r^{(l)}$ as given in Section 3.5.3.

   iii. Update
   Set $\hat{\beta}^{(l)} = \hat{\beta}^{(l-1)} + \hat{\beta}$,
   and

   $$\hat{\alpha}_r^{(l)} = \begin{cases} \hat{\alpha}_r^{(l-1)} & \text{if } r \neq j \\ \hat{\alpha}_r^{(l-1)} + \hat{\alpha}_r & \text{if } r = j, \end{cases}$$

   $$\hat{\delta}^{(l)} = ((\hat{\beta}^{(l)})^T,\ (\hat{\alpha}_1^{(l)})^T,\ \ldots (\hat{\alpha}_m^{(l)})^T)^T.$$

Update for $i = 1, \ldots, n$

$$\eta_i^{(l)} = \eta_i^{(l-1)} + X_{i(j)}\hat{\delta}_j$$

and set $R_{i.}^{(l)} := R_{i.}(\hat{\alpha}_1^{(l)}, \ldots, \hat{\alpha}_k^{(l)}) = (\Phi_{i1}\hat{\alpha}_1^{(l)}, \ldots, \Phi_{ik}\hat{\alpha}_k^{(l)})$, $V_i^{(l)}(\theta) = (\sigma)^2 I + \tilde{Z}_i^{(l)} Q(\rho)(\tilde{Z}_i^{(l)})^T$, where $\tilde{Z}_i^{(l)} = (Z_i, R_{i.}^{(l)})$.

(b) *Computation of Variance Components*
The computation is based on the penalized log-likelihood

$$l_p(\theta|\eta^{(l)}; \delta_l) = -\tfrac{1}{2}\sum_{i=1}^n \log(|V_i^{(l)}(\theta)|) + \sum_{i=1}^n (y_i - \eta^{(l)})^T V_i^{(l)}(\theta)^{-1}(y_i - \eta^{(l)})$$

$$-\tfrac{1}{2}(\hat{\delta}^{(l)})^T K \hat{\delta}^{(l)}.$$

Maximization yields $\hat{\theta}^{(l)}$. Set $V_i^{(l)}(\theta^{(l)}) = (\sigma^{(l)})^2 I + \tilde{Z}_i^{(l)} Q(\rho^{(l)})(\tilde{Z}_i^{(l)})^T$, where $\tilde{Z}_i^{(l)} = (Z_i, R_{i.}^{(l)})$.

We chose componentwise boosting techniques since they turn out to be very stable in the high dimensional case where many potential predictors are under consideration. In this case the procedure automatically selects the relevant variables and may be seen as a tool for variable selection with respect to unspecified smooth functions. In the case of few predictors one may also use boosting techniques without the selection step by refitting the residuals for the full model with design matrix $[X_i \Phi_{i1} \ldots \Phi_{im}]$.

4.2 Simulation

We present part of a simulation study in which the performance of semiparametric mixed models with cluster-specific splines is compared to semiparametric mixed models. The underlying model is the random effects model

$$y_{it} = x_{it1} * \beta_1 + x_{it2} * \beta_2 + \sum_{j=1}^{30} c * \alpha_{(j)}(u_{it}) + b_{i0} + c * \alpha_{(1)}(u_{it}) b_{i1} + \epsilon_{it}, i = 1, \ldots, 66, t = 1, \ldots, 15$$

with the smooth components given by

$$\begin{aligned}
\alpha_{(1)}(u) &= sin(u) \quad u \in [-3,3], \\
\alpha_{(2)}(u) &= cos(u) \quad u \in [-2,8], \\
\alpha_{(3)}(u) &= cos(u) \quad u \in [-3,3], \\
\alpha_{(j)}(u) &= 0 \qquad\quad u \in [-3,3], j = 4,\ldots,30.
\end{aligned} \tag{4.6}$$

The vectors $u_{it}^T = (u_{it1},\ldots,u_{it30})$ have been drawn independently with components following a uniform distribution within the specified interval. For the covariates constant correlation is assumed, i.e. $corr(y_{itr}, y_{its}) = 0.2$. The constant c determines the signal strength of the covariates. The random effect and the noise variable have been specified by $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 0.6$ and $b_i = (b_{i0}, b_{i1})^T \sim N(0, Q)$ with

$$Q = \begin{bmatrix} 8 & 0.1 \\ 0.1 & 4 \end{bmatrix}.$$

In the part of the study which is presented the number of observations has been chosen by $n = 66, T = 15$.

The fit of the model is based on B-splines of degree 3 with 15 equidistant knots. The performance of estimators is evaluated separately for the structural components and the variance. The variance component for the random effects matrix Q is assumed to be unstructured.

To show the effect of using cluster-specific splines, one dataset with setting $c = 1$ and $p = 3$ was chosen. Figure 4.2 shows the 66 clusters with their cluster-specific splines (random intercept and modified spline curve), which are modifications of $\alpha_{(1)}(.)$. Figure 4.2 show the estimated and true modified cluster-specific spline functions (modified $\alpha_{(1)}(.)$ without random intercept. It is very characteristic for this curve that it has joint cut points.

Figure 4.2 shows that cluster-specific splines can improve the mean squared error for the predictor. If the cluster-specific spline is neglected, the variation is captured for small signals in the random effect and for huge signals in the error term and the random effect. The model with cluster-specific splines seem to be more sensitive in the variable selection. Nevertheless the model with cluster-specific splines delivers the original variances as shown in Figure 4.1 nearly independent form signals and smooth effects. For the computation of these mean matrices the 100 estimated covariance matrices were summed up and scaled by 100.

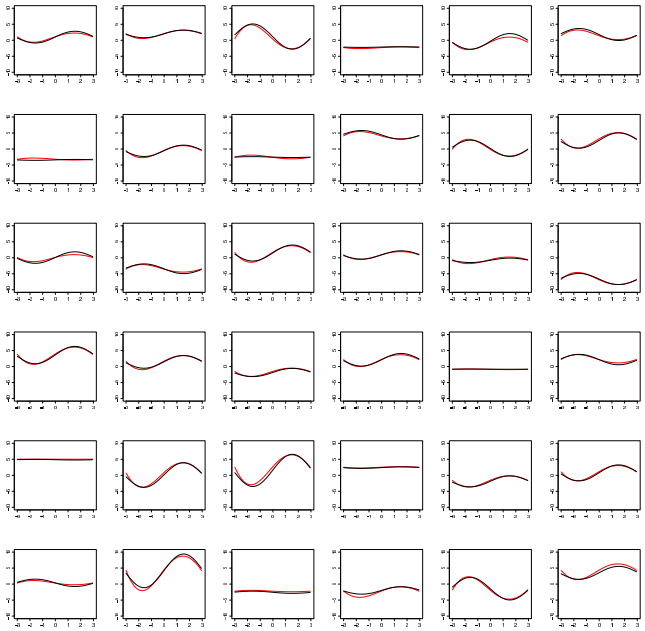


Figure 4.1: Plots of cluster-specific splines with random intercept with respect to the different clusters. The black lines are the estimated splines, the red ones are the true functions

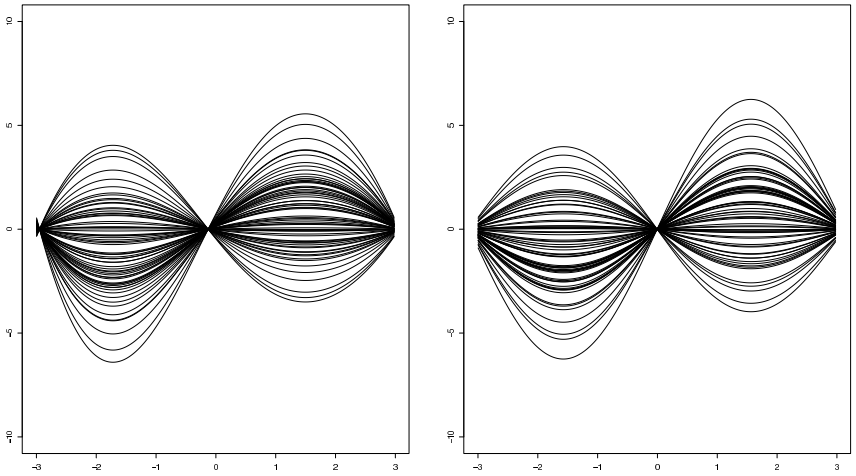


Figure 4.2: Plots of cluster-specific splines without random intercept. Left side are the estimated functions, the right side are the true functions.

	$p = 3$	$p = 5$	$p = 15$	$p = 25$
$c = 0.5$	$\begin{bmatrix} 8.008 & 0.209 \\ 0.209 & 4.735 \end{bmatrix}$	$\begin{bmatrix} 8.002 & 0.212 \\ 0.212 & 4.752 \end{bmatrix}$	$\begin{bmatrix} 7.986 & 0.195 \\ 0.195 & 4.715 \end{bmatrix}$	$\begin{bmatrix} 7.837 & 0.165 \\ 0.165 & 4.393 \end{bmatrix}$
$c = 1$	$\begin{bmatrix} 7.738 & 0.124 \\ 0.124 & 4.480 \end{bmatrix}$	$\begin{bmatrix} 7.736 & 0.124 \\ 0.124 & 4.482 \end{bmatrix}$	$\begin{bmatrix} 7.725 & 0.123 \\ 0.123 & 4.491 \end{bmatrix}$	$\begin{bmatrix} 7.717 & 0.124 \\ 0.124 & 4.515 \end{bmatrix}$
$c = 5$	$\begin{bmatrix} 7.794 & 0.134 \\ 0.134 & 6.687 \end{bmatrix}$	$\begin{bmatrix} 7.779 & 0.134 \\ 0.134 & 6.691 \end{bmatrix}$	$\begin{bmatrix} 7.714 & 0.123 \\ 0.123 & 6.643 \end{bmatrix}$	$\begin{bmatrix} 7.638 & 0.109 \\ 0.109 & 6.676 \end{bmatrix}$

Table 4.1: Mean of the estimated covariance matrices $\hat{Q} := Q(\hat{\rho})$ for the random effects covariance matrix Q

		cluster-specific splines					BoostMixed					
c	par	MSE_η	σ^2_ϵ	steps	falsepos	falseneg	MSE_η	σ^2_ϵ	σ^2_b	steps	falsepos	falseneg
0.5	3	138.611	0.603	14	0.00	0.00	143.502	1.099	8.047	16	0.00	0.00
0.5	5	142.035	0.605	15	1.12	0.85	146.897	1.096	8.039	17	0.66	0.00
0.5	15	148.847	0.610	15	1.73	0.94	155.453	1.089	8.018	20	1.96	0.00
0.5	25	161.973	0.631	15	2.08	0.97	160.488	1.085	8.003	23	2.55	0.01
1.0	3	173.448	0.610	38	0.00	0.00	201.067	2.596	7.781	59	0.00	0.00
1.0	5	173.962	0.609	41	1.11	0.91	205.673	2.593	7.773	61	0.32	0.00
1.0	15	177.910	0.607	42	1.98	0.94	228.118	2.572	7.735	64	1.53	0.00
1.0	25	179.547	0.606	43	2.46	0.94	240.204	2.561	7.708	67	2.16	0.00
5.0	3	1505.018	1.006	328	0.00	0.00	2031.959	50.802	7.776	971	0.00	0.00
5.0	5	1552.813	1.058	341	1.75	0.19	2257.905	50.473	7.759	984	1.44	0.00
5.0	15	1719.956	1.181	358	9.53	0.23	3424.553	49.162	7.585	984	2.89	0.00
5.0	25	2056.678	1.424	376	16.69	0.27	4538.329	47.894	7.452	985	3.46	0.00

Table 4.2: Comparision of MSE_η for BoostMixed and cluster-specific splines

What is getting clear in Table 4.1 that is not a problem to get the true variances from the model. It is also useful to use cluster-specific splines what can be seen in the MSE_η. Neglecting the cluster-specific splines lead with increasing signal to large estimates for the variance of the error component. However the cluster-specific splines tend to disregard relevant variables. Except for large signals the number of irrelevant variables in the model is quite comparable.

4.3 Application of Cluster-Specific Splines

4.3.1 Jimma Data: Description

The Jimma Infant Survival Differential Longitudinal Study which is extensively described in Lesaffre, Asefa & Verbeke (1999) is a cohort study examining the live births which took place during a one year period from September 1992 until September 1993 in Ethiopia. The study involves about 8000 households with live births in that period. The children were followed up for one year to determine the risk factors for infant mortality. Following Lesaffre, Asefa & Verbeke (1999) we consider 495 singleton live births from the town of Jimma and look for the determinants of growth of the children in terms of body weight (in kg). Weight has been measured at delivery and repeatedly afterwards. In addition we consider the socio-economic and demographic covariates age of mother in years (AGEM), educational level of mother (0-5: illiterate, read and write, elementary school, junior high school, high school, college and above), place of delivery (DELIV,1-3: hospital, health center, home), number of antenatal visits (VISIT, 0,$\geq$1), month of birth (TIME,1:Jan.-June, 0:July-Dec.), sex of child (1:male, 0:female). For more details and motivation of the study see Lesaffre, Asefa & Verbeke (1999). Figure 4.3 shows the overall evolution of weight and Figure 4.4 shows the growth curve of four children (observations and fitted curves) for an additive mixed model with random slopes on the additive age effect. It is seen that random slopes are definitely necessary for modelling since speed of growth varies strongly across children.

4.3.2 Jimma Data: Analysis with Cluster-Specific Splines

For the Jimma data we focus on the effect of age (in days) on the weight of children. Since growth measurements usually do not evolve linearly in time the use of a linear mixed model involves to find an appropriate scale of age. Lesaffre, Asefa & Verbeke (1999) found that weight is approximately linearly related with the square root of age. An even better approximation, they actually used in their analysis is the transformation $(age - log(age + 1) - 0.02 \times age)^{1/2}$. Since in growth curve analysis random slopes are needed , they had to find the scale before using mixed model methodology. The big advantage of the approach proposed here is that the scale of age has not to be found separately but is determined by the (flexible) mixed model itself. The model we consider includes random slopes on the age effects, smooth effect of age of mother and several parametric terms for the categorical variables. It has predictor

$$\eta_{it} = \beta_0 + \alpha_A(Age_i) + b_{i0} + b_{i1}\alpha_A(Age_i) + \alpha_{AM}(Age\,of\,Mother_i) + \,parametric\,term.$$

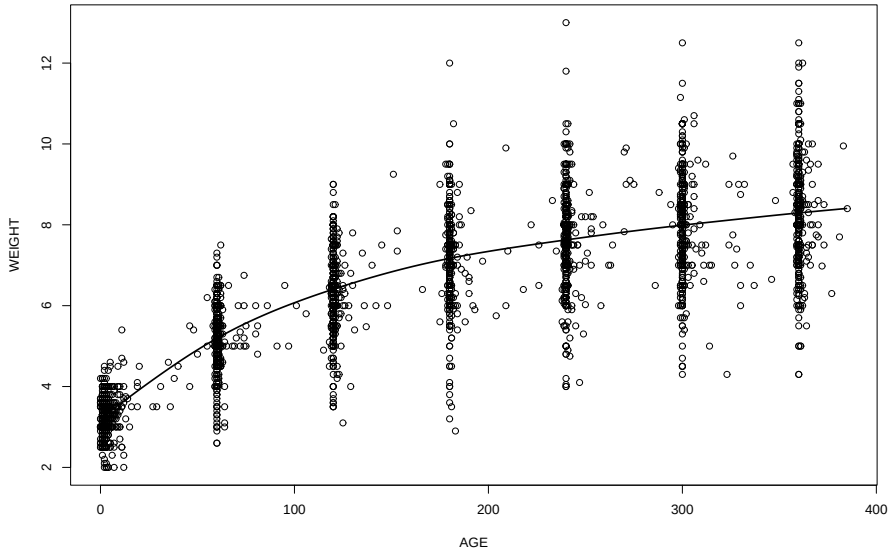


Figure 4.3: Evolution of average weight(kg) as function of age

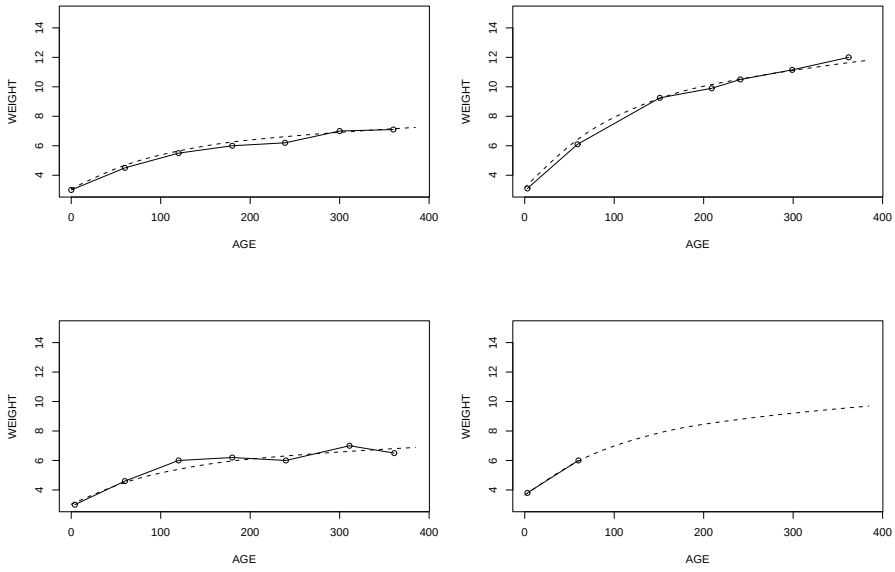


Figure 4.4: Individual infant curves (observed and predicted)

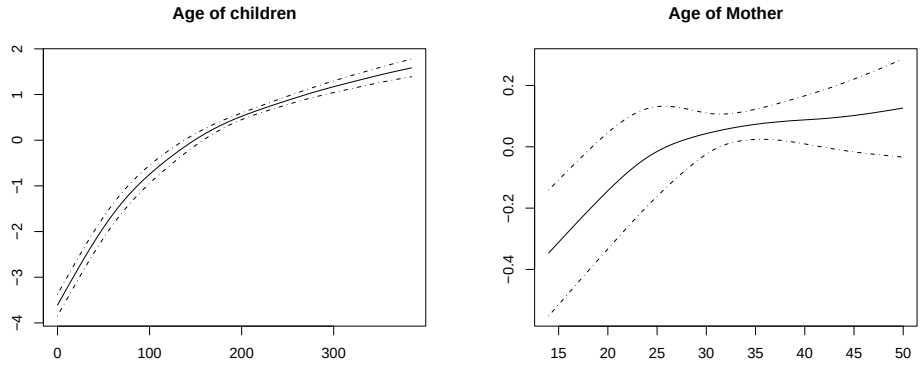


Figure 4.5: Effects of age of children (in days) and age of the mother (in years) in the Jimma study

Figure 4.3 shows the overall dependence (of children). Figure 4.5 shows the (centered) dependence on age and age of mother. It is seen that the effect of age of mothers is hardly linear (as assumed in the linear mixed models). Body weight of children seems to increase with age of mother up to about 30 years, then the effect remains rather stable. Table 4.3 gives the estimates of the parametric terms. For comparison the estimates for the linear mixed model with random slopes on the transformed age and linear effect of age of mother are given in Table 4.3 . As transformed age we use $(age - log(age + 1) - 0.02 \times age)^{1/2}$ as suggested by Lesaffre, Asefa & Verbeke (1999). It is seen that the effects of the categorical covariates are quite comparable. The differing intercepts are due to centering of variables. For age of mother the linear model shows a distinct increase (0.014 with standard deviation 0.004).

Table 4.4 shows the estimated variance of (b_{i0}, b_{i1}) for the flexible model and the linear mixed model with transformed age.

	BoostMixed		Mixed Model	
INTER	6.819	0.174	2.664	0.176
SEX	0.304	0.049	0.296	0.081
EDUC0	-0.051	0.066	-0.085	0.118
EDUC1	-0.021	0.151	-0.044	0.236
EDUC2	0.041	0.051	0.009	0.093
EDUC3	0.036	0.029	-0.005	0.060
EDUC4	-0.005	0.019	-0.042	0.042
VISIT	-0.078	0.072	-0.078	0.117
TIME	-0.177	0.065	-0.169	0.107
DELIV1	-0.027	0.007	-0.019	0.010
DELIV2	-0.148	0.031	-0.141	0.052
AGE			0.886	0.004
AGEM			0.014	0.004

Table 4.3: Effects of categorical covariates in Jimma study

BoostMixed		Mixed Model	
0.825962	0.196618	0.171369	-0.017506
0.196618	0.057253	-0.017506	0.045134

Table 4.4: Covariance matrix for random intercept and slope for Jimma data

4.3.3 Jimma Data: Visualizing Variable Selection

The models compared is the semi-parametric mixed model with cluster-specific splines given by

$$\eta_{it} = \beta_0 + \alpha_A(Age_i) + b_{i0} + b_{i1}\alpha_A(Age_i) + \alpha_{AM}(Age\,of\,Mother_i) + \, parametric\, term. \tag{4.7}$$

where the parametric term contains the categorical variables place of delivery (DELIV1-DELIV2), education (EDUC1-EDUC4), antenatal visits (ant), the interactions of age and delivery, as well as the interactions of sex (SEX.EDUC1-SEX.EDUC2) and education (SEX.EDUC1-SEX.EDUC4). The competitor is the linear mixed model with same parametric terms, but linear and quadratic age and age of the mother. The parametric terms where shrinked with $\lambda_{par} = 20$, the hyperparameter for smooth effects was set to $\lambda_{smooth} = 1000$. The x-axis of Figures 4.6 reflect the effective degrees of freedom for the computed model which is another expression for the needed iterations. On the y-axis one can see the development of the covariates with increasing iterations. The black vertical line indicates where the algorithm stops. For the semi-parametric and the linear mixed model the criterion stops around 6.5 degrees of freedom. In booth models is age the most relevant variable. Important in both models are also the SEX, the interactions AGE.DELIV1 and AGE:DELIV2 and the antenatal visits (ant) in the model. The only difference is that in the semi-parametric model deliv2 was taken and in the linear mixed model educ0.

The generalized build-up graphic is a nice tool to visualize the relevance of variables in both cases, linear and semi-parametric mixed models. It shows also information when variables with small relevance enters the model.

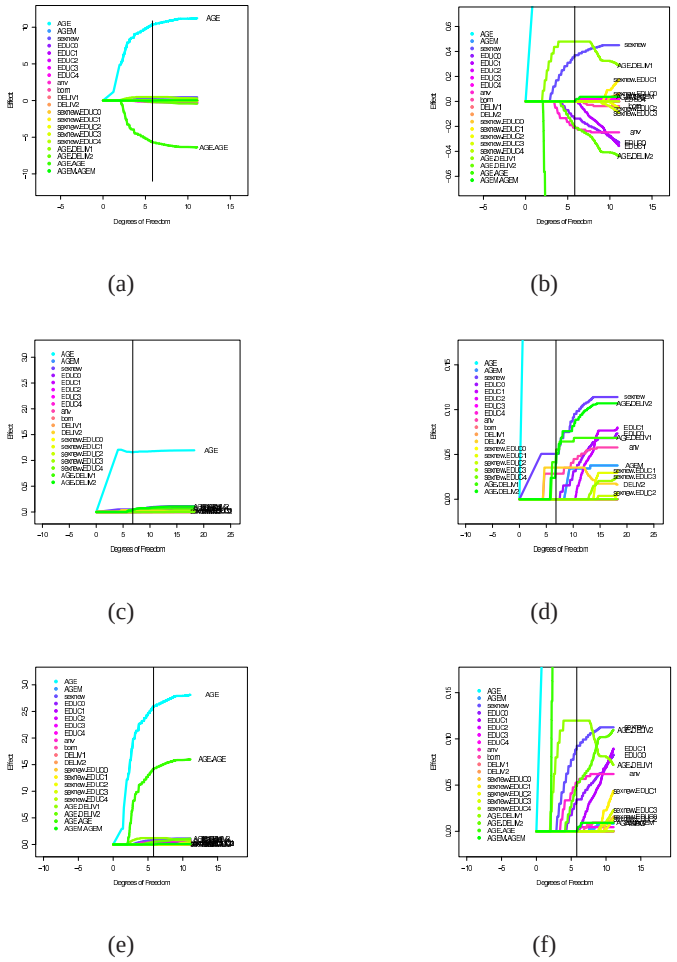


Figure 4.6: Coefficient build up for parametric model in **(a)** and zoomed version in **(b)**. Coefficient build up for semi-parametric model (age and agem are modeled with splines) in **(c)** and zoomed version in **(d)**. **(e)** and zoomed version in **(f)** shows the parametric model for rescaled coefficients.

4.3.4 Ebay-Auctions: Description

The technological advancements in measurement, collection, and storage of data have led to more and more complex data-structures. Examples include measurements of individuals' behavior over time, digitized 2- or 3-dimensional images of the brain, and recordings of 3- or even 4-dimensional movements of objects traveling through space and time. Such data, although recorded in a discrete fashion, are usually thought of as continuous objects represented by functional relationships. This gives rise to functional data analysis (FDA). In FDA Ramsay & Silverman (2002, Ramsey, & Silverman (2005) the center of interest is a set of curves, shapes, objects, or, more generally, a set of *functional observations*. This is in contrast to classical statistics where the interest centers around a set of data vectors.

There is only little other work that addresses the issue of sparse and unevenly spaced functional data. James & Sugar (2003) propose a model-based clustering approach that, similar to our approach, borrows information from neighboring functional objects and thus results in a more representative partitioning of the data.

In the following we motivate the problem of recovering sparsely and unevenly sampled curves by considering eBay's online auctions (see `www.ebay.com`). We describe eBay's auction mechanism, the data that it generates, and the challenges involved in taking a functional approach to analyzing online auction data. eBay's Auction Mechanism eBay is one of the biggest and most popular online marketplaces. In 2004, eBay had 135.5 million registered users, of which over 56 million bid, bought, or sold an item, resulting in over 1.4 billion listings for the year. Part of its success can be attributed to the way in which items are being sold on eBay. The dominant form of sale is the auction and eBay's auction format is a variant of the second price sealed-bid auction Krishna (2002) with "proxy bidding". This means that individuals submit a "proxy bid", which is the maximum value they are willing to pay for the item. The auction mechanism automates the bidding process to ensure that the person with the highest proxy bid is in the lead of the auction. The winner is the highest bidder and pays the second highest bid. For example, suppose that bidder A is the first bidder to submit a proxy bid on an item with a minimum bid of $10 and a minimum bid-increment of $0.50. Suppose that bidder A places a proxy bid of $25. Then eBay's web page automatically displays A as the highest bidder, with a bid of $10. Next, suppose that bidder B enters the auction with a proxy bid of $13. eBay still displays A as the highest bidder, however it raises the displayed high-bid to $13.50, one bid increment above the second-highest bid. If another bidder submits a proxy bid above $25.50, bidder A is no longer in the lead. However, if bidder A wishes, he or she can submit a new proxy bid. This process continues until the auction ends. Unlike other auctions, eBay has strict ending times, ranging between 1 and 10 days from the opening

of the auction, as determined by the seller.

eBay is a rich source of high-quality – and publicly available – bidding data. eBay posts complete bid histories of closed auctions for a duration of at least 15 days on its web site[1]. One implication of this is that eBay-data do not arrive in the traditional form of tables or spreadsheets; rather, it arrives in the form of HTML pages.

Figure 4.7 shows an example of eBay's auction data. The top of Figure 4.7 displays a summary of the auction attributes such as information about the item for sale, the seller, the opening bid, the duration of the auction, and the winner. The bottom of Figure 4.7 displays the bid history, that is the temporal sequence of bids placed by the individual bidders. Figure 4.8 shows the scatter of these bids over the auction duration (a 7-day auction in this example). We can see that only 6 bidders participated in this auction and that most bids were placed towards the auction end, with the earlier part of the auction only receiving one bid. Thus, if we conceptualize the evolution of price as a continuous curve between the start and the end of the auction, then Figure 4.8 shows an example of a very sparsely and unevenly sampled price-curve.

"Does price remain low throughout most of the early auction only to experience sharp increases at the end? And if so, is this price pattern the same for auctions of all types? Or does the pattern differ between, say, electronics and antiques?" Jank & Shmueli (2005) show that answering these questions can help profiling auction dynamics. Wang, Jank & Shmueli (2005) build upon similar ideas to develop a dynamic forecasting system for live auctions. (See also Shmueli, Jank, Aris, Plaisant & Shneiderman (2005) for an interactive visualization tool for online auctions.)

One way around this problem is to borrow information from other auctions. Consider Figure 4.9. It shows the bid histories for three individual auctions, labeled #2, #121 and #173. We can see that the price curve in auction #6 is only sampled at the end. Conversely, in auction #121 the price is sampled mostly at the beginning, with no information from the middle of the auction. And finally, auction #173 contains price information from the auction middle but only little from its start and end. While every auction individually only contains partial information about the price curve, if we put the information from all three auctions together, we obtain a more complete picture. This is illustrated in the bottom right corner of Figure 4.9. The idea of semiparametric mixed model smoothing is now to borrow from this combined information whenever an individual auction contains only incomplete information about its price evolution. We describe the methods more formally next.

Our data consist of 183 closed auctions for Palm M515 personal digital assistants (PDAs)

[1] See `http://listings.ebay.com/pool1/listings/list/completed.html`

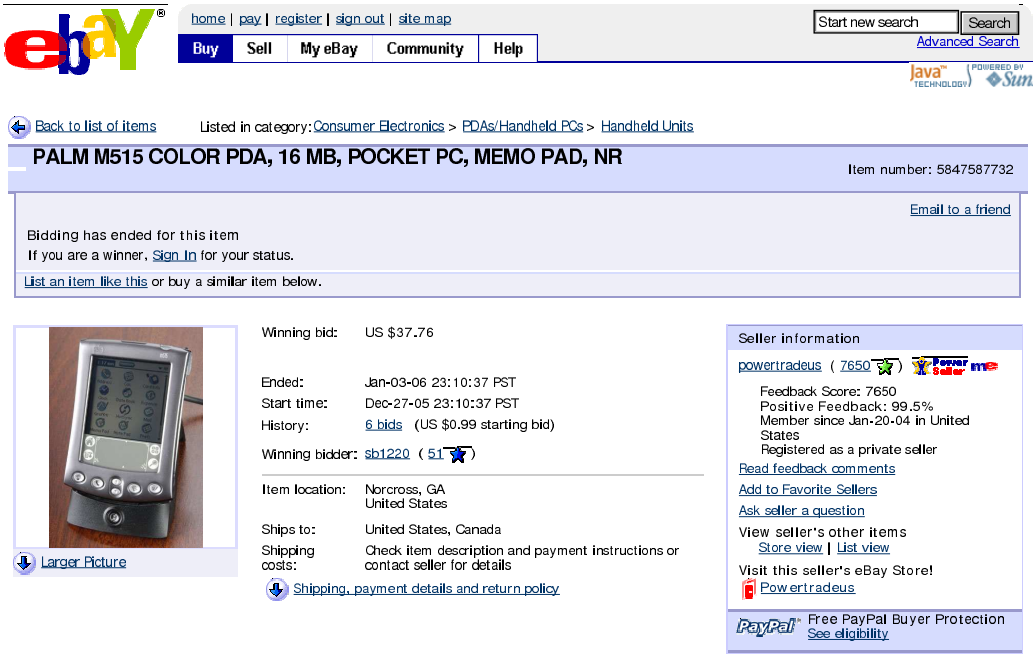


Figure 4.7: Bid history for a completed eBay auction. The top part displays auction attributes and includes information on the auction format, the seller and the item sold; the bottom part displays the detailed history of the bidders and their bids.

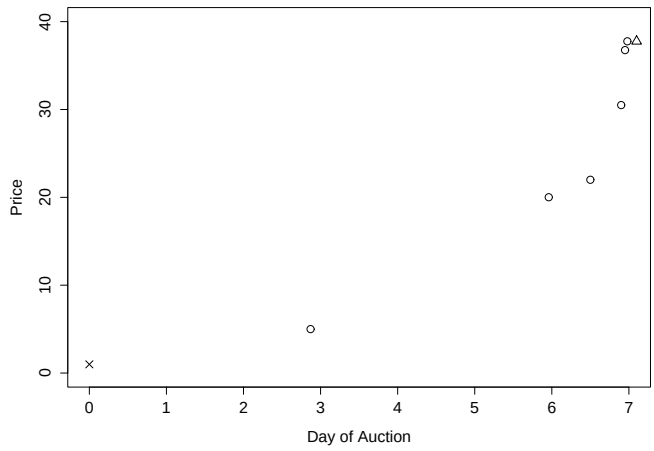


Figure 4.8: Scatterplot for bid history in Figure 4.7. The "×" marks the opening bid; the "△" marks the final price. Of the total of 6 bids, only one arrives before day 6.

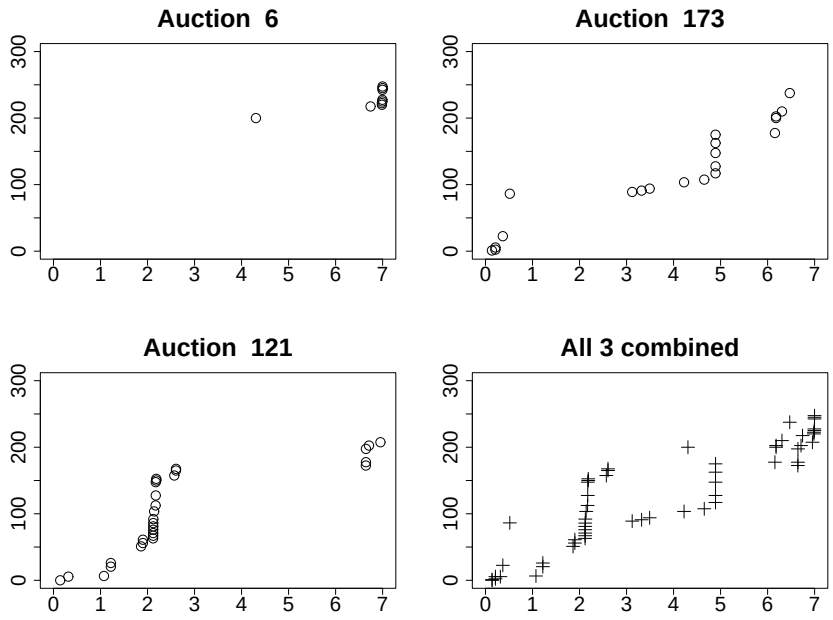


Figure 4.9: Three individual bid histories and their combined bids (bottom right panel).

that took place between March 14 and May 25 of 2003. In an effort to reduce as many external sources of variability as possible, we included data only on 7-day auctions, transacted in US Dollars, for completely new (not used) items with no added features, and

where the seller did not set a secret reserve price. Furthermore, we limited the data to competitive auctions, where there were at least two bids. These data are publicly available at `http://www.smith.umd.edu/ceme/statistics/`.

The data for each auction include its opening price, closing price, and the entire series of bids (amounts and time-stamps) that were placed during the auction. This information is found in the Bid history, as shown in Figure 4.7.

Note that the bid values that appear in the bid history are not the actual price shown by eBay during the auction. The reason is that eBay uses a second-price mechanism, where the highest bidder wins and pays the second highest bid. Therefore, at each point in time the displayed current price is the second highest bid. For this reason, we converted the actual bids into "current price", and therefore our final data are indeed monotone increasing.

4.3.5 Ebay-Data: Mixed Model Approach vs. Penalized Splines: Prognostic Performance

Although it is seen from Figure 4.11 that the more parsimonious mixed model yields better results we wanted to investigate the two procedures with respect to prognostic performance. Therefore the original data were splitter into a training dataset and a validation dataset. for each auction the data were split into bids, which come in within $2/3$ of the time and the rest. The first part of the data is considered as training data, the second part as validation data for the specific auction. One get data pairs $\{(t_{is}, \text{Price}_{is}^{(1)})|t_{is} < \frac{2}{3} * 7 \quad days\}$ for the training data and $\{(t_{is}, \text{Price}_{is}^{(2)})|t_{is} \geq \frac{2}{3} * 7 \quad days\}$ for the test data. The number of observations for auction i in the training dataset is $S_i^{(1)}$, for the test data $S_i^{(2)}$. Auctions with less than 3 bids were removed and not taken into the analysis. Thereby the data set reduces to 132 auctions. This reduction is necessary because in some auctions not enough data were available to fit a penalized spline. For the computation of the separate splines the set of knots were reduced to 3 since numerical problems arise in the computation. For the flexible spline solution 14 knots were taken. For both methods differences of order 2 and B-Splines of degree 2 were used. The estimates of the training dataset was then used to predict the values of the test dataset. For comparison the predicted mean squared errors on the validation set have been computed. In the flexible splines case boosting techniques as described were taken to get estimates. The square root of the price was taken since estimation lead to rather huge variance estimations. The log transformation was also considered but this transformation comprises a stronger reduction of information in the data.

The computed model using separately fitted penalized splines and the mixed model approach for auction i were

$$s(\text{Price}_{is}^{(1)}) = \alpha_0 + \phi^T(t_{is}^{(1)})\alpha_i$$
and
$$s(\text{Price}_{is}^{(1)}) = \tilde{\alpha}_0 + \phi^T(t_{is}^{(1)})\alpha + b_{i0} + \phi^T(t_{is}^{(1)})\tilde{\alpha}b_i.$$

respectively. Computation of mean squared error in the validation set yields 1701507 for separately fitted splines and 28352.5 for the mixed model approach. There the separately fitted splines have mse that is about 60 times larger.

It is obvious that the mixed model approach yields much better prediction than the penalized splines approach. Since the data are sparse in some auctions it is rather restrictive to limit the number of knots only to 3 knots. Another nice feature of the mixed model approach is that the monotonicity holds for all auctions without the implementation of restrictions that guarantee monotonicity.

4.3.6 Ebay Data: Final model

The following mixed effects model was used for all 183 auctions

$$s(Price_{is}) = \alpha_0 + \alpha(t_{is}) + b_{i0} + b_{i1}\alpha(t_{is}) + \epsilon_{is}$$

to model the data. Figure 4.11 shows for the first 36 auctions the estimates resulting from separate spline fitting and from using the mixed model approach. It is seen that the separate spline fitting approach might behave erratically. When data are sparse it may produce decreasing functions or very steep functions. In the case with one observation the estimate does not exist. On the other hand the mixed model approach yields sensible estimates even in sparse data situations. Even for one observation, i.e. auction 16 in figure 4.11, the price evolution can be modeled using all other auctions. If, as is the case auction 11 there is small but important information (bid at start, end and one some where in between), this information is enough to fix the level of the auction (random intercept) and the evolution of the auction (random slope for splines). In the case of auction 20 the random slope is estimated very close to the expectation of the random slope. Here information from other auctions is borrowed to get an idea what could have happened. But still the individuality of this auction is reflected in the random intercept, which allows variation also using the expected price evolution curve. The restriction to monotonicity is unnecessary then since for all auctions nondecreasing functions are estimated.

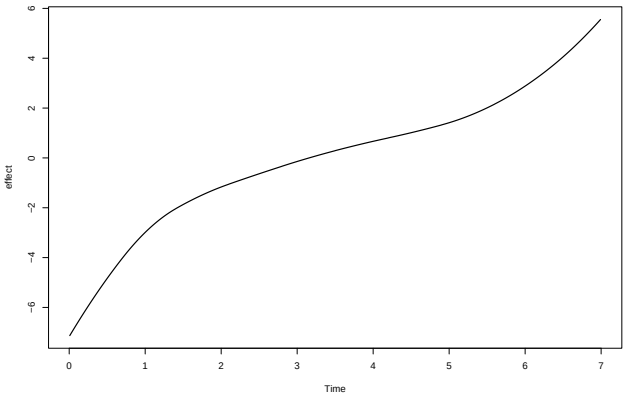


Figure 4.10: spline function for all auctions for Time

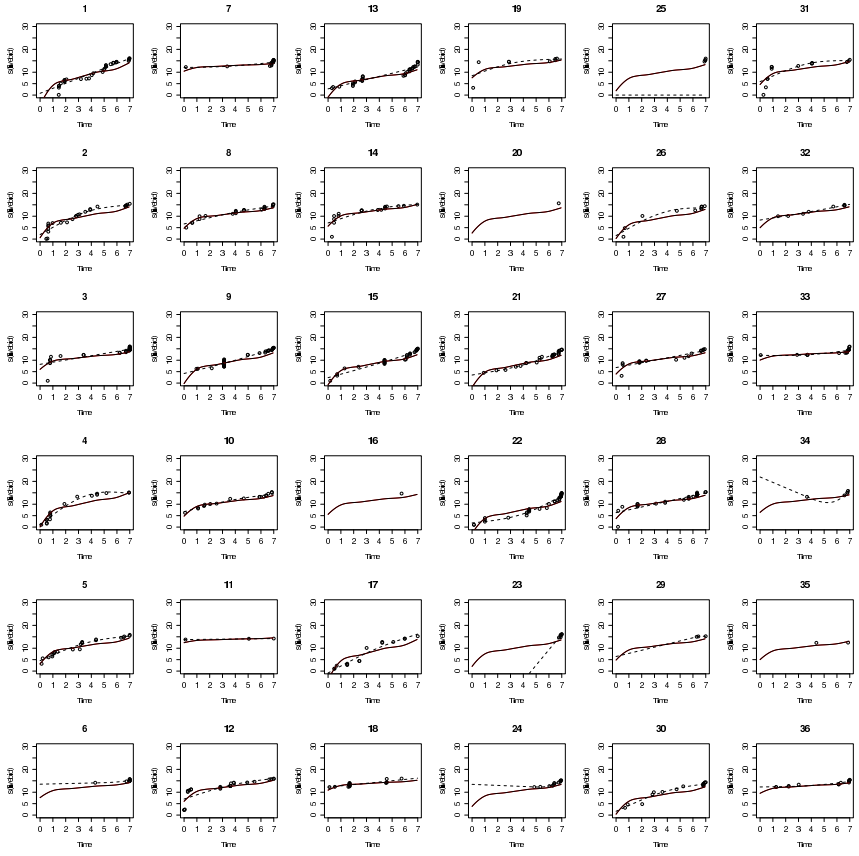


Figure 4.11: Smoothed Time: The first 36 auctions with their specific behavior regarding price and Time. Mixed model approach is shown by the solid lines, separately fitted penalized splines are the dotted lines.

	b_0	b_1
b_0	4.536 (1)	-0.619 (-0.847)
b_1	-0.619 (-0.847)	0.117 (1)

Table 4.5: Estimated covariance matrix $Q(\hat{\rho})$ for random intercept and slope for Ebay data. Correlation is given in brackets.

