Clustering in Linear and Additive Mixed Models

Felix Heinzl

München 2013

Clustering in Linear and Additive Mixed Models

Felix Heinzl

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

vorgelegt von
Felix Heinzl
aus München

München, den 31. Januar 2013

Erster Berichterstatter: Prof. Dr. Gerhard Tutz
Zweiter Berichterstatter: Prof. Dr. Thomas Kneib
Tag der Disputation: 26. März 2013

Zusammenfassung

Zur Modellierung longitudinaler Daten sind gemischte Modelle weit verbreitet. Diese Modelle sind einerseits durch die Verteilungsannahmen über die Zielgröße und die zufälligen Effekte sowie andererseits durch die Annahme über die Strukturkomponente bestimmt, die den Prädiktor mit dem Erwartungswert der Zielgröße verknüpft. In der vorliegenden Arbeit werden lineare gemischte Modelle und additive gemischte Modelle betrachtet, um entweder den linearen oder den nichtlinearen zeitlichen Effekt auf eine bestimmte Zielgröße zu untersuchen. Somit ist die Strukturkomponente durch den identischen Link bestimmt, während für die bedingte Verteilung der Zielgröße gegeben die Kovariablen eine Normalverteilung angenommen wird.

Es werden für die Spezifikation und Schätzung der zufälligen Effekte zwei Ansätze vorgestellt, die die klassische Annahme von normalverteilten zufälligen Effekten durch flexiblere Verteilungsannahmen ersetzen und dadurch im Besonderen Gruppen von Individuen bilden können. Im ersten Ansatz wird eine penalisierte Mischung aus Normalverteilungen für die Verteilung der zufälligen Effekte angenommen. Der hierbei vorgestellte Strafterm schrumpft die paarweisen Distanzen der Gruppenzentren basierend auf dem "group lasso"- und dem "fused lasso"-Ansatz. Dies hat den Effekt, dass Individuen mit ähnlichen zeitlichen Verläufen der Zielgröße derselben Gruppe zugeordnet werden. In einem alternativen Ansatz wird eine approximierte Dirichlet-Prozess-Mischung als Verteilung der zufälligen Effekte herangezogen, die die Clustereigenschaft des Dirichlet-Prozesses zum Aufdecken einer Gruppenstruktur ausnützt. Hierbei basiert die Approximation auf der trunkierten Variante der Stabbruch-Darstellung des Dirichlet-Prozesses.

Zum Schätzen beider Ansätze im Rahmen linearer gemischter Modelle werden EM-Algorithmen detailiert entwickelt. Da Dirichlet-Prozesse geeignet sind, um Priori-Annahmen für Verteilungen anzugeben, werden Modelle mit Dirichlet-Prozessen hauptsächlich in der Bayes-Inferenz verwendet und typischerweise mit Markov-Ketten-Monte-Carlo-Methoden geschätzt. In der vorliegenden Arbeit wird das Konzept der Dirichlet-Prozesse in die Likelihood-Inferenz übertragen, indem ein EM-Algorithmus zum Schätzen von linearen gemischten Modellen mit approximierter Dirichlet-Prozess-Mischung vorgestellt wird. Des Weiteren wird dieser Ansatz auf den Fall additiv gemischter Modelle erweitert, wobei hier ein penalisierter Spline zur Modellierung des Zeiteffekts verwendet wird. Für diese Modellklasse wird außerdem eine rein bayesianische Schätzung basierend auf Markov-Ketten-Monte-Carlo-Methoden vorgestellt.

In zahlreichen Anwendungsbeispielen wird gezeigt, wie die verschiedenen Methoden zum Aufdecken von Gruppen verwendet werden können. Simulationsstudien liefern den Nachweis, dass die Prädiktionsgüte von zufälligen Effekten durch die vorgestellten Ansätze verbessert werden kann.

Summary

Mixed models are a classical tool for the modeling of longitudinal data. They are specified by the assumption on the distributions of the response variable and the random effects as well as by the assumed structural component, that is, the link between the mean responses and the predictors. In this thesis, linear mixed models and additive mixed models are considered to examine either a linear or a nonlinear effect of time on a specific response variable. Thus, the structural component is determined by the identity link whereas the conditional distribution of the predictor is assumed to be normal given the covariates.

For the specification and estimation of the random component two approaches are proposed that replace the classical assumption of normally distributed random effects by more flexible distributions and that, in particular, are able to identify clusters of individuals. In the first approach a penalized normal mixture as random effects distribution is assumed. Here, the proposed penalty term shrinks the pairwise distances of cluster centers in terms of the group lasso and the fused lasso method. The intended effect is that individuals with similar time trends are merged into the same cluster. Alternatively an approximate Dirichlet process mixture as random effects distribution is considered. It is able to exploit the cluster property of the Dirichlet process for finding clusters in the data. Here, the truncated version of the stick breaking presentation of the Dirichlet process provides a basis for the approximation.

For fitting these approaches Expectation-Maximization algorithms within the framework of linear mixed models are developed. Since a Dirichlet process allows to specify a prior on probability measures, models concerning Dirichlet processes have been mainly used within the Bayesian inference framework and have been typically estimated by Markov chain Monte Carlo methods. In this thesis, the concept of Dirichlet processes is transferred to the likelihood inference approach by providing an Expectation-Maximization algorithm for fitting linear mixed models with approximate Dirichlet process mixtures. In addition, the concept is extended to additive mixed models, where a penalized spline is used for the fitting of the time trend. For this kind of model a fully Bayesian approach based on Markov chain Monte Carlo simulation techniques is also given.

In several real data examples it is shown how these different approaches can be used for finding clusters in longitudinal data. Simulation studies provide the evidence that prediction accuracy of random effects can be improved by considering the proposed approaches.

Vorwort und Danksagung

Die vorliegende Arbeit entstand im Rahmen meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Statistik der Ludwig-Maximilians-Universität München. In erster Linie möchte ich meinem Doktorvater Herrn Prof. Dr. Gerhard Tutz dafür danken, dass er mir diese Stelle ermöglichte und meine Arbeit über die Jahre hinweg in hervorragender Weise betreute. Mein Dank gilt außerdem Herrn Prof. Dr. Thomas Kneib, der sich freundlicherweise bereit erklärt hat, die Aufgabe des zweiten Berichterstatters zu übernehmen.

Ebenso möchte ich meinen Kollegen am Institut für Statistik danken, die mich mit Hilfestellungen, Problemlösungen und Ratschlägen unterstützten und während meiner Zeit am Institut stets für eine besonders angenehme Arbeitsatmosphäre sorgten. Vor allem bei meinen aktuellen und früheren Kollegen am Lehrstuhl "Seminar für angewandte Stochastik", die darüber hinaus Teile dieser Arbeit Korrektur gelesen haben und bei denen ich immer eine offene Tür vorgefunden habe, möchte ich mich bedanken. Danken will ich auch meiner Schwester und meinen Freunden, insbesondere Stefan Cieczynski, für die wohltuende Abwechslung in schöpferischen Pausen.

Schließlich gilt mein aufrichtiger Dank meinen Eltern, die mir das Statistik-Studium ermöglicht haben, auf dem diese Promotion basiert.

München, im Januar 2013 *Felix Heinzl*

Notation

In this thesis, the following abbreviations are used:

Short form	**Long form**
AIC	Akaike information criterion
BMI	Body mass index
DPM	Dirichlet process mixture
EM	Expectation-Maximization
MCMC	Markov chain Monte Carlo
P-splines	Penalized splines

Furthermore, we utilize the following mathematical abbreviations and symbols:

i.i.d.	Independent and identically distributed,
ind.	Independent,
$\mathbb{N}$	Set of natural numbers,
$\mathbb{R}$	Set of real numbers,
$\mathbb{S}$	Simplex of probabilities,
$\mathbb{E}(X)$	Mean of random variable X,
$\mathrm{Var}(X)$	Variance of random variable X,
$\mathrm{Cov}(X, Y)$	Covariance of random variables X and Y,
$\Gamma(a)$	Gamma function: $\Gamma(a) = \int_0^\infty x^{a-1} \exp(-x)\, dx$,
$B(a, b)$	Beta function: $B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}\, dx$,
$\mathbb{1}(x)$	Indicator function: $\mathbb{1}(x)$ is 1 if x is true and 0 otherwise,
δ_x	Dirac measure for x.

The following common notation is used:

Mathematical object	**Font**	**Examples**
Scalar	Italic	n, N
Vector	Italic, small and bold	$\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$
Matrix	Italic, large and bold	$\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$
σ-field	Calligraphy	$\mathcal{A}$, $\mathcal{C}$, $\mathcal{F}$
Set of distributions	Fraktur	$\mathfrak{G}$

Only the Borel σ-field makes an exception from this scheme and is denoted by $\mathfrak{B}$ as usual. Note that generally all vectors are defined as column vectors.

Here, an overview of all distributions used in this thesis is given:

Distribution	**Short form**	**Parameter**
Normal distribution	$N(\mu, \sigma^2)$	Mean μ, variance σ^2
Multivariate normal distribution	$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$
Uniform distribution	$U(a, b)$	Lower bound a, upper bound b
Multinomial distribution	$M(n, \boldsymbol{\pi})$	Sample size n, probability vector $\boldsymbol{\pi}$
Poisson distribution	$Po(\nu)$	Rate ν
Gamma distribution	$Ga(a, b)$	Shape a, rate b
Inverse-gamma distribution	$IG(a, b)$	Shape a, rate b
Beta distribution	$Be(a, b)$	Shape parameters a and b
Dirichlet distribution	$Dir(\boldsymbol{\alpha})$	Concentration parameter $\boldsymbol{\alpha}$
Dirichlet process	$DP(\alpha, G_0)$	Concentration parameter α, base distribution G_0

Contents

1. Introduction

Mixed Models

The analysis of longitudinal data is a popular task in statistics (Diggle et al., 2002; Fitzmaurice et al., 2004). This kind of data is given when statistical units like individuals are examined at several observation times with regard to some variables of interest. When regression models are applied for investigating the influence of different covariates on a response variable, one has to incorporate the dependence structure in repeated measurements that arises from the fact that measurements belonging to the same individual are typically correlated. This can either be achieved by considering *mixed models* also known as *random effects models* (Laird and Ware, 1982) or by using the *generalized estimation equation* approach proposed by Liang and Zeger (1986). While in the framework of generalized estimation equations the response values are modeled marginally by using only population-specific effects, mixed models contain population-specific fixed effects as well as individual-specific random effects and focus on the conditional distribution of each response value conditional on the corresponding random effect. Mixed models, which were introduced by Fisher (1918), assign each subject i its own random effect $\boldsymbol{b}_i$. For longitudinal data random effects facilitate the modeling of individual deviations from the population trend of the response variable over time. In contrast to the fixed effects, for the random effects a distribution assumption is specified that is typically given by a normal distribution. A more flexible approach has been proposed by Verbeke and Lesaffre (1996). They consider a mixture of normal distributions as random effects distribution:

$$\boldsymbol{b}_i \sim \sum_{h=1}^{N} \pi_h \, N(\boldsymbol{\mu}_h, \boldsymbol{D}), \quad i = 1, \ldots, n.$$

This offers a possibility for clustering individuals due to their time-dependent trend curves of the response variable: If the number of clusters N is smaller than the number of subjects n, several subjects share the same cluster center $\boldsymbol{\mu}_h$ and form a cluster. The covariance matrix $\boldsymbol{D}$ indicates the dispersion of the random effects around their cluster centers. However, this raises the question how to choose the number of clusters. In this dissertation, two penalization approaches are proposed to determine the number of clusters in a data driven way. One is based on the fusion of cluster centers: If the differences of cluster centers are penalized by an appropriate penalty term, some differences are shrunk to zero. Consequently some clusters are fused and the effective number of clusters is reduced. An alternative possibility consists in the penalization of the amounts of the weights π_h. If some weights are shrunk to zero, the corresponding clusters drop out.

Discussion of Penalization Ideas

In regression models regularization approaches are widely used that aim at penalization of the fixed effects of predictors on a response variable. The fundamental papers of Hoerl and Kennard (1970) and Tibshirani (1996) introduced the penalized regression techniques *ridge* respectively *lasso* based on a L_2-norm respectively L_1-norm penalty. The latter one is particulary characterized by the possibility to shrink parameters *and* to set some of them to zero if the corresponding covariates have no impact on the response variable. In the following, it will be shortly discussed in which extent the lasso method could be used for the two penalization goals mentioned in the previous section: On the one hand, we want to shrink differences of cluster centers to zero. However, for fusion of parameters the *fused lasso* idea of Tibshirani et al. (2005) is much more helpful than the direct lasso approach. Furthermore, for incorporating multivariate random effects the fusion concept has to be combined with the *group lasso* approach by Yuan and Lin (2006), which also is an extension of the lasso idea. On the other hand, at first sight the lasso approach seems to be appropriate to shrink weights to zero. But note that probabilities with the range $[0, 1]$ and the restriction that the sum of all probabilities is one cannot be handled in the same way as usual regression coefficients. Thus, we prefer a completely different approach that is based on a Dirichlet process. In this approach, all restrictions are fulfilled automatically and we get rather a shift than a penalization of the weights: High weights become higher and small weights become nearly zero.

Guideline through the Thesis

The main part of this dissertation consists of four chapters, which show different possibilities of clustering in linear and additive mixed models. In Chapters 3 and 4 the two different methods for penalizing the number of clusters introduced in the previous sections are elaborated and applied within the framework of linear mixed models. An Expectation-Maximization (EM) algorithm is used for inference in each case. A comparison of both methods with regard to simulation results and applications can be found in Sections 4.3.3 and 4.4. Chapters 5 and 6 deal with additive mixed models using an approximate Dirichlet process mixture (DPM) as random effects distribution. While in Chapter 5 the model is fitted by using Markov chain Monte Carlo (MCMC) methods, in Chapter 6 the EM algorithm of Chapter 4 is extended to additive mixed models and compared to the MCMC approach of Chapter 5. Chapter 2 takes a special role in the thesis. Here, the theoretical concepts of Dirichlet processes are explained for a better understanding of the methods in chapters using Dirichlet processes. Nevertheless, the single chapters can be read independently of each other. Just for background knowledge or comparisons to other approaches cross references are helpful. Short summaries of the chapters are given in the following:

Chapter 2: Dirichlet Processes

In this chapter we want to depict the idea as well as the highly praised cluster property of the Dirichlet process. The stick breaking representation of the Dirichlet process and the concept of DPMs play a central role in thesis and are also outlined in this chapter.

Chapter 3: Linear Mixed Models with a Group Fused Lasso Penalty

A method is proposed that aims at identifying clusters of individuals that show similar patterns when observed repeatedly. We consider linear mixed models, which are widely used for the modeling of longitudinal data. In contrast to the classical assumption of a normal distribution for the random effects a finite mixture of normal distributions is assumed. Typically, the number of mixture components is unknown and has to be chosen, ideally by data driven tools. For this purpose an EM algorithm-based approach is considered, that uses a penalized normal mixture as random effects distribution. The penalty term shrinks the pairwise distances of cluster centers based on the group lasso and the fused lasso method with the effect that individuals with similar time trends are merged into the same cluster. The strength of regularization is determined by one penalization parameter. For finding the optimal penalization parameter, a new model choice criterion is proposed. The usefulness of this method is illustrated in three applications and in a simulation study.

Chapter 4: Linear Mixed Models with DPMs using EM Algorithm

For the same goal as in the previous chapter an alternative clustering approach is considered. Note that in linear mixed models the assumption of normally distributed random effects is often inappropriate and unnecessary restrictive. The proposed approximate DPM assumes a hierarchical Gaussian mixture that is based on the truncated version of the stick breaking presentation of the Dirichlet process. In addition to the weakening of distributional assumptions, the specification allows to identify clusters of observations with a similar random effects structure. An EM algorithm is given, that solves the estimation problem and that, in certain respects, may exhibit advantages over Markov chain Monte Carlo approaches when modeling with Dirichlet processes. The method is evaluated in a simulation study and applied to the dynamics of unemployment in Germany as well as lung function growth data.

Chapter 5: Additive Mixed Models with DPMs using MCMC methods

When the population time trend is nonlinear, the methods of Chapters 3 and 4 cannot be used, and more flexible approaches like additive mixed models are necessary. For the handling of nonlinearity and heterogeneity in the data, a combination of flexible time trends and individual-specific random effects is required. We propose a fully Bayesian approach based on MCMC simulation techniques that allows for

the semiparametric specification of both the trend function and the random effects distribution. Bayesian penalized splines (P-splines) are considered for the former while a DPM specification allows for an adaptive amount of deviations from normality for the latter. The advantages of such DPM prior structures for random effects are investigated in terms of a simulation study to improve understanding of the model specification before analyzing childhood obesity data.

Chapter 6: Additive Mixed Models with DPMs using EM Algorithm

As in the previous chapter, additive mixed models with a DPM as random effects distribution are considered, that are based on the truncated version of the stick breaking presentation of the Dirichlet process. In contrast to Chapter 5 an EM algorithm is given, that solves the estimation problem and that exhibits advantages over MCMC approaches, which are typically used when modeling with Dirichlet processes. For handling the trend curve the mixed model representation of P-splines is used. The method is evaluated in a simulation study and applied to theophylline data and childhood obesity data.

An important technical fact concerning regression models in general should be mentioned at this stage. Regression models are among other things specified by an assumption for the conditional distribution of the response variable given all covariates. Formally, we abstain from conditioning on the covariates in the model equations of this thesis for a clearer notation. Nevertheless, this condition is implied.

Publications

As research is a dynamic process, parts of this dissertation have already been published in peer reviewed journals or as technical reports and have been done in cooperation with supervising coauthors. Parts of this thesis can be found in

- Heinzl, F. and G. Tutz (2012). Clustering in linear mixed models with a group fused lasso penalty. Technical Report 123, Ludwig-Maximilians-University Munich. (resubmitted). (Chapter 3)
- Heinzl, F. and G. Tutz (2013). Clustering in linear mixed models with approximate Dirichlet process mixtures using EM algorithm. *Statistical Modelling 13*, 41-67. (Chapter 4)
- Heinzl, F., L. Fahrmeir, and T. Kneib (2012). Additive mixed models with Dirichlet process mixture and P-spline priors. *Advances in Statistical Analysis 96*, 47–68. (Chapter 5)

See the corresponding chapters for more details.

Software

For most of the computations in this thesis the programming language C++ (Stroustrup, 1997) and the statistical software R (R Development Core Team, 2012) were used. All new proposed methods are implemented in C++ for a computing time as low as possible. These C++ functions make use of the libraries `ASA047` (Burkhardt, 2008) and `Newmat` (Davies, 2008) and are embedded in R wrapper functions, that are made available by the self-implemented R add-on package `clustmixed` (Heinzl, 2012), which will presumably be made publicly accessible via CRAN (see `http://www.r-project.org`). A test version of the package can be downloaded from `http://www.statistik.lmu.de/~heinzl/research.html`. This package imports the packages `Matrix` (Bates and Maechler, 2012), `lme4` (Bates et al., 2012), `splines` (Bates and Venables, 2011), `ellipse` (Murdoch and Chow, 2012), and `coda` (Plummer et al., 2012). For comparison to other approaches in the simulation studies the R package `lme4` of Bates et al. (2012) and the software BayesX (Belitz et al., 2012) were used. More information can be found in the corresponding sections.

2. Dirichlet Processes

Dirichlet processes as proposed by Ferguson (1973) represent an essential part of this dissertation. The approaches for fitting mixed models in the subsequent Chapters 4, 5 and 6 make use of the so-called cluster property of the Dirichlet process to detect clusters in longitudinal data. This cluster property is the main advantage of Dirichlet processes and the reason for the increase of their popularity over the last years (Hjort et al., 2010). However, the theory of Dirichlet processes is difficult to access at first sight. Hence, in this chapter the concept of Dirichlet processes is outlined in detail to provide a general description of the basic idea of the Dirichlet process. In addition, several representations of the Dirichlet process are given. The structure of this chapter is inspired by Heinzl (2009) and Fahrmeir and Kneib (2011).

2.1. Definition of the Dirichlet Process

In general, a Dirichlet process is used when a prior distribution on a probability measure is needed. First, as an introduction to the definition of the Dirichlet process it is illustrated how priors on spaces of probability measures can be generated. Let $G \in \mathfrak{G}$ denote a probability measure on a measurable space $(\Theta, \mathcal{A})$, where $\mathfrak{G}$ is the set of all probability measures on this measurable space. Thus, a probability space $(\Theta, \mathcal{A}, G)$ is considered with

$$G: \quad \mathcal{A} \rightarrow [0,1].$$

First, let Θ be a finite set $\{\theta_1, \ldots, \theta_m\}$ and let the probability mass function of G be given by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)^T$ with $\pi_j = G(\theta_j),\ j = 1, \ldots, m$. In order to postulate a prior on the unknown vector $\boldsymbol{\pi}$, this vector is assumed to be a random variable. More concretely, let $(\Omega, \mathcal{F}, P)$ be a probability space and $([0,1]^m, \mathfrak{B}^m_{[0,1]})$ be a measurable space. Here, $\mathfrak{B}^m_{[0,1]}$ denotes the m-dimensional Borel σ-field on the m-fold product of the real set $[0,1]$. Then, the random variable $\boldsymbol{\pi}$ is assumed to be a $(\mathcal{F}, \mathfrak{B}^m_{[0,1]})$-measurable function given by

$$\boldsymbol{\pi}: \quad \Omega \rightarrow [0,1]^m.$$

A distribution assumption for $\boldsymbol{\pi}$, and thereby for the probability measure G, can be expressed by a Dirichlet distribution:

Definition: Dirichlet Distribution

A random vector $\boldsymbol{\pi}$ is said to be Dirichlet distributed of order $m \geq 2$ with parameter vector $(\alpha_1, \ldots, \alpha_m)^T$ and $\alpha_j > 0,\ \forall j = 1, \ldots, m$, written $\boldsymbol{\pi} \sim Dir(\alpha_1, \ldots, \alpha_m)$, if it has a probability mass function with respect to the Lebesgue measure on the $(m-1)$-dimensional simplex $\mathbb{S} = \{\boldsymbol{\pi} : 0 \leq \pi_j \leq 1,\ j = 1, \ldots, m,\ \sum_{j=1}^m \pi_j = 1\}$ given by

$$f(\boldsymbol{\pi}) = \frac{\Gamma(\sum_{j=1}^m \alpha_j)}{\prod_{j=1}^m \Gamma(\alpha_j)} \prod_{j=1}^m \pi_j^{\alpha_j - 1}\, \mathbb{1}(\boldsymbol{\pi} \in \mathbb{S}).$$

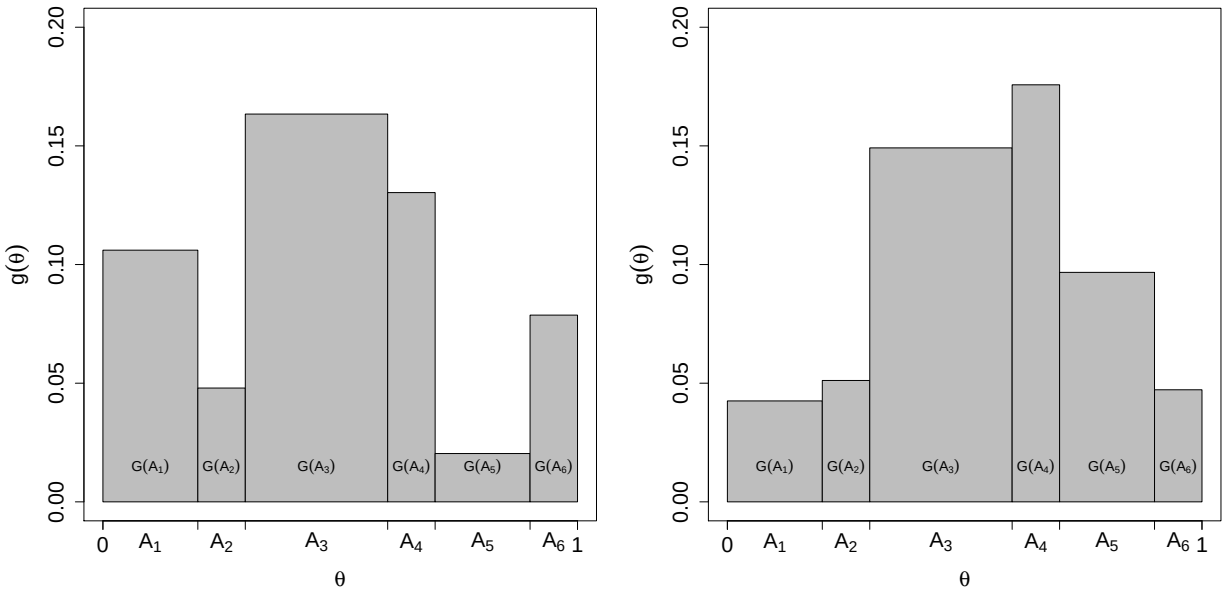
Here $\Gamma(\cdot)$ denotes the gamma function $\Gamma(a) = \int_0^\infty x^{a-1} \exp(-x)\, dx$ and $\mathbb{1}(x)$ is the indicator function that is one if the condition x is true and zero otherwise. Next, Θ is assumed to be a real set. By dividing this set into a finite partition, again, the Dirichlet distribution can be used for formulating a distribution assumption on G. This is also possible for multivariate spaces Θ. Generally, Ferguson (1973) defined the Dirichlet process as follows:

Definition: Dirichlet Process

Let $(\Theta, \mathcal{A})$ be a measurable space and G_0 a probability measure on $(\Theta, \mathcal{A})$, and let $\alpha > 0$.

A stochastic process G indexed by elements A_j of $\mathcal{A}$, is said to be a Dirichlet process on $(\Theta, \mathcal{A})$ with parameters α and G_0, written $G \sim DP(\alpha, G_0)$, if for any measurable partition $\{A_1, \ldots, A_m\}$ of Θ the random vector $(G(A_1), \ldots, G(A_m))^T$ has a Dirichlet distribution with parameter vector $(\alpha\, G_0(A_1), \ldots, \alpha\, G_0(A_m))^T$.

Thus, the Dirichlet process can be seen as generalization of the Dirichlet distribution. In Figure 2.1 two simulated realizations of $G \sim DP(\alpha, G_0)$ with $\alpha = 10$ and $G_0 = U(0, 1)$ for the same fixed partition on [0,1] are visualized by their corresponding probability functions $g(\theta)$. Here, $U(a, b)$ denotes the uniform distribution with the lower bound a and the upper bound b. In this figure the randomness of $G(A_j)$, $j = 1, \ldots, 6$, is illustrated.



(a) Example 1 (b) Example 2

Figure 2.1.: Realizations of $G \sim DP(\alpha, G_0)$ with $\alpha = 10$ and $G_0 = U(0,1)$ for the same fixed partition.

The following paragraph deals with some measure theoretical background of the Dirichlet process. First, in the definition of the Dirichlet process it is seen that a Dirichlet process is formally a stochastic process. In general, a stochastic process is represented by a family of random variables on a specific probability space. Here, the family of random variables is determined by $G(A_j)$, $j = 1, \ldots, m$, for every m and every measurable partition $\{A_1, \ldots, A_m\}$ of Θ. Since these random variables are probabilities, the Dirichlet process has the state space $[0,1]$ with the associated Borel σ-field $\mathfrak{B}_{[0,1]}$. For a fixed partition $\{A_1, \ldots, A_m\}$ the index set of this stochastic process is given by exactly this partition. However, every partition of sets $A_j \in \mathcal{A}$ can be considered. So the index set of the Dirichlet process is formed by $\mathcal{A}$ (Frigyik et al., 2010). Second, note that in the definition of the Dirichlet process a joint distribution of random variables $G(A_1), \ldots, G(A_m)$ for every m and every measurable partition is defined. Ferguson (1973) showed that thus a consistent system of finite dimensional distributions is given verifying the Kolmogorov consistency conditions. Thereby according to the Kolmogorov existence theorem it exists a probability measure on the resulting measurable space $([0,1]^{\mathcal{A}}, \mathfrak{B}_{[0,1]}^{\mathcal{A}})$ yielding these distributions. Here, $[0,1]^{\mathcal{A}}$ symbolizes the space of all functions from $\mathcal{A}$ into $[0,1]$, and $\mathfrak{B}_{[0,1]}^{\mathcal{A}}$ denotes the product σ-field generated by the field of cylinder sets. However, by defining the joint distribution of random variables $G(A_1), \ldots, G(A_m)$ for every measurable partition, a random probability measure G is defined, too. Thus, the probability measure G itself is called *Dirichlet process* just as the family of random variables $G(A_j)$. Now a prior distribution on G is expressed by a specific Dirichlet measure $DP(\alpha, G_0)$, that is a probability measure on $(\mathfrak{G}, \mathcal{C})$. Here, $\mathcal{C}$ is the smallest σ-field generated by sets of the form $\{G : G(A) < r\}$ with $A \in \mathcal{A}$ and

$r \in [0,1]$ (Sethuraman, 1994). Thus, one gets a probability space $(\mathfrak{G}, \mathcal{C}, DP(\alpha, G_0))$ with the probability measure

$$DP(\alpha, G_0): \quad \mathcal{C} \to [0,1].$$

Such a Dirichlet measure can be used for prior assumptions on probability measures. A concrete prior assumption is specified by the choice of the parameters α and G_0. The meaning of the base distribution G_0 is mainly given by $\mathbb{E}(G) = G_0$ that can be proved by using the calculation rule for the mean of Dirichlet distributed random vectors:

$$\mathbb{E}(G(A_j)) = \frac{\alpha G_0(A_j)}{\sum_{l=1}^{m} \alpha G_0(A_l)} = G_0(A_j), \quad j = 1, \ldots, m.$$

Similarly, the variances can be calculated by

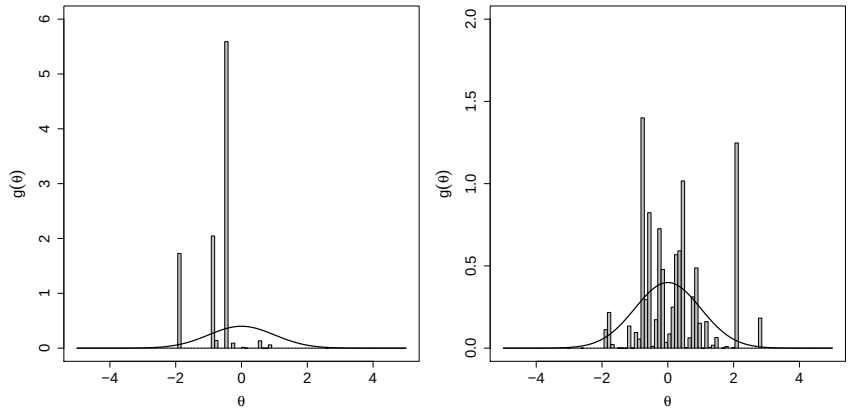
$$\mathrm{Var}(G(A_j)) = \frac{\alpha G_0(A_j) \cdot (\alpha - \alpha G_0(A_j))}{\alpha^2 \cdot (\alpha + 1)} = \frac{G_0(A_j) \cdot (1 - G_0(A_j))}{\alpha + 1}, \quad j = 1, \ldots, m.$$

Obviously, the concentration parameter α acts as inverse variance parameter and controls the confidence in the base distribution G_0. For visualizing this feature, in Figure 2.2 realizations of $G \sim DP(\alpha, G_0)$ with $G_0 = N(0,1)$ and with different values for α are drawn, where $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. In each case the grey boxes correspond to the joint distribution of $G(A_1), \ldots, G(A_m)$ for a given partition and so to the probability measure G while the black line symbolizes the density function of $G_0 = N(0,1)$. As it can be seen, the larger α is, the more similar G is to G_0. Note that instead of representing the parameters of the Dirichlet process by the scalar α and the probability measure G_0, it is also possible to use the finite measure $\alpha\, G_0$, which is nothing else than the product of α and G_0.

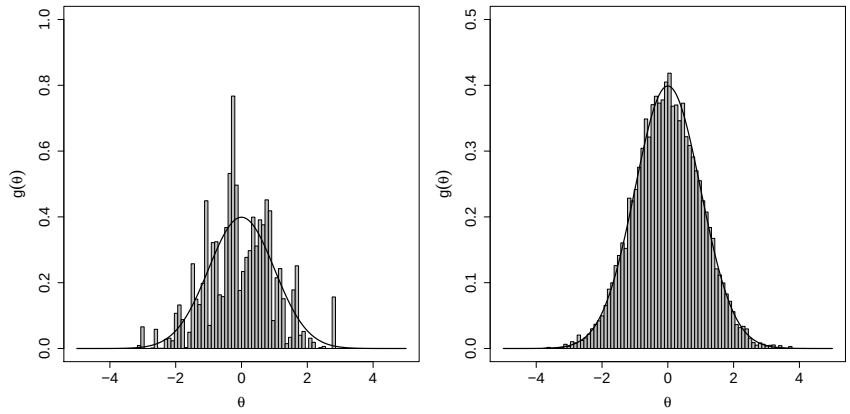
As shown by Blackwell (1973), assuming $G \sim DP(\alpha, G_0)$ selects a discrete distribution G with probability one even if the base distribution G_0 is a continuous distribution. A consequence of this feature is the cluster property of the Dirichlet process: Suppose that an i.i.d. sample $\theta_1, \ldots, \theta_n$ is drawn according to

$$\begin{array}{rcll} G & \sim & DP(\alpha, G_0), & \\ \theta_i | G & \overset{i.i.d.}{\sim} & G, & i = 1, \ldots, n. \end{array} \tag{2.1}$$

While in the first step a realization of G is obtained by simulation from the Dirichlet measure $DP(\alpha, G_0)$, in the second step realizations from G are drawn. The condition on G is necessary here to emphasize that the random measure G is given in this step. Due to the discreteness of G some realizations θ_i can be identical and form a cluster: $\theta_i = \theta_j$ with $i \neq j$. Thus, the parameters θ_i, $i = 1, \ldots, n$, can also be represented by cluster locations μ_h, $h = 1, \ldots, k$, with $k \leq n$ and cluster allocation variables $c_i \in \{1, \ldots, k\}$, $i = 1, \ldots, n$. This relationship is given by $\theta_i = \mu_{c_i}$. Here $c_i = h$ holds if object i belongs to cluster h. The cluster property will be illustrated in the following sections. There, it will be shown that



(a) $\alpha = 1$ (b) $\alpha = 10$



(c) $\alpha = 100$ (d) $\alpha = 10000$

Figure 2.2.: Influence of α on realizations of $G \sim DP(\alpha, G_0)$ with $G_0 = N(0, 1)$. In each case the grey boxes correspond to G while the black line symbolizes the density function of G_0 (Fahrmeir and Kneib, 2011).

the number of clusters is controlled by the concentration parameter α: For large values of α the number of clusters is large while for a low α one gets only few clusters.

Another important property is the conjugacy of the Dirichlet process. That means that if an i.i.d. sample $\theta_1, \ldots, \theta_n$ is drawn from G and if the prior distribution for G is assumed to be a Dirichlet measure according to the equations (2.1), the posterior distribution of $G|\theta_1, \ldots, \theta_n$ is a Dirichlet measure again. More concretely, it follows that

$$G|\theta_1, \ldots, \theta_n \sim DP\left(n + \alpha, \frac{1}{n+\alpha}\sum_{i=1}^{n}\delta_{\theta_i} + \frac{\alpha}{n+\alpha}G_0\right),$$

which is proved in Appendix A.1. Here, δ_{θ_i} denotes the Dirac measure on θ_i. This conjugacy property is used in Section 2.3 for deriving the predictive distribution $\theta_{n+1}|\theta_1, \ldots, \theta_n$.

2.2. Stick Breaking Representation

For understanding the nature of the Dirichlet process the constructive definition of the Dirichlet process by Sethuraman (1994) is even more helpful than the definition in Section 2.1 itself. This stick breaking representation implies that $G \sim DP(\alpha, G_0)$ is equivalent to

$$G = \sum_{h=1}^{\infty} \pi_h \, \delta_{\mu_h}, \tag{2.2}$$

with locations $\mu_h \in \Theta$ and weights $\pi_h \in [0, 1]$. The locations are simulated by $\mu_h \overset{i.i.d.}{\sim} G_0$ and serve as possible cluster locations. The weights are constructed through the stick breaking procedure

$$\begin{array}{rcll} \pi_h & = & v_h \prod_{l<h}(1 - v_l), & h \in \mathbb{N}, \\ v_h & \overset{i.i.d.}{\sim} & Be(1, \alpha), & h \in \mathbb{N}, \end{array}$$

where $Be(\cdot, \cdot)$ denotes the beta distribution and $v_h,\ h \in \mathbb{N}$, are reparameterized weights. Thus, the random measure G is represented as a weighted sum of point masses with random weights π_h linked to the locations μ_h. According to

$$\begin{aligned} \prod_{h=1}^{N}(1 - v_h) = (1 - v_N) \prod_{h=1}^{N-1}(1 - v_h) = \prod_{h=1}^{N-1}(1 - v_h) - \underbrace{v_N \prod_{h=1}^{N-1}(1 - v_h)}_{\pi_N} = \ldots = \\ = (1 - v_1) - \sum_{h=2}^{N} \pi_h = 1 - \sum_{h=1}^{N} \pi_h, \end{aligned} \tag{2.3}$$

one gets a recursive definition of weights $\pi_h = v_h \left(1 - \sum_{l<h} \pi_l\right)$, $h \in \mathbb{N}$, that is visualized in Figure 2.3 and that gives the procedure its name. It works as follows: First, for getting π_1 a piece is broken away from a stick of length one. Next, from the remainder of the stick, $1 - \pi_1$, breaks a further piece away, called π_2 and so on. So the random weights decrease stochastically as the index h grows (Ishwaran and James, 2001). More concretely, $\mathbb{E}(\sum_{h=N+1}^{\infty} \pi_h)$ converges to zero exponentially with $N \to \infty$:

$$\begin{aligned} \mathbb{E}\left(\sum_{h=N+1}^{\infty} \pi_h\right) = \mathbb{E}\left(1 - \sum_{h=1}^{N} \pi_h\right) \overset{(2.3)}{=} \mathbb{E}\left(\prod_{h=1}^{N}(1 - v_h)\right) = \prod_{h=1}^{N} \mathbb{E}(1 - v_h) = \\ = \prod_{h=1}^{N}(1 - \mathbb{E}(v_h)) = \prod_{h=1}^{N}\left(1 - \frac{1}{\alpha + 1}\right) = \left(\frac{\alpha}{\alpha + 1}\right)^N \overset{N \to \infty}{\longrightarrow} 0. \end{aligned} \tag{2.4}$$

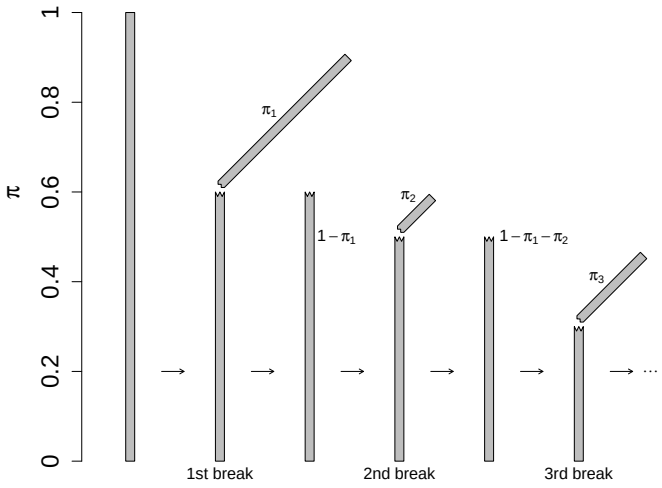


Figure 2.3.: Construction of $\pi_1, \pi_2, \ldots$ by stick breaking (Heinzl, 2009).

So an established concept to make the stick breaking procedure applicable in practice is to approximate the Dirichlet process by considering

$$G = \sum_{h=1}^{N} \pi_h \, \delta_{\mu_h},$$

with large enough N. Here, all locations μ_h and all weights v_h and π_h are constructed as before with the exception of $v_N = 1$. Because of the recursive definition of weights the restriction $v_N = 1$ ensures that $\sum_{h=1}^{N} \pi_h = 1$ and implicates that in the truncated version of the stick breaking presentation the last weight π_N absorbs all the remaining probabilities $\pi_N, \ldots, \pi_\infty$ of the untruncated Dirichlet process. So the truncation would be only admissible if these weights are very low. Therefore, Ohlssen et al. (2007) proposed to set N so that the condition

$$\mathbb{E}\left(1 - \sum_{h=1}^{N-1} \pi_h\right) < \varepsilon, \tag{2.5}$$

is fulfilled for $\varepsilon > 0$. According to equation (2.4) the condition (2.5) can be transformed into

$$\begin{aligned} \left(\frac{\alpha}{\alpha+1}\right)^{N-1} &< \varepsilon \\ (N-1)\log\left(\frac{\alpha}{\alpha+1}\right) &< \log(\varepsilon) \\ N &> 1+\frac{\log(\varepsilon)}{\log\left(\frac{\alpha}{\alpha+1}\right)}. \end{aligned} \tag{2.6}$$

Thus, for a given α and ε a lower bound for N can be determined. This strategy is applied in the Chapters 4 and 6.

In summary, the stick breaking representation has several benefits. First, it yields a possibility to simulate a realization of G as well as realizations from G. Second, it is useful for inference with Dirichlet processes like in the Chapters 4, 5 and 6. Third, theoretical features of the Dirichlet process can be seen easily. According to equation (2.2) G has a countably infinite support. Thus, a realization of $G \sim DP(\alpha, G_0)$ is a discrete probability measure with probability one. The proof of the discreteness without using the stick breaking representation is considerably more complicate (Blackwell, 1973).

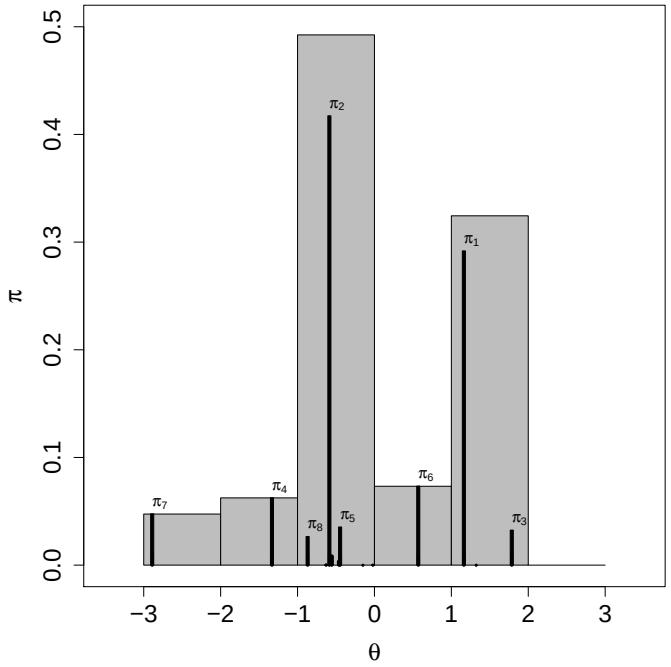
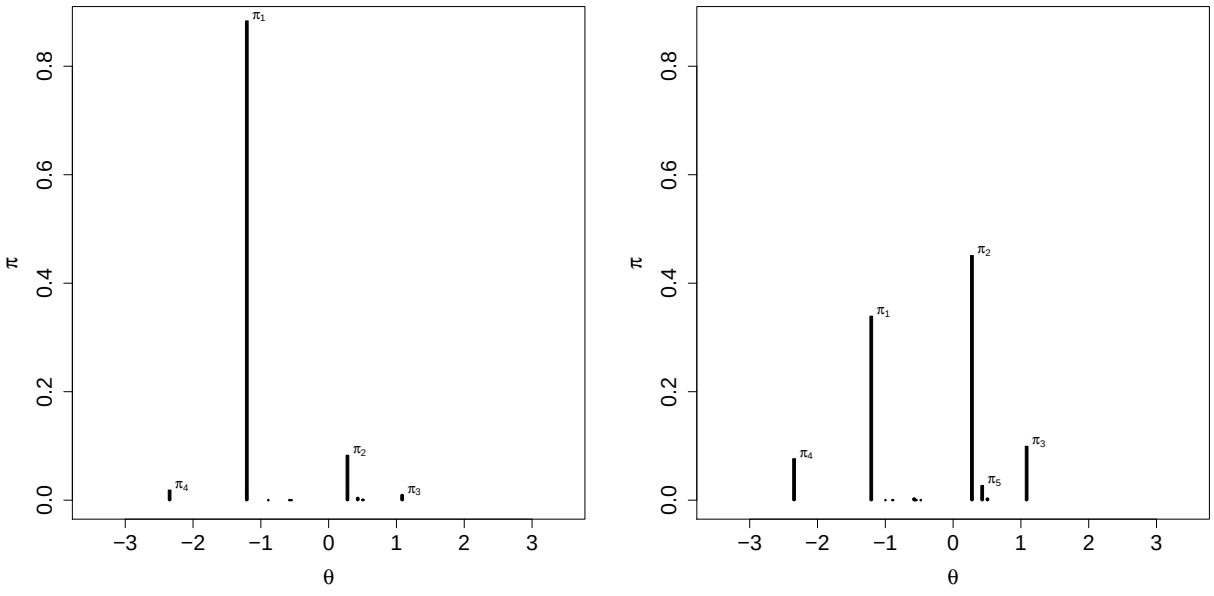


Figure 2.4.: Realization of $G \sim DP(\alpha, G_0)$ with $\alpha = 1$ and $G_0 = N(0,1)$. The black bars symbolize the probability function of G. The grey boxes show the realization of G for the given partition.

See Figure 2.4 for an example of a simulated realization of $G \sim DP(\alpha, G_0)$ with $\alpha = 1$ and $G_0 = N(0,1)$. In this figure the black bars symbolize the weights π_h,

$h = 1, \ldots, N$, and thus the probability function of G. Choosing the partition $\{A_1, \ldots, A_8\}$ $= \{(-\infty, -3], (-3, -2], (-2, -1], (-1, 0], (0, 1], (1, 2], (2, 3], (3, \infty)\}$ the corresponding realization of $(G(A_1), \ldots, G(A_8))^T$ as noted in the definition of the Dirichlet process is visualized in the same figure by grey boxes. For example, $G((1, 2])$ is mainly the sum of π_1 and π_3 whereas $G((0, 1])$ is almost equivalent to π_6. While in Figure 2.1 the discreteness of G is hidden, Figure 2.4 reveals this feature as well as the connection between these two different representations of the Dirichlet process. Another interesting aspect of Figure 2.4 is that clearly just about the first ten weights are visually higher than zero. Regarding again the stick breaking idea, obviously after ten breaks the remaining stick has a length of almost zero. So no more pieces can be broken away from this stick. Here, the cluster property of the Dirichlet process is seen: Suppose that $\theta_i | G \overset{i.i.d.}{\sim} G,\ i = 1, \ldots, n$ with G simulated by $G \sim DP(\alpha, G_0)$. Imagine that G looks like the realization in Figure 2.4. Since the support of G with weights that are clearly different from zero consists of just a few elements, some realizations θ_i are identical and form a cluster: $\theta_i = \theta_j = \mu_h$ with $i \neq j$. By this a natural clustering of similar objects can be realized.

The number of clusters is determined by α, which also controls the confidence in the base distribution G_0. See Figure 2.5 for two realizations of $G \sim DP(\alpha, G_0)$ with $G_0 = N(0, 1)$ and two different confidence parameters $\alpha = 0.5$ and $\alpha = 1$. It can be stated: The larger the confidence parameter α, the more clusters are available.



(a) $\alpha = 0.5$ (b) $\alpha = 1$

Figure 2.5.: Realizations of $G \sim DP(\alpha, G_0)$ with $G_0 = N(0, 1)$.

2.3. Pólya Urn Scheme

Apart from the stick breaking procedure there exists another strategy for making inference possible when Dirichlet processes are considered. Instead of truncating the random probability G in the stick breaking representation of the Dirichlet process (equation (2.2)) as it is done in the Chapters 4, 5 and 6, another possibility to handle the unknown probability measure G consists in the marginalization of G. However, the marginalization of G has the unwished side effect that only realizations from G but not of G can be simulated in contrast to the stick breaking procedure where realizations of G can be simulated, too. Because of this and other disadvantages of the Pólya urn scheme compared to the stick breaking representation, that are discussed by Ishwaran and James (2001), only the latter one is considered in this thesis. Nevertheless, the theoretical background of the Pólya urn scheme, namely the connection between the Dirichlet process and the extended Pólya urn scheme (Blackwell and MacQueen, 1973), will be explained in the following for a better understanding of the character of the Dirichlet process. In general, the classical Pólya urn model works as follows: An urn contains a finite number of colored balls. Step by step one ball is drawn randomly from the urn. After each drawing the drawn ball is put back along with a new ball of the same color. Now, this Pólya urn scheme is extended by allowing an uncountable infinite set of colors. Formally this procedure is described by a so-called Pólya sequence, which is defined as follows:

Definition: Pólya Sequence

Let $(\Theta, \mathcal{A})$ be a Polish space, i.e. a complete separable metric space. A sequence of random variables $(\theta_n)_{n\in\mathbb{N}}$ with $\theta_n \in \Theta$ is said to be a Pólya sequence with parameter $\alpha\, G_0$ if for every $A \in \mathcal{A}$ the following two properties are fulfilled:

(a) $P(\theta_1 \in A) = \frac{\alpha\, G_0(A)}{\alpha\, G_0(\Theta)} = G_0(A)$,

(b) $P(\theta_{n+1} \in A|\theta_1, \ldots, \theta_n) = \frac{\sum_{i=1}^{n} \delta_{\theta_i}(A) + \alpha\, G_0(A)}{\sum_{i=1}^{n} \delta_{\theta_i}(\Theta) + \alpha\, G_0(\Theta)} = \frac{\sum_{i=1}^{n} \delta_{\theta_i}(A) + \alpha\, G_0(A)}{n+\alpha}$.

Here, Θ denotes the set of colored balls. The distribution of the balls at the beginning is given by G_0. By considering the counting measure for $\alpha\, G_0$ one gets again the classical Pólya urn scheme with a finite number of balls. Note that in the definition of the Pólya sequence a Polish space is considered whereas in the definition of the Dirichlet process any measurable space is allowed. However, this theoretical restriction for the relationship between Dirichlet process and Pólya sequence is negligible in practice. On the one hand, this relationship is that a Pólya sequence with parameter $\alpha\, G_0$ converges with probability one as $n \to \infty$ to a limiting discrete distribution G with

(i) $G \sim DP(\alpha, G_0)$,

(ii) $\theta_n|G \overset{i.i.d.}{\sim} G, \quad n \in \mathbb{N}$.

This is proved by Blackwell and MacQueen (1973). On the other hand, assuming (i) and (ii) one gets a Pólya sequence with parameter $\alpha\, G_0$. See Appendix A.2 for the corresponding proof. The importance of this proof is that marginalization of G yields the predictive distribution

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{1}{n+\alpha} \sum_{i=1}^{n} \delta_{\theta_i} + \frac{\alpha}{n+\alpha} G_0, \tag{2.7}$$

and thus a possibility to simulate realizations from G. The Pólya urn simulation idea works as follows: Given $\theta_1, \ldots, \theta_n$ one gets θ_{n+1} by drawing (with replacement) from the urn containing the "balls" $\theta_1, \ldots, \theta_n$ with probability $\frac{n}{n+\alpha}$ or by drawing from G_0 with probability $\frac{\alpha}{n+\alpha}$. The new ball θ_{n+1} is put into the urn. This procedure is visualized for $n = 6$ in Figure 2.6. Again, the cluster property of the Dirichlet process becomes apparent: If θ_7 is drawn from the urn, obviously θ_7 has to be identical to at least one other ball. For a low value of α the probability for drawing from the urn is relatively high. Note that G_0 is typically assumed to be continuous. So in Figure 2.6 G_0 is marked by a continuum of colors.

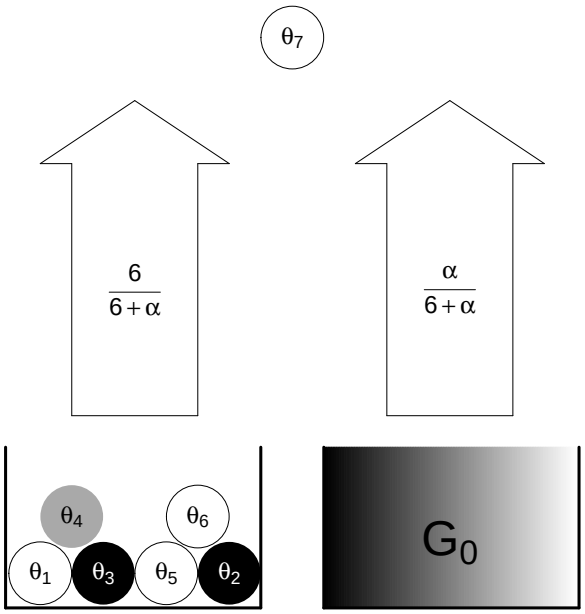


Figure 2.6.: Illustration of the Pólya sequence.