4.3.7 Canadian Weather Stations: Description and Model

The data were collected from 35 Canadian weather stations. Jim Ramsay offers the monthly temperature data for canadian weather stations on his web site[2]. The raw data were supplied by the Atmospheric Environment Service, Canadian Climate Centre, Data Management Division, 4905 Dufferin Street, Downsview, Ontario, M3H 5T4. The study includes 12 monthly measurements for each of the 35 weather station. The observed covariates are mean temperature (temp) in degree celsius, month (month), weather station and precipitation (prec) in mm^3. Let $temp_{it}$ denote the temperature for the i-th weather station with t-th measurement.

The following model was used

$$temp_{it} = \alpha_P(prec_{it}) + \alpha_M(month_{it}) + b_{i0} + b_{i1}\alpha_M(month_{it}) + e_{it}$$

to model the data. The assumption on the random effects b_{i0}, b_{i1} is that they are Gaussian, independent between clusters and conditional independent for the different measures within the cluster. Table 4.6 shows the estimated covariance structure for the random effects. Figure 4.12 shows 16 estimated mean temperatures for weather station 20 to station 35 modeled by cluster-specific spline curves. Figure 4.13 shows the estimated smooth effect for the precipitation.

30.184988	-1.012974
-1.012974	0.066423

Table 4.6: Covariance matrix for random intercept and slope for Canadian Weather Stations data

[2]See `ftp://ego.psych.mcgill.ca/pub/ramsay/FDAfuns/SPLUS/README.txt`

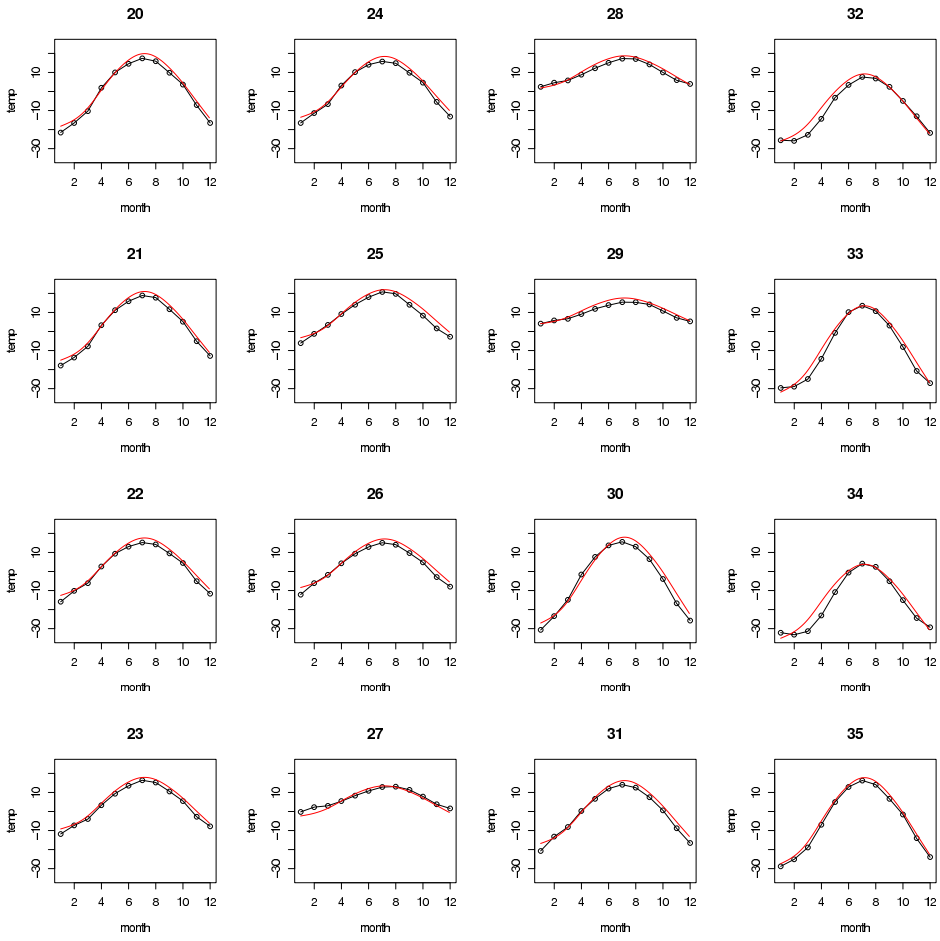


Figure 4.12: Monthly temperatures for 16 selected Canadian weather stations.

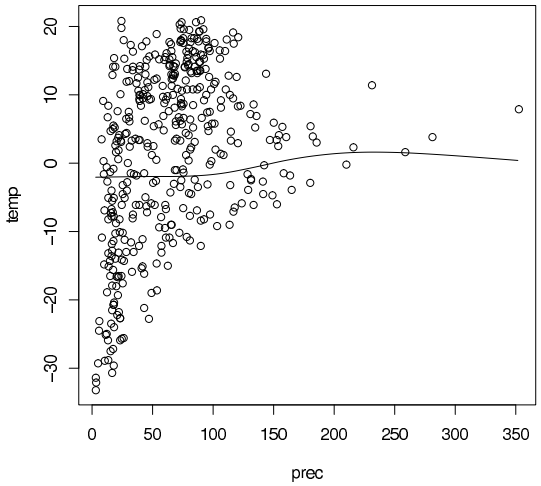


Figure 4.13: Temperatures for the Canadian weather stations depending on precipitation

Chapter 5

Generalized Linear Mixed Models

5.1 Motivation: The European patent data

The used data come form two sources. The one source is the *Online European Patent Register* provided by the European Patent Office at `http://www.epline.org`. The database covers published European patent applications as well as published international patent applications. The second source deals with characteristics on the different companies. Here 107 European firms were observed from 1993 to 2000 collecting variables like number of employees, research and development expenses. This dataset derives from the Global Vantage Database supported by Compustat. The information of both datasets were put together in a panel structured dataset. The objective is the analysis in the behavior of firms according to their preference to do outsourcing. Using all information one get the pooled data on 107 European firms in 856 firm years. Since the research and development data lacks of 261 firm years (missing values), only 595 firm years are remaining for the analysis. So the total number of yearly patent applications (PAT_YEAR) is part of the study as well as the patent applications that were sourced out. The response is the number of patents that were sourced out (OUT).Moreover, a variable that operationalizes the technical diversity of patent applications is collected. It is a measure for the technological breadth where firms show activity (BREADTH). If the applicant is focused only on few technological fields one gets smaller measurements than if an applicant is active in a wide range of few fields. Moreover the volatility (VOL) of patent applications is given by another measure which takes account of the changes and fluctuations in the activity of patent applications. The other variables collected are the firm size measured in employees (EMP), the expenses for research and development in Euro (R_D_EUR), the expenses for research and development adjusted to employee (R_D_EMP) and patent (R_D_PAT), the patent portfolio of the company (PAT_PORT). Since the data derive from

Europe, dummy variables for the region (Germany (GER), France (FRA), United Kingdom (GBR) and others (OTH)) were introduced. For details on the construction of these measures see Wagner (2006). For this study only companies that had less than 20 000 employees over the observation period are considered. Applying these restrictions the hole dataset was reduced to 184 observations in total for 34 different companies. The response is the number of patents that were sourced out. One may assume that the response is Poisson distributed.

A simple model with only some covariates is given by

$$\begin{aligned} \eta_{it} &= 1\beta_0 + PAT\_YEAR_{it}\beta_1 + GER_{it}\beta_2 + FRA_{it}\beta_3 + GBR_{it}\beta_4 + b_i \\ OUT_{it}|\lambda_{it} &= Poisson(\lambda_{it}) \\ \lambda_{it} &= \mathbb{E}(OUT_{it}) = \exp(\eta_{it}) \end{aligned} \tag{5.1}$$

where the index PAT_YEAR_{it} addresses company i with measurement t and b_i is the random intercept for company i. z_{it} is set to 1 since a simple random intercept model is considered. For example, the first company of the dataset has measurements of all covariates in the years 1996-1998. The measurements in total for this company is $T_i = 3$. A common assumption on random intercepts is that they are Gausssian distributed with $b_i \sim N(0, \sigma_b^2)$. σ_b^2 is the random intercept variance.

Since one may use a short notation without addressing the variable names PA_YEAR, GER, FRA, GBR one set generally the response to $y_{it} := OUT_{it}$. The variables that are responsible for the fixed effects are packed into the vector $x_{it}^T := (1, PA\_YEAR_{it}, GER_{it}, FRA_{it}, GBR_{it})$. $z_{it}^T = 1$. The variables associated with the random effect are stacked in blockdiagonal entries in the matrix $\mathbb{Z} = bdiag(Z_1, \ldots, Z_n)$, where $Z_i^T = (z_{i1}, \ldots, z_{iT_i})$. The short term notation is with $X_i^T = (x_{i1}, \ldots, x_{iT})$, $X^T = (X_1^T, \ldots, X_n^T)$, $y_i^T = (y_{i1}, \ldots, y_{iT})$, $y^T = (y_1^T, \ldots, y_n^T)$ and $\beta^T = (\beta_0, \ldots, \beta_4)$, $b^T = (b_1^T, \ldots, b_n^T)$ and $\eta_i^T = (\eta_{i1}, \ldots, \eta_{iT})$ and $\eta^T = (\eta_1^T, \ldots, \eta_n^T)$ for clustered data representation

$$\eta_i = X_i\beta_i + Z_i b_i,$$

or in matrix representation

$$\eta = X\beta + \mathbb{Z}b.$$

There are 595 observations in the dataset derived from 35 companies, so we set $N = 595$ and $n = 35$. In this case the dimension of b which is denoted by q is n ($q := n$) and the

random design matrix has only one component (intercept), so the number of components are set to $c = 1$.

The model (5.1) can be extended to a random slope model

$$\eta_{it} = 1\beta_0 + PA\_YEAR_{it}\beta_1 + GER_{it}\beta_2 + FRA_{it}\beta_3 + GBR_{it}\beta_4 + b_i^{(1)} + PA\_YEAR_{it}b_i^{(2)}$$

In this case $z_{it}^T = (1, PA\_YEAR_{it})$, the number of random components are two (c=2), the dimension for the random intercept is $q_1 = n$ and for the slope $q_2 = n$. This is in short notation

$$\eta = X\beta + \mathbb{Z}b.$$

The dimension of b is 2*n. One can use the ordered design matrix for random effects with $Z_{i(1)} = 1$, where $Z_{i(1)}$ is a T_i dimensional vector of ones, and $Z_{i(2)}^T = [PA\_YEAR_{i1}, \ldots, PA\_YEAR_{iT_i}]$. The ordered random design matrix is then $\tilde{\mathbb{Z}} = [\text{bdiag}(Z_{1(1)}, \ldots, Z_{n(1)}), \text{bdiag}(Z_{1(2)}, \ldots, Z_{n(2)})]$, where

$$\eta = X\beta + \tilde{\mathbb{Z}}\tilde{b}$$

with $\tilde{b}^T = ({b^{(1)}}^T, {b^{(2)}}^T)$. In this representation the clustered structure of the data may be neglected, since the order of random effects are important. One may talk about crossed random effects if one has more than one component in the random design matrix ($c \geq 2$) and it is not possible to build a clustered structure from the random design matrix.

5.2 The Model

First we consider the longitudinal formulation of a GLMM with its assumptions.

Longitudinal formulation (clustered structure) Let the data be of the form $(y_{it}, x_{it}), i = 1, \ldots, n, t = 1, \ldots, T$, with y_{it} denoting a univariate response connected to observation t in cluster i and x_{it} denoting a vector of covariates, that may vary across the observations within one cluster.

Often the cluster corresponds to individuals and the observations to repeated measurements. For a more simpler presentation, the number of observations within one cluster T does not depend on the cluster.

A GLMM is specified by two components. The first assumption is, that the conditional density of y_{it}, given the explanatory variable x_{it} and the random effect b_i is of exponential family type

$$f(y_{it} \mid x_{it}, b_i) = \{\exp \frac{(y_{it}^T \theta_{it} - \kappa(\theta_{it}))}{\phi} + c(y_{it}, \phi)\}, \tag{5.2}$$

where θ_{it} denotes the natural parameter and $c(.)$ the log normalization constant. The second component specifies the link between response and the covariates. The structural assumption is based on the conditional mean

$$\mu_{it} = E(y_{it} \mid x_{it}, b_i) = h(x_{it}^T \beta + z_{it}^T b_i), \tag{5.3}$$

where $h : \mathbb{R}^1 \mapsto \mathbb{R}^1$ is the response function and z_{it} is a design vector composed from x_{it}.

The specification of the random effect model is completed by specifying the distribution $p(b_i, \rho)$ of the random effect b_i where ρ is a vector of structural parameters. The necessity of this assumption, in particular for the maximum likelihood theory follows from

$$f(y_i \mid X_i) = \int f(y_i \mid X_i, b_i) p(b_i; \rho) db_i \tag{5.4}$$

with $X_i^T = (x_{i1}, \ldots, x_{iT})^T$ where $y_i^T = (y_{i1}, \ldots, y_{it})$ and $f(y_i \mid X_i, b_i)$ is assumed to be given by

$$f(y_i \mid X_i, b_i) = \prod_{t=1}^{T} f(y_{it} \mid x_{it}, b_i).$$

General formulation (crossed random effects) In the literature (i.e. Schall (1991), Breslow & Clayton (1993) or Lin & Breslow (1996)) a more general notation for generalized linear mixed models is used. This notation allows the incorporation of crossed random effects and is not limited to a clustered structure of the data. Let $y = (y_{(1)}, \ldots, y_{(N)})$ be a vector of N observations.

Here $X_{N \times p}$ is a known design matrix, β is a vector of fixed effects, the Z_i are known $T_i \times q$ matrices, where q is the dimension of the random effects vector b and T_i is the number for observations in cluster i. The random effects b are assumed to be Gaussian with expectation zero and covariance $cov(b) = \mathbb{Q}(\rho)$, where ρ are structural parameters. $z_{(i)}$ is the design vector for random effects, which corresponds to measurement i, $x_{(i)}$ is the design vector for fixed effects corresponding to measurement i, $i \in 1 \ldots, N$

Then the conditional density of an exponential family is

$$f(y_{(i)} \mid x_{(i)}, b) = \{\exp \frac{(y_{(i)}^T \theta_{(i)} - \kappa(\theta_{(i)}))}{\phi} + c(y_{(i)}, \phi)\}, \tag{5.5}$$

where $\theta_{(i)}$ denotes the natural parameter and $c(.)$ the log normalization constant.

Let $g(.)$ be a monotonic function, the link (McCullagh & Nelder (1989)), such that $g(\mu)$ can be written as the linear model

$$g(\mu_{(i)})) = \eta_{(i)} = x_{(i)}\beta + z_{(i)}b, i = 1, \ldots, N$$

The matrix notation with $\mu^T = (\mu_{(1)}, \ldots, \mu_{(N)}$, $\eta^T = (\eta_{(1)}, \ldots, \eta_{(N)})$ is given by

$$g(\mu) = \eta = X\beta + \mathbb{Z}b \tag{5.6}$$

with $g(\mu) = (g(\mu_{(1)}), \ldots, g(\mu_{(N)}))$.

5.3 Numerical Integration Tools

In the following it is assumed that $y^T = (y_{(1)}, \ldots, y_{(N)})$ has covariance matrix $\mathbb{Q}(\rho)$ where ρ is a vector which parameterizes the covariance matrix.

The integration of marginal densities $\int f(y|b) * \tilde{p}(b; \rho)db$ for Gaussian mixtures with densities of exponential families is usually based on the integration

$$\int f(y|b)p(b; \rho)db = \int f(y|\mathbb{Q}(\rho)^{1/2}a) * \tilde{p}(a)da,$$

where $\tilde{p}(.)$ is the standard normal distribution, $p(.)$ is the normal distribution with expectation zero and covariance matrix $\mathbb{Q}(\rho)$ and $b = \mathbb{Q}(\rho)^{1/2}a$. $\mathbb{Q}(\rho)^{1/2}$ is the left Cholesky root of $\mathbb{Q}(\rho)$.

Most integration methods may be seen as a problem

$$I = \int_{-\infty}^{\infty} f(a)g(a)da.$$

$f(.)$ is a continuous function and $g(.)$ is the integration function (often a density). The functional form behind the integral is reduced to only two functions $g(.)$ and $f(.)$. I is then approximated by the arithmetic mean

$$\hat{I} = \sum_{j=1}^{m} f(a_j)w_j,$$

where $a_j, j = 1, \ldots, m$ are integration knots and $w_j, j = 1, \ldots, m$ are integration weights. The value of a_j and w_j depend on the integration method and on $g(.)$ that is used. They can be deterministic (Gauss-Hermite) or random (Monte Carlo). In the following the set of integration knots (integration points) $a_j, j = 1, \ldots, m$ are called grid of integration knots $a^T = (a_1, \ldots, a_m)$. $\hat{I}$ is called the approximation of the integral I with $I \approx \hat{I}$.

Riemann's sums For integration with Riemann's sums, a_i is deterministic, $w_j = \frac{1}{m}$, $g(a_j)$ is $g(a_j) = 1$. Riemann's sums can be extended to the trapezoid rule, which now uses special weights w_j, but the grid of integration knots a is the same as for Riemann's sums.

Gauss quadratures Since the accuracy of Riemann's sums is often bad one may take Gauss-Hermite quadrature, which is described in detail in the appendix. For more information on quadrature formulas see Deuflhard & Hohmann (1993), Davis & Rabinowitz (1975), Stroud (1971). The tables for nodes and weights can be found in Stroud & Secrest (1966) and Abramowitz & Stegun (1972). For Gauss-Hermite quadrature w_j are the quadrature weights and a_j are the quadrature points, which are arranged by optimizing Hermite polynomials. One problem of Gauss-Hermite quadrature is, that the integral is only sufficient if $f(x)$ is centered around zero. This problem can be usually solved by using adaptive quadrature schemes for Gauss-Hermite quadrature.

Riemann's sums and Gauss-Hermite quadrature operates in d-dimensional integration problems on complete integration grids which is the result of a Tensor product of one dimensional integration grids. Therefore the d-dimensional tensor product is used. The integration points have then an exponential order in the used dimension. For dimensional problems of more than five the curse of dimensionality makes computation not applicable.

Sparse grids for quadrature rules Smolyak's formula on quadrature rules thins out the grid in a way that quadrature points are combined together. For details see Smolyak (1963), Petras (2000), Petras (2001), Gerstner & Griebel (1998) and Gerstner & Griebel (2003). This is often called integration using sparse grids. That is an trade off between the goodness of accuracy and number of integration points. The so called deepness of Smolyak's quadrature is responsible for the number of points. For a deepness of the size of onedimensional quadrature points one obtains the described full grid. For poor deepness one obtains an logarithmic order of quadrature points in the dimension.

Monte Carlo Integration For Monte Carlo integration the function $g(a)$ is a Gaussian density, a_i are i.i.d. drawings from $g(a)$. For more information see Robert & Casella (2004), Calflisch (1998) and Ripley (1987). Problem of this integration method is to assess the goodness of accuracy in dependence of needed integration points. Usually one uses an adaptive integrations scheme where the integration points are increased since many times the same result is delivered.

Quasi Monte Carlo Integration The inverse cumulative d-dimensional standard normal distribution distribution function is uniformly distributed on the d-dimensional cube. That is why one may take low discrepancy sequences (Niederreiter (1992)) , which deliver highly uniform distributed, but deterministic, points in the d-dimensional unit cube. If elementwise the one dimensional inverse normal transformation is applied on these sequences on the unit cube, one obtains quasi monte carlo integration points. The empirical frequencies for small integration points are much closer to the uniform distribution functions, than random drawing on the unit cube. For more information see Judd (1992), Calflisch (1998). On Halton's sequence see Niederreiter (1960). On Sobol's sequence see Antonov & Saleev (1979) and Bratley & Fox (1988).

Since the d-dimensional standard normal distribution is a product of d one-dimensional standard normal distributions, the d-dimensional integration grid can be visualized for d=2 with the first two cumulative normal distribution functions. See Figure 5.1.

Laplace Approximation The marginal log-likelihood for $y^T = (y_{(1)}, \ldots, y_{(N)})$ is specified by

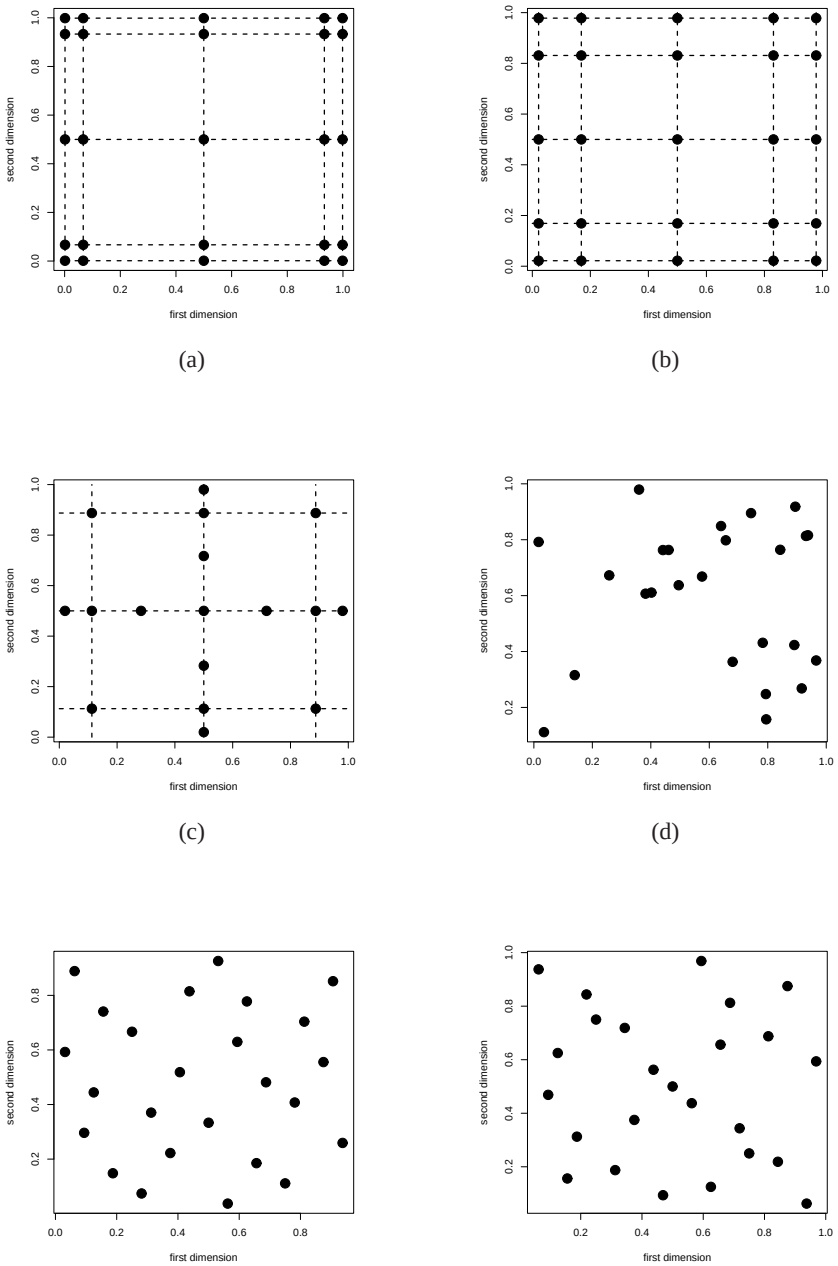
$$l(\beta, \rho) = \log(\int f(y|\tilde{b}; \beta) * p(\tilde{b}, \rho) d\tilde{b}), \tag{5.7}$$

where $f(y|\tilde{b}; \beta) = \prod_{i=1}^{N} f(y_{(i)}|\tilde{b}; \beta)$ and $f(y_{(i)}|\tilde{b}; \beta), i = 1, \ldots, N$ is a density from the exponential family and the mixing distribution $p(\tilde{b}, \rho)$ is the Gaussian density with expectation zero and unknown covariance $\mathbb{Q}(\rho)$. Since this log-likelihood is nasty to handle we try to find an easy approximation for the integral to do further computation. For the Laplace Approximation two likelihood functions for y are needed. The first one is the joint log-likelihood function

$$L_{\text{joint}}(\tilde{b}; \rho) = -k(\tilde{b}) \tag{5.8}$$

with $k(\tilde{b}) = -\log(f(y|\tilde{b}; \beta) * p(\tilde{b}; \rho))$. The second one is the marginal likelihood

$$L(\tilde{b}, \rho) = \int \exp\{-k(\tilde{b})\} d\tilde{b} \tag{5.9}$$



(a) (b)

(c) (d)

(e) (f)

Figure 5.1: Integration points evaluated for the 2-dimensional standard normal distribution. (a) are 25 integration points for Riemann's sums from $[-3, 3] \times [-3, 3]$, (b) are 25 quadrature points from Gauss-Hermite by tensor products, (c) 17 points from Smolyak's rule for Gaussian quadrature, (d) 25 Monte Carlo points, (e) 25 Quasi Monte Carlo points by Hobol's sequence and (f) quasi Monte Carlo points by Sobol's sequence

The basic idea is to make a quadratic expansion of $-k(\tilde{b})$ about its maximum point b before integrating. Therefore we solve

$$\frac{\partial k(\tilde{b})}{\partial \tilde{b}} = 0 = k'(b) \tag{5.10}$$

yielding b. The relation $\frac{\partial k(\tilde{b})}{\partial \tilde{b}} \int \exp\{-k(\tilde{b})\} = \int \frac{\partial k(\tilde{b})}{\partial \tilde{b}} \exp\{-k(\tilde{b})\} = 0$ indicates that b maximizes also the marginal likelihood (5.9) with respect to $\tilde{b}$. Then compute the curvature of equation (5.8)

$$\frac{\partial^2 k(\tilde{b})}{\partial \tilde{b} \partial \tilde{b}^T} = k''(\tilde{b}). \tag{5.11}$$

A first-order Taylor-Approximation of $k(b)$ in b with $k(\tilde{b}) \approx k(b) + k'(b)(\tilde{b} - b) + \frac{1}{2}(\tilde{b} - b)^T k''(b)(\tilde{b} - b)$is now applied to

$$\begin{aligned} \int \exp\{-k(\tilde{b})\} d\tilde{b} &\approx \int \exp\{-k(b) - \tfrac{1}{2}(\tilde{b} - b)^T k''(b)(\tilde{b} - b)\} d\tilde{b} \\ &= \int \exp\{-k(b)\} * \exp\{(\tilde{b} - b)^T k''(b)(\tilde{b} - b)\} d\tilde{b} \\ &= \exp\{-k(b)\} * (\sqrt{2\pi})^{p/2} |k''(b)^{-1}|^{1/2} \end{aligned} \tag{5.12}$$

since $k'(b) = 0$. $k'(b)$ and $k''(b)$ are computed in detail using log-likelihood (5.8) with $\delta^T = (\beta, b)^T$ and $\Sigma_{(i)} = \text{var}(y_{(i)})$

$$\begin{aligned} k'(b) &= -\textstyle\sum_{i=1}^N z_{(i)} D_{(i)}(\delta) \Sigma_{(i)}^{-1}(\delta)(y_{(i)} - \mu_{(i)}(\delta)) + \mathbb{Q}^{-1}(\rho) b, \\ k''(b) &= \textstyle\sum_{i=1}^N z_{(i)} D_{(i)}(\delta) \Sigma_{(i)}^{-1}(\delta) D_{(i)}^T(\delta) z_{(i)}^T + \mathbb{Q}^{-1}(\rho) + R(\delta), \end{aligned} \tag{5.13}$$

with

$$R(\delta) = -\sum_{i=1}^N [\frac{\partial}{\partial b^T}(z_{(i)} D_{(i)}(\delta) \Sigma_{(i)}(\delta))](y_{(i} - \mu_{(i)}(\delta)), \tag{5.14}$$

where

$$\begin{aligned} \eta_{(i)}(\delta) &= x_{(i)}^T \beta + z_{(i)}^T b + \text{offset}_{(i)}, \\ \mu_{(i)}(\delta) &= h(\eta_{(i)}(\delta)), \\ D_{(i)}(\delta) &= \tfrac{\partial h(\eta_{(i)}(\delta))}{\partial \eta_{(i)}(\delta)}. \end{aligned} \tag{5.15}$$

For canonical link-function one has $R(\beta, b) = 0$. Generally can be assumed that $\mathbb{E}(R(\delta)) = 0$. We set

$$k''(b) = \sum_{i=1}^{N} z_{(i)} D_{(i)}(\delta) \Sigma_i^{-1}(\delta) D_{(i)}^T(\delta) z_{(i)}^T + \mathbb{Q}^{-1}(\rho). \tag{5.16}$$

Applying the results of (5.12) to (5.9) we get the Laplace approximated log-likelihood with $\delta^T = (\beta, b)^T$

$$\begin{aligned}
l_{\text{Laplace}}(\delta, \rho) &= -\tfrac{1}{2}\log(|k''(b)|) + \tfrac{p}{2}\log(2\pi) - k(b) \\
&= -\tfrac{1}{2}\log(|k''(\tilde{b})|) + \tfrac{p}{2}\log(2\pi) + \log(f(y|\tilde{b};\beta)) - \log(p(\tilde{b};\rho)) \\
&= -\tfrac{1}{2}\log(|k''(b)|) + \tfrac{p}{2}\log(2\pi) \\
&\quad + \log(f(y|b;\beta)) - \tfrac{p}{2}\log(2\pi) - \tfrac{1}{2}\log(|\mathbb{Q}(\rho)|) - \tfrac{1}{2} b^T \mathbb{Q}(\rho)^{-1} b \\
&= -\tfrac{1}{2}\log(|k''(b)| * |\mathbb{Q}(\rho)|) + \log(f(y|b;\beta)) - \tfrac{1}{2} b^T \mathbb{Q}(\rho)^{-1} b \\
&= -\tfrac{1}{2}\log(|\tilde{R}(\delta, \rho)|) + \log(f(y|b;\beta)) - \tfrac{1}{2} b^T \mathbb{Q}(\rho)^{-1} b
\end{aligned} \tag{5.17}$$

with $\tilde{R}(\delta, \rho) = \sum_{i=1}^{N} z_{(i)} D_{(i)}(\delta) \Sigma_{(i)}^{-1}(\delta) D_{(i)}^T(\delta) z_{(i)})^T \mathbb{Q}(\rho) + I$.

5.4 Methods for Crossed Random Effects

5.4.1 Penalized Quasi Likelihood Concept

There is much confusion on the terminology of penalized quasi likelihood (PQL). This term was used by Green (1987) in a semiparametric context. The PQL is a Laplace Approximation around b, e.g. Lin & Breslow (1996), Breslow & Lin (1995a) and Breslow & Clayton (1993). This is the most popular method to maximize Generalized Linear Mixed Models. The Laplace Approximation around b and β is implemented in the macro GLIMMIX and proc GLIMMIX in SAS (Wolfinger (1994)). It is just a slight modification since $k(\tilde{\delta}) = -\log(f(y|\tilde{b};\tilde{\beta})p(\tilde{b};\rho)$ instead of $k(\tilde{b})$ is used. In the glmmPQL-function in the r-package nlme the Laplace Approximation around b is implemented. Further notes are in Wolfinger & O'Connell (1993), Littell, Milliken, Stroup & Wolfinger (1996) and Vonesh (1996).

In penalized based concepts the joint likelihood-function $L(\delta, \rho)$, described in (5.9), is specified by the parameters of the covariance structure ρ and parameter $\delta^T = (\beta^T, b^T)$.

The idea of the penalized quasi-likelihood is now to ignore the first term in (5.17), hoping that there is small variation in these terms within the iterative estimation. So

$$l_p(\delta, \rho) = \sum_{i=1}^{N} \log(f(y_{(i)}|\delta)) - \frac{1}{2} b^T \mathbb{Q}(\rho)^{-1} b. \tag{5.18}$$

These equations can also be derived via the log-posterior. The posteriori distribution for δ given the data y is

$$g(\delta|y; Q(\rho)) := \frac{f(y|\delta)p(\delta; \mathbb{Q}(\rho))}{\int p(y|\delta)p(\delta; \mathbb{Q}(\rho))d\delta}.$$

The normalization constant $\int p(y|\delta)p(\delta; \rho)d\delta$ is not needed for maximizing the posterior regarding δ. A more convenient representation in comparison to the posterior is the log-posterior without normalization constant, which is more easy to derive

$$l_p(\delta; \rho) = \sum_{i=1}^{N} (\log(f(y_{(i)}|\delta))) - \frac{1}{2} b^T \mathbb{Q}(\rho) b.$$

PQL usually works within the profile likelihood concept. So we can distinguish between the estimation of δ given the plugged in estimation $\hat{\rho}$ resulting in the profile-likelihood

$$l_p(\delta, \hat{\rho})$$

and the estimation of ρ given the plugged in estimator $\hat{\delta}$ resulting in the profile-likelihood

$$l_p(\hat{\delta}, \rho).$$

Estimation of β and b for fixed ρ: First we consider the maximation of $l_p(\delta, \rho)$, where β and b_i are estimated.

$$\begin{aligned} s_\beta &= \tfrac{\partial l_p(\delta,\rho)}{\partial \beta} = \textstyle\sum_{i=1}^{N} x_{(i)} D_{(i)}(\delta) \Sigma_{(i)}^{-1}(\delta)(y_{(i)} - \mu_{(i)}(\delta), \\ s_b &= \tfrac{l_p(\delta,\rho)}{\partial b} = \textstyle\sum_{i=1}^{N} z_{(i)} D_{(i)}(\delta) \Sigma_{(i)}^{-1}(\delta)(y_{(i)} - \mu_{(i)}(\delta)) - \mathbb{Q}^{-1}(\rho) b. \end{aligned} \tag{5.19}$$

As described in Breslow & Clayton (1993) the solution of $s(\delta) = s(\beta, b) = (s_\beta, s_b)^T = 0$ via Fisher-Scoring is equivalently to iteratively solving the BLUP-equations with a linearized version. The aspect of a linearized $\tilde{y}$

$$\tilde{y}(\delta) = X\beta + \mathbb{Z}b + (D(\delta)^{-1})^T(y - \mu(\delta)) \tag{5.20}$$

with

$$W = W(\delta) = D(\delta)\Sigma^{-1}(\delta)D^T(\delta),$$
$$D(\delta) = \text{bdiag}(D(\delta)_{(i)})_{i=1,\ldots,N},$$
$$\Sigma(\delta) = \text{bdiag}(\Sigma(\delta)_{(i)})_{i=1,\ldots,N}.$$

The corresponding BLUP-equations, which are iteratively solved in s, are

$$\begin{bmatrix} X^TWX & XW\mathbb{Z} \\ \mathbb{Z}^TWX & \mathbb{Z}^TW\mathbb{Z} + \mathbb{Q}(\rho)^{-1} \end{bmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{bmatrix} X^TW\tilde{y}(\delta^{(s)}) \\ \mathbb{Z}^TW\tilde{y}(\delta^{(s)}) \end{bmatrix} \tag{5.21}$$

with $(\delta^{(s+1)})^T = (\beta^T, b^T)$, where $\delta^{(s)}$ is the estimate in the s-th Fisher-Scoring cycle.

Example 5.1 : Special case: Estimation of β and b in clustered data
The components of the score function $s(\delta) = \frac{\partial l(\delta,\rho)}{\partial \delta} = (s_\beta, s_{b_1}, \ldots, s_{b_n})^T$ for fixed ρ are then given by

$$s_\beta = \frac{\partial l(\delta)}{\partial \beta} = \sum_{i=1}^n \sum_{t=1}^T X_{it}^T D_{it} \Sigma_{it}^{-1}(\delta)(y_{it} - \mu_{it}),$$
$$s_{b_i} = \frac{\partial l(\delta)}{\partial b_i} = \sum_{t=1}^T Z_{it}^T D_{it} \Sigma_{it}^{-1}(\delta)(y_{it} - \mu_{it}) - Q^{-1}(\rho) b_i$$

with $D_{it} = \frac{\partial h(\eta_{it})}{\partial \eta}, \Sigma_{it} = cov(y_{it}|\beta, b_i)$ and $\mu_{it} = h(\eta_{it})$. The expected conditional Fisher matrix has the shape

$$F(\delta) = \begin{bmatrix} F_{\beta\beta} & F_{\beta 1} & F_{\beta 2} & \ldots & F_{\beta n} \\ F_{1\beta} & F_{11} & & & 0 \\ F_{\beta 2} & & F_{22} & & \\ \vdots & & & & \\ F_{n\beta} & 0 & & & F_{nn} \end{bmatrix}$$

with

$$F_{\beta\beta} = \sum_{i=1}^n \sum_{t=1}^T X_{it}^T D_{it} \Sigma_{it}^{-1} D_{it}^T X_{it},$$

$$F_{\beta i} = F_{i\beta}^T = \sum_{t=1}^T X_{it}^T D_{it} \Sigma_{it}^{-1} D_{it}^T Z_{it},$$

$$F_{ii} = \sum_{t=1}^T Z_{it}^T D_{it} \Sigma_{it}^{-1} D_{it}^T Z_{it} + Q(\rho)^{-1}.$$

The estimator $\hat{\delta}$ can be calculated by the equation

$$F(\delta^{(k)})\delta^{(k+1)} = F(\delta^{(k)})\delta^{(k)} + s(\delta^{(k)}). \tag{5.22}$$

The problem 5.22 can be rewritten by linearized version

$$\tilde{y}(\delta) \quad = X\beta + \mathbb{Z}b + (D(\delta)^{-1})^T(y - \mu(\delta)) \tag{5.23}$$

to BLUP-equations

$$\begin{bmatrix} X^T D\Sigma D^T X & XD\Sigma D^T\mathbb{Z} \\ \mathbb{Z}^T D\Sigma D^T X & \mathbb{Z}D\Sigma D^T\mathbb{Z} + \mathbb{Q}(\rho)^{-1} \end{bmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{bmatrix} X^T D\Sigma D^T \tilde{y}(\delta^{(s)}) \\ \mathbb{Z}^T D\Sigma D^T \tilde{y}(\delta^{(s)}) \end{bmatrix}. \tag{5.24}$$

□

Estimation of $\mathbb{Q}(\rho)$ for fixed β and b: The theory of linear mixed models within the REML framework can be applied to estimate the variance parameters. So a $V(\delta, \rho)$ can be constructed with

$$V(\rho) := V(\delta, \rho) = D(\delta)\Sigma^{-1}(\delta)D^T(\delta) + \mathbb{Z}\mathbb{Q}(\rho)\mathbb{Z}^T$$

which may be seen as an approximation to cov$(h(X\beta + \mathbb{Z}b) + e)$. The penalized quasi likelihood can then be optimized with respect to ρ using the weighted least squares equations

$$l_p(\delta, \rho) \quad \approx -\tfrac{1}{2}\log(|V(\rho)|) + (\tilde{y}(\delta) - X\beta)^T V(\rho)^{-1}(\tilde{y}(\delta) - X\beta).$$

with $\tilde{y}(\delta) = X\beta + \mathbb{Z}b + (D(\delta)^{-1})^T(y - \mu(\delta))$. The restricted maximum log-likelihood is obtained by adding the term $-\frac{1}{2}\log(|X^T V(\rho)X|)$

$$l_r(\delta, \rho) \quad \approx -\tfrac{1}{2}\log(|V(\rho)|) + (\tilde{y}(\delta) - X\beta)^T V(\rho)^{-1}(\tilde{y}(\delta) - X\beta) - \tfrac{1}{2}\log(|X^T V(\rho)X|).$$

Differentiation with respect to $\rho^T = (\rho_1, \ldots, \rho_d)$ yields

$$\begin{aligned} &s(\beta, \rho) = \tfrac{\partial l_r(\beta,\rho)}{\rho} = (s(\rho)_i)_{i=1,\ldots,d} \\ &\text{and} \\ &F(\beta, \rho) = -E(\tfrac{\partial^2 l_r(\beta,\rho)}{\partial\rho\partial\rho^T}) = (F(\rho)_{i,j})_{i,j=1,\ldots,d}. \end{aligned}$$

The score function has elements

$$\begin{aligned} s(\rho)_i = \tfrac{\partial l_r(\rho)}{\rho_i} = \quad & -\tfrac{1}{2}\text{trace}\left(P(\rho)\tfrac{\partial V(\rho)}{\rho_i}\right) \\ & +\tfrac{1}{2}(\tilde{y}(\beta, b) - X\beta)^T V(\rho)^{-1}\tfrac{\partial V(\rho)}{\rho_i}V(\rho)^{-1}(\tilde{y}(\beta, b) - X\beta) \end{aligned}$$

with P defined in Harville (1977) and Breslow & Clayton (1993)

$$P(\rho) = V(\rho)^{-1} - V(\rho)^{-1}X\left(X^T V(\rho)^{-1}X\right)^{-1}X^T V(\rho)^{-1}. \tag{5.25}$$

The Fisher function has elements

$$F(\rho)_{i,j} = \frac{1}{2}\text{trace}\left(P\frac{\partial V(\rho)}{\rho_i}P\frac{\partial V(\rho)}{\rho_j}\right).$$

If ML is preferred to REML then $P(\rho)$ from (5.25) is replaced with $P(\rho) = V(\rho)^{-1}$.

The penalized quasi likelihood is maximized by the following algorithm.

Compute starting values $\hat{\beta}_0$ and $\hat{\theta}_0$. $\hat{\beta}_0$ can be the estimator of a linear model. The elements of θ_0 are set to be small values, i.e. 0.1.

1. set $k = 0$
2. compute $\hat{\beta}^{(k+1)}$ by solving the equation $l(\beta, \hat{\theta})$ above with plugged in $\hat{\theta}^{(k)}$
3. compute $\hat{\theta}^{(k+1)}$ in $l(\hat{\beta}, \theta)$ by running a Fisher scoring algorithm with plugged in $\hat{\theta}^{(k+1)}$.
4. stop, if all stopping criteria are reached, else start in 1 with $k = k + 1$.

Example 5.2 : Special case:Estimation of $Q(\rho)$ in clustered data

In this case computation is simplified since one works on blockdiagonal structures.

$$V_i(\rho) := V_i(\delta, \rho) = D_i(\delta)\Sigma_i^{-1}(\delta)D_i^T(\delta) + Z_iQ(\rho)Z_i^T.$$

The corresponding restricted maximum loglikelihood looks like

$$l_r(\delta, \rho) \approx -\frac{1}{2}\sum_{i=1}^{n}\log(|V_i(\rho)|) + \sum_{i=1}^{n}(\tilde{y}_i(\delta) - X_i\beta)^T V_i(\rho)^{-1}(\tilde{y}_i(\delta) - X_i\beta) - \frac{1}{2}\sum_{i=1}^{n}\log(|X_i^T V_i(\rho)X_i|).$$

The Score function simplifies to

$$\begin{aligned} s(\rho)_i = \frac{\partial l_r(\rho)}{\rho_i} = \quad & -\tfrac{1}{2}\textstyle\sum_{k=1}^{n} \textit{spur}\left(P_k(\rho)\frac{\partial V_k(\rho)}{\rho_i}\right) \\ & +\tfrac{1}{2}\textstyle\sum_{k=1}^{n}(\tilde{y}(\delta^{(k)}) - X_k\beta)^T V_k(\rho)^{-1}\frac{\partial V_k(\rho)}{\rho_i}V_k(\rho)^{-1}(\tilde{y}(\delta^{(k)}) - X_k\beta) \end{aligned}$$

with P_k

$$P_k(\rho) = V_k(\rho)^{-1} - V_k(\rho)^{-1}X_k\left(\sum_{k=1}^{n}X_k^T V_k(\rho)^{-1}X_k\right)^{-1}X_k^T V_k(\rho)^{-1}. \tag{5.26}$$

The Fisher function simplifies to

$$F(\rho)_{i,j} = \frac{1}{2}\sum_{k=1}^{n}\textit{trace}\left(P_k\frac{\partial V_k(\rho)}{\rho_i}P_k\frac{\partial V_k(\rho)}{\rho_j}\right).$$

□

Schall (1991) used the idea of estimating mixed models based on working observations. Breslow & Clayton (1993) put this idea in the framework of Laplace-Approximation and viewed the relationship to PQL, which is often used in semiparametric context. An alternative introduction to PQL is given by McGilchrist (1994) or Engel & Keen (1994). For binomial data PQL was applied by Gilmour, Anderson & Rae (1985). For ordered data see Harville & Mee (1984). Exact covariance in logistic mixed models has been proposed by Drum & McCullagh (1993). Their method may be seen as a method for marginal modelling. In fact using PQL or using methods for marginal modelling is based on the same equations in this context.

5.4.2 Bias Correction in Penalized Quasi Likelihood

Since PQL has been criticize in models with binary response, Breslow & Lin (1995a) and Lin & Breslow (1996) found a method to improve the bias in PQL. An analogous bias-corrected procedure was considered by Goldstein & Rasbash (1996) who suggested using an adjusted second-order Laplace approximation. Lin & Breslow (1996) studies on the bias are based on the Solomon-Cox-Approximation Solomon & Cox (1992), which is used to find correction terms for the PQL. Therefore the integrated quasi likelihood can be written as

$$L(\beta,\rho) = \exp^{l(\beta,\rho)} \propto |\mathbb{Q}(\rho)|^{-1/2} \int \exp\left\{\sum_{i=1}^{N} l_{(i)}(\beta,b) - \frac{1}{2} b^T \mathbb{Q}(\rho)^{-1} b\right\} db, \quad (5.27)$$

where $l_{(i)}(\beta,b) \propto \int_{y_{(i)}}^{\mu_{(i)}(\beta,b)} \Sigma_{(i)}^{-1}(\beta,b)(y_{(i)} - b)db$. Solomon and Cox approximated 5.27 by expanding $\sum_{i=1}^{N} l_{(i)}(\beta,b)$ in Taylor series about $b = 0$ before integration. The assumptions are

$$g(\mu) = \eta = X\beta + \mathbb{Z}b \quad (5.28)$$

with $\mathbb{Z}$, which is a partitioned matrix with $\mathbb{Z} = [Z_{.(1)}, \ldots, Z_{.(c)}]$, where $Z_{.(i)}$ is the design matrix associated with the i-th random effect b_i. b is assumed to be $\text{cov}(b) = \text{bdiag}(\rho_1^2 I_{q_1}, \ldots, \rho_c^2 I_{q_c})$.

The Solomon-Cox approximation is given by

$$\begin{aligned} l_{sol}(\beta,\rho) = \quad & -\tfrac{1}{2}\log|I + \mathbb{Z}^T\Sigma(\beta,b)\mathbb{Z}\mathbb{Q}(\rho)| \\ & + \textstyle\sum_{i=1}^{N} l_i(\beta,0) + \tfrac{1}{2} r(\beta,0)\mathbb{Z}\mathbb{Q}(\rho)(I + \mathbb{Z}^T\Sigma(\beta,0)\mathbb{Z}\mathbb{Q}(\rho))^{-1}\mathbb{Z}^T r(\beta,0) \end{aligned} \quad (5.29)$$

where $r(\beta,b) = \Sigma(\beta,b)^{-1}(y - X\beta - \mathbb{Z}b)$ may be seen as residuals. We denote $H^{(2)} = \{h_{ij}^2\}$ for any matrix H.

Important for latter computations are

$$\begin{aligned}
\tilde{\Sigma}(\beta, b) &= \text{diag}(\tfrac{\partial v(\mu_i(\beta,b))}{\partial \mu_i(\beta,b)} v(\mu_i(\beta, b)), \\
J &= \text{diag}(1_{q_1}, \ldots, 1_{q_c}), \\
\tilde{\tilde{\Sigma}}(\beta, b) &= \text{diag}(\tfrac{\partial^2 v(\mu_i(\beta,b))}{\partial^2 \mu_i(\beta,b)} (v(\mu_i(\beta, b)))^2 + (\tfrac{v(\mu_i(\beta,b))}{\partial \mu_i(\beta,b)})^2 v(\beta, b)), \\
B &= -\tfrac{1}{2} X^T \tilde{\Sigma}(\beta, 0) \mathbb{Z}^{(2)} J, \\
C &= \tfrac{1}{2} J^T \left(\mathbb{Z}^T \Sigma(\beta, 0) \mathbb{Z}\right)^{(2)} J + \tfrac{1}{4} J^T \mathbb{Z}^{(2)^T} \tilde{\tilde{\Sigma}}(\beta, 0) \mathbb{Z}^{(2)} J, \\
&\quad -B^T \left(X^T \Sigma(\beta, 0) X\right)^{-1} B, \\
C_P &= \tfrac{1}{2} J^T \left(\mathbb{Z}^T \Sigma(\beta, 0) \mathbb{Z}\right)^{(2)}, \\
G &= C^{-1} C_P.
\end{aligned} \tag{5.30}$$

Lin & Breslow (1996) propose the following algorithm

1. Get estimates $\hat{\beta}^{(0)}$ and $\hat{\rho}^{(0)}$ from penalized quasi likelihood estimation as described in subsection 5.4.1.
2. Correct $\hat{\rho}^{(0)}$ by $\hat{\rho}^{(1)} = G\hat{\rho}^{(0)}$
3. Use $\hat{\rho}^{(1)}$ to estimate β by solving the PQL-equations for β, which leads to $\hat{\beta}^{(1)}$.
4. Correct $\hat{\beta}^{(1)}$ by
$$\hat{\beta}^{(2)} = \hat{\beta}^{(1)} - (X^T \Sigma(\hat{\beta}^{(1)}, 0) X)^{-1} B \hat{\rho}^{(1)}$$
and
$$\hat{\beta}^{(3)} = \hat{\beta}^{(2)} + X^T \Sigma(\hat{\beta}^{(1)}, 0) X)^{-1} A(\hat{\beta}^{(1)}, 0)$$
where
$$A(\beta, b) = \frac{1}{2} \sum_{j=1}^{d} \sum_{k=1}^{d} X^T \tilde{\Sigma}(\beta, b)[(X_j X_j^T \Sigma(\beta, b) X_k) X_k] 1_{q_k} \hat{\rho}_j \hat{\rho}_k.$$

$\beta^{(3)}$ is called the bias corrected estimator of a generalized linear mixed model.

5.4.3 Alternative Direct Maximization Methods

MCMC integration based methods In the general case one can use Monte Carlo Marcov Chain based on a Metropolis-Hasting Algorithm as suggested by McCulloch (1997). Alternative the Gibbs-Sampler proposed by McCulloch (1994) in connection with numerical integration can be used. The main difference to Gauss-Hermite Quadrature is that the

points d_k are not fixed in the Newton-Raphson-Algorithm. The points must be computed new in every step.

To specify a Metropolis algorithm, the candidate distribution $c(b)$ must be specified, from which potential new values are drawn, as well as the acceptance function that gives the probability of accepting the new value.

The analysis is based on

$$l(\beta, \rho) = \log \int f(y|\beta; b)p(b; \rho)db, \tag{5.31}$$

where b is a q-dimensional vector, $p(b; \rho)$ is the density of a q-dimensional normal distribution with covariance $\mathbb{Q}(\rho)$. The idea is now to generate m drawings $b^1, \ldots, b^m$ from $f_{b|y}(b|y; \beta, \rho) \propto f(y|\beta; b) * p(b; \rho)$.

Since

$$\begin{aligned}
\tfrac{\partial l(\beta,\rho)}{\partial \beta} &= \tfrac{\partial}{\partial \beta} \log \textstyle\int f(y|\beta; b)p(b; \rho)du \\
&= \frac{\int \left[\frac{\partial}{\partial \beta} \log(f(y|\beta;b))\right] f(y|\beta;b)p(b;\rho)du}{\int f(y|\beta;b)p(b;\rho)db} \\
&\propto \textstyle\int (\frac{\partial}{\partial \beta} \log(f(y|\beta, b)))f(y|\beta; b)p(b; \rho)db \\
&\propto \textstyle\int (\frac{\partial}{\partial \beta} \log(f(y|\beta; b)f_{b|y}(b|y; \beta; \rho)db),))db
\end{aligned} \tag{5.32}$$

the integral of (5.32) may be approximated by

$$s(\beta) = \frac{\partial l(\beta, \rho)}{\partial \beta} = \sum_{k=1}^{m} \frac{1}{m} \frac{\partial}{\partial \beta} \log(f(y|\beta; b^k)). \tag{5.33}$$

The difficulty now is to find a good set $b^1, \ldots, b^m$. This problem is solved by the Metropolis algorithm. Let b^k denote $b^k = (b_1^k, \ldots, b_q^k)^T$. Generate a new value b_j^{k*} for the j-th component and accept $b^{k*} = (b_1^k, \ldots, b_{j-1}^k, b_j^{k*}, b_{j+1}^k, \ldots, b_q^k)^T$ as a new value with probability $A_j(b^k, b^{k*})$; otherwise retain b. $A_j(b^k, b^{k*})$ is given by

$$A_j(b^k, b^{k*}) = \min \left\{ 1, \frac{f_{b^{k*}|y}(b^{k*}|y; \beta^{(p)}, \rho^{(p)}) * c(b^k)}{f_{b^k|y}(b^k|y; \beta^{(p)}, \rho^{(p)}) * c(b^{k*})} \right\}.$$

If choosing $p(b; \rho)$ as the candidate distribution $c(b)$, then

$$A_j(b^k, b^{k*}) = \min \left\{ 1, \frac{f(y|\beta^{(p)}; b^{k*})}{f(y|\beta^{(p)}; b^k)} \right\}.$$

This procedure has to repeated for every component $b_j^k, j = 1, \ldots, q$. For small q b^k might be drawn and updated directly in a block. For larg q the acceptance probabilities may become very small, therefore componentwise drawings and updates as described should be preferred. There are only small modifications in the score Function and observed Fisher function to do when using Monte Carlo.

Since the vector ρ must be determined, $\mathbb{Q}(\rho)$ is chosen to maximize $\sum_{k=1}^m \frac{1}{m} f_b(b^k|\mathbb{Q}(\rho))$. This is done by a fisher scoring with

$$\begin{aligned} s(\rho) &= \frac{\partial \log \sum_{k=1}^m f_b(b^k|\mathbb{Q}(\rho)}{\partial \rho} \\ &= \sum_{k=1}^m \left(-\frac{1}{2} \text{trace} \left(\mathbb{Q}^{-1}(\rho) \frac{\partial \mathbb{Q}(\rho)}{\partial \rho} \right) + \frac{1}{2} (b^k)^T \mathbb{Q}^{-1}(\rho) \frac{\partial \mathbb{Q}(\rho)}{\partial \rho} \mathbb{Q}^{-1}(\rho) b^k \right) \end{aligned}$$

and

$$F(\rho)_{i,j} = \frac{\partial \log \sum_{k=1}^m f_b(b^k|\rho)}{\partial \rho_i \rho_j} = \sum_{k=1}^m \frac{1}{2} \text{trace} \left(\mathbb{Q}^{-1}(\rho) \frac{\partial \mathbb{Q}(\rho_i)}{\partial \rho_i} \mathbb{Q}^{-1}(\rho) \frac{\partial \mathbb{Q}(\rho_j)}{\partial \rho_j} \right)$$

The MCMC-Newton-Raphson algorithm has the form

1. Generate starting values β^0 and $\rho^{(0)}$. Set $s = 0$.
2. Generate m values $b^1, \ldots, b^m$ from $f_{b|y}(b|y; \beta^{(s)}, \rho^{(s)})$ and run Fisher-Scoring with respect to β.
3. Maximize $\sum_{k=1}^m f_b(b^k|\mathbb{Q}(\rho))$ with respect to ρ.
4. If convergence is achieved, then declare $\beta^{(s+1)}$ and $\rho^{(s+1)}$ to be MLE. Otherwise start in (2).

Another idea is based on Gelfand & Carlin (1993) and Geyer & Thompson (1992) which suggested to simulate the likelihood directly instead of using the log-likelihood. The simulated likelihood is then maximized directly. This methods is known under the name SML (simulated maximium likelihood).

5.4.4 Indirect Maximization using EM-Algorithm

MC-EM Algorithm - Booth and Hobert's method This method is based on importance sampling. Important for the latter analysis is

$$l(\beta, \rho) = \log \int -k(\tilde{b}) d\tilde{b}$$

with $k(\tilde{b}) = -\log(f(y|\tilde{b};\beta) * p(\tilde{b};\rho))$. The first moment with respect to b is $k'(b)$ where $k'(b)$ is described in (5.10) and $k''(b)$ in (5.11). The second moment is $k''(b)$.

The problem in EM-algorithm is to evaluate

$$\begin{aligned} M\left(\delta|\delta^{(p)}\right) &= \mathbb{E}\left\{\log f\left(y,b;\ \delta)|Y;\ \delta^{(p)}\right)\right\} \\ &= \int \log\left(f(y,b;\ \delta)\right) f\left(b|y,\ \delta^{(p)}\right) db, \end{aligned} \tag{5.34}$$

where b is a q-dimensional vector and $\delta^T = (\beta^T, \rho^T)$. A natural choice for this case is to use $N(k'(b|\delta^{(p)}), k''(b|\delta^{(p)}))$ for the importance sampling density $c(b;\delta^{(p)})$. More information is given in Wei & Tanner (1990). We approximate

$$M\left(\delta|\delta^{(p)}\right) \approx \sum_{k=1}^{m} w_k(\delta^{(p)}, b^k)\left\{\log f\left(y, b^k;\delta\right)\right\} \tag{5.35}$$

with importance weights

$$w_k(\delta^{(p)}, b^k) = \frac{f\left(y|\beta^{(p)}, b^k\right) p(b^k;\rho^{(p)})}{c(b^k;\delta^{(p)})}$$

by drawing vectors $(b^k)^T = ((b_1^k)^T, \ldots, (b_q^k)^T)$, $k = 1, \ldots, m$ from $N(k'(b|\delta^{(p)}), k''(b|\delta^{(p)}))$. Since $f(b|y;\delta)$ involves an unknown normalization constant, so do the weights. Details can be found in Booth & Hobert (1999) However, the normalization constant depends on the known value $\delta^{(p)}$ and not on δ, which means that it has no effect on the M-Step and is therefore irrelevant (see Sinha, Tanner & Hall (1994)). The score functions with $D = D(\beta;b^k) = \text{diag}\left(\frac{\partial h(\eta_{(i)})}{\partial \eta_{(i)}}\right)_{i=1,\ldots,n}$, $\eta^T = (\eta_{(1)}, \ldots, \eta_{(N)})^T$, $\eta_{(i)} = x_{(i)}^T\beta + z_{(i)}^T b^k$, $\Sigma = \text{bdiag}\left(\Sigma_{(i)}\right)_{i=1,\ldots,n}$, $\Sigma_{(i)} = \text{cov}(y_{(i)}|b^k)$ are given by

$$\begin{aligned} \frac{\partial \tilde{M}\left(\delta|\delta^{(p)}\right)}{\partial \beta} &= \sum_{k=1}^m w_k(\delta^{(p)}, b^k)\left\{\log f\left(y|b^k;\beta\right)\right\} \\ &= \sum_{k=1}^m w_k(\delta^{(p)}, b^k) X D \Sigma^{-1}(y-\mu), \end{aligned}$$

$$\begin{aligned} \frac{\partial \tilde{M}\left(\delta|\delta^{(p)}\right)}{\partial \rho} &= \sum_{k=1}^m w_k(\delta^{(p)}, b^k)\left\{\log p\left(b^k;\rho\right)\right\} \\ &= \sum_{k=1}^m w_k(\delta^{(p)}, b^k)\left(-\tfrac{1}{2}\text{trace}\left(\mathbb{Q}^{-1}(\rho)\frac{\partial \mathbb{Q}(\rho)}{\partial \rho}\right) + \tfrac{1}{2}(b^k)^T\mathbb{Q}^{-1}(\rho)\frac{\partial \mathbb{Q}(\rho)}{\partial \rho}\mathbb{Q}^{-1}(\rho)b^k\right). \end{aligned}$$

For the following set $s(\delta|\delta^{(p)})^T = \left(\left(\frac{\partial \tilde{M}\left(\delta|\delta^{(p)}\right)}{\partial \beta}\right)^T, \left(\frac{\partial \tilde{M}\left(\delta|\delta^{(p)}\right)}{\partial \rho}^T\right)^T\right)$.

Booth & Hobert (1999) suggest the following algorithm

1. Choose starting values $\delta^{(0)}$ and initial sample size m. Set p=0.
2. At iteration $p+1$ generate $u^k, k = 1, \ldots, m$ from the importance distribution
3. Maximize $M(\delta, \delta^{(p)})$ using the samples $b^k, k = 1, \ldots, m$
4. Construct a 100(1-α)% confidence ellipsoid for $\delta^{*(p+1)}$. If $(\delta)^{(p)}$ is inside of the region, set $m = m + [m/l]$, where [ ] denotes integer part
5. If convergence is achieved, set $\delta^{(p+1)}$ to be the maximum likelihood estimate $\hat{\delta}$; otherwise, set $p = p + 1$ and return to 2.

Usually the values $\alpha = 0.25$,$l = 3$ and $m = 50$ are chosen. $\delta^{*(p+1)}$ is the theoretical value which maximizes $\frac{\partial M\left((\delta|\delta^{(p)}\right)}{\partial \delta} = 0$ with exact integration. Booth & Hobert (1999) show that $\delta^{(p+1)}$ is asymptotic normally distributed with $\delta^{*(p+1)}$ and cov$(\delta^{(p+1)}|\delta^{(p)})$, which is approximated by

$$\text{cov}(\delta^{(p+1)}|\delta^{(p)}) \approx F(\delta^{(p+1)}|\delta^{(p)})^{-1} \widehat{\text{cov}}\left(s(\delta^{*(p+1)}|\delta^{(p)})\right) F(\delta^{(p+1)}|\delta^{(p)})^{-1}$$

with

$$\begin{aligned}\widehat{\text{cov}}\left(s(\delta^{*(p+1)}|\delta^{(p)})\right) &= \tfrac{1}{m^2} \textstyle\sum_{k=1}^{m} \left(w_k(\delta^{(p)}, b^k) \tfrac{\partial}{\partial \delta} \log\{f(y, b^k|\delta^{(p)}\}\right) \\ &\quad * \left(w_k(\delta^{(p)}, b^k) \tfrac{\partial}{\partial \delta} \log\{f(y, b^k|\delta^{(p)}\}\right)^T.\end{aligned}$$

Booth & Hobert (1999) propose using a multivariate Student t importance density with the same moments as the normal importance distribution $c(b; \delta)$.

McCulloch's Method - MCMC-EM-Algorithm Instead of using Gauss-Hermite Quadrature one can use Monte Carlo Markov Chain based on a Metropolis-Hasting Algorithm as suggested by McCulloch (1997) or Chan & Kuk (1997). Alternative the Gibbs-Sampler described in McCulloch (1994) can be used. The integration points must be computed new in every expectation step.

To specify a Metropolis algorithm Tanner (1993), the candidate distribution $c(b)$ must be specified, from which potential new values are drawn, as well as the acceptance function that gives the probability of accepting the new value.

The analysis is based on

$$\begin{aligned} M\left(\delta|\delta^{(p)}\right) &= \mathbb{E}\left\{\log f\left(y, b;\ \delta)|y;\ \delta^{(p)}\right)\right\} \\ &= \int \log\left(f(y, b;\ \delta)\right) f\left(b|y,\ \delta^{(p)}\right) db, \end{aligned} \tag{5.36}$$

where b is a q-dimensional vector, $p(b;\rho)$ is the density of a q-dimensional normal distribution with covariance $\mathbb{Q}(\rho)$. The idea is now to generate m drawings $b^1,\ldots,b^m$ from $f_{b|y}(b|y;\beta^{(p)},\rho^{(p)}) \propto f(y|\beta^{(p)};b) * p(b;\rho^{(p)})$.

Then

$$\begin{array}{ll} \frac{\partial M(\delta|\delta^{(p)})}{\partial\delta} & = \frac{\partial}{\partial\delta}\int \log(f(y_i|\beta,b)p(b;\rho))f_{b|y}(b|y;\beta^{(p)},\rho^{(p)})db \\ & = \int \frac{\partial}{\partial\delta}\left(\log(f(y|\beta,b)p(b;\rho))\right) f_{b|y}(b|y;\beta^{(p)},\rho^{(p)})db \end{array} \tag{5.37}$$

may be approximated by

$$\frac{\partial \tilde{M}(\delta|\delta^{(p)})}{\partial\delta} = \sum_{k=1}^{m}\frac{1}{m}\frac{\partial}{\partial\delta}\left(\log(f(y|\beta,b^k)p(b^k;\rho))\right). \tag{5.38}$$

The difficulty now is to find a good set $b^1,\ldots,b^m$. This problem is solved by the Metropolis algorithm. Let b^k denote $b^k = (b_1^k,\ldots,b_q^k)^T$. Generate a new value b_j^{k*} for the j-th component and accept $b^{k*} = (b_1^k,\ldots,b_{j-1}^k,b_j^{k*},b_{j+1}^k,\ldots,b_q^k)^T$ as a new value with probability $A_j(b^k,b^{k*})$; otherwise retain b. $A_j(b^k,b^{k*})$ is given by

$$A_j(b^k,b^{k*}) = \min\left\{1,\frac{f_{b^{k*}|y}(b^{k*}|y;\beta^{(p)},\rho^{(p)}) * c(b^k)}{f_{b^k|y}(b^k|y;\beta^{(p)},\rho^{(p)}) * c(b^{k*})}\right\}.$$

If choosing $c(u) = p(u)$ then

$$A_j(b^k,b^{k*}) = \min\left\{1,\frac{f(y|\beta^{(p)};b^{k*})}{f(y|\beta^{(p)};b^k)}\right\}.$$

This procedure has to repeated for every component $b_j^k, j = 1,\ldots,q$. For small q b^k might be drawn and updated directly in a block. For larger q the acceptance probabilities become very small so componentwise drawings and updates as described should be preferred. There are only small modifications in the Score Function and observed Fisher Information Matrix to do when using Monte Carlo.

The MCMC-EM algorithm has the form

1. Generate starting values β^0 and $\rho^{(0)}$. Set $s = 0$.
2. Generate m values $b^1,\ldots,b^m$ from $f_{b|y}(b|y;\beta^{(p)},\rho^{(p)})$ to do the expectation step with modifications described above.
3. Run a Newton-Raphson algorithm with modifications described above.
4. If convergence is achieved, then declare $\beta^{(p+1)}$ and $\rho^{p+1)}$ to be MLE. Otherwise start in (2).

5.5 Methods for Clustered Data

5.5.1 Gauss-Hermite-Quadrature

This method is limited to the case of clustered data. Gauss-Hermite quadrature is one of the most commonly used techinques in integration theory and also applied widely to statistics (e.g. Naylor & Smith (1982). Hedeker & Gibbons (1996) developed a programme called MIXOR to get estimators within the Gauss-Hermite framework. The SAS procedure NLMIXED (SAS Institute Inc. (1999)) uses Gauss-Hermite quadrature. Information on the Gauss-Hermite Quadrature in the statistical context can be found in Liu & Pierce (1994). These computer programmes apply Fisher-Scoring algorithms, with no analytical form of the expected Fisher matrix. Then Gauss-Hermite quadrature has to be used once again to approximate the expectation of the second order derivatives. It is known that in some circumstances Fisher-Scoring algorithm may lead to invalid statistical inferences due to the use of the expected information matrix. This point was illustrated by Lesaffre & Spiessens (2001). According to Gilmour, Thompson & Cullis (1995) the observed information matrix is preferable in GLMM.

For GLMM, the integrated likelihood can be written as

$$L(\beta,\rho)=\prod_{i=1}^{n}\int f(y_i|b_i)p(b_i;\rho)db_i=\prod_{i=1}^{n}\int f(y_i|\beta,a_i)\tilde{p}(a_i)da_i \tag{5.39}$$

with $a_i = Q(\rho)^{-1/2}b_i$. $p(a_i)$ is the density function of a $N_c(0,I_c)$, c is the number of random components. First one has to build sets of Gauss-Hermite quadrature points and weights

$$\{d_k=\left(d_{k_1}^{(1)},\ldots,d_{k_{(c)}}^{(c)}\right)^T:1\le k_1\le m_1;\ldots;1\le k_c\le m_c\} \tag{5.40}$$

and

$$\{v_k=\left(v_{k_1}^{(1)},\ldots,v_{k_c}^{c}\right)^T:1\le k_1\le m_1;\ldots;1\le k_c\le m_c\}, \tag{5.41}$$

where $d_{k_j}^{(j)}$ and $v_{k_j}^{(j)}$ denote the univariate quadrature points and weights for component j and m_j is the number of quadrature points for the j-th component, $j=1,\ldots,c$. Then the Gauss-Hermite-Approximation to the log-likelihood has the form

$$l_{GH}(\beta,\rho)=\sum_{i=1}^{n}\log\left[\sum_{k_1=1}^{m_1}\cdots\sum_{k_c=1}^{m_c}\left(\frac{v_{k_1}^{(1)}}{\sqrt{\pi}}\right)\ldots\left(\frac{v_{k_c}^{(c)}}{\sqrt{\pi}}\right)f(y_i|\beta,\rho,d_k)\right] \tag{5.42}$$

with $\eta_i(d_k)=\eta_i(\beta,\rho,d_k)=X_i^T\beta+Z_i\sqrt{2}Q(\rho)^{1/2}d_k$. $\eta_{it}(d_k)=\eta_{it}(\beta,\rho,d_k)=x_{it}^T\beta+z_{it}^T\sqrt{2}Q(\rho)^{1/2}d_k$.

Let $\rho = vech(Q^{1/2})$ be the *symmetric direct operator* for the matrix $Q^{1/2}$, that is a vector formed by all the lower triangular entries of $Q^{1/2}$ through column by column.Denote $vec(Q^{1/2})$ as the *direct operator* for the matrix $Q^{1/2}$, in other words, the $c^2 \times 1$ vector formed by stacking the columns of Q under each other. According to Nel (1980) and Pan, Fang & van Rosen (1997) there must exist a $c^2 \times c^*$ matrix S_c with $vec(Q) = S_c * vech(Q)$ where $c^* = c(c+1)/2$.

$$\begin{aligned} \frac{\partial l(\beta,\rho)}{\partial \beta} &= \sum_{i=1}^{n} \sum_{k_1=1}^{m_1} \cdots \sum_{k_c=1}^{m_c} w_i(d_k) * \\ &\quad \left[\sum_{t=1}^{T} x_{it} D_{it} \Sigma_{it}^{-1} (y_{it} - h(\eta_{it}(d_k))\right] \end{aligned} \tag{5.43}$$

where $\Sigma_{it} = \text{cov}(y_{it}|\eta_{it}(d_k))$ and

$$w_i(d_k) := w_i(\beta, \rho, d_k) = \frac{\left(\frac{v_{k_1}^{(1)}}{\sqrt{\pi}}\right) \cdots \left(\frac{v_{k_c}^{(c)}}{\sqrt{\pi}}\right) f(y_i|\eta_i(d_k))}{\sum_{k_1=1}^{m_1} \cdots \sum_{k_c=1}^{m_c} \left(\frac{v_{k_1}^{(1)}}{\sqrt{\pi}}\right) \cdots \left(\frac{v_{k_c}^{(c)}}{\sqrt{\pi}}\right) f(y_i|\eta_i(d_k))}.$$

Similarly

$$\begin{aligned} \frac{\partial l(\beta,\rho)}{\partial \rho} &= \sum_{i=1}^{n} \sum_{k_1=1}^{m_1} \cdots \sum_{k_c=1}^{m_c} w_i(d_k) \\ &\quad \left[\sum_{t=1}^{T} \frac{\partial \eta_{it}}{\partial \rho} D_{it} \Sigma_{it}^{-1} (y_{it} - h(\eta_{it}(d_k))\right] \end{aligned} \tag{5.44}$$

with $\frac{\partial \eta_{ij}}{\partial \rho} = S_c^T (d_k^T \otimes z_{ij}^T)$.

For simplicity we suppress the notation $w_i(d_k)$ to w_i and $\eta_{it}(d_k)$ to η_{it} for the computation of the second derivatives.

$$\begin{aligned} \frac{\partial^2 l(\beta,\rho)}{\partial \beta \partial \beta^T} &= \sum_{i=1}^{n} [\sum_{k_1}^{m_1} \cdots \sum_{k_c}^{m_c} w_i \left(\sum_{t=1}^{T_i} x_{it} D_{it} \Sigma_{it}^{-1} (y_{it} - h(\eta_{it}))\right) * \\ &\qquad \left(\sum_{t=1}^{T_i} x_{it} D_{it} \Sigma_{it}^{-1} (y_{it} - h(\eta_{it}))\right)^T] \\ &- \sum_{i=1}^{n} [\sum_{k_1}^{m_1} \cdots \sum_{k_c}^{m_c} w_i \left(\sum_{t=1}^{T_i} x_{it} D_{it} \Sigma_{it}^{-1} (y_{it} - h(\eta_{it}))\right)] * \\ &\qquad [\sum_{k_1}^{m_1} \cdots \sum_{k_c}^{m_c} w_i \left(\sum_{t=1}^{T_i} x_{it} D_{it} \Sigma_{it}^{-1} (y_{it} - h(\eta_{it}))\right)]^T \\ &- \sum_{i=1}^{n} [\sum_{k_1}^{m_1} \cdots \sum_{k_c}^{m_c} w_i \left(\sum_{t=1}^{T_i} x_{it} D_{it} \Sigma_{it}^{-1} D_{it}^T x_{it}^T\right)], \end{aligned} \tag{5.45}$$

$$
\begin{aligned}
\frac{\partial^2 l(\beta,\rho)}{\partial\rho\partial\beta^T} &= \sum_{i=1}^{n} \quad [\sum_{k_1}^{m_1}\dots\sum_{k_c}^{m_c} w_i \quad \left(\sum_{t=1}^{T_i} S_c^T (d_k^T \otimes z_{it})^T D_{it}\Sigma_{it}^{-1}(y_{it}-h(\eta_{it}))\right) * \\
&\qquad\qquad \left(\sum_{t=1}^{T_i} x_{it} D_{it}\Sigma_{it}^{-1}(y_{it}-h(\eta_{it}))\right)^T] \\
&- \sum_{i=1}^{n} \quad [\sum_{k_1}^{m_1}\dots\sum_{k_c}^{m_c} w_i \quad \left(\sum_{t=1}^{T_i} S_c^T (d_k^T \otimes z_{it}^T) D_{it}\Sigma_{it}^{-1}(y_{it}-h(\eta_{it}))\right)] * \\
&\qquad\qquad [\sum_{k_1}^{m_1}\dots\sum_{k_c}^{m_c} w_i \quad \left(\sum_{t=1}^{T_i} x_{it} D_{it}\Sigma_{it}^{-1}(y_{it}-h(\eta_{it}))\right)]^T \\
&- \sum_{i=1}^{n} \quad [\sum_{k_1}^{m_1}\dots\sum_{k_c}^{m_c} w_i \quad \left(\sum_{t=1}^{T_i} S_c^T (d_k^T \otimes z_{it}^T) D_{it}\Sigma_{it}^{-1} D_{it}^T x_{it}^T\right)]
\end{aligned} \tag{5.46}
$$

and

$$
\begin{aligned}
\frac{\partial^2 l(\beta,\rho)}{\partial\rho\partial\beta^T} &= \sum_{i=1}^{n} \quad [\sum_{k_1}^{m_1}\dots\sum_{k_c}^{m_c} w_i \quad \left(\sum_{t=1}^{T_i} S_c^T (d_k^T \otimes z_{it}^T) D_{it}\Sigma_{it}^{-1}(y_{it}-h(\eta_{it}))\right) * \\
&\qquad\qquad \left(\sum_{t=1}^{T_i} S_c^T (d_k^T \otimes z_{it}^T) D_{it}\Sigma_{it}^{-1}(y_{it}-h(\eta_{it}))\right)^T] \\
&- \sum_{i=1}^{n} \quad [\sum_{k_1}^{m_1}\dots\sum_{k_c}^{m_c} w_i \quad \left(\sum_{t=1}^{T_i} S_c^T (d_k^T \otimes z_{it}^T) D_{it}\Sigma_{it}^{-1}(y_{it}-h(\eta_{it}))\right)] * \\
&\qquad\qquad [\sum_{k_1}^{m_1}\dots\sum_{k_c}^{m_c} w_i \quad \left(\sum_{t=1}^{T_i} S_c^T (d_k^T \otimes z_{it}^T) D_{it}\Sigma_{it}^{-1}(y_{it}-h(\eta_{it}))\right)]^T \\
&- \sum_{i=1}^{n} \quad [\sum_{k_1}^{m_1}\dots\sum_{k_c}^{m_c} w_i \quad \left(\sum_{t=1}^{T_i} S_c^T (d_k^T \otimes z_{it}^T) D_{it}\Sigma_{it}^{-1} D_{it}^T (d_k^T \otimes z_{it}^T)^T S_c\right)].
\end{aligned} \tag{5.47}
$$

The Fisher-Scoring method is based on the work of Hedeker & Gibbons (1994) and Hedeker & Gibbons (1996). The second derivatives of the marginal likelihood (5.42) are substituted with their expectations. Since $w_i(\beta, \rho, d_k)$ depends on the parameters β and ρ this is very cumbersome. In this case Gauss-Hermite-Quadrature has to be used once again to solve the integral. A more straight forward way is to parameterize $w_i(\beta, \rho, d_k)$ by $w_i(\tilde{\beta}, \tilde{\rho}, d_k)$, where $\tilde{\beta}$ and $\tilde{\rho}$ are the estimates of the previous Fisher-Scoring-step.

$$
\begin{aligned}
F_{\beta\beta} = -\mathbb{E}(\tfrac{\partial^2 l(\beta,\beta)}{\partial\beta\partial\beta^T}) = & \sum_{i=1}^{n} [\sum_{k_1}^{m_1}\dots\sum_{k_c}^{m_c} w_i(\tilde{\beta},\tilde{\rho},d_k) \left(\sum_{t=1}^{T_i} x_{it} D_{it}\Sigma_{it}^{-1} D_{it}^T x_{it}^T\right)],\\
F_{\rho\beta} = -\mathbb{E}(\tfrac{\partial^2 l(\beta,\rho)}{\partial\rho\partial\beta^T}) = & \sum_{i=1}^{n} [\sum_{k_1}^{m_1}\dots\sum_{k_c}^{m_c} w_i(\tilde{\beta},\tilde{\rho},d_k) \left(\sum_{t=1}^{T_i} S_c^T(d_k^T\otimes z_{it}^T) D_{it}\Sigma_{it}^{-1} D_{it}^T x_{it}^T\right)],\\
F_{\rho\rho} = -\mathbb{E}(\tfrac{\partial^2 l(\rho,\rho^T)}{\partial\rho\partial\beta^T}) = & \sum_{i=1}^{n} [\sum_{k_1}^{m_1}\dots\sum_{k_c}^{m_c} w_i(\tilde{\beta},\tilde{\rho},d_k) \left(\sum_{t=1}^{T_i} S_c^T(d_k^T\otimes z_{it}^T) D_{it}\Sigma_{it}^{-1} D_{it}^T x_{it}^T\right)].
\end{aligned}
\tag{5.48}
$$

The Newton-Raphson Algorithm becomes now the Fisher-Scoring Algorithm.

5.5.2 Adaptive Gauss-Hermite Quadrature

The analysis starts with

$$
l(\delta;\rho) = \sum_{i=1}^{n} \log \int f(y_i|\tilde{b}_i;\beta) * p(\tilde{b}_i;\rho) d\tilde{b}_i = \sum_{i=1}^{n} \log \int \exp\{-k_i(\tilde{b}_i)\} d\tilde{b}_i. \tag{5.49}
$$

Basic idea is to combine Gauss-Hermite Quadrature and the equations in the Laplace-approximation for an adaptive approach. The terms (5.10) and (5.11) are used to refine the grid for integration. In this approach the grid of abscissas on the $\tilde{b}_i$ scale is centered around the conditional modes b_i rather than 0. Therefore we need

$$
\frac{\partial k_i(\tilde{b}_i)}{\partial \tilde{b}_i} = 0 = k_i'(b_i) \tag{5.50}
$$

yielding b_i.