However, because of the cluster property the parameters $\theta_1, \ldots, \theta_n$ can also be identified by the cluster locations $\mu_1, \ldots, \mu_k$ with $k \leq n$. Thereby the predictive distribution (2.7) can be rewritten by

$$\theta_{n+1}|\mu_1, \ldots, \mu_k \sim \frac{1}{n+\alpha} \sum_{h=1}^{k} n_h \, \delta_{\mu_h} + \frac{\alpha}{n+\alpha} G_0. \tag{2.8}$$

Here, n_h denotes the absolute frequency of elements in cluster h. This simulation idea is known as Chinese restaurant process. Imagine that guests arrive one after the other at a Chinese restaurant. The first guest θ_1 chooses a free table labeled with μ_1. When later on the guest θ_{n+1} arrives, k tables are occupied, each with n_h persons. He can choose to join other guests at an occupied table μ_h, $h = 1, \ldots, k$, with probability $\frac{n_h}{n+\alpha}$ or to sit down at a new table $\mu_{k+1} \sim G_0$ with probability $\frac{\alpha}{n+\alpha}$ (see Figure 2.7). The lower α, the lower is the probability for choosing an empty table. Again, the cluster property of the Dirichlet process is evident. People sitting at the same table form a cluster. Theoretically infinitely many tables are necessary for infinite many guests. But especially for low values of α a finite number of tables N is sufficient. This number of tables corresponds with the truncation in the stick breaking representation. Note that at each table any number of guests can sit. The typical round tables in Chinese restaurants should illustrate this.

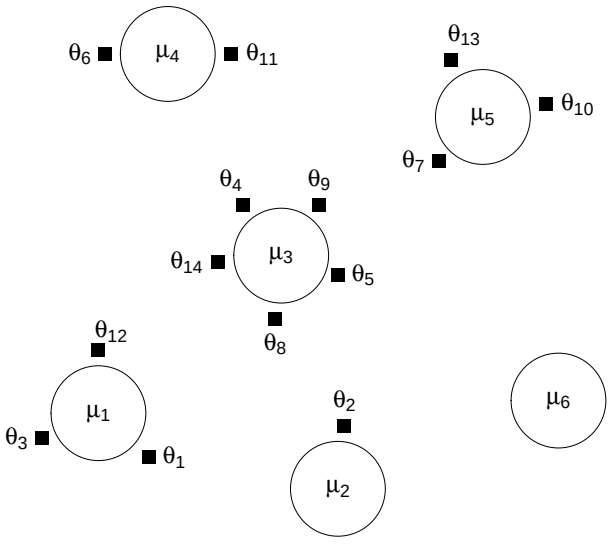


Figure 2.7.: Illustration of the Chinese restaurant process (Heinzl, 2009)

2.4. Dirichlet Process Mixtures

The almost sure discreteness of the Dirichlet process seems to be cumbersome if one wants a prior on a class of continuous distributions (Antoniak, 1974) like, for example, in the mixed models of the Chapters 4, 5 and 6 where a more flexible distribution assumption for the random effects distribution than the normal distribution is aspired. In general, random effects $\boldsymbol{b}_i$, $i = 1, \ldots, n$, are individual-specific vectors containing, for example, a random intercept b_{i0} and a random slope b_{i1}. Assuming a Dirichlet process for the random effects distribution creates ties among the random effects due to the discreteness and especially due to the cluster property of the Dirichlet process. On the one hand, this has the advantage that clusters of subjects can be formed. But on the other hand, the prediction accuracy of the random effects suffers from the restriction that some individuals must have the same random effect in comparison to the normal distribution, that allows different random effects. Thus, usually for the random effects distribution a DPM is considered

$$\begin{array}{rcll} \boldsymbol{b}_i|\boldsymbol{\theta}_i & \overset{ind.}{\sim} & F(\boldsymbol{\theta}_i), & i = 1, \ldots, n, \\ \boldsymbol{\theta}_i|G & \overset{i.i.d.}{\sim} & G, & i = 1, \ldots, n, \\ G & \sim & DP(\alpha, G_0), & \end{array}$$

where $F(\cdot)$ is an arbitrary continuous distribution with parameter vector $\boldsymbol{\theta}_i$. Then each subject has its own random effect and clusters can still be identified by the parameters $\boldsymbol{\theta}_i$. Remember that for $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ only $k \leq n$ different values are given due to the cluster property of the Dirichlet process while all values of $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ are different because of the continuity of $F(\cdot)$. Finally, given the probability measure G, for the random effects density function one gets

$$p(\boldsymbol{b}_i|G) = \int f(\boldsymbol{b}_i|\boldsymbol{\theta}_i) dG(\boldsymbol{\theta}_i), \tag{2.9}$$

where G acts as mixing distribution and $f(\cdot)$ denotes the density function of the continuous distribution $F(\cdot)$. Because of the discreteness of G the integral in equation (2.9) can be rewritten as sum. More concretely, by using the stick breaking procedure (2.2), one obtains

$$\begin{array}{rcll} p(\boldsymbol{b}_i|\boldsymbol{\pi}, \boldsymbol{\mu}) & = & \sum_{h=1}^{\infty} \pi_h \, f(\boldsymbol{b}_i|\boldsymbol{\mu}_h), & i = 1, \ldots, n, \\ \pi_h & = & v_h \prod_{l<h}(1 - v_l), & h \in \mathbb{N}, \\ v_h & \overset{i.i.d.}{\sim} & Be(1, \alpha), & h \in \mathbb{N}, \\ \boldsymbol{\mu}_h & \overset{i.i.d.}{\sim} & G_0, & h \in \mathbb{N}. \end{array} \tag{2.10}$$

Here, the vectors $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)^T$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T, \ldots)^T$ determine the probability measure G. Note that $\boldsymbol{\theta}_i = \boldsymbol{\mu}_h$ if individual i belongs to cluster h. In summary, equation (2.10) yields an infinite mixture distribution as random effects distribution. Using the truncated version of the stick breaking presentation from Section 2.2, one gets formally a finite mixture distribution for the random effects but with the special characteristic that the weights are constructed by the stick breaking procedure. In practice, for $F(\cdot)$ typically (multi-

variate) normal distributions are assumed, in which the means are in accordance with the parameters $\boldsymbol{\theta}_i$, $i = 1, \ldots, n$, respectively $\boldsymbol{\mu}_h$, $h = 1, \ldots, N$, that follow the unknown distribution $G \sim DP(\alpha, G_0)$. Thus, one obtains formally a finite normal mixture as random effects distribution

$$\begin{array}{rcll} \boldsymbol{b}_i|\boldsymbol{\pi}, \boldsymbol{\mu} & \overset{i.i.d.}{\sim} & \sum_{h=1}^{N} \pi_h \, N(\boldsymbol{\mu}_h, \boldsymbol{D}), & i = 1, \ldots, n, \\ \pi_h & = & v_h \prod_{l<h}(1 - v_l), & h = 1, \ldots, N, \\ v_h & \overset{i.i.d.}{\sim} & Be(1, \alpha), & h = 1, \ldots, N-1, \\ \boldsymbol{\mu}_h & \overset{i.i.d.}{\sim} & G_0, & h = 1, \ldots, N, \end{array}$$

with the stick breaking weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^T$, the means $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_N^T)^T$ and the covariance matrix $\boldsymbol{D}$. This random effects distribution will be picked up in the Chapters 4, 5 and 6. There, some strategies will be explained how to estimate all the unknown parameters.

3. Linear Mixed Models with a Group Fused Lasso Penalty

3.1. Introduction

Linear mixed models, which were proposed by Laird and Ware (1982), are a common tool for the modeling of longitudinal data. The model can be written as

$$\boldsymbol{y}_i|\boldsymbol{b}_i \overset{ind.}{\sim} N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i,\, \sigma^2\boldsymbol{I}_{n_i}), \quad i = 1, \ldots, n, \tag{3.1}$$

where $\boldsymbol{y}_i$ contains the response values y_{ij} observed for subject i at observation times t_{ij}, $j = 1, \ldots, n_i$. Here, $\boldsymbol{I}_{n_i}$ is the identity matrix with dimension n_i. Population effects are included in the parameter $\boldsymbol{\beta}$ whereas $\boldsymbol{b}_i$ represents the individual-specific effects. $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ denote the corresponding individual design matrices. All observations y_{ij} are normally distributed conditioned on the random effects and are regarded as independent with the same variance σ^2. The classical assumption in (3.1) is a Gaussian distribution for the random effects, i.e. $\boldsymbol{b}_i$ i.i.d. $N(\boldsymbol{0}, \boldsymbol{D})$, see, for example, Verbeke and Molenberghs (2000) and Ruppert et al. (2003). While this choice is mathematically convenient, it is often questionable in applications for several reasons. The normal distribution is symmetric, unimodal and has light tails. Since the distributional assumption is made on unobserved quantities, it is typically hard to validate these properties based on estimates. Possible skewness and multimodality (arising, for example, from an unconsidered grouping structure in the data) may be masked when checking the normal distribution based on estimated random effects. In contrast to this homogeneity model the heterogeneity model introduced by Verbeke and Lesaffre (1996) is much more flexible. It assumes

$$\boldsymbol{b}_i \sim \sum_{h=1}^{N} \pi_h\, N(\boldsymbol{\mu}_h, \boldsymbol{D}), \quad i = 1, \ldots, n, \tag{3.2}$$

where $\pi_1, \ldots, \pi_N$ are mixture weights. Several extensions and alternatives to this heterogeneity model have been proposed. For example, Gaffney and Smyth (2003) used random effects regression mixtures in the context of curve clustering. Approaches for clustering functional data were proposed by James and Sugar (2003) and Liu and Yang (2009). Celeux et al. (2005), Ng et al. (2006) and Scharl et al. (2010) dealt with mixtures of linear mixed effects models. In these approaches the mixture weights, the variance parameters and all fixed effects are cluster-specific whereas in equation (3.2) just the mixture weights and the locations corresponding to the time trend depend on the cluster. While Booth et al.

(2008) extended this concept by proposing a stochastic search algorithm for finding the partition, that maximizes an objective function based on the classification likelihood, De la Cruz-Mesía et al. (2008) generalized the approach to a mixture of non-linear hierarchical models. Villarroel et al. (2009) extended the heterogeneity model to allow for a multivariate response variable. In addition, heteroscedastic normal mixtures in the random effect distribution for multiple longitudinal markers were considered by Komárek et al. (2010) for linear mixed models and by Komárek and Komárková (2013) for generalized linear mixed models. However, in all these approaches it is necessary to fix the number of mixture components for estimation even though in most applications the number of mixture components is unknown. Additional procedures are typically provided for selecting this number, which are usually based on information criteria. A data driven choice of this number can be achieved by penalization of pairwise distances of cluster centers by a group fused lasso penalty term. In contrast to the approaches in Chapter 4 and Chapter 6, that aim at penalizing the reparameterized mixture weights, the "penalized heterogeneity approach" introduced here reduces the number of clusters by penalizing the cluster centers in the form

$$\sqrt{N \cdot q} \sum_{h<l} \|\boldsymbol{\mu}_h - \boldsymbol{\mu}_l\|. \tag{3.3}$$

The idea of the penalty term is as follows: If two cluster locations are very similar in terms of the Euclidean distance $\|\cdot\|$, these clusters should be fused. Therefore only the relevant clusters are expected to remain in the model. Fusion methods in regression modeling, but with quite differing penalty terms, have been proposed by Tibshirani et al. (2005). Penalty terms that include vectors, as is needed here, have been considered by Yuan and Lin (2006) but not in a fusion context. It should be noted that the factor $\sqrt{N \cdot q}$, where q denotes the dimension of random effects, is used for incorporating the number of parameters to be estimated. For inference, we extend the traditional EM algorithm (Dempster et al., 1977) used in the heterogeneity model of Verbeke and Lesaffre (1996) by adding the penalty term (3.3) multiplied by a penalty parameter to the logarithm of the complete but not fully observed likelihood (see Section 3.2.1). To find the optimal penalty parameter we introduce a new model choice criterion, which is based on the concept of Braun et al. (2012) (see Section 3.2.2). The usefulness of our approach is demonstrated by three applications (see Section 3.3) and a simulation study (see Section 3.4). Large parts of this chapter are based on Heinzl and Tutz (2012).

It will be shown that our penalized heterogeneity approach is much more flexible than the conventional homogeneity model and allows to determine the number of clusters automatically. Regularization allows to identify the underlying clusters and cluster individuals in longitudinal studies.

3.2. Linear Mixed Models with a Group Fused Lasso Penalty

3.2.1. Estimation

For the model introduced in Section 3.1 we give an EM algorithm, which is based on derivations by McLachlan and Krishnan (1997) and McLachlan and Peel (2000) and is similar to the algorithm used by Verbeke and Molenberghs (2000) but includes the penalty term (3.3). Let the parameters be collected in $\boldsymbol{\xi} = (\boldsymbol{\pi}, \boldsymbol{\psi})^T$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^T$ comprises the mixture weights and $\boldsymbol{\psi}$ is the vector containing all the remaining parameters $\boldsymbol{\beta}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N, \boldsymbol{D}, \sigma^2$. In the following the order of $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ is determined by the corresponding weights in decreasing order under the restrictions $\sum_{h=1}^N \pi_h = 1$ and $\sum_{h=1}^N \pi_h \boldsymbol{\mu}_h = \mathbf{0}$. The latter ensures $\mathbb{E}(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$. The cluster membership of each individual can be described by latent variables $\boldsymbol{w}_i := (w_{i1}, \ldots, w_{iN})^T$, where $w_{ih} = 1$ if subject i belongs to cluster h and 0 otherwise. Marginalization over the random effects yields the complete model with observed data $\boldsymbol{y}_i$ as well as unobserved data $\boldsymbol{w}_i$:

$$\begin{array}{rcll} \boldsymbol{y}_i|\boldsymbol{w}_i & \overset{ind.}{\sim} & N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\mu}_h,\, \boldsymbol{V}_i), & i = 1, \ldots, n, \\ \boldsymbol{w}_i & \overset{i.i.d.}{\sim} & M(1, \boldsymbol{\pi}), & i = 1, \ldots, n, \end{array} \tag{3.4}$$

with $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \sigma^2\boldsymbol{I}_{n_i}$ and $M(\cdot, \cdot)$ representing the multinomial distribution. The likelihood function corresponding to (3.4) is given by

$$L(\boldsymbol{\xi}) = \prod_{i=1}^{n}\prod_{h=1}^{N}[\pi_h\, f_{ih}(\boldsymbol{y}_i; \boldsymbol{\psi})]^{w_{ih}},$$

where $f_{ih}(\cdot)$ denotes the density function of $N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\mu}_h,\, \boldsymbol{V}_i)$. The penalized log-likelihood we propose is

$$l_P(\boldsymbol{\xi}) = \sum_{i=1}^{n}\sum_{h=1}^{N} w_{ih}[\log\pi_h + \log f_{ih}(\boldsymbol{y}_i; \boldsymbol{\psi})] - \lambda\sqrt{N \cdot q}\sum_{h<l}\|\boldsymbol{\mu}_h - \boldsymbol{\mu}_l\|, \tag{3.5}$$

where λ indicates the penalty parameter. Obviously for $\lambda = 0$ the penalization term drops out. We will use an EM algorithm procedure, which alternates between taking the expectation of $l_P(\boldsymbol{\xi})$ over all unobserved w_{ih} in the E-step and maximization of the expected value in the M-step instead of directly maximizing the penalized incomplete likelihood function based only on the observed data. The steps have the following form.

E-step

Collecting all observed data in $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$, we obtain at iteration $t+1$ the E-step

$$Q(\boldsymbol{\xi}) = \mathbb{E}\left(l_P(\boldsymbol{\xi})|\boldsymbol{y}, \boldsymbol{\xi}^{(t)}\right) = \sum_{i=1}^{n}\sum_{h=1}^{N} \pi_{ih}(\boldsymbol{\xi}^{(t)})[\log \pi_h + \log f_{ih}(\boldsymbol{y}_i; \boldsymbol{\psi})] - \lambda\sqrt{N \cdot q} \sum_{h<l} \|\boldsymbol{\mu}_h - \boldsymbol{\mu}_l\|,$$

where $\pi_{ih}(\boldsymbol{\xi}^{(t)})$ is the probability at iteration t that subject i belongs to cluster h and is given by

$$\pi_{ih}(\boldsymbol{\xi}^{(t)}) = \frac{f_{ih}(\boldsymbol{y}_i; \boldsymbol{\psi}^{(t)})\pi_h^{(t)}}{\sum_{l=1}^{N} f_{il}(\boldsymbol{y}_i; \boldsymbol{\psi}^{(t)})\pi_l^{(t)}}.$$

M-step

For simplicity, we write $\pi_{ih} := \pi_{ih}(\boldsymbol{\xi}^{(t)})$, but it should be noted that for the M-step it is essential that π_{ih} is fixed from the last iteration t because then one can use that $Q(\boldsymbol{\xi})$ is the sum of two components, $Q(\boldsymbol{\pi})$ and $Q(\boldsymbol{\psi})$, and the optimization problem in the M-step can be separated into two parts: The maximization of

$$Q(\boldsymbol{\pi}) = \sum_{i=1}^{n}\sum_{h=1}^{N} \pi_{ih} \log \pi_h,$$

with respect to $\boldsymbol{\pi}$ and the maximization of

$$Q(\boldsymbol{\psi}) = \sum_{i=1}^{n}\sum_{h=1}^{N} \pi_{ih}\, \log f_{ih}(\boldsymbol{y}_i; \boldsymbol{\psi}) - \lambda\, \sqrt{N \cdot q} \sum_{h<l} \|\boldsymbol{\mu}_h - \boldsymbol{\mu}_l\|,$$

with respect to $\boldsymbol{\psi}$. The first optimization problem yields

$$\pi_h = \frac{1}{n}\sum_{i=1}^{n} \pi_{ih}, \qquad h = 1, \ldots, N.$$

In the second part of the M-step one obtains the current state for $\boldsymbol{\psi}$ by alternating between the maximization of $Q(\boldsymbol{\psi})$ with respect to $\boldsymbol{\beta}$, to $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ and to the variance parameters $\boldsymbol{D}$ and σ^2. Conditional on the current state of the other parameters the maximization of $\boldsymbol{\beta}$ results in

$$\boldsymbol{\beta} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\right)^{-1} \left(\sum_{i=1}^{n} \left(\boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{y}_i - \sum_{h=1}^{N} \pi_{ih} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{Z}_i \boldsymbol{\mu}_h\right)\right).$$

The corresponding proof is given in Appendix A.3.2. For the maximization of $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ given $\boldsymbol{\beta}$ and the variance parameters as well as for the maximization of the variance pa-

rameters given $\boldsymbol{\beta}$ and $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ a numerical procedure like the Nelder-Mead method is necessary. More information about this procedure is given in the paragraph "Implementation" below. In each M-step deviations from the constraint $\sum_{h=1}^{N} \pi_h \boldsymbol{\mu}_h = \mathbf{0}$ are subtracted from $\boldsymbol{\mu}_h$, $h = 1, \ldots, N$, and included in $\boldsymbol{\beta}$ so that this constraint holds. The second restriction $\sum_{h=1}^{N} \pi_h = 1$ is fulfilled by ensuring that the row sums of the matrix (π_{ih}) are one.

Start and stop of the algorithm

For EM algorithms it is essential how to choose the starting values because the (penalized) incomplete log-likelihood is ascending at each step and the algorithm can converge to a local maximum. Because in each M-step the fusion of clusters is investigated, it is sensible to choose starting values for an agglomerative clustering method. Therefore each subject starts in its own cluster. Thus, in the beginning there are $N = n$ clusters with weights $\pi_h = 1/N$, $h = 1, \ldots, N$. As starting values for the cluster locations $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ we consider the predicted random effects $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ of the previously fitted linear mixed model with Gaussian random effects distribution. This fit yields starting values for $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{D}$, too. To reduce computation time it is sometimes advisable to choose $N < n$ if the number of individuals is high. Then one obtains starting values for the cluster centers by a k-means clustering of predicted random effects of the former fitted linear mixed model. However, the algorithm starts with N clusters and successively merges clusters until there is no further ascent of the penalized incomplete log-likelihood. If two cluster centers $\boldsymbol{\mu}_h$ and $\boldsymbol{\mu}_l$ are fused, only one of these parameters is kept and the other one is deleted with the effect that the number of clusters N is reduced by one. In general, our penalized heterogeneity approach can be seen as an agglomerative cluster analysis but based on a regression model. After convergence we get the cluster membership by the matrix of estimated π_{ih}. Individual i is assigned to that cluster h for which $\hat{\pi}_{ih}$ is maximal. Based on the weights of all clusters the prediction of the random effects has the form

$$\hat{\boldsymbol{b}}_i = \hat{\boldsymbol{D}} \boldsymbol{Z}_i^T \hat{\boldsymbol{V}}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}) + (\boldsymbol{I}_q - \hat{\boldsymbol{D}} \boldsymbol{Z}_i^T \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{Z}_i) \sum_{h=1}^{N} \hat{\pi}_{ih} \hat{\boldsymbol{\mu}}_h, \quad i = 1, \ldots, n,$$

which is shown in Appendix A.4.

Implementation

All computations are implemented in C++ (Stroustrup, 1997), allowing for an efficient treatment of loop-intensive calculations, which is needed because of the slow convergence of the EM algorithm. They are made easily accessible by the function `lmmLASSO()` in the R package `clustmixed` (Heinzl, 2012) using the statistical software R (R Development Core Team, 2012). For computation, all variables are internally standardized, which is explained in more detail in Appendix A.5. For updating variance parameters we use the C++ library `ASA047` (Burkhardt, 2008), an implementation of the Nelder-Mead algorithm in C++, which was used by Papageorgiou and Hinde (2012) for similar tasks. For reflection,

extension and contraction coefficients we choose the common settings 1.0, 2.0 and 0.5 respectively. See Nelder and Mead (1965) and O'Neill (1971) for more technical details of the algorithm. Note that for ensuring that the covariance matrix $\boldsymbol{D}$ is nonnegative-definite, we parameterize the corresponding variance parameters by the entries of a lower triangular matrix $\boldsymbol{L}$ according to the Cholesky decomposition $\boldsymbol{D} = \boldsymbol{L}\boldsymbol{L}^T$. Then $\boldsymbol{D}$ is nonnegative-definite for each $\boldsymbol{L}$ and positive-definite (and so invertible, too) if $\boldsymbol{L}$ is a matrix with exclusively nonzero diagonal entries (Lindstrom and Bates, 1988).

Standard errors and confidence intervals

Users typically are interested in standard errors for the unknown parameters, in particular for the fixed effects and the variance parameters. Several strategies have been suggested in the EM literature for providing standard errors, especially information-based approaches and bootstrap methods (McLachlan et al., 2004). Basford et al. (1997) found in a comparative study that information-based approaches tend to be too unstable to be recommended in the case of normal mixture models. Therefore we will use bootstrap methods. In the applications in Section 3.3 we will use the nonparametric bootstrap method proposed by Efron (1979): Let F denote the true probability function of $\boldsymbol{y}_i$, $i = 1, \ldots, n$, and $\hat{F}$ the corresponding empirical probability distribution. Then we draw U random samples of size n from $\hat{F}$ with replacement. For each bootstrap replication $u = 1, \ldots, U$ the rth fixed effect, for example, is estimated by $\hat{\beta}_r^{(u)}$. Thus, the associated standard error can be estimated by

$$\widehat{se}(\hat{\beta}_r) = \sqrt{\frac{1}{U-1}\sum_{u=1}^{U}\left(\hat{\beta}_r^{(u)} - \overline{\hat{\beta}_r^{(u)}}\right)^2}. \tag{3.6}$$

Efron and Tibshirani (1993) showed that 50 to 100 bootstrap replications are generally sufficient for standard error estimation.

The bootstrap estimates can also be used for deriving confidence intervals. For example, the 95% confidence interval for β_r is given by the 0.025- and the 0.975-quantile of the empirical distribution of $\hat{\beta}_r^{(1)}, \ldots, \hat{\beta}_r^{(U)}$. According to Efron and Tibshirani (1993) the number of bootstrap replications should be 500 to 1000 for estimating confidence intervals based on bootstrap percentiles. The large number of bootstrap samples is necessary because the percentiles depend on the tails of the distribution where fewer samples occur. Investigating if covariates show a significant effect on the response variable, that is, testing $H_0 : \beta_r = 0$ against $H_1 : \beta_r \neq 0$, can be done by checking if the corresponding confidence intervals include zero. An example is given in Section 3.3.2. Alternatively the statistic $\hat{\beta}_r / \widehat{se}(\hat{\beta}_r)$ can be considered, which follows approximatively a standard normal distribution for a large sample size. Based on this test statistic also approximative confidence intervals can be developed. We pursue this strategy in the simulation studies.

3.2.2. Model Choice: Predictive Cross-Validation

In general, optimal penalization parameters can be chosen by cross-validation or information criteria such as Akaike information criterion (AIC) or Bayesian information criterion. In normal linear mixed models the AIC is not as straightforward as in normal linear models (compare Vaida and Blanchard (2005) and Greven and Kneib (2010)). For the penalized heterogeneity approach, the evaluation of the marginal or conditional AIC is even more complicated. Hence we prefer a cross-validation approach. Braun et al. (2012) introduced a new predictive cross-validation approach for model choice in linear mixed models with Gaussian random effects. The approach is based on the "mixed" cross-validation method proposed by Marshall and Spiegelhalter (2003). An advantage of this approach is that in contrast to full cross-validation the model must be fitted only once, which saves computing time. In general, each observed response value y_{obs} is compared to the corresponding predictive distribution, for example, by the continuous ranked probability score

$$CRPS(y_{obs}) = -\int_{-\infty}^{\infty} \left(P(Y_{obs} \leq r) - \mathbb{1}(y_{obs} \leq r)\right)^2 dr,$$

where P symbolizes the predictive distribution of the random variable Y_{obs} and $\mathbb{1}(x)$ denotes the indicator function that is one if the condition x is true and zero otherwise. If the predictive distribution is a normal distribution with estimated mean μ_{pre} and estimated standard deviation σ_{pre}, the continuous ranked probability score will take the form

$$CRPS(y_{obs}) = \hat{\sigma}_{pre}\left[\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{y_{obs} - \hat{\mu}_{pre}}{\hat{\sigma}_{pre}}\right) - \frac{y_{obs} - \hat{\mu}_{pre}}{\hat{\sigma}_{pre}}\left(2\Phi\left(\frac{y_{obs} - \hat{\mu}_{pre}}{\hat{\sigma}_{pre}}\right) - 1\right)\right]. \tag{3.7}$$

Here $\varphi(\cdot)$ denotes the density function and $\Phi(\cdot)$ the distribution function of the standard normal distribution. For linear mixed models Braun et al. (2012) consider the predictive distribution of the random variable y_{ij} conditional on the other given response values of the same subject $\boldsymbol{y}_{i,-j} := (y_{i1}, \ldots, y_{i,j-1}, y_{i,j+1}, \ldots, y_{in_i})^T$ for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. They argue that there is only small danger of conservatism due to ignoring the individual random effect as well as the real response value even though the model choice criterion is based on full data. When assuming normally distributed random effects one also obtains normal $y_{ij}|\boldsymbol{y}_{i,-j}$. Unfortunately in our case this distribution is not normal. Thus, we extend the approach of Braun et al. (2012) to our scenario. We exploit that in the case of known cluster membership the conditional distribution is normal. Because the cluster membership is not known the continuous ranked probability score is weighted by the estimated weights

$$WCRPS(y_{ij}) = \sum_{h=1}^{N} \hat{\pi}_h \, CRPS_h(y_{ij}),$$

where $CRPS_h(y_{ij})$ is given by formula (3.7) with $y_{obs} = y_{ij}$ and

$$\hat{\mu}_{pre} = \boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}^T\hat{\boldsymbol{\mu}}_h + \boldsymbol{z}_{ij}^T\hat{\boldsymbol{D}}\boldsymbol{Z}_{i,-j}^T\left(\hat{\sigma}^2\boldsymbol{I}_{n_i-1} + \boldsymbol{Z}_{i,-j}\hat{\boldsymbol{D}}\boldsymbol{Z}_{i,-j}^T\right)^{-1}(\boldsymbol{y}_{i,-j} - \boldsymbol{X}_{i,-j}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}_{i,-j}\hat{\boldsymbol{\mu}}_h),$$

$$\hat{\sigma}_{pre} = \left(\boldsymbol{z}_{ij}^T\hat{\boldsymbol{D}}\boldsymbol{z}_{ij} - \boldsymbol{z}_{ij}^T\hat{\boldsymbol{D}}\boldsymbol{Z}_{i,-j}^T\left(\hat{\sigma}^2\boldsymbol{I}_{n_i-1} + \boldsymbol{Z}_{i,-j}\hat{\boldsymbol{D}}\boldsymbol{Z}_{i,-j}^T\right)^{-1}\boldsymbol{Z}_{i,-j}\hat{\boldsymbol{D}}\boldsymbol{z}_{ij} + \hat{\sigma}^2\right)^{1/2}.$$

The parameters $\hat{\mu}_{pre}$ and $\hat{\sigma}_{pre}$ are the moments of the distribution of $y_{ij}|\boldsymbol{y}_{i,-j}, w_{ih} = 1$, which is proved in Appendix A.6. Here, $\boldsymbol{x}_{ij}^T$ is the jth row of $\boldsymbol{X}_i$ while $\boldsymbol{X}_{i,-j}$ symbolizes the matrix $\boldsymbol{X}_i$ without row j (analog for $\boldsymbol{z}_{ij}^T$ and $\boldsymbol{Z}_{i,-j}$). Finally, the mean of the weighted continuous ranked probability score is taken over all measurement points. The best value for the penalization parameter λ is the one maximizing the mean of the weighted continuous ranked probability score.

3.3. Applications

3.3.1. Unemployment

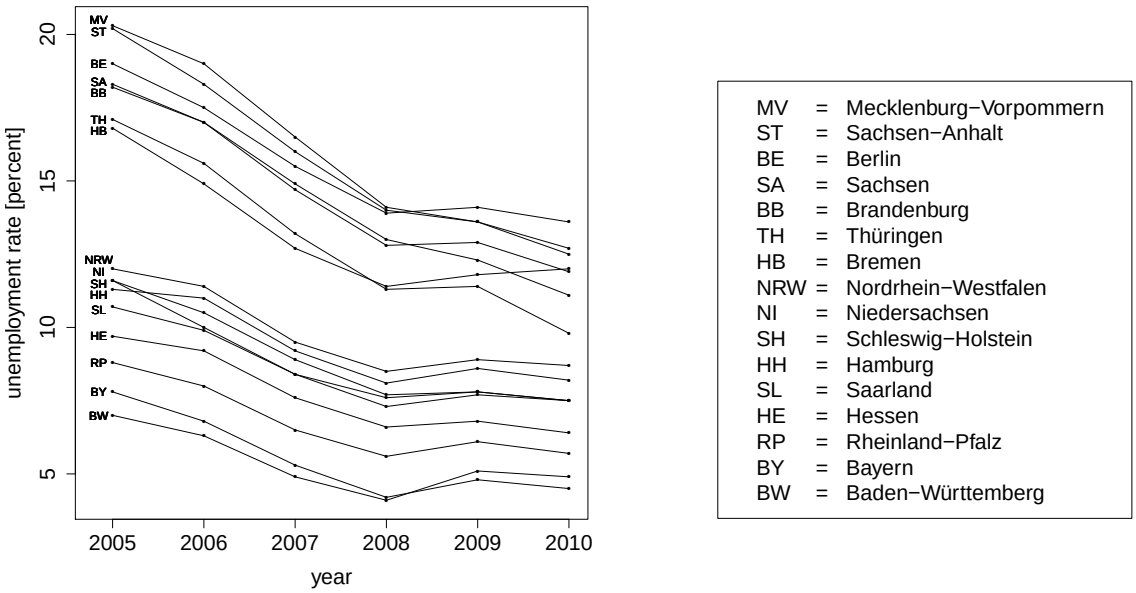


Figure 3.1.: Unemployment rates of the federal states of Germany across time.

We will illustrate the practical use of our model by considering the unemployment rates of the federal states of Germany from 2005 until 2010 (Weise et al., 2011). We aim at pointing out which states feature a similar development. Figure 3.1 shows different levels of the unemployment rates and a negative time trend, which can be regarded as approxi-

mately linear. Therefore we consider a random slope model for the annual average of the unemployment rate y_{ij} of state i in year j

$$y_{ij}|\boldsymbol{b}_i \overset{ind.}{\sim} N(\beta_0 + b_{i0} + (\beta_1 + b_{i1})\texttt{year}_{ij}, \sigma^2), \quad i = 1, \ldots, 16, \quad j = 0, \ldots, 5.$$

For the centered i.i.d. random effects $\boldsymbol{b}_i = (b_{i0}, b_{i1})^T$ we assume a mixture distribution of Gaussian components with penalized cluster centers (see Section 3.1). The covariance matrix in the random effects distribution (3.2) is denoted by

$$\boldsymbol{D} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}.$$

Note that only for a better interpretability the zero point of the time variable is changed to 2005. Thus, when computing estimates, the time variable is labeled by 0, 1, ..., 5 for the years 2005, 2006, ..., 2010. According to the considerations in Section 3.2.1 the algorithm starts with 16 clusters. Figure 3.2 suggests to choose the penalization parameter $\lambda = 0.01$. The resulting fit can be seen in Figure 3.3 and Table 3.1.

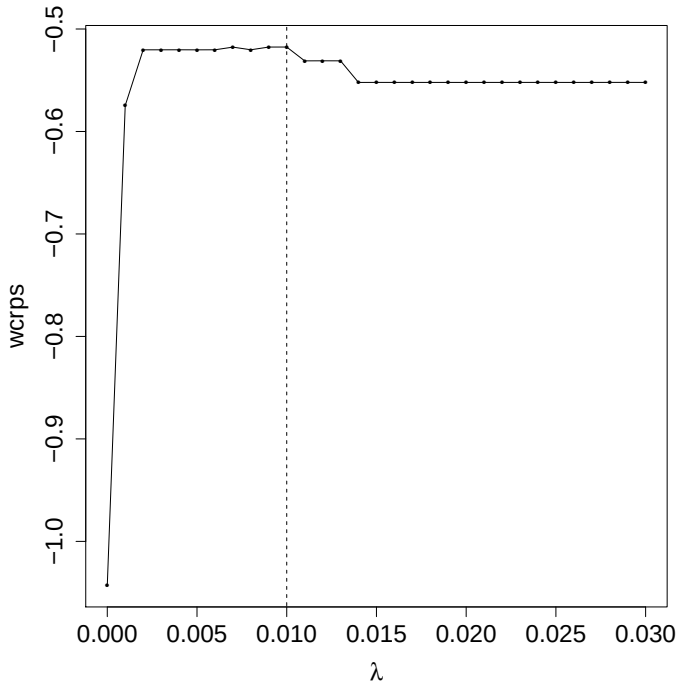


Figure 3.2.: Weighted continuous ranked probability score for the unemployment data depending on λ.

Table 3.1 shows the estimates as well as the estimated standard errors and the bootstrap confidence intervals based on 1000 bootstrap replications for the fixed effects and the variance parameters. The estimated global intercept $\hat{\beta}_0 = 13.315$ can be seen as the mean unemployment rate in the year 2005 and the global slope $\hat{\beta}_1 = -0.980$ is the mean decrease of unemployment in the years from 2005 until 2010. The standard errors are

	estimate	standard error	95%-CI lower	95%-CI upper
β_0	13.315	0.917	11.386	15.066
β_1	-0.980	0.076	-1.131	-0.819
σ^2	0.480	0.050	0.390	0.571
σ_0^2	2.299	0.770	0.000	2.601
σ_1^2	0.005	0.003	0.000	0.010
σ_{01}	-0.107	0.045	-0.149	0.009

Table 3.1.: Estimation results for the fixed effects and the variance parameters by the penalized heterogeneity approach with $\lambda = 0.01$ for the unemployment data.

relatively small, which promises precise estimations. They are even smaller than those in Section 4.4.1, where a competing approach is considered.

The clustering of the penalized heterogeneity approach is shown in Figure 3.3. Here, the dashed line symbolizes the population effect whereas the solid lines display the cluster centers. Observations belonging to the same cluster are marked with the same symbol. To each solid line the corresponding symbol is also added to visualize which cluster center belongs to which cluster. Three clusters are detected by our model: cluster 1 ($\bigcirc$) has weight $\hat{\pi}_1 = 0.563$ and is characterized by a comparably low unemployment. In 2005 the level is $\hat{\mu}_{10} = -3.702$ lower that the base level whereas the decrease $\hat{\mu}_{11} = 0.296$ is not so powerful than in the two other clusters. All western states are in cluster 1 with the exception of the city states Berlin and Bremen. These states form cluster 3 ($+$) with $\hat{\pi}_3 = 0.129$, $\hat{\mu}_{30} = 3.838$ and $\hat{\mu}_{31} = -0.111$. In Figure 3.1 it seems that Berlin and Bremen show a similar behavior like the eastern states in clusters 2 ($\triangle$, $\hat{\pi}_2 = 0.309$, $\hat{\mu}_{20} = 5.145$, $\hat{\mu}_{21} = -0.492$) but on closer inspection it can be seen that these states show a worse development of unemployment than the states in cluster 2 and slipped some notches in the ranking. The fit of our model highlights this feature.

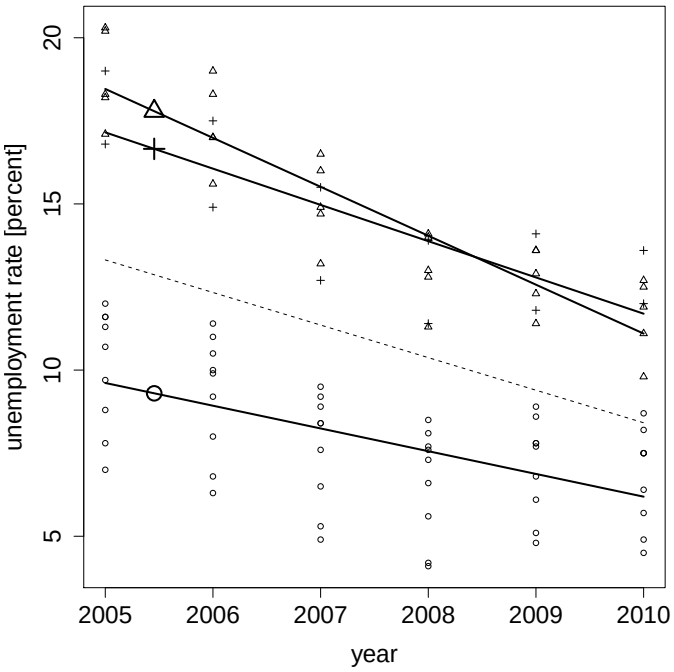


Figure 3.3.: Clustering of the unemployment data by the penalized heterogeneity approach with $\lambda = 0.01$. Observations belonging to the same cluster are marked with the same symbol. The dashed line represents the population effect, the solid lines symbolize the cluster effects.

3.3.2. Hormonotherapy

As another example the craniofacial growth of male rats is analyzed by the penalized heterogeneity model. The data were collected in an experiment at the Catholic University of Leuven with the aim to analyze the effect of testosterone on the growth of rats (Verdonck et al., 1998). Therefore 50 male rats have been randomized to either a control group or to one of the two treatment groups that differ in the dose of the drug Decapeptyl, which inhibits the testosterone production. The response of interest is the distance (in pixels) between well-defined points of the skull that characterize the height of skull. These heights were measured for each rat every 10 days starting at the age of 50 days and the treatment began at the age of 45 days. See Verbeke and Molenberghs (2000) for more information about the data. Figure 3.4 shows different levels of heights of the skulls and a positive time trend, which varies from rat to rat. According to Figure 3.5, where the heights of rat skulls across age are shown for each treatment group separately, there seems to be a negative effect of the drug Decapeptyl on the growth of rats, but the three groups are relatively mixed and cannot be clearly separated.

To examine how many and which clusters can be found in these data the penalized heterogeneity approach with a group fused lasso penalty is used. As suggested by Verbeke and Lesaffre (1999), and also done by Verbeke and Molenberghs (2000) and Fahrmeir et al. (2007), the age of rat i at measurement j is transformed by $t_{ij} = \log(1 + \text{age})_{ij}$ to get

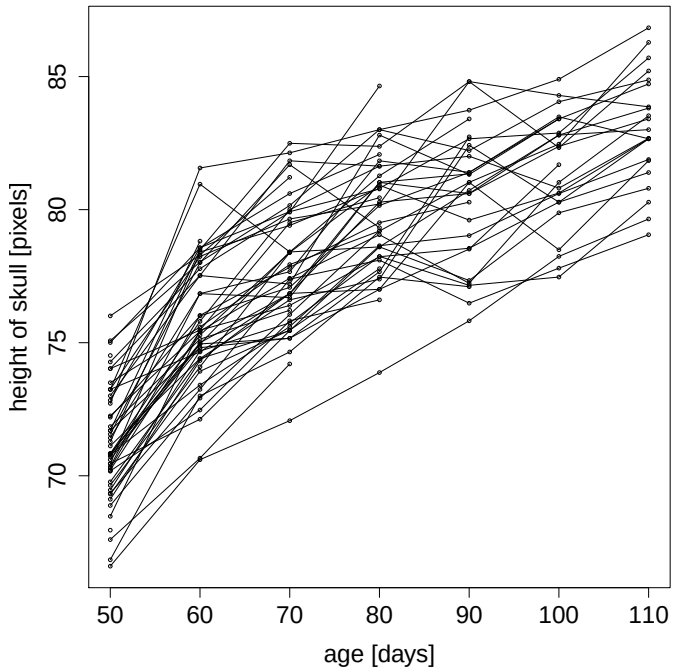
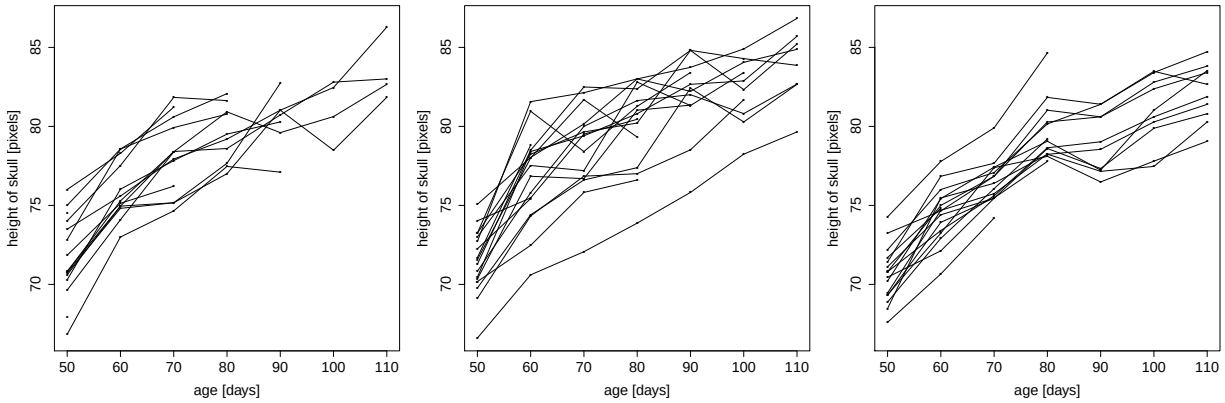


Figure 3.4.: Heights of rat skulls across age.

a linear time trend. In analogy to Verbeke and Molenberghs (2000) and Fahrmeir et al. (2007) the time trends in each group are modeled as fixed effects and random intercepts and slopes are included to incorporate individual deviations of the time trend. In summary, we consider the following model for the height y of the skull of rat i at measurement j

$$y_{ij}|\boldsymbol{b}_i \overset{ind.}{\sim} N(\beta_0 + b_{i0} + (\beta_1 + \beta_2 L_i + \beta_3 H_i + b_{i1})t_{ij},\, \sigma^2), \quad i = 1, \ldots, 50, \quad j = 1, \ldots, n_i,$$

with effect-coded variables L and H for a low and high dose of drug, respectively. For the centered i.i.d. random effects $\boldsymbol{b}_i = (b_{i0}, b_{i1})^T$ we assume a mixture distribution of Gaussian components with penalized cluster centers (see Section 3.1). The four rats for which only one measurement was available were excluded because for these no reasonable random slope can be predicted. For faster computations the algorithm starts with 20 clusters. Figure 3.6 suggests to choose the penalization parameter $\lambda = 0.011$. This yields three clusters as it can be seen in the resulting fit in Figure 3.7. For larger values than $\lambda = 0.0115$ the estimated number of clusters would be two, for much larger values of λ only one cluster would be detected. The large jumps in Figure 3.6 are seen when the number of clusters changes. Otherwise there are only small, but not negligible differences in the weighted continuous ranked probability score. Due to the large scale these small differences are hard to see.



(a) Control group (b) Low dose (c) High dose

Figure 3.5.: Heights of rat skulls across age separately for each treatment group.

	estimate	standard error	95%-CI lower	95%-CI upper
β_0	68.658	0.312	68.044	69.263
β_1	7.248	0.149	6.964	7.543
β_2	0.082	0.311	-0.442	0.738
β_3	-0.459	0.296	-0.879	0.249
σ^2	1.425	0.145	1.087	1.654
σ_0^2	0.282	0.867	0.000	3.122
σ_1^2	0.019	0.037	0.000	0.138
σ_{01}	0.073	0.118	-0.349	0.125

Table 3.2.: Estimation results for fixed effects and variance parameters by the penalized heterogeneity approach with $\lambda = 0.011$ for the hormonotherapy data.

In Table 3.2 the estimates, the standard errors, and the confidence intervals for the fixed effects and the variance parameters are given for $\lambda = 0.011$. The standard errors and confidence intervals have been estimated by 1000 bootstrap replications, see Section 3.2.1 for more details. The estimated global intercept $\hat{\beta}_0 = 68.658$ can be interpreted as the mean height at the beginning of the treatment. Since the covariates L and H are effect-coded, the global slope $\hat{\beta}_1 = 7.248$ represents the overall mean growth of all rat skulls in the considered time period. The expected negative effect of the drug Decapeptyl can be seen from the estimates $\hat{\beta}_2 = 0.082$ and $\hat{\beta}_3 = -0.459$, which can be interpreted as deviations from the overall time trend. For rats which had been exposed to a high dose of the drug the growth ($\hat{\beta}_3 = -0.459$) is considerably slower than in low dose group ($\hat{\beta}_2 = 0.082$), differences to the control group are even bigger ($-\hat{\beta}_2 - \hat{\beta}_3 = 0.376$). These results are more intuitive than the results obtained by Verbeke and Molenberghs (2000). In their analysis the rats which had been exposed to a low dose show a higher growth than those in the control group, though the drug has a negative effect on the growth for a high dose. It

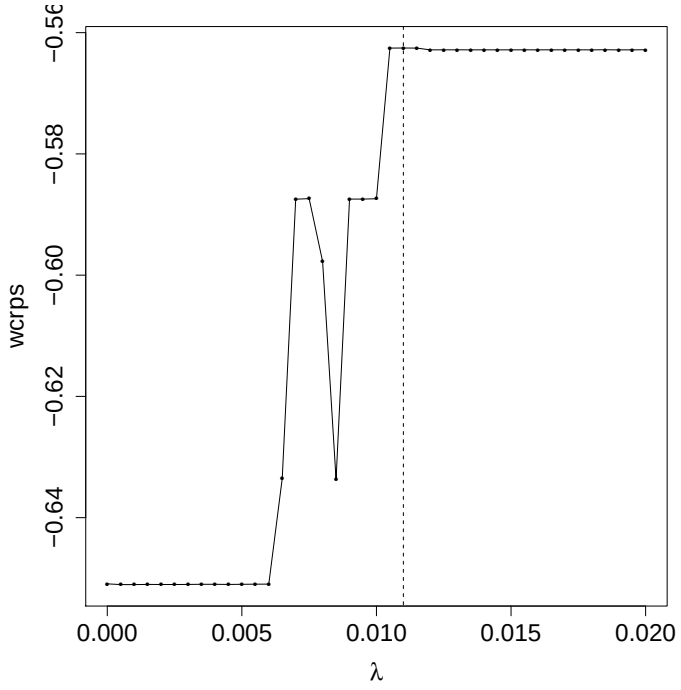


Figure 3.6.: Weighted continuous ranked probability score for the hormonotherapy data depending on λ.

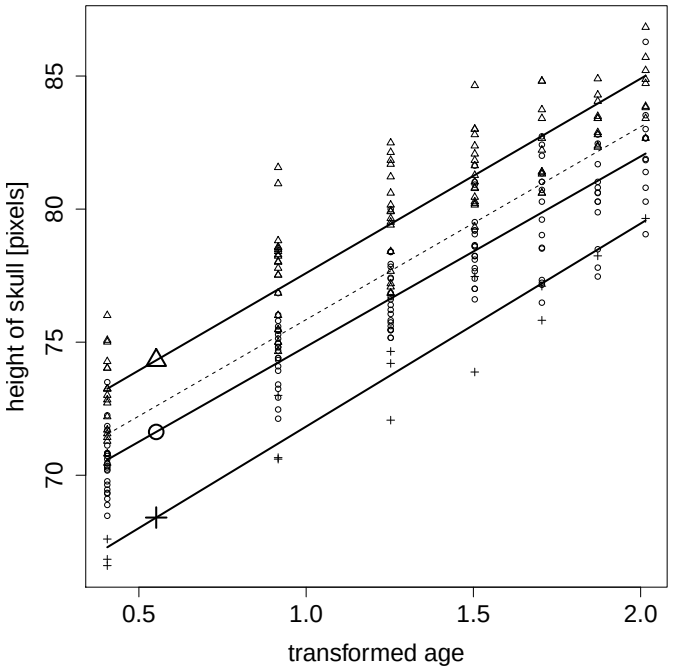
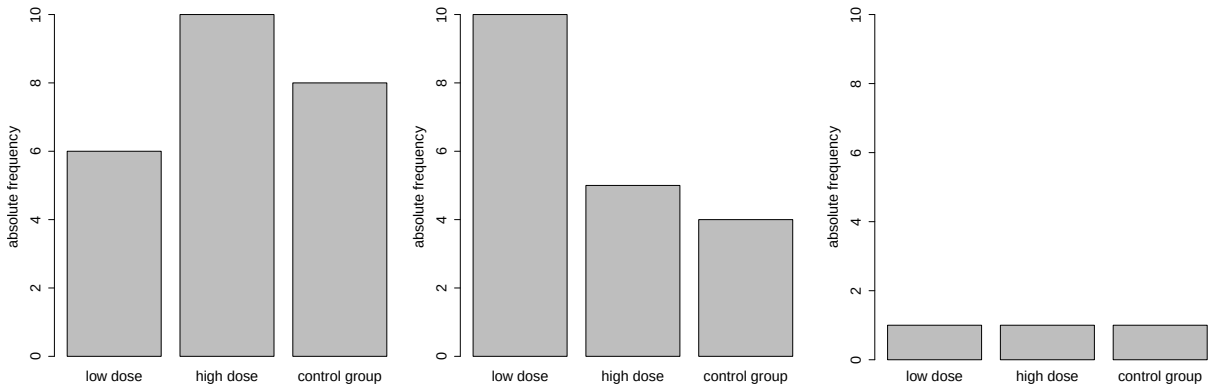


Figure 3.7.: Clustering of the hormonotherapy data by the penalized heterogeneity approach with $\lambda = 0.011$. Observations belonging to the same cluster are marked with the same symbol. The dashed line represents the population effect, the solid lines symbolize the cluster effects.

seems that the penalized mixture of normal distributions as random effects distribution is much more adequate than a simple normal distribution for these data with an underlying grouping structure. However, from the confidence intervals for the fixed effects we can see that the interactions between *time* and L respectively H have no significant effect on the height on the five percent significance level while the variable *time* itself shows a significant effect. The estimated standard errors are quite similar to those of Verbeke and Molenberghs (2000) and Fahrmeir et al. (2007). For example, the standard error for the intercept (0.312) is only a bit smaller in the penalized heterogeneity approach than in the models by Verbeke and Molenberghs (2000) (0.325) and Fahrmeir et al. (2007) (0.338). For the variance parameters, differences are much larger. The variance of the random intercept is considerably smaller (0.282) than in the reference approaches: 3.369 (Verbeke and Molenberghs, 2000) and 3.739 (Fahrmeir et al., 2007) since the variation of the data is for the most part incorporated by the penalized normal mixture for the random effects.



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

Figure 3.8.: Distribution of the two treatment groups respectively the control group in the three clusters corresponding to the penalized heterogeneity approach with $\lambda = 0.011$.

According to Figure 3.7 three clusters are detected. The dashed line symbolizes the population effect whereas the solid lines display the cluster centers. Observations belonging to the same cluster are marked with the same symbol. To each solid line the corresponding symbol is added to visualize which cluster center belongs to which cluster. While there are only low discrepancies in the random slopes ($\hat{\mu}_{11} = -0.100$, $\hat{\mu}_{21} = 0.061$, $\hat{\mu}_{31} = 0.382$), the base levels are quite different. Cluster 2 ($\triangle$) with weight $\hat{\pi}_2 = 0.435$ has the largest intercept, which is about $\hat{\mu}_{20} = 1.706$ larger than the overall intercept. By comparison, in Cluster 1 ($\bigcirc$) with $\hat{\pi}_1 = 0.503$ the base level is considerably lower ($\hat{\mu}_{10} = -0.912$). Cluster 3 ($+$) has the estimated weight $\hat{\pi}_3 = 0.062$ and contains the three rats with the lowest base level ($\hat{\mu}_{30} = -4.578$). As can be seen from Figure 3.8 the response types collected in the clusters come from all groups. In cluster 1 rats of the high dose group are in the majority followed by rats of the control group. In cluster 2 in particular rats which had been exposed to a low dose of the drug are found.

3.3.3. Lung Function Growth

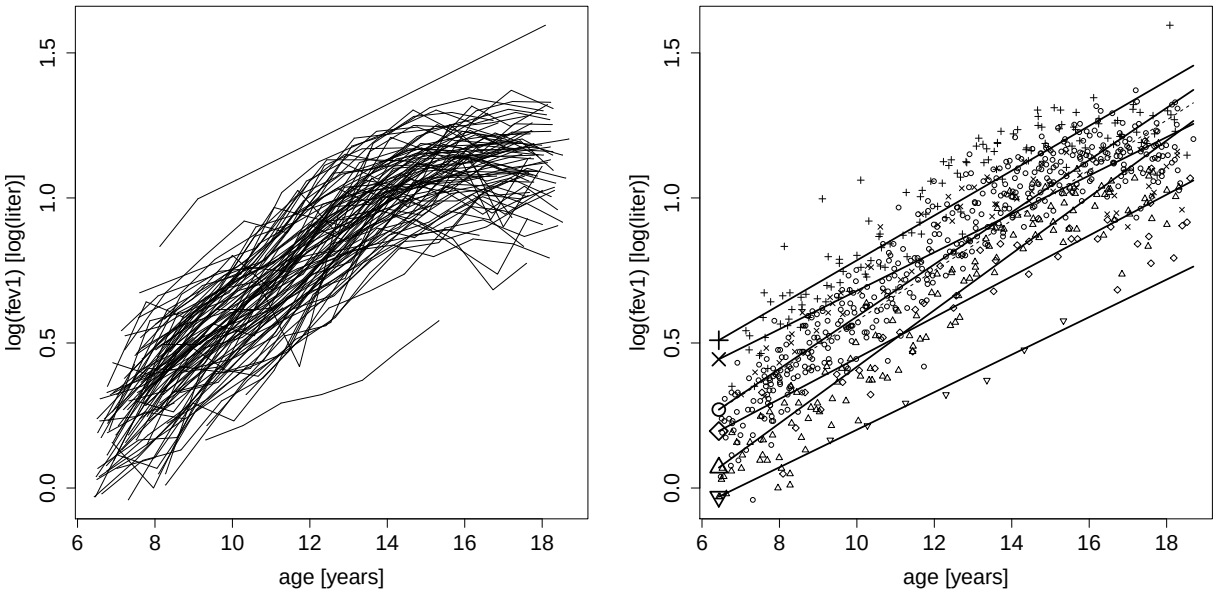


Figure 3.9.: Logarithmic forced expiratory volume in one second of girls across age: raw data (left) and clustering by the penalized heterogeneity approach with $\lambda = 0.0175$ (right). On the right observations belonging to the same cluster are marked with the same symbol. The dashed line represents the population effect, the solid lines symbolize the cluster effects.

The third data example deals with lung function growth of girls in Topeka (USA). These data are a subsample from the six cities study of air pollution and health in Dockery et al. (1983). In this study a cohort of 13,379 children born in or after 1967 was enrolled in six communities in the United States: Watertown (Massachusetts), Kingston and Harriman (Tennessee), a section of St. Louis (Missouri), Steubenville (Ohio), Portage (Wisconsin), and Topeka (Kansas). The study aims at characterizing the lung function growth of the children. One important indicator for the pulmonary function is the logarithmic forced expiratory volume in one second (`fev1`), i.e. the quantity of air a person breathes out in one second as fast and powerful as possible. Our sample consists of 100 girls, with a minimum of two and a maximum of twelve observations over time. Although a cluster structure is not evident from looking at the raw data (Figure 3.9, left) our approach is able to identify clusters in the data. Again we consider a random slope model

$$\texttt{log(fev1)}_{ij} | \boldsymbol{b}_i \overset{ind.}{\sim} N(\beta_0 + b_{i0} + (\beta_1 + b_{i1})\texttt{age}_{ij}, \sigma^2), \quad i = 1, \ldots, 100, \quad j = 1, \ldots, n_i,$$

for modeling the logarithm of `fev1` subject to `age` and use a finite mixture as random effects distribution with a group fused lasso penalty. Because of the comparably large number of individuals we start with $N = 30$ clusters instead of 100.

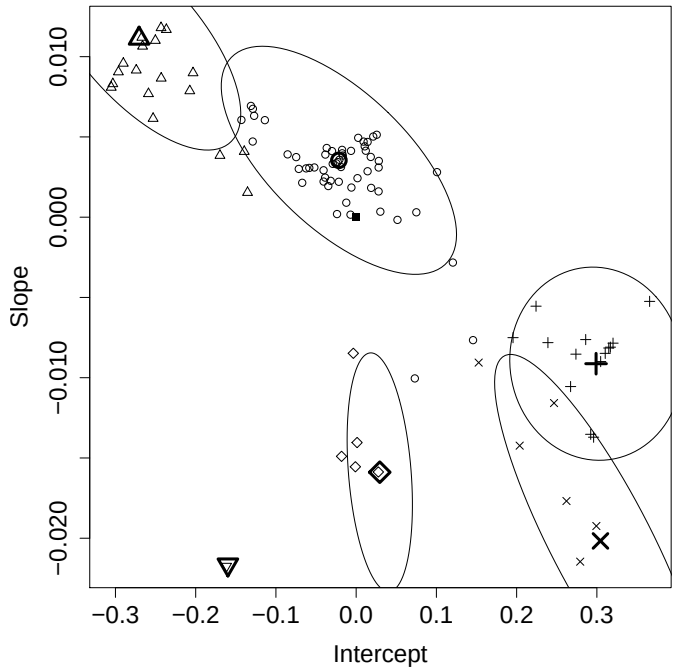


Figure 3.10.: Cluster centers and random effects of the penalized heterogeneity approach with $\lambda = 0.0175$ for the lung function growth data: The thick big icons symbolize the cluster centers $\hat{\boldsymbol{\mu}}_h$, the thin small ones the random effects $\hat{\boldsymbol{b}}_i$. The square at coordinates (0,0) marks the population effect. Ellipses with level 0.95 visualize the estimated conditional distribution of random effects in the clusters.

In Figure 3.9 (right) the clustering structure is visualized but it is hard to see which girls are merged to the same cluster. Figure 3.10 makes clear how the clustering works. On the axes the intercepts and the slopes are drawn. The filled square at coordinates (0,0) symbolizes the population effect. All other icons represent deviations from the population effect. The big bold ones represent the cluster locations $\hat{\boldsymbol{\mu}}_h$ and the thin small ones the random effects $\hat{\boldsymbol{b}}_i$. Girls that are assigned to the same cluster are marked with the same symbol and are arranged around the cluster locations in the form of ellipses. It is easily seen that subjects with random effects that are similar in terms of a low Euclidean distance belong to the same cluster. For visualizing the variation of the random effects around the cluster centers ellipses are added in Figure 3.10. The ellipses are constructed as follows: If it was known that an arbitrary subject i belongs to cluster h, the conditional distribution of random effect $\boldsymbol{b}_i$ would be given by $\boldsymbol{b}_i | w_{ih} = 1 \sim N(\boldsymbol{\mu}_h, \boldsymbol{D})$. However, the cluster membership is not known in practice. Thus, this conditional distribution has to be estimated from the data. While the estimate for the mean is directly given from the estimation results of the model, the covariance matrix in cluster h is estimated on the basis of the individuals with

$$\arg\max_{l=1,\ldots,N} \hat{\pi}_{il} = h.$$

3.4. Simulation Study

In the following simulation study the performance of our penalized heterogeneity approach is evaluated. The study aims at clarifying in which data situations our approach improves estimation compared to the commonly used linear mixed model with Gaussian random effects distribution and the heterogeneity model by Verbeke and Lesaffre (1996). Note that the estimated number of clusters and the estimated clustering in general have an essential impact on the prediction accuracy of the random effects. Of course, for the prediction of $\boldsymbol{b}_i$ it is reasonable to borrow information from other subjects which show a similar behavior and so belong to the same cluster while incorporating dissimilar individuals impairs the prediction accuracy. For examining this trade-off we compare the usual linear mixed model with normal random effects distribution (one cluster model) using the R function `lmer()` from the `lme4` package by Bates et al. (2012) to our penalized heterogeneity approach with the penalization parameter λ being determined by predictive cross-validation (see Section 3.2.2). In addition, the heterogeneity model by Verbeke and Lesaffre (1996) with a finite unpenalized mixture of normal distributions as random effects distribution is considered. In the latter approach, the number of mixture components is identified by the predictive cross-validation criterion, too. In addition to an elaborate examination of the prediction accuracy of the random effects and the goodness of the estimates for the fixed effects and variance parameters in Section 3.4.2 two further aspects are investigated: In Section 3.4.3 the impact of the considered methods on characteristics of tests concerning the significance of covariates is evaluated while Section 3.4.4 deals with the cause and the meaning of heterogeneity in the random effects distribution.

3.4.1. Settings

In the following simulation study we investigate the impact of the number of observations per unit and the separation between clusters. We generated data sets assuming a simple linear trend model

$$y_{ij} = \beta_0 + b_{i0} + (\beta_1 + b_{i1})t_{ij} + \beta_2 x_i + \varepsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i, \tag{3.8}$$

with i.i.d. errors $\varepsilon_{ij} \sim N(0, \sigma^2)$. The values x_i are generated from a standard normal distribution. The centered i.i.d. random effects $\boldsymbol{b}_i = (b_{i0}, b_{i1})^T$ follow a mixture distribution with three Gaussian components:

$$\boldsymbol{b}_i \sim 0.4 \cdot N(\boldsymbol{\mu}_1, \boldsymbol{D}) + 0.3 \cdot N(\boldsymbol{\mu}_2, \boldsymbol{D}) + 0.3 \cdot N(\boldsymbol{\mu}_3, \boldsymbol{D}), \quad i = 1, \ldots, n,$$

imitating a population consisting of three clusters of overlapping subpopulations. Throughout the simulations, we set $n = 20$ and

$$\sigma^2 = 0.25, \qquad \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 0.2 \\ 0.1 \\ 0.05 \end{pmatrix}, \qquad \boldsymbol{D} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix} = \begin{pmatrix} 0.02 & 0.01 \\ 0.01 & 0.02 \end{pmatrix}.$$

We vary, however, the number of individual observations n_i, the centers $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ of the clusters and the locations of observation times t_{ij}. To produce longitudinal data with varying numbers of repeated observations per unit i, we set $n_i = 2 + X_i$, where X_i follows a Poisson distribution with rate ν. Setting $\nu = 1$ corresponds to longitudinal data with only *few individual observations* (3 on average), $\nu = 3$ to a *medium number of individual observations* and $\nu = 5$ to comparably *many individual observations*. For given n_i, observation times are generated from

$$\begin{aligned} t_{i1} &\sim U(0,1), \quad i = 1,\ldots,n, \\ t_{ij} &\sim U(t_{i,j-1} + 0.5, t_{i,j-1} + 1.5), \quad i = 1,\ldots,n, \quad j = 2,\ldots,n_i, \end{aligned}$$

where $U(\cdot,\cdot)$ denotes the uniform distribution. In this way, different numbers $n_i(s)$ and measuring times $t_{ij}(s)$ are generated in each simulation run $s = 1,\ldots,100$. Similarly, different "true" random effects $\boldsymbol{b}_i(s)$ are drawn from the Gaussian mixture distribution in each simulation run. For the cluster locations, we chose

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -2.25 \\ 1 \end{pmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 0.75 \\ -1.2 \end{pmatrix}, \qquad \boldsymbol{\mu}_3 = \begin{pmatrix} 2.25 \\ -2/15 \end{pmatrix},$$

corresponding to *clearly separated clusters*,

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -1.5 \\ 0.75 \end{pmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 0.5 \\ -0.9 \end{pmatrix}, \qquad \boldsymbol{\mu}_3 = \begin{pmatrix} 1.5 \\ -0.1 \end{pmatrix},$$

corresponding to *moderately separated clusters*, and

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

corresponding to *only one cluster*.

Combining these different settings for observation times and clusters results in nine different scenarios. For each of them, the prediction accuracy of the random effects as well as the estimation results of the fixed effects and the variance parameters are compared for all considered models. More concretely, in each simulation run s, we calculate the average prediction error

$$PE_r(s) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{b}^*_{ir}(s) - b^*_{ir}(s)\right)^2, \quad r = 0, 1,$$

for uncentered random intercepts $b^*_{i0} = \beta_0 + b_{i0}$ and random slopes $b^*_{i1} = \beta_1 + b_{i1}$. In addition, the estimation accuracy of the fixed effects is investigated by the estimated mean squared errors $\widehat{MSE}_r = \widehat{Var}(\hat{\beta}_r) + (\overline{\hat{\beta}}_r - \beta_r)^2$ and the medians of the relative biases $RB_r = (\hat{\beta}_r - \beta_r)/\beta_r$, $r = 0, 1, 2$. The standard errors for the fixed effects are estimated by the standard deviations of the estimates $\hat{\beta}_r$ from 100 simulation runs using formula (3.6).

Furthermore, we examine the estimated variance parameters, especially the variances of the random effects $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$.

3.4.2. Results

The results for the nine combinations are summarized below. For all scenarios we illustrate the empirical distribution of $PE_0(s)$ values obtained from simulation run $s = 1, \ldots, 100$ by box plots. The corresponding figures of the random slopes are not shown because these are very similar to those of the random intercepts. Tables show the estimation results of the fixed effects, random effects and variance parameters. In addition, we demonstrate the clustering related characteristics.

Clearly separated clusters

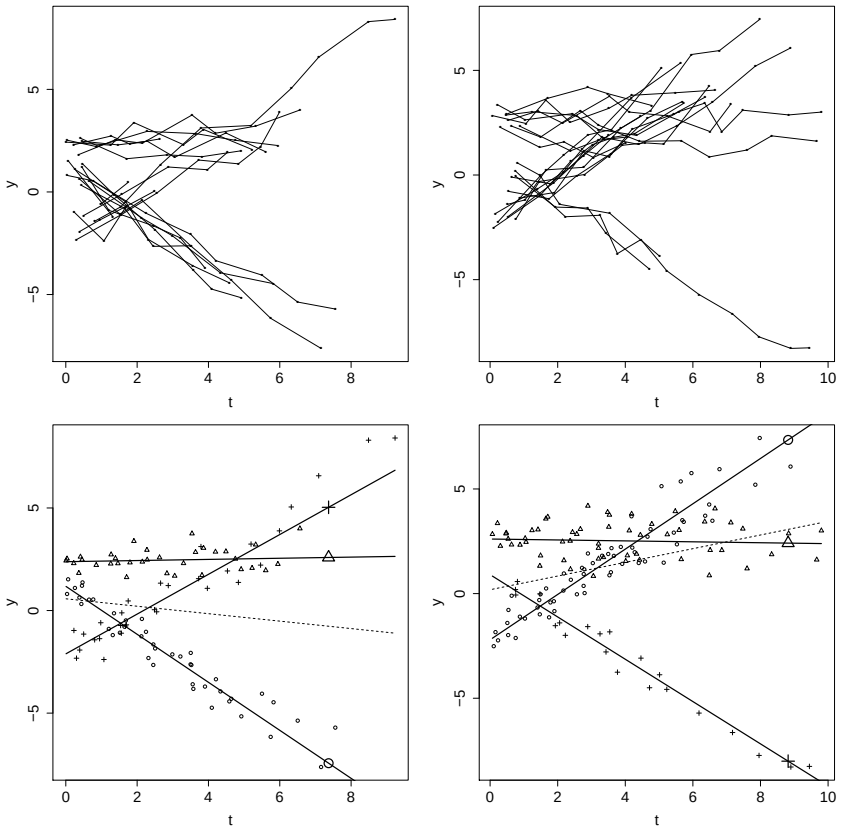
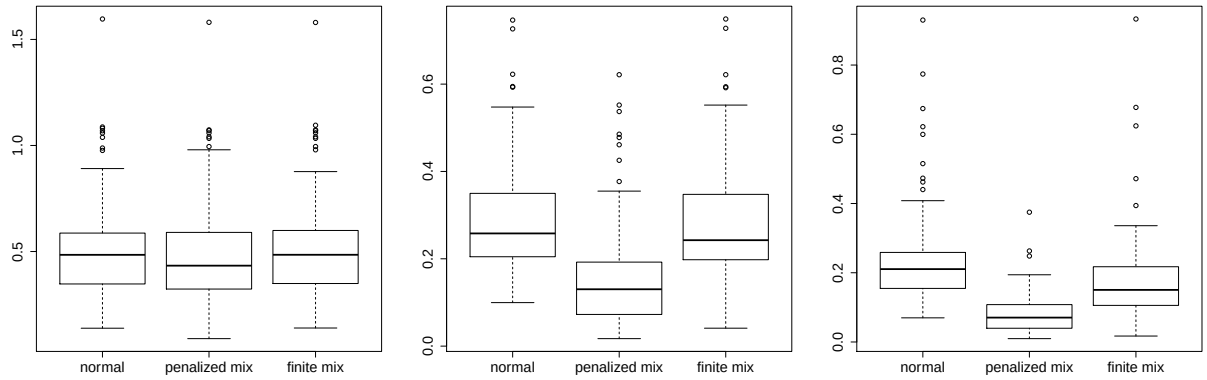


Figure 3.11.: Trace plots (top) and clustering by the penalized heterogeneity approach (below) with clearly separated clusters for a medium number of individual observations ($\nu = 3$) (left) and many individual observations ($\nu = 5$) (right).

Figure 3.11 (top) displays trace plots of typical longitudinal data generated in the setting of clearly separated clusters. Cluster effects can easily be seen. On the left, there is

a medium number of observations for each subject while on the right the mean of the number of repeated measurements is seven resulting in more observation times. Figure 3.11 (bottom) demonstrates that in both cases the penalized heterogeneity approach detects three clusters. Again, in this type of plot the dashed line shows the overall effect, and the solid lines visualize the means of the resulting clusters. The assignment to clusters is visualized by differing symbols.



(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

Figure 3.12.: Box plots of PE_0 with clearly separated clusters for few individual observations (left), a medium number of individual observations (middle) and many individual observations (right).

Table 3.3 and Figure 3.12 show the simulation results in the setting of clearly separated clusters. The denotation "normal" labels the homogeneity model with normally distributed random effects. In the heterogeneity model the random effects follow a "finite mixture" as specified in equation (3.2), where the number of mixture components has been determined by predictive cross-validation. In contrast to this discrete optimization the approach proposed in this paper uses the penalty term (3.3) multiplied by a tuning parameter, which is also determined by predictive cross-validation. In Figure 3.12 it can be seen that the penalization approach outperforms the homogeneity model and the heterogeneity model for few observations as well as for a medium number of individual observations and many observations. It is especially remarkable that the "penalized mixture" yields a better prediction accuracy than the "finite mixture" although in both cases the same criterion for finding the best number of clusters is used. The reason for that is that for optimization in our penalized heterogeneity approach a denser grid is used. This is the main justification for our model. Apart from that it can be seen that the more repeated measurements per unit are given the better is the prediction accuracy of the penalized heterogeneity approach.

Table 3.3 shows several features. First, it is seen that for $\hat{\beta}_0$ and $\hat{\beta}_1$ the mean squared errors, the relative biases, and the standard errors tend to be a bit smaller for the penalized heterogeneity model. Second, huge differences can be seen for the estimation results for $\hat{\beta}_2$. In particular, the mean squared errors and the standard errors are considerably smaller for the penalized heterogeneity model, especially in the case of a medium number of individual

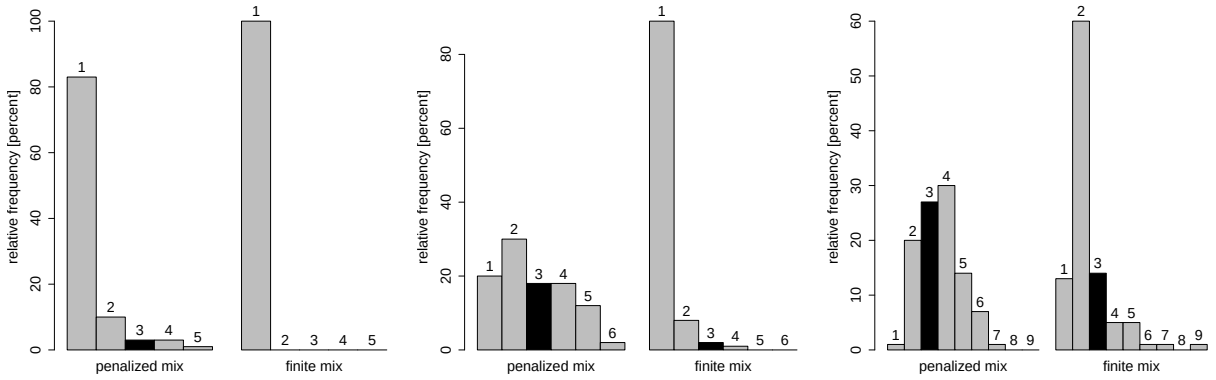
observations ($\nu = 3$) and many individual observations ($\nu = 5$) due to a general variance reduction. This is seen mainly in the estimates for σ_0^2 and σ_1^2, which are clearly smaller for the penalized heterogeneity model. The reason for these small variances is that the heterogeneity in the data is partially accounted for by the penalized mixture. On the other hand, we can see that in the linear mixed model with normally distributed random effects the true variances $\sigma_0^2 = 0.02$ and $\sigma_1^2 = 0.02$ are overestimated.

			$\widehat{MSE}_r$	RB_r	$\widehat{se}(\hat{\beta}_r)$	PE_r	$\hat{\sigma}_r^2$
$\nu = 1$	$r = 0$	normal	0.219	-0.147	0.467	0.484	3.779
		penalized mix	**0.209**	**-0.045**	**0.456**	**0.433**	**3.432**
		finite mix	0.219	-0.156	0.467	0.485	3.437
	$r = 1$	normal	**0.054**	-0.073	0.232	0.186	0.842
		penalized mix	**0.054**	**-0.027**	**0.231**	**0.169**	**0.769**
		finite mix	0.055	-0.075	0.233	0.187	0.793
	$r = 2$	normal	0.106	1.032	0.322		
		penalized mix	**0.097**	**0.801**	**0.308**		
		finite mix	0.106	1.006	0.322		
$\nu = 3$	$r = 0$	normal	0.255	-0.072	0.504	0.258	3.787
		penalized mix	**0.253**	-0.094	**0.502**	**0.130**	**1.246**
		finite mix	0.259	**-0.069**	0.508	0.243	3.373
	$r = 1$	normal	0.049	0.154	0.222	0.044	0.867
		penalized mix	**0.048**	**0.143**	**0.219**	**0.027**	**0.611**
		finite mix	0.050	0.146	0.224	0.044	0.812
	$r = 2$	normal	0.104	-0.422	0.322		
		penalized mix	**0.034**	0.308	**0.185**		
		finite mix	0.097	**-0.014**	0.312		
$\nu = 5$	$r = 0$	normal	**0.210**	**0.076**	**0.459**	0.211	3.661
		penalized mix	0.212	0.150	0.460	**0.070**	**0.011**
		finite mix	0.212	0.180	0.460	0.150	2.948
	$r = 1$	normal	0.041	0.478	0.202	0.015	0.841
		penalized mix	**0.040**	**0.415**	0.201	**0.007**	**0.020**
		finite mix	**0.040**	0.540	**0.199**	0.012	0.211
	$r = 2$	normal	0.112	-0.794	0.333		
		penalized mix	**0.012**	**-0.057**	**0.111**		
		finite mix	0.048	-0.162	0.218		

Table 3.3.: For clearly separated clusters the estimated mean squared errors MSE_r, the medians of the relative biases RB_r and the estimated standard errors $se(\hat{\beta}_r)$ for the fixed effects are shown. In addition, the medians of PE_r and $\hat{\sigma}_r^2$ are given. Bold values indicate the best value in each case.

In Figure 3.13 the estimated number of clusters of the mixture models are seen. Obviously the penalized mixture model tends to detect more clusters than the finite mixture model. The larger the number of repeated measurements per unit the higher is the estimated number of clusters. In Figure 3.13 the bar corresponding to three clusters is

highlighted by black color because in the simulation setting three clusters are used. As it could expected, the number of clusters is hard to identify, in particular in the case of few repeated observations since not enough information is available. Here, it can be seen that three individual observations on average are not enough to discriminate between possible clusters. In this case it is hard to determine whether the membership to different clusters or random deviations are responsible for different time trends. For many observations the performance of the penalized heterogeneity approach is much better and outperforms the finite mixture approach.



(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

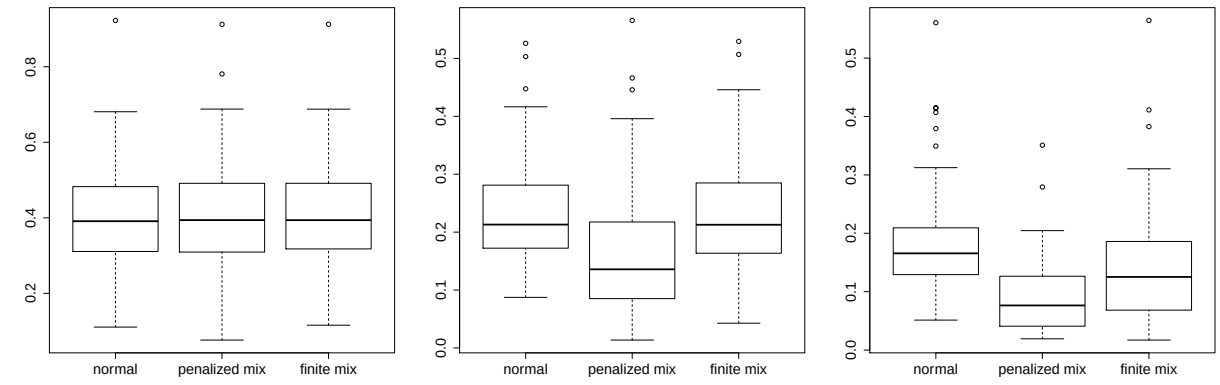
Figure 3.13.: Bar plots of the number of clusters with clearly separated clusters for few individual observations (left), a medium number of individual observations (middle) and many individual observations (right).

Moderately separated clusters

When the differences between clusters become smaller, the penalized heterogeneity approach still outperforms the homogeneity model and the heterogeneity model in the case of a medium number of individual observations and many individual observations. As it can be seen in Figure 3.14 and Table 3.4 the prediction errors for the random effects are considerably smaller. With regard to the accuracy of the estimated fixed effects we obtain the same results as in the case of clearly separated clusters: In particular, for β_2 the mean squared errors, the relative biases and the standard errors are clearly smaller for the penalized heterogeneity model. Again, for few individual observations the results of the three models are quite similar since mostly only one cluster is detected by the mixture approaches (Figure 3.15). As in the case of clearly separated clusters the penalized mixture model tends to detect more clusters than the finite mixture model.

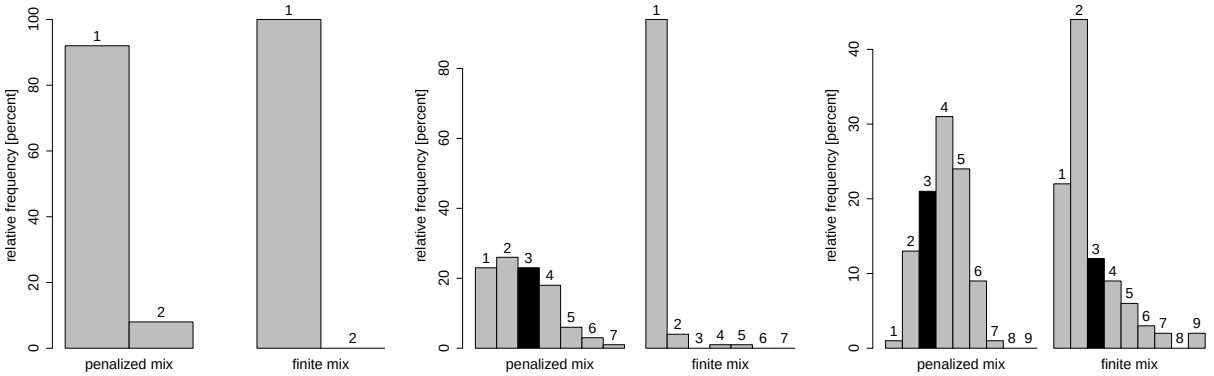
			$\widehat{MSE}_r$	RB_r	$\widehat{se}(\hat{\beta}_r)$	PE_r	$\hat{\sigma}_r^2$
$\nu = 1$	$r = 0$	normal	0.111	-0.264	0.333	**0.391**	1.689
		penalized mix	**0.107**	**-0.225**	**0.327**	0.394	**1.493**
		finite mix	0.111	-0.275	0.332	0.394	1.550
	$r = 1$	normal	0.035	-0.017	0.187	**0.155**	0.493
		penalized mix	**0.034**	0.098	**0.184**	0.160	**0.448**
		finite mix	0.035	**-0.005**	0.186	0.160	0.453
	$r = 2$	normal	**0.051**	**0.657**	**0.224**		
		penalized mix	0.052	0.758	0.225		
		finite mix	0.052	0.713	0.225		
$\nu = 3$	$r = 0$	normal	**0.125**	0.103	**0.353**	0.213	1.726
		penalized mix	**0.125**	**0.031**	**0.353**	**0.136**	**0.554**
		finite mix	0.126	0.104	0.355	0.213	1.516
	$r = 1$	normal	0.029	**0.047**	0.171	0.041	0.498
		penalized mix	**0.028**	0.075	**0.168**	**0.030**	**0.374**
		finite mix	0.029	0.071	0.172	0.040	0.463
	$r = 2$	normal	0.052	-0.103	0.227		
		penalized mix	**0.024**	**0.060**	**0.153**		
		finite mix	0.051	0.088	0.225		
$\nu = 5$	$r = 0$	normal	**0.101**	0.065	**0.318**	0.166	1.606
		penalized mix	0.102	**0.024**	0.319	**0.076**	**0.009**
		finite mix	0.104	0.087	0.322	0.125	1.208
	$r = 1$	normal	**0.024**	**0.328**	0.156	0.015	0.481
		penalized mix	**0.024**	0.386	**0.154**	**0.008**	**0.022**
		finite mix	**0.024**	0.480	0.155	0.012	0.125
	$r = 2$	normal	0.054	-0.481	0.230		
		penalized mix	**0.011**	**-0.228**	**0.105**		
		finite mix	0.032	-0.394	0.178		

Table 3.4.: For moderately separated clusters the estimated mean squared errors MSE_r, the medians of the relative biases RB_r and the estimated standard errors $se(\hat{\beta}_r)$ for the fixed effects are shown. In addition, the medians of PE_r and $\hat{\sigma}_r^2$ are given. Bold values indicate the best value in each case.



(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

Figure 3.14.: Box plots of PE_0 with moderately separated clusters for few individual observations (left), a medium number of individual observations (middle) and many individual observations (right).



(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

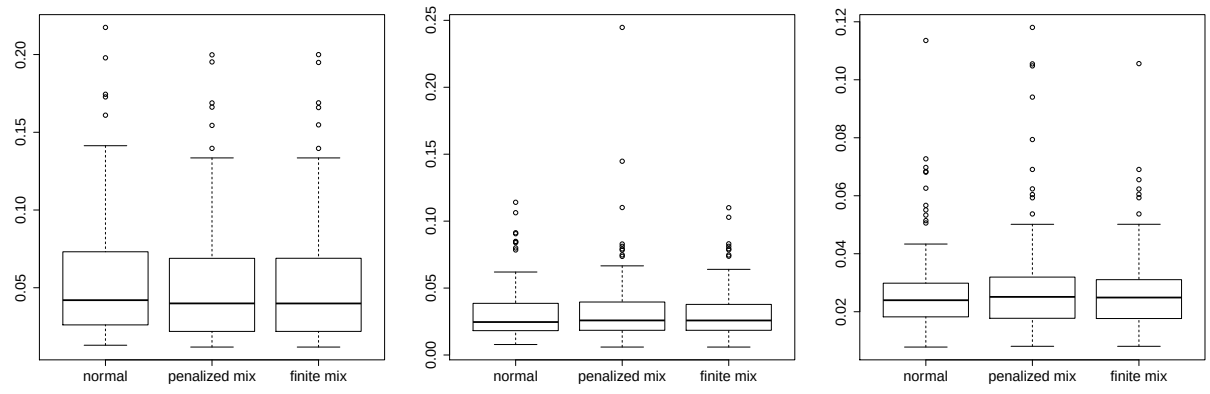
Figure 3.15.: Bar plots of the number of clusters with moderately separated clusters for few individual observations (left), a medium number of individual observations (middle) and many individual observations (right).

One cluster

If the random effects are sampled from a normal distribution, then the classical linear mixed model assumes exactly the correct model. However, as seen in Figure 3.17 also for the mixture approaches mostly all subjects are assigned to the same cluster. So the prediction accuracy of the random effects as well as the accuracy of the estimated fixed effects are almost identical for the three models (Figure 3.16 and Table 3.5).

			$\widehat{MSE}_r$	RB_r	$\widehat{se}(\hat{\beta}_r)$	PE_r	$\hat{\sigma}_r^2$
$\nu = 1$	$r = 0$	normal	0.044	0.055	0.209	**0.151**	0.407
		penalized mix	**0.043**	**0.052**	**0.208**	0.156	**0.243**
		finite mix	0.044	0.062	0.209	0.154	0.357
	$r = 1$	normal	**0.014**	-0.023	**0.12**	0.033	0.227
		penalized mix	**0.014**	**0.004**	**0.12**	**0.032**	**0.204**
		finite mix	**0.014**	-0.016	**0.12**	**0.032**	0.209
	$r = 2$	normal	**0.020**	**0.028**	**0.142**		
		penalized mix	**0.020**	0.204	**0.142**		
		finite mix	**0.020**	0.110	0.143		
$\nu = 3$	$r = 0$	normal	**0.009**	0.094	**0.091**	**0.025**	0.030
		penalized mix	**0.009**	**0.077**	0.092	0.026	**0.019**
		finite mix	**0.009**	0.085	**0.091**	0.026	0.022
	$r = 1$	normal	**0.002**	-0.156	**0.042**	0.008	0.021
		penalized mix	**0.002**	**-0.144**	0.043	0.008	**0.019**
		finite mix	**0.002**	**-0.144**	**0.042**	**0.007**	**0.019**
	$r = 2$	normal	**0.008**	-0.010	0.089		
		penalized mix	**0.008**	**-0.009**	**0.088**		
		finite mix	**0.008**	**-0.009**	0.090		
$\nu = 5$	$r = 0$	normal	**0.008**	-0.050	**0.088**	**0.024**	0.023
		penalized mix	**0.008**	**-0.038**	0.089	0.025	**0.014**
		finite mix	**0.008**	-0.046	**0.088**	0.025	0.016
	$r = 1$	normal	**0.001**	0.019	**0.038**	**0.004**	0.020
		penalized mix	**0.001**	**0.017**	**0.038**	**0.004**	**0.019**
		finite mix	**0.001**	**0.017**	**0.038**	**0.004**	**0.019**
	$r = 2$	normal	**0.008**	-0.181	**0.089**		
		penalized mix	0.009	-0.177	0.094		
		finite mix	**0.008**	**-0.151**	0.090		

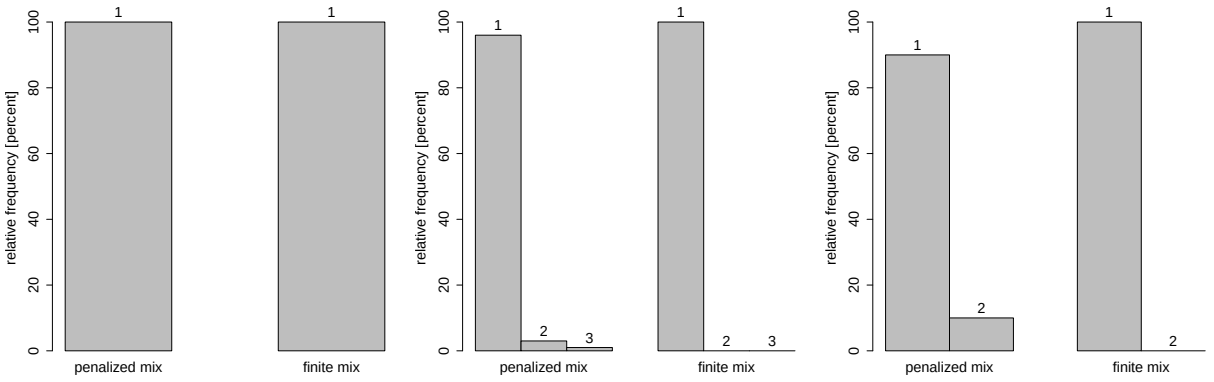
Table 3.5.: For only one cluster the estimated mean squared errors MSE_r, the medians of the relative biases RB_r and the estimated standard errors $se(\hat{\beta}_r)$ for the fixed effects are shown. In addition, the medians of PE_r and $\hat{\sigma}_r^2$ are given. Bold values indicate the best value in each case.



(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

Figure 3.16.: Box plots of PE_0 with only one cluster for few individual observations (left), a medium number of individual observations (middle) and many individual observations (right).

In summary, we can draw the following conclusion: The penalized heterogeneity approach performs well in terms of prediction errors if the clusters are well separated and enough observations are available. We found that for few repeated measurements per subject the discrimination between clusters is harder than for a medium number of individual observations or many individual observations. Nevertheless, there is no loss in efficiency in using the penalized heterogeneity model in the case of few repeated measurements per subject, even in the extreme situation that the true random effects are a sample from a homogeneous Gaussian population.



(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

Figure 3.17.: Bar plots of the number of clusters with only one cluster for few individual observations (left), a medium number of individual observations (middle) and many individual observations (right).

3.4.3. Impact on Test Characteristics

In applications one is often interested in testing the significance of some variables. In what follows, the widely-used statistic $\hat{\beta}_r / \widehat{se}(\hat{\beta}_r)$, which frequently can be approximated by a standard normal distribution, is used as test statistic for the hypothesis $H_0 : \beta_r = 0$ against $H_1 : \beta_r \neq 0$. Typically the performance of tests is checked by characteristics like the type I error, the power or the coverage rate of the corresponding confidence intervals. Although we do not primarily aim at evaluating the test procedure itself, it is quite interesting if the three methods considered in Section 3.4.2 have an impact on these test characteristics. For examining these effects we pick up the simulation setting of clearly separated clusters with many individual observations. In addition, we consider three further settings by varying the choice for the fixed effects. In each setting 100 data sets are generated following the linear trend model described in Section 3.4.1. Altogether we choose $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T = (0.2, 0.1, 0.05)^T$, $\boldsymbol{\beta} = (0.5, 0.5, 0.5)^T$, and $\boldsymbol{\beta} = (3, 2, 1)^T$ for different kinds of effects to check the power of the test and $\boldsymbol{\beta} = (0, 0, 0)^T$ for no effects to examine the type I error. In addition, it is examined in how many times the approximative 95% confidence intervals for β_r based on the statistic $\hat{\beta}_r / \widehat{se}(\hat{\beta}_r)$ covers the true parameters. The standard errors are estimated for each data set by the bootstrap method (3.6) with 50 replications.

	$\boldsymbol{\beta}^T$	$(0,0,0)$	$(0.2,0.1,0.05)$		$(0.5,0.5,0.5)$		$(3,2,1)$	
		Type I error	95%-CI coverage	Power	95%-CI coverage	Power	95%-CI coverage	Power
$r=0$	normal	0.07	0.93	0.11	0.93	0.24	0.93	1.00
	penalized mix	0.09	0.92	0.15	0.91	0.23	0.92	1.00
	finite mix	0.07	0.93	0.14	0.93	0.22	0.92	1.00
$r=1$	normal	0.07	0.93	0.07	0.93	0.74	0.93	1.00
	penalized mix	0.06	0.93	0.08	0.94	0.73	0.93	1.00
	finite mix	0.07	0.93	0.08	0.93	0.73	0.94	1.00
$r=2$	normal	0.06	0.94	0.05	0.94	0.30	0.94	0.75
	penalized mix	0.01	0.99	0.03	0.99	0.81	0.99	0.98
	finite mix	0.05	0.94	0.07	0.97	0.78	0.99	0.98

Table 3.6.: Type I error rates and power rates of approximative tests for $H_0 : \beta_r = 0$ versus $H_1 : \beta_r \neq 0$ for the significance level 5% as well as coverage rates of 95% confidence intervals for different settings of $\boldsymbol{\beta}$. The data are generated according to the setting of clearly separated clusters with many individual observations.

According to Table 3.6 the type I error rates, the power rates and the coverage rates for the intercept β_0 as well as the slope parameter β_1 are quite similar for the three different random effects distributions "normal", "penalized mixture", and "finite mixture". In contrast, we get very different test characteristics for β_2. For the finite mixture approach and especially for the penalized mixture approach high coverage rates and low error rates are observed. In particular, the power rates in the settings $\beta_2 = 0.5$ and $\beta_2 = 1$ are considerably higher for the mixture approaches than in the case of a normal random effects

distribution. This comes along with the results in Section 3.4.2, where the standard errors and the mean squared errors for β_2 are clearly smaller for the finite mixture and especially for the penalized mixture. In summary, we detect positive impacts of the penalized heterogeneity approach on significance tests concerning the fixed effects. The performance of tests is considerably improved for variables which are modeled exclusively by fixed effects.

3.4.4. Heterogeneity of Random Effects

This section deals with the background of the mixture approach used in our penalized heterogeneity model. The reason for the heterogeneity in the distribution of the random effects is a grouping structure with respect to the time trends of the response variable. However, it could be that some unobserved covariates or some interactions of latent covariates with the time variable are actually responsible for the multimodality in the random effects distribution and the question arises if a mixture distribution for the random effects is not needed any more when these covariates are incorporated in the model. For illustrating this topic, we consider data generated by

$$y_{ij} = \beta_0 + \underbrace{\tilde{b}_{i0} + \beta_2 x_i}_{b_{i0}} + (\beta_1 + \underbrace{\tilde{b}_{i1} + \beta_3 x_i}_{b_{i1}}) t_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i, \tag{3.9}$$

with i.i.d. errors $\varepsilon_{ij} \sim N(0, \sigma^2)$ and i.i.d. normally distributed random effects $\tilde{\boldsymbol{b}}_i = (\tilde{b}_{i0}, \tilde{b}_{i1})^T \sim N(\boldsymbol{0}, \boldsymbol{D})$. Let $x_i \in \{-1, 1\}$ be an effect-coded binary variable like, for example, gender. On the one hand, these data can be modeled by including the effect-coded variable x_i and the interaction of x_i with *time* in the model with normally distributed random effects. On the other hand, if these variables are not included in the model, a normal mixture for the random effects would be able to account for this hidden grouping structure. Then the random effects are given by

$$\boldsymbol{b}_i = (b_{i0}, b_{i1})^T = \begin{cases} \boldsymbol{\mu}_1 + \tilde{\boldsymbol{b}}_i = (-\beta_2, -\beta_3)^T + \tilde{\boldsymbol{b}}_i, & i \in \text{cluster 1}, \\ \boldsymbol{\mu}_2 + \tilde{\boldsymbol{b}}_i = (\beta_2, \beta_3)^T + \tilde{\boldsymbol{b}}_i, & i \in \text{cluster 2}. \end{cases}$$

This means that if the grouping structure comes from variables with a prior known cluster membership, it could be accounted for by fixed effects directly and a mixture as random effects distribution is not necessary. However, our approach aims at detecting clusters in cases where a prior known cluster membership is not given.

It should be noted that no identifiability problem arises for data generated by equation (3.9) if the variable x_i and the interaction term are included in the model *and* the penalized normal mixture as random effects distribution is assumed because of the restriction $\sum_{h=1}^{N} \pi_h \boldsymbol{\mu}_h = \boldsymbol{0}$ and the penalty term (3.3). Therefore the cluster centers are always devi-

ations from the trend curve with lengths as small as possible. This feature is illustrated in more detail now. We assume data generated as in equation (3.9) with $n = 20$ and

$$\sigma^2 = 0.25, \qquad \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0.5 \\ 1 \\ 0.2 \end{pmatrix}, \qquad \boldsymbol{D} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix} = \begin{pmatrix} 0.02 & 0.01 \\ 0.01 & 0.02 \end{pmatrix}.$$

Observation times are simulated as in Section 3.4.1 with $\nu = 3$ and x_i are sampled from $\{-1, 1\}$ with equal weights. We consider two different models: Model I denotes the random slope model with a penalized normal mixture as random effects distribution including the variable x_i and the interaction term as in equation (3.9). Model II is the same model but without the variable x_i and the interaction term. As expected, in model II two clusters are detected with cluster centers $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2 \neq \mathbf{0}$ while in model I all individuals are assigned to the same cluster. So model I equates to the homogeneity model with only one cluster center $\hat{\boldsymbol{\mu}} = \mathbf{0}$ and coefficients $\hat{\beta}_2, \hat{\beta}_3 \neq 0$. As it is seen in Table 3.7 the uncentered estimates for intercepts and slopes are almost equal for model I and model II. Since both models obviously yield the same fit and model II is a special case of model I with $\beta_2 = \beta_3 = 0$, one may wonder if a identifiability problem arises in model I and if it is possible that model I yields estimates $\hat{\beta}_2 = \hat{\beta}_3 = 0$ and two clusters $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2 \neq \mathbf{0}$ as in model II. In this case, however, the penalty term (3.3) would increase from zero to $\sqrt{2 \cdot 2}\|\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2\|$ while the likelihood function is still the same. Hence such a solution would not be optimal. In general, our penalized heterogeneity model prefers always the unique solution where the differences between the clusters are as small as possible.

In summary, it has been seen that if the grouping structure comes from variables with a prior known cluster membership it could be accounted for by fixed effects directly or otherwise by a penalized normal mixture as random effects distribution. But note that typically such covariates are not available. In addition, there are many instances in which the heterogeneity in the random effects distribution cannot be accounted for by normally distributed random effects even if these covariates which explain the heterogeneity are available and included in the model: Imagine that the fixed effect β_{2h} of a covariate x_i differs for several groups $h = 1, \ldots, N$ and let the data generating process be given by

$$y_{ij} = \beta_0 + \tilde{b}_{i0} + (\beta_1 + \tilde{b}_{i1})t_{ij} + \beta_{2c_i}x_i + \varepsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i, \tag{3.10}$$

with i.i.d. errors $\varepsilon_{ij} \sim N(0, \sigma^2)$ and i.i.d. normally distributed random effects $\tilde{\boldsymbol{b}}_i = (\tilde{b}_{i0}, \tilde{b}_{i1})^T \sim N(\mathbf{0}, \boldsymbol{D})$. The cluster allocation variable c_i is h if subject i belongs to cluster h. With $\beta_{2c_i} = \beta_2 + \tilde{\beta}_{2c_i}$ equation (3.10) can be rewritten as

$$y_{ij} = \beta_0 + \underbrace{\tilde{b}_{i0} + \tilde{\beta}_{2c_i}x_i}_{b_{i0}} + (\beta_1 + \underbrace{\tilde{b}_{i1}}_{b_{i1}})t_{ij} + \beta_2 x_i + \varepsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i.$$

	Intercept		Slope	
Subject	Model I	Model II	Model I	Model II
i	$\hat{\beta}_0 + \hat{\tilde{b}}_{i0} + \hat{\beta}_2 x_i$	$\hat{\beta}_0 + \hat{b}_{i0}$	$\hat{\beta}_1 + \hat{\tilde{b}}_{i1} + \hat{\beta}_3 x_i$	$\hat{\beta}_1 + \hat{b}_{i1}$
1	2.2784	2.2790	0.8548	0.8547
2	-0.2558	-0.2550	0.2528	0.2526
3	0.1297	0.1303	0.3865	0.3864
4	1.8761	1.8757	0.5744	0.5744
5	2.0698	2.0690	0.6139	0.6141
6	1.8142	1.8141	0.5900	0.5899
7	1.6784	1.6792	0.6278	0.6274
8	1.9397	1.9391	0.5929	0.5930
9	0.4439	0.4444	0.3960	0.3962
10	-0.3305	-0.3298	0.2158	0.2156
11	-0.3957	-0.3954	0.1741	0.1736
12	-0.2097	-0.2097	0.1842	0.1841
13	2.0671	2.0667	0.6645	0.6646
14	-0.3740	-0.3735	0.1711	0.1708
15	-0.0634	-0.0633	0.2524	0.2523
16	0.0002	-0.0008	0.1608	0.1610
17	0.2381	0.2385	0.4028	0.4027
18	-0.0720	-0.0724	0.1987	0.1988
19	0.3572	0.3563	0.2921	0.2926
20	0.1490	0.1486	0.2939	0.2940

Table 3.7.: Estimation results of the intercepts and slopes for model I and model II.

Thus, the covariate x_i causes heterogeneity in the distribution of $\boldsymbol{b}_i = (b_{i0}, b_{i1})^T$ which cannot be accounted for by normally distributed random effects even if the covariate x_i is included in the model. The heterogeneity can only be accounted for by fixed effects if it is known a priori which subject belongs to which cluster. But typically a prior known cluster membership is not given. In cases like this our penalized heterogeneity model can be used for clustering.

3.5. Summary and Discussion

We introduced a penalized heterogeneity approach for linear mixed models, which assumes a finite mixture of normal distributions for the random effects distribution and penalizes the number of mixture components by fusing the cluster centers via a group fused lasso penalty term. The approach aims at clustering individuals for longitudinal data. We presented an EM algorithm for estimating all parameters. A simulation study showed that our approach basically outperforms the classical linear mixed model with normal random effects distribution and the heterogeneity model. Furthermore, the usefulness of our model

was demonstrated in three data examples: We identified similarities in the development of unemployment rates in Germany as well as of the growth of rats depending on the treatment group and showed that our model is able to detect a underlying cluster structure in the lung function growth data, which is hardly seen in the raw data. Extensions of our approach to additive mixed models or to linear mixed models with multiple levels of grouping are possible and seem to be feasible without major difficulties.

4. Linear Mixed Models with DPMs using EM Algorithm

4.1. Introduction

In the following chapter linear mixed models are considered as in Chapter 3. These models are a common tool for the modeling of longitudinal data. The classical model has the form

$$y_{ij}|\boldsymbol{b}_i \overset{ind.}{\sim} N(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i,\, \sigma^2), \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i, \tag{4.1}$$

where y_{ij} denotes the response observed for subject i at observation time t_{ij} with $t_{i1} < \ldots < t_{ij} < \ldots < t_{in_i}$. Population effects of covariates $\boldsymbol{x}_{ij}$ are collected in the parameter vector $\boldsymbol{\beta}$ whereas individual-specific effects of covariates $\boldsymbol{z}_{ij}$ are represented in the parameter vector $\boldsymbol{b}_i$. Typically, in linear mixed models (4.1) normally distributed random effects are assumed, i.e. $\boldsymbol{b}_i$ is i.i.d. $N(\boldsymbol{0}, \boldsymbol{D})$, see for example Verbeke and Molenberghs (2000) and Ruppert et al. (2003). While this choice features some mathematical benefits, in applications it is often questionable because of special properties of the normal distribution like symmetry or unimodality. Since the distributional assumption is made on unobserved quantities, it is typically hard to validate these properties. Especially in the case of a grouping structure in the data the unimodal normal distribution is very restrictive. A finite mixture of normal distributions as a random effects distribution as suggested by Verbeke and Lesaffre (1996) is much more flexible. One assumes

$$\boldsymbol{b}_i \sim \sum_{h=1}^{N} \pi_h\, N(\boldsymbol{\mu}_h, \boldsymbol{D}), \quad i = 1, \ldots, n, \tag{4.2}$$

where $\pi_1, \ldots, \pi_N$ are mixture weights, which add up to one. See Section 3.1 for an detailed overview on extensions and alternatives to this heterogeneity model. A data driven choice of the number of mixture components is desirable. In contrast to the approach in Chapter 3, where the cluster centers are fused by a group fused lasso penalty, this could also be achieved by a penalization of the mixture weights π_h. For example, Komárek and Lesaffre (2008) penalized differences between reparameterized weights. In contrast, Magder and Zeger (1996) used component specific covariance matrices subject to the constraint that their determinants are greater than or equal to some minimum value.

In this chapter we present an alternative penalization approach. The basic concept is to shrink the weights π_h towards zero in order to reduce the number of clusters. We consider an approximate DPM for the random effects distribution by using the truncated version

of the stick breaking presentation of the Dirichlet process; see Ferguson (1973) for the theory of the Dirichlet process and Sethuraman (1994) for the stick breaking presentation of the Dirichlet process. Chapter 2 gives an elaborate outline of the features of the Dirichlet process. The main advantage of Dirichlet processes is the cluster property: by using a DPM for the random effects distribution we automatically obtain a clustering of individuals. Under the assumption that the population can be described by few clusters we want to identify and interpret them. Since a Dirichlet process allows to specify a prior on probability measures, it has been widely used in Bayesian inference. For linear mixed models, Dirichlet process priors for random effects were first proposed by Bush and MacEachern (1996). The first application of a DPM of Gaussian distributions to random effects was given by Müller and Rosner (1997).

We aim at establishing the Dirichlet process as a tool for frequentist modeling. Therefore, instead of using MCMC methods, which are usually applied for estimation in random effects models with Dirichlet processes like in Chapter 5, we extend the traditional EM algorithm of Dempster et al. (1977) used in the heterogeneity model of Verbeke and Lesaffre (1996) and refer to it as DPM-EM model. We will illustrate that the EM algorithm has an essential advantage over MCMC methods, as far as Dirichlet processes are concerned. In summary, on the one hand, our DPM-EM model provides a regularization approach for the number of mixture components in equation (4.2). On the other hand, our model is a method to obtain clustering of individuals in longitudinal data.