Then compute

$$
\frac{\partial^2 k_i(\tilde{b}_i)}{\partial \tilde{b}_i \tilde{b}_i^T} = k''(\tilde{b}_i). \tag{5.51}
$$

$k_i'(b_i)$ and $k_i''(b_i)$ are in detail

$$
\begin{aligned}
k_i'(b_i) &= Z_i^T D_i \Sigma_i (y_i - \mu_i) - Q^{-1}(\rho) b,\\
k_i''(b_i) &= Z_i^T D_i \Sigma_i^{-1} D_i^T Z_i + Q^{-1}(\rho) + R_i
\end{aligned}
\tag{5.52}
$$

with $D_i = D_i(\beta, b_i)$,$\Sigma_i = \Sigma_i(\beta, b_i)$ and $\mu_i = \mu_i(\beta, b_i)$ and $\mathbb{E}(R_i) = 0$. Then we set $\tau_i^{-1} = \mathbb{E}\left(k''(b_i)\right)$ and $\tilde{b}_i = \tau_i^{1/2} a_i + b_i$, $\frac{\partial a_i}{\partial \tilde{b}_i} = |\tau_i^{-1/2}|$, where a_i is standard normal

distributed. A modification is the use of $k''(b_i)$ instead of $Q(\rho)$ in the scaling of the a_i. So the predictor is $\eta_i = X_i\beta + Z_i\tilde{b}_i = X_i\beta + Z_i(b_i + \tau_i^{1/2}a_i)$. So (5.49) can be rewritten to

$$\begin{aligned} l(\beta,\rho) &= \sum_{i=1}^n \log \int |\tau_i^{1/2}| \exp\{-k_i(b_i + \tau^{1/2}a_i\}da_i \\ &= \sum_{i=1}^n \int |\tau_i^{1/2}| \exp\{-k_i(b_i + \tau^{1/2}a_i) + \tfrac{p}{2}\log(2\pi) + \tfrac{1}{2}a_i^T a_i\} \qquad (5.53) \\ &\quad * \tfrac{1}{(2\pi)^{p/2}} \exp\{-\tfrac{1}{2}a_i^T a_i\}da_i. \end{aligned}$$

Taking the quadrature points (5.40) and quadrature weights (5.41) one obtains with $\eta_i(d_k) = X_i\beta + Z_i\tilde{b}_i = X_i\beta Z_i(b_i + \sqrt{2}\tau_i^{1/2}d_k)$ an approximation

$$\begin{aligned} \tilde{l}(\beta,\rho) &= \sum_{i=1}^n \log \left[\sum_{k_1}^{m_1}\ldots\sum_{k_c}^{m_c} |\tau_i^{1/2}| \exp\{-k_i(b_i + \sqrt{2}\tau_i^{1/2}d_k) + \tfrac{p}{2}\log(2\pi) + \tfrac{1}{2}||d_k||\}\right] \\ &\quad * \left(\tfrac{v_k}{(2\pi)^{p/2}}\right). \end{aligned} \qquad (5.54)$$

Since the parameters β and b_i should be obtained by maximizing (5.54) and not by solving the Laplace-Approximation iteratively, we replace b_i by b_i in (5.54). The scaling matrix τ_i depends on b_i,β and ρ, which causes computational problems for getting the score functions. That is why τ_i is computed using provisorial estimates $\hat{\beta}$, $\hat{b}_{i,}$ and $\hat{\rho}$, i.e. the estimates of the last iteration cycle. The score functions with $D_i = D_i(\eta_i(d_k))$,$\Sigma_i = \Sigma_i(\eta_i(d_k))$ and $\mu_i = \mu_i(d_k)$ have the form

$$s_\beta = \tfrac{\partial\tilde{l}(\delta,\rho)}{\partial\beta} = \sum_{i=1}^n\sum_{k_1}^{m_1}\ldots\sum_{k_c}^{m_c} w_i(\beta,\rho,d_k,b_i)X_iD_i\Sigma_i(y_i-\mu_i),$$

$$s_{b_i} = \tfrac{\partial\tilde{l}(\delta,\rho)}{\partial b_i} = \sum_{k_1}^{m_1}\ldots\sum_{k_c}^{m_c} w_i(\beta,\rho,d_k,b_i)Z_iD_i\Sigma_i^{-1}(y_i-\mu_i) - Q^{-1}(\rho)(b_i+\sqrt{2}\tau_i^{1/2}d_k) \qquad (5.55)$$

with

$$w_i(\beta,\rho,d_k,b_i) = \frac{\left(\frac{v_{k_1}^{(1)}}{\sqrt{\pi}}\right)\ldots\left(\frac{v_{k_c}^{(c)}}{\sqrt{\pi}}\right) f(y_i|\eta(d_k))w_{i,\text{corr}}(\beta,\rho,d_k)}{\sum_{k_1=1}^{m_1}\ldots\sum_{k_c=1}^{m_c}\left(\frac{v_{k_1}^{(1)}}{\sqrt{\pi}}\right)\ldots\left(\frac{v_{k_c}^{(c)}}{\sqrt{\pi}}\right) f(y_i|\eta(d_k))w_{i,\text{corr}}(\beta,\rho,d_k,b_i)}$$

and

$$w_{i,\text{corr}}(\beta,\rho,d_k,b_i) = \exp\{(b_i+\sqrt{2}\tau_i^{1/2}d_k)^TQ^{-1}(\rho)(b_i+\sqrt{2}\tau_i^{1/2}d_k) + ||d_k||\}.$$

For estimation $l(\delta, \rho)$ is profiled on the $\hat{\rho}$ on the one side to get estimates for δ and profiled on $\hat{\delta}$ on the other side to get estimates for ρ. This is usually realized by IWLS.

The use of Fisher scoring algorithm or Newton-Raphson given $\hat{\rho}$ is difficult because the weights $w_i(\beta, \hat{\rho}, d_k, b_i)$ depends on β,b_i . Nevertheless, if β and b_i in $w_i(\beta, \hat{\rho}_{pr}, d_k, b_i)$ are replaced with their provisorial estimates $\hat{b}_{i,}$, $\hat{\beta}$, then the dependence of $w_i(\hat{\beta}, \hat{\rho}, d_k, \hat{b}_i)$ may be ignored when calculating the second-order derivatives or the expected fisher information matrix. Instead an equation system can be solved iteratively. Therefore we need

$$\begin{aligned}
W_i &= \sum_{i=1}^{n}\sum_{k_1}^{m_1}\ldots\sum_{k_c}^{m_c} w_i(\beta, \rho, d_k, b_i) D_i \Sigma_i^{-1} D_i^T, \\
W &= \text{bdiag}(W_1 \ldots, W_n), \\
b_i^* &= \sum_{i=1}^{n}\sum_{k_1}^{m_1}\ldots\sum_{k_c}^{m_c} w_i(\beta, \rho, d_k, b_i)(\hat{b}_i + \sqrt{2}\tau^{1/2} d_k), \\
(b^*)^T &= ((b_1^*)^T, \ldots, (b_n^*)^T)^T \\
R_i &= w_i(\hat{\beta}, \rho, d_k, b_i) D_i \Sigma_i^{-1}, \\
R &= \text{bdiag}(R_1, \ldots, R_n).
\end{aligned} \tag{5.56}$$

By denoting $\mu^* = \sum_{k_1}^{m_1}\ldots\sum_{k_c}^{m_c} R\mu$ and $y^* = \sum_{k_1}^{m_1}\ldots\sum_{k_c}^{m_c} Ry$ we get the working vector

$$\tilde{y} = X\beta + \mathbb{Z}b + W^{-1}(y^* - \mu^*).$$

The solutions of δ must satisfy

$$\begin{bmatrix} X^T W X & X^T W \mathbb{Z} \\ \mathbb{Z}^T W X & \mathbb{Z}^T W \mathbb{Z} + \mathbb{Q}^{-1} \end{bmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{bmatrix} X^T W \tilde{y} \\ \mathbb{Z}^T W y + \mathbb{Q}^{-1}(b - b^*) \end{bmatrix} \tag{5.57}$$

The estimator for ρ is obtained by using the linearized version

$$y^* = X\beta + \mathbb{Z}b^* + W^{-1}(y^* - \mu^*).$$

It should be mentioned that for using only one quadrature point, the equation is the same that is solved in the Laplace-Approximation case. The estimation of the variance components can be done by maximizing a linear mixed model in ML or REML equations. More

details and refinements can be found in Pinheiro & Bates (1995),Pinheiro & Bates (2000) and SAS Institute Inc. (1999), since SAS uses this form of adaptive Gauss-Hermite-Quadrature in the procedure proc nlmixed.

5.5.3 Gauss-Hermite-Quadrature using EM-Algorithm

The marginal likelihood depends only on the structural parameters of the model. These are given by β and $Q = cov(b_i)$. Let Q be decomposed by $Q = Q^{1/2}Q^{T/2}$ where $Q^{1/2}$ denotes the left Cholesky factor. By simple matrix algebra the linear predictor may be written in the usual linear form with $b_i = Q^{1/2}a_i$

$$\begin{aligned} \eta_{it} &= x_{it}^T\beta + z_{it}^T Q^{1/2} a_i \\ &= \left[x_{it}, a_i^T \otimes z_{it}^T\right] \begin{bmatrix} \beta \\ \theta \end{bmatrix} \end{aligned} \tag{5.58}$$

where $a_i \sim N(0, I)$ is the standardized random variable and $\theta = vec\left(Q^{1/2}\right)$ is the symmetric diagonal operator. For univariate random effects the Kronecker product simplifies to $a_i z_{it}^T$ and $\theta = \sqrt{var(b_i)}$. By utilizing all of the parameters are collected in $\delta^T = (\beta^T\ \theta^T)$.

The indirect approach which is based on the EM algorithm avoids calculation of difficult derivatives in the case of Newton-Raphson algorithm or another approximation of the expected Fisher matrix in direct approaches. Since it is often used in literature is given more explicitly. In the $\mathbb{E}$-step of the $(p+1)$th cycle one has to determine

$$\begin{aligned} M\left(\delta|\delta^{(p)}\right) &= \mathbb{E}\left\{\log f\left(y, b;\ \delta)|y;\ \delta^{(p)}\right)\right\} \\ &= \int \log\left(f(y, b;\ \delta)\right) f\left(b|y,\ \delta^{(p)}\right) db^{'} \end{aligned}$$

where

$$\log f(y, b;\ \delta) = \sum_{i=1}^{n} \log f(y_i|a_i,\ \delta) + \sum_{i=1}^{n} \log\left(\tilde{p}(a_i)\right)$$

is the complete penalized data log likelihood with $y = (y_1, \ldots, y_n)$ denoting the observed data and $b = (b_1, \ldots, b_n)$ denoting the unobserved data and $a = (a_1, \ldots, a_n)$ the standardized unobserved data. $\tilde{p}(.)$ is the mixture distribution of the standardized random effects a_i. Basic idea is to use the theorem of Bayes

$$\tilde{p}(a|y, \delta^{(p)}) * \int f(y, a; \delta^{(p)}) da = f(y|a, \delta^{(p)}) * \tilde{p}(a).$$

Rewriting the problem the posterior has the simple form

$$f\left(a|y,\delta^{(p)}\right)=\prod_{i=1}^{n}f\left(y_i|a_i,\delta^{(p)}\right)\prod_{i=1}^{n}\tilde{p}(a_i)\Big/\prod_{i=1}^{n}\int f\left(y_i|a_i,\ \delta^{(p)}\right)\tilde{p}(a_i)da_i$$

$M\left(\delta|\delta^{(p)}\right)$ simplifies to

$$\begin{aligned}
M\left(\delta|\delta^{(p)}\right) &= \int \log\left(f(y,a;\ \delta)\right)p\left(a|y,\ \delta^{(p)}\right)da \\
&= \int \log\left(\textstyle\prod_{i=1}^{n}f(y_i,a_i;\ \delta)\right)\frac{\prod_{j=1}^{n}f\left(y_j|a_j,\delta^{(p)}\right)\prod_{j=1}^{n}\tilde{p}(a_j)}{\prod_{j=1}^{n}\int f\left(y_j|a_j,\ \delta^{(p)}\right)\tilde{p}(a_j)da_j}da \\
&= \textstyle\sum_{i=1}^{n}\int\ldots\int \log\left(f(y_i,a_i;\ \delta)\right)\frac{\prod_{j=1}^{n}f\left(y_j|a_j,\delta^{(p)}\right)\prod_{j=1}^{n}\tilde{p}(a_j)}{\prod_{j=1}^{n}\int f\left(y_j|a_j,\ \delta^{(p)}\right)\tilde{p}(a_j)da_j}da_1\ldots da_n \\
&= \textstyle\sum_{i=1}^{n}\int \log\left(f(y_i,a_i;\ \delta)\right)\frac{f\left(y_i|a_i,\delta^{(p)}\right)\tilde{p}(a_i)}{\int f\left(y_j|a_j,\ \delta^{(p)}\right)\tilde{p}(a_j)da_j}da_i \\
&= \textstyle\sum_{i=1}^{n}\int \left[\log f(y_i|a_i,\delta)+\log\tilde{p}(a_i)\right]\frac{f\left(y_i|a_i,\delta^{(p)}\right)\tilde{p}(a_i)}{\int f\left(y_j|a_j,\ \delta^{(p)}\right)\tilde{p}(a_j)da_j}da_i.
\end{aligned}$$

Then we need sets of Gauss-Hermite quadrature points and weights

$$\{d_k=\left(d_{k_1}^{(1)},\ldots,d_{k_{(c)}}^{c}\right)^T:1\le k_1\le m_1;\ldots;1\le k_c\le m_c\}$$

and

$$\{v_k=\left(v_{k_1}^{(1)},\ldots,v_{k_c}^{(c)}\right)^T:1\le k_1\le m_1;\ldots;1\le k_c\le m_c\}$$

where $d_{k_j}^{(j)}$ and $v_{k_j}^{(j)}$ denote the univariate quadrature points and weights for component j and m_j is the number of quadrature points for the j-th component, $j=1,\ldots,c$.

In a Gauss-Hermite type approximation which is used in the following one has the approximation

$$M\left(\delta|\delta^{(p)}\right)\approx\tilde{M}\left(\delta|\delta^{(p)}\right), \tag{5.59}$$

where

$$\tilde{M}\left(\delta|\delta^{(p)}\right)=\sum_{i=1}^{n}\left[\sum_{k_1=1}^{m_1}\cdots\sum_{k_c=1}^{m_c}w_{ik}\left[\log\left(f(y_i|\delta,d_k)\right)+\log(\tilde{p}(a_i))\right]\right] \tag{5.60}$$

with $\eta_{itk}=x_{it}^T\beta+[\sqrt{2}d_k^T\otimes z_{it}]\theta$, and $\eta_{i.k}^T=(\eta_{i1k},\ldots,\eta_{iT_ik})$

$$w_{ik}:=w_i(\delta^{(p)},d_k)=\frac{\left(\frac{v_{k_1}^{(1)}}{\sqrt{\pi}}\right)\ldots\left(\frac{v_{k_c}^{(c)}}{\sqrt{\pi}}\right)f\left(y_i|\delta,d_k\right))}{\sum_{k_1=1}^{m_1}\cdots\sum_{k_c=1}^{m_c}\left(\frac{v_{k_1}^{(1)}}{\sqrt{\pi}}\right)\ldots\left(\frac{v_{k_c}^{(c)}}{\sqrt{\pi}}\right)f(y_i|\delta,d_k)}.$$

The beauty of this approximation is that $M(\delta|\delta^{(p)})$ again corresponds to the penalized weighted log-likelihood of a generalized linear model and therefore maximization (the M step of the EM algorithm) is simply realized within the framework of GLMs. For simplicity we drop the notation of δ in brackets, so $\eta_{itk} = x_{it}\beta + \sqrt{2}z_{it}^T Q(\varrho)^{-1/2} d_k$, $D_{itk} = \frac{\partial \eta_{itk}}{\partial \delta}$ and $\Sigma_{itk} = \text{cov}(y_{it}|\delta, d_k)$. According to Nel (1980) and Pan, Fang & van Rosen (1997) there must exist a $c^2 \times c^*$ matrix S_c with $vec(Q) = S_c * vech(Q)$ where $c^* = c(c+1)/2$. The score functions are

$$\begin{aligned} s(\beta|\delta^{(p)}) &= \frac{\partial \tilde{M}(\delta|\delta^{(p)})}{\partial \beta} = \sum_{i=1}^n \sum_{k_1=1}^{m_1} \cdots \sum_{k_c=1}^{m_c} w_{ik} \sum_{t=1}^{T_j} x_{it} D_{itk} \Sigma_{itk}^{-1} (y_{it} - h(\eta_{itk})) \\ &\text{and} \\ s(\varrho|\delta^{(p)}) &= \frac{\partial \tilde{M}(\delta|\delta^{(p)})}{\partial \varrho} = \sum_{i=1}^n \sum_{k_1=1}^{m_1} \cdots \sum_{k_c=1}^{m_c} w_{ik} \sum_{t=1}^{T_j} S_c^T (d_k^T \otimes z_{it}^T) D_{itk} \Sigma_{itk}^{-1} (y_{it} - h(\eta_{itk})). \end{aligned} \tag{5.61}$$

The corresponding expected Fisher matrices are

$$\begin{aligned} \frac{\partial^2 \tilde{M}(\delta|\delta^{(p)})}{\partial \beta \partial \beta^T} &= \sum_{i=1}^n \sum_{k_1}^{m_1} \cdots \sum_{k_c}^{m_c} w_{ik} \sum_{t=1}^{T_j} x_{it} D_{it} \Sigma_{itk}^{-1} D_{itk}^T x_{itk}^T, \\ \frac{\partial^2 \tilde{M}(\delta|\delta^{(p)})}{\partial \beta \partial \varrho^T} &= \sum_{i=1}^n \sum_{k_1}^{m_1} \cdots \sum_{k_c}^{m_c} w_{ik} \sum_{t=1}^{T_j} x_{it} D_{itk} \Sigma_{itk}^{-1} D_{itk}^T (d_k^T \otimes z_{it})^T S_c \\ &\text{and} \\ \frac{\partial^2 \tilde{M}(\delta|\delta^{(p)})}{\partial \varrho \partial \varrho^T} &= \sum_{i=1}^n \sum_{k_1}^{m_1} \cdots \sum_{k_c}^{m_c} w_{ik} \sum_{t=1}^{T_j} S_c^T (d_k^T \otimes z_{it}^T) D_{itk} \Sigma_{itk}^{-1} D_{itk}^T (d_k^T \otimes z_{it}^T)^T S_c. \end{aligned} \tag{5.62}$$

The EM-Fisher-Scoring-Algorithm in this case is characterized by the form

1. calculate startvalue $\delta^{(0)}$.

2. for $p = 1, 2, \ldots$ approximate $M(\delta|\delta^{(p)})$ by $\tilde{M}(\delta|\delta^{(p)})$ and compute weights $w_{ik}(\delta^{(p)})$ with respect of $\delta^{(p)}$

3. for $s = 1, 2, \ldots$ run the Fisher-Scoring-Algorithm till
$$\frac{||\delta^{(s+1)} - \delta^{(s)}||}{||\delta^{(s)}||}$$

4. if the condition
$$\frac{||\delta^{(s+1)} - \delta^{(p)}||}{||\delta^{(p)}||}$$
is accomplished, convergence of the EM-Algorithm is achieved. If not start in step 2 with $\delta^{(p+1)} = \delta^{(s+1)}$ as update for $\delta^{(p)}$.

Im & Gianola (1988) used the EM-Algorithm with Gaussian-Quadrature in binary data, as well as Bock & Aitkin (1981) and Anderson & Aitkin (1985). Stiratelli, Laird & Ware (1984) and Steele (1996) use the Laplace-Approximation to approximate the conditional expectation. Stiratelli, Laird & Ware (1984) use the first-order Laplace-Approximation, Steele (1996) uses a modified second-order Laplace-Approximation with the Newton-Raphson Algorithm. Meng & Rubin (1993) introduced the ECM Algoritm (Expectation/Conditional Maximation) which is a generalization of the EM-Algorithm. This algorithm takes advantage of the simplicity of complete data conditional maximum likelhood estimation by replacing a complicated M-Step of the EM with several computationally simpler CM-Steps (Conditional Maximization Steps). The ECME-Algorithm (Expectation/Conditional Maximization Either) based on Liu & Rubin (1994) replaces some CM-Steps of the ECM, which maximize the constrained expected complete-data likelihood function, with steps that maximize the correspondingly constrained actual likelihood function. Other variants are described in Rai & Matthews (1993), McLachlan & Krishnan (1997) and Meng & van Dyk (1997).

Not mentioned in this context is the full Bayesian approach as mentioned in Zeger & Karim (1991), which is based on MCMC. Further Waclawiw & Liang (1993) use modified GEE to estimate GLMM's. In the recent years specifying a random effects density with no further restrictions became popular. One approach is the nonparametric maximum likelihood for finite mixtures as described in Aitkin & Francis (1998) and Aitkin (1999). Another way of modeling smooth random effects is given by Davidian & Gallant (1993), Chen, Zhang & Davidian (2002) and Gallant & Nychka (1987), who use a modified Monte Carlo EM algorithm for estimation. Similarly is the approach of Chiou, Müller & Wang (2003). Ghidey, Lesaffres & Eilers (2004) use penalized Gaussian Mixtures in a similar way as P-spline smoothing for the estimation of a linear mixed model with smooth random effects distribution.

Chapter 6

Generalized Semi-Structured Mixed Models

6.1 The Model

It is simpler to derive the generalized semi-structured mixed model in the notation of general model, since the representation of clusters together with basis function expansions is not easy.

Suppose that the data are composed of N observations, with response $y_{(i)}$, covariate vectors $x_{(i)}$ associated with fixed effects, covariate vectors $u_{(i)}$ associated with non-parametric effects covariate vectors $z_{(i)}$ associated with random effects. Let $u_{(i)}^T = (u_{(i)1}, \ldots, u_{(i)m})^T$ consists of m different covariates. It is assumed that the observations $y_{(i)}$ are conditionally independent with means $\mu_{(i)} = E(y_{(i)}|b)$ and variances $var(y_{(i)}|b) = \phi v(\mu_{(i)})$, where $v(.)$ is a known variance function and ϕ is a scale parameter. The generalized semiparametric mixed model that is considered in the following has the form

$$g(\mu_{(i)}) = x_{(i)}^T\beta + \sum_{j=1}^{m} \alpha_{(j)}(u_{(i)j}) + z_{(i)}^T b \tag{6.1}$$

$$= \eta_{(i)}^{par} + \eta_{(i)}^{add} + \eta_{(i)}^{rand}, \tag{6.2}$$

where $g(.)$ is a monotonic differentiable link function,

$\eta_{(i)}^{par} = x_{(i)}^T\beta$ is a linear parametric term,

$\eta_{(i)}^{add} = \sum_{j=1}^{m} \alpha_{(j)}(u_{(i)j})$ is an additive term with unspecified influence functions $\alpha_{(1)}, \ldots, \alpha_{(m)}$,

$\eta_{(i)}^{rand} = z_{(i)}^T b$ contains the cluster-specific random effect $b \sim N(0, \mathbb{Q}(\varrho))$, where $\mathbb{Q}(\varrho)$ is a parameterized covariance matrix.

An alternative form that is used in the following is $\mu_{(i)} = h(\eta_{(i)}), \eta_{(i)} = \eta_{(i)}^{par} + \eta_{(i)}^{add} + \eta_{(i)}^{rand}$, where $h(.) = g^{-1}(.)$ is the inverse link function. If the functions $\alpha_{(j)}(.)$ are linear , the model reduces to the generalized mixed model of Breslow & Clayton (1993). Versions of the additive model (6.1) have been considered by Zeger & Diggle (1994) and Lin & Zhang (1999), Zhang, Lin, Raz & Sowers (1998).

While Lin & Zhang (1999) used natural cubic smoothing splines for the estimation of the unknown functions $\alpha_{(j)}$, in the following regression splines are used. In recent years regression splines have been used widely for the estimation of additive structures, see Marx & Eilers (1998), Wood (2004) and Wand (2000).

In regression spline methodology the unknown functions $\alpha_{(j)}(.)$ are approximated by basis functions. A simple basis is known as the truncated power series basis of degree d, yielding

$$\alpha_{(j)}(u_{(i)j}) = \gamma_0^{(j)} + \gamma_1^{(j)} u_{(i)j} + \ldots \gamma_d^{(j)} u_{(i)j}^d + \sum_{s=1}^{M} \alpha_s^{(j)} (u_{(i)j} - k_s^{(j)})_+^d \ ,$$

where $k_1^{(j)} < \ldots < k_M^{(j)}$ are distinct knots. More generally one uses

$$\alpha_{(j)}(u_{(i)j}) = \sum_{s=1}^{M} \alpha_s^{(j)} \phi_s^j(u_{(i)j}) = \alpha_j^T \phi_{(i)j}, \tag{6.3}$$

where $\phi_s^{(j)}$ denotes the s-th basis function for variable j, $\alpha_j^T = (\alpha_1^{(j)}, \ldots, \alpha_M^{(j)})$ are unknown parameters and $\phi_{(i)j}^T = (\phi_1^{(j)}(u_{(i)j}), \ldots, \phi_M^{(j)}(u_{(i)j}))$ represents the vector-valued evaluations of the basis functions.
The parameterized model for (6.1) is given in the form

$$g(\mu_{(i)}) = x_{(i)}^T \beta + \phi_{(i)1}^T \alpha_1 + \cdots + \phi_{(i)m}^T \alpha_m + z_{(i)}^T b$$

or the matrix form

$$g(\mu) = X\beta + \Phi_{.1}\alpha_1 + \cdots + \Phi_{.m}\alpha_m + Zb$$

where the matrices X and Z have rows $x_{(i)}^T$ and $z_{(i)}^T$, and $\Phi_{.j}$ has rows $\phi_{(i)j}^T$, which again can be reduced to

$$g(\mu) = X\beta + \Phi\alpha + Zb$$

with $\alpha^T = (\alpha_1^T, \ldots, \alpha_m^T))$ and $\Phi = (\Phi_{.1}, \ldots, \Phi_{.m})$ where Φ has rows $\phi_{(i)}^T = (\phi_{(i)1}^T, \ldots, \phi_{(i)m}^T)$.

6.1.1 The Penalized Likelihood Approach

Focusing on generalized semiparametric mixed models we assume that the conditional density of $y_{(i)}$, given the explanatory variable $x_{(i)}$ and the random effect b is of exponential family type

$$f(y_{(i)} \mid x_{(i)}, b) = \{\exp \frac{(y_{(i)}^T \gamma_{(i)} - \kappa(\gamma_{(i)}))}{\phi} + c(y_{(i)}, \phi)\}, \tag{6.4}$$

where γ_i denotes the natural parameter, $c(.)$ the log normalization constant and ϕ the dispersion parameter.

The most popular method to maximize generalized linear mixed models is penalized quasi likelihood (PQL), which has been suggested by Breslow & Clayton (1993), Breslow & Lin (1995b) and Breslow & Lin (1995a). It is implemented in the macro GLIMMIX and proc GLIMMIX in SAS (Wolfinger (1994)) or the gamm-function in the R-package mgcv. Further notes are in Wolfinger & O'Connell (1993), Littell, Milliken, Stroup & Wolfinger (1996) and Vonesh (1996).

In penalized based concepts the joint likelihood-function is specified by the parameters of the covariance structure ϱ together with the dispersion parameter ϕ which are collected in $\theta^T = (\phi, \varrho^T)$ and parameter vector $\delta^T = (\beta^T, \alpha^T, b^T)$. The corresponding log-likelihood is

$$l(\delta, \theta) = \sum \log(\int f(y_{(i)}|\delta) * p(b, \varrho) db). \tag{6.5}$$

where $p(b, \varrho)$ denotes the density of the random effects.

For the case of few basis functions and therefore low-dimensional parameter vector α, the log-likelihood may be approximated as proposed by Breslow & Clayton (1993). However, the form of the unknown functions $\alpha(.)$ is severely restricted. A more flexible approach which is advocated here is to use many basis functions, say about 20 for each function $\alpha_{(j)}$, and add a penalty term to the log-likelihood. Then one obtains the penalized log-likelihood

$$l_p(\delta, \theta) = \sum_{i=1}^{N} \log(\int f(y_{(i)}|\delta) * p(b; \varrho) db) - \frac{1}{2} \sum_{j=1}^{m} \lambda_j \alpha_j^T K_j \alpha_j. \tag{6.6}$$

where K_j penalizes the parameters α_j. When using P splines one penalizes the difference between adjacent categories in the form $\lambda_j \alpha^T K_j \alpha = \lambda_j \sum_j \{\triangle^d \alpha_j\}^2$ where $\triangle$ is the

difference operator with $\triangle\alpha_j = \alpha_{j+1} - \alpha_j$, $\triangle^2\alpha_j = \triangle(\triangle\alpha_j)$ etc., for details see Eilers & Marx (1996). The log-likelihood (6.6) has also been considered by Lin & Zhang (1999) but with K_j referring to smoothing splines. For smoothing splines the dimension of α_j increases with sample size whereas the low rank smoother used here does not depend on n.

Approximation of (6.6) along the lines of Breslow & Clayton (1993) yields the double penalized likelihood

$$l_p(\delta, \theta) = \sum_{i=1}^{N} \log(f(y_{(i)}|\delta)) - \frac{1}{2} b^T \mathbb{Q}(\varrho)^{-1} b - \frac{1}{2} \sum_{j=1}^{m} \lambda \alpha_j^T K_j \alpha_j. \tag{6.7}$$

The first penalty term $b^T\mathbb{Q}(\varrho)^{-1}b$ is due to the approximation based on the Laplace method, the second penalty term $\sum_{j=1}^{m} \lambda_j \alpha_j^T K_j \alpha_j$ determines the smoothness of the functions $\alpha_{(j)}(.)$ depending on the chosen smoothing parameter λ_j.

PQL usually works within the profile likelihood concept. So we can distinguish between the estimation of δ given the plugged in estimation $\hat{\theta}$ resulting in the profile-likelihood $l_p(\delta, \hat{\theta})$ and the estimation of θ given the plugged in estimator $\hat{\delta}$ resulting in the profile-likelihood $l_p(\hat{\delta}, \theta)$.

Estimation of β, α and b for fixed θ: First we consider the maximation of $l_p(\delta, \theta)$ with respect to $\delta = (\beta^T, \alpha^T, b^T)$. As described in Breslow & Clayton (1993) the solution of the score function $s(\delta) = \frac{\partial l_p(\delta,\theta)}{\partial \delta} = 0$ for (6.7) via Fisher-Scoring is equivalent to iteratively solving the BLUP-equations with a linearized version. For derivations to follow the motivation $\Sigma(\delta, \theta)_i = \text{cov}(y_{(i)}|\delta, \theta)$ and $D_{(i)}(\delta) = \frac{h(\eta_{(i)})}{\partial \eta_{(i)}}$ are necessary. The matrix versions are $D(\delta) = \text{diag}(D_{(i)}(\delta))_{i=1,\ldots,N}$ and $\Sigma(\delta, \theta) = \text{diag}(\Sigma_{(i)}(\delta, \theta))_{i=1,\ldots,N}$. The linearized version is given by

$$\tilde{y}_{(i)} = x_{(i)}^T \beta + \phi_{(i)}^T \alpha + z_{(i)}^T b + D_{(i)}^{-1}(\delta)(y_{(i)} - \mu_{(i)}).$$

In matrix notation one obtains

$$\tilde{y} = X\beta + \Phi\alpha + Zb + D^{-1}(\delta)(y - \mu).$$

For the linearized version the approximated covariance is given by

$$W = W(\delta) = D(\delta)\Sigma^{-1}(\delta)D^T(\delta).$$

The estimation problem using weighted least squares is equivalent to the estimation problem of the mixed model

$$\tilde{y}|b \overset{approx}{\sim} N(X\beta + \Phi\alpha + Zb, W^{-1}). \tag{6.8}$$

Estimation of θ for fixed β, α and b: If we assume (6.8) and if b is normally distributed the random effect can integrated out analytically. The theory of linear mixed models within the REML framework can be applied to estimate the variance parameters. So a $V(\delta, \theta)$ can be constructed with

$$V(\theta) := V(\delta, \theta) = W^{-1} + Z\mathbb{Q}(\varrho)Z^T$$

The corresponding REML-equation has the form

$$\begin{aligned} l_p(\delta, \theta) &\approx \int \tilde{f}(\tilde{y}|b) * p(b; \varrho) \\ &\approx -\tfrac{1}{2}\log(|V(\theta)|) + (\tilde{y} - X\beta - \Phi\alpha)^T V(\theta)^{-1}(\tilde{y} - X\beta - \Phi\alpha) - \tfrac{1}{2}\log(|X^T V(\theta) X|) \end{aligned} \tag{6.9}$$

where $\tilde{f}(.|b)$ and $p(.)$ are Gaussian densities for $\tilde{y}$ and b as described in (6.8).

6.2 Boosted Generalized Additive Mixed Models - bGAMM

The following algorithm uses componentwise boosting. Componentwise boosting means that only one component of the predictor, in our case one smooth term $\phi_{(i)j}\alpha_j$, is refitted at a time. That means that a model containing the linear term and only one smooth component is fitted in one iteration step. For simplicity we will use the notation

$$x^T_{(i)(r)} := [x^T_{(i)},\ \phi^T_{(i)r},\ z^T_{(i)}] \quad , \quad \delta^T_r = (\beta^T, \alpha^T_r, b^T)$$

for the design matrix. For the predictor without random part we denote $\tilde{\eta}^T_{(i)r} = x^T_{(i)}\beta + \phi^T_{(i)r}\alpha_r$.

bGAMM

1. *Initialization*
   Compute starting values $\hat{\beta}^{(0)}, \hat{\alpha}^{(0)}_1, \ldots \hat{\alpha}^{(0)}_m$, $b^{(0)}$ and set $\tilde{\eta}^{(0)}_{(i)} = x^T_i\hat{\beta}^{(0)} + \phi^T_{(i)1}\hat{\alpha}^{(0)}_1 + \ldots + \phi^T_{(i)m}\hat{\alpha}^{(0)}_m$.

2. *Iteration*
   For l=1,2,...

   (a) *Refitting of residuals*

   i. Computation of parameters
   For $r \in \{1, \ldots, m\}$ fit the model

$$g(\mu_{(i)r}) = \tilde{\eta}^{(l-1)}_{(i)} + x_{(i)}\beta + \phi_{(i)r}\alpha_r + z^T_{(i)}b$$

yielding $\delta_r^T = (\beta^T, \alpha_r^T, b^T)$ where $\tilde{\eta}_{(i)}^{(l-1)}$ is treated as an offset using $\tilde{y}_{(i)} = \eta_{(i)}^{(l)} + z_{(i)}^T b + D_i^{-1}(\delta)(y_i - \tilde{\eta}_i^{(l)} - z_{(i)}^T b)$ with only one iteration.

ii. Selection step
Select from $r \in \{1, \ldots, m\}$ the component j that leads to the smallest $BIC_r^{(l)}$.

iii. Update
Set $\hat{\beta}^{(l)} = \hat{\beta}^{(l-1)} + \hat{\beta}$,
and

$$\hat{\alpha}_r^{(l)} = \begin{cases} \hat{\alpha}_r^{(l-1)} & \text{if } r \neq j \\ \hat{\alpha}_r^{(l-1)} + \hat{\alpha}_r & \text{if } r = j, \end{cases}$$

$$\hat{\delta}^{(l)} = ((\hat{\beta}^{(l)})^T,\ (\hat{\alpha}_1^{(l)})^T,\ \ldots (\hat{\alpha}_m^{(l)})^T, (\hat{b}^{(l)})^T).$$

Update for $i = 1, \ldots, N$

$$\tilde{\eta}_{(i)}^{(l)} = \tilde{\eta}_{(i)}^{(l-1)} + x_{(i)}^T \beta + \phi_{(i)j}^T \alpha_j.$$

(b) *Computation of Variance Components*
The computation is based on the penalized quasi likelhood and its score and fisher functions 6.9

$$\begin{aligned} l_p(\theta|\eta^{(l)}; \delta_l) &= -\tfrac{1}{2}\log(|V(\theta)|) - \tfrac{1}{2}(\tilde{y} - \tilde{\eta}^{(l)})^T V(\theta)^{-1}(\tilde{y} - \tilde{\eta}^{(l)}) \\ &\quad -\tfrac{1}{2}(\hat{\delta}^{(l)})^T K \hat{\delta}^{(l)}. \end{aligned}$$

The corresponding penalty matrix is denoted by K_r, which for the truncated power series has the form

$$K_r = Diag(0,\ \lambda I,\ 0).$$

Maximization yields $\hat{\theta}^{(l)}$.

6.2.1 Stopping Criteria

With starting value $\hat{\delta}^{(0)}$ and $\bar{W}^{(0)} = W(\hat{\delta}^{(0)}, \theta^{(0)})$, $\Sigma^{(0)} = \Sigma(\hat{\delta}^{(0)}, \theta^{(0)})$, $D^{(0)} = D(\hat{\delta}^{(0)}, \theta^{(0)})$ denoting evaluations at value $\hat{\eta}^{(0)} + X\hat{\delta}^{(0)}$ one step Fisher Scoring is given by

$$\begin{aligned} \hat{\delta}^{(1)} &= F(\hat{\delta})^{-1} s(\hat{\delta}^{(0)}) \\ &= (X^T \bar{W}^{(0)} X + K)^{-1} X \bar{W}^{(0)} {D^{(0)}}^{-1} (y - \hat{\mu}^{(0)}). \end{aligned}$$

Setting $\hat{\mu}^{(l)} = h(\hat{\eta}^{(l)} + X_j\delta^{(l)})$ one obtains

$$\begin{aligned} \hat{\eta}^{(l+1)} + Zb^{(l)} &= X_j\hat{\delta}_j + \hat{\eta}^{(l)} \\ \hat{\eta}^{(l+1)} + Zb^{(l)} - \hat{\eta}^{(l)} &= X_j\hat{\delta}_j \\ &= X_j(X_j^T W^{(l)} X_j + K)^{-1} X_j^T W^{(l)} {D^{(l)}}^{-1}(y - \hat{\mu}^{(l)}). \end{aligned}$$

Taylor approximation of first order $h(\hat{\eta} + Zb) = h(\eta) + \frac{\partial h(\eta)}{\partial \eta^T}(\hat{\eta} + Zb - \eta)$ yields

$$\begin{aligned} \hat{\mu}^{(l+1)} &\approx \hat{\tilde{\mu}}^{(l)} + \tilde{D}_l(\hat{\eta}_{(l+1)} + Zb^{(l)} - \hat{\eta}_{(l)}) \\ \hat{\eta}^{(l+1)} + Zb^{(l)} - \hat{\eta}^{(l)} &\approx \tilde{D}^{(l)})^{-1}(\hat{\mu}^{(l+1)} - \hat{\mu}^{(l)}) \end{aligned}$$

and therefore

$$(\tilde{W}^{(l)})^{1/2}(\tilde{D}^{(l)})^{-1}(\hat{\mu}^{(l+1)} - \hat{\mu}^{(l)}) \approx (\tilde{W}^{(l)})^{1/2} X_j (X_j^T W^{(l)} X_j + K)^{-1} X_j^T W^{(l)} {D^{(l)}}^{-1}(y - \hat{\mu}^{(l)}).$$

Since $(W^{(l)})^{1/2}(D^{(l)})^{-1} = (\Sigma^{(l)}))^{1/2}$ and $(\tilde{W}^{(l)})^{1/2}(\tilde{D}^{(l)})^{-1} = \tilde{(}\Sigma^{(l)})^{1/2}$ this can be transformed to

$$\hat{\mu}^{(l+1)} - \hat{\tilde{\mu}}^{(l)} \approx M^{(l+1)}(y - \hat{\mu}^{(l)})$$

with $M^{(l+1)} = (\tilde{\Sigma}^{(l)})^{1/2}(\tilde{W}^{(l)})^{1/2} X_j (X_j^T \bar{W}^{(l)} X_j + K)^{-1} X_j^T (W^{(l)})^{1/2} {\Sigma^{(l)}}^{1/2}$.

Define $\hat{\mu}^{(l)} = \hat{\tilde{\mu}}^{(l)} + C^{(l)}$. For simplicity one can use

$$\hat{\mu}^{(l+1)} - \hat{\mu}^{(l)} \approx M^{(l+1)}(y - \hat{\mu}^{(l)}) + C^{(l)}.$$

So one obtains

$$\begin{aligned} \hat{\mu}^{(l+1)} - \hat{\mu}^{(l)} &\approx M^{(l+1)}(y - \hat{\mu}^{(l)} + C^{(l)} \\ &= M^{(l+1)}(y - \hat{\mu}^{(l-1)} - (\hat{\mu}^{(l)} - \hat{\mu}^{(l-1)})) + C^{(l)} - M^{(l+1)} C^{(l-1)} \\ &\approx M^{(l+1)}(y - \hat{\mu}^{(l-1)} - M^{(l)}(y - \hat{\mu}^{(l-1)})) + C^{(l)} - M^{(l+1)} C^{(l-1)} \\ &= M^{(l+1)}(I - M^{(l)})(y - \hat{\mu}^{(l-1)}) + C^{(l)} - M^{(l+1)} C^{(l-1)}. \end{aligned}$$

So one gets

$$\hat{\mu}^{(m)} \approx \sum_{j=0}^{m} M^{(j)} \prod_{i=0}^{j-1} (I - M^{(i)})\ y + R^{(m)}$$

with $R^{(m)} = \sum_{j=1}^{m} S^{(j)}$. $S^{(j)}$ is defined by

$$S^{(j)} = C^{(j-1)} - \sum_{k=1}^{j} M^{(k)} \prod_{i=1}^{k-1} (I - M^{(k-i)}) C^{(k-i-1)}.$$

For interpretation the version

$$\hat{\mu}^{(m)} - R^{(m)} \quad \approx \quad \sum_{j=0}^{m} M^{(j)} \prod_{i=0}^{j-1} (I - M^{(i)}) \; y$$

should be used where $\hat{\mu}^{(m)} - R^{(m)}$ is the result of the projection of y. $R^{(m)}$ is the correction term associated with the random effects. So one can write

$$\hat{\mu}^{(m)} - R^{(m)} \quad \approx \quad H^{(m)} y$$

The corresponding projection matrix is given by

$$H^{(m)} = \sum_{j=0}^{m} M^{(j)} \prod_{i=0}^{j-1} (I - M^{(i)}). \tag{6.10}$$

6.2.2 Simulation Study

Poisson Link We present part of a simulation study in which the performance of Boost-Mixed models is compared to alternative approaches. The underlying model is the random intercept model

$$\eta_{it} = b_i + \textstyle\sum_{j=1}^{20} c * \alpha_{(j)}(u_{it}), \quad i = 1, \ldots, 40, \quad t = 1, \ldots, 5$$

$$E(y_{it}) = \exp(\eta_{it})$$

with the smooth components given by

$$\alpha_{(1)}(u) = sin(u) \quad u \in [-3, 3],$$

$$\alpha_{(2)}(u) = cos(u) \quad u \in [-2, 8],$$

$$\alpha_{(3)}(u) = u^2 \qquad u \in [-1, 1],$$

$$\alpha_{(4)}(u) = u^3 \qquad u \in [-1, 1],$$

$$\alpha_{(5)}(u) = -u^2 \qquad u \in [-1, 1],$$

$$\alpha_{(j)}(u) = 0 \qquad u \in [-3, 3], j = 6, \ldots, 20.$$

The vectors $u_{it}^T = (u_{it1}, \ldots, u_{it20})$ have been drawn independently with components following a uniform distribution within the specified interval. For the covariates constant correlation is assumed, i.e. $corr(u_{itr}, u_{its}) = 0.1$. The constant c determines the signal strength of the covariates. The random effect and the noise variable have been specified by $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 0.6$. In the part of the study which is presented the number of observations has been chosen by $n = 40, T = 5$.

The fit of the model is based on B-splines of degree 3 with 15 equidistant knots. The performance of estimators is evaluated separately for the structural components and the variance. By averaging across 100 datasets we consider mean squared errors for η, σ_b^2 given by

$$\text{mse}_\eta \quad = \sum_{i=1}^n \sum_{t=1}^T (\eta_{it} - \hat{\eta}_{it})^2, \hat{\eta}_{it} = x_{it}^T \hat{\beta},$$

$$\text{mse}_{\sigma_b^2} \quad = ||\sigma_b^2 - \hat{\sigma}_b^2||^2$$

Additional information on the stability of the algorithms is collected in $notconv$, which indicates the sum over the datasets, where numerical problems occurred during estimation. $falseneg$ is the mean over the count of variables $\alpha_{(i)}(u), i = 1, \ldots, 5$, that were not selected. $falsepos$ is the mean over the count of variables $\alpha_{(i)}(u), i = 6, \ldots, 20$, that were selected.

In Table 6.1 the resulting mean squared errors are given for increasing signals and increasing number of parameters. Since for a large number of covariates the Generalized Additive Mixed Model strategy (GAMM) did not converge for many cases, i.e. for $c = 0.7$ and $p = 15$ only 18 of 100 datasets lead to feasible results using GAMM. Only the cases

that lead to convergence were compared with the boosted Generalized Additive Mixed Model (bGAMM) on the one side and the cases that lead to convergence using bGAMM were compared to GAMM on the other side. That means only datasets which lead on both sides to convergence were chosen to be compared. It becomes obvious that for many parameters (p>=10) GAMM is not a suitable method to handle many unspecified parameters. FalsePositive (FalsePos) are the unspecified variables that were selected by the algorithm but have no real effect on the response. Instead FalseNegative (FalseNeg) are those variables that should have been selected by the algorithm but were not selected. For Table 6.1 the BIC-Criterion was chosen to be the stopping and selection criterion.

		GAMM			bGAMM					
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg
0.5	5	15.417	0.020	4	15.403	0.015	0	72.2	0.0	0.4
0.5	10	18.503	0.007	71	18.271	0.020	0	63.6	1.2	0.5
0.5	15	22.694	0.009	88	21.772	0.006	0	71.3	1.6	1.0
0.5	20				22.116	0.017	0	63.1	2.2	0.7
0.7	5	14.537	0.027	1	13.415	0.018	0	87.4	0.0	0.0
0.7	10	16.702	0.016	72	15.427	0.026	0	126.3	1.2	0.0
0.7	15	22.466	0.009	92	17.799	0.012	0	66.8	1.7	0.1
0.7	20				20.496	0.016	0	99.6	2.4	0.1
1.0	5	15.746	0.025	0	14.123	0.015	0	104.5	0.0	0.0
1.0	10	18.121	0.006	68	16.399	0.009	0	104.4	1.3	0.0
1.0	15	19.626	0.001	95	13.758	0.017	0	118.0	2.0	0.0
1.0	20				22.138	0.012	0	108.7	2.9	0.0

Table 6.1: Generalized additive mixed model and boosted generalized additive mixed model on poisson data

For a more extensive analysis of BoostMixed six simulation studies with different settings were made. In all studies 100 datasets were generated. AIC-Criterion and BIC-Criterion were compared.

Study 15 - small clusters and small random effect
The random effect and the noise variable have been specified by $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 0.6$. In the part of the study which is presented the number of observations has been chosen by $n = 100, T = 2$. Pairwise correlation was taken to be $corr(u_{itr}, u_{its}) = 0.1$. Details can be found in Table C.14 and Table C.15.

Study 16 - few clusters and large random effect
The random effect and the noise variable have been specified by $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 1.2$. In the part of the study which is presented the number of observations has been

chosen by $n = 40, T = 5$. Pairwise correlation was taken to be $corr(u_{itr}, u_{its}) = 0.1$. Details can be found in Table C.16 and Table C.17.

Study 17 - big clusters, few clusters
The random effect and the noise variable have been specified by $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 0.6$. In the part of the study which is presented the number of observations has been chosen by $n = 20, T = 10$. Pairwise correlation was taken to be $corr(u_{itr}, u_{its}) = 0.1$. Details can be found in Table C.18 and Table C.19.

Study 18 - many clusters and small random effect
The random effect and the noise variable have been specified by $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 0.6$. In the part of the study which is presented the number of observations has been chosen by $n = 80, T = 5$. Pairwise correlation was taken to be $corr(u_{itr}, u_{its}) = 0.1$. Details can be found in Table C.20 and Table C.21.

Study 19 - many clusters and huge random effect
The random effect and the noise variable have been specified by $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 1.2$. In the part of the study which is presented the number of observations has been chosen by $n = 80, T = 5$. Pairwise correlation was taken to be $corr(u_{itr}, u_{its}) = 0.1$.Details can be found in Table C.22 and Table C.23.

Study 20 - big clusters, many clusters, correlated data
The random effect and the noise variable have been specified by $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 0.6$. In the part of the study which is presented the number of observations has been chosen by $n = 40, T = 10$. Pairwise correlation was taken to be $corr(u_{itr}, u_{its}) = 0.5$. Details can be found in Table C.24 and Table C.25.

If one wants to summarize the results of study 15 to study 20 the boosted GAMM (bGAMM) seems to be a good competitor to the generalized additive mixed model (GAMM) for cases with more than 400 observations in total, see study 17 to study 20. Only the cases that lead to convergence were compared with the boosted Generalized Additive Mixed Model (bGAMM) on the one side and the cases that lead to convergence using bGAMM were compared to GAMM on the other side. Nevertheless it is remarked that for small dataset with small clusters (200 observations in total, study 15 and 16) that numerical problem affects the GAMM method. In study 15 the GAMM method did not converge in 17 of 100 cases for strength $c = 0.5$ and five variables. For more than 15 variables GAMM did not lead to convergence in at least 84 of 100 datasets for strength $c = 0.5$, $c = 0.7$ and $c = 1$ for AIC. These problems also arise in all studies. In almost all studies the BIC criterion delivered better MSE_η than the AIC criterion in cases with many irrelevant variables ($p \geq 10$). In cases with just relevant variables AIC was in most cases superior to the BIC criterion. Responsible for the MSE_η in studies 15 and 16

may be the selection of relevant variables. In these studies not all relevant variables were selected. In study 16 averaged 0.71 relevant variables of 5 possible were not selected in the case of AIC ($c = 0.5$ and $p = 5$) , 1.37 in the case of BIC. In study 17 to 20 nearly all relevant variables were selected using AIC or BIC but with more irrelevant variable in the case of AIC. The problem of AIC is that it allows to select more irrelevant variables which is reflected in a remarkable downgrade in terms of MSE_η. In most of the studies bGAMM has better MSE_b than GAMM.

Binomial Link We present part of a simulation study in which the performance of BoostMixed models is compared to alternative approaches. The underlying model is the random intercept model

$$\eta_{it} = b_i + \sum_{j=1}^{20} c * \beta_{(j)} * u_{itj}, i = 1, \ldots, 80, t = 1, \ldots, 5,$$

$$E(y_{it}) = h(\eta_{it})$$

with the smooth components given by $\beta_{(1)} = 2.0, \beta_{(2)} = 2.5, \beta_{(3)} = 3.0, \beta_{(4)}(u) = 3.5, \beta_{(5)}(u) = 4.0, \beta_{(j)} = 0 \quad j = 6, \ldots, 20.$, where $h(.)$ is the logistic function.

The vectors $u_{it}^T = (u_{it1}, \ldots, u_{it20})$ have been drawn independently with components following a uniform distribution within the specified interval. For the covariates constant correlation is assumed, i.e. $corr(u_{itr}, u_{its}) = \rho$. The constant c determines the signal strength of the covariates. The random effect and the noise variable have been specified by $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 0.6$. In the part of the study which is presented the number of observations has been chosen by $n = 60, T = 5$. For Table 6.2 the AIC-Criterion was used.

In Table 6.2 the resulting mean squared errors are given for increasing signals and increasing number of parameters. In this case an implicit variable selection procedure makes sense since for increasing number of parameters the Generalized Mixed Model strategy (GLMM) deliver very instable estimates or a dramatic loss in the accuracy of the predictions. FalsePositive (FalsePos) are the unspecified variables that were selected by the algorithm but have no real effect on the response. Instead FalseNegative (FalseNeg) are those variables that should have been selected by the algorithm but were not selected. Nevertheless there a some datasets where the boosted Mixed Model (bGLMM) did not find all relevant variables. On the other side the boosted Mixed Model method helps to reduce the irrelevant variables. In the case for signal c=1.0 and 15 parameters only averaged 1.49 from 10 possible irrelevant variables were selected which have no effect on the response. Remarkable is that for small signals in this study the mean squared errors for the random effects variance are quite smaller.

		GLMM			bGLMM					
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg
0.5	5	72.003	0.031	0	87.475	0.056	0	100.5	0.00	0.02
0.5	10	146.845	0.038	0	102.632	0.066	0	102.6	0.23	0.02
0.5	15	210.027	0.058	0	119.176	0.059	0	98.2	0.36	0.02
0.5	20	283.818	0.071	0	123.898	0.077	0	106.2	0.55	0.02
0.7	5	141.793	0.142	0	141.322	0.123	0	123.5	0.00	0.01
0.7	10	279.357	0.161	0	170.764	0.152	0	112.5	0.25	0.01
0.7	15	416.436	0.165	0	220.872	0.161	0	106.4	0.57	0.01
0.7	20	696.907	0.187	0	244.113	0.161	0	120.7	0.83	0.01
1.0	5	673.332	0.256	0	532.380	0.336	1	128.6	0.00	0.02
1.0	10	1906.076	0.251	0	535.680	0.353	0	114.1	0.64	0.02
1.0	15	3563.036	0.277	0	636.291	0.504	0	105.7	1.49	0.02
1.0	20	4198.591	0.301	0	698.534	0.509	0	139.6	2.88	0.02

Table 6.2: Generalized mixed model and boosted generalized mixed model on binomial data

For a more extensive analysis of BoostMixed six simulation studies with different settings were made. In all studies 100 datasets were generated. AIC-Criterion and BIC-Criterion were compared.

Study 21 - small dataset and small random effect
The random effect and the noise variable have been specified by $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 0.6$. In the part of the study which is presented the number of observations has been chosen by $n = 159, T = 2$. Pairwise correlation was taken to be $corr(u_{itr}, u_{its}) = 0.1$. Details can be found in Table C.26 and Table C.27.

Study 22 - small dataset and large random effect
The random effect and the noise variable have been specified by $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 1.2$. In the part of the study which is presented the number of observations has been chosen by $n = 60, T = 5$. Pairwise correlation was taken to be $corr(u_{itr}, u_{its}) = 0.1$. Details can be found in Table C.28 and Table C.29.

Study 23 - big clusters, small dataset
The random effect and the noise variable have been specified by $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 0.6$. In the part of the study which is presented the number of observations has been chosen by $n = 30, T = 10$. Pairwise correlation was taken to be $corr(u_{itr}, u_{its}) = 0.1$. Details can be found in Table C.30 and Table C.31.

Study 24 - many clusters and small random effect
The random effect and the noise variable have been specified by $b_i \sim N(0, \sigma_b^2)$ with

$\sigma_b^2 = 0.6$. In the part of the study which is presented the number of observations has been chosen by $n = 80, T = 5$. Pairwise correlation was taken to be $corr(u_{itr}, u_{its}) = 0.1$. Details can be found in Table C.32 and Table C.33.

Study 25 - many clusters and big dataset
The random effect and the noise variable have been specified by $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 0.6$. In the part of the study which is presented the number of observations has been chosen by $n = 100, T = 5$. Pairwise correlation was taken to be $corr(u_{itr}, u_{its}) = 0.1$.Details can be found in Table C.34 and Table (C.35.

Study 26 - big clusters and big dataset
The random effect and the noise variable have been specified by $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 0.6$. In the part of the study which is presented the number of observations has been chosen by $n = 50, T = 10$. Pairwise correlation was taken to be $corr(u_{itr}, u_{its}) = 0.1$. Details can be found in Table C.36 and Table C.37.

The results can be summarized as follows. In all studies except study 26 the boosted generalized linear mixed model (bGLMM) was superior in the MSE_η for signals $c = 1$. For signal $c = 0.5$ and 5 relevant variables the generalized linear mixed model could not be further improved by the boosted variant. In only two cases (Study 21 with AIC and Study 23 with BIC) the MSE_η could be improved for signal $c = 0.7$ and 5 variables in the model. Except study 21 the right amount of relevant variables were found by the boosted version of the generalized linear mixed model. For models based on just relevant variables the AIC criterion seems to perform best. In models with many irrelevant variables the BIC seems to deliver better results in the MSE_η. However in models with large signals the accuracy of the adjustment is decreasing using the generalized linear mixed model. Quite impressing is the influence of irrelevant variables on the MSE_η which is reflected in study 22 (for $c = 0.7$ and $p = 10$) which has double the value of the model without irrelevant variables ($c = 0.7$ and $p = 5$). In the context of binary data the boosted generalized linear mixed model may be a suitable tool to do variable selection in datasets with many covariates.

6.3 Application of the European Patent Data

For a detailed description of the dataset see Chapter 5.1. Descriptive statistics for the response (OUT) are given in the Table 6.3 and for the covariates in Table 6.4. The estimates can be found in Table 6.5 and the smooth estimates in Figure 6.1.

The variables BREADTH, PA_EMP, EMP and R_D_PAT were not selected by the

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	1.000	5.000	9.701	12.000	169.000

Table 6.3: Summary statistics for the response considering small campanies

Covariate	Mean	Minimum	Maximum
YEAR		1993	2000
PA_YEA	20.21	1.00	202.00
BREADTH	0.58	0.12	0.90
PAT_PORT	144.47	0.00	1836.00
VOLATILITY	0.44	0.00	2.00
EMP (000s)	6.11	0.07	17.71
COUNTRY	2.74	1.00	4.00
R_D_EUR (Mio. EUR)	44.04	0.23	454.69
R_D_PAT (Mio. EUR/ Patent)	3.67	0.00	26.48
R_D_EMP (Mio. EUR/ Employee)	16.30	0.02	215.69
PA_EMP (PAT / EMP)	20.76	0.19	989.58
GER	0.29		
FRA	0.07		
GBR	0.23		
OTH	0.40		

Table 6.4: Summary statistics for the covariates considering small campanies

boosted generalized semi-structured mixed model (bgssmm). An huge number of patents a year seems to influence the outsourcing process positive. On the other side an increasing number of research and development expenses shortens the tendency to source out. The effect of the time in the study may be neglected. Companies which are very volatile in their patent portfolio seem to fancy with outsourcing.

The model computed is given by

Covariate	Estimated Effect
Intercept	3.749
GER:	-0.236
FRA:	0.329
GBR:	-0.794

Random Effect	Estimate
σ_b^2	2.574

Table 6.5: Estimated Fixed Effects and Random Effects Variance

$$
\begin{aligned}
\eta_{it} &= \eta_{it}^{add} + \eta_{it}^{par} + b_i, \\
\eta_{it}^{add} &= \alpha_{(1)}(PA\_YEAR_{it}) + \alpha_{(2)}(BREADTH_{it}) + \alpha_{(3)}(PAT\_PORT_{it}) + \alpha_{(4)}(EMP_{it}) \\
&\quad + \alpha_{(5)}(R\_D\_EUR_{it}) + \alpha_{(6)}(R\_D\_PAT_{it}) + \alpha_{(7)}(PA\_EMP_{it}) + \alpha_{(8)}(VOL_{it}) \\
&\quad + \alpha_{(9)}(YEAR_{it}) + \alpha_{(10)}(R\_D\_EMP_{it}), \\
\eta_{it}^{par} &= GER_{it}\beta_1 + FRA_{it}\beta_2 + GBR_{it}\beta_2, \\
OUT_{it}|\lambda_{it} &= Poisson(\lambda_{it}), \\
\lambda_{it} &= \mathbb{E}(OUT_{it}) = \exp(\eta_{it})
\end{aligned}
\tag{6.11}
$$

with $h(\eta_{it}) = \log(\eta_{it}$.

The mixed model method was not applicable since numerical problems occurred in the estimation.

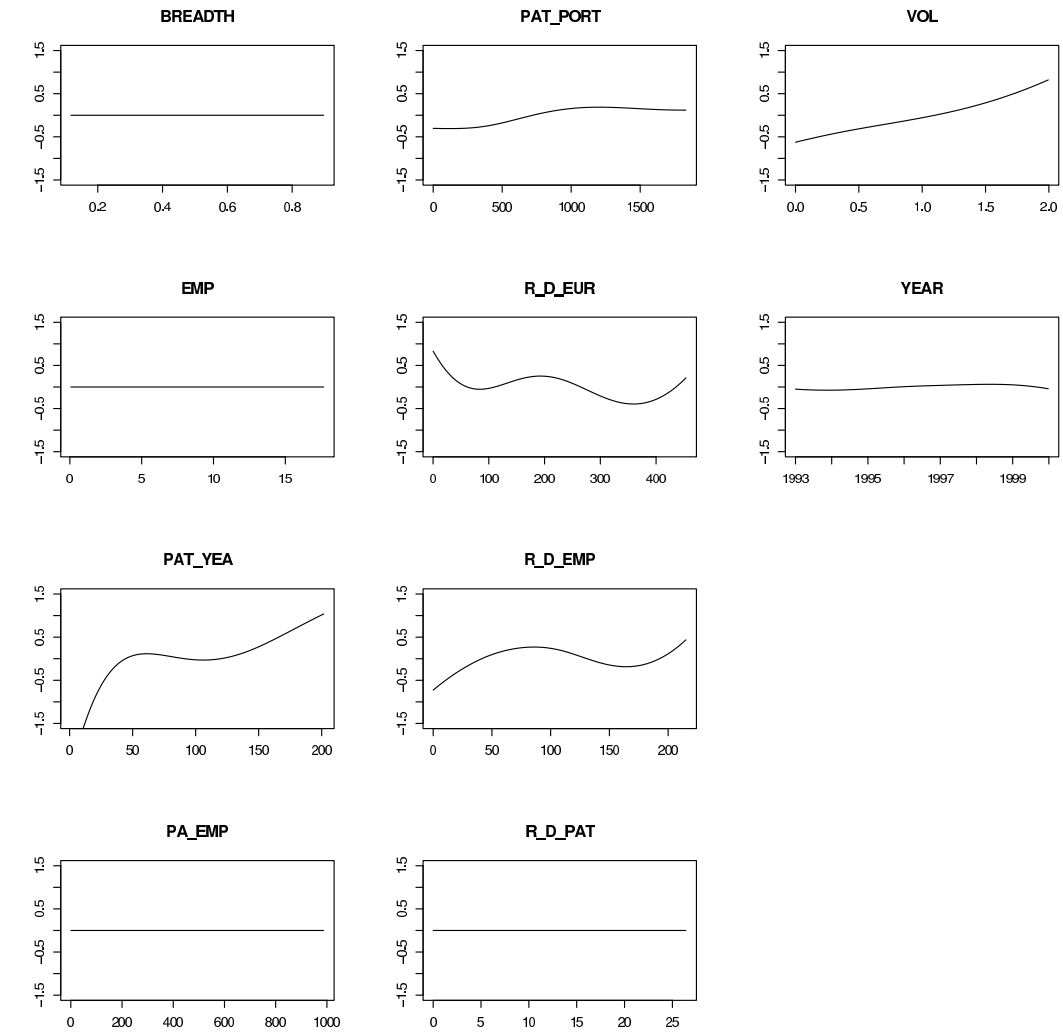


Figure 6.1: Patent data: estimated smooth effects for the patent data

Chapter 7

Summary and Perspectives

One can observe that likelihood boosting in mixed models is an important competitor to the mixed model strategy for getting estimates of additive smooth functions or linear parameters.

One conceptual disadvantage of the mixed model approach for smooth covariates is that the estimation of the smoothing parameter is quite sensitive to the signal strength in the data. For a large signal, which means only a small perturbation, the mixed model approach delivers only a rough approximation. The variance that has to be estimated is inversely proportional to the smoothing parameter which causes this phenomenon. So in fact, for a very small smoothing parameter, the variance of the random effects is very large and therefore good estimates for the variance are difficult to obtain. This effect is not restricted to the semi-structured mixed model but also for the generalized semi-structured mixed model which is shown in the simulation studies for additive covariates.

Another nice aspect of boosting in mixed models is that one can optimize many additive functions using the componentwise selection strategy. So high-dimensional function spaces can be optimized without the lack of stability and time which is a part of many classical simultaneous optimization procedures. This property of boosting is based on the similarity to the functional gradient descend algorithm, where the step-size for the direction are adjusted by the data. For L_2-loss the likelihood boosting concept can be described by a fixed step-size for the optimization direction. In the generalized semi-structured mixed models using Poisson data and many additive covariates the mixed model based optimization scheme did not converge for most of the datasets.

However for small signals, the mixed model approach provides better estimates of the parameters in many cases. One reason might be that the variables with small signals can not easily be separated from the error in the data. Since componentwise boosting is

an integrated selection algorithm for relevant variables, it might happen, that important variables that should have been selected, are not selected by the boosting algorithm. On the one side, this is bad for prediction which is visible in the mean squared error for the predictor. One the other side, if one is interested in relevant variables then some important variables are suppressed. The difference to classical selection strategies based on p-values is that the selection is based on the improvement of the likelihood. The improvement is corrected downward with a large penalty if the variable enters the model for the first time and only a small penalty if the variable has already been chosen. The penalty term originates in the selection criterion. So each variable has to compete with all other variables given the variable was selected or not. According to this idea one gets a natural order in increasing boosting steps from highly relevant to less relevant variables. Critics might be expressed by the choice of the selection and stopping criterion.

In this thesis, AIC and BIC were used to model the complexity of the data. In this connection, the objective is to find a model with relevant variables but not to much since one might over-parameterize the model. The boosting algorithm is stopped if the complexity criterion can not be improved any more by increasing boosting iterations. The problem now to choose a suitable complexity criterion. This is done empirically in this thesis . In the semi-structured mixed model cases, BIC showed generally better results in terms of MSE, AIC did not converge for settings with many additive covariates. One should mention that there exists no theory-based definition of the AIC-Criterion or BIC-Criterion in the mixed model methodology. The context is just transferred from the theory of cross sectional experiments and P-splines. But in this context, these criterions were only used as a crude criterion to scan for relevant variables and to stop the scanning process.

It is remarkable that componentwise boosting is a nice way to check complexity criteria in simulation studies. If one neglects the selection aspect, the job of the complexity criterion is to stop the algorithm at the right boosting step. In boosting the complexity is increasing from boosting step to another boosting step with small improvements which guarantees the weak learner concept. So once again one can use relevant and irrelevant variables to check if the complexity criterion finds out the right amount of relevant variables and rejects the irrelevant ones. For the additive models BIC was quite a good complexity or stopping criterion. But in this case a complexity criterion might be found via boosting that improves the results of BIC. Especially the BIC criterion provide for comparable results, one obtains by the mixed model approach for settings, that have only relevant additive variables. This idea of boosted information criterion follows the idea of Bühlmann and Yu, where the complexity of the model should be optimized. Just from parametric approaches, the AIC criterion showed better results than BIC for both types, mixed model and generalized linear mixed models.

So one idea might be to distinguish the selection criterion from the stopping criterion. One might think of threshold concepts or information criterion based concepts which seem to have a connection among each other. The selection concept might be improvable in the first step. Here, thee aspects of concurvity or multicollinearity can be plugged in the selection criterion as another penalty on the likelihood. The complexity criterion reflects just the best adjustment to the data which can be clearly separated from the former question.

Another nice point to be mentioned is that the componentwise selection strategy is especially suitable for high dimensional covariates. It combines the idea of a forward selection strategy without the iteration until convergence. Instead, the relevance of a candidate variable is judged by the selection criterion given all other covariates. The computation of the next candidate variables is based on the variables that were already selected. Effects on other variables by taking in a new variable is corrected in the consecutive boosting steps. A forward selection strategy is highly sensitive to the variables that enters the model. On the other side for high dimensional covariates, the computational effort is almost unbearable. For just a few covariates the forward selection strategy delivers comparable results, but take more time.

One may criticize the use of the Laplacian approximation for generalized semi-structured mixed models. For small datasets and binary data, one gets heavy biased estimates sometimes and another point for the accuracy of the estimates is the number of measurements in the cluster. The less measurements one has, the harder it is to compute the random effects variances. What is getting evident is that, if one studies the literature for generalized linear mixed models that one operates in areas where matrix algebra is just a small part to solve estimation problems. Concepts like quadrature or Monte-Carlo-Integration use weighted version of linear equations which are computer intense to solve. Moreover getting a hint on effective degrees of the computed model is only possible in some very special cases. Therefore the Laplacian approximation uses the idea of a linearized generalized mixed models. The computations are made using this framework but they are just necessary approximations to utilize the already developed concept. These approximations might be improved by better ones. But this also affects the mixed model approach to generalized semi-structured mixed models which uses the same approximation to get estimates. It should be noted, that the mixed model approach need not to compute a quasi hat-matrix, which is costly in computational effort. For the semi-structured mixed model, fast decompositions of the hat-matrix can be found. For the generalized semi-structured mixed model, efficient decompositions of hat-matrices in boosting are not known. On the other hand a crossed random effects model has to be computed where the marginal variances are not diagonal any more. This problem makes the mixed model approach also

very computer intensive.

This thesis encompasses only covariates that have a metric or binary covariates. Variables that have an ordinal and categorial scheme or cause interactions with metric variables are not handled. But further research on these aspects would be precious. Stratified variables are also a problem in mixed models so one can do research about this as well as on variable selection in varying coefficient models. Is a varying coefficient model necessary for getting additional information or is just a normal mixed model suitable to the problem. Variable selection strategies and special complexity criteria have to be developed in these cases. Boosting may be a nice toolkit in further research.

Last aspect to summarize is the idea of flexible splines. In the literature, one can find proposals where each cluster is characterized by its individual function in semi-parametric mixed models. So the individuality grows by allowing separate developments of these functions in the same covariate. Another interest focused in this context is to reduce the individuality to a common spline function and detached cluster specific function. The parameters in the last case are estimated by fitting the unknown random effects vector. The assumption here is that the mean of all these coefficients are derived from a density function with unknown diagonal variance. In the example of Ebay data where only a few, sometimes only one observation was collected, this idea is hard to implement, because limited observations are available to estimate the already described random coefficients of a random effects model. On the other side, one gets a large number of parameters to estimate. A sparse alternative is suggested in this thesis . The common spline function is modified by one random effect which disperses the spline function from the zero function or shorten the spline function towards zero. It may be seen as a generalization of random slopes to smooth functions. In this case only the coefficients for the common effects and additionally a random effects matrix for intercept and modifications on the functions have to be estimated. Since for this concept one has to optimized multiplicative effects it became apparent that using boosting techniques may be a way of handling such problems.

Appendix A: Splines

A.1 Solving Singularities

The problem is given by

$$\eta_{(i)} = \beta_0 + \Phi_{(i)}^T \alpha,$$

where $\Phi_{(i)} = \phi^T(u_{(i)}) = [\phi^{(1)}(u_{(i)}), \ldots, \phi^{(M)}(u_{(i)})$. Here Φ has dimension $N \times M$. In matrix notation one can write with $\eta^T = (\eta_{(1)}, \ldots, \eta_{(N)})$, $\Phi^T = [\Phi_{(1)}, \ldots, \Phi_{(N)}]$

$$\eta = \begin{bmatrix} 1 & \Phi \end{bmatrix} \begin{bmatrix} \beta_0 \\ \alpha \end{bmatrix} = X\delta.$$

The spline matrix Φ has to be reparametrized by a matrix T to a nonsingular $\tilde{X} = \begin{bmatrix} 1 & \tilde{\Phi}(u) \end{bmatrix}$.

A.1.1 Truncated Power Series for Semi-Parametric Models

Since for Truncated Power Series the Spline basis $\mathcal{B}$ has an element Φ_1 which consists of ones, the necessary transformation has simply to delete the first entry of this basis.

The transformation matrix doing this job has the form

$$T = \begin{bmatrix} 0_{(M-1)\times(1)} | I_{(M-1)} \end{bmatrix}^T.$$

So one gets

$$\alpha = T\tilde{\alpha},$$

$$\tilde{\Phi} = \Phi T,$$

$$\tilde{K} = T^T \breve{K} T.$$

A.1.2 Parametrization of α and Φ Using Restrictions

Identification problems and singularities may be solved by a suitable transformations of the centered basis coefficients.

$$\sum_{i=1}^{M} \alpha_i = 0 \text{ can be expressed by } \alpha_M = -\sum_{i=1}^{M-1} \alpha_i$$

The consequence of this representation is that designmatrix and difference penalty have to be modified accordingly. So one estimates with $M-1$ parameters $\tilde{\alpha}_j, j \in \{1, \ldots, M-1\}$ which are collected in $\tilde{\alpha}$. So the difference matrix D^d has to rewritten in $\tilde{D}^d$.

The transformation matrix doing this job has the form

$$T = \left[I_{(M-1)} | -1_{(M-1)} \right]^T .$$

So one gets

$$\alpha = T\tilde{\alpha},$$
$$\tilde{\Phi} = \Phi T,$$
$$\tilde{K} = T^T K T = (\tilde{D}^D)^T \tilde{D}^d$$

Detailed information of reparametrization by incorporating restrictions on P-splines is given in Scholz (2003) for one and more dimensional B-Splines. So incorporating the described restriction delivers

$$\Phi\alpha = \tilde{\Phi}\tilde{\alpha}$$

A.1.3 Parametrization of α and Φ Using Mixed Models

The use of B-Splines is sketched in the following . For simplicity, only one smooth component is considered with $\Phi_1(u), \ldots, \Phi_M(u)$ denoting the B-Splines for equidistant knots $k_1, \ldots, k_M$. First the spline basis $\mathcal{B}$ is transformed by an orthogonal decomposition to another spline basis $\breve{\mathcal{B}}$, consisting of $\breve{\Phi}_i, i = 1, \ldots, M$.

Example A.1 : Changing the B-Spline basis
First the difference matrix D^d is considered corresponding to B-Spline penalization (see Eilers & Marx (1996)). With D being the $(M-1) \times M$ contrast matrix

$$D = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}$$

one obtains higher order differences by the recursion $D^d = DD^{d-1}$ which is a $(M-d) \times M$ matrix. The penalty term is based on $\tilde{K} = (D^d)^T D^d$. New matrices $\tilde{X}_{(d)}$, depending on the order of the penalized differences are defined by

$$\tilde{X}_{(1)} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \tilde{X}_{(2)} = \begin{pmatrix} 1 & k_1 \\ \vdots & \vdots \\ 1 & k_M \end{pmatrix}, \tilde{X}_{(3)} = \begin{pmatrix} 1 & k_1 & k_1^2 \\ \vdots & \vdots & \vdots \\ 1 & k_M & k_M^2 \end{pmatrix}.$$

For differences of order d one consider the $(M-d)\times M$ *matrix* $\tilde{Z}^T_{(d)} = (D^d(D^d)^T)^{-1}D^d$. *In the following we drop the notation of d and set* $D := D^d$,$\tilde{Z} := \tilde{Z}_{(d)}$ *and* $\tilde{X} := \tilde{X}_{(d)}$. *So* $\tilde{Z}$ *and* $\tilde{X}$ *have the properties* $D\tilde{X} = 0$, $\tilde{Z}^T\tilde{X} = (DD^T)^{-1}D\tilde{X} = 0$,$\tilde{X}^TK\tilde{X} = 0 = \tilde{X}^TD^TD\tilde{X} = (D\tilde{X})^T(D\tilde{X})$. *Important is the equation*

$$\tilde{Z}^TK\tilde{Z} = (DD^T)^{-1}DD^TDD^T(DD^T)^{-1} = I_{(M-d)}.$$

since α *can be decomposed into* $\alpha = \tilde{X}\breve{\alpha}_1 + \tilde{Z}\breve{\alpha}_2$. *The orthogonal matrices* $\tilde{X}$ *and* Z *are used in the following way*

$$\Phi\alpha = \Phi[\tilde{X}\breve{\alpha}_1 + \tilde{Z}\breve{\alpha}_2] = [\Phi\tilde{X}, \Phi Z]\breve{\alpha} = \breve{\Phi}\breve{\alpha}$$

with $\breve{\alpha}^T = (\breve{\alpha}_1^T, \breve{\alpha}_2^T)$. *The new spline basis* $\breve{\mathcal{B}} = \{\breve{\Phi}_1, \ldots, \breve{\Phi}_M\}$ *consists of the columns of* $\breve{\Phi}$. *The corresponding penalty matrix is* $\breve{K} = bdiag(0_{(d)\times(d)}, I_{(M-d)\times(M-d)})$. □

Benefit of using the spline basis $\breve{\mathcal{B}}$ is that singularities can be avoided by deleting $\breve{\Phi}_1$, which holds $\breve{\Phi}_1 = 1$.

The transformation matrix doing this job has the form

$$T = \left[0_{(M-1)\times(1)}|I_{(M-1)}\right]^T.$$

So one gets

$$\breve{\alpha} = T\tilde{\alpha},$$
$$\tilde{\Phi}(u) = \breve{\Phi}(u)T,$$
$$\tilde{K} = T^T\breve{K}T.$$

For details on this reparametrization see Green (1987).

A.2 Smoothing with Mixed Models

The use of B-Splines is sketched in the following . For simplicity only one smooth component is considered with $\Phi_{(1)}(u), \ldots, \Phi_{(M)}(u)$ denoting the B-Splines for equidistant knots $k_1, \ldots, k_M$ and $y_i = X_i\beta + \Phi_i\alpha$ denoting the predictor.

We use the transformed spline basis $\breve{\mathcal{B}}$ as described in example A.1

The predictor can now be rewritten in the form

$$\begin{aligned} y_i &= \Big[X_i, \Phi_i\Big] \begin{bmatrix}\beta \\ \alpha\end{bmatrix} + Z_i b_i = \Big[X_i, \Phi_i\Big] \begin{bmatrix}\beta \\ \tilde{X}\breve{\alpha}_1 + \tilde{Z}\breve{\alpha}_2\end{bmatrix} + Z_i b_i \\ &= \Big[X_i, \Phi(u_i)\tilde{X}, \Phi(u_i)\tilde{Z}\Big] \begin{bmatrix}\beta \\ \breve{\alpha}_1 \\ \breve{\alpha}_2\end{bmatrix} + Z_i b_i \\ &= \Big[X_i, \Phi_i\tilde{X}\Big] \begin{bmatrix}\beta \\ \breve{\alpha}_1\end{bmatrix} + \Big[\Phi_i\tilde{Z}, Z_i\Big] \begin{bmatrix}\breve{\alpha}_2 \\ b_i\end{bmatrix} \end{aligned}$$

with $\Phi(u_i)$ as a matrix for a vector $u_i^T = (u_{i1}, \ldots, u_{it})$. $\Phi(u_i)$ has rows $\phi(u_{ij})^T = (\phi_1(u_{ij}), \ldots, \phi_M(u_{ij}))$.

The penalized log-likelihood of the linear mixed model simplifies to

$$\begin{aligned} l_p(\delta) &= \textstyle\sum_{i=1}^n \log(f(y_i|\delta; b_i)p(b_i)) - \lambda\delta^T Diag(0_{(p\times p)}, \lambda K)\delta \\ &= \textstyle\sum_{i=1}^n \log(f(y_i|\delta; b_i)p(b_i)) - \lambda((\tilde{X}\breve{\alpha}_1 + \tilde{Z}\breve{\alpha}_2)^T K(\tilde{X}\breve{\alpha}_1 + \tilde{Z}\breve{\alpha}_2 \\ &= \textstyle\sum_{i=1}^n \log(f(y_i|\delta; b_i)p(b_i)) - \frac{1}{2}\breve{\alpha}_2^T 2 * \lambda I_{(M-d)}\breve{\alpha}_2. \end{aligned}$$

with $\delta^T = (\beta, \alpha)$.

This corresponds to the BLUP criterion of the mixed model

$$y_i = \tilde{X}_i\tilde{\beta} + \Big[\Phi(u_i)\tilde{Z} \quad Z\Big] \begin{pmatrix}\breve{\alpha}_1 \\ b_i\end{pmatrix} + \epsilon_i$$

$$\text{with} \qquad \begin{pmatrix}\tilde{\alpha} \\ b_i \\ \epsilon\end{pmatrix} \sim N\left(\begin{pmatrix}0 \\ 0 \\ 0\end{pmatrix}, \begin{pmatrix}\frac{1}{2\lambda}I & 0 & 0 \\ 0 & Q(\rho) & 0 \\ 0 & 0 & \sigma_\epsilon^2 I\end{pmatrix}\right)$$

and $\tilde{\beta}^T = (\beta^T, \breve{\alpha}_1)$, $\tilde{X}_i = [X_i, \Phi(u_i)\tilde{X}]$. Thus, from decomposition $\alpha = \tilde{X}\breve{\alpha}_1 + \tilde{Z}\breve{\alpha}_2$ one obtains a mixed model with uncorrelated parameters $\breve{\alpha}_2$.

Appendix B: Parametrization of covariance structures

To make sure, that the notation is clear in all parts of the paper, a short sketch of handling covariances and its parametrization is proposed.

B.1 Independent Identical

This structure is has only one parameter, so $\rho^T = (\rho_1)^T$. So

$$Q(\rho) = \rho_1^2 * I$$

The elementwise derivative is

$$\frac{\partial Q(\rho)}{\partial \rho_1} = 2\rho_1 * I$$

B.2 Independent but Not Identical

If d is the dimension of the covariance matrix, then the structure has d parameters, so $\rho^T = (\rho_1, \ldots, \rho_d)^T$. So

$$Q(\rho) = \begin{bmatrix} \rho_1^2 & & \\ & \ddots & \\ & & \rho_d^2 \end{bmatrix}$$

The elementwise derivative is

$$\frac{\partial Q(\rho)}{\partial \rho_i} = DQ_i = \begin{cases} (DQ_i)_{jj} = 2 * \rho_j & \text{if } j = i \\ 0 & \text{sonst} \end{cases}$$

B.3 Unstructured

Since $Q(\rho)$ is a symmetric, positive semidefinite Matrix, $Q(\rho)$ can be parametrisized

$$Q(\rho) = L * L^T$$

where L is the Cholesky root of $Q(\rho)$. So $\rho = vec(L)$ is the adequate parametrisation of $Q(\rho)$.