The chapter is organized as follows: In Section 4.2.1 the model hierarchy as well as the cluster property of Dirichlet processes are illustrated. In Section 4.2.2 we present our DPM-EM algorithm in detail. Simulation results can be seen in Section 4.3 while applications are shown in Section 4.4. Finally Section 4.5 subsumes the main aspects of our approach. Large parts of this chapter can also be found in Heinzl and Tutz (2013).

4.2. Linear Mixed Models with Dirichlet Process Mixtures

4.2.1. Model Hierarchy

Collecting observations y_{ij}, $j = 1, \ldots, n_i$, for individual i in the vector $\boldsymbol{y}_i$, model (4.1) can be written in matrix notation as

$$\boldsymbol{y}_i|\boldsymbol{b}_i \overset{ind.}{\sim} N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i,\, \sigma^2\boldsymbol{I}_{n_i}), \quad i = 1, \ldots, n,$$

where $\boldsymbol{I}_{n_i}$ is the identity matrix with dimension n_i and $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ denote the individual design matrices constructed from covariates $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$, respectively. For the random effects distribution, we assume a hierarchical Gaussian mixture

$$\begin{array}{rcll} \boldsymbol{b}_i|\boldsymbol{\theta}_i & \overset{ind.}{\sim} & N(\boldsymbol{\theta}_i, \boldsymbol{D}), & i = 1, \ldots, n, \\ \boldsymbol{\theta}_i|G & \overset{i.i.d.}{\sim} & G, & i = 1, \ldots, n, \\ G & \sim & DP(\alpha, G_0). & \end{array} \tag{4.3}$$

Here, $DP(\alpha, G_0)$ is a distributional assumption for the unknown mixing distribution G. See Section 2.4 for a general description of DPMs. A special feature of the Dirichlet process is that each realization of G is a discrete probability measure (Blackwell, 1973). So in the DPM specification, choosing a Dirichlet process for the $\boldsymbol{\theta}_i,\ i = 1, \ldots, n$, creates ties among these and therefore forms clusters of subjects whereas each subject still has its own unique random effects value. In general, there are $k \leq n$ clusters and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ can be represented by cluster locations $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ and cluster allocation variables. The strength of clustering and therefore the number of clusters is determined by the parameter α, which controls the confidence in the base distribution G_0. This cluster property is illustrated in Chapter 2. According to the relationship between Bayesian and likelihood inference we choose a diffuse uniform distribution on $(-\infty, \infty)$ for G_0. So, in principle, no cluster location is preferred over others. Although in theory an automatic clustering structure is induced by the Dirichlet process, a severe practical problem arises within the Bayesian framework when using MCMC methods, namely how to obtain a single clustering estimate $\hat{c}$ based on an MCMC sample of clusterings $c^{(1)}, \ldots, c^{(M)}$, where $c^{(m)},\ m = 1, \ldots, M$, describes the cluster allocation at iteration m and $\hat{c}$ the final cluster allocation. By using MCMC methods in each iteration ties among the $\boldsymbol{\theta}_i,\ i = 1, \ldots, n$, are created and clusters are formed. But when approximating the posterior means by the means over MCMC samples $\hat{\boldsymbol{\theta}}_i = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\theta}_i^{(m)},\ i = 1, \ldots, n$, the clustering of subjects gets lost. Fritsch and Ickstadt (2009) gave an overview on operations how the MCMC sample of clusterings $c^{(1)}, \ldots, c^{(M)}$ can aggregate to a single clustering $\hat{c}$, but due to the high number of possible clusterings, these methods are typically not feasible in larger problems. By using EM type algorithms all these strategies for rescuing the cluster property of the Dirichlet process are unnecessary. The reason is that the EM algorithm converges to fixed values whereas MCMC methods converge to distributions. So with EM type algorithms the cluster property of the Dirichlet process can be used more directly. While other alternatives to the MCMC methods as the recursive algorithm of Newton and Zhang (1999) or the variational method of Blei and Jordan (2006) are based on approximative posterior distributions, our EM algorithm aims at maximizing the posterior given in Section 4.2.2 directly.

In practice, inference with Dirichlet processes can be built on the stick breaking representation of the Dirichlet process by Sethuraman (1994), which is explained in Section 2.2. In its truncated version G is given by

$$G = \sum_{h=1}^{N} \pi_h \, \delta_{\boldsymbol{\mu}_h},$$

with sufficient large N. Here, $\delta_{\boldsymbol{\mu}_h}$ denotes the Dirac measure on $\boldsymbol{\mu}_h$ and π_h is the corresponding random weight. In summary, by using the stick breaking procedure the distribution assumption for the random effects (4.3) can be rewritten as

$$\begin{array}{rcll} \boldsymbol{b}_i|\boldsymbol{v} & \overset{i.i.d.}{\sim} & \sum_{h=1}^{N} \pi_h \, N(\boldsymbol{\mu}_h, \boldsymbol{D}), & i = 1, \ldots, n, \\ \pi_h & = & v_h \prod_{l<h}(1 - v_l), & h = 1, \ldots, N, \\ v_h & \overset{i.i.d.}{\sim} & Be(1, \alpha), & h = 1, \ldots, N-1, \end{array} \tag{4.4}$$

with $\boldsymbol{v} = (v_1, \ldots, v_{N-1})^T$ and beta distribution $Be(\cdot, \cdot)$. Therefore, for the random effects distribution we get a finite mixture of normal distributions as in equation (4.2), in which the number of mixture components with $\pi_h \neq 0$ is penalized. The concentration parameter α controls the number of cluster locations $\boldsymbol{\mu}_h$ with weights $\pi_h \neq 0$ and thus the effective number of clusters. Figure 2.5 illustrates two discrete probability measures simulated by Dirichlet processes with different values of α. It should be noted that a generalization to a heteroscedastic normal mixture with different covariance matrices over components is also possible − following, for example, the approach of Yao and Holmes (2011). Nevertheless, the assumption (4.4) seems to be sufficiently flexible and avoids numerical problems, which arise in the case of a heteroscedastic normal mixture (Verbeke and Molenberghs, 2000). In the following the order of $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ is given by the corresponding weights in decreasing order under the restrictions $\sum_{h=1}^{N} \pi_h \boldsymbol{\mu}_h = \mathbf{0}$ and $\sum_{h=1}^{N} \pi_h = 1$. The first restriction ensures $\mathbb{E}(\boldsymbol{y}_i) = \boldsymbol{X}_i \boldsymbol{\beta}$. The second constraint is standard and is automatically fulfilled by $v_N = 1$.

For example, the truncated Dirichlet process was used by Muliere and Tardella (1998), Ishwaran and James (2002), Kottas and Gelfand (2001), Gelfand and Kottas (2002) and Ohlssen et al. (2007); see Section 4.2.2 for a strategy of choosing N. Even though other methods exist that are based on the stick breaking representation and that avoid the truncation (see, for example, Walker (2007) and Papaspiliopoulos and Roberts (2008)) the truncated version distinguishes itself by simplicity and theoretical justifications as shown in Muliere and Tardella (1998), Ishwaran and James (2001) as well as Ishwaran and James (2002). In our case, this truncation is still more attractive because our approach is formally similar to the heterogeneity model of Verbeke and Lesaffre (1996) but with "penalized" weights referred to the stick breaking procedure, which induces that only the relevant clusters get comparably high weights. Inference is possible by extending the EM algorithm of the heterogeneity model. Another inference approach within the framework of Dirichlet processes is based on the Pólya urn scheme (Blackwell and MacQueen, 1973) and thus on integrating out the unknown distribution G (compare Escobar (1994), MacEachern (1994), Escobar and West (1995) as well as MacEachern and Müller (1998)). A description of the Pólya urn scheme can be found in Section 2.3. Nevertheless, when using this marginal method instead of the stick breaking procedure the connection between the Dirichlet process and the heterogeneity model of Verbeke and Lesaffre (1996) is hidden. This is the main reason why the stick breaking presentation is much more appealing to us and seems to be more user-friendly than the Pólya urn inference scheme, which also has other drawbacks (see, for example, Ishwaran and James (2001)). In the next section, we will explain how Dirichlet processes can be embedded in the EM framework. It can be seen that an elaborate handling of the Dirichlet process's parameters is necessary.

4.2.2. Inference

In the following, we give an EM algorithm for the linear mixed model described in Section 4.2.1. The algorithm is based on derivations by McLachlan and Krishnan (1997) and McLachlan and Peel (2000) and is similar to the algorithm used by Verbeke and Lesaffre (1996) but includes a penalty term. The following approach can either be parameterized

by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^T$ or by $\boldsymbol{v}$. Since the latter parametrization simplifies calculations, it is used in the following. Nevertheless, only for a compact presentation, we write π_h instead of $v_h \prod_{l<h}(1 - v_l)$. Let $\boldsymbol{\xi} = (\alpha, \boldsymbol{v}, \boldsymbol{\psi})^T$, where $\boldsymbol{\psi}$ is the vector containing all the remaining parameters $\boldsymbol{\beta}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N, \boldsymbol{D}, \sigma^2$. The cluster membership of each individual can be described by the latent variable $\boldsymbol{w}_i := (w_{i1}, \ldots, w_{iN})^T$ where $w_{ih} = 1$ if subject i belongs to cluster h and 0 otherwise. Marginalization over the random effects yields the complete model with observed data $\boldsymbol{y}_i$ as well as unobserved data $\boldsymbol{w}_i$ and $\boldsymbol{v}$:

$$\begin{aligned} \boldsymbol{y}_i|\boldsymbol{w}_i &\overset{ind.}{\sim} N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\mu}_h,\, \boldsymbol{V}_i), && i = 1, \ldots, n, \\ \boldsymbol{w}_i|\boldsymbol{v} &\overset{i.i.d.}{\sim} M(1, \boldsymbol{\pi}), && i = 1, \ldots, n, \\ v_h &\overset{i.i.d.}{\sim} Be(1, \alpha), && h = 1, \ldots, N-1, \end{aligned} \tag{4.5}$$

with $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \sigma^2\boldsymbol{I}_{n_i}$ and $M(\cdot, \cdot)$ denoting the multinomial distribution. Equation (4.5) describes the data generating process for the data ($\boldsymbol{y}_i$,$\boldsymbol{w}_i$,$\boldsymbol{v}$) given the parameters (α,$\boldsymbol{\psi}$), i.e.,

$$p(\boldsymbol{y}_i, \boldsymbol{w}_i, \boldsymbol{v}; \alpha, \boldsymbol{\psi}) = p(\boldsymbol{y}_i|\boldsymbol{w}_i; \boldsymbol{\psi}) \cdot p(\boldsymbol{w}_i|\boldsymbol{v}) \cdot p(\boldsymbol{v}; \alpha), \quad i = 1, \ldots, n.$$

This can also be viewed as product of $p(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{v}; \boldsymbol{\psi})$ with the prior $p(\boldsymbol{v}; \alpha)$. Following this formulation, the posterior for $\boldsymbol{\xi}$ is proportional to the product of the likelihood and the prior, which is given by

$$L_P(\boldsymbol{\xi}) = \prod_{i=1}^{n}\prod_{h=1}^{N}[\pi_h\, f_{ih}(\boldsymbol{y}_i; \boldsymbol{\psi})]^{w_{ih}} \cdot \alpha^{N-1} \prod_{h=1}^{N-1}(1 - v_h)^{\alpha-1},$$

when assuming a flat prior for α and $\boldsymbol{\psi}$. Here $f_{ih}(\cdot)$ denotes the density function of $N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\mu}_h,\, \boldsymbol{V}_i)$. Note that from a Bayesian point of view $\boldsymbol{\psi}$ and $\boldsymbol{v}$ are parameters whereas α is the hyperparameter for the prior on $\boldsymbol{v}$. In an empirical Bayes context such a hyperparameter would be estimated by maximizing the marginal incomplete likelihood (Maritz and Lwin, 1989). However, in the present case the marginalization is analytically not feasible. Following the strategy of McAuliffe et al. (2006), in the case of a DPM model such an integration could be avoided by many alternations between an inference phase where the parameters $\boldsymbol{v}$ and $\boldsymbol{\psi}$ are estimated and an estimation phase where the hyperparameter α is estimated. This procedure would be very time-consuming in our case. Thus, we prefer to handle α like any other parameter and to estimate α conditionally on the actual state of the other parameters during the algorithm. In general, vague priors like our diffuse prior for α are an alternative to empirical Bayes inference for achieving robustness (McAuliffe et al., 2006).

Finally, as log-posterior one obtains

$$l_P(\boldsymbol{\xi}) = \sum_{i=1}^{n}\sum_{h=1}^{N} w_{ih}[\log \pi_h + \log f_{ih}(\boldsymbol{y}_i; \boldsymbol{\psi})] + (N-1)\log\alpha + (\alpha - 1)\sum_{h=1}^{N-1}\log(1 - v_h).$$

This function can be seen either as log-posterior in the Bayesian context or as penalized log-likelihood, whose penalization term results from the stick breaking procedure of the Dirichlet process. Obviously for $\alpha = 1$ the penalization term drops out. According to the general EM algorithm procedure we alternate between taking the expectation of $l_P(\boldsymbol{\xi})$ over all unobserved w_{ih} in the E-step and maximization of this expected value in the M-step instead of maximizing the penalized incomplete likelihood function based only on the observed data directly.

E-step

Collecting all observed data in $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$, for the E-step of iteration $t+1$ we get

$$\begin{aligned} Q(\boldsymbol{\xi}) &= \mathbb{E}\left(l_P(\boldsymbol{\xi})|\boldsymbol{y}, \boldsymbol{\xi}^{(t)}\right) = \\ &= \sum_{i=1}^{n}\sum_{h=1}^{N} \pi_{ih}(\boldsymbol{\xi}^{(t)})[\log \pi_h + \log f_{ih}(\boldsymbol{y}_i; \boldsymbol{\psi})] + (N-1)\log\alpha + (\alpha - 1)\sum_{h=1}^{N-1}\log(1-v_h), \end{aligned}$$

where $\pi_{ih}(\boldsymbol{\xi}^{(t)})$ is the probability at iteration t that subject i belongs to cluster h and is given by

$$\pi_{ih}(\boldsymbol{\xi}^{(t)}) = \frac{f_{ih}(\boldsymbol{y}_i; \boldsymbol{\psi}^{(t)})\pi_h^{(t)}}{\sum_{l=1}^{N} f_{il}(\boldsymbol{y}_i; \boldsymbol{\psi}^{(t)})\pi_l^{(t)}}.$$

M-step

For clarity, in the following we write $\pi_{ih} := \pi_{ih}(\boldsymbol{\xi}^{(t)})$, but note that for the M-step it is essential that π_{ih} is fixed from the last iteration t. As $Q(\boldsymbol{\xi}) = Q(\alpha, \boldsymbol{v}) + Q(\boldsymbol{\psi})$ holds, the optimization problem in the M-step can be separated into two parts: The maximization of

$$Q(\alpha, \boldsymbol{v}) = \sum_{i=1}^{n}\sum_{h=1}^{N} \pi_{ih}\log\pi_h + (N-1)\log\alpha + (\alpha-1)\sum_{h=1}^{N-1}\log(1-v_h),$$

with respect to α and $\boldsymbol{v}$ and the maximization of

$$Q(\boldsymbol{\psi}) = \sum_{i=1}^{n}\sum_{h=1}^{N} \pi_{ih}\ \log f_{ih}(\boldsymbol{y}_i; \boldsymbol{\psi}),$$

with respect to $\boldsymbol{\psi}$. The first optimization problem is solved by alternating updates of the first order conditions

$$v_h = \frac{\sum_{i=1}^{n}\pi_{ih}}{\sum_{i=1}^{n}\sum_{l=h}^{N}\pi_{il} + \alpha - 1}, \qquad h = 1, \ldots, N-1, \tag{4.6}$$

and

$$\alpha = \frac{1-N}{\sum_{h=1}^{N-1} \log(1-v_h)},$$

that are proved in Appendix A.3.1. Without further restrictions it could happen that $v_h \notin [0,1]$ if $\alpha \in (0,1)$. To avoid this we use the following correction approach: Update v_h by (4.6) for increasing h. If $v_{h^*} > 1$, set v_h to 1 for $h = h^*, \ldots, N-1$. This constraint for $\boldsymbol{v}$ is equivalent to the following restriction on $\boldsymbol{\pi}$ by using the stick breaking procedure:

$$\pi_h = \begin{cases} \frac{1}{n+\alpha-1} \sum_{i=1}^{n} \pi_{ih}, & \text{for } h < h^*, \\ 1 - \sum_{l=1}^{h-1} \pi_l & \text{for } h = h^*, \\ 0 & \text{for } h > h^*, \end{cases}$$

where h^* is the lowest index h for which the cumulative sum of the original weights π_l° exceeds one: $\sum_{l=1}^{h} \pi_l^\circ > 1$. See Appendix A.3.1 for more technical details about this correction step. Finally, the idea of the penalization approach becomes evident. First note that for $\alpha = 1$ we get the usual estimators for π_h and no restrictions are needed. Compared to these estimators, for $\alpha \in (0,1)$, all weights π_h for $h < h^*$ are stretched by the factor $\frac{n}{n+\alpha-1}$ while all weights π_h for $h > h^*$ are set to zero. The amount of stretching is controlled by the parameter α. If $\alpha \approx 0$ a very strong clustering is achieved while for larger values of α only few clusters drop out. In order to avoid $\log(0)$ we choose $v_h = 1 - 10^{-300}$ instead of $v_h = 1$ in the algorithm. Then $\pi_h \approx 0$ for $h > h^*$.

In the second part of the M-step,we get the current state for $\boldsymbol{\psi}$ by alternating separate maximization of $Q(\boldsymbol{\psi})$ to $\boldsymbol{\beta}$, to $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ and to the variance parameters $\boldsymbol{D}$ and σ^2. Conditional on the actual state of the other parameters the maximization of $\boldsymbol{\beta}$ results in

$$\boldsymbol{\beta} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^{n} \left(\boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{y}_i - \sum_{h=1}^{N} \pi_{ih} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{Z}_i \boldsymbol{\mu}_h \right) \right).$$

Setting the derivative of $Q(\boldsymbol{\psi})$ with respect to $\boldsymbol{\mu}_h$, $h = 1, \ldots, N$, given $\boldsymbol{\beta}$, $\boldsymbol{D}$ and σ^2 to zero yields

$$\boldsymbol{\mu}_h = \left(\sum_{i=1}^{n} \pi_{ih} \boldsymbol{Z}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{Z}_i \right)^{-1} \left(\sum_{i=1}^{n} \pi_{ih} \boldsymbol{Z}_i^T \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}) \right).$$

The corresponding proofs are shown in Appendix A.3.2. For the simultaneous maximization of the variance parameters given $\boldsymbol{\beta}$ and $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ a numerical procedure like the Nelder-Mead method is necessary. More information about this procedure is given in the paragraph "Implementation" in this section.

Choice of N

By truncation of the Dirichlet process the originally infinite constraints $\sum_{h=1}^{\infty} \pi_h \boldsymbol{\mu}_h = \boldsymbol{0}$ and $\sum_{h=1}^{\infty} \pi_h = 1$ are converted into finite ones $\sum_{h=1}^{N} \pi_h \boldsymbol{\mu}_h = \boldsymbol{0}$ and $\sum_{h=1}^{N} \pi_h = 1$, which can be handled easily. Concretely, the first constraint is obtained by a correction at each M-

step. Deviations from this constraint are subtracted from $\boldsymbol{\mu}_h$, $h = 1, \ldots, N$, and included into $\boldsymbol{\beta}$. The second constraint is fulfilled by $v_N = 1$. On the one hand, this idea avoids reparameterizations as in Jara et al. (2009) or post-processing strategies as in Li et al. (2011). On the other hand, $v_N = 1$ actually means that the last weight π_N absorbs all the remaining probabilities $\pi_N, \ldots, \pi_\infty$ of the untruncated Dirichlet process. So it is important that N is chosen adequately. This is still more challenging because the choice of N depends on α, which itself is estimated. Ohlssen et al. (2007) proposed to set N so that

$$N > 1 + \frac{\log(\varepsilon)}{\log\left(\frac{\alpha}{\alpha+1}\right)},$$

with $\varepsilon > 0$. This condition is derived in the equations (2.6). Thus, for a given range of α a lower bound for N can be determined. For inducing a very strong clustering and according to the previous considerations within this section we restrict α to the range $\alpha \in (0, 1)$ which is automatically fulfilled by a very low starting value for α. This means that even for $N \geq 15$ a good approximation can be achieved ($\varepsilon = 0.0001$). So in the majority of cases $N = \min\{n, 100\}$ is a satisfying choice.

Start and stop of the algorithm

For EM algorithms it is essential how to choose the starting values because the (penalized) incomplete log-likelihood is ascending at each step and the algorithm can converge to a local but not a global maximum. Because there is an agglomerative attempt in each M-step it is reasonable to choose starting values for an agglomerative clustering method generally. Therefore, each subject starts in its own cluster. So there are $n = N$ clusters with weights $\pi_h = 1/N$, $h = 1, \ldots, N$ in the beginning. As cluster locations $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ we consider the predicted random effects $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ of the former fitted linear mixed model with Gaussian random effects distribution. This fit yields starting values for $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{D}$, too. For α we use zero as starting value to induce a very strong clustering.

The algorithm starts with $N = n$ clusters and successively merges clusters during the iterations. Rearranging the weights after each step has the effect that only the relevant clusters keep positive probabilities. So the linear mixed model with DPM as a random effects distribution can be seen as an agglomerative cluster analysis.

The EM algorithm stops if the penalized incomplete log-likelihood is not ascending any more. After convergence we get the cluster membership by the matrix of estimated π_{ih}. An individual i is assigned to that cluster h for which $\hat{\pi}_{ih}$ is maximal. If there are a lot of small weights $\hat{\pi}_h$ we get only few relevant clusters k. Based on the weights of all clusters the random effects are predicted by using the mean of the posterior $\boldsymbol{b}_i|\boldsymbol{y}_i$, which is given by

$$\hat{\boldsymbol{b}}_i = \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) + (\boldsymbol{I}_q - \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{Z}_i)\sum_{h=1}^{N}\hat{\pi}_{ih}\hat{\boldsymbol{\mu}}_h, \quad i = 1, \ldots, n,$$

where q denotes the dimension of the random effects. A proof of this formula is given in Appendix A.4.

Implementation

All computations are implemented in C++ (Stroustrup, 1997), allowing for an efficient treatment of loop-intensive calculations and with regard to slow convergence of the EM algorithm. They are made accessible by the function `lmmDPMEM()` within the R package `clustmixed` (Heinzl, 2012) using the statistical software R (R Development Core Team, 2012). All variables are standardized internally for calculations. See Appendix A.5 for more details about the used standardization. For updating variance parameters we use the C++ library `ASA047` (Burkhardt, 2008), an implementation of the Nelder-Mead algorithm in C++, which was used by Papageorgiou and Hinde (2012) for similar tasks. For the reflection, extension and contraction coefficients we choose the common settings 1.0, 2.0 and 0.5 respectively. See Nelder and Mead (1965) and O'Neill (1971) for more technical details about the algorithm. Note that for ensuring that the covariance matrix $\boldsymbol{D}$ is nonnegative-definite we parameterize the concerning variance parameters by the entries of a lower triangular matrix $\boldsymbol{L}$ according to the Cholesky decomposition $\boldsymbol{D} = \boldsymbol{L}\boldsymbol{L}^T$. Then $\boldsymbol{D}$ is nonnegative-definite for each $\boldsymbol{L}$ and positive-definite (and so invertible, too) if $\boldsymbol{L}$ is a matrix with exclusively nonzero diagonal entries (Lindstrom and Bates, 1988).

4.3. Simulation Study

In the following simulation study the estimation results of our DPM-EM model are examined and compared to competing approaches. In general, for prediction accuracy of random effects there is a trade-off with regard to the assumed number of clusters: On the one hand, for prediction of $\boldsymbol{b}_i$ it makes sense to borrow information from other similar subjects. On the other hand, it is not reasonable to incorporate individuals which show a basically different behavior. First, in Section 4.3.2 this trade-off is analyzed by comparing the commonly used linear mixed model with Gaussian random effects distribution (one cluster model) as well as the three, five, and ten cluster model to our DPM-EM model with a data driven choice for the number of clusters. For fitting linear mixed models with Gaussian random effects the R function `lmer()` from the `lme4` package of Bates et al. (2012) is used. Unpenalized finite normal mixture as random effects distribution are estimated by the function `lmmLASSO()` in the R package `clustmixed` of Heinzl (2012) with $\lambda = 0$ (Section 3.2). Second, in Section 4.3.3 the simulation results of Section 4.3.2 are compared to these of the penalized heterogeneity model based on the group fused lasso penalty from Chapter 3 to see if the penalized heterogeneity model or the DPM-EM model works better.

4.3.1. Settings

In the simulation study we investigate the impact of the number of observations within clusters and the separation between clusters. We generated data sets assuming a simple linear trend model

$$y_{ij}|\boldsymbol{b}_i \overset{ind.}{\sim} N(\beta_0 + b_{i0} + (\beta_1 + b_{i1})t_{ij},\, \sigma^2), \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i.$$

The centered i.i.d. random effects $\boldsymbol{b}_i = (b_{i0}, b_{i1})^T$ follow a mixture distribution with three Gaussian components:

$$\boldsymbol{b}_i \sim 0.4\, N(\boldsymbol{\mu}_1, \boldsymbol{D}) + 0.3\, N(\boldsymbol{\mu}_2, \boldsymbol{D}) + 0.3\, N(\boldsymbol{\mu}_3, \boldsymbol{D}), \quad i = 1, \ldots, n,$$

imitating a population consisting of three clusters of overlapping subpopulations. Throughout the simulations, we set $n = 20$ and

$$\sigma^2 = 0.25, \qquad \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \qquad \boldsymbol{D} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix} = \begin{pmatrix} 0.02 & 0.01 \\ 0.01 & 0.02 \end{pmatrix}.$$

We vary, however, the number of individual observations n_i, the centers $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ of the clusters and the locations of observation times t_{ij}. To produce longitudinal data with varying numbers of repeated observations per unit i, we set $n_i = 2 + X_i$, where X_i follows a Poisson distribution with rate ν. Setting $\nu = 1$ corresponds to longitudinal data with only *few individual observations* (3 on average), $\nu = 3$ to a *medium number of individual observations* and $\nu = 5$ to comparably *many individual observations*. For given n_i, observation times are generated from

$$\begin{aligned} t_{i1} &\sim U(0, 1), \quad i = 1, \ldots, n, \\ t_{ij} &\sim U(t_{i,j-1} + 0.5, t_{i,j-1} + 1.5), \quad i = 1, \ldots, n, \quad j = 2, \ldots, n_i, \end{aligned}$$

where $U(\cdot, \cdot)$ denotes the uniform distribution. Thus, different numbers $n_i(s)$ and $t_{ij}(s)$ are generated in each simulation run $s = 1, \ldots, 100$. Similarly, different "true" random effects $\boldsymbol{b}_i(s)$ are drawn from the Gaussian mixture distribution in each simulation run. For the cluster locations, we chose

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -2.25 \\ 1 \end{pmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 0.75 \\ -1.2 \end{pmatrix}, \qquad \boldsymbol{\mu}_3 = \begin{pmatrix} 2.25 \\ -2/15 \end{pmatrix},$$

corresponding to *clearly separated clusters*,

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -1.5 \\ 0.75 \end{pmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 0.5 \\ -0.9 \end{pmatrix}, \qquad \boldsymbol{\mu}_3 = \begin{pmatrix} 1.5 \\ -0.1 \end{pmatrix},$$

corresponding to *moderately separated clusters*,

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -0.75 \\ 0.5 \end{pmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 0.25 \\ -0.6 \end{pmatrix}, \qquad \boldsymbol{\mu}_3 = \begin{pmatrix} 0.75 \\ -1/15 \end{pmatrix},$$

corresponding to *substantially overlapping clusters*.

Combining these different settings for observations times and clusters results in nine different scenarios. For each of them, we compare the estimation results from the DPM-EM algorithm to results based on Gaussian random effects using the R function `lmer()` from the `lme4` package by Bates et al. (2012) ("normal") and to results of models using an unpenalized finite normal mixture as random effects distribution with different numbers of mixture components ($N = 3$, $N = 5$, $N = 10$). In each simulation run s, we calculate the average prediction error

$$PE_r(s) = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{b}^*_{ir}(s) - b^*_{ir}(s) \right)^2, \quad r = 0, 1, \tag{4.7}$$

for uncentered random intercepts $b^*_{i0} = \beta_0 + b_{i0}$ and random slopes $b^*_{i1} = \beta_1 + b_{i1}$. In addition, the estimation accuracy of the fixed effects is investigated by the relative bias $RB_r = (\hat{\beta}_r - \beta_r)/\beta_r$, $r = 0, 1$.

4.3.2. Results

In the following, we summarize results of the nine combinations. For some scenarios the empirical distribution of $PE_0(s)$ values obtained from simulation run $s = 1, \ldots, 100$ is represented through box plots. Since the figures for $PE_1(s)$ look very similar to those of the random intercepts, they are not shown.

Clearly separated clusters

Figure 4.1 (top) displays trace plots of typical longitudinal data generated in the setting of clearly separated clusters, that show that cluster effects can easily be detected visually. On the left, there are only a few observations for each subject while on the right the mean of the number of repeated measurements is five corresponding to a medium number of observations. Not surprisingly the DPM-EM model detects three clusters in both cases (Figure 4.1 (bottom)). The dashed line shows the overall effect and the solid lines visualize the means of the resulting clusters. Observations from the same cluster are represented by the same symbol.

Linear mixed models with DPM penalty substantially improve upon results based on a misspecified Gaussian random effects assumption, especially in the case of a medium number of individual observations and many observations (see Table 4.1 and, for example, Figure 4.2). In general, models with a finite mixture as random effects distribution yield better predictions for random effects than the classical linear mixed model with normally distributed random effects. Of course, the best prediction can be observed for the model with fixed $N = 3$ clusters because this model is exactly the same as in the data generating

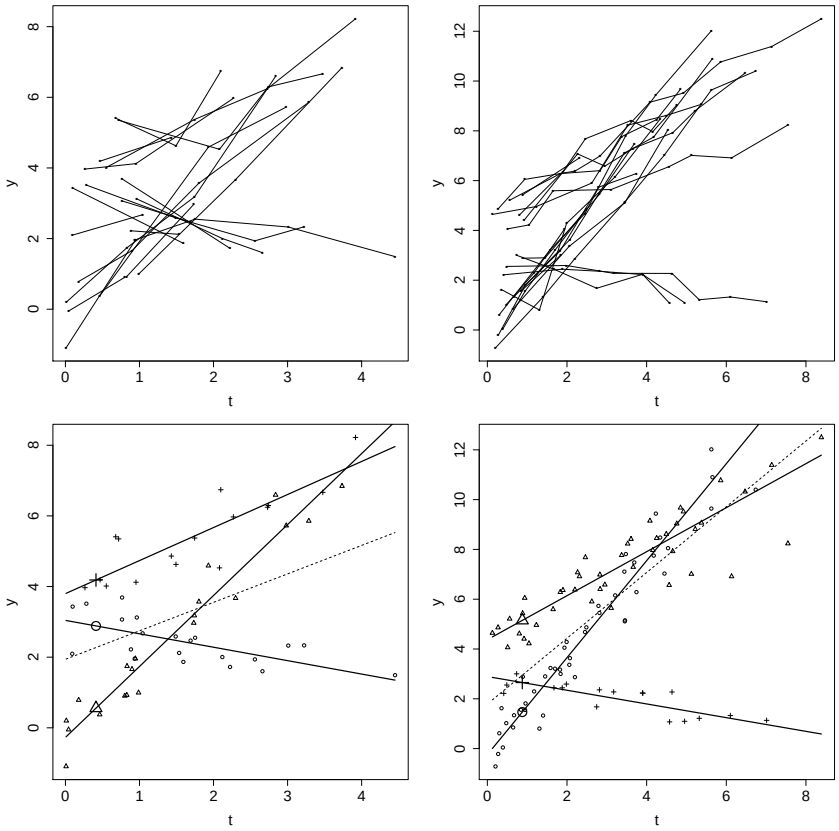


Figure 4.1.: Trace plots (top) and clustering by the DPM-EM model (bottom) with clearly separated clusters for few individual observations ($\nu = 1$) (left) and a medium number of individual observations ($\nu = 3$) (right).

process. However, the DPM-EM model shows quite similar results although in this case the number of clusters was determined by the model itself. The DPM-EM model as well as the other models show a small bias concerning the estimation of fixed effects. The bias tends to be a bit higher in the DPM-EM model.

		PE_0	PE_1	RB_0	RB_1
$\nu = 1$	normal	0.373	0.185	-0.041	**0.021**
	DPM-EM	0.135	0.063	-0.068	0.047
	$N = 3$	**0.111**	**0.058**	-0.037	0.043
	$N = 5$	0.145	0.062	**-0.029**	0.048
	$N = 10$	0.222	0.112	-0.033	**0.021**
$\nu = 3$	normal	0.222	0.054	**-0.010**	0.047
	DPM-EM	0.060	0.012	-0.052	0.069
	$N = 3$	**0.054**	**0.011**	-0.029	0.052
	$N = 5$	0.072	0.015	-0.028	**0.044**
	$N = 10$	0.101	0.020	-0.022	0.063
$\nu = 5$	normal	0.148	0.015	-0.021	0.010
	DPM-EM	0.048	0.006	-0.014	**0.009**
	$N = 3$	**0.045**	**0.005**	-0.005	0.017
	$N = 5$	0.050	0.006	**-0.002**	0.020
	$N = 10$	0.080	0.008	**-0.002**	0.015

Table 4.1.: Medians of PE_r and RB_r with $r = 0, 1$ for clearly separated clusters. Bold values indicate the best value in each case.

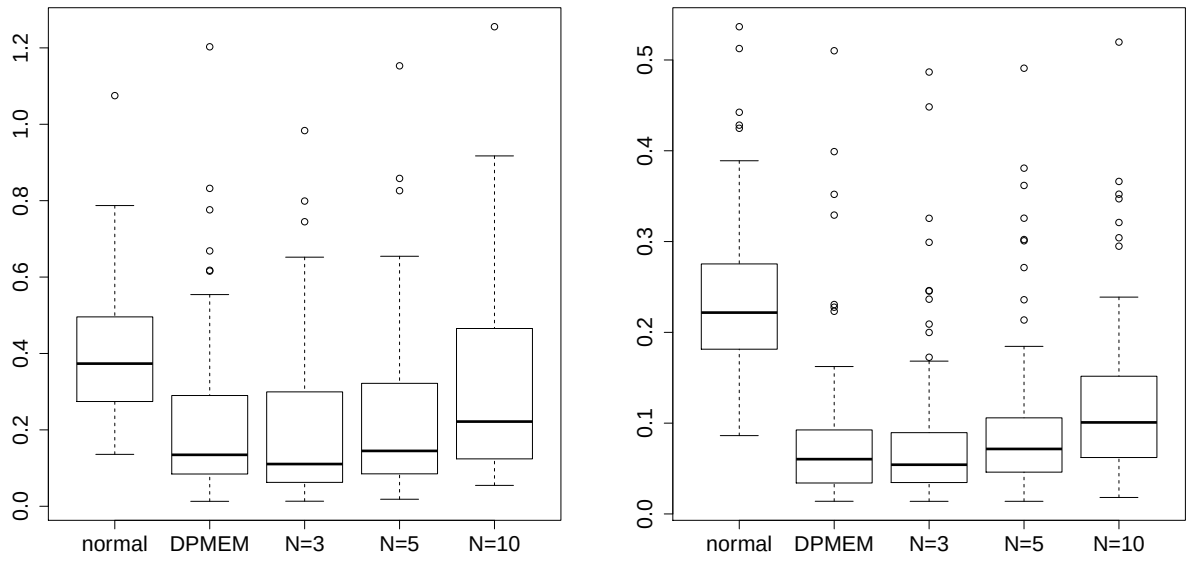


Figure 4.2.: Box plots of PE_0 with clearly separated clusters for few individual observations ($\nu = 1$) (left) and a medium number of individual observations ($\nu = 3$) (right).

Moderately separated clusters

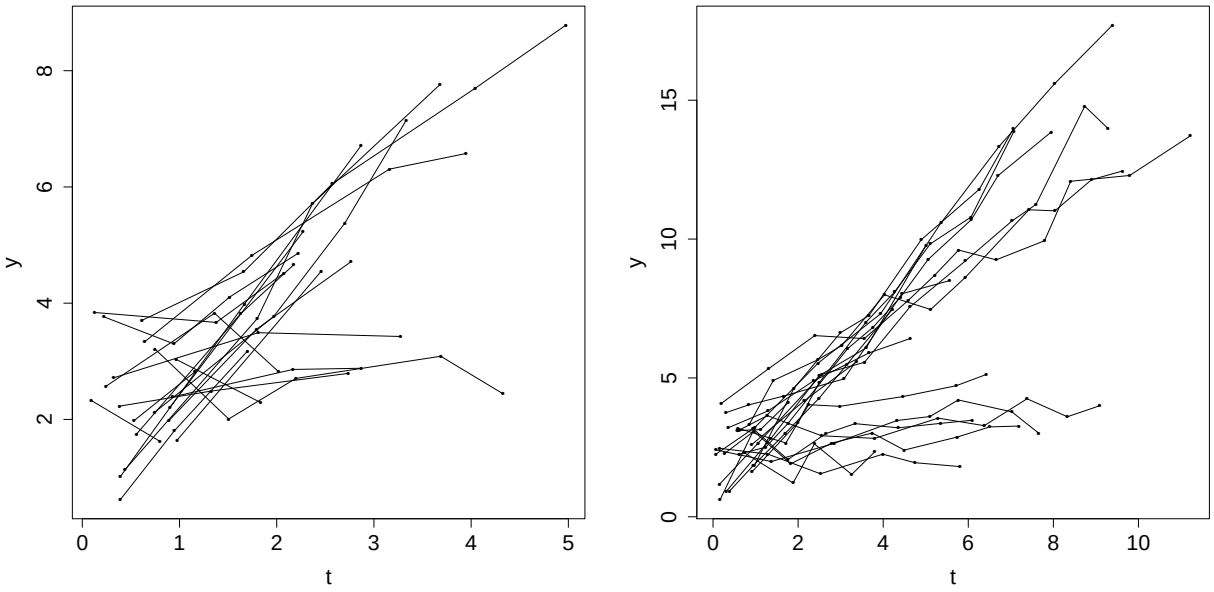


Figure 4.3.: Trace plots with moderately separated clusters for few individual observations ($\nu = 1$) (left) respectively many individual observations ($\nu = 5$) (right).

In the following the differences between the clusters get smaller. See Figure 4.3 for two typical trace plots in the case of few respectively many individual observations. Still the DPM-EM model outperforms both the homogeneity model (linear mixed model with normal random effects distribution) and the unpenalized heterogeneity model with $N = 5$ and $N = 10$ clusters (Figure 4.4 and Table 4.2). Only the true model with $N = 3$ clusters is able to feature a lower error in predicting the random effects. Note that the superiority of the DPM-EM model over the classical linear mixed model with normal random effects distribution is even higher in the case of many individual observations.

		PE_0	PE_1	RB_0	RB_1
$\nu = 1$	normal	0.335	0.164	**-0.021**	0.019
	DPM-EM	0.204	0.114	-0.061	0.047
	$N = 3$	**0.175**	**0.097**	-0.038	0.021
	$N = 5$	0.224	0.122	-0.031	0.020
	$N = 10$	0.274	0.140	-0.030	**0.014**
$\nu = 3$	normal	0.207	0.046	-0.008	**0.022**
	DPM-EM	0.082	0.018	-0.031	0.023
	$N = 3$	**0.063**	**0.014**	**-0.001**	0.032
	$N = 5$	0.082	0.018	**-0.001**	0.031
	$N = 10$	0.126	0.025	-0.003	0.031
$\nu = 5$	normal	0.138	0.015	-0.011	0.008
	DPM-EM	0.048	**0.005**	**-0.009**	0.011
	$N = 3$	**0.043**	**0.005**	-0.013	0.009
	$N = 5$	0.050	0.006	-0.012	**0.007**
	$N = 10$	0.082	0.008	-0.013	0.015

Table 4.2.: Medians of PE_r and RB_r with $r = 0, 1$ for moderately separated clusters. Bold values indicate the best value in each case.

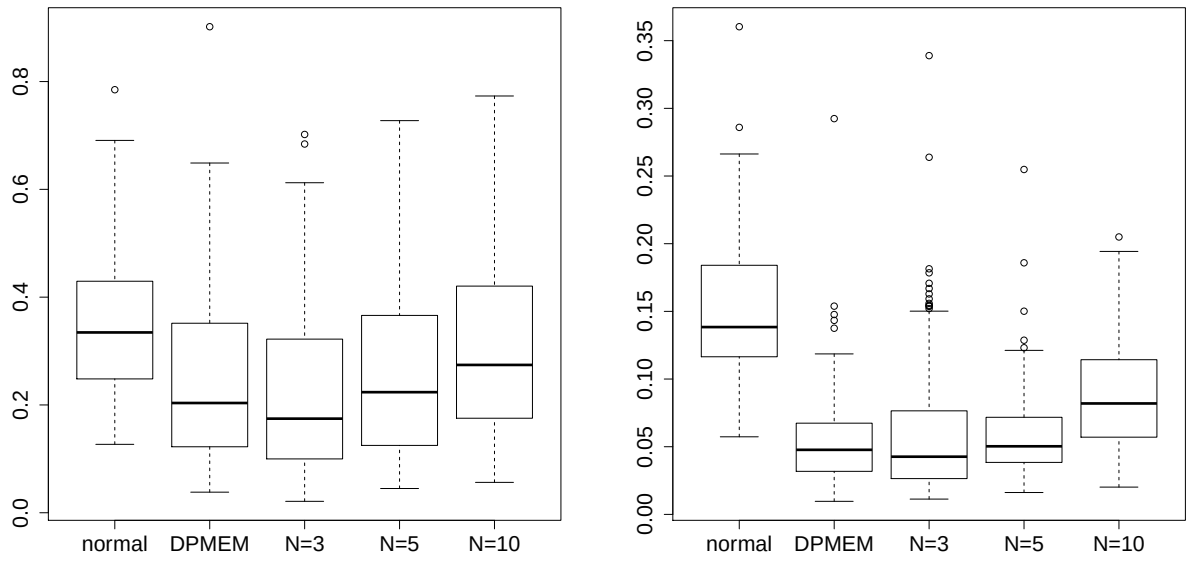


Figure 4.4.: Box plots of PE_0 with moderately separated clusters for few individual observations ($\nu = 1$) (left) respectively many individual observations ($\nu = 5$) (right).

Substantially overlapping clusters

When regarding Figure 4.5 for substantially overlapping clusters, we draw similar conclusion as in the case of moderately separated clusters. Again the true model with $N = 3$ clusters features a lower error in predicting the random effects for a medium number of individual observations. However, for many observations the DPM-EM model exhibits lower prediction errors than all other models. It can be seen that the supremacy of the DPM-EM model gets smaller the less observations are given. For few individual observations the linear mixed model with a normal random effects distribution is actually a bit better than the DPM-EM model (Table 4.3). The background for this feature is that the DPM-EM model detects sometimes more than one cluster in the data. Different patterns in the data are taken seriously.

		PE_0	PE_1	RB_0	RB_1
$\nu = 1$	normal	0.245	**0.111**	-0.012	0.016
	DPM-EM	0.273	0.123	-0.029	0.017
	$N = 3$	**0.236**	0.112	-0.019	0.010
	$N = 5$	0.271	0.125	-0.019	**0.001**
	$N = 10$	0.303	0.142	**-0.002**	0.004
$\nu = 3$	normal	0.160	0.037	**0.000**	0.023
	DPM-EM	0.153	0.036	-0.011	0.019
	$N = 3$	**0.129**	**0.030**	0.001	**0.014**
	$N = 5$	0.147	0.035	-0.001	0.017
	$N = 10$	0.153	0.037	-0.002	0.016
$\nu = 5$	normal	0.114	0.013	-0.002	0.010
	DPM-EM	**0.073**	0.009	0.001	0.028
	$N = 3$	0.076	**0.008**	-0.004	0.011
	$N = 5$	0.078	**0.008**	**0.000**	0.012
	$N = 10$	0.102	0.010	-0.003	**0.008**

Table 4.3.: Medians of PE_r and RB_r with $r = 0, 1$ for substantially overlapping clusters. Bold values indicate the best value in each case.

In summary, we draw the following conclusion: The DPM-EM models yield the better estimates for random effects − in terms of prediction errors − the clearer the clusters differ and the more observations are in the data. Especially in the case of many individual observations per subject it can only be outperformed by the model with $N = 3$ clusters that is the same as in the data generating process. Thus, the DPM-EM model turns out to be very flexible without risk of misspecifying the model like it can happen for the homogeneity model and the unpenalized heterogeneity model.

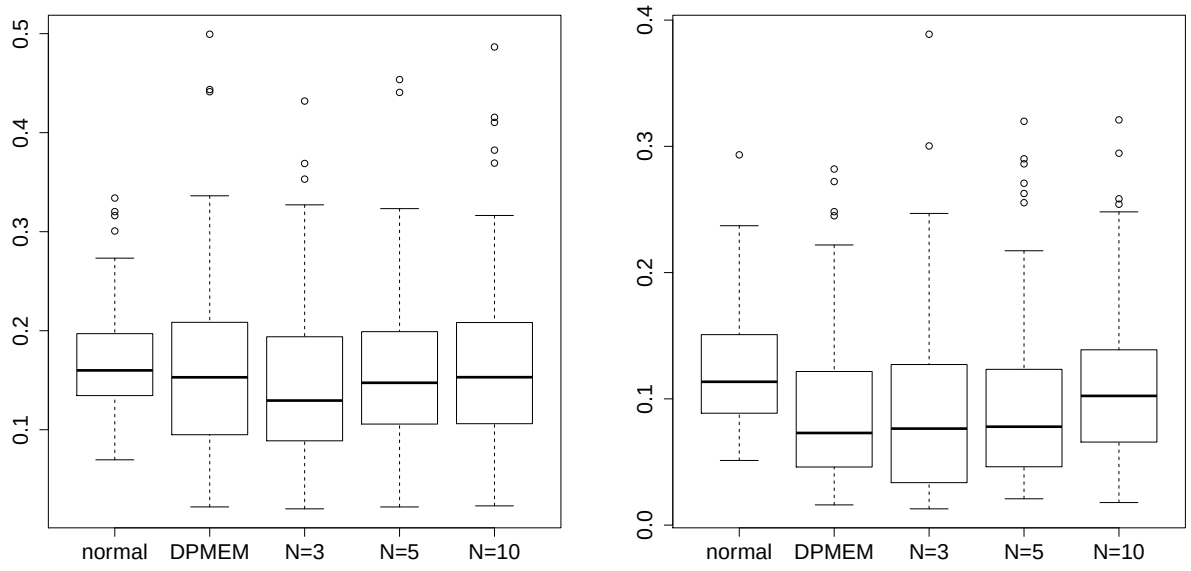
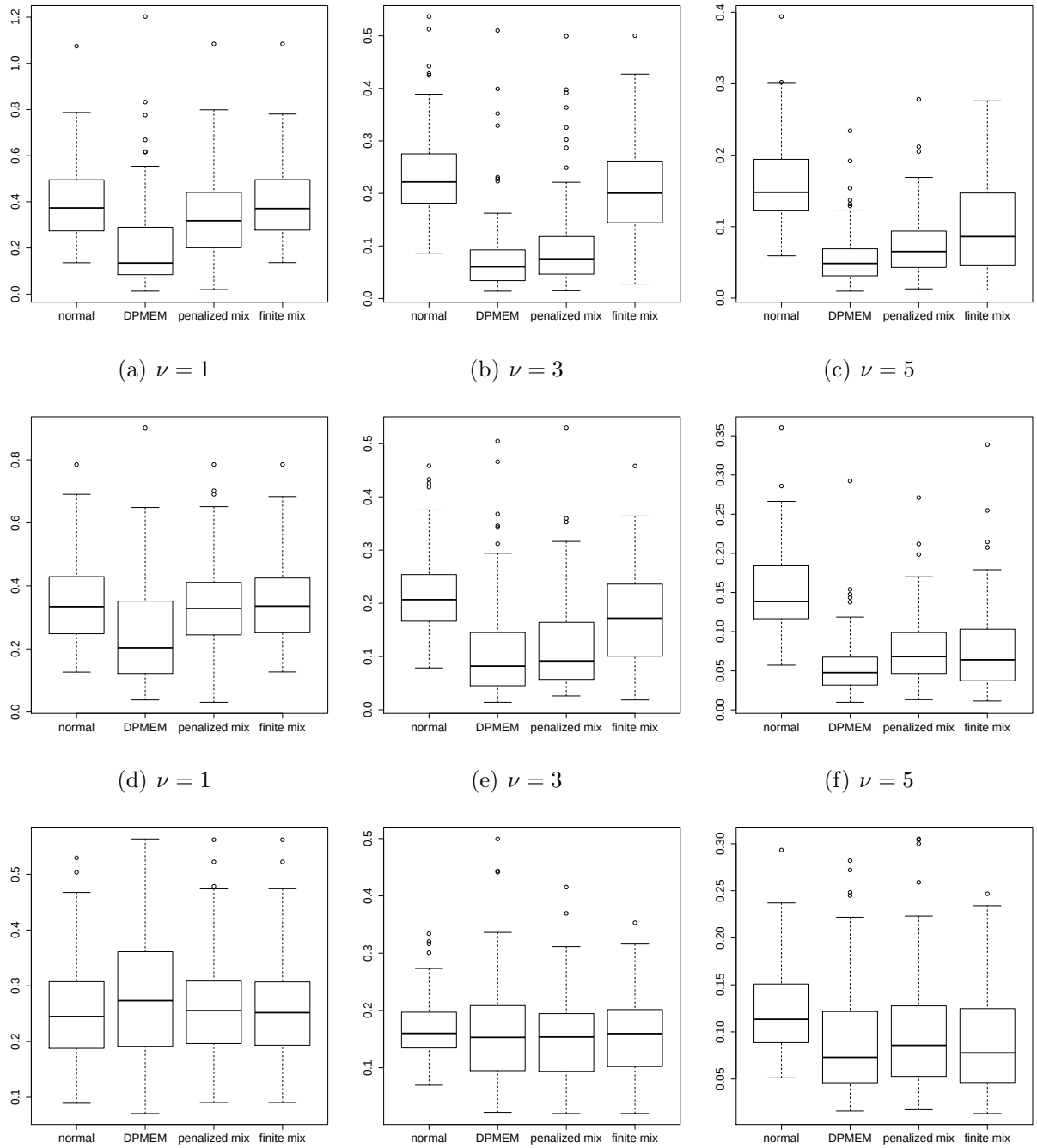


Figure 4.5.: Box plots of PE_0 with substantially overlapping clusters for a medium number of individual observations ($\nu = 3$) (left) respectively many individual observations ($\nu = 5$) (right).

4.3.3. Comparison of Simulation Results

In the following section, the DPM-EM model explained in Section 4.2 and the penalized heterogeneity model (penalized mix) based on the group fused lasso penalty from Chapter 3 are compared with regard to the prediction accuracy for the random effects and the clustering related characteristics. The simulation study is based on the settings in Section 4.3.1.

For comparison of the prediction accuracy of the random effects we use again the average prediction error (4.7) as criterion (Figure 4.6). For the sake of completeness the prediction errors of the homogeneity model with normally distributed random effects (normal) and of the heterogeneity model with a finite mixture distribution for the random effects (finite mix) are visualized, too. It should be noted that for the penalized mixture respectively the finite mixture approach the predictive cross-validation from Section 3.2.2 is used to determine the penalization parameter λ respectively the number of mixture components.

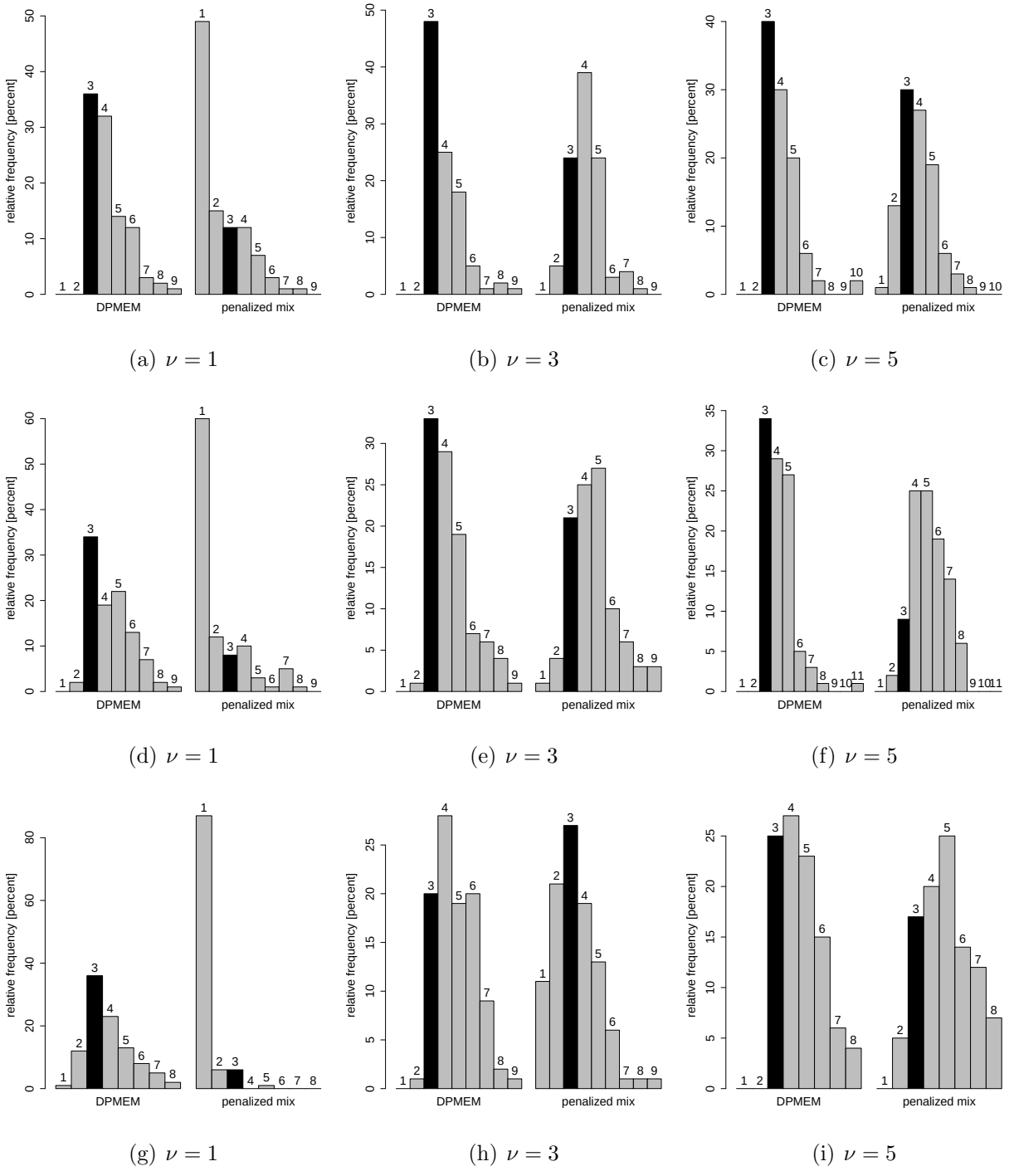


(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

(d) $\nu = 1$ (e) $\nu = 3$ (f) $\nu = 5$

(g) $\nu = 1$ (h) $\nu = 3$ (i) $\nu = 5$

Figure 4.6.: Box plots of PE_0 with clearly separated clusters ((a)-(c)), moderately separated clusters ((d)-(f)), and substantially overlapping clusters ((g)-(i)) for few individual observations (left), a medium number of individual observations (middle) and many individual observations (right).



(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

(d) $\nu = 1$ (e) $\nu = 3$ (f) $\nu = 5$

(g) $\nu = 1$ (h) $\nu = 3$ (i) $\nu = 5$

Figure 4.7.: Bar plots of the number of clusters with clearly separated clusters ((a)-(c)), moderately separated clusters ((d)-(f)), and substantially overlapping clusters ((g)-(i)) for few individual observations (left), a medium number of individual observations (middle) and many individual observations (right).

Regarding Figure 4.6 it can be seen that the DPM-EM model mostly yields considerably better predictions for the random effects than the penalized mixture approach based on the group fused lasso penalty term. Only in the case (g) a worse prediction accuracy of the DPM-EM model can be stated whereas apart from that especially for few individual observations the predominance of the DPM-EM model is evident ((a) and (d)). For a medium number of individual observations the performance of the DPM-EM model is quite similar to that of the penalized heterogeneity model ((b), (e) and (h)). In this setting, both approaches clearly outperform the classical linear mixed model with normally distributed random effects and the finite mixture approach, especially in the case of clearly and moderately separated clusters. Note that for many individual observations the assumption of a normal distribution for the random effects yields a considerably worse prediction accuracy than the DPM-EM model and the penalized heterogeneity model even for substantially overlapping clusters ((c), (f) and (i)).

In Figure 4.7 the estimated numbers of clusters for the 100 simulation runs are visualized by bar plots. Remember that the true random effects distribution is a normal mixture with three mixture components. For this reason the bar corresponding to three clusters is highlighted by black color. Except from the settings (h) and (i) most frequently three clusters are detected by the DPM-EM approach. In particular, the estimated number of clusters is hardly never one or two. The penalized heterogeneity approach is much more affected by the number of repeated measurements or the separation between the clusters than the DPM-EM approach. For few individual observation mostly all subjects are assigned to the same cluster while for many individual observations often more than three clusters are found. Generally, the variance of the number of clusters seems to be higher for the penalized heterogeneity approach in comparison to the DPM-EM model.

4.4. Applications

4.4.1. Unemployment

The proposed method is applied to two data examples. First, the variation of the unemployment over the federal states of Germany across time is considered (Weise et al., 2011). We examine the unemployment rate of each federal state from 2005 to 2010 in order to identify differences between states (Figure 3.1). Like in Section 3.3.1, where the unemployment data are analyzed by the penalized heterogeneity model based on a group fused lasso penalty, we consider a random slope model for the annual average of the unemployment rate y_{ij} of state i and measurement j

$$y_{ij}|\boldsymbol{b}_i \overset{ind.}{\sim} N(\beta_0 + b_{i0} + (\beta_1 + b_{i1})\texttt{year}_{ij},\, \sigma^2), \quad i = 1, \ldots, 16, \quad j = 0, \ldots, 5.$$

Since there is no symmetric unimodal variation of the individual intercepts around the overall mean it would not be appropriate to assume a Gaussian random effects distribution. Instead, the centered i.i.d. random effects $\boldsymbol{b}_i = (b_{i0}, b_{i1})^T$ follow a mixture distribution of Gaussian components with penalized mixture weights (4.4). The aim is to cluster the federal states in order to expose which states show similar behavior. Like in Section 3.3.1 only for a better interpretability we change the zero point of the time variable to 2005. Thus, during calculations the time variable is labeled by 0, 1, ..., 5 for the years 2005, 2006, ..., 2010.

	estimate	standard error	95%-CI lower	95%-CI upper
β_0	13.718	1.370	10.558	15.898
β_1	-1.007	0.111	-1.201	-0.765
σ^2	0.521	0.063	0.388	0.632
σ_0^2	1.084	0.883	0.036	2.813
σ_1^2	0.004	0.005	0.000	0.017
σ_{01}	-0.062	0.063	-0.203	0.013

Table 4.4.: Estimation results for the fixed effects and variance parameters by the DPM-EM model for the unemployment data.

First, Table 4.4 shows the estimated fixed effects and variance parameters. The associated standard errors and confidence intervals have been estimated by the nonparametric bootstrap method proposed by Efron (1979) using the Monte Carlo approximation with 1000 replications. In comparison to the results of the penalized heterogeneity approach in Section 3.3.1 it can be seen that the estimated standard errors of the DPM-EM model are somewhat larger in each case.

Our DPM-EM model detects three clusters with estimated weights $\hat{\pi}_1 = 0.467$, $\hat{\pi}_2 = 0.425$ and $\hat{\pi}_3 = 0.108$. Figure 4.8 shows the population effect (dashed line) as well as the cluster effects (solid lines). Observations belonging to the same cluster are marked with

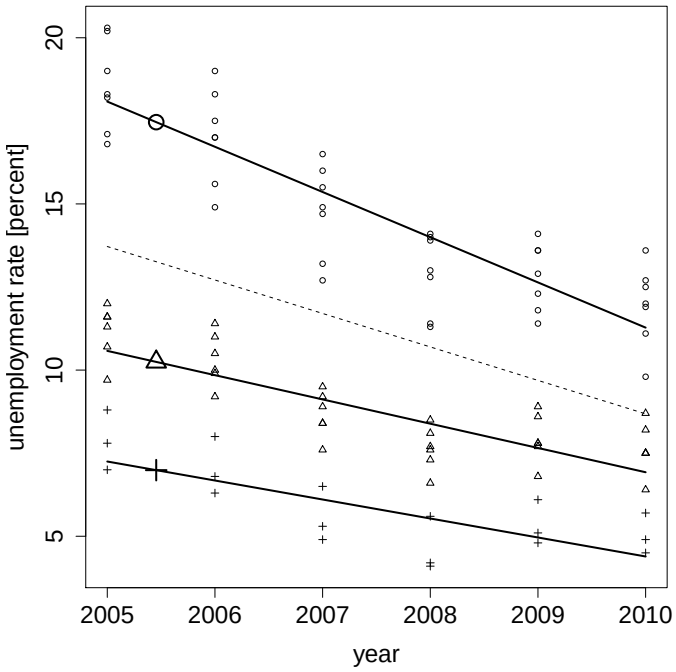


Figure 4.8.: Clustering of the unemployment data by the DPM-EM model. Observations belonging to the same cluster are marked with the same symbol. The dashed line represents the population effect, the solid lines symbolize the cluster effects.

the same symbol. For identification these symbols are also added to the corresponding solid lines. The southern federal states Bayern, Baden-Württemberg and Rheinland-Pfalz are assigned to cluster 3 (+) which features the lowest unemployment rate and the weakest decrease over time. As Table 4.5 shows, the base level in 2005 is -6.469 lower compared to the overall unemployment rate 13.718. In the south also the decrease of the unemployment rate is less distinct than in the other states. A similar effect can be observed in cluster 2 (△). Here, the gap to the global intercept is considerably smaller. Furthermore, there is one cluster (○) with a much higher base level and a stronger decrease of the unemployment rates. It is remarkable that these states are all in Eastern Germany or city states. Only the city state Hamburg makes an exception to that feature and belongs to cluster 2.

	$\hat{\boldsymbol{\mu}}_1$	$\hat{\boldsymbol{\mu}}_2$	$\hat{\boldsymbol{\mu}}_3$
Intercept	4.361	-3.140	-6.469
Slope	-0.353	0.277	0.436

Table 4.5.: Estimates of the cluster centers by the DPM-EM model for the unemployment data.

In summary, the clustering of the DPM-EM model differs from the result of penalized heterogeneity model based on a group fused lasso penalty in Section 3.3.1. Even though the penalized heterogeneity model detects three clusters, too, the partition of federals states is not the same. Obviously the DPM-EM model takes the different developments of the federal states Bayern, Baden-Württemberg and Rheinland-Pfalz compared to the other western states Schleswig-Holstein, Hamburg, Niedersachsen, Nordrhein-Westfalen, Hessen, and Saarland more seriously. In contrast, the penalized heterogeneity model emphasizes the special role of the city states Berlin and Bremen.

			cluster j		
			1	2	3
	1	Schleswig-Holstein	0	0.998	0.002
	2	Hamburg	0	1	0
	3	Niedersachsen	0	0.999	0.001
	4	Bremen	1	0	0
	5	Nordrhein-Westfalen	0	1	0
	6	Hessen	0	0.942	0.058
	7	Rheinland-Pfalz	0	0.424	0.576
state i	8	Baden-Württemberg	0	0.008	0.992
	9	Bayern	0	0.012	0.988
	10	Saarland	0	0.997	0.003
	11	Berlin	1	0	0
	12	Brandenburg	1	0	0
	13	Mecklenburg-Vorpommern	1	0	0
	14	Sachsen	1	0	0
	15	Sachsen-Anhalt	1	0	0
	16	Thüringen	1	0	0

Table 4.6.: Estimates for π_{ij} by the DPM-EM model for the unemployment data.

Table 4.6 shows the estimated probabilities $\hat{\pi}_{ij}$. Here, it can be seen that for most of the states the assignment to a specific cluster is very distinct. Only for Rheinland-Pfalz the probability for cluster 3 and cluster 2 is similar. The parameter α, which controls the number of clusters, is estimated by $\hat{\alpha} = 0.00155$. It is a typical feature that estimated αs are very small. This means that the strongest clustering as allowed by the data is the best one. Figure 4.9 visualizes the cluster history of the EM algorithm. Here, each cluster has its own symbol and its own shade. On the ordinate, the federal states of Germany are listed. See Table 4.6 for an overview which number represents which state. On the abscissa, the iterations of the EM algorithm are numbered. As mentioned in 4.2.2, at the beginning of the algorithm each state forms its own cluster. During the algorithm the clusters are successively fused according to an agglomerative clustering method. The final clustering after convergence can also be seen in Figure 4.8.

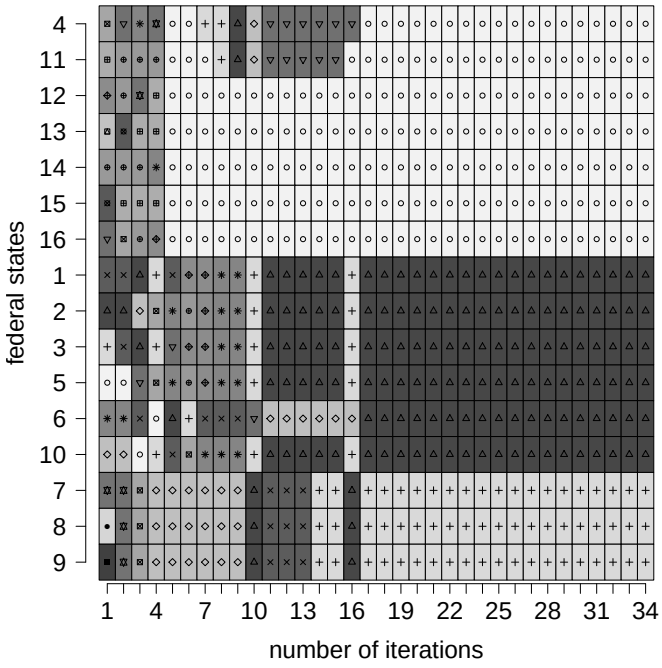


Figure 4.9.: Clustering history for the unemployment data during the DPM-EM algorithm. Each cluster has its own symbol and its own shade.

4.4.2. Lung Function Growth

In the second application, the lung function growth of girls in Topeka (USA) is examined by our DPM-EM model. These data are a subsample from the six cities study of air pollution and health in Dockery et al. (1983). The response variable is the logarithmic forced expiratory volume in one second (`fev1`). Our sample consists of 100 girls, with a minimum of two and a maximum of twelve observations over time. See Section 3.3.3 for more details on the data that are illustrated in Figure 3.9 (left). Like in Section 3.3.3 we use a linear mixed model with random intercepts and random slopes

$$\texttt{log(fev1)}_{ij}|\boldsymbol{b}_i \overset{ind.}{\sim} N(\beta_0 + b_{i0} + (\beta_1 + b_{i1})\texttt{age}_{ij}, \sigma^2), \quad i = 1, \ldots, 100, \quad j = 1, \ldots, n_i,$$

for modeling the logarithmic `fev1` subject to `age` and a DPM as random effects distribution as in equation (4.4). While the plot of all measurements over time (Figure 4.10) is not very informative because of the large number of measurements, the clustering effect of the DPM-EM model can be seen more easily from Figure 4.11. Here the axes represent the intercepts and slopes respectively. The square at coordinates (0,0) marks the population effect. All other icons are interpreted as deviations from the population effect. The thick big ones symbolize the cluster locations $\hat{\boldsymbol{\mu}}_h$, the thin small ones the random effects $\hat{\boldsymbol{b}}_i$.

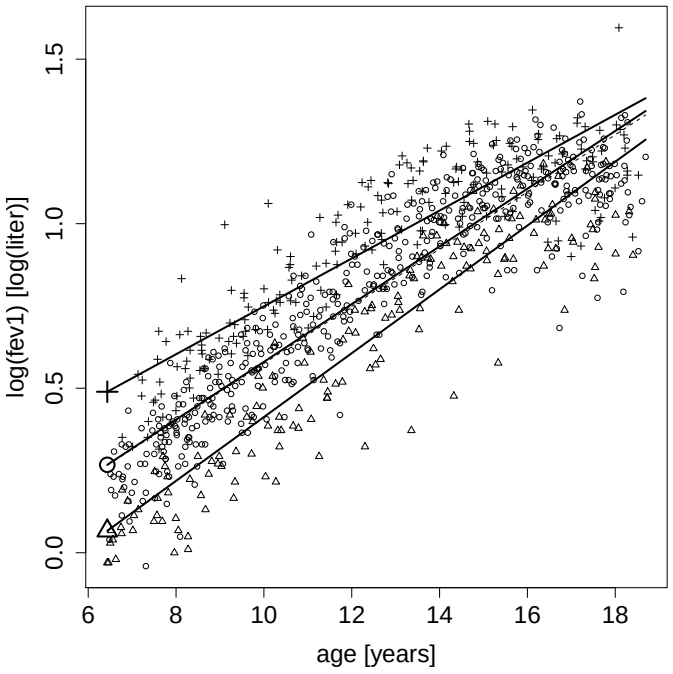


Figure 4.10.: Clustering of the lung function growth data by the DPM-EM model: Observations belonging to the same cluster are marked with the same symbol. The dashed line represents the population effect, the solid lines symbolize the cluster effects.

Girls which are assigned to the same cluster are marked with the same symbol and are arranged around the three cluster locations in the form of ellipses. See Section 3.3.3 for more information about the construction of the ellipses. While the penalized heterogeneity model in Figure 3.10 yields six clusters, by the DPM-EM model the number of clusters is reduced to three. It can also be seen that, while the penalized heterogeneity model tends to assign outliers to individual clusters due to the general lasso approach, the DPM-EM model rather forms clusters of comparable spatial extent. Furthermore the ellipses in Figure 4.11 are more "circular" than in Figure 3.10.

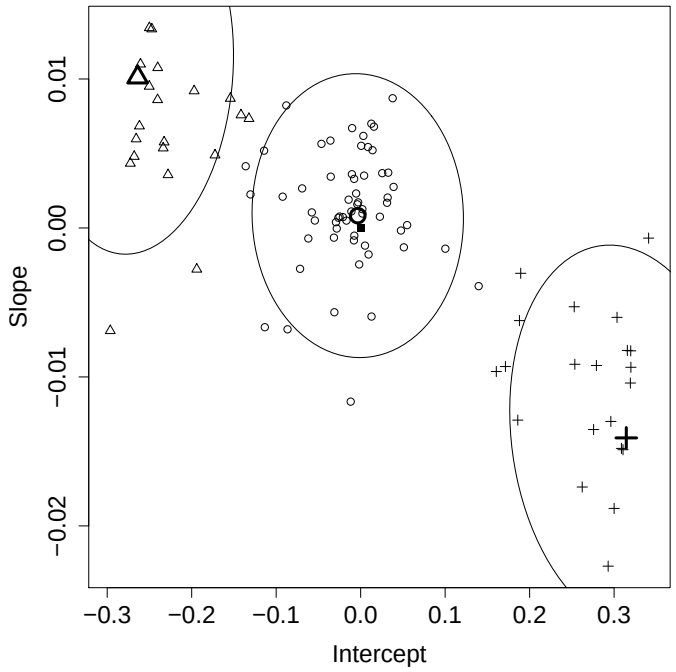


Figure 4.11.: Cluster centers and random effects of the DPM-EM model for the lung function growth data: The thick big icons symbolize the cluster centers $\hat{\boldsymbol{\mu}}_h$, the thin small ones the random effects $\hat{\boldsymbol{b}}_i$. The square at coordinates (0,0) marks the population effect. Ellipses with level 0.95 visualize the estimated conditional distribution of random effects in the clusters.

4.5. Summary and Discussion

We introduced linear mixed models with a DPM for the random effects distribution in order to penalize the number of clusters in the finite mixture of normal distributions. While models with Dirichlet processes are typically fitted by Bayesian methods like MCMC we used the EM algorithm because then the cluster property of the Dirichlet process can be used directly. So our method can be called an agglomerative clustering approach of individuals for longitudinal data. The DPM-EM algorithm itself was presented in detail. Furthermore, we showed in a simulation study that our approach outperforms the classical linear mixed model in the case of a underlying grouping structure. The DPM-EM model yielded even better prediction results than the penalized heterogeneity model in Chapter 3. Applications of this DPM-EM algorithm were demonstrated by considering unemployment data and lung function growth data. Extensions of this DPM-EM algorithm to additive mixed models follow in Chapter 6.

5. Additive Mixed Models with DPMs using MCMC methods

5.1. Introduction

In the previous two chapters linear mixed models were used for clustering longitudinal data. In this chapter and in Chapter 6 we consider the extension to additive mixed models to incorporate also nonlinear time trends. The methods proposed in the following chapters are mainly motivated by a study about childhood obesity, which has become a major public health issue in industrialized countries. We will analyze data from the LISA study (Influences of **L**ife-style factors on the development of the **I**mmune **S**ystem and **A**llergies in East and West Germany), a prospective birth cohort study conducted in four cities in Germany (Bad Honnef, Leipzig, Munich, Wesel) including 3097 healthy neonates born between 11/1997 and 01/1999. In this study longitudinal data on the body mass index (BMI) of children are collected along with covariates supposed to influence the nutritional status to gain a better understanding of factors determining the nutritional status of children. The data are collected in connection with nine mandatory medical examinations starting at birth and ending at 60 months. As a consequence, we are faced with a huge data set with complex structure comprising highly nonlinear growth patterns, long individual time series, clustered individual-specific deviations from the population trend and irregular time points.

While simple longitudinal data can often be fitted sufficiently well with growth curve models comprising individual-specific random effects, complex longitudinal data as in our application on childhood obesity often require a combination of semiparametric modeling of nonlinear trends and a flexible non-Gaussian random effects distribution, allowing to detect deviations from normality and clusters of individuals. This yields the longitudinal semiparametric regression model

$$y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + f(t_{ij}) + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i + \varepsilon_{ij}, \quad i = 1,\ldots,n, \quad j = 1,\ldots,n_i, \tag{5.1}$$

where y_{ij} denotes the BMI observed for a subject i, $i = 1,\ldots,n$, at observation times t_{ij}, $j = 1,\ldots,n_i$ with $t_{i1} < \ldots < t_{ij} < \ldots < t_{in_i}$, and $\varepsilon_{ij} \sim N(0,\sigma^2)$ are independent Gaussian measurement errors. Population effects of covariates $\boldsymbol{x}_{ij}$ such as gender or maternal smoking behavior are collected in the parameter vector $\boldsymbol{\beta}$ whereas individual-specific effects of covariates $\boldsymbol{z}_{ij}$ are represented in the parameter vector $\boldsymbol{b}_i$. In this chapter, we combine an efficient low-rank smoothing approach based on P-splines (Brezger and Lang, 2006; Jullion and Lambert, 2007) for the nonlinear trend function with a flexible DPM prior specifica-

tions for the random effects (Kleinman and Ibrahim, 1998; Jara, 2007). The specification of a P-spline for $f(t)$ considerably reduces the numerical complexity as compared to Bayesian smoothing splines since a much smaller number of parameters is involved and therefore systems of equations with a much smaller dimension have to solved iteratively during the MCMC run. P-splines also allow for a flexible exploration of possible trend patterns as opposed to restrictive parametric growth models. The DPM prior yields continuous random effects distributions (as compared to pure Dirichlet process approaches) and clustering of individuals can be achieved based on the truncated stick breaking prior representation of the Dirichlet process (Sethuraman, 1994). More information about Dirichlet processes can be found in Chapter 2. Our simulation results indicate that a DPM prior specification can safely be used even when the true data generating mechanism involves a parametric random effects distribution, but may yield a considerable improvement in estimation accuracy when deviations from a parametric distribution are present in the data.

In our application, the random effects part of the predictor will capture individual-specific deviations from the trend function $f(t)$, leading for example to $\boldsymbol{z}_{ij}^T\boldsymbol{b}_i = b_{i0} + t_{ij}b_{i1}$ for individual-specific linear deviations or

$$\boldsymbol{z}_{ij}^T\boldsymbol{b}_i = b_{i0} + t_{ij}b_{i1} + h(t_{ij})b_{i2}, \tag{5.2}$$

with a known nonlinear transformation $h(t)$ to gain additional flexibility. The specific form of the transformation $h(t)$ can for example be derived from exploratory analyzes of the data in a population model. We will primarily make use of model (5.2) to adapt individual-specific deviations to the structure of the trend observed in the obesity data. Models of the form (5.1) are mostly used in combination with a Gaussian (prior) distribution for the random effects, i.e. $\boldsymbol{b}_i$ i.i.d. $N(\boldsymbol{0}, \boldsymbol{D})$, see for example Lin and Zhang (1999), Fahrmeir and Lang (2001), Ruppert et al. (2003), Fahrmeir et al. (2004), Durban et al. (2005). While this choice is mathematically convenient, it may be questionable for several reasons in applied work. The normal distribution is symmetric and unimodal and has light tails. Since the distributional assumption is made on unobserved quantities, it is typically hard to validate these properties based on estimates. Possible skewness and multimodality (arising for example from an unconsidered grouping structure in the data) may be masked when checking the normal distribution in terms of estimated random effects.

Replacing the Gaussian random effects prior with a DPM allows to specify a hyperprior on the random effects (see Ferguson (1973) for the theory of Dirichlet processes). For linear mixed models, Dirichlet process priors for random effects were first proposed by Bush and MacEachern (1996), and the `DPpackage` in R (Jara, 2007) has options for Dirichlet process and DPM priors. As a consequence, the model becomes generally more robust since the Gaussian random effects model is encompassed in a hypermodel that allows to take deviations from normality into account. Moreover, the DPM prior specification naturally leads to clustering of the individuals in the data set with respect to their individual-specific effects. This is of particular interest in our application, where specific patterns of deviations from the population model shall be identified.

With a DPM prior specification, the random effects distribution is a parameter itself and, thus, a random measure in terms of the Bayesian paradigm. Simple Dirichlet processes will lead to a discrete distribution almost surely (Ferguson, 1974), but adding a mixing distribution stage allows to overcome this limitation. More specifically, consider $\theta_1, \ldots, \theta_n$ to be generated from a probability measure G with a Dirichlet process prior $G \sim DP$ as latent parameters of continuous random effects priors $p(\boldsymbol{b}_i|\theta_i)$, and, given θ_i, draw $\boldsymbol{b}_i$ from $p(\boldsymbol{b}_i|\theta_i)$. In hierarchical form we then have

$$\begin{aligned} G &\sim DP(\alpha, G_0), \\ \theta_i|G &\overset{i.i.d.}{\sim} G, & i = 1, \ldots, n, \\ \boldsymbol{b}_i|\theta_i &\overset{ind.}{\sim} p(\boldsymbol{b}_i|\theta_i), & i = 1, \ldots, n. \end{aligned}$$

As a consequence, the random effects distribution is a mixture of distributions with a Dirichlet process for the mixing distribution: $p(\boldsymbol{b}_i|G) = \int p(\boldsymbol{b}_i|\theta_i) dG(\theta_i)$.

In this DPM specification, each subject still has its own unique random effects value whereas choosing a Dirichlet process for the θ_i, $i = 1, \ldots, n$, creates ties among these and will therefore form clusters of subjects. In general, there are $k \leq n$ clusters and $\theta_1, \ldots, \theta_n$ can be represented by cluster locations $\mu_1, \ldots, \mu_k$ and cluster allocation variables $c_1, \ldots, c_n$. More specifically, $c_i \in \{1, \ldots, k\}$ denotes the cluster subject i belongs to, so that $\theta_i = \mu_{c_i}$. The strength of clustering is determined by the concentration parameter α which controls the confidence in the base distribution G_0. To match the standard assumption of mixed models, we will utilize Gaussian base distributions.

Li et al. (2010) consider a model that is comparable to (5.1), but differs from our specification with respect to some important points. First, they assume a Bayesian smoothing spline for the time trend $f(t)$ while we use a low-rank Bayesian P-spline yielding a more efficient representation of the nonlinear trend in terms of a manageable number of parameters. Second, they employ a Dirichlet process prior for the random effects distribution while our DPM prior allows to overcome the restriction to discrete random effects distributions imposed by the Dirichlet process prior. Third, our MCMC simulation algorithm is based on a truncated stick breaking representation of the Dirichlet process according to Sethuraman (1994) as compared to the Pólya urn scheme used by Li et al. (2010). All computations are implemented in C++, allowing for an efficient treatment of loop-intensive calculations, and are made easily accessible by providing the R wrapper function `ammDPMMCMC()` in the R package `clustmixed` (Heinzl, 2012).

The rest of this chapter is organized as follows: Section 5.2 considers the additive mixed model with DPM priors for the random effects in more detail. Section 5.2.1 deals with the model hierarchy of the additive mixed model and describes prior specifications for all model parameters as well as associated hyperparameter choices. The Gibbs sampler we use for inference is discussed in Section 5.2.2 and described in Section 5.2.3. The impact of deviations from a Gaussian random effects distribution is investigated in a simulation study in Section 5.3, focusing on the impact of the number of individual observations and the presence of more or less overlapping clusters. The main aim of this simulation study is

to detect situations in which DPM modeling is required to avoid considerable impact on the random effects estimation by misspecifying the prior as being Gaussian. Section 5.4 applies additive mixed models to the childhood obesity data with the specific aim to investigate specific patterns in the data utilizing the cluster property of the Dirichlet process prior. Section 5.5 concludes with a short summary of our main findings. Large parts of this chapter can be found in Heinzl et al. (2012).

5.2. Additive Mixed Models with DPM Priors

5.2.1. Model Hierarchy

Additive mixed models for longitudinal data extend linear mixed models by including smooth functions of time or of other continuous covariates as additional nonparametric effects in the predictor. We focus on modeling only a nonlinear time effect as in (5.1), but extensions to additive mixed models with additional nonlinear effects are conceptually straightforward. The prior choices discussed in the following and in particular choices for hyperparameters are to be understood as default values, but may have to be adapted to specific data situations. In most cases, the prior settings are designed to be either noninformative or to enforce specific properties of the posterior estimates such as the number of clusters for the random effects.

For fixed effects, we make the common assumption of a (weakly informative) Gaussian distribution $\boldsymbol{\beta}|\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$. Further hyperpriors are assigned to the parameters of the Gaussian distribution, yielding $\boldsymbol{\mu}_\beta \sim N(\boldsymbol{m}_\beta, \boldsymbol{S}_\beta)$ and $\boldsymbol{\Sigma}_\beta = \text{diag}(\sigma^2_{\beta_1}, \ldots, \sigma^2_{\beta_p})$ with inverse gamma priors $\sigma^2_{\beta_r} \overset{i.i.d.}{\sim} IG(a_\beta, b_\beta)$ for $r = 1, \ldots, p$. In our experience, the restriction to a diagonal matrix for $\boldsymbol{\Sigma}_\beta$ is more robust than assuming an inverse Wishart prior for a non-diagonal covariance matrix, in particular if the dimension p of fixed effects is large. Therefore, we will also use this restriction in the specification of the random effects prior later on. Gelman (2006) provides a critical discussion of the inverse gamma distribution as a conjugate prior in Gaussian prior structures and also considers a number of alternatives. To complete the prior specification for fixed effects, we suggest the following parameter defaults: $\boldsymbol{m}_\beta = \boldsymbol{0}_p$, $\boldsymbol{S}_\beta = 1000\boldsymbol{I}_p$ and $a_\beta = b_\beta = 0.005$.

For the random effects distribution $p(\boldsymbol{b}_i|\boldsymbol{\theta}_i)$, we assume a hierarchical Gaussian mixture prior

$$\begin{aligned} \boldsymbol{b}_i|\boldsymbol{\theta}_i, \boldsymbol{D} &\overset{ind.}{\sim} N(\boldsymbol{\theta}_i, \boldsymbol{D}), & i = 1, \ldots, n, \\ \boldsymbol{\theta}_i|G &\overset{i.i.d.}{\sim} G, & i = 1, \ldots, n, \\ G &\sim DP(\alpha, G_0). \end{aligned}$$

Inference for the latent parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ is based on the stick breaking representation of the Dirichlet process (Sethuraman, 1994) in its truncated version (Ishwaran and James, 2001), where

$$G = \sum_{h=1}^{N} \pi_h \, \delta_{\boldsymbol{\mu}_h},$$

and $\delta_{\boldsymbol{\mu}_h}$ denotes the Dirac measure in $\boldsymbol{\mu}_h$. Hence, the unknown distribution G is represented as a weighted sum of point masses with random weights π_h, which are independent of locations $\boldsymbol{\mu}_h$. The locations are i.i.d. random variables from the base distribution G_0, i.e. $\boldsymbol{\mu}_h \overset{i.i.d.}{\sim} G_0, \quad h = 1, \ldots, N$, while weights are constructed through a stick breaking procedure with beta distributed weights:

$$\begin{array}{rcll} \pi_h & = & v_h \prod_{l<h}(1 - v_l), & h = 1, \ldots, N, \\ v_h & \overset{i.i.d.}{\sim} & Be(1, \alpha), & h = 1, \ldots, N-1, \quad v_N = 1. \end{array}$$

Sethuraman (1994) showed that (in the limit $N \to \infty$) the probability measure of G is given by $DP(\alpha, G_0)$. See Section 2.2 for more details about the stick breaking procedure. The truncated version still is a good approximation for G because the random weights decrease stochastically as the index h grows, compare for example Ishwaran and James (2001, Theorem 1). In particular, $\mathbb{E}(\sum_{h=N+1}^{\infty} \pi_h)$ converges to zero exponentially with $N \to \infty$ (see equation (2.4)), i.e. the tail probability converges to zero with increasing N. Following the arguments in Ohlssen et al. (2007) and in Section 2.2 based on this formula, we truncate the stick breaking procedure at $N = 100$ in our simulations and applications.

The main advantage of the truncated representation is that the number of resulting parameters is high-dimensional but finite, enabling the construction of a blocked Gibbs sampler for $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_N^T)^T$, $\boldsymbol{v} = (v_1, \ldots, v_{N-1})^T$ and $\boldsymbol{c} = (c_1, \ldots, c_n)^T$. In addition, one obtains estimates for $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ via $\boldsymbol{\theta}_i = \boldsymbol{\mu}_{c_i}$. Contrary to a Pólya urn Gibbs sampling scheme, the stick breaking representation also offers the possibility to estimate G itself (see Ishwaran and James (2001) for other advantages of blocked Gibbs sampling over Pólya urn Gibbs sampling).

For the base distribution, we assume a multivariate normal distribution $G_0 = N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ with hyperpriors $\boldsymbol{\mu}_0 \sim N(\boldsymbol{m}_0, \boldsymbol{S}_0)$ and $\boldsymbol{\Sigma}_0 = \text{diag}(\sigma^2_{0_0}, \ldots, \sigma^2_{0_{q-1}})$ with $\sigma^2_{0_r} \overset{i.i.d.}{\sim} IG(a_0, b_0)$ for $r = 0, \ldots, q-1$. In analogy to the specifications for the fixed effects hyperpriors, we suggest the following default values for hyperparameters: $\boldsymbol{m}_0 = \boldsymbol{0}_q$, $\boldsymbol{S}_0 = 100\boldsymbol{I}_q$, $a_0 = b_0 = 0.5$. For the prior covariance of the random effects, we also assume a diagonal structure, leading to $\boldsymbol{D} = \text{diag}(\sigma^2_{b_0}, \ldots, \sigma^2_{b_{q-1}})$ with $\sigma^2_{b_r} \overset{i.i.d.}{\sim} IG(a_b, b_b)$ for $r = 0, \ldots, q-1$ and $a_b = b_b = 0.0001$. The different prior choices for a_0 and b_0 on the one hand and a_b and b_b on the other hand reflect our prior preference for a small variance within clusters in contrast to a high variance between clusters. In Section 5.4, it can be seen that this prior structure yields a higher power for detecting clusters in the data than other prior settings.

For the concentration parameter α, we consider a discrete prior $\alpha \sim \sum_{\omega \in \Omega} P(\alpha = \omega)\, \delta_\omega$ with support $\Omega = \{0.5, 0.6, \ldots, 99.9, 100\}$ and probabilities which resemble a gamma distribution. This specification avoids difficulties with too small values for α that would naturally appear in a blocked Gibbs sampler with a gamma prior. For illustration, assume that $\alpha \sim Ga(a_\alpha, b_\alpha)$. In this case, the full conditionals for v_h, $h = 1, \ldots, N-1$ are given by

$$\begin{array}{rcl} v_h | \boldsymbol{c}, \alpha & \sim & Be(1 + n_h,\ \alpha + \sum_{l=h+1}^{N} n_l), \\ \alpha | \boldsymbol{v} & \sim & Ga(N - 1 + a_\alpha,\ b_\alpha - \sum_{h=1}^{N-1} \log(1 - v_h)), \end{array}$$

where n_h denotes the number of subjects in cluster h. Updating v_h in stick breaking representation for a small value of α near zero could lead to $v_h^* = 1$, where h^* represents an empty cluster followed by further empty clusters. This results in $\alpha = 0$ in the next step and so there is at least one improper $Be(\cdot, 0)$ full conditional for v_h^* in the next update if the last cluster N is empty. Excluding small values for α allows us to avoid such problems.

The prior for the concentration parameter of course influences the resulting number of clusters. Liu (1996) derived a direct relation between the expected value $\mathbb{E}(k)$ and the variance $\mathrm{Var}(k)$ of the number of clusters k and the concentration parameter α. If there is prior information about $\mathbb{E}(k)$ and $\mathrm{Var}(k)$, this relation may be used in specifying a prior for α. Without such knowledge, however, it is in general not possible to define some optimal prior. We follow the recommendation of Ishwaran and James (2002) and choose a discrete prior that resembles a $Ga(2, 2)$ prior as a standard option. In our application in Section 5.4, we investigate sensitivity with regard to the hyperparameters. Additional information about the prior for α can also be found in Section 4.3 of Ohlssen et al. (2007).

The nonlinear time trend $f(t)$ is modeled through a Bayesian P-spline. That is we assume that $f(t)$ can be represented through $f(t) = \sum_{s=1}^{d} \gamma_s B_s^l(t)$, where $B_s^l(t)$ are B-spline basis functions of degree l defined for a grid of knots on the time scale. Collecting observations y_{ij}, $j = 1, \ldots, n_i$, for individual i in the vector $\boldsymbol{y}_i$, model (5.1) can be written in matrix notation as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{\gamma} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n, \tag{5.3}$$

with $\boldsymbol{\varepsilon}_i \overset{i.i.d.}{\sim} \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Here, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ denote the individual design matrices constructed from covariates $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$, $\boldsymbol{B}_i$ denotes the matrix of B-spline basis functions of subject i and $\boldsymbol{\gamma}$ denotes the vector of basis function coefficients. In our setting, there are m equidistant inner knots and $d = m + l - 1$ B-spline basis functions of degree l.

For Bayesian P-splines (Lang and Brezger, 2004), a Gaussian smoothness prior $p(\boldsymbol{\gamma}|\tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}^T\boldsymbol{K}\boldsymbol{\gamma}\right)$ is assumed for the vector of basis coefficients. The precision matrix acts as a penalty matrix to enforce smoothness and is defined through $\boldsymbol{K} = \boldsymbol{\Delta}^T\boldsymbol{\Delta}$, where $\boldsymbol{\Delta}$ is a first or second order difference matrix for adjacent B-spline coefficients. The variance (or inverse smoothing) parameter τ^2 controls the amount of smoothness. The negative log-penalty $\boldsymbol{\gamma}^T\boldsymbol{K}\boldsymbol{\gamma}$ corresponds exactly to the penalty term introduced by Eilers and Marx (1996) in a frequentist penalized likelihood setting. For the variance parameter, we assume an inverse gamma prior $\tau^2 \sim IG(a_\gamma, b_\gamma)$. As a standard option, we use $a_\gamma = b_\gamma = 0.0001$, $m = 12$, $l = 3$ and a second order difference penalty for the spline function.

Finally, the error variance σ^2 is also assigned an inverse gamma distribution, i.e. $\sigma^2 \sim IG(a_\varepsilon, b_\varepsilon)$ with default values $a_\varepsilon = b_\varepsilon = 0.005$.

5.2.2. Inference

From the prior choices we have employed in the previous section, a convenient MCMC sampler results since all full conditionals are available in closed form such that a blocked Gibbs sampler can be utilized. Before full details of this sampler are provided in Section 5.2.3,

some comments on special properties and peculiarities of the algorithm will be given in the following.

In general, inference in additive mixed models has to deal with an identifiability problem. Consider, as a typical example, the additive mixed model (5.1) with population trend $f(t)$ modeled through a Bayesian P-spline with second order random walk prior or a Bayesian cubic smoothing spline. Suppose, we want to include individual-specific linear trends $b_{i0} + b_{i1}t$ in addition to the population trend, then we are faced with the following problem: Without further restrictions, P-splines and smoothing splines comprise linear trends as special cases. There are (at least) two possible strategies: Either the population trend $f(t)$ models only deviations from a linear population trend or random intercepts and slopes have to be centered around zero, modeling only individual linear deviations from $f(t)$. Li et al. (2010) deal with this problem in a post-processing step while Dunson et al. (2007) introduce a centered Dirichlet process prior. We suggest centering random effects about zero in the MCMC algorithm as a simple but effective device. Specifying $f(t)$ as a non-linear deviation from a linear population trend would require additional linear constraints for B-splines with a second order random walk prior or for cubic smoothing splines, see Panagiotelis and Smith (2008). For linear regression splines represented through a truncated power series basis, a simple approach is to delete the "fixed" linear effect corresponding to the basis functions 1 and t. Obviously the same strategies for assuring identification are relevant for nonlinear functions of other continuous covariates.

Note that for Gibbs sampling it is necessary to define working responses as partial residuals in order to take the remaining parameters into account when updating parameters corresponding to the P-spline, the fixed effects and the random effects. In this context centering random effects implies that these parts of the model can no longer be updated in arbitrary order. It is essential that updating P-spline parameters follows updating random effects so that the basis function coefficients can absorb the general time trend. Moreover, centering random effects has another important implication: For updating the observation variance σ^2 the uncentered random effects have to be used. Otherwise drawn values for σ^2 would be too high in the beginning of the Markov chain and the convergence of the samples would slow down.

In all computations in this paper, we used 55,000 iteration with 5,000 burn in and thinned the Markov chains by a factor of 50, resulting in samples of size 1,000 for inference. The convergence and mixing was assessed via empirical autocorrelations and the inspection of sampling paths.

5.2.3. Block Gibbs Algorithm

For the description of the MCMC-sampler, the additive mixed model (5.3) is written in matrix notation as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon},$$

where all individual vectors and matrices are stacked, e.g. $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$, except for $\boldsymbol{Z} = \text{diag}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$. The derivations of all full conditionals are given in Appendix A.7. Let the current state of the Markov chain consist of $\boldsymbol{\gamma}$, τ^2, $\boldsymbol{\beta}$, $\boldsymbol{\mu}_\beta$, $\boldsymbol{\Sigma}_\beta$, $\boldsymbol{b}$, $\boldsymbol{D}$, $\boldsymbol{\mu}$, $\boldsymbol{c}$, $\boldsymbol{v}$, α, $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$ and σ^2 and iterate through the following steps:

1. Update parameters referring to the P-spline:
   - Create working response $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b}$.
   - Draw new values for $\boldsymbol{\gamma}$ and τ^2 using:
     - $\boldsymbol{\gamma}|\tau^2, \boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{y}, \sigma^2 \sim N(\boldsymbol{\mu}_\gamma^*, \boldsymbol{\Sigma}_\gamma^*)$,
       $\boldsymbol{\mu}_\gamma^* = (\frac{1}{\tau^2}\boldsymbol{K} + \frac{1}{\sigma^2}\boldsymbol{B}^T\boldsymbol{B})^{-1}\frac{1}{\sigma^2}\boldsymbol{B}^T\tilde{\boldsymbol{y}}$,
       $\boldsymbol{\Sigma}_\gamma^* = (\frac{1}{\tau^2}\boldsymbol{K} + \frac{1}{\sigma^2}\boldsymbol{B}^T\boldsymbol{B})^{-1}$,
     - $\tau^2|\boldsymbol{\gamma} \sim IG\left(a_\gamma + 0.5\,(d-k), b_\gamma + 0.5\,\boldsymbol{\gamma}^T\boldsymbol{K}\boldsymbol{\gamma}\right)$.
2. Update parameters referring to fixed effects:
   - Create working response $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{B}\boldsymbol{\gamma} - \boldsymbol{Z}\boldsymbol{b}$.
   - Draw new values for $\boldsymbol{\beta}$, $\boldsymbol{\mu}_\beta$ and $\boldsymbol{\Sigma}_\beta$ using:
     - $\boldsymbol{\beta}|\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta, \boldsymbol{\gamma}, \boldsymbol{b}, \boldsymbol{y}, \sigma^2 \sim N(\boldsymbol{\mu}_\beta^*, \boldsymbol{\Sigma}_\beta^*)$,
       $\boldsymbol{\mu}_\beta^* = (\boldsymbol{\Sigma}_\beta^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \frac{1}{\sigma^2}\boldsymbol{X}^T\tilde{\boldsymbol{y}})$,
       $\boldsymbol{\Sigma}_\beta^* = (\boldsymbol{\Sigma}_\beta^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X})^{-1}$,
     - For $r = 1, \ldots, p$:
       - $\mu_{\beta r}|\sigma^2_{\beta_r}, \beta_r \sim N\left((\frac{1}{\sigma^2_{\beta_r}} + \frac{1}{s^2_{\beta_r}})^{-1}(\frac{\beta_r}{\sigma^2_{\beta_r}} + \frac{m_{\beta r}}{s^2_{\beta_r}}), (\frac{1}{\sigma^2_{\beta_r}} + \frac{1}{s^2_{\beta_r}})^{-1}\right)$,
       - $\sigma^2_{\beta_r}|\mu_{\beta r}, \beta_r \sim IG\left(a_\beta + 0.5,\, b_\beta + 0.5(\beta_r - \mu_{\beta r})^2\right)$.
3. Update parameters referring to random effects:
   - Create working response $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{B}\boldsymbol{\gamma}$.
   - For $i = 1, \ldots, n$: Draw a new value for $\boldsymbol{b}_i$ using:
     $\boldsymbol{b}_i|\boldsymbol{\theta}_i, \boldsymbol{D}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{y}_i, \sigma^2 \sim N(\boldsymbol{\theta}_i^*, \boldsymbol{D}_i^*)$,
     $\boldsymbol{\theta}_i^* = (\boldsymbol{D}^{-1} + \frac{1}{\sigma^2}\boldsymbol{Z}_i^T\boldsymbol{Z}_i)^{-1}(\boldsymbol{D}^{-1}\boldsymbol{\theta}_i + \frac{1}{\sigma^2}\boldsymbol{Z}_i^T\tilde{\boldsymbol{y}}_i)$,
     $\boldsymbol{D}_i^* = (\boldsymbol{D}^{-1} + \frac{1}{\sigma^2}\boldsymbol{Z}_i^T\boldsymbol{Z}_i)^{-1}$.
   - Centering:
     - Create mean $\bar{\boldsymbol{b}}$.
     - For $i = 1, \ldots, n$: Replace $\boldsymbol{b}_i$ by $\boldsymbol{b}_i - \bar{\boldsymbol{b}}$.

- For $h = 1, \ldots, N$: Draw a new value for $\boldsymbol{\mu}_h$ using:
  - If $\nexists\, i : c_i = h$:
    $$\boldsymbol{\mu}_h | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0 \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).$$
  - If $\exists\, i : c_i = h$: For $r = 0, \ldots, q-1$:
    $$\mu_{hr} | \sigma^2_{b_r}, \mu_{0r}, \sigma^2_{0_r}, \boldsymbol{b}, \boldsymbol{c} \sim N(\mu^*_{0r}, \sigma^{2*}_{0_r}),$$
    $$\mu^*_{0r} = (\tfrac{n_h}{\sigma^2_{b_r}} + \tfrac{1}{\sigma^2_{0_r}})^{-1} (\tfrac{n_h}{\sigma^2_{b_r}} \bar{b}_{r,h} + \tfrac{\mu_{0r}}{\sigma^2_{0_r}}),$$
    $$\sigma^{2*}_{0_r} = (\tfrac{n_h}{\sigma^2_{b_r}} + \tfrac{1}{\sigma^2_{0_r}})^{-1}.$$
- For $i = 1, \ldots, n$:
  - Draw a new value for c_i using:
    $$c_i | \boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{b}_i, \boldsymbol{D} \sim \textstyle\sum_{h=1}^{N} c^\star f(\boldsymbol{b}_i | \boldsymbol{\mu}_h, \boldsymbol{D}) \pi_h \delta_h,$$
    $f \mathrel{\widehat{=}}$ multivariate normal density,
    $c^\star \mathrel{\widehat{=}}$ constant so that the sum of probabilities is 1,
  - Set $\boldsymbol{\theta}_i = \boldsymbol{\mu}_{c_i}$.
- For $h = 1, \ldots, N$:
  - Draw a new value for v_h (except for $h = N$) using:
    $$v_h | \boldsymbol{c} \sim Be(1 + n_h, \alpha + \textstyle\sum_{l=h+1}^{N} n_l),$$
  - Create π_h using:
    $$\pi_h = v_h \textstyle\prod_{l<h} (1 - v_l).$$
- Draw a new value $\omega \in \Omega = \{0.5, 0.6, \ldots, 99.9, 100\}$ for α using:
  $$\alpha | \boldsymbol{v} \sim \textstyle\sum_{\omega \in \Omega} \exp\left((N-1) \log \omega + (\omega - 1) \sum_{h=1}^{N-1} \log(1 - v_h) \right) \cdot P(\alpha = \omega) \delta_\omega.$$
- For $r = 0, \ldots, q-1$: Draw new values for μ_{0r}, $\sigma^2_{0_r}$ and $\sigma^2_{b_r}$ using:
  - $\mu_{0r} | \sigma^2_{0_r}, \boldsymbol{\theta} \sim N\left((\tfrac{n}{\sigma^2_{0_r}} + \tfrac{1}{s^2_{0_r}})^{-1} (\tfrac{n}{\sigma^2_{0_r}} \bar{\theta}_r + \tfrac{m_{0r}}{s^2_{0_r}}), (\tfrac{n}{\sigma^2_{0_r}} + \tfrac{1}{s^2_{0_r}})^{-1} \right)$,
  - $\sigma^2_{0_r} | \mu_{0r}, \boldsymbol{\theta} \sim IG\left(a_0 + 0.5\, n,\ b_0 + 0.5 \textstyle\sum_{i=1}^{n} (\theta_{ir} - \mu_{0r})^2\right)$,
  - $\sigma^2_{b_r} | \boldsymbol{\theta}, \boldsymbol{b} \sim IG\left(a_b + 0.5\, n,\ b_b + 0.5 \textstyle\sum_{i=1}^{n} (b_{ir} - \theta_{ir})^2\right)$.

4. Update the error variance:
   - Create working response $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{B}\boldsymbol{\gamma} - \boldsymbol{Z}\bar{\boldsymbol{b}} - \boldsymbol{Z}\boldsymbol{b}$.
   - Draw a new value for σ^2 using:
     $$\sigma^2 | \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}, \boldsymbol{y} \sim IG\left(a_\varepsilon + 0.5\, n_d, b_\varepsilon + 0.5\, \tilde{\boldsymbol{y}}^T \tilde{\boldsymbol{y}}\right).$$

5.3. Simulation Study

5.3.1. Settings

The following simulation study aims at clarifying in which data situations a DPM random effects prior substantially improves estimation compared to the commonly used Gaussian random effects prior. More specifically, we are interested in the ability of the DPM specification to detect deviations from normality and to identify clusters of random effects in the data. In fact, it has been observed that in some situations the empirical distribution of estimated random effects based on a Gaussian prior is quite close to the empirical distribution obtained with Dirichlet process or DPM priors. As an illustration, Figure 5.1 shows a kernel density estimate for the estimated random intercepts in an additive mixed model with random slopes for simulated data, where the true random effects distribution is a Gaussian mixture (the generation of these data will be explained later on in this section). Even in the case of a Gaussian random effects prior, the kernel density estimate shows a bimodal form that reflects the true mixing distribution quite well. The reason for this surprising flexibility is that each random effect has its own posterior density – linked only by variance parameters – even if all random effects have the same unimodal Gaussian prior. In this context, traditional random effects assumptions are less restrictive than one would generally expect. Hence, the question arises in which situations DPM priors actually improve upon Gaussian priors and whether we fit overly flexible models when the true data generating model is close to Gaussian. We will therefore investigate the impact of the number of observations within clusters and the separation between clusters.

We generated data sets assuming an additive mixed model

$$y_{ij} = f(t_{ij}) + b_{i0} + t_{ij} b_{i1} + \varepsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i,$$

with i.i.d. errors $\varepsilon_{ij} \sim N(0, \sigma^2)$ and excluding any fixed effects. As nonlinear trend function we consider $f(t) = \frac{50 \cdot \log(0.2\,t+1)}{(0.2\,t+1)^2}$ which looks similar to the trend curve in equation (5.4) from Section 5.4. For estimation, the function is approximated by a cubic P-spline with twelve inner knots and second order random walk penalty. The i.i.d. random effects $\boldsymbol{b}_i = (b_{i0}, b_{i1})^T$ follow a mixture distribution with three Gaussian components:

$$\boldsymbol{b}_i \sim 0.4\, N(\boldsymbol{\mu}_1, \boldsymbol{D}) + 0.3\, N(\boldsymbol{\mu}_2, \boldsymbol{D}) + 0.3\, N(\boldsymbol{\mu}_3, \boldsymbol{D}), \quad i = 1, \ldots, n,$$

imitating a population consisting of three clusters of overlapping subpopulations.