For example

$$L = \begin{bmatrix} L_{11} & 0 \\ L_{12} & L_{22} \end{bmatrix}$$

So $vec(L)^T = (L_{11}, L_{12}, L_{22})^T = \rho^T$. The zeros are omitted.

d is the dimension of the covariance matrix. If $\rho_i = (L)_{jj}, j \in \{1, \ldots, d\}$ (is diagonalelement of L) the elementwise derivative are

$$\frac{\partial Q(\rho)}{\partial \rho_i} = \frac{\partial Q(\rho)}{\partial L_{jj}} = DQ_i = \begin{cases} (DQ_i)_{jj} = 2 * L_{jj} & \text{if } k = j \\ (DQ_i)_{kj} = (DQ_i)_{jk} = L_{kj} & \text{if } k > j \\ 0 & \text{else} \end{cases}$$

If $\rho_i \in (L)_{ij}, i = 1, \ldots, d, i \neq j$ (is not diagonal element of L) the elementwise derivative are

$$\frac{\partial Q(\rho)}{\partial \rho_i} = \frac{\partial Q(\rho)}{\partial L_{jk}} = DQ_i = \begin{cases} (DQ_i)_{ll} = 2 * L_{jk} & \text{if } l = j \\ (DQ)_{lj} = (DQ)_{jl} = L_{lk} & \text{if } l \neq j \\ 0 & \text{else} \end{cases}$$

Appendix C: Simulation Studies

C.1 Mixed Model Approach vs. BoostMixed

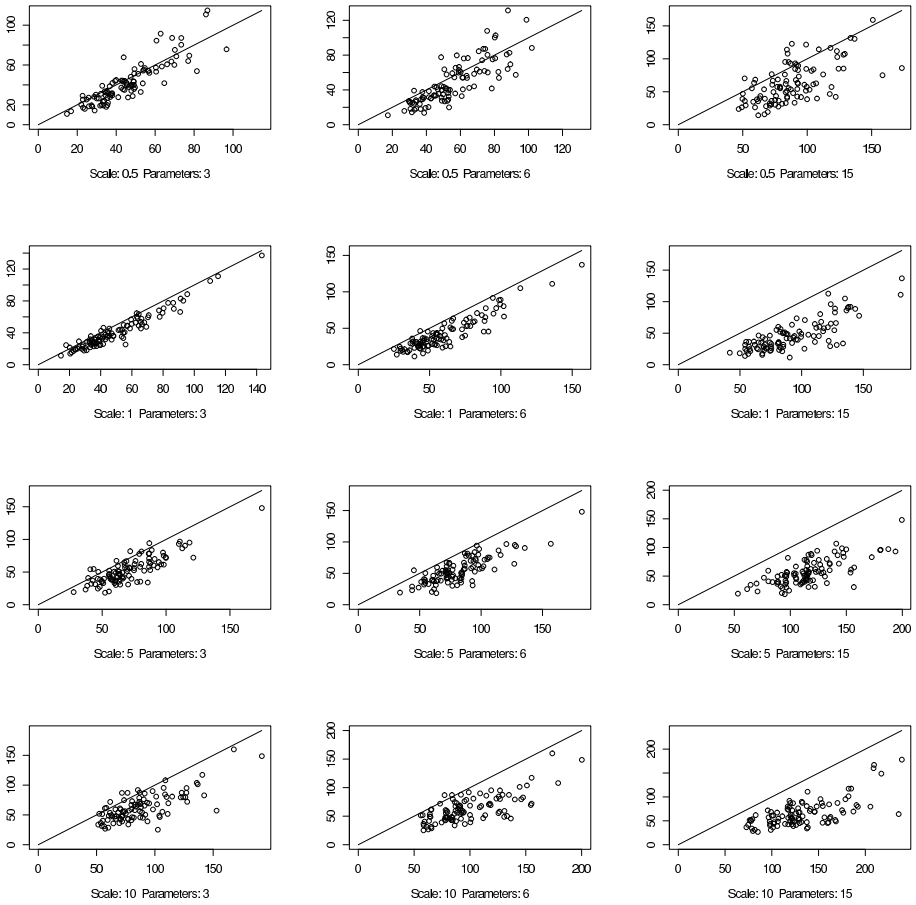


Figure C.1: Simulation study 5: MSE_η of BoostMixed (y-axis)and mixed model approach (x-axis)

		MM						BoostMixed									
c	p	mse_η	mse_f	mse_{σ_b}	$\text{mse}_{\sigma_\epsilon}$	Steps	Time	mse_η	mse_f	mse_{σ_b}	$\text{mse}_{\sigma_\epsilon}$	Steps	Time	FalsePos	FalseNeg	Initial	Selected
0.5	3	45.791	37.701	0.026	0.117	13	0.09	42.079	36.867	0.026	0.115	9.9	0.0	0.0	0.1	2.0	2.9
0.5	6	55.721	48.399	0.030	0.117	18	0.41	48.666	45.112	0.028	0.114	10.2	0.0	0.4	0.1	2.0	3.3
0.5	15	88.005	85.470	0.031	0.129	25	7.03	62.501	62.270	0.029	0.114	9.7	0.1	0.9	0.2	2.0	3.7
0.5	25							73.134	74.790	0.030	0.116	9.8	0.1	1.2	0.3	2.0	3.9
1.0	3	50.448	37.422	0.024	0.126	8	0.06	41.946	31.226	0.026	0.119	19.7	0.0	0.0	0.0	2.0	3.0
1.0	6	60.520	48.547	0.024	0.120	15	0.33	42.773	32.237	0.026	0.120	19.7	0.1	0.0	0.0	2.0	3.0
1.0	15	92.705	85.021	0.028	0.120	21	6.05	46.662	36.725	0.029	0.120	20.0	0.2	0.2	0.0	2.0	3.2
1.0	25							50.440	41.102	0.028	0.118	20.2	0.3	0.3	0.0	2.0	3.3
5.0	3	71.243	60.651	0.032	0.187	12	0.08	53.399	47.592	0.031	0.181	144.6	0.4	0.0	0.0	1.9	3.0
5.0	6	82.051	72.296	0.031	0.185	14	0.32	55.396	49.947	0.031	0.182	146.9	0.4	0.1	0.0	1.9	3.1
5.0	15	116.472	113.781	0.036	0.190	20	5.87	57.510	52.545	0.032	0.182	145.2	2.3	0.2	0.0	1.9	3.2
5.0	25							58.533	53.910	0.034	0.182	145.5	3.4	0.2	0.0	1.9	3.2
10.0	3	88.045	71.694	0.027	0.264	14	0.10	62.981	59.701	0.029	0.139	495.6	1.1	0.0	0.0	3.0	3.0
10.0	6	98.669	84.396	0.026	0.226	17	0.40	62.981	59.701	0.029	0.139	495.6	2.6	0.0	0.0	3.0	3.0
10.0	15	132.549	125.730	0.033	0.239	24	7.11	65.726	62.807	0.033	0.139	492.1	6.7	0.1	0.0	3.0	3.1
10.0	25							66.588	63.895	0.033	0.139	490.9	12.0	0.1	0.0	3.0	3.1

Table C.1: Study 5

		MM						BoostMixed									
c	p	mse_η	mse_f	mse_{σ_b}	$\text{mse}_{\sigma_\epsilon}$	Steps	Time	mse_η	mse_f	mse_{σ_b}	$\text{mse}_{\sigma_\epsilon}$	Steps	Time	FalsePos	FalseNeg	Initial	Selected
0.5	3	46.503	34.107	0.022	0.133	13	0.09	45.416	36.576	0.026	0.136	9.9	0.0	0.0	0.1	2.0	2.9
0.5	6	57.421	48.626	0.024	0.133	18	0.42	50.530	43.280	0.028	0.139	10.3	0.0	0.3	0.1	2.0	3.2
0.5	15	90.615	92.066	0.029	0.135	28	8.30	64.707	61.314	0.032	0.140	11.0	0.1	0.8	0.2	2.0	3.7
0.5	25							72.285	70.857	0.035	0.141	11.5	0.2	1.1	0.2	2.0	3.9
1.0	3	49.449	40.515	0.033	0.146	9	0.06	40.716	34.440	0.035	0.145	17.4	0.0	0.0	0.0	2.0	3.0
1.0	6	60.771	54.728	0.037	0.148	16	0.37	42.105	36.107	0.037	0.143	17.6	0.1	0.1	0.0	2.0	3.0
1.0	15	93.651	97.541	0.038	0.151	21	6.41	43.327	37.663	0.037	0.144	17.7	0.2	0.1	0.0	2.0	3.1
1.0	25							46.404	41.527	0.036	0.145	17.9	0.4	0.2	0.0	2.0	3.2
5.0	3	72.155	62.797	0.023	0.153	12	0.09	53.174	49.862	0.025	0.153	109.6	0.3	0.0	0.0	3.0	3.0
5.0	6	82.856	77.115	0.025	0.157	14	0.33	53.663	50.515	0.026	0.154	109.5	0.6	0.0	0.0	3.0	3.0
5.0	15	114.390	118.645	0.028	0.156	18	5.25	54.918	51.990	0.026	0.154	109.4	1.5	0.1	0.0	3.0	3.1
5.0	25							56.471	53.814	0.027	0.154	109.1	2.6	0.1	0.0	3.0	3.1
10.0	3	93.000	77.369	0.029	0.230	14	0.09	68.369	63.423	0.030	0.184	430.2	1.1	0.0	0.0	3.0	3.0
10.0	6	103.896	92.147	0.028	0.225	15	0.34	69.027	64.432	0.030	0.184	430.0	2.2	0.0	0.0	3.0	3.0
10.0	15	136.460	137.261	0.035	0.184	20	5.81	70.142	65.935	0.031	0.180	428.9	5.7	0.1	0.0	3.0	3.1
10.0	25							73.504	70.497	0.031	0.181	427.1	7.9	0.2	0.0	3.0	3.2

Table C.2: Study 2

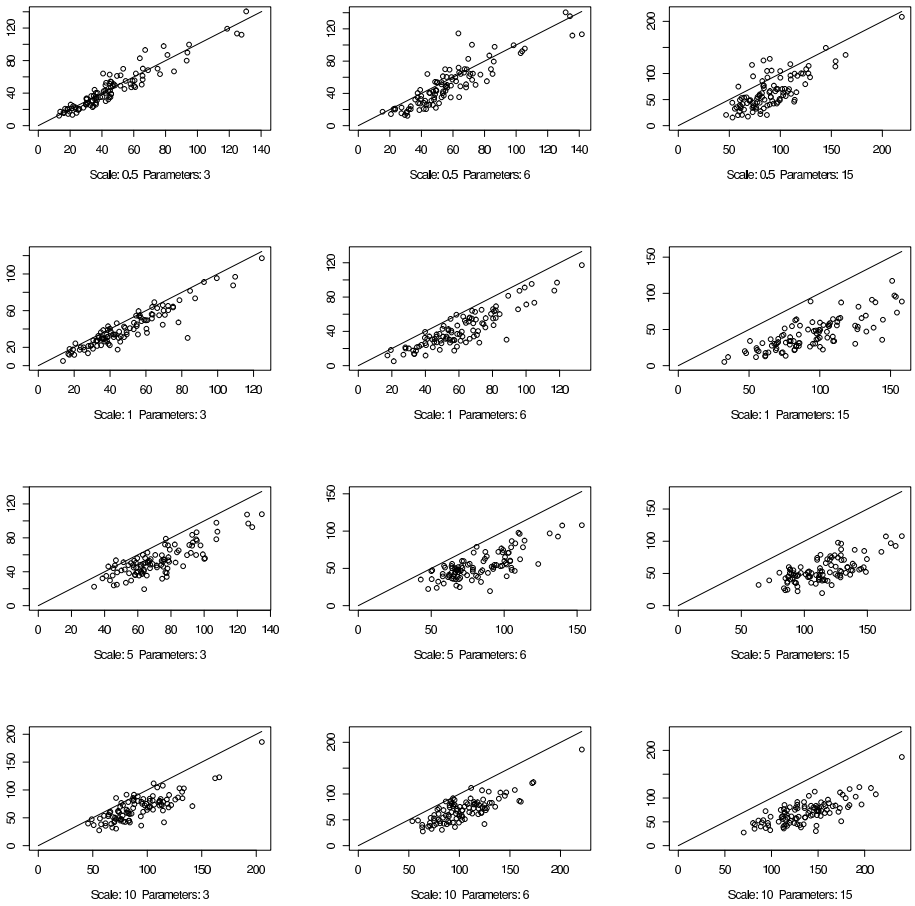


Figure C.2: Simulation study 6: MSE_η of BoostMixed (y-axis)and mixed model approach (x-axis)

		MM						BoostMixed									
c	p	mse_η	mse_f	mse_{σ_b}	$\text{mse}_{\sigma_\epsilon}$	Steps	Time	mse_η	mse_f	mse_{σ_b}	$\text{mse}_{\sigma_\epsilon}$	Steps	Time	FalsePos	FalseNeg	Initial	Selected
0.5	3	59.357	41.040	0.150	0.862	17	0.04	61.277	46.459	0.154	0.811	15.3	0.1	0.0	0.3	2.0	2.7
0.5	6	73.025	56.516	0.170	0.860	21	0.20	76.980	65.766	0.173	0.813	13.6	0.5	0.9	0.4	2.0	3.5
0.5	15	127.888	127.512	0.188	0.947	28	6.92	106.143	103.125	0.182	0.848	11.4	0.9	1.8	0.8	2.0	4.0
0.5	25							121.971	123.515	0.203	0.841	12.4	1.6	2.2	1.0	2.0	4.2
1.0	3	72.348	56.262	0.158	0.702	13	0.03	64.580	51.397	0.172	0.672	28.8	0.5	0.0	0.0	2.0	3.0
1.0	6	90.224	77.686	0.173	0.714	18	0.17	81.472	73.019	0.191	0.697	28.9	1.1	0.7	0.0	2.0	3.6
1.0	15	150.190	158.350	0.256	0.710	26	6.32	102.419	99.349	0.253	0.715	36.5	2.9	1.2	0.1	2.0	4.2
1.0	25							112.858	112.811	0.299	0.713	34.0	3.9	1.5	0.1	2.0	4.4
5.0	3	96.755	82.750	0.123	0.797	13	0.03	70.340	58.043	0.156	0.607	202.0	3.3	0.0	0.0	3.0	3.0
5.0	6	112.757	102.820	0.128	0.738	15	0.14	71.819	59.890	0.159	0.609	203.2	2.1	0.1	0.0	3.0	3.0
5.0	15	167.118	179.655	0.186	0.779	19	4.72	83.092	75.498	0.202	0.613	206.0	6.7	0.4	0.0	3.0	3.4
5.0	25							94.376	90.400	0.261	0.643	212.6	11.7	0.7	0.0	3.0	3.7

Table C.3: Study 3

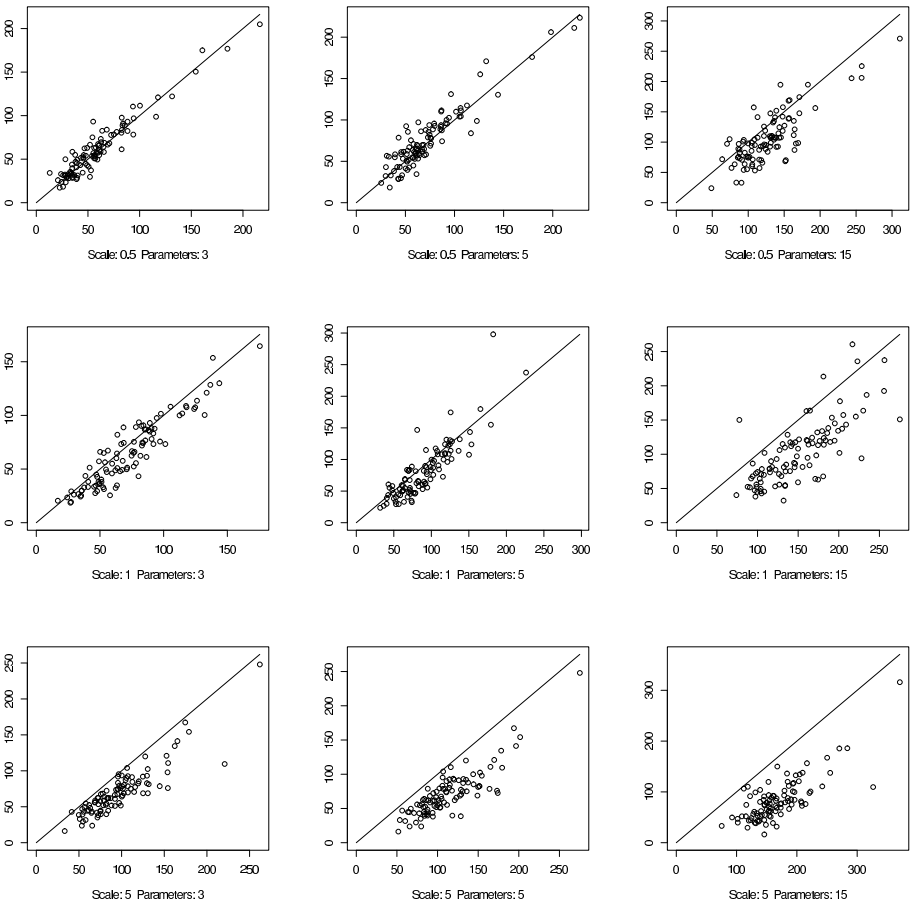


Figure C.3: Simulation study 1: MSE_η of BoostMixed (y-axis)and mixed model approach (x-axis)

		MM						BoostMixed									
c	p	mse_η	mse_f	mse_{σ_b}	mse_{σ_ϵ}	Steps	Time	mse_η	mse_f	mse_{σ_b}	mse_{σ_ϵ}	Steps	Time	FalsePos	FalseNeg	Initial	Selected
0.5	3	98.413	38.248	0.114	5.338	15	0.02	101.372	45.198	0.139	4.780	46.1	0.8	0.0	0.3	2.0	2.7
0.5	6	113.736	54.938	0.118	5.381	19	0.13	121.564	69.512	0.166	4.833	37.6	1.2	1.1	0.5	2.0	3.6
0.5	15	160.835	116.182	0.132	5.380	30	5.66	146.019	101.852	0.211	4.834	27.2	2.0	2.1	0.8	2.0	4.2
0.5	25							166.688	130.291	0.297	4.817	37.5	4.8	2.8	0.9	2.0	4.8
1.0	3	99.531	51.153	0.108	4.211	12	0.02	89.402	43.097	0.115	3.805	48.8	0.8	0.0	0.0	2.0	3.0
1.0	6	113.800	68.266	0.120	4.206	16	0.11	100.060	56.704	0.135	3.810	51.6	1.4	0.6	0.0	2.0	3.6
1.0	15	163.859	133.089	0.123	4.335	28	5.34	123.632	86.472	0.191	3.812	57.8	3.9	1.4	0.0	2.0	4.4
1.0	25							141.182	110.096	0.256	3.821	58.3	2.2	2.0	0.1	2.0	5.0
5.0	3	143.293	78.221	0.102	4.096	13	0.02	120.386	57.729	0.146	3.747	303.8	1.9	0.0	0.0	2.8	3.0
5.0	6	156.224	93.300	0.108	4.077	15	0.10	124.271	62.650	0.158	3.733	303.4	3.5	0.2	0.0	2.8	3.2
5.0	15	205.228	160.784	0.135	4.363	24	4.66	138.146	82.855	0.224	3.809	308.5	9.0	0.7	0.0	2.9	3.7
5.0	25							157.296	109.001	0.317	3.767	304.0	11.3	1.5	0.0	2.9	4.5

Table C.4: Study 4

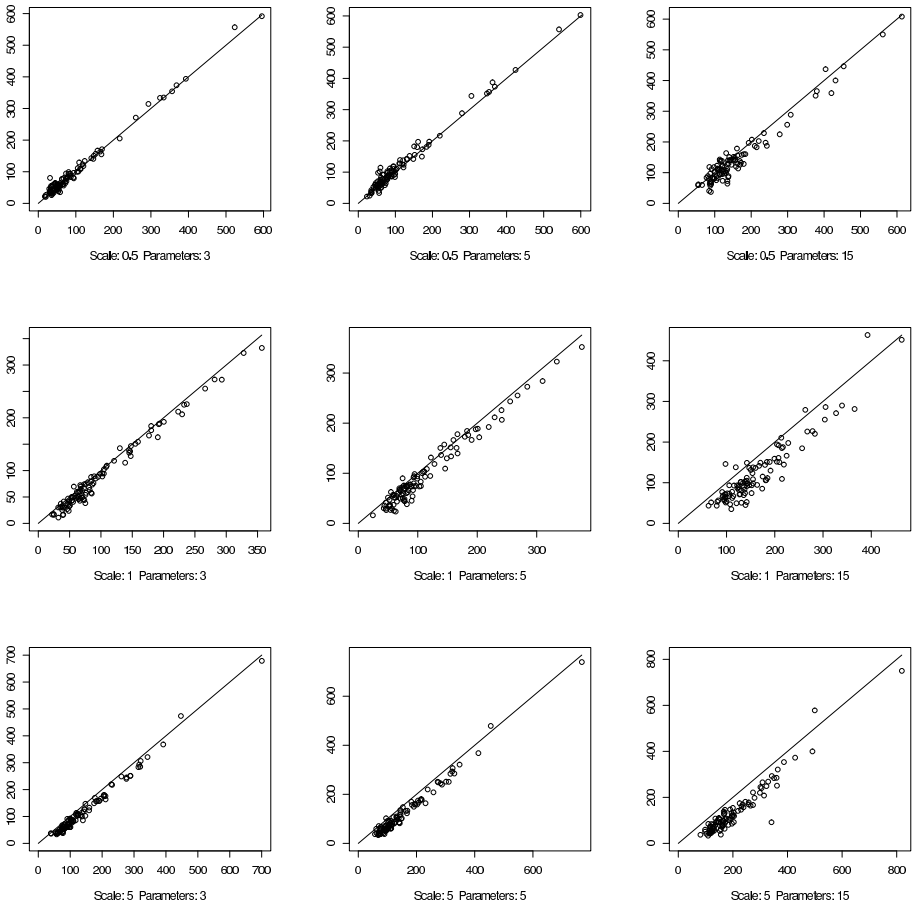


Figure C.4: Simulation study 2: MSE_η of BoostMixed (y-axis)and mixed model approach (x-axis)

		MM						BoostMixed									
c	p	mse_η	mse_f	mse_{σ_b}	mse_{σ_ϵ}	Steps	Time	mse_η	mse_f	mse_{σ_b}	mse_{σ_ϵ}	Steps	Time	FalsePos	FalseNeg	Initial	Selected
0.5	3	66.286	52.069	0.209	1.722	16	0.11	71.142	62.030	0.210	2.096	8.3	0.0	0.0	0.3	2.0	2.7
0.5	6	86.792	75.244	0.230	1.772	22	0.49	95.563	92.137	0.230	2.398	7.2	0.4	1.1	0.5	2.0	3.6
0.5	15	167.329	174.776	0.337	2.046	30	13.06	136.678	144.482	0.204	2.979	8.1	0.9	2.2	1.2	2.0	4.0
0.5	25							158.212	171.764	0.219	3.297	6.5	1.2	2.5	1.4	2.0	4.1
1.0	3	91.187	78.666	0.200	2.097	14	0.10	81.414	74.290	0.209	2.379	30.9	0.7	0.0	0.0	2.0	3.0
1.0	6	112.376	104.646	0.229	2.107	19	0.43	107.823	107.660	0.216	2.590	20.9	1.0	0.8	0.1	2.0	3.8
1.0	15	189.637	205.956	0.310	2.124	30	13.18	140.558	150.621	0.241	2.905	20.6	2.3	1.3	0.2	2.0	4.0
1.0	25							157.894	174.234	0.281	3.182	20.2	3.3	1.5	0.4	2.0	4.1
5.0	3	125.484	121.703	0.261	2.343	13	0.09	81.755	77.670	0.251	2.031	167.6	3.2	0.0	0.0	2.8	3.0
5.0	6	150.929	152.391	0.285	2.704	15	0.34	86.438	83.302	0.276	2.008	170.6	0.7	0.1	0.0	2.8	3.1
5.0	15	234.111	267.276	0.364	2.104	23	10.08	97.044	97.570	0.308	2.045	166.8	1.7	0.3	0.0	2.8	3.3
5.0	25							100.519	102.959	0.314	2.005	166.9	3.5	0.4	0.0	2.8	3.4

Table C.5: Study 5

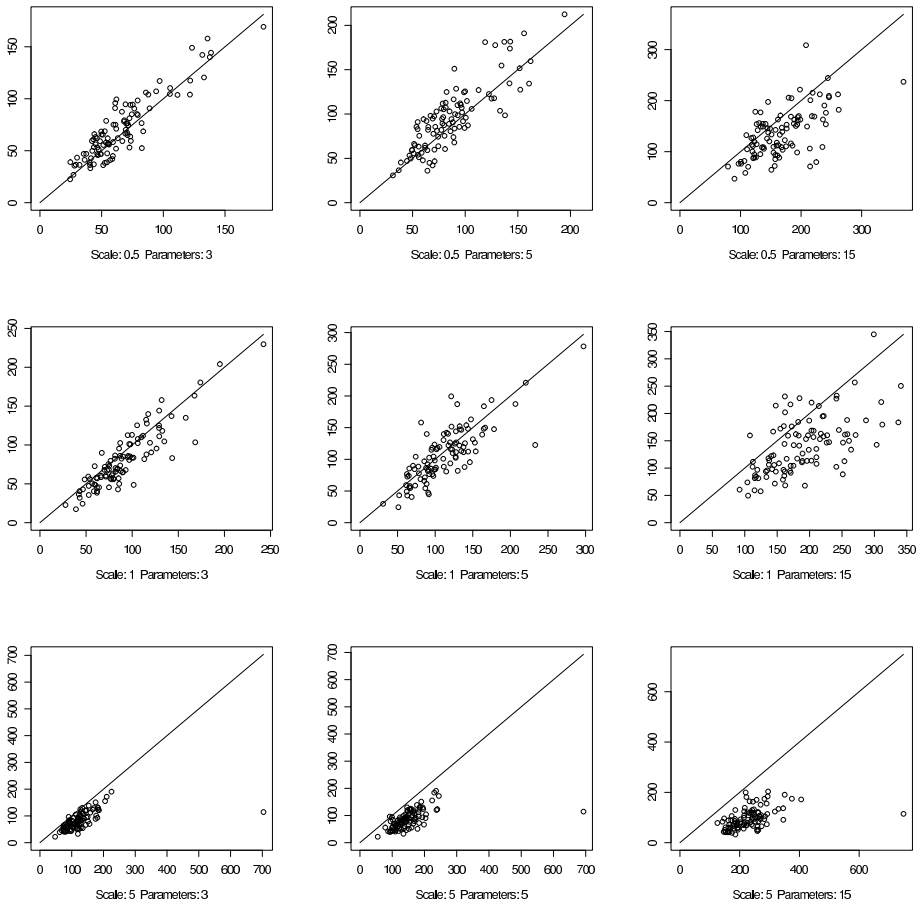


Figure C.5: Simulation study 3: MSE_η of BoostMixed (y-axis)and mixed model approach (x-axis)

		MM						BoostMixed									
c	p	mse_η	mse_f	mse_{σ_b}	$\text{mse}_{\sigma_\epsilon}$	Steps	Time	mse_η	mse_f	mse_{σ_b}	$\text{mse}_{\sigma_\epsilon}$	Steps	Time	FalsePos	FalseNeg	Initial	Selected
0.5	3	406.115	134.189	0.027	1.153	8	0.79	399.614	137.878	0.027	1.134	5.1	2.9	0.0	0.0	2.0	3.0
0.5	6	449.401	177.529	0.027	1.155	12	2.77	451.231	195.191	0.028	1.136	5.5	5.1	0.8	0.0	2.0	3.8
0.5	15	590.300	320.097	0.026	1.155	18	28.04	482.218	229.798	0.028	1.134	5.8	0.4	1.0	0.0	2.0	4.0
0.5	25							496.580	245.561	0.029	1.131	5.9	0.6	1.0	0.0	2.0	4.0
1.0	3	409.284	167.122	0.037	1.442	7	0.75	378.237	150.048	0.038	1.460	5.3	0.1	0.0	0.0	2.0	3.0
1.0	6	454.819	213.754	0.038	1.444	12	2.64	403.424	178.154	0.039	1.462	6.2	0.2	0.3	0.0	2.0	3.3
1.0	15	592.514	355.240	0.038	1.443	16	25.53	445.306	223.659	0.040	1.461	7.5	0.4	0.7	0.0	2.0	3.7
1.0	25							465.749	245.175	0.040	1.459	8.0	0.8	0.9	0.0	2.0	3.9
5.0	3	499.925	253.122	0.031	1.442	11	1.03	432.461	232.640	0.032	1.450	74.2	0.9	0.0	0.0	3.0	3.0
5.0	6	541.312	295.061	0.031	1.445	12	2.76	446.733	248.923	0.033	1.450	75.4	1.5	0.1	0.0	3.0	3.1
5.0	15	672.337	428.405	0.031	1.443	16	24.61	462.807	266.004	0.033	1.452	76.1	3.3	0.3	0.0	3.0	3.3
5.0	25							481.958	288.545	0.034	1.451	77.3	5.8	0.4	0.0	3.0	3.4

Table C.6: Study 6

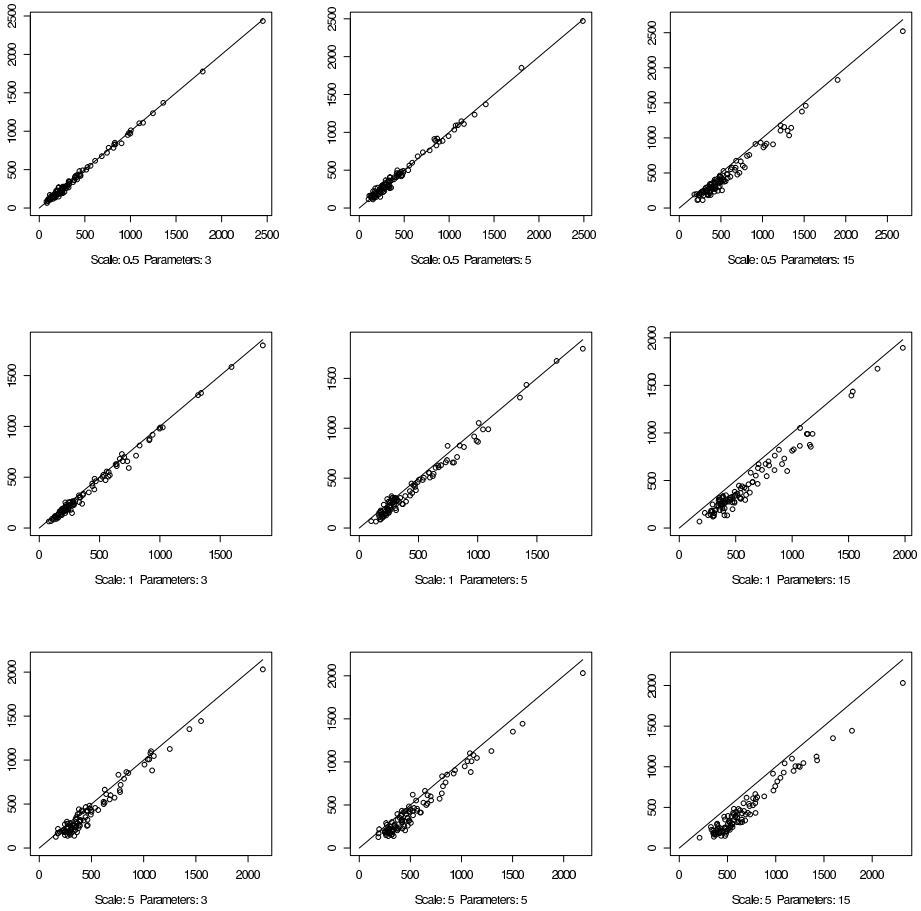


Figure C.6: Simulation study 4: MSE_η of BoostMixed (y-axis)and mixed model approach (x-axis)

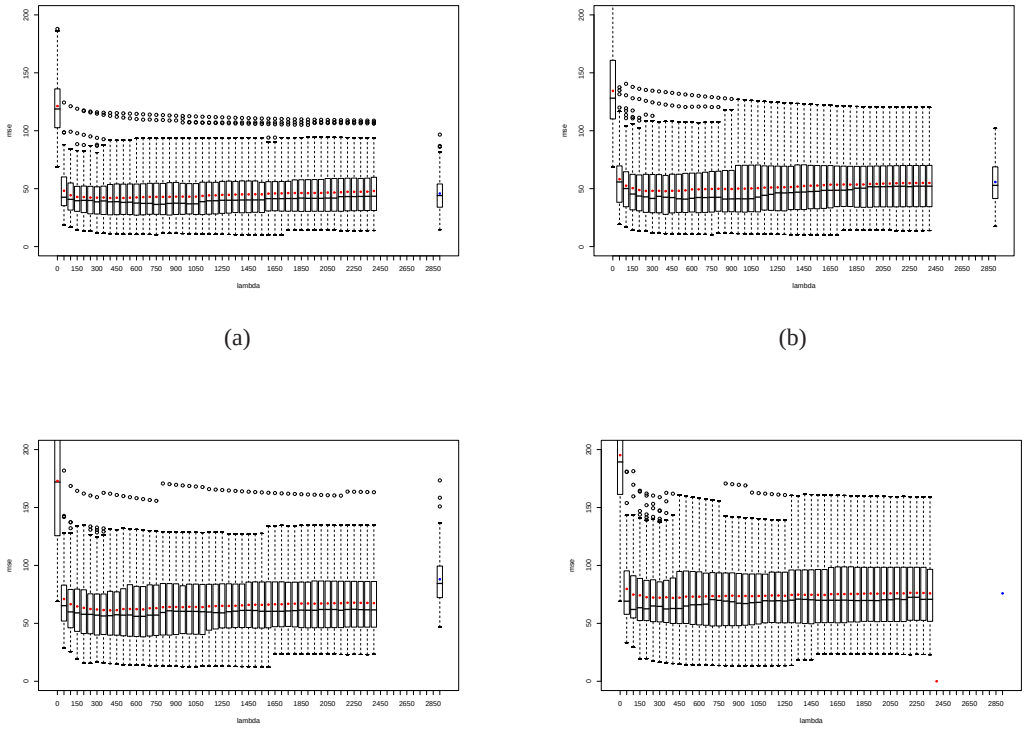
		MM					BoostMixed								
c	p	mse_η	mse_{σ_b}	mse_{σ_ϵ}	Steps	Time	mse_η	mse_{σ_b}	mse_{σ_ϵ}	Steps	Time	FalsePos	FalseNeg	Initial	Selected
0.5	6	55.883	1.049	1.135	23	0.44	63.562	1.941	1.924	4.6	0.0	0.0	2.4	2.0	3.6
0.5	15	86.980	1.077	1.140	30	9.12	78.204	1.924	1.901	5.2	0.1	0.8	2.9	2.0	4.0
0.5	25						85.125	1.918	1.888	5.1	0.1	1.1	3.1	2.0	4.0
1.0	6	71.221	1.079	1.074	18	0.33	79.341	1.945	1.961	11.6	0.1	0.0	1.8	2.0	4.2
1.0	15	105.589	1.117	1.081	29	8.24	87.140	1.935	1.955	11.7	0.2	0.2	1.8	2.0	4.4
1.0	25						91.876	1.931	1.949	11.7	0.3	0.4	1.9	2.0	4.5
5.0	6	94.113	1.136	1.109	11	0.21	78.574	1.872	1.962	79.8	0.3	0.0	0.0	2.9	6.0
5.0	15	125.063	1.152	1.110	17	4.78	80.397	1.866	1.963	79.6	0.9	0.1	0.0	2.9	6.1
5.0	25						81.504	1.862	1.963	79.8	1.4	0.1	0.0	2.9	6.1

		Forward						
c	p	mse_η	mse_{σ_b}	mse_{σ_ϵ}	Time	FalsePos	FalseNeg	Selected
0.5	6	58.894	0.027	0.139	1.084	1.0	4.0	3.0
0.5	15	65.833	0.027	0.140	2.789	1.0	4.0	3.0
0.5	25							
1.0	6	81.499	0.027	0.133	1.932	2.0	3.0	5.0
1.0	15	88.720	0.027	0.136	5.915	2.0	3.0	5.0
1.0	25							
5.0	6	97.554	0.027	0.132	2.699	4.0	3.0	7.0
5.0	15	106.336	0.031	0.135	11.466	4.0	3.0	7.0
5.0	25							

Table C.7: Study 7

C.2 Choosing an Appropriate Smoothing Parameter and an Appropriate Selection Criterion

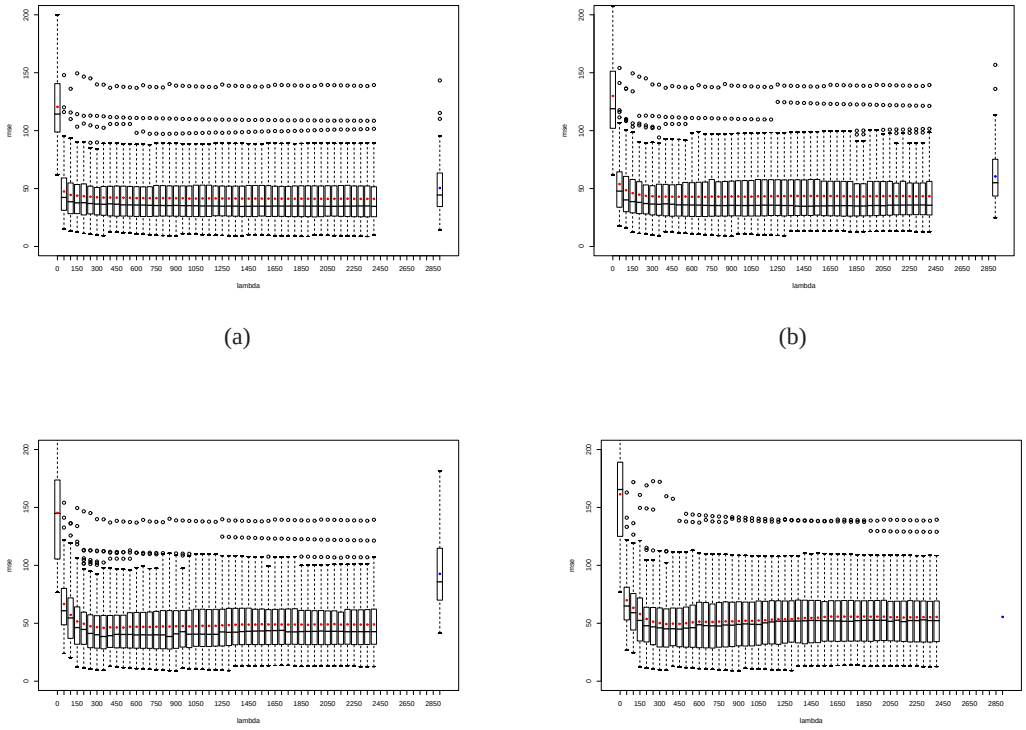
C.2.1 BIC as Selection/Stopping Criterion



(a) (b)

(c) (d)

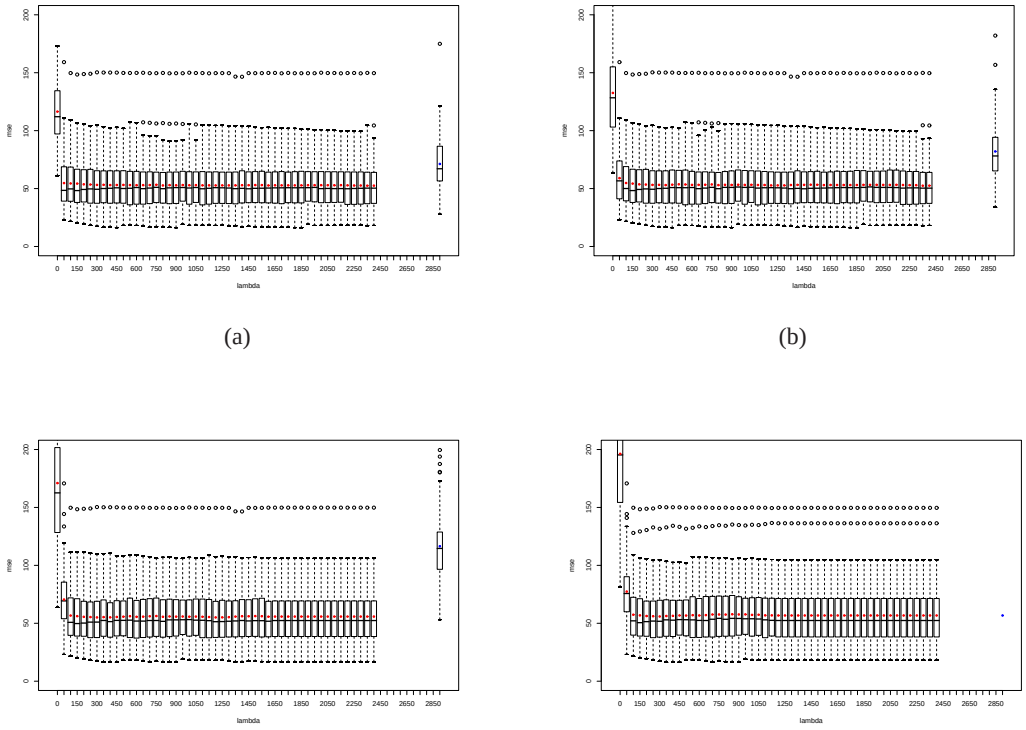
Figure C.7: The distributions of the mean squared errors for different counts of parameters in the model are presented by boxplots. BIC was used as selection and stopping criterion. The red points are the means for the mses depending on different lambdas. On the right side the distribution of the mses of the mixed model approach is plotted. The blue point is the mean of the mses of the mixed model approach. (a) 3 parameters used (b) 5 parameters used (c) 15 parameters used and (d) 25 parameters used. c was chosen to be $c = 0.5$



(a) (b)

(c) (d)

Figure C.8: The distributions of the mean squared errors for different counts of parameters in the model are presented by boxplots. BIC was used as selection and stopping criterion. The red points are the means for the mses depending on different lambdas. On the right side the distribution of the mses of the mixed model approach is plotted. The blue point is the mean of the mses of the mixed model approach. (a) 3 parameters used (b) 5 parameters used (c) 15 parameters used and (d) 25 parameters used. c was chosen to be $c = 0.1$

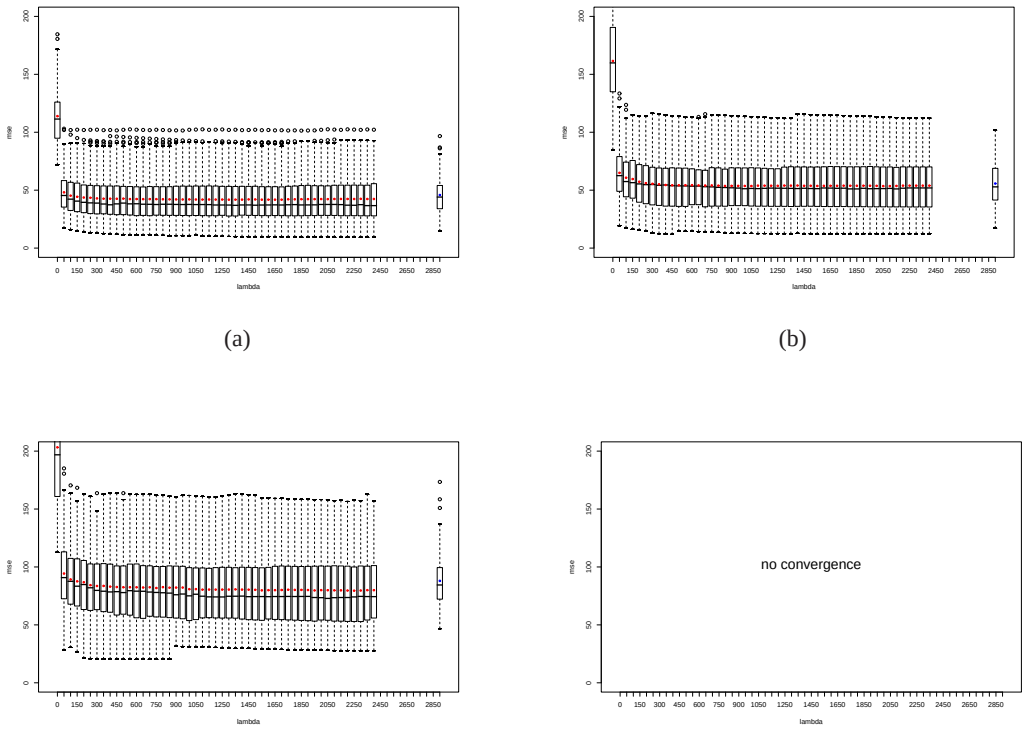


(a) (b)

(c) (d)

Figure C.9: The distributions of the mean squared errors for different counts of parameters in the model are presented by boxplots. BIC was used as selection and stopping criterion. The red points are the means for the mses depending on different lambdas. On the right side the distribution of the mses of the mixed model approach is plotted. The blue point is the mean of the mses of the mixed model approach. (a) 3 parameters used (b) 5 parameters used (c) 15 parameters used and (d) 25 parameters used. c was chosen to be $c = 5$