Throughout the simulations, we set $n = 20$, $\sigma^2 = 0.25$ and $\boldsymbol{D} = \mathrm{diag}(0.1, 0.1)$. We vary, however, the number of individual observations n_i, the centers $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ of the clusters and the locations of observation times t_{ij}. To produce longitudinal data with varying numbers of repeated observations per unit i, we set $n_i = 2 + X_i$, where X_i is Poisson distributed with rate ν. Setting $\nu = 0.5$ corresponds to longitudinal data with only *few individual observations* (2.5 on average), $\nu = 2.5$ to a *medium number of individual*

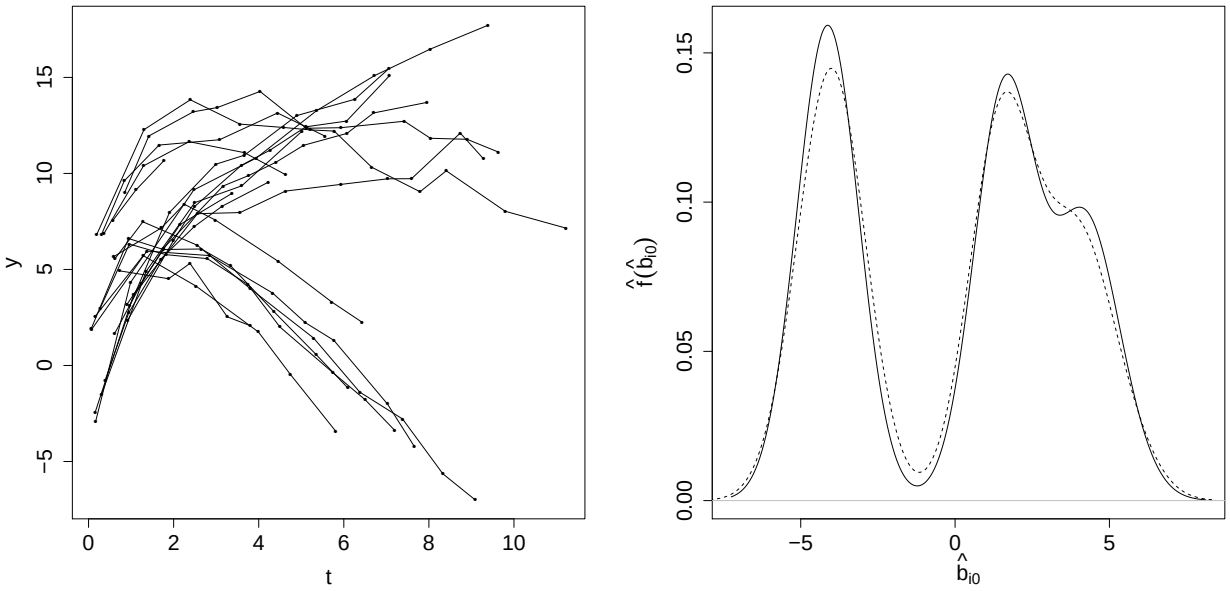


Figure 5.1.: Trace plot of simulated data for many individual observations ($\nu = 5$) with clearly separated clusters (left) and the corresponding kernel density estimate of $\hat{b}_{i0}$ with Gaussian kernel and bandwidth = 1 (right) for a DPM (solid line) and a normal distribution (dashed line) as random effects distribution.

observations and $\nu = 5$ to (comparably) *many individual observations*. For given n_i, observation times are generated from

$$\begin{aligned} t_{i1} &\sim U(0,1), \quad i = 1, \ldots, n, \\ t_{ij} &\sim U(t_{i,j-1} + 0.5, t_{i,j-1} + 1.5), \quad i = 1, \ldots, n, \quad j = 2, \ldots, n_i. \end{aligned}$$

In this way, different numbers $n_i^{(s)}$ and $t_{ij}^{(s)}$ are generated in each simulation run $s = 1, \ldots, 100$. Similarly, different "true" random effects $\boldsymbol{b}_i^{(s)}$ are drawn from the Gaussian mixture distribution in each simulation run. For the cluster locations, we choose

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -4.5 \\ 1.5 \end{pmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 1.5 \\ -1.8 \end{pmatrix}, \qquad \boldsymbol{\mu}_3 = \begin{pmatrix} 4.5 \\ -0.2 \end{pmatrix},$$

corresponding to *clearly separated clusters (case 1)*,

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -1.5 \\ 0.75 \end{pmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 0.5 \\ -0.9 \end{pmatrix}, \qquad \boldsymbol{\mu}_3 = \begin{pmatrix} 1.5 \\ -0.1 \end{pmatrix},$$

corresponding to *moderately separated clusters (case 2)*, and

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -0.3 \\ 0.375 \end{pmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 0.1 \\ -0.45 \end{pmatrix}, \qquad \boldsymbol{\mu}_3 = \begin{pmatrix} 0.3 \\ -0.05 \end{pmatrix},$$

corresponding to *substantially overlapping clusters (case 3).*

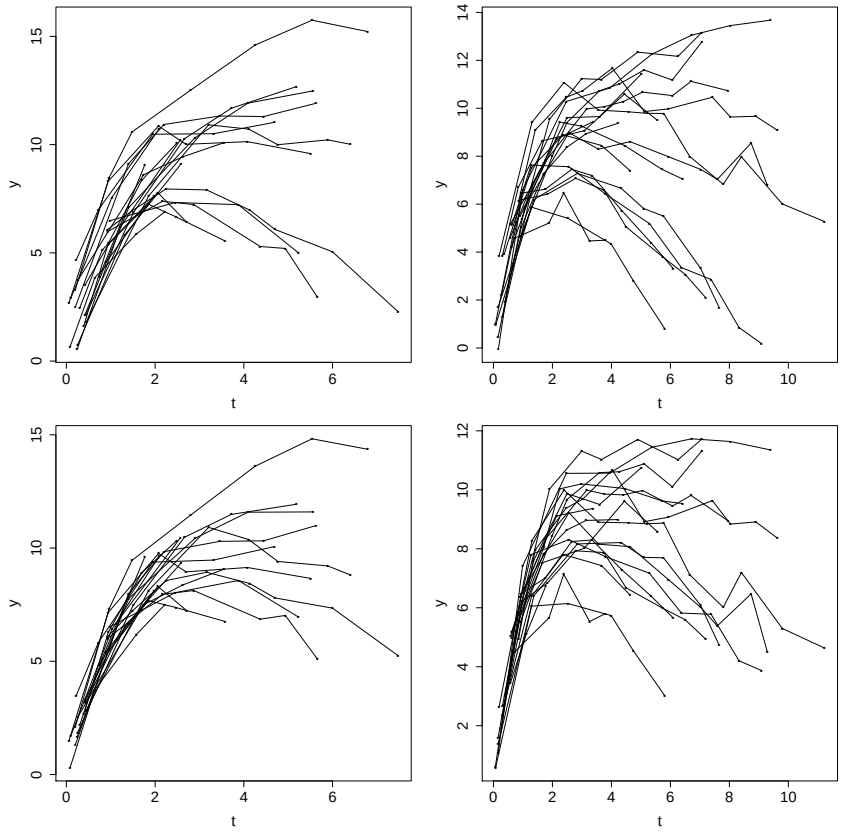


Figure 5.2.: Trace plots for simulated data with a medium number of individual observations ($\nu = 2.5$) (left) and many individual observations ($\nu = 5$) (right) with moderately separated (top) and substantially overlapping clusters (bottom).

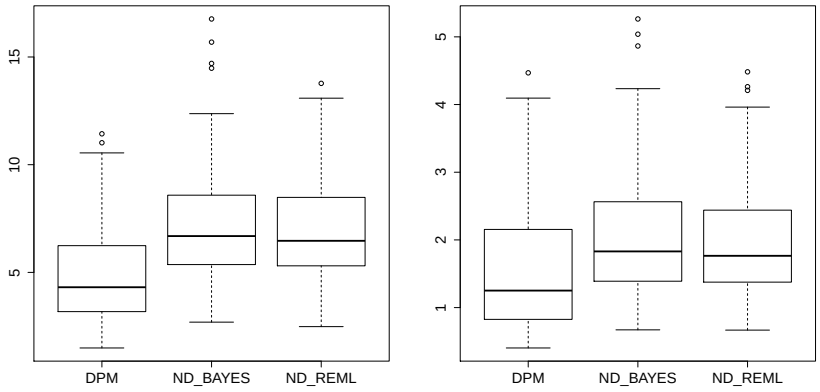
Combining these different settings for observation times and clusters results in nine different scenarios. For each of them, we compare squared errors of estimated random effects obtained from full Bayesian inference based on DPM priors (denoted as DPM in the figures) with estimation results based on Gaussian random effects priors, using full Bayesian (ND_MCMC) or empirical Bayesian (ND_REML) inference as implemented in BayesX (Brezger et al., 2005). In each simulation run s, we compare true random effects with corresponding estimates through squared differences. Specifically, we sum up over the $n = 20$ individual parameters and obtain a sum of squares, $SSQ_r(s) = \sum_{i=1}^{n} \left(\hat{b}_{ir}^{(s)} - b_{ir}^{(s)}\right)^2$, $r = 0, 1$, for random intercepts and slopes. The empirical distributions of $SSQ_r(s)$ values obtained from simulation runs $s = 1, \ldots, 100$ are then represented through box plots. In addition, we also compare estimates $\hat{\sigma}^2$ for the error variance as well as estimated variances $\hat{\sigma}^2_{b_0}$, $\hat{\sigma}^2_{b_1}$ obtained from a Gaussian prior assumption with estimated variances $\hat{\sigma}^2_{0_0}$, $\hat{\sigma}^2_{0_1}$ of the Gaussian base distribution and of the Gaussian prior $\boldsymbol{b}_i|\boldsymbol{\theta}_i, \boldsymbol{D} \sim N(\boldsymbol{\theta}_i, \boldsymbol{D})$ for DPM priors. The results for the nonlinear trend function $f(t)$ were mostly insensitive to the

random effects specification and will therefore be omitted. The summary of results will be grouped along a number of scenarios selected from the total of nine possible combinations.

5.3.2. Results

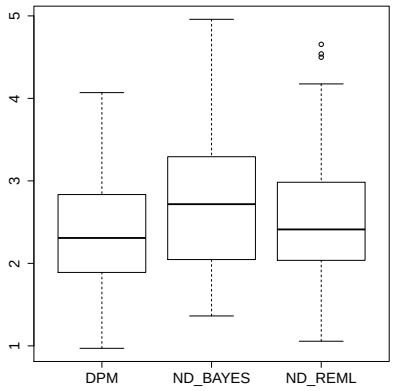
Medium number of individual observations for cases 2 and 3

Figure 5.2 (top left) displays a trace plot of typical longitudinal data generated in the setting of case 2 for a medium number of individual observations, showing that cluster effects can easily be detected visually. As we would expect, additive mixed models with DPM priors substantially improve upon results based on a misspecified Gaussian random effects assumption (Figure 5.3, top) in this case.

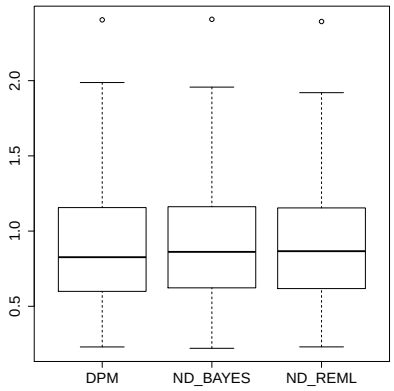


(a) Box plot for $SSQ_0(s)$

(b) Box plot for $SSQ_1(s)$



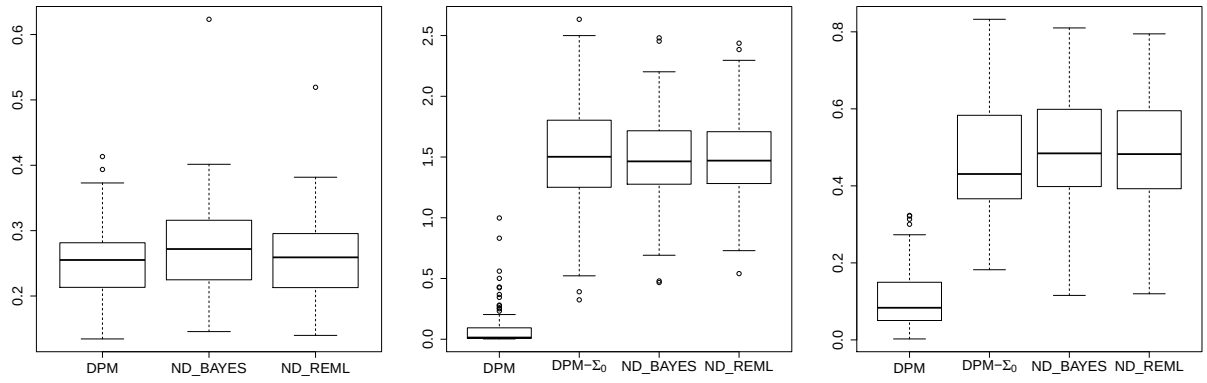
(c) Box plot for $SSQ_0(s)$



(d) Box plot for $SSQ_1(s)$

Figure 5.3.: Box plots for random intercepts (left) and random slopes (right) with moderately separated (top) and substantially overlapping clusters (bottom) for a medium number of individual observations.

When regarding the figure of estimated variances (Figure 5.4), we can conclude the following: First, there are no substantial differences for the estimates of the error variance. Second, estimates for the base variances $\sigma^2_{0_0}$, $\sigma^2_{0_1}$ in the DPM model are similar to the variation of the random effects in the models with Gaussian prior. On the other hand, variation of the random effects $\boldsymbol{b}_i$ in the DPM around their mean $\boldsymbol{\theta}_i$ is always quite small. This result holds generally also in the other settings considered in this simulation study.



(a) $\hat{\sigma}^2$ (b) $\hat{\sigma}^2_{b_0}$ (c) $\hat{\sigma}^2_{b_1}$

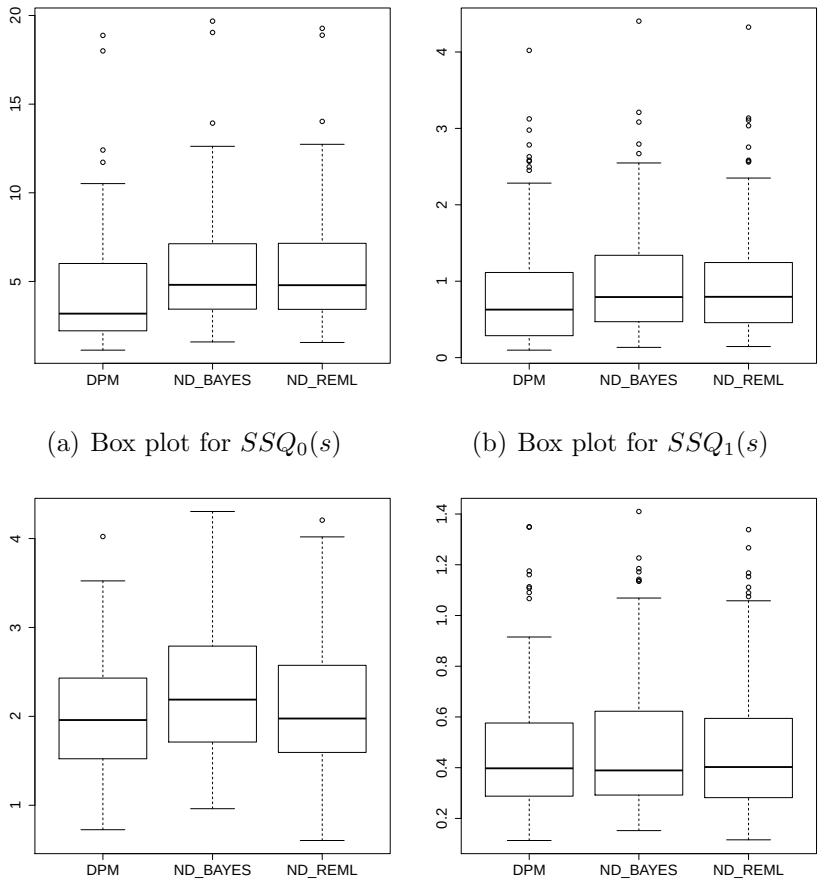
Figure 5.4.: Box plots for the estimated error variance (left), the estimated variance of the random intercept (middle) respectively of the random slope (right) with moderately separated clusters for a medium number of individual observations.

As can be seen from the trace plot (Figure 5.2, bottom left) of typical longitudinal data in case 3, it is not easy to recognize that the random curves come from three different clusters. Hence, it may be much more tempting to apply a mixed model with Gaussian random effects to analyze these data. Still, the SSQ box plots in Figure 5.3 (bottom) demonstrate that DPM priors perform at least comparable or even a bit superior to Gaussian priors.

Many individual observations for cases 2 and 3

For data with comparably many individual observations, improvement of DPM priors relative to Gaussian priors tends to become smaller. As might be expected from visual inspection of the trace plot (Figure 5.2, top right), DPM priors still clearly improve upon Gaussian priors for case 2 (Figure 5.5, top). For case 3 (see Figure 5.2, bottom right, for a typical trace plot) it becomes quite difficult to detect heterogeneity caused by clusters through visual inspection.

The SSQ box plots in Figure 5.5 (bottom) confirm what might be expected: For longitudinal data with many observations and moderate population heterogeneity, DPM priors do not yield substantial improvement upon traditional Gaussian random effects assumptions. Still, the good message is that there is no loss in efficiency in using DPM priors in this situation or even in the theoretically ideal situation that the true random effects are a sample from a homogeneous, approximately Gaussian population. In summary, we



(a) Box plot for $SSQ_0(s)$ (b) Box plot for $SSQ_1(s)$

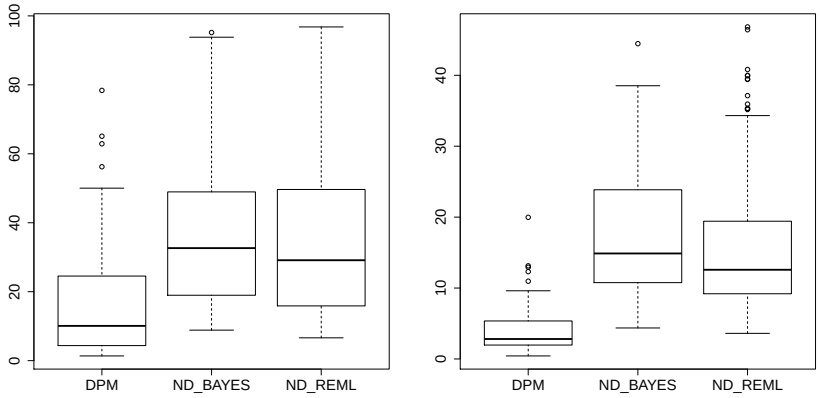
(c) Box plot for $SSQ_0(s)$ (d) Box plot for $SSQ_1(s)$

Figure 5.5.: Box plots for random intercepts (left) and random slopes (right) with moderately separated (top) and substantially overlapping clusters (bottom) for many individual observations.

draw the following conclusion: The improvement in using DPM priors for the estimation of random effects (measured in terms of SSQs) increases either for clusters that are well separated from each other or a small number of observations in the data.

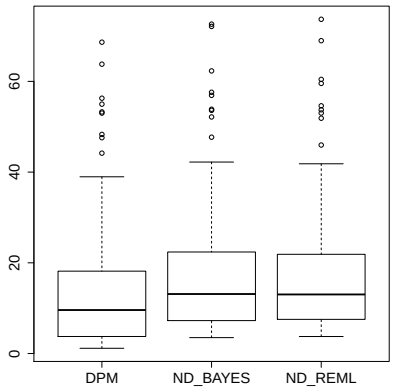
Case 1

In case 1, the DPM model improves considerably upon models with Gaussian priors if the number of individual observations is small. If the number of individual observations is moderate or larger, results become closer to those obtained with moderately separated clusters. The corresponding plots are shown in Figure 5.6.

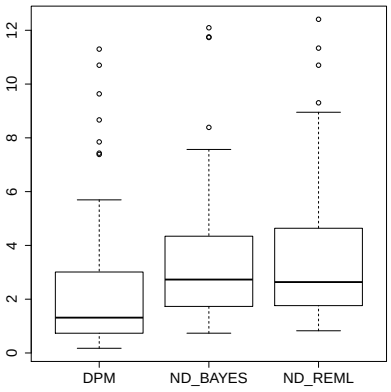


(a) Box plot for $SSQ_0(s)$

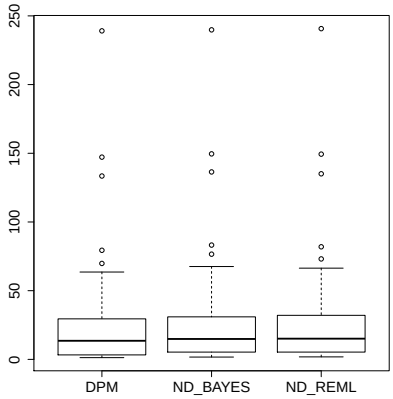
(b) Box plot for $SSQ_1(s)$



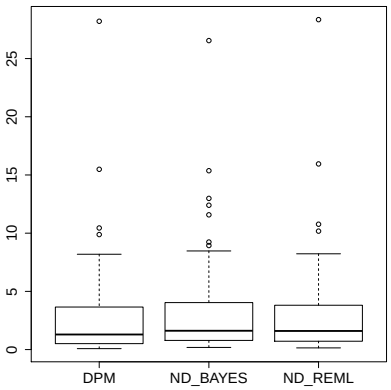
(c) Box plot for $SSQ_0(s)$



(d) Box plot for $SSQ_1(s)$



(e) Box plot for $SSQ_0(s)$



(f) Box plot for $SSQ_1(s)$

Figure 5.6.: Box plots for random intercepts (left) and random slopes (right) with clearly separated clusters for few individual observations (top), a medium number of individual observations (middle) and many individual observations (bottom).

5.4. Application: Childhood Obesity

In the following, we analyze the longitudinal BMI profiles of children collected in the LISA study based on the additive mixed model (5.1). The LISA study is a prospective birth cohort study in four cities in Germany (Bad Honnef, Leipzig, Munich, Wesel), including 3,097 healthy neonates born between 11/1997 and 01/1999. Follow-up time was until the age of six by questionnaires in connection with the nine mandatory examinations at birth and around the age of 2 weeks, 1, 3, 6, 12, 24, 48 and 60 months. Thus, the maximum number of observations per child was nine. Following Fenske et al. (2008), we handle the missing data problems by a complete case analysis. Finally, there are 2,043 children and 17,316 observations. Taking age of children as the basic time scale t, our statistical aim is to assess the influence of a child's age and risk factors on its BMI. Table 5.1 gives an overview of the covariates that are included in the analysis. Note that later on we use centered versions of the continuous covariates `mBMI` and `mDiffBMI` to avoid autocorrelations in the samples that are avoided by making covariates orthogonal to the constant.

Covariate	**Description**	**Categories**	**Relative frequency**	**Absolute frequency**
`sex`	gender	0 = female	47.2%	964
		1 = male	52.8%	1079
`breast`	Nutrition until the age of 4 months	0 = bottle-feeding only or mixture of bottle-feeding and breastfeeding	40.5%	828
		1 = breastfeeding only	59.5%	1215
`mSmoke`	maternal smoking during pregnancy	0 = no	86.0%	1756
		1 = yes	14.0%	287
`area`	region	0 = rural (Bad Honnef, Wesel)	21.5%	439
		1 = urban (Leipzig, Munich)	78.5%	1604

Covariate	**Description**	**Median**	**Mean**	**Sd**
`ageY`	age (in *years*)	0.52	1.39	1.76
`mBMI`	maternal BMI at pregnancy begin (in kg/m^2)	21.72	22.58	3.74
`mDiffBMI`	maternal BMI gain during pregnancy (in kg/m^2)	4.96	5.12	1.63

Table 5.1.: Description of the used categorial and continuous covariates of the LISA data (related to 2043 children).

The effect of age on the BMI is of particular interest in our analyzes. Figure 5.7 (left) shows individual BMI patterns by age for twelve randomly selected children. Here as well as in Figure 5.7 (right) showing the complete data, a nonlinear trend of age is obvious. To fit a smooth age trend, we utilize a cubic Bayesian P-spline with second order random walk penalty and twelve equidistant inner knots. Apart from the general effect of the covariate `ageY`, we are interested in individual deviations from this trend. For this purpose we start with an additive mixed model (5.1) with linear individual deviations $\boldsymbol{z}_{ij}^T \boldsymbol{b}_i = b_{i0} + t_{ij} b_{i1}$.

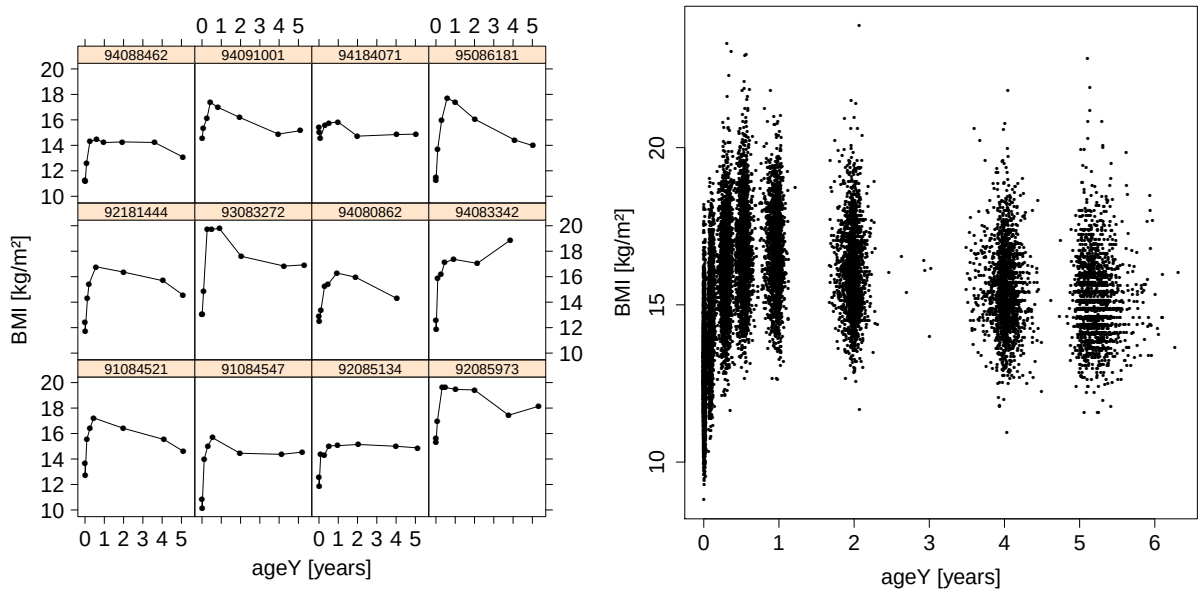


Figure 5.7.: BMI against age: trace plots (left) for twelve randomly selected children and a scatter plot for all children (right) of the LISA data.

Figure 5.8 (left) visualizes the fit of this model. Here one can see the general time trend (solid line) as well as individual fits for four selected subjects. The measurements of these subjects show different peculiar patterns that allow to investigate the ability of the semiparametric additive mixed model to fit individual BMI profiles. One individual (id = 92189214, △) features very low values of BMI while for another individual (id = 95089461, ×) there is a nontypical gain of BMI after the age of one year. We find that the additive mixed model responds to these features sufficiently even with only linear individual-specific deviations from the overall trend. In contrast, the individual with id = 94182011 (+) shows an extremely high value in the age of two years that is not captured well by the model. Similarly, the distinct apex at an age of six months observed for the individual with id = 92185191 (○) is not adequately reflected by the model. Obviously, the deviations for these individuals are too nonlinear to be detectable with linear individual-specific deviations only. Therefore, we extend the model by an additional random effect as in (5.2) with

$$h(t) = \frac{\log(t+1)}{(t+1)^2}, \tag{5.4}$$

that was chosen to reflect the special pattern of the temporal trends in the LISA data. Indeed, the individual fits improve considerably as shown in Figure 5.8 (right).

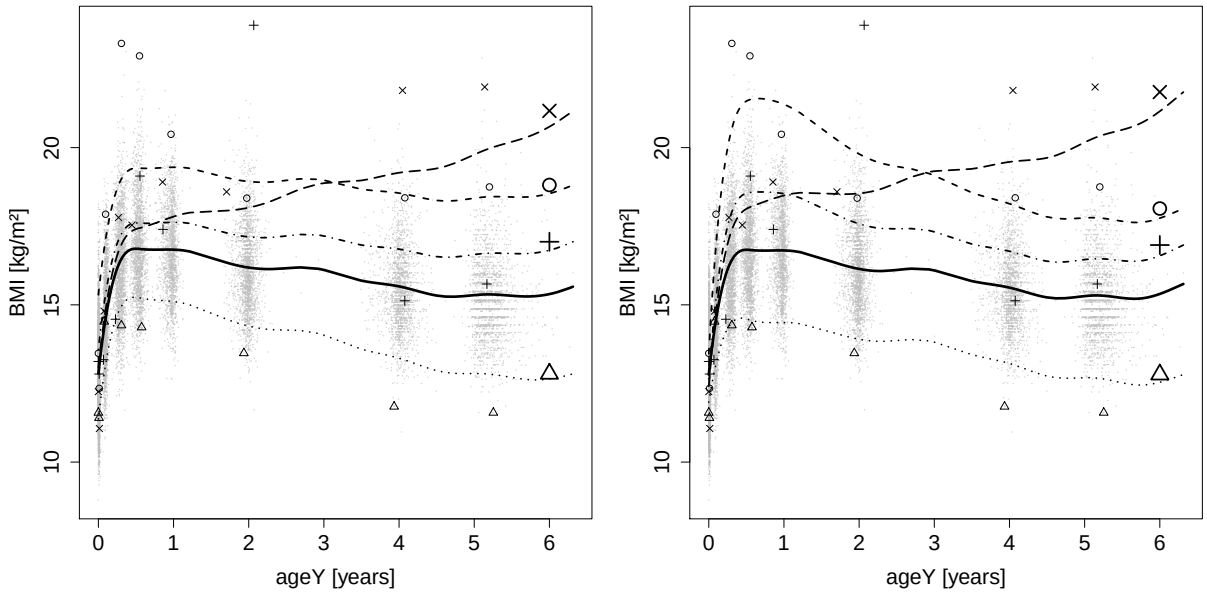


Figure 5.8.: Fit of the DPM model (5.1) for the LISA data with linear individual-specific deviations $\boldsymbol{z}_{ij}^T\boldsymbol{b}_i = b_{i0} + t_{ij}b_{i1}$ (left) and with nonlinear individual-specific deviations $\boldsymbol{z}_{ij}^T\boldsymbol{b}_i = b_{i0} + t_{ij}b_{i1} + h(t_{ij})b_{i2}$ (right). The solid line shows the general effect of age while the dashed lines show individual effects. Observations belonging to the same subject are marked with the same symbol. These symbols are also added to the corresponding individual curves: $\triangle$ (id = 92189214), $\times$ (id = 95089461), $+$ (id = 94182011) and $\bigcirc$ (id = 92185191).

Table 5.2 contains estimation results for fixed effects in the extended model. According to the symmetric 95% credibility intervals, there are three significant effects. The BMI of boys is about 0.2 points larger than that of girls if all other covariates are kept fixed. The maternal BMI and the maternal BMI gain during pregnancy also have a positive impact on the child's BMI while the covariates `breast`, `mSmoke` and `area` show no significant effect on the BMI. This fact is surprising since the impact of breastfeeding is discussed somewhat controversially in the literature, see Beyerlein et al. (2008) and the references therein. Our finding is in agreement with the results in Beyerlein et al. (2008), who used conventional linear regression and quantile regression with the BMI as continuous response and concluded that breastfeeding has no significant effect on the expectation or the median of the BMI distribution. However, breastfeeding can reduce the BMI of children having BMI values in the upper quantile range. On the other hand, previous studies (e.g. Harder et al., 2005; Arenz et al., 2004) observed a protective effect of breastfeeding, using logistic regression with obesity as binary response, where children are defined as obese if their BMI exceeds some predefined threshold. While at first sight our results seem to contradict the protective effect of breastfeeding, more detailed analyzes reveal that the relationship between BMI and breastfeeding depends upon the cluster to which an individual belongs. We now discuss how we assign individuals to clusters.

	Mean	Median	Se	2.5%-quantile	97.5%-quantile
σ^2	0.70258	0.70252	0.00925	0.68464	0.72167
`sex`	0.20103	0.20077	0.03800	0.12439	0.27666
`breast`	0.05282	0.05124	0.03748	-0.02315	0.12654
`mSmoke`	-0.00729	-0.00631	0.05288	-0.10690	0.09506
`area`	-0.05123	-0.05162	0.04809	-0.14603	0.04109
`mBMI` (cent.)	0.04642	0.04650	0.00517	0.03625	0.05662
`mDiffBMI` (cent.)	0.08035	0.08003	0.01197	0.05743	0.10367

Table 5.2.: Estimates for the error variance and the fixed effects of the DPM model for the LISA data.

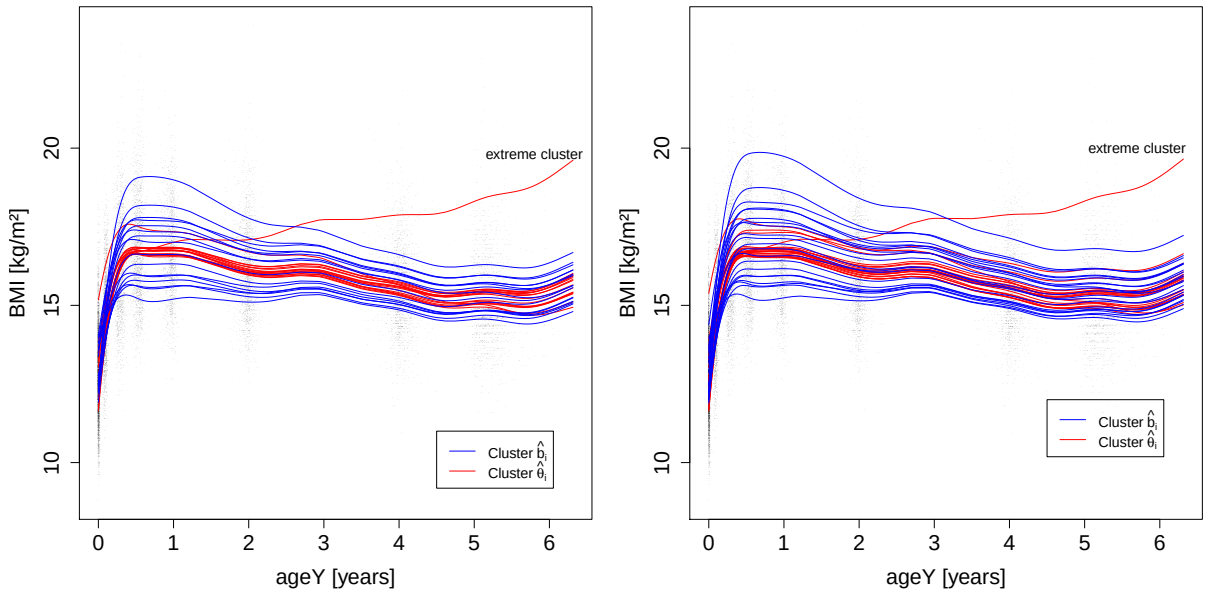
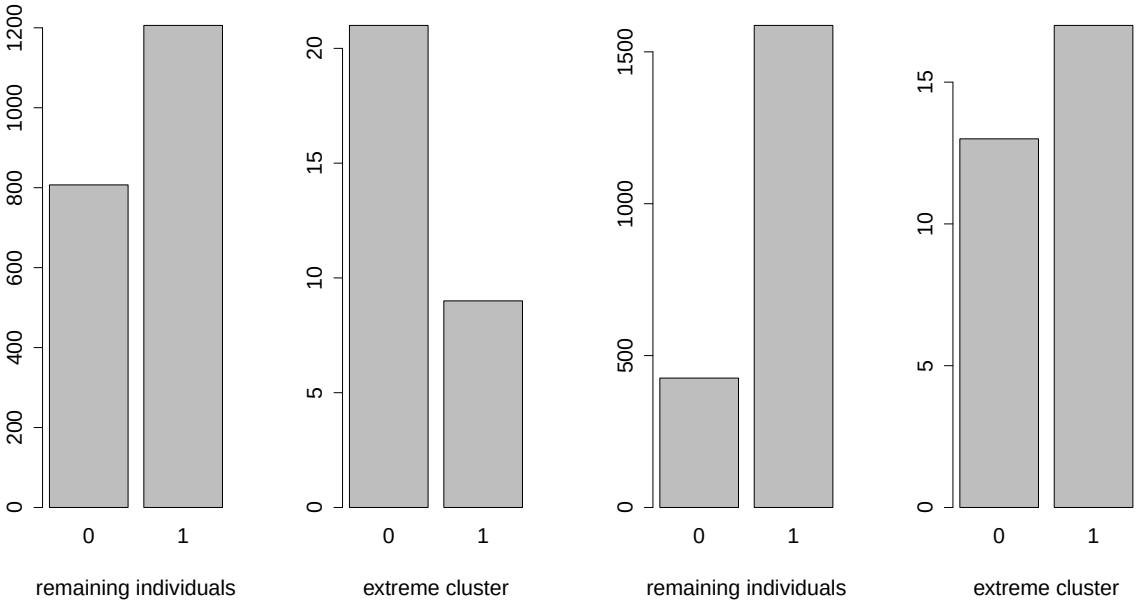


Figure 5.9.: Clustering of $\hat{\boldsymbol{\theta}}_i$ and $\hat{\boldsymbol{b}}_i$ of the DPM model for the LISA data with nonlinear individual-specific deviations $\boldsymbol{z}_{ij}^T\boldsymbol{b}_i = b_{i0} + t_{ij}b_{i1} + h(t_{ij})b_{i2}$. On the left for the prior hyper parameters of the random effects' variances the default settings $a_0 = b_0 = 0.5$ and $a_b = b_b = 0.0001$ are used while on the right these parameters are given by $a_0 = b_0 = a_b = b_b = 0.005$.

Although there is an automatic clustering structure induced by the Dirichlet process in theory, some practical problems arise from the necessity of using MCMC methods: We get a clustering of subjects at each iteration, but how can these be combined to a universal clustering? Diverse operations exist to handle this (see for example Fritsch and Ickstadt, 2009), but concerning the high number of subjects and hence the high number of possible clusterings, these methods are typically not feasible for our model. Therefore we pursue an alternative strategy: First, we get an estimated number of clusters using the median for all numbers of clusters in the MCMC iterations. Second, we allocate the $\hat{\boldsymbol{\theta}}_i$ for $i = 1, \ldots, n$ to clusters by k-means clustering with the number of clusters fixed to this median number of

clusters. In addition, we perform a similar clustering for $\hat{\boldsymbol{b}}_i$ with $i = 1, \ldots, n$. Still, we are primarily interested in the clustering of $\hat{\boldsymbol{\theta}}_i$, since these are the parameters on which the clustering property of the Dirichlet process works in the prior specification.

The cluster results shown in Figure 5.9 (left) reveal that there are mainly level shifts between clusters both for $\hat{\boldsymbol{\theta}}_i$ and for $\hat{\boldsymbol{b}}_i$. However, for $\hat{\boldsymbol{\theta}}_i$ one cluster that differs from all other clusters attracts special attention: While the BMI steeply increased until an age of about six months followed by a steady decrease for most of the children, individuals in this cluster show an almost permanent ascent of the BMI. Note that this extreme cluster can only be detected for $\hat{\boldsymbol{\theta}}_i$ and not for $\hat{\boldsymbol{b}}_i$. The extreme cluster consists of thirty individuals, including one of the individuals that we have picked for the individual profiles shown in Figure 5.8 (id = 95089461, $\times$).



(a) breastfeeding (b) area of residence

Figure 5.10.: Bar plots of the covariates `breast` (left) and `area` (right), each for the subjects of the extreme cluster (on the right hand) and for the others (on the left hand) corresponding to the clustering by the DPM model.

It is now of particular interest to detect whether there are differences in the covariate values between individuals assigned to the extreme clusters and the remaining individuals. Indeed, especially the covariates relating to breastfeeding and area of residence show noticeable differences between these two groups (see Figure 5.10). Obviously, most of the children in the extreme cluster were bottlefed or bottle- and breastfed until the age of four months. In contrast, the majority of the remaining children were breastfed only. As a consequence, breastfeeding essentially serves as an indicator for a normal and a lower development of the BMI, although there is no general protective effect of breastfeeding. This points into the same direction as the quantile regression result of Beyerlein et al. (2008).

Similarly, the ratio of children living in an urban area as compared to children living in a rural area is almost balanced in the extreme cluster while most of the remaining children live in urban areas.

	$Ga(2, 0.5)$	$Ga(2, 1)$	$Ga(2, 2)$	$Ga(2, 4)$
number of clusters	24	20	16	12
σ^2	0.70207	0.70228	0.70252	0.70294
`sex`	0.20111	0.20180	0.20077	0.20120
`breast`	0.05783	0.05194	0.05124	0.05650
`mSmoke`	-0.00254	0.00006	-0.00631	0.00058
`area`	-0.04942	-0.04912	-0.05162	-0.05202
`mBMI (cent.)`	0.04654	0.04685	0.04650	0.04653
`mDiffBMI (cent.)`	0.07955	0.07900	0.08003	0.08066

Table 5.3.: The effect of the prior for the concentration parameter α on the number of clusters and the estimates for σ^2 and $\boldsymbol{\beta}$. Note that we assume a discrete prior with probabilities which resemble a gamma distribution.

To investigate prior sensitivity of our results, we re-ran the analyzes with different choices for the hyperparameters of the inverse gamma priors for the variance parameters. Essentially, the basic conclusions remain unchanged and the identification of the extreme cluster is very robust with respect to prior assumptions but there is some variation in the number of clusters. For example, when choosing $a_b = b_b = 0.005$ and $a_0 = b_0 = 0.005$ it is still possible to identify the cluster of extreme observations, but the total number of clusters increases from 16 to 21 compared to our original analysis, as it can be seen in Figure 5.9 (right). Table 5.3 provides some additional information on the impact of the prior for the concentration parameter α on the number of clusters, the parametric effects and the error variance.

5.5. Summary and Discussion

The semiparametric mixed model considered in this chapter combines the advantages of Bayesian smoothing of nonlinear time trends and other nonlinear covariate effects with the flexibility of DPM priors in order to deal with heterogeneity of random effects. Our simulation study provides evidence, under which circumstances DPM random effects priors really lead to substantial improvement compared to conventional Gaussian random effects priors. DPM priors can also be used as an exploratory tool to check sensitivity of parametric assumptions on random effects. In particular, as illustrated in our BMI application, Dirichlet processes and DPM priors allow to detect hidden clusters in the data.

In summary, using a DPM prior as random effects distribution implies a (Gaussian) mixture prior for random effects with a data driven choice of the number of mixture components. Hence, using DPM priors is not only a more flexible modeling opportunity

without restrictive assumptions for the random effects distribution, but also provides additional insights into the hidden pattern of clusters in given data, which are not possible in the case of a Gaussian random effects model. In general, this knowledge can be used to detect indicators for this pattern as we demonstrated for the clusters with deviating BMI profiles in the obesity applications. Note that detecting clusters is not possible in the classical Gaussian random effects model, where always a single cluster for all individuals is assumed.

The development of Gibbs samplers for Gaussian additive mixed models with more complicated structured additive predictors, comprising for example spatial effects or varying coefficient terms as in Fahrmeir et al. (2004), is conceptually straightforward due to the modular hierarchical structure of the MCMC sampler. Extensions to models with non-Gaussian responses require additional computational effort, involving hybrid Metropolis Hastings algorithms unless a latent Gaussian formulation can be obtained (as for binary or multinomial probit models).

ACKNOWLEDGEMENTS: We thank the German National Science Foundation (DFG) for financial support from the project "Bayesian Regularisation in Regression Models with High-Dimensional Predictors" (Grants FA 128/5-1, FA 128/5-2) and Prof. Dr. Erich Wichmann (Helmholtz Zentrum, LMU Munich and Munich Center of Health) for providing the data of the LISA study. Special thanks also to Daniel Sabanés Bové for his idea to use the nonlinear function (5.4) for modeling BMI profiles.

6. Additive Mixed Models with DPMs using EM Algorithm

6.1. Introduction

For the modeling of longitudinal data with a nonlinear time trend, additive mixed models are useful. The model considered in this chapter assumes an additive structure for the nonparametric term of the time variable and parametric terms for the random effects as well as the fixed effects for other covariates. Due to this combination of nonparametric and parametric terms the model is called *semiparametric mixed model*. Here, the conditional distribution of the response y_{ij} observed for subject i at observation time t_{ij} can be written as

$$y_{ij}|\boldsymbol{b}_i \overset{ind.}{\sim} N(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + f(t_{ij}) + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i,\, \sigma^2), \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i. \tag{6.1}$$

Fixed effects $\boldsymbol{\beta}$ describe the influence of covariates $\boldsymbol{x}_{ij}$ whereas individual-specific deviations from the population time trend $f(\cdot)$ are modeled in the product of random effects $\boldsymbol{b}_i$ and time-dependent variables $\boldsymbol{z}_{ij}$. For example, in a so-called *random slope model* one specifies $\boldsymbol{z}_{ij}^T\boldsymbol{b}_i = b_{i0} + t_{ij} \cdot b_{i1}$, which means that the variation over time is given by $f(\cdot)$ but with individual shift and slope. The approach proposed in this chapter combines an approximate DPM for the random effects with a P-spline for approximating the trend function $f(\cdot)$ and uses an EM algorithm for estimation. In model (6.1) typically normally distributed random effects are assumed (see, for example, Zeger and Diggle (1994), Zhang et al. (1998), Verbyla et al. (1999), Ruppert et al. (2003), Fan and Li (2004), and Wang et al. (2005)). In contrast to these approaches we consider in analogue to Chapter 5 a DPM as random effects distribution because the cluster property of the Dirichlet process allows to find clusters in longitudinal data (Ferguson, 1973). More concretely, we make use of the stick breaking representation of the Dirichlet process (Sethuraman, 1994). See Chapter 2 for more information about the theory behind Dirichlet processes. The most innovative aspect of our method is that we introduce an EM algorithm for inference instead of the popular MCMC methods, which are used, for example, in Chapter 5 and Li et al. (2010). The advantage of the EM algorithm over MCMC methods is, as far as Dirichlet processes are concerned, that it provides a pointwise convergence instead of a distributional convergence. One consequence is that the cluster property of the Dirichlet process can be used directly. More details about this property are given in Chapter 4, where *linear* mixed models with approximate DPMs for incorporating a linear time trend are estimated by the

EM algorithm. This algorithm will be extended to additive mixed models in the present chapter for clustering nonlinear longitudinal data.

The chapter is organized as follows: In Section 6.2, the model hierarchy and the according EM algorithm for fitting the proposed model is presented in detail. In addition, a short discussion of reparameterizations of P-spline coefficients and the choice of knots is given. The simulation study in Section 6.3 compares our approach to the MCMC-method in Chapter 5 and to additive mixed models with normally distributed random effects. In Section 6.4, the theophylline data and BMI profiles of children are analyzed.

6.2. Additive Mixed Models with Dirichlet Process Mixtures

6.2.1. Model Hierarchy

Let the time trend in model (6.1) be specified by B-splines (De Boor, 1978) yielding $f(t_{ij}) = \sum_{s=1}^{d} \gamma_s B_s^l(t_{ij})$, where γ_s denotes the basis coefficient corresponding to the B-spline basis function B_s^l of degree l. For m inner knots $\kappa_1, \ldots, \kappa_m$ one obtains all in all $m + 2 \cdot l$ knots and $d = m + l - 1$ basis coefficients which are collected in the vector $\boldsymbol{\gamma}$. Generally, in order to get a smooth trend curve, the curvature is penalized by considering the penalty term $\lambda \cdot \int (f''(t))^2 dt$ as is customary also for smoothing splines (Reinsch, 1967), where λ denotes a tuning parameter. Using B-splines this penalty term may be written as $\boldsymbol{\gamma}^T \boldsymbol{K} \boldsymbol{\gamma}$, where $\boldsymbol{K}$ denotes a singular penalty matrix with rank $d - k$ and whose element in the rth row and the sth column is given by $\int B_r''(t) B_s''(t) dt$ (O'Sullivan, 1986). The integer k describes the rank deficiency of the penalty matrix. Eilers and Marx (1996) introduced the so-called P-splines that penalize the differences between the basis coefficients by considering the penalty matrix $\boldsymbol{K} = \boldsymbol{\Delta}^T \boldsymbol{\Delta}$ based on the difference matrix $\boldsymbol{\Delta}$ of order k. In the following, these P-splines are considered for estimating the trend curve. In addition, we make use of the mixed model representation of the P-spline term to avoid time-consuming methods like cross-validation when determining the tuning parameter: Let the basis coefficient vector be decomposed in the form of $\boldsymbol{\gamma} = \boldsymbol{T}\boldsymbol{\gamma}_0 + \boldsymbol{W}\boldsymbol{\gamma}_p$ into an unpenalized vector $\boldsymbol{\gamma}_0$ and a penalized vector $\boldsymbol{\gamma}_p$ for suitable matrices $\boldsymbol{T}$ and $\boldsymbol{W}$ (Green, 1987). See Section 6.2.3 for more details about this decomposition and other facts concerning the P-spline term. In Fahrmeir et al. (2007) it is clarified that $\boldsymbol{\gamma}_p$ can be interpreted as a normally distributed random effect in a classical mixed model. Thus, the conditional distribution (6.1) can be rewritten in matrix notation as

$$\begin{aligned} \boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{\gamma}_p &\overset{ind.}{\sim} N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 + \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p + \boldsymbol{Z}_i\boldsymbol{b}_i,\, \sigma^2\boldsymbol{I}_{n_i}), \quad i = 1, \ldots, n, \\ \boldsymbol{\gamma}_p &\sim N(\boldsymbol{0}, \tau^2\boldsymbol{I}_{d-k}), \end{aligned}$$

where $\boldsymbol{I}_{d-k}$ symbolizes the identity matrix with dimension $d - k$. $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ denote the individual design matrices constructed from covariates $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ whereas the matrix $\boldsymbol{B}_i$

contains the B-spline basis functions of subject i. In $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ the response values of subject i are collected. The variance parameter τ^2 acts as an inverse smoothing parameter and will be estimated in the inference procedure. While large values of τ^2 yield a rough spline, for $\tau^2 \to 0$ the coefficients in $\boldsymbol{\gamma}_p$ are shrunk to zero and thus, the spline converges to a polynomial of degree $k-1$.

In our approach, instead of a normal distribution as random effects distribution a DPM is considered:

$$\begin{array}{rcll} \boldsymbol{b}_i|\boldsymbol{\theta}_i & \overset{ind.}{\sim} & N(\boldsymbol{\theta}_i, \boldsymbol{D}), & i = 1, \ldots, n, \\ \boldsymbol{\theta}_i|G & \overset{i.i.d.}{\sim} & G, & i = 1, \ldots, n, \\ G & \sim & DP(\alpha, G_0). & \end{array} \tag{6.2}$$

Here, $DP(\alpha, G_0)$ is a distributional assumption for the unknown mixing distribution G. Given G, the means of the normal distribution are drawn from the distribution G, which is a discrete distribution and that has $-$ in the case of a low α $-$ a set of just a few elements with probabilities that are considerably larger than zero. Thus, the marginal random effects distribution is a normal mixture with a data driven and typically low number of mixture components. Thereby, a natural clustering of individuals can be achieved: Subjects with the same mean $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$, $i \neq j$, belong to the same cluster. By using the stick breaking procedure of the Dirichlet process in its truncated version, inference for the unknown distribution G becomes possible and the distributional assumption for the random effects (6.2) can be rewritten as

$$\begin{array}{rcll} \boldsymbol{b}_i|\boldsymbol{v} & \overset{i.i.d.}{\sim} & \sum_{h=1}^{N} \pi_h \, N(\boldsymbol{\mu}_h, \boldsymbol{D}), & i = 1, \ldots, n, \\ \pi_h & = & v_h \prod_{l<h}(1 - v_l), & h = 1, \ldots, N, \\ v_h & \overset{i.i.d.}{\sim} & Be(1, \alpha), & h = 1, \ldots, N-1, \end{array} \tag{6.3}$$

where $Be(\cdot, \cdot)$ denotes the beta distribution and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^T$ respectively $\boldsymbol{v} = (v_1, \ldots, v_{N-1})^T$ are vectors of weights respectively reparameterized weights. See Section 2.2 for an extended discussion of the stick breaking representation of the Dirichlet process and Section 6.2.2 for a recommendation how to choose N. As customary, in this context two constraints have to hold: $\sum_{h=1}^{N} \pi_h \boldsymbol{\mu}_h = \boldsymbol{0}$ and $\sum_{h=1}^{N} \pi_h = 1$. The first ensures $\mathbb{E}(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{\gamma}$ and therefore the identifiability of the P-spline. Note that the order of $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ is given by the corresponding weights in decreasing order. The second constraint $\sum_{h=1}^{N} \pi_h = 1$ is automatically fulfilled by $v_N = 1$.