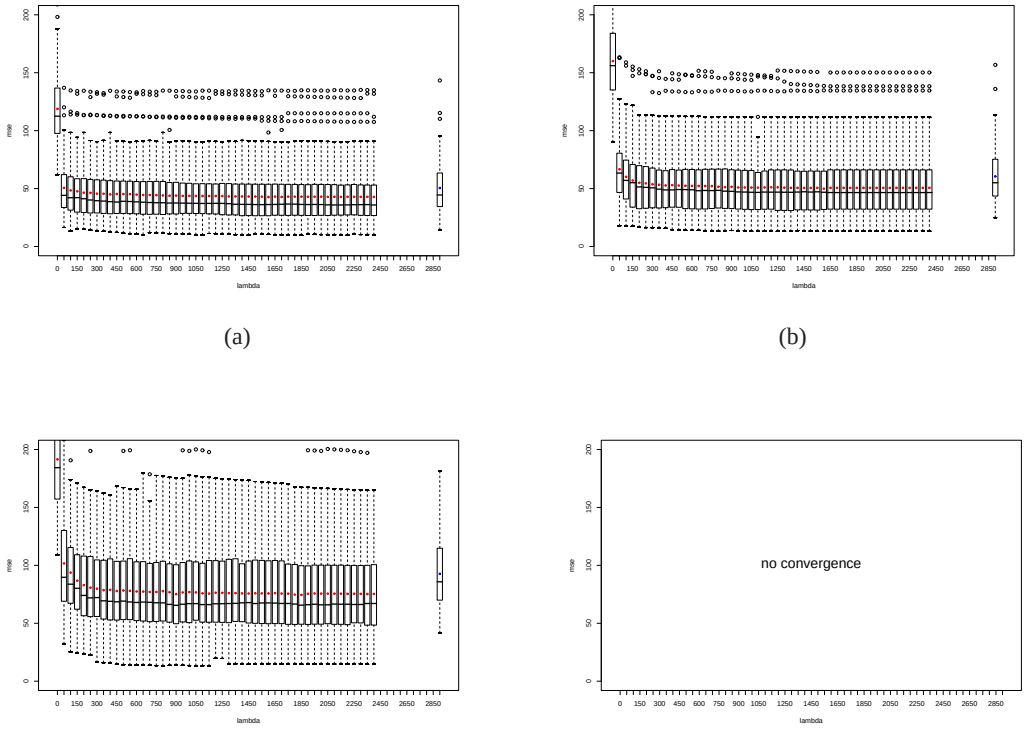
C.2.2 AIC as Selection/Stopping Criterion



(a) (b)

(c) (d)

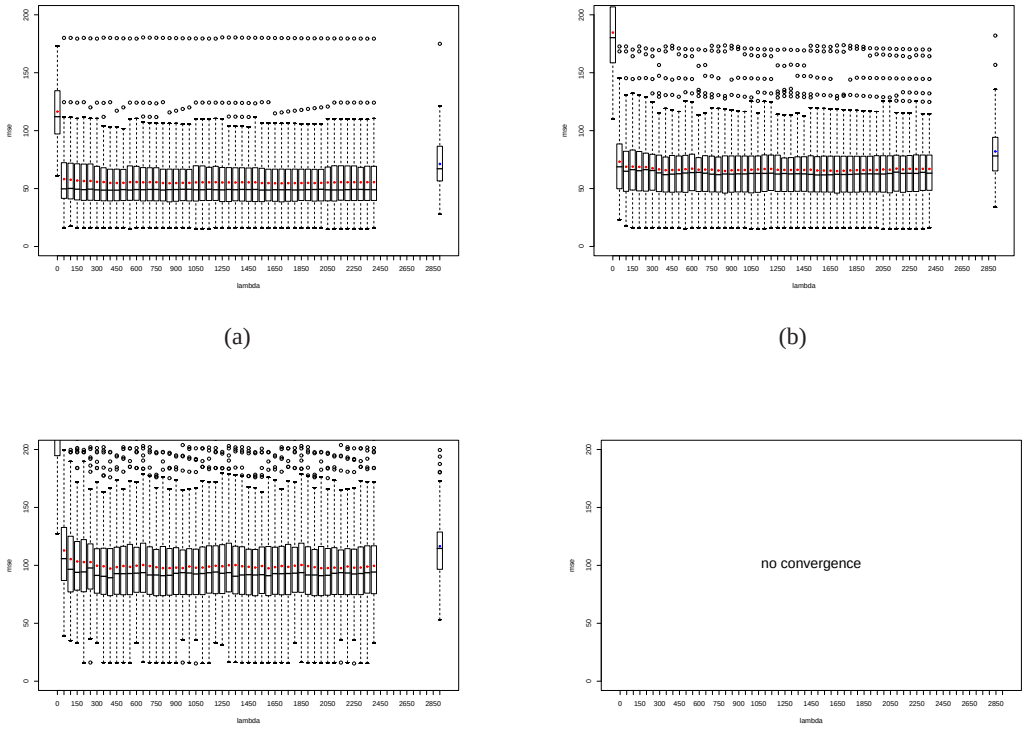
Figure C.10: The distributions of the mean squared errors for different counts of parameters in the model are presented by boxplots. BIC was used as selection and stopping criterion. The red points are the means for the mses depending on different lambdas. On the right side the distribution of the mses of the mixed model approach is plotted. The blue point is the mean of the mses of the mixed model approach. (a) 3 parameters used (b) 5 parameters used (c) 15 parameters used and (d) 25 parameters used. c was chosen to be $c = 0.5$



(a) (b)

(c) (d)

Figure C.11: The distributions of the mean squared errors for different counts of parameters in the model are presented by boxplots. AIC was used as selection and stopping criterion. The red points are the means for the mses depending on different lambdas. On the right side the distribution of the mses of the mixed model approach is plotted. The blue point is the mean of the mses of the mixed model approach. (a) 3 parameters used (b) 5 parameters used (c) 15 parameters used and (d) 25 parameters used. c was chosen to be $c = 0.1$



(a) (b)

(c) (d)

Figure C.12: The distributions of the mean squared errors for different counts of parameters in the model are presented by boxplots. AIC was used as selection and stopping criterion. The red points are the means for the mses depending on different lambdas. On the right side the distribution of the mses of the mixed model approch is plotted. The blue point is the mean of the mses of the mixed model approach. (a) 3 parameters used (b) 5 parameters used (c) 15 parameters used and (d) 25 parameters used. c was chosen to be $c = 5$

C.3 Linear BoostMixed

We present simulation studies in which the performance of BoostMixed is compared the the common mixed model. The underlying model is the random intercept model

$$y_{it} = b_i + x_{it}^T\beta + \epsilon_{it}, t = 1, \ldots, 5, i = 1, \ldots, 80$$

with $x_{it}^T = (x_{it1}, \ldots, x_{itp})$, where $x_{its}, s = 1, \ldots, p$ a realizations of a random variable X_{it} with a uniform distribution with variance 10 for each component and $p = 40$. The elements of $\beta^T = (\beta_1, \ldots, \beta_p)$ are set to

$$\beta_i = \begin{cases} c * \frac{5}{i} & \text{if}, i \leq 5 \\ 0 & \text{else} \end{cases}.$$

For the covariates constant pairwise correlation is assumed, i.e. x_{it} has the correlation structure, i.e.

$$cor(X_{it}) = \begin{bmatrix} 1 & \varrho & \ldots & \varrho \\ \varrho & 1 & \ldots & \\ \ddots & \ddots & \ddots & \varrho \\ \ldots & \ldots & \varrho & 1 \end{bmatrix}.$$

The constant signal c determines the signal of the covariates. The random effect and the noise variable have been specified by $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 2$ and $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 2$. The shrinkage parameter λ was set to 100 . The performance of estimators is evaluated separately for the structural components and variance. By averaging across 100 datasets we consider mean squared errors for $\eta, \sigma_\epsilon^2, \sigma_b^2$ given by

$$\text{mse}_\eta = \sum_{i=1}^n \sum_{t=1}^T (\eta_{it} - \hat{\eta}_{it})^2, \hat{\eta}_{it} = x_{it}^T\hat{\beta}, \qquad \text{mse}_\beta = ||\beta - \hat{\beta}||^2,$$

$$\text{mse}_{\sigma_b^2} = ||\sigma_b^2 - \hat{\sigma}_b^2||^2, \qquad \text{mse}_{\sigma_\epsilon^2} = ||\sigma_\epsilon^2 - \hat{\sigma}_\epsilon^2||^2.$$

For a more extensive analysis of BoostMixed six simulation studies with different settings were made. In all studies 100 datasets were generated

Study 9 - Start setting
The random effect and the noise variable have been specified by $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 3$ and $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 3$. In the part of the study which is presented the number of observations has been chosen by $n = 100, T = 5$. Pairwise correlation was taken to be $\varrho = 0.1$. Details can be found in Table C.8.

Study 10 - small variances
The random effect and the noise variable have been specified by $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 2$ and $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 1$. In the part of the study which is presented the number of observations has been chosen by $n = 80, T = 5$. Pairwise correlation was taken to be $\varrho = 0.1$. Details can be found in Table C.9.

Study 11 - big clusters
The random effect and the noise variable have been specified by $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 1$ and $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 2$. In the part of the study which is presented the number of observations has been chosen by $n = 50, T = 10$. Pairwise correlation was taken to be $\varrho = 0.1$. Details can be found in Table C.10.

Study 12 - big dataset, small variances
The random effect and the noise variable have been specified by $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 2$ and $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 1$. In the part of the study which is presented the number of observations has been chosen by $n = 200, T = 5$. Pairwise correlation was taken to be $\varrho = 0.1$. Details can be found in Table C.11.

Study 13 - big dataset, huge variances
The random effect and the noise variable have been specified by $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 3$ and $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 3$. In the part of the study which is presented the number of observations has been chosen by $n = 200, T = 5$. Pairwise correlation was taken to be $\varrho = 0.1$. Details can be found in Table C.12.

Study 14 - correlated data
The random effect and the noise variable have been specified by $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 3$ and $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 3$. In the part of the study which is presented the number of observations has been chosen by $n = 100, T = 5$. Pairwise correlation was taken to be $\varrho = 0.5$. Details can be found in Table C.13.

BoostMixed is compared to the classical mixed model with all covariates (MM) and to the mixed model with an integrated forward selection (forward). It is quite similar to the BoostMixed algorithm since one starts with the intercept model. In every step all remaining covariates are fitted separately. The covariate characterized by the best improvement of the AIC-Criterion is taken into the model and seen as relevant. The selection is stopped if the complexity criterion can not improved any more. So the extreme case is that for 25 covariates with 25 relevant covariates. Here $\sum_{i=1}^{20} i = 210$ models have to be computed for the forward selection. For the simulation study with 20 covariates the number of computed models is quite moderate with up to $\sum_{i=14}^{20} i = 119$ (6 variables selected). For 100 covariates with 5 relevant covariates nearly 585 models have to be computed if 6 variables are selected. It seen in Tables C.8 - C.13 that forward selection procedures

take a very long time. For example in Table C.8 it took averaged 1.8 minutes for 20 covariates (strength=5) to get an estimate. In comparison the BoostMixed approach took 3.6 seconds to find the relevant variables. But unfortunately a small mount of irrelevant variables were selected which downgrade the mse_η. Along the mean squared error for the predictor mse_η, the mean squared errors for the parameters mse_β, for the noise and random variance $\text{mse}_{\sigma_\epsilon}$ and mse_{σ_b}, the steps (Steps) until convergence, the variables that were selected but have no relevance (FalsePos) and the variables that have relevance but were not selected (FalseNeg) were collected and averaged.

		MM						BoostMixed								
c	p	mse_η	mse_β	mse_{σ_b}	mse_{σ_ϵ}	Steps	Time	mse_η	mse_β	mse_{σ_b}	mse_{σ_ϵ}	Selected	Steps	Time	FalsePos	FalseNeg
0.5	5	35.812	0.001	0.042	0.257	8.0	0.020	35.804	0.004	0.038	0.248	5.0	10.9	0.010	0.0	0.0
0.5	10	52.825	0.001	0.043	0.257	8.0	0.021	46.832	0.006	0.037	0.248	6.1	12.9	0.163	1.1	0.0
0.5	15	72.676	0.001	0.044	0.264	8.0	0.023	59.206	0.008	0.039	0.255	7.0	13.9	0.025	2.0	0.0
0.5	20	90.250	0.001	0.044	0.262	8.0	0.025	68.894	0.010	0.039	0.252	7.7	15.5	0.043	2.7	0.0
1.0	5	41.067	0.001	0.050	0.227	9.0	0.021	41.107	0.004	0.049	0.227	5.0	10.4	0.505	0.0	0.0
1.0	10	57.843	0.001	0.050	0.226	9.0	0.024	51.628	0.006	0.049	0.224	6.1	12.5	0.011	1.1	0.0
1.0	15	74.845	0.001	0.049	0.224	9.0	0.026	62.178	0.008	0.051	0.223	7.0	13.8	0.018	2.0	0.0
1.0	20	92.257	0.001	0.050	0.223	9.0	0.027	71.030	0.010	0.054	0.227	7.7	15.1	0.034	2.7	0.0
5.0	5	35.824	0.001	0.048	0.238	10.9	0.029	35.820	0.004	0.047	0.231	5.0	12.3	0.008	0.0	0.0
5.0	10	54.534	0.001	0.047	0.242	11.0	0.031	47.547	0.006	0.048	0.235	6.0	15.0	0.021	1.1	0.0
5.0	15	72.608	0.001	0.046	0.243	11.0	0.034	58.047	0.009	0.050	0.234	6.9	15.7	0.023	1.9	0.0
5.0	20	90.507	0.001	0.046	0.241	10.9	0.037	67.930	0.011	0.053	0.231	7.7	16.5	0.060	2.7	0.0

		Forward					
c	p	mse_η	mse_{σ_b}	mse_{σ_ϵ}	Time	FalsePos	FalseNeg
0.5	5.0	35.812	0.042	0.257	0.234	0.0	0.0
0.5	10.0	45.693	0.041	0.259	0.776	1.0	0.0
0.5	15.0	51.182	0.041	0.264	1.326	1.0	0.0
0.5	20.0	53.436	0.040	0.261	1.869	1.0	0.0
1.0	5.0	41.067	0.050	0.227	0.253	0.0	0.0
1.0	10.0	50.052	0.049	0.226	0.837	1.0	0.0
1.0	15.0	53.781	0.048	0.225	1.427	1.0	0.0
1.0	20.0	56.495	0.048	0.225	2.007	1.0	0.0
5.0	5.0	35.824	0.048	0.238	0.229	0.0	0.0
5.0	10.0	46.174	0.047	0.241	0.767	1.0	0.0
5.0	15.0	50.244	0.046	0.243	1.301	1.0	0.0
5.0	20.0	52.772	0.046	0.238	1.840	1.0	0.0

Table C.8: Study 9

		MM						BoostMixed								
c	p	mse_η	mse_β	mse_{σ_b}	mse_{σ_ϵ}	Steps	Time	mse_η	mse_β	mse_{σ_b}	mse_{σ_ϵ}	Selected	Steps	Time	FalsePos	FalseNeg
0.5	5	17.634	0.000	0.020	0.040	8.1	0.021	17.644	0.002	0.020	0.038	5.0	11.4	0.479	0.0	0.0
0.5	10	28.488	0.000	0.020	0.040	8.1	0.024	24.482	0.004	0.022	0.038	6.0	12.5	0.014	1.0	0.0
0.5	15	39.756	0.000	0.021	0.040	8.1	0.025	30.840	0.005	0.024	0.038	6.8	13.6	0.024	1.8	0.0
0.5	20	52.100	0.000	0.022	0.040	8.1	0.026	37.945	0.007	0.026	0.038	7.6	14.2	0.021	2.6	0.0
1.0	5	18.083	0.000	0.018	0.041	9.0	0.025	18.092	0.002	0.018	0.039	5.0	11.1	0.010	0.0	0.0
1.0	10	29.109	0.000	0.018	0.041	9.0	0.027	25.384	0.004	0.020	0.040	6.0	11.8	0.019	1.0	0.0
1.0	15	40.983	0.000	0.019	0.043	9.1	0.029	32.865	0.005	0.021	0.040	6.9	13.3	0.059	1.9	0.0
1.0	20	52.068	0.000	0.018	0.043	9.2	0.031	39.054	0.007	0.023	0.041	7.6	13.9	0.037	2.6	0.0
5.0	5	17.603	0.000	0.019	0.037	11.1	0.029	17.605	0.002	0.019	0.036	5.0	12.7	0.007	0.0	0.0
5.0	10	29.539	0.000	0.019	0.037	11.2	0.032	25.041	0.004	0.020	0.037	6.0	14.2	0.013	0.9	0.0
5.0	15	40.185	0.000	0.019	0.037	11.2	0.034	30.561	0.005	0.022	0.037	6.7	15.6	0.019	1.7	0.0
5.0	20	52.018	0.000	0.020	0.038	11.2	0.035	37.263	0.007	0.024	0.036	7.5	17.1	0.030	2.5	0.0

		Forward					
c	p	mse_η	mse_{σ_b}	mse_{σ_ϵ}	Time	FalsePos	FalseNeg
0.5	5.0	17.634	0.020	0.040	0.237	0.0	0.0
0.5	10.0	23.661	0.020	0.040	0.790	1.0	0.0
0.5	15.0	26.033	0.020	0.039	1.348	1.0	0.0
0.5	20.0	27.875	0.020	0.040	1.896	1.0	0.0
1.0	5.0	18.083	0.018	0.041	0.256	0.0	0.0
1.0	10.0	24.537	0.018	0.041	0.843	1.0	0.0
1.0	15.0	27.463	0.018	0.041	1.431	1.0	0.0
1.0	20.0	28.871	0.018	0.041	2.015	1.0	0.0
5.0	5.0	17.603	0.019	0.037	0.229	0.0	0.0
5.0	10.0	24.130	0.019	0.037	0.766	1.0	0.0
5.0	15.0	26.375	0.019	0.037	1.305	1.0	0.0
5.0	20.0	28.188	0.019	0.037	1.846	1.0	0.0

Table C.9: Study 10

		MM						BoostMixed								
c	p	mse_η	mse_β	mse_{σ_b}	mse_{σ_ϵ}	Steps	Time	mse_η	mse_β	mse_{σ_b}	mse_{σ_ϵ}	Selected	Steps	Time	FalsePos	TrueNeg
0.5	5	25.682	0.000	0.021	0.056	8.0	0.011	25.673	0.002	0.021	0.055	5.0	10.8	0.026	0.0	0.0
0.5	10	37.220	0.000	0.020	0.056	8.0	0.012	33.229	0.004	0.022	0.054	6.1	12.2	0.013	1.1	0.0
0.5	15	48.282	0.000	0.021	0.056	8.0	0.012	39.882	0.005	0.023	0.055	7.0	13.0	0.029	2.0	0.0
0.5	20	58.691	0.000	0.021	0.057	8.0	0.013	45.756	0.006	0.024	0.055	7.7	13.5	0.031	2.7	0.0
1.0	5	23.124	0.000	0.024	0.055	9.0	0.012	23.121	0.002	0.024	0.053	5.0	10.2	0.007	0.0	0.0
1.0	10	33.922	0.000	0.025	0.056	9.0	0.012	30.497	0.004	0.025	0.053	6.1	11.6	0.021	1.1	0.0
1.0	15	43.586	0.000	0.025	0.056	9.0	0.013	35.471	0.005	0.026	0.053	6.8	12.7	0.016	1.8	0.0
1.0	20	55.047	0.000	0.025	0.055	9.0	0.015	42.103	0.006	0.027	0.053	7.6	14.0	0.037	2.6	0.0
5.0	5	23.958	0.000	0.019	0.050	11.0	0.014	23.963	0.002	0.018	0.049	5.0	12.7	0.010	0.0	0.0
5.0	10	34.041	0.000	0.020	0.050	11.0	0.015	29.627	0.003	0.018	0.048	5.8	14.0	0.015	0.8	0.0
5.0	15	44.475	0.000	0.019	0.050	11.0	0.016	35.380	0.005	0.017	0.048	6.5	15.2	0.022	1.5	0.0
5.0	20	54.941	0.000	0.019	0.050	11.0	0.018	41.185	0.006	0.018	0.049	7.3	16.6	0.049	2.4	0.0

		Forward					
c	p	mse_η	mse_{σ_b}	mse_{σ_ϵ}	Time	FalsePos	FalseNeg
0.5	5.0	25.682	0.021	0.056	0.152	0.0	0.0
0.5	10.0	31.776	0.021	0.056	0.502	1.0	0.0
0.5	15.0	33.995	0.021	0.056	0.854	1.0	0.0
0.5	20.0	35.080	0.021	0.056	1.209	1.0	0.0
1.0	5.0	23.124	0.024	0.055	0.160	0.0	0.0
1.0	10.0	29.118	0.024	0.055	0.532	1.0	0.0
1.0	15.0	31.024	0.024	0.055	0.900	1.0	0.0
1.0	20.0	32.848	0.024	0.055	1.269	1.0	0.0
5.0	5.0	23.958	0.019	0.050	0.151	0.0	0.0
5.0	10.0	29.508	0.019	0.050	0.500	1.0	0.0
5.0	15.0	32.372	0.018	0.050	0.856	1.0	0.0
5.0	20.0	33.442	0.018	0.050	1.203	1.0	0.0

Table C.10: Study 11

		MM						BoostMixed								
c	p	mse_η	mse_β	mse_{σ_b}	mse_{σ_ϵ}	Steps	Time	mse_η	mse_β	mse_{σ_b}	mse_{σ_ϵ}	Selected	Steps	Time	FalsePos	TrueNeg
0.5	5	17.906	0.000	0.012	0.022	8.0	0.086	17.904	0.001	0.013	0.022	5.0	11.7	0.010	0.0	0.0
0.5	10	29.418	0.000	0.012	0.023	8.0	0.086	25.838	0.002	0.013	0.023	6.2	12.8	0.018	1.2	0.0
0.5	15	39.788	0.000	0.012	0.023	8.0	0.101	30.654	0.003	0.014	0.023	6.8	13.2	0.027	1.8	0.0
0.5	20	50.718	0.000	0.012	0.023	8.1	0.109	36.056	0.003	0.014	0.023	7.4	13.9	0.037	2.4	0.0
1.0	5	18.852	0.000	0.009	0.014	9.0	0.097	18.863	0.001	0.009	0.014	5.0	11.9	0.011	0.0	0.0
1.0	10	31.039	0.000	0.009	0.014	9.0	0.105	27.166	0.002	0.009	0.014	6.1	12.8	0.018	1.1	0.0
1.0	15	43.521	0.000	0.009	0.014	9.0	0.112	34.495	0.003	0.009	0.014	7.0	13.6	0.028	2.0	0.0
1.0	20	54.969	0.000	0.009	0.014	9.0	0.118	41.612	0.004	0.009	0.014	7.9	15.2	0.041	2.9	0.0
5.0	5	19.249	0.000	0.010	0.018	11.0	0.108	19.249	0.001	0.011	0.018	5.0	13.3	0.011	0.0	0.0
5.0	10	30.618	0.000	0.010	0.018	11.0	0.114	25.986	0.002	0.011	0.018	5.9	14.4	0.019	0.9	0.0
5.0	15	41.515	0.000	0.011	0.018	11.1	0.127	31.936	0.003	0.011	0.018	6.6	15.5	0.031	1.6	0.0
5.0	20	52.582	0.000	0.011	0.018	11.1	0.131	38.552	0.003	0.012	0.018	7.4	16.8	0.046	2.4	0.0

		Forward					
c	p	mse_η	mse_{σ_b}	mse_{σ_ϵ}	Time	FalsePos	FalseNeg
0.5	5.0	17.906	0.012	0.022	1.061	0.0	0.0
0.5	10.0	24.354	0.012	0.023	3.559	1.0	0.0
0.5	15.0	26.470	0.013	0.023	6.051	1.0	0.0
0.5	20.0	28.035	0.012	0.023	8.427	1.0	0.0
1.0	5.0	18.852	0.009	0.014	1.092	0.0	0.0
1.0	10.0	25.735	0.009	0.014	3.656	1.0	0.0
1.0	15.0	28.345	0.009	0.014	6.268	1.0	0.0
1.0	20.0	29.621	0.009	0.015	8.835	1.0	0.0
5.0	5.0	19.249	0.010	0.018	1.028	0.0	0.0
5.0	10.0	25.667	0.010	0.018	3.453	1.0	0.0
5.0	15.0	27.608	0.010	0.018	5.922	1.0	0.0
5.0	20.0	29.297	0.011	0.018	8.312	1.0	0.0

Table C.11: Study 12

		MM						BoostMixed								
c	p	mse_η	mse_β	mse_{σ_b}	mse_{σ_ϵ}	Steps	Time	mse_η	mse_β	mse_{σ_b}	mse_{σ_ϵ}	Selected	Steps	Time	FalsePos	FalseNeg
0.5	5	35.544	0.000	0.020	0.130	8.0	0.086	35.550	0.002	0.020	0.131	5.0	11.3	0.037	0.0	0.0
0.5	10	52.303	0.000	0.019	0.129	8.0	0.095	45.491	0.003	0.021	0.129	6.0	12.7	0.039	1.0	0.0
0.5	15	71.293	0.000	0.020	0.129	8.0	0.099	56.130	0.004	0.022	0.131	6.8	13.9	0.042	1.8	0.0
0.5	20	92.994	0.000	0.019	0.130	8.0	0.107	69.648	0.005	0.023	0.131	7.7	15.3	0.079	2.7	0.0
1.0	5	32.546	0.000	0.023	0.121	9.0	0.098	32.550	0.002	0.022	0.119	5.0	11.8	0.021	0.0	0.0
1.0	10	50.598	0.000	0.023	0.121	9.0	0.102	44.052	0.003	0.023	0.119	6.0	13.2	0.081	1.0	0.0
1.0	15	70.189	0.000	0.024	0.122	9.0	0.110	55.896	0.004	0.024	0.119	6.9	14.3	0.049	1.9	0.0
1.0	20	89.646	0.000	0.023	0.125	9.0	0.117	68.165	0.006	0.024	0.121	7.9	15.8	0.063	2.9	0.0
5.0	5	36.670	0.000	0.022	0.150	11.0	0.110	36.668	0.002	0.021	0.149	5.0	11.9	0.017	0.0	0.0
5.0	10	55.584	0.000	0.022	0.151	10.9	0.115	48.853	0.003	0.022	0.150	6.1	13.8	0.029	1.1	0.0
5.0	15	73.733	0.000	0.022	0.151	11.0	0.123	59.184	0.004	0.022	0.150	6.9	15.7	0.046	1.9	0.0
5.0	20	91.585	0.000	0.022	0.151	11.0	0.130	69.471	0.005	0.023	0.150	7.8	16.3	0.073	2.8	0.0

		Forward					
c	p	mse_η	mse_{σ_b}	mse_{σ_ϵ}	Time	FalsePos	FalseNeg
0.5	5.0	35.544	0.020	0.130	1.054	0.0	0.0
0.5	10.0	44.564	0.020	0.129	3.512	1.0	0.0
0.5	15.0	49.073	0.020	0.130	5.960	1.0	0.0
0.5	20.0	52.238	0.020	0.131	8.308	1.0	0.0
1.0	5.0	32.546	0.023	0.121	1.084	0.0	0.0
1.0	10.0	42.060	0.023	0.121	3.615	1.0	0.0
1.0	15.0	47.380	0.023	0.121	6.232	1.0	0.0
1.0	20.0	50.036	0.023	0.122	8.804	1.0	0.0
5.0	5.0	36.670	0.022	0.150	1.026	0.0	0.0
5.0	10.0	46.972	0.022	0.150	3.446	1.0	0.0
5.0	15.0	50.664	0.021	0.149	5.902	1.0	0.0
5.0	20.0	52.276	0.021	0.150	8.284	1.0	0.0

Table C.12: Study 13

		MM						BoostMixed								
c	p	mse_η	mse_β	mse_{σ_b}	mse_{σ_ϵ}	Steps	Time	mse_η	mse_β	mse_{σ_b}	mse_{σ_ϵ}	Selected	Steps	Time	FalsePos	FalseNeg
0.5	5	38.899	0.001	0.052	0.324	8.0	0.021	38.905	0.005	0.054	0.308	5.0	14.8	0.006	0.0	0.0
0.5	10	54.340	0.001	0.051	0.322	8.0	0.021	47.435	0.007	0.055	0.308	5.9	16.2	0.010	0.9	0.0
0.5	15	74.793	0.001	0.052	0.328	8.0	0.023	59.391	0.010	0.060	0.312	6.9	17.9	0.016	1.9	0.0
0.5	20	92.172	0.001	0.052	0.331	8.0	0.024	67.828	0.012	0.065	0.310	7.5	19.3	0.022	2.5	0.0
1.0	5	36.129	0.001	0.041	0.261	9.0	0.023	36.098	0.004	0.040	0.260	5.0	14.8	0.006	0.0	0.0
1.0	10	53.242	0.001	0.043	0.267	9.0	0.026	45.719	0.007	0.042	0.262	6.0	16.1	0.011	1.0	0.0
1.0	15	70.668	0.001	0.045	0.266	9.0	0.028	55.420	0.009	0.045	0.256	6.8	17.4	0.017	1.8	0.0
1.0	20	87.263	0.001	0.047	0.268	9.0	0.031	63.323	0.011	0.049	0.259	7.5	18.9	0.024	2.5	0.0
5.0	5	39.595	0.001	0.039	0.255	11.0	0.029	39.596	0.005	0.042	0.248	5.0	16.9	0.008	0.0	0.0
5.0	10	57.738	0.001	0.040	0.253	11.0	0.032	49.302	0.008	0.046	0.248	5.9	18.4	0.013	0.9	0.0
5.0	15	74.932	0.001	0.041	0.252	11.0	0.034	58.447	0.010	0.049	0.246	6.7	19.9	0.020	1.7	0.0
5.0	20	94.285	0.001	0.043	0.260	11.0	0.038	68.972	0.013	0.057	0.251	7.6	21.4	0.027	2.6	0.0

		Forward					
c	p	mse_η	mse_{σ_b}	mse_{σ_ϵ}	Time	FalsePos	FalseNeg
0.5	5.0	38.899	0.052	0.324	0.236	0.0	0.0
0.5	10.0	47.058	0.051	0.325	0.783	1.0	0.0
0.5	15.0	52.103	0.053	0.325	1.335	1.0	0.0
0.5	20.0	54.950	0.054	0.326	1.885	1.0	0.0
1.0	5.0	36.129	0.041	0.261	0.258	0.0	0.0
1.0	10.0	44.834	0.041	0.264	0.858	1.0	0.0
1.0	15.0	48.623	0.041	0.258	1.455	1.0	0.0
1.0	20.0	50.367	0.041	0.261	2.045	1.0	0.0
5.0	5.0	39.595	0.039	0.255	0.235	0.0	0.0
5.0	10.0	48.836	0.040	0.252	0.791	1.0	0.0
5.0	15.0	52.163	0.041	0.252	1.340	1.0	0.0
5.0	20.0	54.772	0.042	0.254	1.879	1.0	0.0

Table C.13: Study 14

C.4 Boosted GAMM - Poisson

		GAMM			bGAMM						Reference		
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg	mse_η	mse_{σ_b}	notconv
0.5	5	17.922	0.013	17	21.952	0.013	0	428.1	0.00	0.87	78.807	1.516	0
0.5	10	22.258	0.016	63	30.760	0.010	3	237.5	1.89	1.54	78.807	1.516	0
0.5	15	29.117	0.013	87	33.894	0.003	1	247.8	3.17	1.67	78.807	1.516	0
0.5	20				44.925	0.012	1	351.2	4.59	1.52	78.807	1.516	0
0.7	5	18.312	0.011	4	19.536	0.013	3	353.3	0.00	0.28	156.869	2.233	0
0.7	10	22.804	0.013	52	28.404	0.010	1	365.3	2.06	0.44	156.869	2.233	0
0.7	15	31.009	0.012	90	35.508	0.011	5	263.4	2.89	0.67	156.869	2.233	0
0.7	20				45.287	0.010	3	310.2	4.23	0.81	156.869	2.233	0
1.0	5	19.699	0.017	1	24.235	0.009	6	280.9	0.00	0.09	344.090	4.480	0
1.0	10	25.488	0.023	69	39.790	0.009	6	325.0	2.48	0.23	344.090	4.480	0
1.0	15	31.870	0.011	84	60.806	0.040	10	266.9	4.21	0.50	344.090	4.480	0
1.0	20				62.585	0.016	7	285.8	5.54	0.42	344.090	4.480	0

Table C.14: Study 15 - AIC

		GAMM			bGAMM						Reference		
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg	mse_η	mse_{σ_b}	notconv
0.5	5	17.922	0.013	17	27.936	0.007	0	86.8	0.00	1.78	78.742	1.516	0
0.5	10	22.258	0.016	63	34.864	0.006	0	42.6	1.03	2.32	78.742	1.516	0
0.5	15	29.117	0.013	87	36.501	0.002	0	81.8	1.54	2.46	78.742	1.516	0
0.5	20				39.921	0.006	0	71.8	1.82	2.49	78.742	1.516	0
0.7	5	18.312	0.011	4	27.485	0.010	0	112.7	0.00	0.72	155.711	2.199	0
0.7	10	22.804	0.013	52	31.346	0.008	0	121.3	1.13	0.96	155.711	2.199	0
0.7	15	31.009	0.012	90	41.412	0.022	0	76.7	1.60	1.30	155.711	2.199	0
0.7	20				45.684	0.009	0	91.3	1.92	1.50	155.711	2.199	0
1.0	5	19.699	0.017	1	29.894	0.007	0	136.3	0.00	0.27	342.556	4.417	0
1.0	10	25.488	0.023	69	49.314	0.009	0	111.5	1.45	0.58	342.556	4.417	0
1.0	15	31.870	0.011	84	51.388	0.018	0	108.1	2.25	0.75	342.556	4.417	0
1.0	20				58.984	0.011	0	124.4	2.97	0.73	342.556	4.417	0

Table C.15: Study 15 - BIC

		GAMM			bGAMM						Reference		
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg	mse_η	mse_{σ_b}	notconv
0.5	5	21.392	0.301	16	22.149	0.027	12	252.0	0.00	0.71	87.632	1.729	0
0.5	10	28.089	0.428	77	31.072	0.017	8	226.2	2.09	0.82	87.632	1.729	0
0.5	15	29.445	0.512	88	46.032	0.027	6	220.0	4.33	0.92	87.632	1.729	0
0.5	20				49.440	0.025	7	229.1	4.81	1.43	87.632	1.729	0
0.7	5	19.956	0.386	5	21.600	0.054	12	305.0	0.00	0.18	170.181	2.630	0
0.7	10	26.556	0.512	75	32.915	0.020	7	261.1	2.20	0.24	170.181	2.630	0
0.7	15	32.136	0.442	83	43.682	0.028	6	300.1	3.81	0.50	170.181	2.630	0
0.7	20				53.041	0.036	5	292.5	5.24	0.63	170.181	2.630	0
1.0	5	17.939	0.230	0	17.107	0.042	24	407.4	0.00	0.00	386.764	7.731	0
1.0	10	20.273	0.313	67	21.215	0.021	24	418.3	2.46	0.00	386.764	7.731	0
1.0	15	21.729	0.391	96	29.423	0.074	18	292.8	5.50	0.00	386.764	7.731	0
1.0	20				38.480	0.039	13	47.2	9.01	0.00	386.764	7.731	0

Table C.16: Study 16 - AIC

		GAMM			bGAMM						Reference		
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg	mse_η	mse_{σ_b}	notconv
0.5	5	21.392	0.301	16	27.438	0.028	2	80.0	0.00	1.37	87.632	1.729	0
0.5	10	28.089	0.428	77	33.513	0.021	1	75.0	1.22	1.70	87.632	1.729	0
0.5	15	29.445	0.512	88	35.072	0.050	1	43.3	1.67	1.75	87.632	1.729	0
0.5	20				42.534	0.019	1	53.5	2.15	2.30	87.632	1.729	0
0.7	5	19.956	0.386	5	25.978	0.042	1	99.0	0.00	0.54	170.181	2.630	0
0.7	10	26.556	0.512	75	35.349	0.037	1	120.4	1.24	0.88	170.181	2.630	0
0.7	15	32.136	0.442	83	43.612	0.033	1	114.1	1.94	1.35	170.181	2.630	0
0.7	20				48.953	0.040	2	84.3	2.55	1.31	170.181	2.630	0
1.0	5	17.939	0.230	0	15.600	0.038	12	163.1	0.00	0.00	386.764	7.731	0
1.0	10	20.273	0.313	67	17.748	0.022	13	175.2	1.63	0.00	386.764	7.731	0
1.0	15	21.729	0.391	96	22.958	0.078	12	165.0	4.00	0.00	386.764	7.731	0
1.0	20				24.512	0.040	13	208.5	3.89	0.00	386.764	7.731	0

Table C.17: Study 16 - BIC

		GAMM			bGAMM						Reference		
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg	mse_η	mse_{σ_b}	notconv
0.5	5	17.519	0.049	3	15.776	0.027	0	68.8	0.00	0.00	76.730	1.813	0
0.5	10	24.663	0.021	72	26.474	0.020	0	44.3	4.21	0.07	76.730	1.813	0
0.5	15	22.629	0.020	86	28.097	0.037	0	31.3	6.79	0.00	76.730	1.813	0
0.5	20				31.836	0.023	0	23.5	7.50	0.20	76.730	1.813	0
0.7	5	15.272	0.041	0	14.470	0.014	1	97.2	0.00	0.00	160.487	3.107	0
0.7	10	14.201	0.015	73	16.964	0.014	1	82.4	4.59	0.00	160.487	3.107	0
0.7	15	23.223	0.026	88	29.946	0.010	1	58.8	8.25	0.00	160.487	3.107	0
0.7	20				35.095	0.013	1	41.3	9.42	0.02	160.487	3.107	0
1.0	5	16.079	0.043	0	14.995	0.024	0	67.9	0.00	0.00	356.200	7.012	0
1.0	10	18.816	0.022	63	21.749	0.028	0	53.3	4.70	0.00	356.200	7.012	0
1.0	15	17.332	0.004	93	23.969	0.003	0	36.4	6.43	0.00	356.200	7.012	0
1.0	20				31.237	0.021	0	36.3	7.92	0.00	356.200	7.012	0

Table C.18: Study 17 - AIC

		GAMM			bGAMM						Reference		
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg	mse_η	mse_{σ_b}	notconv
0.5	5	17.519	0.049	3	16.277	0.028	0	99.8	0.00	0.20	76.730	1.813	0
0.5	10	24.663	0.021	72	22.178	0.022	0	93.1	1.29	0.29	76.730	1.813	0
0.5	15	22.629	0.020	86	19.210	0.038	0	75.3	2.50	0.21	76.730	1.813	0
0.5	20				23.106	0.023	0	109.6	2.97	0.39	76.730	1.813	0
0.7	5	15.272	0.041	0	14.480	0.014	1	130.5	0.00	0.01	160.487	3.107	0
0.7	10	14.201	0.015	73	14.367	0.012	1	158.1	1.56	0.00	160.487	3.107	0
0.7	15	23.223	0.026	88	24.335	0.010	1	178.1	2.83	0.00	160.487	3.107	0
0.7	20				21.650	0.014	1	139.9	3.35	0.05	160.487	3.107	0
1.0	5	16.079	0.043	0	14.695	0.023	0	184.5	0.00	0.00	356.200	7.012	0
1.0	10	18.816	0.022	63	18.487	0.032	0	190.1	1.70	0.00	356.200	7.012	0
1.0	15	17.332	0.004	93	15.105	0.005	0	179.3	2.57	0.00	356.200	7.012	0
1.0	20				21.838	0.017	0	179.5	3.80	0.00	356.200	7.012	0

Table C.19: Study 17 - BIC

		GAMM			bGAMM						Reference		
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg	mse_η	mse_{σ_b}	notconv
0.5	5	15.200	0.042	0	13.935	0.009	0	44.0	0.00	0.00	152.314	1.749	0
0.5	10	18.149	0.030	62	21.544	0.010	0	39.5	4.37	0.00	152.314	1.749	0
0.5	15	21.046	0.025	89	27.093	0.008	0	14.0	4.18	0.00	152.314	1.749	0
0.5	20				28.871	0.011	0	14.5	4.55	0.08	152.314	1.749	0
0.7	5	14.347	0.038	0	13.339	0.006	0	29.0	0.00	0.00	314.264	2.788	0
0.7	10	17.605	0.030	46	21.334	0.007	0	23.4	3.81	0.00	314.264	2.788	0
0.7	15	20.467	0.010	91	26.793	0.012	0	18.0	4.33	0.00	314.264	2.788	0
0.7	20				29.873	0.006	0	19.4	5.05	0.00	314.264	2.788	0
1.0	5	14.625	0.024	0	13.316	0.010	0	69.2	0.00	0.00	704.219	5.822	0
1.0	10	17.759	0.015	48	16.007	0.009	0	67.4	1.04	0.00	704.219	5.822	0
1.0	15	19.041	0.009	91	16.132	0.003	0	76.9	1.22	0.00	704.219	5.822	0
1.0	20				15.915	0.010	0	71.8	1.30	0.00	704.219	5.822	0

Table C.20: Study 18 - AIC

		GAMM			bGAMM						Reference		
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg	mse_η	mse_{σ_b}	notconv
0.5	5	15.200	0.042	0	14.083	0.009	0	46.4	0.00	0.08	152.314	1.749	0
0.5	10	18.149	0.030	62	15.573	0.008	0	52.8	0.92	0.11	152.314	1.749	0
0.5	15	21.046	0.025	89	17.631	0.007	0	45.5	1.27	0.27	152.314	1.749	0
0.5	20				18.380	0.009	0	50.0	1.60	0.13	152.314	1.749	0
0.7	5	14.347	0.038	0	12.962	0.006	0	61.4	0.00	0.00	314.264	2.788	0
0.7	10	17.605	0.030	46	14.587	0.007	0	69.8	1.04	0.00	314.264	2.788	0
0.7	15	20.467	0.010	91	16.040	0.011	0	71.2	1.33	0.00	314.264	2.788	0
0.7	20				16.577	0.006	0	67.0	1.73	0.00	314.264	2.788	0
1.0	5	14.625	0.024	0	13.132	0.010	0	85.7	0.00	0.00	704.219	5.822	0
1.0	10	17.759	0.015	48	16.043	0.009	0	86.1	1.15	0.00	704.219	5.822	0
1.0	15	19.041	0.009	91	15.228	0.003	0	104.6	1.33	0.00	704.219	5.822	0
1.0	20				16.152	0.010	0	89.0	1.50	0.00	704.219	5.822	0