6.2.2. Inference

In what follows, an EM algorithm for the additive mixed model described in Section 6.2.1 is given. The algorithm is based on the estimation procedure of the heterogeneity model by Verbeke and Lesaffre (1996). In general, the EM algorithm is an useful inference tool in the case of unobserved data (McLachlan and Krishnan, 1997). In finite mixture models, the unknown cluster membership of each individual can be expressed by the latent variable

$\boldsymbol{w}_i := (w_{i1}, \ldots, w_{iN})^T$, where $w_{ih} = 1$ if subject i belongs to cluster h and 0 otherwise (McLachlan and Peel, 2000). For our approach, the marginalization over the random effects yields the following complete model with observed data $\boldsymbol{y}_i$ and unobserved data $\boldsymbol{w}_i$

$$\begin{array}{rcll} \boldsymbol{y}_i|\boldsymbol{w}_i, \boldsymbol{\gamma}_p & \overset{ind.}{\sim} & N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 + \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p + \boldsymbol{Z}_i\boldsymbol{\mu}_h,\, \boldsymbol{V}_i), & i = 1, \ldots, n, \\ \boldsymbol{w}_i|\boldsymbol{v} & \overset{i.i.d.}{\sim} & M(1, \boldsymbol{\pi}), & i = 1, \ldots, n, \\ v_h & \overset{i.i.d.}{\sim} & Be(1, \alpha), & h = 1, \ldots, N-1, \\ \boldsymbol{\gamma}_p & \sim & N(\boldsymbol{0}, \tau^2\boldsymbol{I}_{d-k}), & \end{array} \tag{6.4}$$

with $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \sigma^2\boldsymbol{I}_{n_i}$ and $M(\cdot, \cdot)$ symbolizing the multinomial distribution. The first two lines in model (6.4) determine the likelihood function of the independent observations $(\boldsymbol{y}_i, \boldsymbol{w}_i)$, $i = 1, \ldots, n$. The third and the fourth line correspond to prior distributions that can also be seen as penalty terms. As customary in the likelihood inference, for the other parameters diffuse priors are assumed. All parameters are collected in the vector $\boldsymbol{\xi} = (\alpha, \boldsymbol{v}, \boldsymbol{\psi})^T$, where $\boldsymbol{\psi}$ is the vector containing all the remaining parameters $\boldsymbol{\beta}, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}_p, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N, \boldsymbol{D}, \sigma^2$ and τ^2. Note that model (6.4) can either be parameterized by $\boldsymbol{\pi}$ or by $\boldsymbol{v}$. Since the latter parametrization simplifies calculations, it is used in the following. Nevertheless, only for a simpler presentation, we write π_h instead of $v_h \prod_{l<h}(1 - v_l)$. Omitting multiplicative constants, the posterior function respectively the penalized likelihood function corresponding to the complete model (6.4) is given by

$$L_P(\boldsymbol{\xi}) = \prod_{i=1}^{n}\prod_{h=1}^{N}[\pi_h\, f_{ih}(\boldsymbol{y}_i; \boldsymbol{\psi})]^{w_{ih}} \cdot (\tau^2)^{-\frac{d-k}{2}} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}_p^T\boldsymbol{\gamma}_p\right) \cdot \alpha^{N-1} \prod_{h=1}^{N-1}(1 - v_h)^{\alpha-1}.$$

Here, $f_{ih}(\cdot)$ denotes the density function of $N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 + \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p + \boldsymbol{Z}_i\boldsymbol{\mu}_h,\, \boldsymbol{V}_i)$. Finally, one obtains the penalized log-likelihood

$$\begin{aligned} l_P(\boldsymbol{\xi}) = \sum_{i=1}^{n}\sum_{h=1}^{N} w_{ih}[\log \pi_h + \log f_{ih}(\boldsymbol{y}_i; \boldsymbol{\psi})] - \frac{1}{2}\left((d-k)\log(\tau^2) + \frac{1}{\tau^2}\boldsymbol{\gamma}_p^T\boldsymbol{\gamma}_p\right) + \\ + (N-1)\log\alpha + (\alpha - 1)\sum_{h=1}^{N-1}\log(1 - v_h). \end{aligned}$$

Also the parameter α can be seen as penalization parameter as τ^2. For $\alpha \in (0, 1)$ a penalization of the number of clusters is achieved whereas for $\alpha = 1$ the penalty term in $l_P(\boldsymbol{\xi})$ drops pout. For $\alpha \to 0$ the number of clusters converges to one. See Section 4.2.2 for more information about the meaning of α in this context. Instead of maximizing the penalized incomplete likelihood function

$$l_{PI}(\boldsymbol{\xi}) = \sum_{i=1}^{n} \log \left(\sum_{h=1}^{N} \pi_h f_{ih}(\boldsymbol{y}_i; \boldsymbol{\psi}) \right) - \frac{1}{2} \left((d-k) \log(\tau^2) + \frac{1}{\tau^2} \boldsymbol{\gamma}_p^T \boldsymbol{\gamma}_p \right) + \\ + (N-1) \log \alpha + (\alpha - 1) \sum_{h=1}^{N-1} \log(1 - v_h).$$

based only on the observed data directly, an EM algorithm is used for estimation of parameters. Here, we alternate between E-step and M-step until $l_{PI}(\boldsymbol{\xi})$ does not change any more.

E-step

In the E-step, we take the expectation of the penalized likelihood $l_P(\boldsymbol{\xi})$ based on the complete model over all unobserved w_{ih}. Collecting all observed data in $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$, we get for the E-step of iteration $t+1$

$$Q(\boldsymbol{\xi}) = \mathbb{E}\left(l_P(\boldsymbol{\xi}) | \boldsymbol{y}, \boldsymbol{\xi}^{(t)} \right) = \sum_{i=1}^{n} \sum_{h=1}^{N} \pi_{ih}(\boldsymbol{\xi}^{(t)}) [\log \pi_h + \log f_{ih}(\boldsymbol{y}_i; \boldsymbol{\psi})] - \\ - \frac{1}{2} \left((d-k) \log(\tau^2) + \frac{1}{\tau^2} \boldsymbol{\gamma}_p^T \boldsymbol{\gamma}_p \right) + (N-1) \log \alpha + (\alpha - 1) \sum_{h=1}^{N-1} \log(1 - v_h),$$

where $\pi_{ih}(\boldsymbol{\xi}^{(t)})$ is the probability at iteration t that subject i belongs to cluster h and is given by

$$\pi_{ih}(\boldsymbol{\xi}^{(t)}) = \frac{f_{ih}(\boldsymbol{y}_i; \boldsymbol{\psi}^{(t)}) \pi_h^{(t)}}{\sum_{l=1}^{N} f_{il}(\boldsymbol{y}_i; \boldsymbol{\psi}^{(t)}) \pi_l^{(t)}}.$$

For clarity, in the following we write $\pi_{ih} := \pi_{ih}(\boldsymbol{\xi}^{(t)})$.

M-step

In the M-step, $Q(\boldsymbol{\xi})$ is maximized with respect to all unknown parameters. Due to $Q(\boldsymbol{\xi}) = Q(\alpha, \boldsymbol{v}) + Q(\boldsymbol{\psi})$ the M-step can be separated into two parts: The maximization of

$$Q(\alpha, \boldsymbol{v}) = \sum_{i=1}^{n} \sum_{h=1}^{N} \pi_{ih} \log \pi_h + (N-1) \log \alpha + (\alpha - 1) \sum_{h=1}^{N-1} \log(1 - v_h),$$

with respect to α and $\boldsymbol{v}$ and the maximization of

$$Q(\boldsymbol{\psi}) = \sum_{i=1}^{n}\sum_{h=1}^{N} \pi_{ih} \log f_{ih}(\boldsymbol{y}_i;\boldsymbol{\psi}) - \frac{1}{2}\left((d-k)\log(\tau^2) + \frac{1}{\tau^2}\boldsymbol{\gamma}_p^T\boldsymbol{\gamma}_p\right),$$

with respect to $\boldsymbol{\psi}$. The first optimization problem is solved by alternating updates of the first order conditions

$$v_h = \frac{\sum_{i=1}^{n}\pi_{ih}}{\sum_{i=1}^{n}\sum_{l=h}^{N}\pi_{il} + \alpha - 1}, \qquad h = 1,\ldots,N-1, \tag{6.5}$$

and

$$\alpha = \frac{1-N}{\sum_{h=1}^{N-1}\log(1-v_h)},$$

that are proved in Appendix A.3.1. Without further restrictions it could happen that $v_h \notin [0,1]$ if $\alpha \in (0,1)$. For preventing this we use the same correction approach as in Section 4.2.2: Update v_h by (6.5) for increasing h. If $v_{h^*} > 1$ set v_h to 1 for $h = h^*,\ldots,N-1$. This constraint for $\boldsymbol{v}$ is equivalent to the following restriction on $\boldsymbol{\pi}$ by using the stick breaking procedure:

$$\pi_h = \begin{cases} \frac{1}{n+\alpha-1}\sum_{i=1}^{n}\pi_{ih}, & \text{for } h < h^*, \\ 1 - \sum_{l=1}^{h-1}\pi_l & \text{for } h = h^*, \\ 0 & \text{for } h > h^*, \end{cases}$$

where h^* is the lowest index h for which the cumulative sum of the original weights π_l° exceeds one: $\sum_{l=1}^{h}\pi_l^\circ > 1$. See Appendix A.3.1 for more technical details about this correction step. Finally, it can be seen that for $\alpha \in (0,1)$, all weights π_h for $h < h^*$ are stretched by the factor $\frac{n}{n+\alpha-1}$ compared to the unpenalized estimators for π_h as in Verbeke and Molenberghs (2000), which we get for $\alpha = 1$. The amount of stretching is controlled by the parameter α. If $\alpha \approx 0$, a very strong clustering is achieved while for larger values of α only few clusters drop out. It should be noted that during the computations $v_h = 1-10^{-300}$ instead of $v_h = 1$ is used to avoid $\log(0)$. Then one gets $\pi_h \approx 0$ for $h > h^*$. For $\alpha > 1$ no correction is needed, but especially in this case it is important that N is large enough due to the considerations in Section 2.2. As proposed by Ohlssen et al. (2007) and as shown in the equations (2.6) N should be chosen such that

$$N > 1 + \frac{\log(\varepsilon)}{\log\left(\frac{\alpha}{\alpha+1}\right)},$$

with $\varepsilon > 0$. Thus, for a given range on α a lower bound for N can be determined. Since in practice a very strong clustering with a low number of clusters is generally desirable, we propose to allow only the range $\alpha \in (0,1)$. In our experience, this can be achieved by a very low starting value like $\alpha = 0$. This means that for $\varepsilon = 0.001$ even $N = 11$ is sufficiently large for an adequate approximation of the distribution G.

In the second part of the M-step, we get the current state for $\boldsymbol{\psi}$ by alternating separate maximization of $Q(\boldsymbol{\psi})$ to $\boldsymbol{\beta}$, $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_p$, $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ and to the variance parameters τ^2, σ^2 and $\boldsymbol{D}$. Conditional on the actual state of the other parameters, the maximization of $\boldsymbol{\beta}$ results in

$$\boldsymbol{\beta} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \left(\boldsymbol{y}_i - \boldsymbol{B}_i \boldsymbol{T} \boldsymbol{\gamma}_0 - \boldsymbol{B}_i \boldsymbol{W} \boldsymbol{\gamma}_p - \sum_{h=1}^{N} \pi_{ih} \boldsymbol{Z}_i \boldsymbol{\mu}_h \right) \right).$$

The first order condition for $\boldsymbol{\gamma}_0$, given all the other parameters, yields

$$\begin{aligned} \boldsymbol{\gamma}_0 &= \left(\sum_{i=1}^{n} \boldsymbol{T}^T \boldsymbol{B}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{B}_i \boldsymbol{T} \right)^{-1} \\ &\quad \left(\sum_{i=1}^{n} \boldsymbol{T}^T \boldsymbol{B}_i^T \boldsymbol{V}_i^{-1} \left(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta} - \boldsymbol{B}_i \boldsymbol{W} \boldsymbol{\gamma}_p - \sum_{h=1}^{N} \pi_{ih} \boldsymbol{Z}_i \boldsymbol{\mu}_h \right) \right), \end{aligned}$$

whereas the penalized basis coefficients are updated by

$$\begin{aligned} \boldsymbol{\gamma}_p &= \left(\sum_{i=1}^{n} \boldsymbol{W}^T \boldsymbol{B}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{B}_i \boldsymbol{W} + \frac{1}{\tau^2} \boldsymbol{I}_{d-k} \right)^{-1} \\ &\quad \left(\sum_{i=1}^{n} \boldsymbol{W}^T \boldsymbol{B}_i^T \boldsymbol{V}_i^{-1} \left(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta} - \boldsymbol{B}_i \boldsymbol{T} \boldsymbol{\gamma}_0 - \sum_{h=1}^{N} \pi_{ih} \boldsymbol{Z}_i \boldsymbol{\mu}_h \right) \right). \end{aligned}$$

Given the other parameters, setting the derivative of $Q(\boldsymbol{\psi})$ with respect to $\boldsymbol{\mu}_h$, $h = 1, \ldots, N$, to zero yields

$$\boldsymbol{\mu}_h = \left(\sum_{i=1}^{n} \pi_{ih} \boldsymbol{Z}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{Z}_i \right)^{-1} \left(\sum_{i=1}^{n} \pi_{ih} \boldsymbol{Z}_i^T \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta} - \boldsymbol{B}_i \boldsymbol{T} \boldsymbol{\gamma}_0 - \boldsymbol{B}_i \boldsymbol{W} \boldsymbol{\gamma}_p) \right).$$

For the inverse smoothing parameter τ^2 one gets the update

$$\tau^2 = \frac{1}{d-k} \boldsymbol{\gamma}_p^T \boldsymbol{\gamma}_p.$$

The corresponding proofs are shown in Appendix A.3.2. For holding the constraint $\sum_{h=1}^{N} \pi_h \boldsymbol{\mu}_h = \boldsymbol{0}$, in each M-step deviations from this restriction are subtracted from $\boldsymbol{\mu}_h$, $h = 1, \ldots, N$. But it should be noted that these deviations could only be added to the unpenalized spline coefficients $\boldsymbol{\gamma}_0$ in the case of the decomposition (6.6) with equidistant knots and if $q \leq k$, i.e. if the dimension of the random effects is equal to or smaller than

the order k of the penalty matrix. For other cases we propose the following simple but effective strategy: Similar to the procedure in Section 5.2.2, we just center the cluster centers followed by an immediate update of the basis coefficients so that the P-spline parameters can absorb the general time trend. For a correct update of the variance parameters the uncentered cluster centers should be used in the working response.

For the simultaneous maximization of the variance parameters σ^2 and $\boldsymbol{D}$, given $\boldsymbol{\beta}$, $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_p$, $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ and τ^2 the algorithm AS 47 of O'Neill (1971) in the C++ version (Burkhardt, 2008) is used, which is an implementation of the Nelder-Mead algorithm (Nelder and Mead, 1965). In this optimization procedure we choose for the reflection, extension and contraction coefficients the common settings 1.0, 2.0 and 0.5 respectively. Note that the covariance matrix $\boldsymbol{D}$ is parameterized by $\boldsymbol{D} = \boldsymbol{L}\boldsymbol{L}^T$ because then the matrix is automatically nonnegative-definite and even positive-definite (and so invertible, too) if $\boldsymbol{L}$ is a matrix with exclusively nonzero diagonal entries (Lindstrom and Bates, 1988). The whole EM algorithm for fitting additive mixed models with a DPM as random effects distribution is implemented in C++ and is accessible by the R wrapper function `ammDPMEM()` in the R package `clustmixed` (Heinzl, 2012). Here, the starting values can be chosen individually. Otherwise, the following starting values are used by default: In the beginning, there are $N = n$ clusters − one for each subject with the same weight $\pi_h = 1/N$, $h = 1, \ldots, N$. Thus, during the iterations clusters are fused step by step until there is no increase of the penalized incomplete log-likelihood $l_{PI}(\boldsymbol{\xi})$ any more. This is the reason why our method can be called an agglomerative cluster approach. Rearranging the weights after each step has the effect that only the relevant clusters keep positive probabilities. As starting values for the basis coefficients least squares estimates of the model $\boldsymbol{y}_i = \boldsymbol{B}_i\hat{\boldsymbol{\gamma}}$, $i = 1, \ldots, n$, are used. With the resulting residuals as response values a linear mixed model with normally distributed random effects is fitted to get starting values for $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{D}$. In addition, cluster centers $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ are initialized by the predicted random effects $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ of this model. If $N < n$ is chosen, a k-means clustering of the predicted random effects is used for determining starting values for the cluster centers. Concerning the "penalization" parameters $\alpha = 0$ and $\tau^2 = 0.1$ are used as starting values to induce a very strong clustering and a smooth trend curve. However, it is advisable to try several different starting values to avoid that the EM algorithm converges to a local but not a global maximum. After convergence we get the cluster membership by the matrix of estimated π_{ih}. Individual i is assigned to that cluster h for which $\hat{\pi}_{ih}$ is maximal. If there are a lot of small weights $\hat{\pi}_h$, we get only few relevant clusters. Based on the weights of all clusters the random effects are predicted by using the mean of the posterior $\boldsymbol{b}_i|\boldsymbol{y}_i$, which is given by

$$\hat{\boldsymbol{b}}_i = \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \boldsymbol{B}_i\boldsymbol{T}\hat{\boldsymbol{\gamma}}_0 - \boldsymbol{B}_i\boldsymbol{W}\hat{\boldsymbol{\gamma}}_p)(\boldsymbol{I}_q - \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{Z}_i)\sum_{h=1}^{N}\hat{\pi}_{ih}\hat{\boldsymbol{\mu}}_h, \quad i = 1, \ldots, n.$$

This is a direct extension of the prediction in the case of linear mixed models, which is proved in Appendix A.4. Note that after convergence all parameters have to be restandard-

ized internally because the algorithm works with standardized variables. See Appendix A.5 for more details about the used standardization.

6.2.3. Discussion of the P-spline Term

In this section, some properties of P-splines will be discussed that are crucial for the EM algorithm presented in Section 6.2.2. First, note that the decomposition of the basis coefficient vector mentioned in Section 6.2.1 is not unique. Two variants for the choice of these matrices are conventional: One yields the matrices $\boldsymbol{T} = \boldsymbol{\Gamma}_0$ and $\boldsymbol{W} = \boldsymbol{\Gamma}_p \boldsymbol{\Omega}_p^{-1/2}$ and is based on the spectral decomposition of the singular penalty matrix

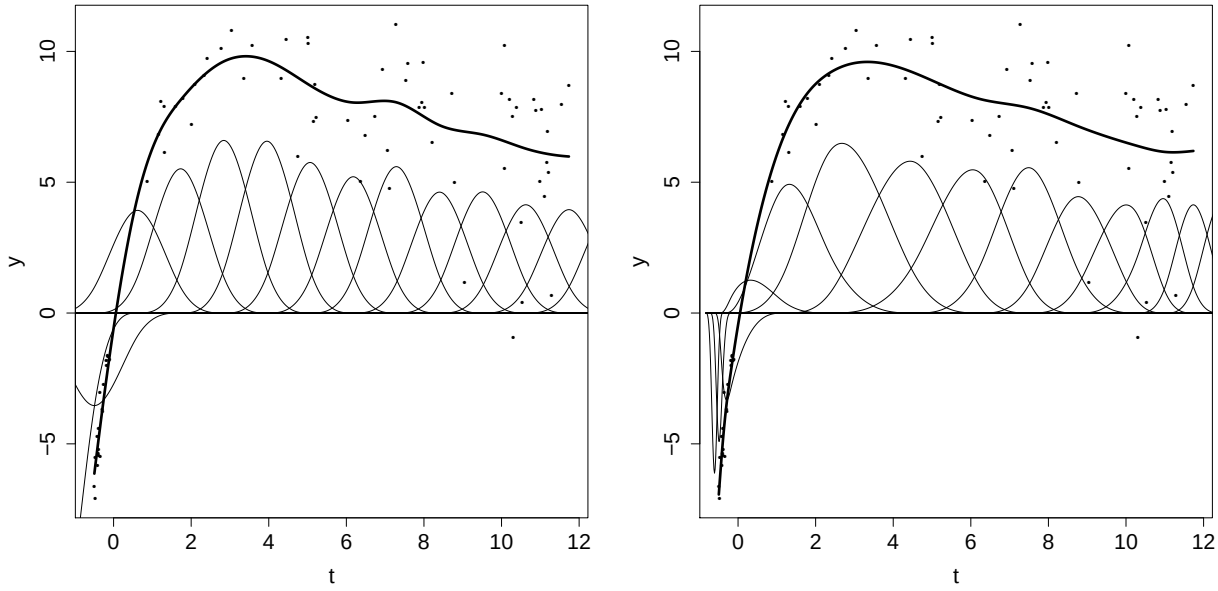
$$\boldsymbol{K} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T = \begin{pmatrix} \boldsymbol{\Gamma}_p & \boldsymbol{\Gamma}_0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\Omega}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Gamma}_p^T \\ \boldsymbol{\Gamma}_0^T \end{pmatrix} = \boldsymbol{\Gamma}_p \boldsymbol{\Omega}_p \boldsymbol{\Gamma}_p^T,$$

where $\boldsymbol{\Omega}$ is a diagonal matrix with the corresponding eigenvalues arranged in descending order on the leading diagonal and where $\boldsymbol{\Omega}_p$ contains only the $d - k$ strictly positive eigenvalues of $\boldsymbol{K}$ (Wood, 2006). The corresponding eigenvectors form the column vectors in the orthogonal matrix $\boldsymbol{\Gamma}$, respectively in the matrix $\boldsymbol{\Gamma}_p$. In the special case of the penalty matrix $\boldsymbol{K} = \boldsymbol{\Delta}^T\boldsymbol{\Delta}$ the choice

$$\boldsymbol{T} = \begin{pmatrix} 1 & \varsigma_1 & \dots & \varsigma_1^{k-1} \\ \vdots & & & \vdots \\ 1 & \varsigma_d & \dots & \varsigma_d^{k-1} \end{pmatrix} \quad \text{and} \quad \boldsymbol{W} = \boldsymbol{\Delta}^T(\boldsymbol{\Delta}\boldsymbol{\Delta}^T)^{-1}, \tag{6.6}$$

is also suitable, where $\varsigma_1, \dots, \varsigma_d$ are equidistant grid points on the relevant range. The proof that both choices for $\boldsymbol{T}$ and $\boldsymbol{W}$ ensures a decomposition into a penalized and an unpenalized part is found in Appendix A.8. There, it can be seen that the grid points $\varsigma_1, \dots, \varsigma_d$ have to be equidistant. When equidistant knots are considered, these can be used as grid points. In addition, equidistant knots offers a further benefit, which is examined in the following. First, note that P-splines based on the difference penalty of order k feature generally the property that they produce polynomials of degree $k - 1$ for a strong penalization − independently of the choice of the knots. For equidistant knots and the decomposition (6.6) the unpenalized part describes exactly this polynomial of degree $k-1$. For example, when second-order differences are used, $\boldsymbol{\gamma}_0$ contains the global intercept and the global slope which are unpenalized. The penalized coefficients $\boldsymbol{\gamma}_p$ correspond to terms of higher degrees. This gives rise to the general discussion whether equidistant knots or knots chosen as quantiles of the time variable should be preferred. While Ruppert and Carroll (2000) recommend knots based on quantiles, Eilers and Marx (2010) emphasize the benefits of equidistant knots. Apart from that, it is generally questionable if the penalty matrix $\boldsymbol{K} = \boldsymbol{\Delta}^T\boldsymbol{\Delta}$ could be used directly when knots based on quantiles are considered. In our opinion this is not only possible but also meaningful. In this case basis coefficients are penalized equally although the corresponding basis functions are unequally spaced and show different shapes. In ranges with lots of data differences between basis coefficients are penalized relatively weakly whereas in ranges with only few data a stronger penalization

can be observed. Thus, we obtain a reasonable "adaptive smoothing" in contrast to the constant smoothing for equidistant knots.



(a) Equidistant knots

(b) Knots based on quantiles

Figure 6.1.: Estimation of the P-spline by the DPM-EM approach for simulated data with substantially overlapping clusters for few individual observations ($\nu = 1$). On the left equidistant knots are considered for the P-spline while on the right the knots are based on quantiles. The thick line symbolizes the P-spline while the thin lines represent the weighted B-spline basis functions $\hat{\gamma}_s B_s^l(t)$.

This feature is demonstrated by an example. It should be noted that the underlying data are based on a setting of the simulation study in the following section. Concretely, the setting of substantially overlapping clusters with only few individual observations is used, which will be explained in Section 6.3.1. The corresponding trace plot is shown later in Figure 6.7 (top left). In Figure 6.1, the estimated P-spline (thick line) by the DPM-EM model for the simulated data can be seen for equidistant knots (left) and for knots based on quantiles (right). Here, we used $m = 12$ inner knots, B-spline basis functions of degree $l = 3$ and a difference penalty of second order. The thin lines represent the weighted B-spline basis functions $\hat{\gamma}_s B_s^l(t)$. For equidistant knots only few B-spline basis functions are available for fitting the strong increase of the spline in $t \in [-0.5, 0]$. For this purpose, a comparatively high inverse smoothing parameter $\hat{\tau}^2 = 0.33$ is necessary to permit relatively high differences between the basis coefficients. But this value seems to be too high in $t \in [4, 12]$. In contrast to equidistant knots for knots based on quantiles the amount of smoothing has not to be the same for the whole range of the time variable. Indeed, the inverse smoothing parameter is the same for all values of t, but it is lower ($\hat{\tau}^2 = 0.09$) than in the case of equidistant knots. The reason for this is that for knots based on quantiles more B-spline basis functions are available in ranges with many data as in $t \in [-0.5, 0]$. Thus, the differences between the basis coefficients can be smaller in

these ranges corresponding to a lower inverse smoothing parameter. This yields a smoother trend curve.

6.3. Simulation Study

In the following section, the settings and the results of a simulation study are presented in which the prediction accuracy of random effects and of the whole individual curves is examined. Here, we are interested in whether additive mixed models considering a DPM as random effects distribution yield better prediction results than additive mixed models with normally distributed random effects when the true random effects distribution is a mixture of three normal distributions. Furthermore, the performances of the proposed EM approach and the MCMC approach of Chapter 5 for fitting additive mixed models with DPMs are compared.

6.3.1. Settings

More concretely, in the simulation study 100 data sets are generated. Each data set consists of $n = 20$ individuals with response values simulated by

$$y_{ij}|\boldsymbol{b}_i \overset{ind.}{\sim} N(f(t_{ij}) + b_{i0} + t_{ij}b_{i1},\, \sigma^2), \quad i = 1, \ldots, 20, \quad j = 1, \ldots, n_i,$$

where $f(t) = \frac{50 \cdot \log(0.2\,t+1)}{(0.2\,t+1)^2}$ represents a nonlinear global time trend. The error variance is fixed on $\sigma^2 = 0.25$. In each simulation run different "true" random effects $\boldsymbol{b}_i = (b_{i0}, b_{i1})^T$ are drawn from a mixture distribution of three normal distributions

$$\boldsymbol{b}_i \sim 0.4\, N(\boldsymbol{\mu}_1, \boldsymbol{D}) + 0.3\, N(\boldsymbol{\mu}_2, \boldsymbol{D}) + 0.3\, N(\boldsymbol{\mu}_3, \boldsymbol{D}), \quad i = 1, \ldots, 20,$$

imitating a population consisting of three clusters of overlapping subpopulations. The covariance matrix in each cluster is given by $\boldsymbol{D} = \text{diag}(0.1, 0.1)$. However, we vary the differences between the clusters and distinguish between three scenarios:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -4.5 \\ 1.5 \end{pmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 1.5 \\ -1.8 \end{pmatrix}, \qquad \boldsymbol{\mu}_3 = \begin{pmatrix} 4.5 \\ -0.2 \end{pmatrix},$$

corresponding to *clearly separated clusters*,

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -1.5 \\ 0.75 \end{pmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 0.5 \\ -0.9 \end{pmatrix}, \qquad \boldsymbol{\mu}_3 = \begin{pmatrix} 1.5 \\ -0.1 \end{pmatrix},$$

corresponding to *moderately separated clusters*, and

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -0.3 \\ 0.375 \end{pmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 0.1 \\ -0.45 \end{pmatrix}, \qquad \boldsymbol{\mu}_3 = \begin{pmatrix} 0.3 \\ -0.05 \end{pmatrix},$$

corresponding to *substantially overlapping clusters*.

In addition, in each of these scenarios three different settings for the individual numbers of observations are considered. To produce longitudinal data with varying numbers of repeated observations per unit i, we set $n_i = 3 + X_i$, where X_i is Poisson distributed with rate ν. Setting $\nu = 1$ corresponds to longitudinal data with only *few individual observations* (4 on average), $\nu = 3$ to a *medium number of individual observations* and $\nu = 5$ to comparably *many individual observations*. For given n_i, the observation times are generated from diverse uniform distributions $U(a, b)$ with lower bound a and upper bound b. For each subject $i = 1, \ldots, n$, the first measuring point t_{i1} is drawn from $U(-0.5, 0)$ while the last measuring point is simulated by $t_{in_i} \sim U(10, 12)$. To generate the remaining time points, first, the medial interval $[0, 10]$ is partitioned into $n_i - 2$ subintervals with equal lengths and corresponding means $\zeta_2, \ldots, \zeta_{n_i-1}$. Then, the observation times are generated from intervals with the same mean but with bisected length: $t_{ij} \sim U(\zeta_j - \frac{2.5}{n_i-2}, \zeta_j + \frac{2.5}{n_i-2})$, $j = 2, \ldots, n_i - 1$. The bisection is used to avoid huge jumps of response values at measuring points which are very close to each other. The simulation concept is visualized in Appendix A.9. In summary, in each simulation run $s = 1, \ldots, 100$ we get different numbers of observations, time points, random effects and response variables for each subject.

Combining these different settings for observations times and clusters, results in nine different scenarios. For each of them we use additive mixed models with random slopes and a cubic P-spline with 12 equidistant inner knots based on a difference penalty of second order for fitting the unknown trend function $f(\cdot)$. However, we vary the assumption for the random effects distribution and the estimation procedure. On the one hand, additive mixed models with normally distributed random effects are considered, estimated via MCMC methods (ND-MCMC) respectively the REML approach (ND-REML) as implemented in BayesX (Brezger et al., 2005). On the other hand, we apply the approach proposed in Section 6.2 with a DPM as random effects distribution estimated via EM algorithm (DPM-EM) and compare it to the corresponding MCMC-approach from Chapter 5 (DPM-MCMC). In addition, based on the considerations in Section 6.2.3 for the DPM-EM approach knots chosen as quantiles of the time variable are also considered for an adaptive smoothing. For these five approaches the fit of individual curves as well as clustering related characteristics are compared. More concretely, in each simulation run s, we calculate the average prediction error of all individual curves

$$PE(s) = \frac{1}{n} \sum_{i=1}^{n} \int_{-0.5}^{12} \left(\hat{f}_{is}(t) - f_{is}(t) \right)^2 dt, \tag{6.7}$$

with $f_{is}(t) = f(t) + b_{i0}^{(s)} + t \cdot b_{i1}^{(s)}$ and with $\hat{f}_{is}(t)$ as the corresponding estimate. In the criterion (6.7) the integral is approximated by the trapezoidal rule. The empirical distribution of the average prediction errors $PE(s)$ obtained from simulation run $s = 1, \ldots, 100$ is then represented through box plots. In addition, the estimated numbers of clusters are examined for the approaches with a DPM as random effects distribution. Of course, for the mixed models with normally distributed random effects we obtain one cluster by construction for all simulation settings.

6.3.2. Results

Clearly separated clusters

In Figure 6.2 (top), two examples of the trace plots in the setting of clearly separated clusters can be seen. On the left only few individual observations are available while on the right in the average six observations per subject are given. In both cases our DPM-EM approach with knots chosen as quantiles of the time variable finds three clusters as it can be seen in Figure 6.2 (below). Here, the solid lines illustrate the three cluster centers while the dashed line represents the general time trend. Observations belonging to the same cluster are marked with the same symbol. To each solid line the corresponding symbol is added to visualize which cluster center belongs to which cluster.

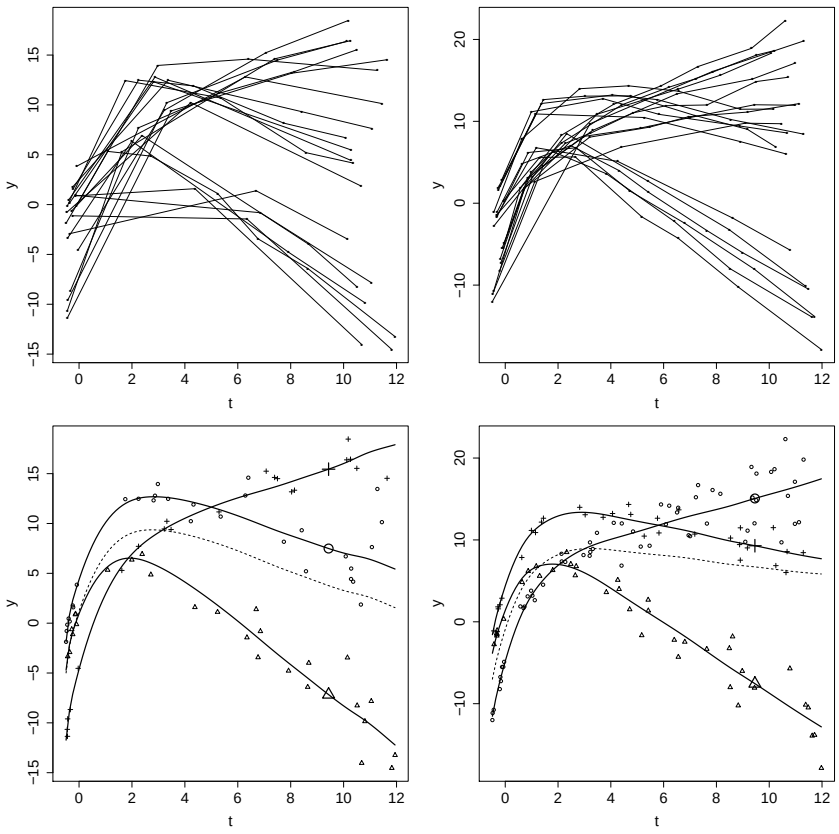
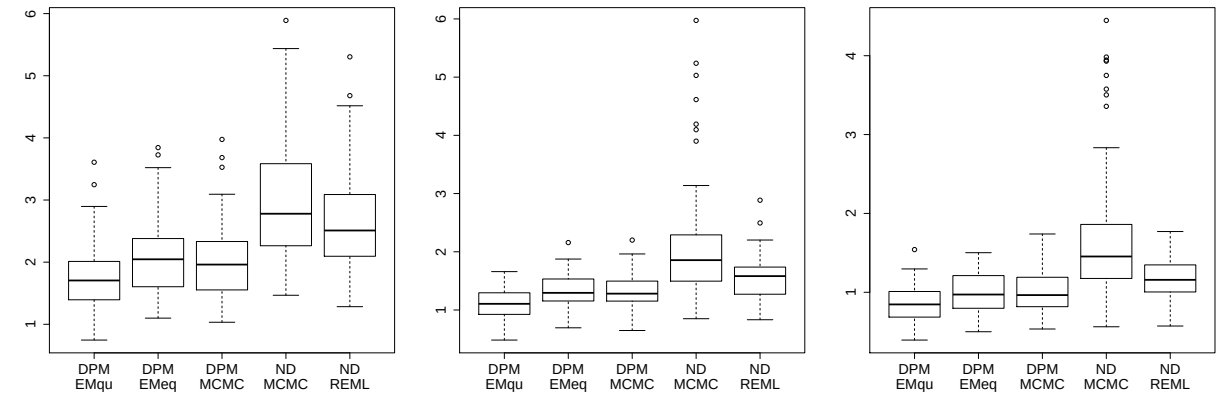


Figure 6.2.: Trace plots (top) and clustering by the DPM-EM approach with knots based on quantiles (below) with clearly separated clusters for few individual observations ($\nu = 1$) (left) and a medium number of individual observations ($\nu = 3$) (right).

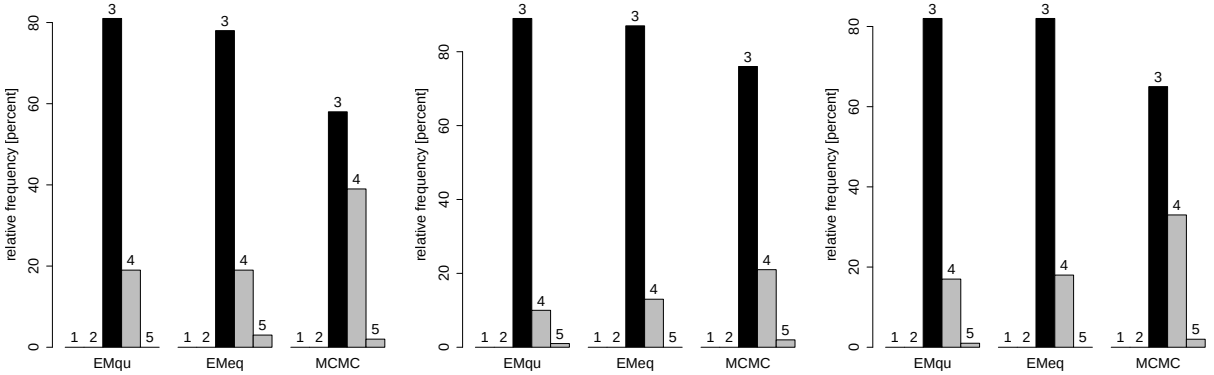
Figure 6.3 shows that the individual curves are fitted much better by the DPM models than by the models using normally distributed random effects. Especially the classical additive mixed model with a normal distribution as random effects distribution using MCMC methods (ND-MCMC) features a higher prediction error than the model using



(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

Figure 6.3.: Box plots of PE with clearly separated clusters for few individual observations (left), a medium number of individual observations (middle) and many individual observations (right).

restricted maximum likelihood as inference tool (ND-REML). The performance of the DPM models with equidistant knots (DPM-EMeq, DPM-MCMC) is quite similar, regardless of the estimation procedure. Using knots based on quantiles (DPM-EMqu), the prediction accuracy can even be improved.



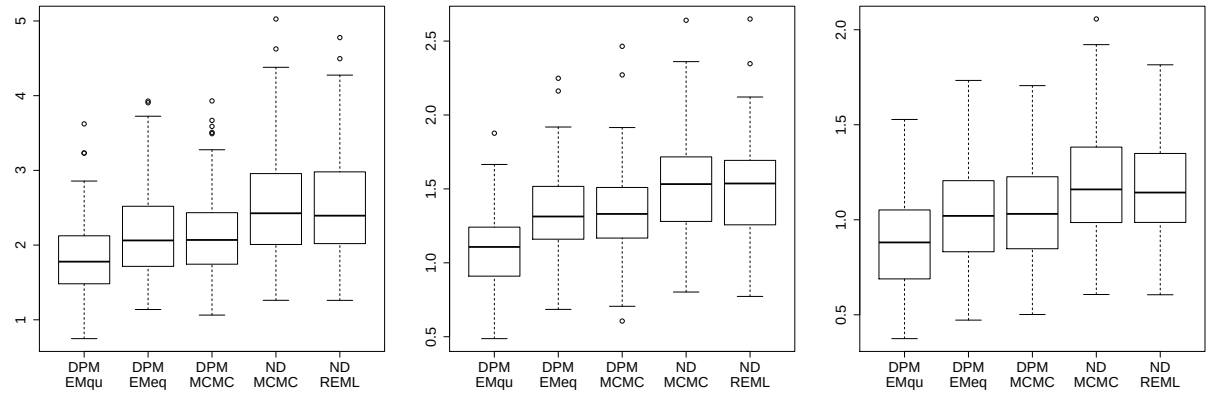
(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

Figure 6.4.: Bar plots of the estimated numbers of clusters by the DPM approaches with clearly separated clusters for few individual observations (left), a medium number of individual observations (middle) and many individual observations (right).

The clustering related characteristics are shown in Figure 6.4. In this figure, the bar corresponding to three clusters is highlighted by black color because in the simulation setting three clusters are used. We get quite similar results for the three scenarios with varying individual observations. Obviously, in the most cases three clusters are detected by the DPM approaches. The DPM approach using MCMC methods (DPM-MCMC) tends

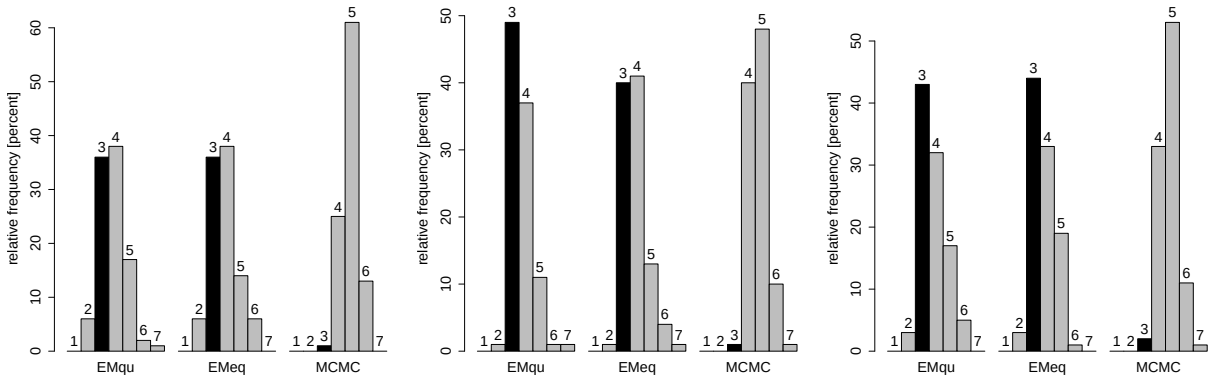
to detect a bit more clusters than the DPM approaches based on the EM algorithm (DPM-EMeq, DPM-EMqu), which show quite similar results with regard to the estimated number of clusters.

Moderately separated clusters



(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

Figure 6.5.: Box plots of PE with moderately separated clusters for few individual observations (left), a medium number of individual observations (middle) and many individual observations (right).



(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

Figure 6.6.: Bar plots of the estimated numbers of clusters by the DPM approaches with moderately separated clusters for few individual observations (left), a medium number of individual observations (middle) and many individual observations (right).

For a smaller separation of the cluster centers the DPM approaches still outperform the classical mixed models with normally distributed random effects (Figure 6.5): Now, the prediction accuracy is nearly the same for the classical methods ND-MCMC and ND-REML. However, lower prediction errors can be achieved by using DPM approaches. Again,

we obtain similar results for the both DPM approaches with equidistant knots (DPM-MCMC, DPM-EMeq). Their prediction error can only be outperformed by the DPM-EM approach with knots chosen as quantiles (DPM-EMqu).

In Figure 6.6, it can be seen that apparently more clusters are detected than in the setting of clearly separated clusters: For the DPM approach using MCMC methods the modus of the distribution for the estimated numbers of clusters is five while for the DPM-EM approaches mostly three or four clusters are found. For few individual observations the estimated number of clusters tends to be a bit higher.

Substantially overlapping clusters

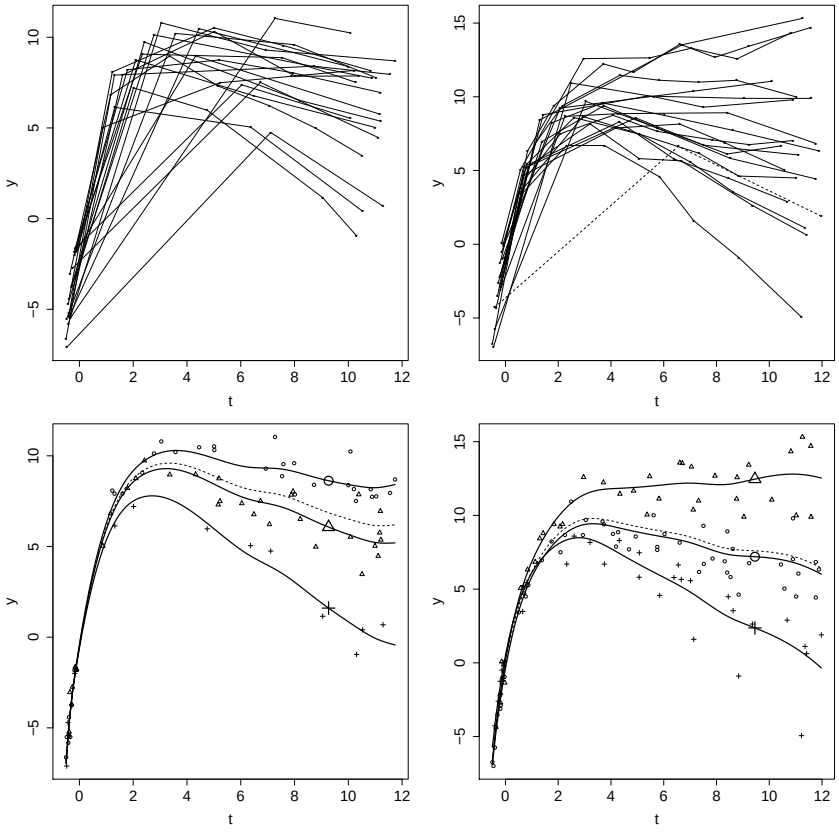
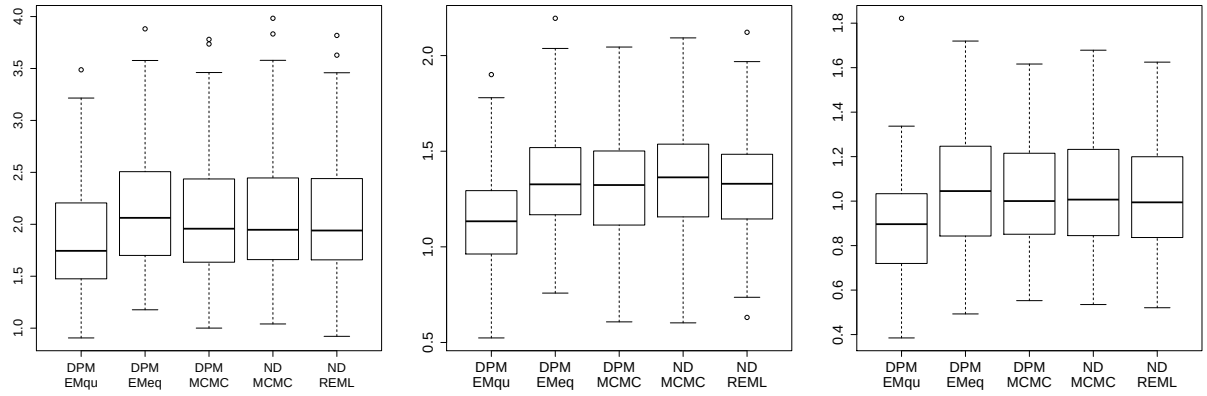


Figure 6.7.: Trace plots (top) and clustering by the DPM-EM approach with knots based on quantiles (below) with substantially overlapping clusters for few individual observations ($\nu = 1$) (left) and a medium number of individual observations ($\nu = 3$) (right).

In the scenario of substantially overlapping clusters we pick up the example of Section 6.2.3 for few individual observations. See Figure 6.7 for the according trace plot (top left) and the clustering by the DPM-EM approach with knots based on quantiles (below left). Obviously, three clusters are detected. For the data with a medium number of individual observations (Figure 6.7, top right) three clusters are found by our DPM-EM

approach (below right), too. Let regard these plots in more detail. In Figure 6.7 (top right), subject 8 (dashed line) seems to have a quite special individual curve and one could expect that this subject forms its own cluster. However, this is just a visual effect because no measurements are available for this subject in the time interval $(-0.427, 6.636)$. Actually subject 8 is assigned to cluster 3 (+) together with four other individuals by the DPM-EM approach. If one is interested in predicting response values for this subject in the concerning interval, for this purpose cluster 3 can be used.



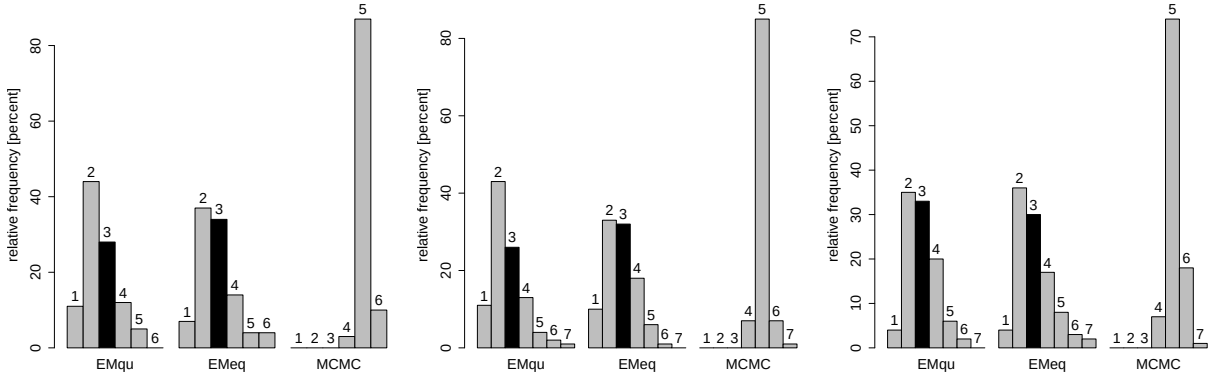
(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

Figure 6.8.: Box plots of PE with substantially overlapping clusters for few individual observations (left), a medium number of individual observations (middle) and many individual observations (right).

With regard to the prediction accuracy we conclude the following: For substantially overlapping clusters the prediction errors are nearly the same for the approaches using equidistant knots regardless the assumption for the random effects distribution (Figure 6.8). Only the prediction accuracy for the DPM-EM approach with equidistant knots is a bit worse. The reason for that is that the estimated splines are considerably rough as it can be seen, for example, on the left side of Figure 6.1. For the DPM-EM approach with knots based on quantiles, however, the best performance can be observed. See Section 6.2.3 for a discussion about the choice of knots.

According to Figure 6.9 for the DPM-EM models mostly two or three clusters are detected. As expected, it is more difficult to distinguish between the clusters in the setting of substantially overlapping clusters. However, for the DPM approach using MCMC methods in the most cases still five clusters are found.

In summary, we conclude that the proposed DPM-EM approach improves the prediction accuracy with regard to the fitted individual curves compared to methods that assume normally distributed random effects. The prediction errors for the DPM approach using MCMC methods tend to be a bit lower than these of the DPM-EM approach, when using equidistant knots. However, the best performance in the meaning of prediction errors can be stated for the DPM-EM approach with knots based on quantiles.



(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

Figure 6.9.: Bar plots of the estimated numbers of clusters by the DPM approaches with substantially overlapping clusters for few individual observations (left), a medium number of individual observations (middle) and many individual observations (right).

6.4. Applications

6.4.1. Theophylline

In the following, the approach introduced in Section 6.2 will be applied to the theophylline data that were reported by Boeckmann et al. (1994). In this study, the anti-asthmatic drug theophylline was administered orally to twelve test persons, and serum concentrations were measured at several time points. Figure 6.10 (left) shows the concentration-time profiles of the considered subsample. It is seen that after the drug administration the theophylline concentration in the sample increases steeply at first, followed by a weak decrease. In addition, the data set contains two further covariates: `weight` and `dose`. These covariates are invariants, i.e. the dose was given on a per-weight basis: lower doses were administered to heavy-weighted people. While Davidian and Giltinan (1995) and Pinheiro and Bates (2000) considered a two-compartment open pharmacokinetic model, we aim to identify clusters by using the DPM-EM model for additive mixed models. Concretely, we consider a random slope model for the theophylline concentration in the sample $\texttt{conc}_{ij}$ of subject i at measurement j

$$\texttt{conc}_{ij}|\boldsymbol{b}_i \overset{ind.}{\sim} N(f(\texttt{time}_{ij}) + b_{i0} + \texttt{time}_{ij} b_{i1} + \texttt{weight}_i \beta_1,\, \sigma^2), \quad i = 1, \ldots, 12, \quad j = 1, \ldots, 10.$$

For the nonlinear term $f(\texttt{time})$ a cubic P-spline with $m = 12$ inner knots based on quantiles of the time variable is used. The basis coefficients are penalized by a difference penalty of second order based on the decomposition (6.6). See Section 6.2 for more details about this choice. The DPM for the random effects allows to identify clusters due to individual deviations from the population trend. Indeed, our approach detects three clusters (Figure

6.10, right) for the estimated concentration parameter $\hat{\alpha} = 0.00164$. The shapes of the trend curves of cluster 2 ($\triangle$) and cluster 3 ($+$) seem to be alike but on different levels. In Cluster 2 the intercept is about $\hat{\mu}_{20} = 0.335$ higher than the base level while in cluster 3 it is about $\hat{\mu}_{30} = -1.748$ lower. The corresponding slopes tend to be a bit higher compared to the global trend curve ($\hat{\mu}_{21} = 0.133$, $\hat{\mu}_{31} = 0.067$). Cluster 1 ($\bigcirc$) is characterized by the strongest decrease ($\hat{\mu}_{11} = -0.100$) after the maximum at two hours. The level of cluster 1 ($\hat{\mu}_{10} = 0.059$) resembles that of the global trend curve.