Table C.21: Study 18 - BIC

		GAMM			bGAMM						Reference		
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg	mse_η	mse_{σ_b}	notconv
0.5	5	15.820	0.178	0	14.894	0.029	4	48.6	0.00	0.00	159.323	1.952	0
0.5	10	17.842	0.243	75	21.295	0.033	4	39.8	4.29	0.00	159.323	1.952	0
0.5	15	22.977	0.300	95	28.048	0.029	3	19.8	4.60	0.00	159.323	1.952	0
0.5	20				28.214	0.028	4	20.5	4.66	0.01	159.323	1.952	0
0.7	5	17.314	0.192	0	16.817	0.025	3	42.7	0.00	0.00	340.879	3.386	0
0.7	10	17.588	0.232	61	23.918	0.035	2	34.3	3.95	0.00	340.879	3.386	0
0.7	15	27.360	0.317	96	34.503	0.148	2	27.8	5.75	0.00	340.879	3.386	0
0.7	20				19.385	0.025	3	65.5	1.47	0.01	340.879	3.386	0
1.0	5	16.982	0.172	0	16.250	0.031	13	111.5	0.00	0.00	766.881	7.406	0
1.0	10	20.017	0.211	58	18.278	0.019	12	89.0	0.82	0.00	766.881	7.406	0
1.0	15	21.525	0.201	94	18.098	0.058	14	65.5	1.00	0.00	766.881	7.406	0
1.0	20				19.415	0.034	13	100.7	1.66	0.00	766.881	7.406	0

Table C.22: Study 19 - AIC

		GAMM			bGAMM						Reference		
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg	mse_η	mse_{σ_b}	notconv
0.5	5	15.820	0.178	0	15.112	0.029	4	56.6	0.00	0.05	159.530	1.952	0
0.5	10	17.842	0.243	75	15.997	0.031	4	62.4	0.83	0.00	159.530	1.952	0
0.5	15	22.977	0.300	95	20.468	0.028	4	52.8	1.00	0.00	159.530	1.952	0
0.5	20				18.542	0.027	4	61.4	1.42	0.05	159.530	1.952	0
0.7	5	17.314	0.192	0	16.201	0.025	2	76.6	0.00	0.00	342.913	3.398	0
0.7	10	17.588	0.232	61	16.462	0.031	2	77.6	0.92	0.00	342.913	3.398	0
0.7	15	27.360	0.317	96	24.600	0.149	1	74.0	2.00	0.00	342.913	3.398	0
0.7	20				19.366	0.024	2	78.6	1.63	0.00	342.913	3.398	0
1.0	5	16.982	0.172	0	16.092	0.035	12	123.8	0.00	0.00	768.000	7.463	0
1.0	10	20.017	0.211	58	18.305	0.021	14	106.8	0.85	0.00	768.000	7.463	0
1.0	15	21.525	0.201	94	18.019	0.059	13	87.3	1.00	0.00	768.000	7.463	0
1.0	20				20.418	0.034	12	114.1	1.92	0.00	768.000	7.463	0

Table C.23: Study 19 - BIC

		GAMM			bGAMM						Reference		
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg	mse_η	mse_{σ_b}	notconv
0.5	5	18.272	0.061	0	16.751	0.009	0	44.1	0.00	0.00	150.653	1.819	0
0.5	10	18.791	0.053	59	21.543	0.008	0	37.6	4.49	0.00	150.653	1.819	0
0.5	15	21.249	0.035	91	28.941	0.008	0	27.9	6.44	0.00	150.653	1.819	0
0.5	20				29.053	0.008	0	15.9	4.74	0.03	150.653	1.819	0
0.7	5	16.751	0.051	0	15.742	0.010	0	30.7	0.00	0.00	308.952	3.025	0
0.7	10	19.534	0.037	50	23.283	0.011	0	27.5	4.14	0.00	308.952	3.025	0
0.7	15	23.794	0.064	94	30.113	0.009	0	16.8	3.83	0.00	308.952	3.025	0
0.7	20				28.612	0.010	0	19.8	4.64	0.00	308.952	3.025	0
1.0	5	14.422	0.061	0	13.952	0.013	1	44.4	0.00	0.00	697.360	6.863	0
1.0	10	16.749	0.033	49	21.135	0.011	1	34.1	3.96	0.00	697.360	6.863	0
1.0	15	14.982	0.007	96	19.403	0.048	3	25.0	5.00	0.00	697.360	6.863	0
1.0	20				28.838	0.012	3	28.9	5.33	0.00	697.360	6.863	0

Table C.24: Study 20 - AIC

		GAMM			bGAMM						Reference		
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg	mse_η	mse_{σ_b}	notconv
0.5	5	18.272	0.061	0	16.798	0.009	0	58.0	0.00	0.04	150.653	1.819	0
0.5	10	18.791	0.053	59	16.984	0.008	0	57.2	1.24	0.00	150.653	1.819	0
0.5	15	21.249	0.035	91	17.373	0.009	0	52.4	1.44	0.00	150.653	1.819	0
0.5	20				21.481	0.009	0	57.7	1.97	0.08	150.653	1.819	0
0.7	5	16.751	0.051	0	15.258	0.010	1	67.4	0.00	0.00	308.952	3.025	0
0.7	10	19.534	0.037	50	17.681	0.012	1	71.9	1.24	0.00	308.952	3.025	0
0.7	15	23.794	0.064	94	17.305	0.010	0	56.5	1.33	0.00	308.952	3.025	0
0.7	20				18.891	0.011	1	75.1	1.93	0.00	308.952	3.025	0
1.0	5	14.422	0.061	0	13.318	0.013	2	94.7	0.00	0.00	697.360	6.863	0
1.0	10	16.749	0.033	49	15.072	0.012	3	89.4	1.26	0.00	697.360	6.863	0
1.0	15	14.982	0.007	96	11.559	0.042	2	82.0	2.00	0.00	697.360	6.863	0
1.0	20				16.651	0.012	3	93.6	1.95	0.00	697.360	6.863	0

Table C.25: Study 20 - BIC

C.5 Boosted GLMM - Binomial

		GLMM			bGLMM					
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg
0.5	5	505.805	0.279	0	609.219	3.830	0	195.1	0.00	1.03
0.5	10	942.978	0.300	0	733.998	6.806	1	178.6	1.20	1.06
0.5	15	904.031	0.307	0	1091.575	12.036	0	157.9	2.35	1.08
0.5	20	652.706	0.324	0	833.776	0.304	0	122.3	3.62	1.05
0.7	5	161.824	0.163	0	158.951	0.272	0	172.1	0.00	0.01
0.7	10	288.621	0.203	0	252.622	0.324	1	173.9	1.04	0.01
0.7	15	630.741	0.215	0	328.352	0.365	1	166.5	2.08	0.03
0.7	20	713.179	0.249	0	401.971	0.967	1	143.3	3.48	0.01
1.0	5	883.756	0.267	0	430.694	0.346	1	196.9	0.00	0.03
1.0	10	1226.259	0.298	0	617.398	0.724	1	138.9	1.46	0.02
1.0	15	1479.220	0.326	0	808.421	1.958	1	123.1	3.31	0.02
1.0	20	2640.851	0.343	1	1102.831	4.932	2	114.8	5.08	0.02

Table C.26: Study 21 - AIC

		GLMM			bGLMM					
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg
0.5	5	505.805	0.279	0	610.542	2.433	0	200.8	0.00	1.44
0.5	10	942.978	0.300	0	650.625	3.250	0	166.6	0.42	1.44
0.5	15	904.031	0.307	0	673.830	3.366	0	155.9	0.69	1.45
0.5	20	652.706	0.324	0	830.624	5.782	1	139.8	1.07	1.46
0.7	5	161.824	0.163	0	163.677	0.282	0	164.1	0.00	0.04
0.7	10	288.621	0.203	0	235.837	0.308	1	172.7	0.55	0.03
0.7	15	630.741	0.215	0	277.287	0.349	1	131.6	1.08	0.03
0.7	20	713.179	0.249	0	342.036	0.349	1	133.4	1.70	0.03
1.0	5	883.756	0.267	0	430.694	0.351	1	220.2	0.00	0.07
1.0	10	1226.259	0.298	0	513.985	0.676	1	197.4	0.81	0.06
1.0	15	1479.220	0.326	0	639.968	1.897	0	127.5	2.03	0.05
1.0	20	2640.851	0.343	1	874.884	3.928	2	104.8	3.41	0.06

Table C.27: Study 21 - BIC

		GLMM			bGLMM					
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg
0.5	5	69.168	0.661	0	83.962	0.248	0	163.5	0.00	0.04
0.5	10	137.148	0.722	0	124.377	0.277	0	167.8	0.74	0.06
0.5	15	225.064	0.768	0	152.428	0.306	0	168.2	1.63	0.06
0.5	20	339.263	0.832	0	195.905	0.373	0	161.6	2.64	0.06
0.7	5	196.078	1.008	0	207.807	0.550	0	255.3	0.00	0.08
0.7	10	437.570	1.042	0	291.958	0.660	0	206.2	1.18	0.03
0.7	15	590.907	1.102	1	353.126	1.012	0	165.7	2.56	0.04
0.7	20	777.063	1.191	0	459.025	2.068	1	134.3	4.36	0.07
1.0	5	890.850	1.237	0	381.598	0.678	0	157.0	0.00	0.03
1.0	10	2177.524	1.291	0	495.011	1.110	0	147.1	1.33	0.03
1.0	15	3095.759	1.336	0	868.701	2.516	0	119.5	3.34	0.03
1.0	20	2829.335	1.387	0	1041.733	2.766	4	117.5	4.88	0.05

Table C.28: Study 22 - AIC

		GLMM			bGLMM					
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg
0.5	5	69.168	0.661	0	92.200	0.213	0	177.4	0.00	0.14
0.5	10	137.148	0.722	0	123.918	0.247	0	177.4	0.26	0.15
0.5	15	225.064	0.768	0	135.733	0.213	0	170.5	0.70	0.17
0.5	20	339.263	0.832	0	152.804	0.281	0	176.4	0.96	0.18
0.7	5	196.078	1.008	0	233.679	0.557	0	256.1	0.00	0.13
0.7	10	437.570	1.042	0	266.187	0.685	0	195.9	0.54	0.09
0.7	15	590.907	1.102	1	319.132	0.811	0	173.5	1.24	0.09
0.7	20	777.063	1.191	0	353.330	1.200	2	155.6	1.95	0.11
1.0	5	890.850	1.237	0	399.608	0.747	0	154.3	0.00	0.05
1.0	10	2177.524	1.291	0	447.823	1.122	0	134.8	0.82	0.04
1.0	15	3095.759	1.336	0	720.313	1.403	0	127.8	2.00	0.05
1.0	20	2829.335	1.387	0	939.573	3.226	1	109.0	3.17	0.04

Table C.29: Study 22 - BIC

		GLMM			bGLMM					
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg
0.5	5	75.410	0.014	0	76.611	0.070	0	113.0	0.00	0.00
0.5	10	121.737	0.018	0	101.686	0.067	0	97.9	0.72	0.00
0.5	15	173.560	0.026	0	119.413	0.067	0	124.2	1.68	0.01
0.5	20	254.561	0.027	0	145.157	0.082	0	111.4	2.73	0.01
0.7	5	103.782	0.049	0	140.968	0.146	0	163.9	0.00	0.02
0.7	10	193.702	0.064	0	183.950	0.148	0	178.1	0.99	0.02
0.7	15	299.595	0.087	0	217.021	0.154	0	161.2	2.01	0.01
0.7	20	463.942	0.106	0	275.414	0.165	0	137.0	3.57	0.00
1.0	5	260.046	0.188	0	277.874	0.202	0	225.2	0.00	0.04
1.0	10	606.771	0.215	0	366.292	0.235	0	153.1	1.28	0.00
1.0	15	1170.769	0.241	0	487.435	0.385	0	135.8	3.13	0.02
1.0	20	2074.060	0.273	1	661.705	0.562	0	104.8	5.00	0.00

Table C.30: Study 23 - AIC

		GLMM			bGLMM					
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg
0.5	5	75.410	0.014	0	79.368	0.065	0	101.0	0.00	0.04
0.5	10	121.737	0.018	0	87.911	0.070	0	99.0	0.29	0.04
0.5	15	173.560	0.026	0	109.611	0.088	0	111.8	0.65	0.05
0.5	20	254.561	0.027	0	117.490	0.087	0	110.0	1.10	0.05
0.7	5	103.782	0.049	0	137.540	0.145	0	163.5	0.00	0.02
0.7	10	193.702	0.064	0	159.741	0.154	0	146.2	0.50	0.02
0.7	15	299.595	0.087	0	190.079	0.153	0	151.2	1.01	0.03
0.7	20	463.942	0.106	0	229.323	0.160	0	118.1	1.74	0.00
1.0	5	260.046	0.188	0	277.252	0.188	0	213.4	0.00	0.04
1.0	10	606.771	0.215	0	332.720	0.236	0	184.2	0.78	0.00
1.0	15	1170.769	0.241	0	475.836	0.270	0	137.5	2.05	0.01
1.0	20	2074.060	0.273	1	533.623	0.344	1	134.2	3.43	0.00

Table C.31: Study 23 - BIC

		GLMM			bGLMM					
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg
0.5	5	72.003	0.031	0	90.400	0.058	0	101.2	0.00	0.00
0.5	10	146.845	0.038	0	120.603	0.058	0	116.6	0.87	0.00
0.5	15	210.027	0.058	0	151.121	0.075	0	115.3	1.51	0.00
0.5	20	283.818	0.071	0	187.371	0.074	0	138.1	2.37	0.00
0.7	5	141.793	0.142	0	146.673	0.126	0	124.4	0.00	0.01
0.7	10	279.357	0.161	0	221.064	0.159	0	114.6	0.76	0.01
0.7	15	416.436	0.165	0	268.429	0.207	0	95.2	1.66	0.01
0.7	20	696.907	0.187	0	304.783	0.283	0	111.9	2.72	0.01
1.0	5	673.332	0.256	0	534.183	0.325	1	134.9	0.00	0.01
1.0	10	1906.076	0.251	0	537.176	0.335	1	138.7	1.32	0.01
1.0	15	3563.036	0.277	0	679.513	0.832	0	130.6	2.77	0.02
1.0	20	4198.591	0.301	0	1056.651	1.328	0	111.4	4.48	0.01

Table C.32: Study 24 - AIC

		GLMM			bGLMM					
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg
0.5	5	72.003	0.031	0	87.475	0.056	0	100.5	0.00	0.02
0.5	10	146.845	0.038	0	102.632	0.066	0	102.6	0.23	0.02
0.5	15	210.027	0.058	0	119.176	0.059	0	98.2	0.36	0.02
0.5	20	283.818	0.071	0	123.898	0.077	0	106.2	0.55	0.02
0.7	5	141.793	0.142	0	141.322	0.123	0	123.5	0.00	0.01
0.7	10	279.357	0.161	0	170.764	0.152	0	112.5	0.25	0.01
0.7	15	416.436	0.165	0	220.872	0.161	0	106.4	0.57	0.01
0.7	20	696.907	0.187	0	244.113	0.161	0	120.7	0.83	0.01
1.0	5	673.332	0.256	0	532.380	0.336	1	128.6	0.00	0.02
1.0	10	1906.076	0.251	0	535.680	0.353	0	114.1	0.64	0.02
1.0	15	3563.036	0.277	0	636.291	0.504	0	105.7	1.49	0.02
1.0	20	4198.591	0.301	0	698.534	0.509	0	139.6	2.88	0.2

Table C.33: Study 24 - BIC

		GLMM			bGLMM					
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg
0.5	5	75.597	0.663	0	89.526	0.217	0	109.8	0.00	0.01
0.5	10	134.710	0.688	0	109.934	0.198	0	113.3	0.67	0.01
0.5	15	200.426	0.694	0	147.711	0.214	0	116.4	1.41	0.01
0.5	20	309.023	0.759	0	176.625	0.231	0	150.9	2.27	0.01
0.7	5	162.585	0.921	0	215.827	0.408	0	198.2	0.00	0.02
0.7	10	279.924	0.952	0	244.586	0.499	0	185.7	0.86	0.02
0.7	15	415.746	0.979	0	286.309	0.526	0	168.7	1.77	0.00
0.7	20	603.542	1.027	0	328.878	0.572	0	164.2	2.84	0.00
1.0	5	996.046	1.203	0	562.277	0.689	0	187.0	0.00	0.00
1.0	10	1944.526	1.244	0	589.615	0.672	0	116.0	1.43	0.00
1.0	15	4028.865	1.259	0	922.077	1.006	2	105.0	3.14	0.00
1.0	20	4950.561	1.299	0	1045.734	1.493	2	106.1	4.43	0.02

Table C.34: Study 25 - AIC

		GLMM			bGLMM					
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg
0.5	5	75.597	0.663	0	94.786	0.215	0	110.4	0.00	0.05
0.5	10	134.710	0.688	0	107.229	0.199	0	113.6	0.20	0.05
0.5	15	200.426	0.694	0	117.090	0.219	0	113.3	0.41	0.05
0.5	20	309.023	0.759	0	134.776	0.205	0	122.3	0.62	0.05
0.7	5	162.585	0.921	0	205.430	0.409	0	199.9	0.00	0.03
0.7	10	279.924	0.952	0	227.562	0.447	0	209.3	0.35	0.01
0.7	15	415.746	0.979	0	248.913	0.521	0	206.7	0.65	0.01
0.7	20	603.542	1.027	0	255.406	0.521	0	187.0	1.04	0.01
1.0	5	996.046	1.203	0	549.455	0.699	0	178.5	0.00	0.00
1.0	10	1944.526	1.244	0	566.077	0.691	0	127.8	0.85	0.00
1.0	15	4028.865	1.259	0	836.432	0.729	2	123.1	1.88	0.00
1.0	20	4950.561	1.299	0	838.410	1.135	1	103.9	2.79	0.00

Table C.35: Study 25 - BIC

		GLMM			bGLMM					
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg
0.5	5	61.184	0.011	0	80.679	0.064	0	60.9	0.00	0.00
0.5	10	115.029	0.009	0	103.594	0.067	0	71.2	0.67	0.00
0.5	15	161.572	0.013	0	118.355	0.067	0	71.8	1.30	0.00
0.5	20	219.242	0.012	0	135.897	0.062	0	63.7	2.01	0.00
0.7	5	102.948	0.067	0	121.804	0.071	0	106.5	0.00	0.00
0.7	10	197.693	0.071	0	194.875	0.079	0	118.3	0.76	0.00
0.7	15	299.296	0.081	0	207.587	0.067	0	123.0	1.53	0.00
0.7	20	391.835	0.088	0	257.841	0.087	0	129.8	2.49	0.02
1.0	5	220.248	0.156	0	250.653	0.156	0	182.7	0.00	0.01
1.0	10	482.936	0.169	0	325.853	0.192	0	160.6	1.17	0.00
1.0	15	858.955	0.188	0	412.780	0.221	0	110.3	2.56	0.00
1.0	20	1116.216	0.203	0	557.111	0.274	0	117.0	4.36	0.00

Table C.36: Study 26 - AIC

		GLMM			bGLMM					
c	p	mse_η	mse_{σ_b}	notconv	mse_η	mse_{σ_b}	notconv	Steps	falsepos	falseneg
0.5	5	61.184	0.011	0	80.679	0.064	0	61.2	0.00	0.00
0.5	10	115.029	0.009	0	90.256	0.068	0	63.1	0.25	0.00
0.5	15	161.572	0.013	0	92.554	0.068	0	64.0	0.41	0.00
0.5	20	219.242	0.012	0	99.579	0.068	0	65.2	0.62	0.00
0.7	5	102.948	0.067	0	136.216	0.075	0	100.6	0.00	0.00
0.7	10	197.693	0.071	0	175.618	0.071	0	85.9	0.27	0.00
0.7	15	299.296	0.081	0	194.787	0.081	0	107.1	0.59	0.00
0.7	20	391.835	0.088	0	210.871	0.092	0	127.1	0.89	0.00
1.0	5	220.248	0.156	0	252.390	0.160	0	196.0	0.00	0.01
1.0	10	482.936	0.169	0	285.406	0.138	0	142.7	0.60	0.00
1.0	15	858.955	0.188	0	316.506	0.178	0	143.0	1.17	0.00
1.0	20	1116.216	0.203	0	361.343	0.258	0	133.6	1.97	0.00

Table C.37: Study 26 - BIC

Bibliography

ABRAMOWITZ, M. AND STEGUN, I. (1972). *Handbook of Mathematical Functions*. New York: Dover.

ADAMS, R. (1975). *Sobolev spaces*. Academic Press.

AITKIN, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics 55*, 117–128.

AITKIN, M. AND FRANCIS, B. J. (1998). Fitting generalized linear variance component models by nonparametric maximum likelihood. *The GLIM Newsletter 29* (in press).

ALT, H. W. (1985). *Lineare Funktionalanalysis*. Springer.

AMEMIYA, T. (1981). Qualitative response models: A survey. *Journal of Economic Literature XIX*, 1483–1536.

ANDERSON, D. A. AND AITKIN, M. (1985). Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society Ser. B 47*, 203–210.

ANTONOV, I. AND SALEEV, V. (1979). An economic method of computing LP_τ-sequences. *USSR computational mathematics and mathematical physics 19*, 252–256.

ARMSTRONG, B. AND SLOAN, M. (1989). Ordinal regression models for epidemiologic data. *American Journal of Epidemiology 129*, 191–204.

BOCK, R. D. AND AITKIN, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika 46*, 443–459.

BOOTH, J. G. AND HOBERT, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Statist. Soc B 61*, 265–285.

BRATLEY, P. AND FOX, B. L. (1988). Algorithm 659 Implementing Sobol's Quasirandom Sequence generator. *ACM Transactions on Mathematical Software 14*, 88–100.

BREIMAN, L. (1999). Prediction games and arcing algorithms. *Neural Computation 11*, 1493–1517.

BRESLOW, N. E. AND CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association 88*, 9–25.

BRESLOW, N. E. AND LIN, X. (1995a). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika 82*, 81–91.

BRESLOW, N. E. AND LIN, X. (1995b). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika 82*, 81–91.

BRUMBACK, B. A. AND RICE, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association 93*, 961–976.

BRUNNER, L., CHAN, A., JAMES, L., AND A.Y.LO (1996). Weighted chinese restaurant processes and bayesian mixture models. *Hong Kong University of Science and Technology Research Report 1*.

BRUNNER, L., CHAN, A., JAMES, L., AND A.Y.LO (2001). Weighted chinese restaurant processes and bayesian mixture models. *unpublishe manuskript*.

BÜHLMANN, P. AND YU, B. (2003). Boosting with l2 loss: Regression and classification. *Journal of the American Statistical Association 98*, 324–339.

BÜHLMANN, P. AND YU, B. (2005). Sparse boosting. *Journal of Machine Learning Research 7*, 1001–1024.

CALFLISCH, R. (1998). Monte carlo and quasi-Monte-Carlo Methods. *Acta Numerica 7*, 1998.

CHAN, K. AND KUK, A. (1997). Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics 53*, 86–97.

CHEN, J., ZHANG, D., AND DAVIDIAN, M. (2002). A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics 3*, 447–360.

CHIOU, J.-M., MÜLLER, H.-G., AND WANG, J.-L. (2003). Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society, Series B 65*, 405–423.

CORNISH, E. A. (1954). The multivariate t-distribution associated with a set of normal standard deviates. *Australian Journal of Physics 7*, 531–542.

CRAVEN, P. AND WAHBA, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik 31*, 377–403.

DAVIDIAN, M. AND GALLANT, A. (1993). The Nonlinear Mixed Effects Model with Smooth Random Effects Density. *Biometrika 80*, 475–488.

DAVIS, P. J. AND RABINOWITZ, P. (1975). *Numerical Integration*. Waltham, MA: Blaisdell.

DE BOOR, C. (1978). *A Practical Guide to Splines*. New York: Springer-Verlag.

DETTLING, M. AND BÜHLMANN, P. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics 19*, 1061–1069.

DEUFLHARD, P. AND HOHMANN, A. (1993). *Numerische Mathematik I*. Berlin: Walter de Gruyter.

DIERCKX, P. (1993). *Curve and Surface Fitting with Splines*. Clarendon Press: Oxford.

DRUM, M. L. AND MCCULLAGH, P. (1993). REML estimation with exact covariance in the logistic mixed model. *Biometrics 49*, 677–689.

EILERS, P., CURRIE, I., AND DURBAN, M. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society, Series B 68*, 259–280.

EILERS, P. H. C. AND MARX, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*, 89–121.

ENGEL, B. AND KEEN, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica 48*, 1–22.

EUBANK, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker.

FAHRMEIR, L., KNEIB, T., AND LANG, S. (2004). Penalized structured additive regression for space-time data: A bayesian perspective. *Statistica Sinica 14*, 731–761.

FAHRMEIR, L. AND LANG, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Applied Statistics 50*, 201–220.

FAHRMEIR, L. AND TUTZ, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models* (2nd ed.). New York: Springer.

FAN, J. AND GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman & Hall.

FERGUSON, T. (1973). A Bayesian Analysis of Some Nonparametric Problems. *Annals of Statistics 2*, 209–230.

FRIEDMAN, J. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist. 19*, 1–14.

FRIEDMAN, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics 29*, 337–407.

FRIEDMAN, J. AND SILVERMAN, B. (1989). Flexible parsimonious smoothing and additive modelling (with discussion). *Technometrics 31*, 3–39.

FRIEDMAN, J. H., HASTIE, T., AND TIBSHIRANI, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics 28*, 337–407.

GALLANT, A. AND NYCHKA, D. (1987). Seminonparametric maximum likelihood estimation. *Econometrica 55*, 363–390.

GASSER, T. AND MÜLLER, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist. 11*, 171–185.

GELFAND, A. E. AND CARLIN, B. (1993). Maximum Likelihood Estimation for Constrained- or Missing-Ddata Problems. *Canadian Journal of Statistics 21*, 303–311.

GERSTNER, T. AND GRIEBEL, M. (1998). Numerical integration using sparse grids. *Numer. Algorithms 18*, 209–232.

GERSTNER, T. AND GRIEBEL, M. (2003). Dimension - adaptive tensor-product quadrature. *Computer 71*, 65–87.

GEYER, C. AND THOMPSON, E. (1992). Constrained Monte Carlo Maximum Likelihood for Dependent Data. *Journal of the Royal Statistical Society, Ser. B 54*, 657–699.

GHIDEY, W., LESAFFRES, E., AND EILERS, P. (2004). Smooth Random Effects Distribution in a Linear Mixed Model. *Biometrics 60*, 945–953.

GILMOUR, A., ANDERSON, R., AND RAE, A. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika 72*, 593–599.

GILMOUR, A., THOMPSON, R., AND CULLIS, B. (1995). Average information reml: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics 51*, 1440–1450.

GOLDSTEIN, H. AND RASBASH, J. (1996). Improved approximation for multilevel models with binary responses. *Journal of the Royal Statistical Society,Series A 159*, 505–513.

GREEN, D. J. AND SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman & Hall.

GREEN, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review 55*, 245–259.

HÄMMERLIN, G. AND HOFFMANN, K. (1992). *Numerische Mathematik*. Springer Berlin.

HARVILLE, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika 61*, 383–385.

HARVILLE, D. A. (1976). Extension of the Gauss-Markov theorem to include the estimation of random effects. *Annals of Statistics 4*, 384–395.

HARVILLE, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association 72*, 320–338.

HARVILLE, D. A. AND MEE, R. W. (1984). A mixed-model procedure for analyzing ordered categorical data. *Biometrics 40*, 393–408.

HASTIE, T. AND LOADER, C. (1993). Local regression: Automatic kernel carpentry. *Statist. Sci. 8*, 120–143.

HASTIE, T. AND TIBSHIRANI, R. (1990). *Generalized Additive Models*. London: Chapman & Hall.

HEDEKER, D. AND GIBBONS, R. (1996). A computer programme for mixed-effects ordinal probit and logistic regression analysis. *Computer Methods and Programs in Biomedicine 49*, 157–176.

HEDEKER, D. AND GIBBONS, R. B. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics 50*, 933–944.

HENDERSON, C. R. (1953). Estimation of variance and covariance components. *Biometrics 9*, 226–252.

IM, S. AND GIANOLA, D. (1988). Mixed models for bionomial data with an application to lamb mortality. *Applied Statistics 37*, 196–204.

ISHWARAN, H. AND TAKAHARA, G. (2002). Independent and identically distributed monte carlo algorithms for semiparametric linear mixed models. *Journal of the American Statistical Association 97*, 1154–1166.

JAMES, G. AND SUGAR, C. (2003). Clustering sparsely sampled functional data. *Journal of the American Statistical Association 98*(462), 397–408.

JANK, W. AND SHMUELI, G. (2005). Profiling price dynamics in online auctions using curve clustering. Technical Report, Smith School of Business, University of Maryland.

JUDD, K. (1992). *Numerical methods in economics*. New York: MIT-Press.

KASLOW, R. A., OSTROW, D. G., DETELS, R., PHAIR, J. P., POLK, B. F., AND RINALDO, C. R. (1987). The multicenter aids cohort study: Rationale, organization and selected characteristic of the participiants. *American Journal of Epidemiology 126*, 310–318.

KAY, R. AND LITTLE, S. (1986). Assessing the fit of the logistic model: A case study of children with the Haemolytic Uraemic Syndrome. *Applied Statistics 35*, 16–30.

KNEIB, T. AND FAHRMEIR, L. (2006). Structured additive regression for categorical spaceŰtime data: A mixed model approach. *Biometrics 62*, 109–118.

KRISHNA, V. (2002). *Auction Theory*. Academic Press, San Diego.

LAIRD, N. M., LANGE, N., AND STRAM, D. (1987). Maximum likelihood computations with repeated measures: Application of the em-algorithm. *Journal of the American Statistical Association 82*, 97–105.

LAIRD, N. M. AND WARE, J. H. (1982). Random effects models for longitudinal data. *Biometrics 38*, 963–974.

LANG, S., ADEBAYO, S., FAHRMEIR, L., AND STEINER, W. (2003). Bayesian geoadditive seemingly unrelated regression. *Computational Statistics 18*, 263–292.

LANGE, K. (1995). A Gradient Algorithm Locally Equivalent to the EM Algorithm. *Journal of the Royal Statistical Society, Series B 57*, 425–437.

LANGE, K., RODERICK, J., LITTLE, J., AND TAYLOR, J. (1989). Robust statistical modeling using the t-distribution. *Journal of the American Statistical Association 84*, 881–896.

LESAFFRE, E., ASEFA, M., AND VERBEKE, G. (1999). Assessing the goodness-of-fit of the laird and ware model - an example: The jimma infant survival differential longitudinal study. *Statistics in Medicine 18*, 835–854.

LESAFFRE, E. AND SPIESSENS, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: an example. *Applied Statistics 50*, 325–335.

LIANG, K.-Y. AND ZEGER, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*, 13–22.

LIN, X. AND BRESLOW, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association 91*, 1007–1016.

LIN, X. AND ZHANG, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society B61*, 381–400.

LINDSTROM, M. AND BATES, D. (1988). Newton-raphson and em-algorithm for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association 83*, 1014–1022.

LINDSTROM, M. AND BATES, D. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics 46*, 673–687.

LITTELL, R., MILLIKEN, G., STROUP, W., AND WOLFINGER, R. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute Inc.

LIU, Q. AND PIERCE, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika 81*, 624–629.

LIU, Q. AND RUBIN, D. (1994). The ECME algorithm: A simple extension to EM and ECM with faster monotone convergence. *Biometrika 81*, 633–648.

MADDALA, G. S. (1983). *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

MARX, B. AND EILERS, P. (2005). Multidimensional penalized signal regression. *Technometrics 47*(1), 13–22.

MARX, D. B. AND EILERS, P. (1998). Direct generalized additive modelling with penalized likelihood. *Comp. Stat. & Data Analysis 28*, 193–209.

MCCULLAGH, P. AND NELDER, J. A. (1989). *Generalized Linear Models* (2nd ed.). New York: Chapman & Hall.

MCCULLOCH, C. E. (1994). Maximum likelihood variance components estimation for binary data. *J. Am. Statist. Assoc. 89*, 330–335.

MCCULLOCH, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association 92*, 162–170.

MCCULLOCH, C. E. AND SEARLE, S. R. (2001). *Generalized, linear and mixed models*. New York: Wiley.

MCGILCHRIST, C. A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society B 55*, 945–955.

MCLACHLAN, G. J. AND KRISHNAN, T. (1997). *The EM algorithm and its extensions*. New York: Wiley.

MENG, X.-L. AND VAN DYK, D. (1997). The EM Algorithm - an old folk-song sung to a fast new tune (with discussion). *Journal of Royal Statistical Society, Series B 59*, 511–567.

MENG, X.-L. AND RUBIN, D. (1993). Maximum likelihood estimation via the emc algorithm: A general framework. *Biometrika 80*, 267–278.

NASKAR, M., DAS, K., AND IBRAHIM, J. G. (2005). A semiparametric mixture model for analyzing clustered competing risks data. *Biometrics 61*, 729–737.

NAYLOR, J. C. AND SMITH, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics 31*, 214–225.

NEL, D. (1980). On matrix differentiation in statistics. *South African Statistical Journal 14*, 137–193.

NELDER, J. A. AND WEDDERBURN, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society A 135*, 370–384.

NIEDERREITER, H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik 2*, 84–90.

NIEDERREITER, H. (1992). Random Number Generation and Quasi-Monte Carlo Methods. *CBMS-NFS Regional Conference Series in Applied Mathematics 63*, 264–269.

PAN, J., FANG, K., AND VAN ROSEN, D. (1997). Local influence assessment in growth curve model with unstructured covariance. *Journal of Statistical Planning and Inference 62*, 263–278.

PARISE, H., WAND, M. P., RUPPERT, D., AND RYAN, L. (2001). Incorporation of historical controls using semiparametric mixed models. *Applied Statistics 50*, 31–42.

PATTERSON, H. AND THOMPSON, R. (1971). Recovery of innerblock information when the block sizes are unequal. *Biometrica 40*, 545–554.

PATTERSON, H. AND THOMPSON, R. (1974). *Maximum Likelihood Estimation of Components of Variance*. Proceedings of the 8th International Biometric Conference.

PETRAS, K. (2000). On the smolyak cubature error for analytic functions. *Advances in Computational Mathematics 12*, 71–93.

PETRAS, K. (2001). Fast calculation of coefficients in the smolyak algorithm. *Numerical Algorithms 26*, 93–109.

PINHEIRO, J. AND BATES, D. (2000). *Mixed Effects models in S and SPLUS*. New York: Springer.

PINHEIRO, J. C. AND BATES, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics 4*, 12–35.

RAI, S. AND MATTHEWS, D. (1993). Improving the EM Algorithm. *Biometrics 49*, 587–591.

RAMSAY, J. O. AND SILVERMAN, B. (2002). *Applied functional data analysis: methods and case studies*. Springer-Verlag, New York.

RAMSEY, J., , AND SILVERMAN, B. (2005). *Functional Data Analysis* (2nd ed.). Springer Series in Statistics. Springer-Verlag New York.

REINSCH, C. (1967). Smoothing by spline functions. *Numerische Mathematik 10*, 177–183.

RIPLEY, B. (1987). *Stochastic Simulation*. New York: Wiley.

ROBERT, C. AND CASELLA, G. (2004). *Monte Carlo statistical methods*. New York: Springer.

ROBINSON, G. K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion). *Statistical Science 6*, 15–51.

RUPPERT, D. AND CARROLL, R. J. (1999). Spatially-adaptive penalties for spline fitting. *Australian Journal of Statistics 42*, 205–223.

RUPPERT, D., WAND, M., AND CARROLL, R. (2003). *Semiparametric Regression*. Cambridge University Press.

SAS Institute Inc. (1999). Cary, NC: SAS Institute Inc.

SCHALL, R. (1991). Estimation in generalised linear models with random effects. *Biometrika 78*, 719–727.

SCHAPIRE, R. E. (1990). The strength of weak learnability. *Machine Learning 5*, 197–227.

SCHIMEK, M. (2000). *Smoothing and Regression. Approaches, Computation and Application*. New York: Wiley.

SCHOLZ, T. (2003). *Flexible Modelierung kategorialer Responsevariablen*. Ph. D. thesis, Universität München.

SHMUELI, G., JANK, W., ARIS, A., PLAISANT, C., AND SHNEIDERMAN, B. (2005). Exploring auction databases through interactive visualization. Technical Report, Smith School of Business, University of Maryland, College Park.

SILVERMAN, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *Annals of Statistics 12*, 898–916.

SIMONOFF, J. S. (1996). *Smoothing Methods in Statistics*. New York: Springer-Verlag.

SINHA, D., TANNER, M., AND HALL, W. (1994). Maximization of the marginal likelihood of grouped survival data. *Biometrika 81*, 53–60.

SMOLYAK, S. A. (1963). Quadrature and interpolation formulas for tensor products of certain classes of functions. *Dokl. Akad. Nauk SSSR 148*, 1042–1043. Russian, Engl. Transl.: Soviet Math. Dokl. 4:240–243, 1963.

SOLOMON, P. AND COX, D. (1992). Nonlinear components of variance models. *Biometrika 79*, 1–11.

SPEED, T. (1991). That BLUP is a good thing: The estimation of random effects: Comment. *Statistical Science 6*, 42–44.

STANISWALIS, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association 84*, 276–283.

STEELE, B. M. (1996). A modified em algorithm for estimation in generalized mixed models. *Biometrics 52*, 1295–1310.

STIRATELLI, R., LAIRD, N., AND WARE, J. H. (1984). Random-effects models for serial observation with binary response. *Biometrics 40*, 961–971.

STONE, C., HANSEN, M., KOOPERBERG, C., AND TRUONG, Y. (1997). Polynomial splines and their tensor products in extended linear modeling. *The Annals of Statistics 25*, 1371–1470.

STROUD, A. (1971). *Approximate Calculation of Multiple Integrals*. Englewood Cliffs, NJ: Prentice-Hall.

STROUD, A. H. AND SECREST, D. (1966). *Gaussian Quadrature Formulas*. Englewood Cliffs, NJ: Prentice-Hall.

TANNER, M. A. (1993). *Tools for Statistical Inference: Observed Data Augmentation*. Berlin: Springer-Verlag.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B 58*, 267–288.

TUTZ, G. AND BINDER, H. (2006). Generalized additive modelling with implicit variable selection by likelihood based boosting. *Biometrics*.

VERBEKE, G. AND LESSAFRE, E. (1996). A linear mixed-effects models with heterogeneity in the random-effects population. *Journal of the American Statistical Association 91*, 217–221.

VERBEKE, G. AND MOLENBERGHS, G. (2001). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

VERBYLA, A. P., CULLIS, B. R., KENWARD, M. G., AND WELHAM, S. J. (1999). The anlysis of designed experiments and longitudinal data by using smoothing splines. *Applied Statistics 48*, 269–311.

VONESH, E. F. (1996). A note on the use of laplace's approximation for nonlinear mixed-effects models. *Biometrika 83*, 447–452.

WACLAWIW, M. AND LIANG, K. Y. (1993). Prediction of random effects in the generalized linear model. *Journal of the American Statistical Association 88*, 171–178.

WAGNER, S. (2006). Make-or-buy decisions in patenting. *preprint*.

WAND, M. P. (2000). A comparison of regression spline smoothing procedures. *Computational Statistics 15*, 443–462.

WAND, M. P. (2003). Smoothing and mixed models. *Computational Statistics 18*, 223–249.

WANG, S., JANK, W., AND SHMUELI, G. (2005). Forecasting ebay's online auction prices using functional data analysis. Technical Report, Smith School of Business, University of Maryland.

WEI, G. AND TANNER, M. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithm. *Journal of the American Statistical Association 90*, 699–704.

WOLFINGER, R. AND O'CONNELL, M. (1993). Generalized linear mixed models; a pseudolikelihood approach. *Journal Statist. Comput. Simulation 48*, 233–243.

WOLFINGER, R. W. (1994). Laplace's approximation for nonlinear mixed models. *Biometrika 80*, 791–795.

WOOD, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additve models. *Journal of American Statistical Association 99*, 673–686.

WU, H. AND LIANG, H. (2004). Backfitting random varying-coefficient models with time-depenendent smoothing covariates. *Scandinavian Journal of Statistics 31*, 3–19.

ZEGER, S. L. AND DIGGLE, P. J. (1994). Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics 50*, 689–699.

ZEGER, S. L. AND KARIM, M. R. (1991). Generalized linear models with random effects; a Gibbs' sampling approach. *Journal of the American Statistical Association 86*, 79–95.

ZHANG, D. AND DAVIDIAN, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics 57*, 795–802.

ZHANG, D., LIN, X., RAZ, J., AND SOWERS, M. (1998). Semi-parametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association 93*, 710–719.

Lebenslauf

Florian Reithinger

Frau-Holle-Str. 20a
81739 München

15.07.1976	geboren in München
1982 bis 1986	Besuch der Grundschule Neubiberg
1986 bis 1997	Besuch des Heinrich-Heine-Gymnasiums in München
1997	Abitur
1994 bis 2004	Verpflichtung im Bayerischen Katastrophenschutz
1998 bis 2003	Studium der Statistik an der LMU München
2003	Diplom
2003 bis 2006	Beschäftigung als wissensch. Mitarbeiter am Institut für Statistik der LMU