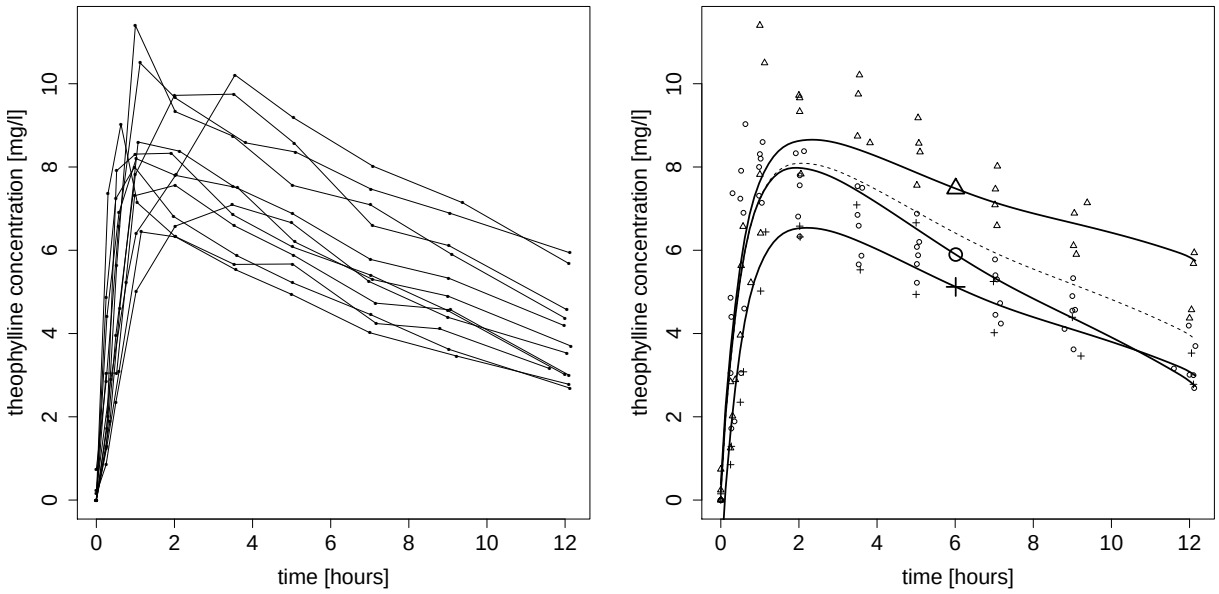


Figure 6.10.: Theophylline concentration in the sample across time: raw data (left) and clustering by the DPM-EM approach (right). On the right observations belonging to the same cluster are marked with the same symbol. The dashed line represents the population effect, the solid lines symbolize the cluster effects.

	estimate	standard error	95%-CI lower	95%-CI upper
`weight`	0.012	0.047	-0.098	0.047
σ^2	1.226	1.557	0.605	1.557
σ_0^2	0.039	0.618	0.000	0.618
σ_1^2	0.003	0.014	0.000	0.014
σ_{01}	-0.010	0.011	-0.071	0.011

Table 6.1.: Estimation results for the fixed effects and variance parameters by the DPM-EM approach for the theophylline data.

Table 6.1 shows the estimated fixed effect and the variance parameters. The corresponding standard errors and confidence intervals have been estimated by the nonparametric

bootstrap method proposed by Efron (1979) with 1000 replications. The confidence intervals are based on the bootstrap quantiles. Since the confidence interval for β_1 includes zero, the covariate `weight` has no general significant effect on the theophylline concentration on the five percent level. However, in Figure 6.11 it is seen that the distribution of the variable `weight` differs between the clusters. In cluster 2 ($\triangle$) mostly lightweight people with considerably high doses of the drug can be found. As expected, people with lower weights and higher doses show a higher trend of the theophylline concentration in the sample (Figure 6.10, right).

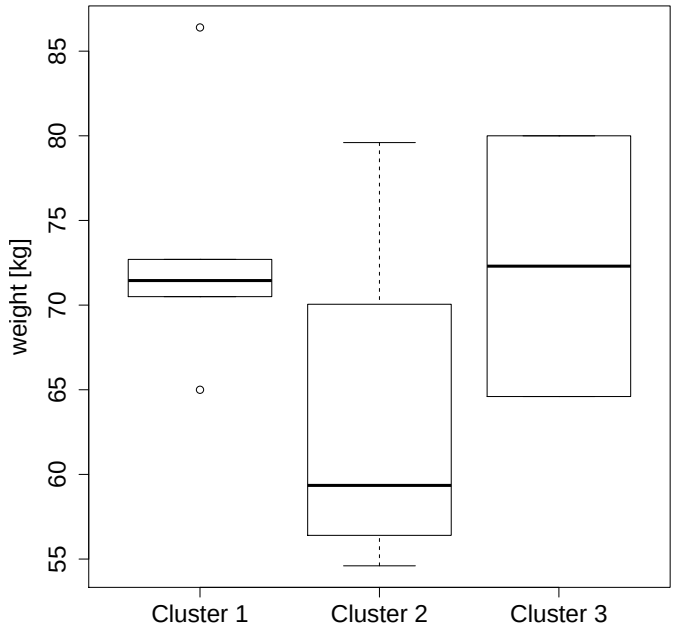


Figure 6.11.: Distribution of the variable `weight` in the three clusters corresponding to the clustering by the DPM-EM approach.

6.4.2. Childhood Obesity

As second application we reanalyze data from the LISA study. In this study the influences of **L**ife-style factors on the development of the **I**mmune **S**ystem and **A**llergies in East and West Germany are examined for 3,097 healthy neonates born between November 1997 and January 1999 in 14 obstetrical clinics in Munich, Leipzig, Wesel, and Bad Honnef. A detailed description of the study can be found, for example, in Chen et al. (2007) and Zutavern et al. (2007). We are mainly interested in the longitudinal BMI profiles of the children and aim to expose clusters in the BMI profiles over time by our DPM-EM approach. In particular, it is of interest whether a cluster of obese children can be detected and if so how the trajectory of this cluster can be described and which indicators can be

found for this childhood obesity. Figure 5.7 (left) shows the development of the BMI for twelve randomly selected children, while in Figure 5.7 (right) all measurements are drawn. In the given data the children have been examined until the age of six by questionnaires at birth and around the age of 2 weeks, 1, 3, 6, 12, 24, 48 and 60 months. Thus, up to 9 measurements are available. We handle missing data problems by a complete case analysis: Following Fenske et al. (2008), children were excluded from the analysis if an observation of a time constant covariate was missing. If only a single observation of age or BMI was missing, only this particular observation was excluded from the analysis. Finally, 2,043 children and 17,316 observations are available.

All in all, one has to deal with a huge data set with highly nonlinear growth patterns, long individual time series, clustered individual-specific deviations from the population trend and irregular time points. We consider the DPM-EM model proposed in Section 6.2. Here, a cubic P-Spline of second order with 12 inner knots based on quantiles is used to achieve a smooth trend curve even in these ranges where almost no data are available. To cluster the BMI trajectories, an approximate DPM as random effects distribution is assumed. Following the argumentations in Section 6.2.2, we truncate the Dirichlet process at $N = 11$. In addition, we use the same predictors as in Section 5.4. See Table 5.1 for an overview of the categorial and continuous covariates included in the analysis. Altogether, for the measurement $j = 1, \ldots, n_i$ of subject $i = 1, \ldots, n$ we consider

$$\begin{aligned} \texttt{BMI}_{ij} | \boldsymbol{b}_i \overset{ind.}{\sim} N(&\texttt{sex}_i\beta_1 + \texttt{breast}_i\beta_2 + \texttt{mSmoke}_i\beta_3 + \texttt{area}_i\beta_4 + \texttt{mBMI}_i\beta_5 + \texttt{mDiffBMI}_i\beta_6 + \\ &+ f(\texttt{ageY}_{ij}) + b_{i0} + \texttt{ageY}_{ij} b_{i1},\, \sigma^2). \end{aligned}$$

Some authors like Beyerlein et al. (2008) and Mayr et al. (2012) argue that the distribution of BMI values is typically skewed depending on the age of children. However, we assume a symmetric distribution since Fenske et al. (2008) found out that for the given data with measurements up to the age of six years the distributional shape of children's BMI is rather symmetric. Solely for the extended LISA study, where one additional measurement per child at about the age of ten years is given, the BMI distribution becomes right-skewed at the age of ten years (Mayr et al., 2012).

With regard to the fixed effects (Table 6.2) we obtain the same significant predictors as in Section 5.4 and quite similar results for the estimated coefficients. The expected BMI of the boys is somewhat larger than that of the girls if all other covariates are kept fixed. The gender has a significant impact on the child's BMI, since the corresponding 95% confidence interval does not include zero. Note that the given confidence intervals are based on the widely-used test statistic $\hat{\beta}_r / \widehat{se}(\hat{\beta}_r)$, whose distribution can be approximated by a standard normal distribution. The standard errors have been estimated by the nonparametric bootstrap method of Efron (1979) with 140 replications. They seem to be a bit higher than in Section 5.4. Positive significant effects can also be stated for the maternal BMI and the maternal BMI gain during pregnancy while the general effects of the covariates `breast`, `mSmoke` and `area` are not significantly different from zero. However, we will see later in this section that the impact of these covariates may depend upon the clusters.

	estimate	standard error	95%-CI lower	95%-CI upper
`sex`	0.300	0.043	0.217	0.383
`breast`	0.054	0.040	-0.059	0.097
`mSmoke`	-0.019	0.059	-0.061	0.169
`area`	0.019	0.055	-0.128	0.090
`mBMI`	0.044	0.006	0.032	0.056
`mDiffBMI`	0.064	0.011	0.042	0.086
σ^2	0.915	0.016	0.883	0.947
σ_0^2	0.259	0.087	0.088	0.430
σ_1^2	0.019	0.007	0.006	0.032
σ_{01}	-0.006	0.023	-0.051	0.039

Table 6.2.: Estimation results for the fixed effects and variance parameters by the DPM-EM approach for the LISA data.

Figure 6.12 shows that five clusters are detected by the DPM-EM model, which was not obvious when looking at the raw data in Figure 5.7. Note that the concentration parameter is estimated by $\hat{\alpha} = 0.00224$. The clusters are highlighted by solid colored lines. Observations belonging to the same cluster are marked with the same color. The dashed black line represents the population effect. As in Section 5.4 a cluster of obese children can be found, which is marked by the light blue color and which we call cluster 5. The probability of this cluster and thus the probability of a child to get obese is given by $\hat{\pi}_5 = 0.023$. Interestingly, this cluster shows a normal trajectory in the first six months. Not till then a strong increase of the BMI is observed. In contrast, for the most children in cluster 1 (green, $\hat{\pi}_1 = 0.476$) and 2 (orange, $\hat{\pi}_2 = 0.401$) the BMI is descending after six months while in cluster 3 (dark blue, $\hat{\pi}_3 = 0.056$) a somewhat constant BMI profile is seen. Due to the trajectory of cluster 4 (violet, $\hat{\pi}_4 = 0.043$) parents do not have to be worried if their child shows plenty of baby fat and a high BMI in the first months because in the age of six years children of the violet cluster show a normal BMI. We conclude that a high value of BMI in the first year of one's life is no sign for obesity.

In Figure 6.13, the random intercepts and the random slopes are drawn for all children. In addition, the two-dimensional cluster centers $\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_5$ are shown. In this plot it is seen, how subjects with similar random effects are assigned to the same cluster. These subjects are marked with the same color. Again, the light blue cluster is eye-catching since it exhibits a considerably high slope ($\hat{\mu}_{51} = 0.729$). The intercept is a bit smaller than that of the population: $\hat{\mu}_{50} = -0.540$. The green ($\hat{\boldsymbol{\mu}}_1 = (-0.679, 0.049)^T$), the orange ($\hat{\boldsymbol{\mu}}_2 = (0.472, -0.090)^T$) and the dark blue cluster ($\hat{\boldsymbol{\mu}}_3 = (1.029, 0.216)^T$) are next to the overall mean, which is highlighted by a black square at coordinates (0,0). A high intercept ($\hat{\mu}_{40} = 2.042$) and a low slope ($\hat{\mu}_{41} = -0.372$) characterize the violet cluster. The estimated conditional distribution of random effects in the clusters is visualized by ellipses with level 0.95. See Section 3.3.3 for more information about the construction of these ellipses.

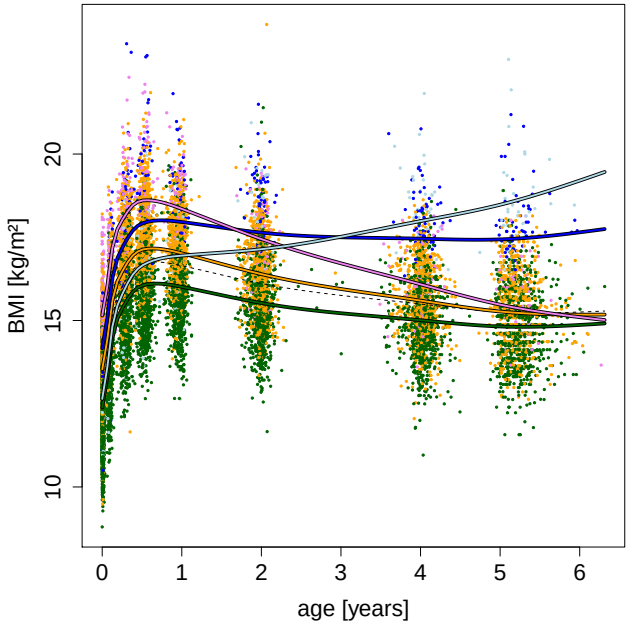


Figure 6.12.: Clustering of the LISA data by the DPM-EM model. Observations belonging to the same cluster are marked with the same color. The dashed black line represents the population effect, the solid colored lines symbolize the cluster effects.

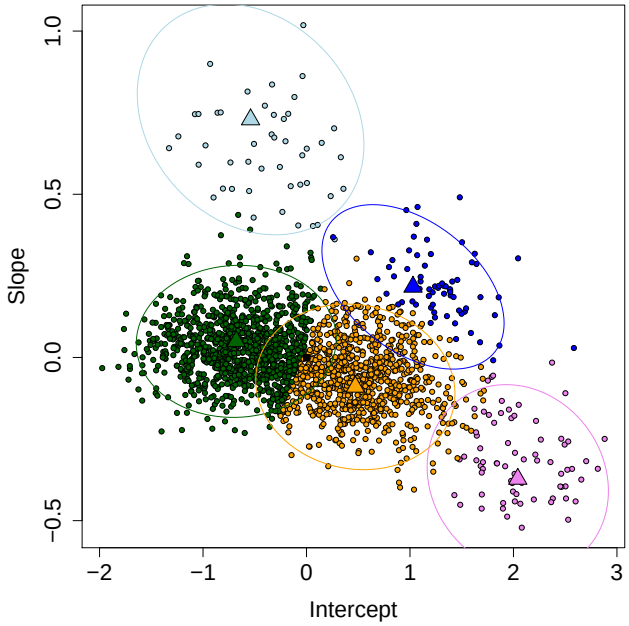
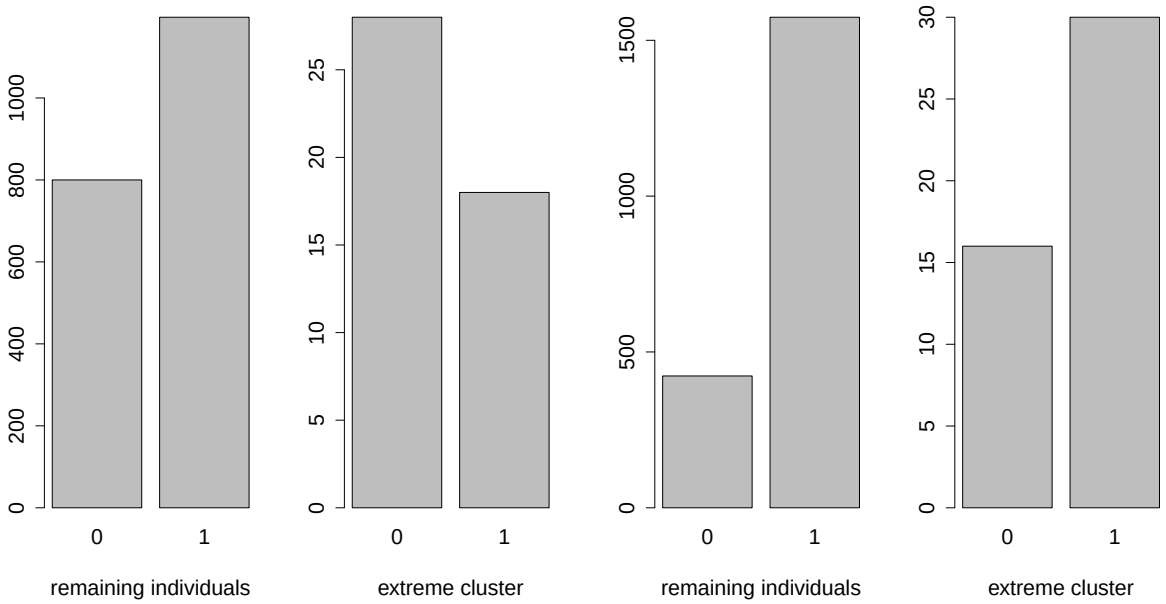


Figure 6.13.: Cluster locations and random effects of DPM-EM model for the LISA data: The big triangles symbolize the cluster locations $\hat{\boldsymbol{\mu}}_h$, the small points the random effects $\hat{\boldsymbol{b}}_i$. Subjects belonging to the same cluster are marked with the same color. The black square at coordinates (0,0) marks the population effect. Ellipses with level 0.95 visualize the estimated conditional distribution of random effects in the clusters.



(a) breastfeeding (b) area of residence

Figure 6.14.: Bar plots of the covariates `breast` (left) and `area` (right), each for the subjects of the extreme cluster (on the right hand) and for the others (on the left hand) corresponding to the clustering by the DPM-EM approach.

In the following, the impacts of the covariates `breast` and `area` are examined in more detail. The effect of breastfeeding is discussed extensively in the literature. For example, Arenz et al. (2004), Harder et al. (2005) and Rzehak et al. (2009) observed a slightly lower risk of being overweight for breastfed children and so a protective effect of breastfeeding. However, a significant effect of breastfeeding on the mean of the BMI distribution could neither be verified in the analyzes of Beyerlein et al. (2008), who used general linear models (McCullagh and Nelder, 1989) and generalized additive models for location, scale and shape (Rigby and Stasinopoulos, 2005), nor in this chapter as well as in Section 5.4. In addition, Fenske et al. (2008) and Mayr et al. (2012) considered additive quantile regression models (Koenker, 2005) and were also unable to provide evidence of a significant effect of breastfeeding on the upper quantiles (0.9, 0.97, 0.975) of the BMI distribution. However, in Figure 6.14 (left) we compare the frequency of children with `breast` $= 1$, i.e. of children that were only breastfed, in the extreme cluster 5 (light blue cluster in Figure 6.12 and Figure 6.13) and in the subpopulation of the remaining individuals. Obviously, the majority of the remaining children were breastfed only, while most of the children in the extreme cluster were bottlefed or bottle- and breastfed. Thus, breastfeeding can be seen an indicator for a normal and a lower development of the BMI. This conclusion is in agreement with the results in Section 5.4. Similarly, the ratio of children living in an urban area (`area` $= 1$) as compared to children living in a rural area is quite different in the two subpopulations:

In the extreme cluster the ratio is about 2:1 while for the remaining children it is given by circa 4:1 (Figure 6.14, right).

6.5. Summary and Discussion

In this chapter, an additive mixed model with a P-spline for the nonlinear time trend and an approximate DPM as random effects distribution is proposed, which is estimated by the EM algorithm. The feature of the EM algorithm of converging to fixed values is an advantage in the context of Dirichlet processes over MCMC methods, which are characterized by convergence to distributions. That is why the cluster property of the Dirichlet process can be used directly. Thus, our DPM-EM algorithm is able to cluster individuals in longitudinal data with a data driven identification of the number of clusters. We illustrated the algorithm in detail and discussed diverse model settings. In a simulation study it is shown that the goodness of fitted individual curves can be improved by the DPM-EM approach compared to the Bayesian approach in Chapter 5 and to methods that use normally distributed random effects. In addition, we showed that the DPM-EM can be used to find clusters in the theophylline data and to the LISA data.

ACKNOWLEDGEMENTS: We thank Elisabeth Thiering and Dr. Joachim Heinrich from the Helmholtz Zentrum Munich for providing the data of the LISA study.

7. Conclusion and Outlook

In this thesis, different concepts of clustering individuals in longitudinal data are proposed for linear and additive mixed models. These concepts are mainly based on a specific distribution assumption for the random effects that replaces the traditional normal distribution. More concretely, the used random effects distributions are basically mixture distributions that aim at accounting for a possible heterogeneity in the random effects. A general discussion about the reasons and consequences of this heterogeneity was given in Section 3.4.4.

One new approach, which is introduced in Chapter 3 within the framework of linear mixed models, assumes a finite normal mixture as random effects distribution, in which the pairwise distances of the means of these normal distributions are penalized by a group fused lasso penalty term. We used an EM algorithm for estimating the model parameters. An alternative approach in Chapter 4 is based on an approximate DPM as random effects distribution and makes use of the cluster property of the Dirichlet process for finding clusters of subjects. This feature is explained in Chapter 2. Again, an EM algorithm is outlined that solves the estimation problem in linear mixed models. In particular, the embedding of Dirichlet processes in the likelihood inference by using the EM algorithm instead of MCMC methods to maximize a penalized log-likelihood is an innovation. The prediction accuracy for the random effects of these approaches and two alternative methods are compared in a simulation study (Section 4.3.3). It is shown that the proposed approaches outperform the classical linear mixed model with normally distributed random effects and the unpenalized heterogeneity model of Verbeke and Lesaffre (1996) in the used settings. The DPM-EM algorithm mostly yields better results than the penalized mixture approach based on the group fused lasso penalty term. Additionally, a lower computation time for the DPM-EM algorithm is observed. In the application examples in Section 3.3 and Section 4.4 among other applications the unemployment data and the lung function growth data are analyzed by the two methods. Here, we saw that the DPM-EM approach apparently tends to detect fewer and more homogeneous clusters. The DPM-EM approach was extended to additive mixed models in Chapter 6. An extension of the penalized heterogeneity model with the group fused lasso penalty to additive mixed models has not been done yet, but seems to be realizable without further difficulties.

In Chapters $5 - 6$ additive mixed models are treated. On the one hand, in Chapter 5 a new combination of a P-spline for the nonlinear time trend and an approximate DPM for the random effects using MCMC methods is proposed. On the other hand, in Chapter 6 the same model is considered but with an EM algorithm as inference tool. In the simulation study in Section 6.3 it can be seen that both DPM approaches yield lower prediction errors for fitting individual curves than models with normally distributed random effects. Using equidistant knots the prediction accuracy of the DPM-MCMC approach is slightly better

than that of the DPM-EM method. We found out that knots based on quantiles can be recommended for the DPM-EM approach (Section 6.2.3). Thereby one obtains smoother splines and lower prediction errors. In the applications in Section 5.4 and in Section 6.4.2 the LISA study is picked up in order to examine the BMI profiles of children. While for the DPM-MCMC method further calculations are necessary to get a clustering for the children, the DPM-EM approach has the huge benefit that the cluster property of the Dirichlet process is used directly. However, we observed a faster computation time for the MCMC approach. A more interpretable clustering with fewer clusters is obtained by the DPM-EM algorithm. Interestingly both approaches detect one extreme cluster with a strong increase of the BMI as of the age of six months. This cluster is suitable to describe the development of childhood obesity.

In the following, some possible generalizations of the methods proposed in this thesis are discussed. Extensions to generalized linear or generalized additive mixed models with a non-normal response distribution are obvious and desirable. These generalizations seem to be realizable, in principle. However, there is need for further research.

Another interesting aspect concerning the models with Dirichlet processes would be to compare the results of Chapters $4-6$, which are based on the two inference procedures EM algorithm and MCMC methods, to results based on another upcoming inference tool called *variational approximations* (Blei and Jordan, 2006). These methods should also yield a possibility to make use of the cluster property of Dirichlet processes directly. However, the implementation of the variational approximations for DPMs in the framework of linear or additive mixed models seems to be challenging.

A further idea for future research could be the combination of the random effects terms in this thesis with penalty terms for the fixed effects. So, for example, Groll and Tutz (2013) proposed a regularization approach for fixed effects in generalized linear mixed models, which is based on the L_1-penalty of Tibshirani (1996). This approach could be extended by replacing the assumption of normally distributed random effects by an approximate DPM or the penalized normal mixture based on the group fused lasso penalty, which were proposed in this thesis. Certainly, alternative regularization concepts like boosting can also be used for variable selection. See, for example, Freund and Schapire (1997) and Bühlmann and Hothorn (2007) for background knowledge about boosting. A likelihood-based boosting approach for generalized additive mixed models was implemented by Groll and Tutz (2012) but with normally distributed random effects. More general random effects distributions as proposed in this thesis could be used.

In summary, several possibilities for model-based clustering in longitudinal data are presented in this dissertation. However, the mentioned ideas for extensions show there is still need for further research in the future.

A. Appendix

A.1. Proof of the Conjugacy of the Dirichlet Process

Let the likelihood model be given by $\theta_i|G \overset{i.i.d.}{\sim} G,\ i = 1, \ldots, n$. As prior $G \sim DP(\alpha, G_0)$ is considered. Due to the definition of the Dirichlet process in Section 2.1 the distribution of $G(A_1), \ldots, G(A_m)$ is given by a Dirichlet distribution with parameter vector $(\alpha\, G_0(A_1), \ldots, \alpha\, G_0(A_m))^T$ for any measurable partition $\{A_1, \ldots, A_m\}$ of Θ. Equally, the posterior $G|\theta_1 \ldots, \theta_n$ can be identified by the distribution of $G(A_1), \ldots, G(A_m)|\theta_1 \ldots, \theta_n$, whose density function is deduced in the following:

$$\begin{aligned} p(G(A_1), \ldots, G(A_m)|\theta_1, \ldots, \theta_n) &\propto p(\theta_1, \ldots, \theta_n|G(A_1), \ldots, G(A_m))\, p(G(A_1), \ldots, G(A_m)) \\ &\propto \prod_{i=1}^{n} p(\theta_i|G(A_1), \ldots, G(A_m)) \prod_{j=1}^{m} G(A_j)^{\alpha\, G_0(A_j)-1} \\ &\propto \prod_{j=1}^{m} G(A_j)^{n_j} \prod_{j=1}^{m} G(A_j)^{\alpha\, G_0(A_j)-1} \\ &= \prod_{j=1}^{m} G(A_j)^{n_j+\alpha\, G_0(A_j)-1}. \end{aligned}$$

In the third line of the proof it is used that $\theta_i|G(A_1), \ldots, G(A_m)$ has a multinomial distribution with probability vector $(G(A_1), \ldots, G(A_m))^T$. Here, n_j denotes the number of elements in the set A_j. Finally, for any measurable partition $\{A_1, \ldots, A_m\}$ of Θ one obtains

$$G(A_1), \ldots, G(A_m)|\theta_1, \ldots, \theta_n \sim Dir(n_1 + \alpha\, G_0(A_1), \ldots, n_m + \alpha\, G_0(A_m)).$$

According to the definition of the Dirichlet process in Section 2.1 it follows that $G|\theta_1, \ldots, \theta_n \sim DP(\alpha^*, G_0^*)$, which proofs the conjugacy of the Dirichlet process. One gets the new updated parameters by standardization:

$$\alpha^* = \textstyle\sum_{j=1}^{m}(n_j + \alpha\, G_0(A_j)) = n + \alpha,$$

$$G_0^*(A_j) = \tfrac{1}{n+\alpha}(n_j + \alpha\, G_0(A_j)) = \tfrac{1}{n+\alpha}(\textstyle\sum_{i=1}^{n} \delta_{\theta_i}(A_j) + \alpha\, G_0(A_j)),$$

$$\Rightarrow G_0^* = \tfrac{1}{n+\alpha} \textstyle\sum_{i=1}^{n} \delta_{\theta_i} + \tfrac{\alpha}{n+\alpha} G_0.$$

A.2. Pólya Sequence via Dirichlet Process

In the following, the properties (a) and (b) in the definition of the Pólya sequence in Section 2.3 are verified based on the assumptions:

(i) $G \sim DP(\alpha, G_0)$,

(ii) $\theta_n | G \overset{i.i.d.}{\sim} G, \quad n \in \mathbb{N}$.

Proof of property (a):

$$\begin{aligned} P(\theta_1 \in A) &= \int P(\theta_1 \in A, G)\, dG \\ &= \int P(\theta_1 \in A|G)\, f(G)\, dG \\ &\overset{(ii)}{=} \int G(A)\, f(G)\, dG \\ &= E(G(A)) \\ &\overset{(i)}{=} G_0(A). \end{aligned}$$

Proof of property (b):

$$\begin{aligned} P(\theta_{n+1} \in A|\theta_1, \ldots, \theta_n) &= \int P(\theta_{n+1} \in A, G|\theta_1, \ldots, \theta_n)\, dG \\ &= \int P(\theta_{n+1} \in A|G, \theta_1, \ldots, \theta_n)\, f(G|\theta_1, \ldots, \theta_n)\, dG \\ &\overset{(ii)}{=} \int P(\theta_{n+1} \in A|G)\, f(G|\theta_1, \ldots, \theta_n)\, dG \\ &\overset{(ii)}{=} \int G(A)\, f(G|\theta_1, \ldots, \theta_n)\, dG \\ &= E(G(A)|\theta_1, \ldots, \theta_n) \\ &\overset{(i)}{=} \frac{1}{n+\alpha} \sum_{i=1}^{n} \delta_{\theta_i}(A) + \frac{\alpha}{n+\alpha} G_0(A). \end{aligned}$$

In the last step the posterior $G|\theta_1 \ldots, \theta_n$ is used, that is proved in Appendix A.1.

A.3. Derivations in the M-step

A.3.1. Derivation of $Q(\alpha, \boldsymbol{v})$

Note that $Q(\alpha, \boldsymbol{v})$ can be written as

$$Q(\alpha, \boldsymbol{v}) = \sum_{i=1}^{n}\sum_{h=1}^{N} \pi_{ih} \log\left(v_h \prod_{l<h}(1-v_l)\right) + (N-1)\log\alpha + (\alpha-1)\sum_{h=1}^{N-1}\log(1-v_h)$$

$$= \sum_{i=1}^{n}\sum_{h=1}^{N} \pi_{ih}\log v_h + \sum_{i=1}^{n}\underbrace{\sum_{h=1}^{N}\pi_{ih}\sum_{l<h}\log(1-v_l)}_{\begin{array}{l}\pi_{i1}+\\ \pi_{i2}\log(1-v_1)+\\ \pi_{i3}(\log(1-v_1)+\log(1-v_2))+\\ \ldots+\\ \pi_{iN}(\log(1-v_1)+\\ \ldots+\log(1-v_{N-1}))\end{array}} + (N-1)\log\alpha + (\alpha-1)\sum_{h=1}^{N-1}\log(1-v_h).$$

Thus, one gets the following derivation for $v_h = 1, \ldots, N-1$:

$$\begin{aligned}
\frac{\partial Q(\alpha, \boldsymbol{v})}{\partial v_h} &= \frac{1}{v_h}\sum_{i=1}^{n}\pi_{ih} - \frac{1}{1-v_h}\sum_{i=1}^{n}\sum_{l=h+1}^{N}\pi_{il} - \frac{\alpha-1}{1-v_h} \stackrel{!}{=} 0 \\
\Leftrightarrow \frac{\sum_{i=1}^{n}\pi_{ih}}{v_h} &= \frac{\alpha - 1 + \sum_{i=1}^{n}\sum_{l=h+1}^{N}\pi_{il}}{1-v_h} \\
\Leftrightarrow \sum_{i=1}^{n}\pi_{ih} - v_h\sum_{i=1}^{n}\pi_{ih} &= v_h\left(\alpha - 1 + \sum_{i=1}^{n}\sum_{l=h+1}^{N}\pi_{il}\right) \\
\Rightarrow v_h &= \frac{\sum_{i=1}^{n}\pi_{ih}}{\alpha - 1 + \sum_{i=1}^{n}\left(\pi_{ih} + \sum_{l=h+1}^{N}\pi_{il}\right)} = \frac{\sum_{i=1}^{n}\pi_{ih}}{\alpha - 1 + \sum_{i=1}^{n}\sum_{l=h}^{N}\pi_{il}}.
\end{aligned} \tag{A.1}$$

The first order condition for α is given by:

$$\begin{aligned}
\frac{\partial Q(\alpha, \boldsymbol{v})}{\partial \alpha} &= \frac{N-1}{\alpha} + \sum_{h=1}^{N-1}\log(1-v_h) \stackrel{!}{=} 0 \\
\Rightarrow \alpha &= \frac{1-N}{\sum_{h=1}^{N-1}\log(1-v_h)}.
\end{aligned}$$

Thus, the updates for $\pi_h = 1, \ldots, N$ are given by:

$$
\begin{aligned}
\pi_1 = v_1 &= \frac{\sum_{i=1}^{n} \pi_{i1}}{\alpha - 1 + \sum_{i=1}^{n} \sum_{l=1}^{N} \pi_{il}} = \frac{1}{\alpha - 1 + n} \sum_{i=1}^{n} \pi_{i1}, \\
\pi_2 = v_2(1 - \pi_1) &= \frac{\sum_{i=1}^{n} \pi_{i2}}{\alpha - 1 + \sum_{i=1}^{n} (1 - \pi_{i1})} \left(1 - \frac{\sum_{i=1}^{n} \pi_{i1}}{\alpha - 1 + n}\right) = \\
&= \frac{\sum_{i=1}^{n} \pi_{i2}}{\alpha - 1 + n - \sum_{i=1}^{n} \pi_{i1}} \frac{\alpha - 1 + n - \sum_{i=1}^{n} \pi_{i1}}{\alpha - 1 + n} = \frac{1}{\alpha - 1 + n} \sum_{i=1}^{n} \pi_{i2}, \\
\pi_3 = v_3(1 - \pi_1 - \pi_2) &= \frac{\sum_{i=1}^{n} \pi_{i3}}{\alpha - 1 + \sum_{i=1}^{n} (1 - \pi_{i1} - \pi_{i2})} \frac{\alpha - 1 + n - \sum_{i=1}^{n} \pi_{i1} - \sum_{i=1}^{n} \pi_{i2}}{\alpha - 1 + n} = \\
&= \frac{1}{\alpha - 1 + n} \sum_{i=1}^{n} \pi_{i3}, \\
&\vdots \\
\pi_N &= 1 - \sum_{h=1}^{N-1} \pi_h.
\end{aligned}
$$

In the following it is examined in which cases updates due to (A.1) yield values $v_h \neq [0, 1]$. Since the numerator of (A.1) is always positive, v_h becomes negative if the denominator is negative, i.e. if

$$
\begin{aligned}
\alpha - 1 + \sum_{i=1}^{n} \sum_{l=h}^{N} \pi_{il} &< 0 \\
\sum_{i=1}^{n} \sum_{l=h}^{N} \pi_{il} &< 1 - \alpha.
\end{aligned} \tag{A.2}
$$

Note that $v_h > 1$ if the numerator of (A.1) is higher than the denominator, i.e. if

$$
\begin{aligned}
\alpha - 1 + \sum_{i=1}^{n} \sum_{l=h}^{N} \pi_{il} &< \sum_{i=1}^{n} \pi_{ih} \\
\sum_{i=1}^{n} \left(\sum_{l=h}^{N} \pi_{il} - \pi_{ih} \right) &< 1 - \alpha \\
\sum_{i=1}^{n} \sum_{l=h+1}^{N} \pi_{il} &< 1 - \alpha.
\end{aligned} \tag{A.3}
$$

First, note that these conditions can never be fulfilled for $\alpha \geq 1$. Second, for increasing h, condition (A.2) holds if the condition (A.3) has been fulfilled in the previous step. Thus, in our correction approach we start with $h = 1$ and increase h stepwise. In each step we update v_h by (A.1) and check condition (A.3) corresponding to $v_h > 1$. Suppose that this condition is fulfilled for $h = h^*$ for the first time, i.e. the index h^* is determined by $\sum_{i=1}^{n} \sum_{l=h^*}^{N} \pi_{il} > 1 - \alpha$ and $\sum_{i=1}^{n} \sum_{l=h^*+1}^{N} \pi_{il} < 1 - \alpha$ as highlighted by grey colors in Table A.1. Then we set v_h to 1 for $h = h^*, \ldots, N - 1$.

		Cluster h						
		1	...	h^*	h^*+1	...	N	$\sum$
	1	π_{11}	...	π_{1h^*}	π_{1,h^*+1}	...	π_{1N}	1
	$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$
Subject	i	π_{i1}	...	π_{ih^*}	π_{i,h^*+1}	...	π_{iN}	1
	$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$
	n	π_{n1}	...	π_{nh^*}	π_{n,h^*+1}	...	π_{nN}	1

Table A.1.: Matrix of probabilities π_{ih}. The index h^* is determined by $\sum_{i=1}^{n} \sum_{l=h^*}^{N} \pi_{il} > 1 - \alpha$ and $\sum_{i=1}^{n} \sum_{l=h^*+1}^{N} \pi_{il} < 1 - \alpha$.

Thus, one obtains the following weights:

$$\pi_h = \begin{cases} \frac{1}{n+\alpha-1} \sum_{i=1}^{n} \pi_{ih}, & \text{for } h < h^*, \\ 1 - \sum_{l=1}^{h-1} \pi_l & \text{for } h = h^*, \\ 0 & \text{for } h > h^*. \end{cases}$$

A.3.2. Derivation of $Q(\boldsymbol{\psi})$

In the following formulas different font colors correspond to the different approaches in the Chapters 3, 4 and 6. So black colored formulas describe the function $Q(\boldsymbol{\psi})$ and the particular deviations for the linear mixed models with DPMs from Chapter 4. For the additive mixed models with DPMs from Chapter 6 the red-colored term has to be added while the blue-colored term appears in the case of the linear mixed models based on the group fused lasso penalty from Chapter 3.

$$
\begin{aligned}
Q(\boldsymbol{\psi}) = &\sum_{i=1}^{n}\sum_{h=1}^{N} \pi_{ih} \log\left((2\pi)^{-\frac{n_i}{2}}|\boldsymbol{V}_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 - \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p - \right.\right.\\
&- \left.\left.\boldsymbol{Z}_i\boldsymbol{\mu}_h)^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 - \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p - \boldsymbol{Z}_i\boldsymbol{\mu}_h)\right)\right) -\\
&- \frac{1}{2}\left((d-k)\log(2\pi\tau^2) + \frac{1}{\tau^2}\boldsymbol{\gamma}_p^T\boldsymbol{\gamma}_p\right) - \lambda\sqrt{N\cdot q}\sum_{h<l}\|\boldsymbol{\mu}_h - \boldsymbol{\mu}_l\|\\
= &-\frac{1}{2}\left(\sum_{i=1}^{n}\sum_{h=1}^{N}\pi_{ih}\left(n_i\log(2\pi) + \log|\boldsymbol{V}_i| + (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 - \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p - \right.\right.\\
&- \left.\boldsymbol{Z}_i\boldsymbol{\mu}_h)^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 - \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p - \boldsymbol{Z}_i\boldsymbol{\mu}_h)\right) +\\
&+ \left.\left.(d-k)\log(2\pi) + (d-k)\log\tau^2 + \frac{1}{\tau^2}\boldsymbol{\gamma}_p^T\boldsymbol{\gamma}_p\right)\right) - \lambda\sqrt{N\cdot q}\sum_{h<l}\|\boldsymbol{\mu}_h - \boldsymbol{\mu}_l\|.
\end{aligned}
$$

The first order condition for $\boldsymbol{\beta}$ is given by

$$
\frac{\partial Q(\boldsymbol{\psi})}{\partial\boldsymbol{\beta}} = -\frac{1}{2}\sum_{i=1}^{n}\sum_{h=1}^{N} -2\pi_{ih}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 - \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p - \boldsymbol{Z}_i\boldsymbol{\mu}_h) \stackrel{!}{=} 0
$$

$$
\Leftrightarrow \sum_{i=1}^{n}\sum_{h=1}^{N}\pi_{ih}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 - \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p - \boldsymbol{Z}_i\boldsymbol{\mu}_h) = \sum_{i=1}^{n}\sum_{h=1}^{N}\pi_{ih}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\boldsymbol{\beta}
$$

$$
\Rightarrow \boldsymbol{\beta} = \left(\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\left(\boldsymbol{y}_i - \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 - \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p - \sum_{h=1}^{N}\pi_{ih}\boldsymbol{Z}_i\boldsymbol{\mu}_h\right)\right).
$$

For the linear and additive mixed models with DPMs from Chapter 4 and Chapter 6 the derivations of $Q(\boldsymbol{\psi})$ with respect to $\boldsymbol{\mu}_h$, $h = 1, \ldots, N$, are given by

$$
\frac{\partial Q(\boldsymbol{\psi})}{\partial\boldsymbol{\mu}_h} = -\frac{1}{2}\sum_{i=1}^{n} -2\pi_{ih}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 - \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p - \boldsymbol{Z}_i\boldsymbol{\mu}_h) \stackrel{!}{=} 0
$$

$$
\Leftrightarrow \sum_{i=1}^{n}\pi_{ih}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 - \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p) = \sum_{i=1}^{n}\pi_{ih}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i\boldsymbol{\mu}_h
$$

$$
\Rightarrow \boldsymbol{\mu}_h = \left(\sum_{i=1}^{n}\pi_{ih}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\pi_{ih}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 - \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p\right)\right).
$$

In the case of additive mixed models with DPMs further derivations for the additional parameters $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_p$ and τ^2 are necessary. As from now we abstain from coloration.

First, for the derivation referred to $\boldsymbol{\gamma}_0$ one gets

$$\frac{\partial Q(\boldsymbol{\psi})}{\partial \boldsymbol{\gamma}_0} = -\frac{1}{2}\sum_{i=1}^{n}\sum_{h=1}^{N} -2\pi_{ih}\boldsymbol{T}^T\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 - \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p - \boldsymbol{Z}_i\boldsymbol{\mu}_h) \stackrel{!}{=} 0$$

$$\Leftrightarrow \sum_{i=1}^{n}\sum_{h=1}^{N}\pi_{ih}\boldsymbol{T}^T\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p - \boldsymbol{Z}_i\boldsymbol{\mu}_h) = \sum_{i=1}^{n}\sum_{h=1}^{N}\pi_{ih}\boldsymbol{T}^T\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0$$

$$\begin{aligned}\Rightarrow \boldsymbol{\gamma}_0 \;=\; & \left(\sum_{i=1}^{n}\boldsymbol{T}^T\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{B}_i\boldsymbol{T}\right)^{-1} \\ & \left(\sum_{i=1}^{n}\boldsymbol{T}^T\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p - \sum_{h=1}^{N}\pi_{ih}\boldsymbol{Z}_i\boldsymbol{\mu}_h\right)\right).\end{aligned}$$

For the penalized spline coefficients the first order condition is given by

$$\frac{\partial Q(\boldsymbol{\psi})}{\partial \boldsymbol{\gamma}_p} = -\frac{1}{2}\left(\sum_{i=1}^{n}\sum_{h=1}^{N} -2\pi_{ih}\boldsymbol{W}^T\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 - \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p - \boldsymbol{Z}_i\boldsymbol{\mu}_h) + \right.$$

$$\left. + \frac{2\boldsymbol{\gamma}_p}{\tau^2}\right) \stackrel{!}{=} 0 \qquad \Leftrightarrow \qquad \sum_{i=1}^{n}\sum_{h=1}^{N}\pi_{ih}\boldsymbol{W}^T\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 - \boldsymbol{Z}_i\boldsymbol{\mu}_h) =$$

$$= \sum_{i=1}^{n}\sum_{h=1}^{N}\pi_{ih}\boldsymbol{W}^T\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p + \frac{\boldsymbol{\gamma}_p}{\tau^2}$$

$$\begin{aligned}\Rightarrow \boldsymbol{\gamma}_p \;=\; & \left(\sum_{i=1}^{n}\boldsymbol{W}^T\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{B}_i\boldsymbol{W} + \frac{1}{\tau^2}\boldsymbol{I}_{d-k}\right)^{-1} \\ & \left(\sum_{i=1}^{n}\boldsymbol{W}^T\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 - \sum_{h=1}^{N}\pi_{ih}\boldsymbol{Z}_i\boldsymbol{\mu}_h\right)\right).\end{aligned}$$

Finally, one obtains the following derivation for τ^2:

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\psi})}{\partial \tau^2} &= -\frac{1}{2}\left(\frac{d-k}{\tau^2} - \frac{1}{\tau^4}\boldsymbol{\gamma}_p^T\boldsymbol{\gamma}_p\right) \stackrel{!}{=} 0 \\ \Leftrightarrow \frac{d-k}{\tau^2} &= \frac{1}{\tau^4}\boldsymbol{\gamma}_p^T\boldsymbol{\gamma}_p \\ \Rightarrow \tau^2 &= \frac{1}{d-k}\boldsymbol{\gamma}_p^T\boldsymbol{\gamma}_p.\end{aligned}$$

A.4. Prediction of Random Effects

Proposition:

$$E(\boldsymbol{b}_i|\boldsymbol{y}_i) = \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) + (\boldsymbol{I}_q - \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{Z}_i)\sum_{h=1}^{N}\hat{\pi}_{ih}\hat{\boldsymbol{\mu}}_h.$$

Proof:
According to (4) – (8) in Lindley and Smith (1972) it follows from

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\theta}_1 &\sim N(\boldsymbol{A}_1\boldsymbol{\theta}_1, \boldsymbol{C}_1), \\ \boldsymbol{\theta}_1 &\sim N(\boldsymbol{A}_2\boldsymbol{\theta}_2, \boldsymbol{C}_2), \end{aligned}$$

that

$$E(\boldsymbol{\theta}_1|\boldsymbol{y}) = (\boldsymbol{C}_2^{-1} + \boldsymbol{A}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{A}_1)^{-1}(\boldsymbol{A}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{y} + \boldsymbol{C}_2^{-1}\boldsymbol{A}_2\boldsymbol{\theta}_2),$$

holds. By defining

$$\begin{aligned} &\boldsymbol{\theta}_1 := \boldsymbol{b}_i, \quad \boldsymbol{A}_1 := \boldsymbol{Z}_i, \quad \boldsymbol{C}_1 := \hat{\boldsymbol{\Sigma}}_i = \hat{\sigma}^2\boldsymbol{I}_{n_i}, \quad \boldsymbol{y} := \boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}, \\ &\boldsymbol{\theta}_2 := \hat{\boldsymbol{\mu}}_h, \quad \boldsymbol{A}_2 := \boldsymbol{I}_q, \quad \boldsymbol{C}_2 := \hat{\boldsymbol{D}}, \end{aligned}$$

and by assuming that individual i belongs to cluster h one obtains

$$\begin{aligned} E(\boldsymbol{b}_i|\boldsymbol{y}_i) &= (\hat{\boldsymbol{D}}^{-1} + \boldsymbol{Z}_i^T\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{Z}_i)^{-1}(\boldsymbol{Z}_i^T\hat{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) + \hat{\boldsymbol{D}}^{-1}\hat{\boldsymbol{\mu}}_h) \\ &\overset{(*)}{=} (\hat{\boldsymbol{D}} - \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T(\underbrace{\boldsymbol{Z}_i\hat{\boldsymbol{D}}\boldsymbol{Z}_i^T + \hat{\boldsymbol{\Sigma}}_i}_{\hat{\boldsymbol{V}}_i})^{-1}\boldsymbol{Z}_i\hat{\boldsymbol{D}})(\boldsymbol{Z}_i^T\hat{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) + \hat{\boldsymbol{D}}^{-1}\hat{\boldsymbol{\mu}}_h) \\ &= \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{Z}_i\hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) + \\ &\quad + \hat{\boldsymbol{D}}\hat{\boldsymbol{D}}^{-1}\hat{\boldsymbol{\mu}}_h - \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{Z}_i\hat{\boldsymbol{D}}\hat{\boldsymbol{D}}^{-1}\hat{\boldsymbol{\mu}}_h \\ &= \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T(\boldsymbol{I}_{n_i} - \hat{\boldsymbol{V}}_i^{-1}\boldsymbol{Z}_i\hat{\boldsymbol{D}}\boldsymbol{Z}_i^T)\hat{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) + (\boldsymbol{I}_q - \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{Z}_i)\hat{\boldsymbol{\mu}}_h \\ &= \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T(\hat{\boldsymbol{V}}_i^{-1}\hat{\boldsymbol{V}}_i - \hat{\boldsymbol{V}}_i^{-1}\boldsymbol{Z}_i\hat{\boldsymbol{D}}\boldsymbol{Z}_i^T)\hat{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) + (\boldsymbol{I}_q - \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{Z}_i)\hat{\boldsymbol{\mu}}_h \\ &= \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{Z}_i\hat{\boldsymbol{D}}\boldsymbol{Z}_i^T + \hat{\boldsymbol{\Sigma}}_i - \boldsymbol{Z}_i\hat{\boldsymbol{D}}\boldsymbol{Z}_i^T)\hat{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) + (\boldsymbol{I}_q - \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{Z}_i)\hat{\boldsymbol{\mu}}_h \\ &= \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) + (\boldsymbol{I}_q - \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{Z}_i)\hat{\boldsymbol{\mu}}_h. \end{aligned}$$

Note that in $(*)$ the matrix lemma (10) in Lindley and Smith (1972) with $\boldsymbol{A}_1 := \boldsymbol{Z}_i^T$, $\boldsymbol{C}_1 := \hat{\boldsymbol{D}}^{-1}$ and $\boldsymbol{C}_2 := \hat{\boldsymbol{\Sigma}}_i^{-1}$ is used.
Thus, without knowing the cluster membership one obtains

$$E(\boldsymbol{b}_i|\boldsymbol{y}_i) = \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) + (\boldsymbol{I}_q - \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{Z}_i)\sum_{h=1}^{N}\hat{\pi}_{ih}\hat{\boldsymbol{\mu}}_h.$$

A.5. Standardization

In general, it is reasonable to standardize the variables before the calculations so that the results are independent from the particular scales. This raises the question how to transform the estimated parameters $\tilde{\boldsymbol{\xi}}$ for the standardized data to parameters $\hat{\boldsymbol{\xi}}$ corresponding to the original data. These transformations are given in the following. Note that in contrast to the notation in the Chapters 3 and 4 in this section the parameter vector $\boldsymbol{\beta}$ does not include the intercept β_0.

For the linear mixed models in the Chapters 3 and 4 the following transformations are used:

$$\begin{array}{rcll} \hat{\boldsymbol{\beta}} & = & \boldsymbol{R}\tilde{\boldsymbol{\beta}}, & \\ \hat{\boldsymbol{b}}_i & = & \boldsymbol{S}\tilde{\boldsymbol{b}}_i, & i = 1, \ldots, n, \\ \hat{\boldsymbol{\mu}}_h & = & \boldsymbol{S}\tilde{\boldsymbol{\mu}}_h, & h = 1, \ldots, N, \\ \hat{\boldsymbol{D}} & = & \boldsymbol{S}\tilde{\boldsymbol{D}}\boldsymbol{S}^T, & \\ \hat{\sigma}^2 & = & s_y^2\tilde{\sigma}^2, & \end{array} \tag{A.4}$$

with

$$\boldsymbol{R} = \begin{pmatrix} \frac{s_y}{s_{x_1}} & 0 & \ldots & 0 \\ 0 & \frac{s_y}{s_{x_2}} & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{s_y}{s_{x_p}} \end{pmatrix}, \qquad \boldsymbol{S} = \begin{pmatrix} s_y & -\bar{t}\frac{s_y}{s_t} & \ldots & -\bar{t^q}\frac{s_y}{s_{t^q}} \\ 0 & \frac{s_y}{s_t} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{s_y}{s_{t^q}} \end{pmatrix}.$$

The intercept is transformed by

$$\hat{\beta}_0 = \tilde{\beta}_0 s_y + \bar{y} - \bar{x}_1 \frac{s_y}{s_{x_1}}\tilde{\beta}_1 - \ldots - \bar{x}_p \frac{s_y}{s_{x_p}}\tilde{\beta}_p.$$

Here, $\bar{x}_1$ and s_{x_1}, for example, denote the mean respectively the standard deviation of the covariate x_1. As illustration of the transformations for the fixed and the random effects we consider the observation y_{ij} for the special case $p = 2$ and $q = 2$. For a clearer notation the indices i and j are omitted:

$$\left(\frac{y - \bar{y}}{s_y}\right) = \tilde{\beta}_0 + \left(\frac{x_1 - \bar{x}_1}{s_{x_1}}\right)\tilde{\beta}_1 + \left(\frac{x_2 - \bar{x}_2}{s_{x_2}}\right)\tilde{\beta}_2 + \tilde{b}_{i0} + \left(\frac{t - \bar{t}}{s_t}\right)\tilde{b}_{i1} + \left(\frac{t^2 - \bar{t^2}}{s_{t^2}}\right)\tilde{b}_{i2}$$

$$y = \underbrace{\tilde{\beta}_0 s_y + \bar{y} - \bar{x}_1 \frac{s_y}{s_{x_1}}\tilde{\beta}_1 - \bar{x}_2 \frac{s_y}{s_{x_2}}\tilde{\beta}_2}_{\hat{\beta}_0} + x_1 \underbrace{\frac{s_y}{s_{x_1}}\tilde{\beta}_1}_{\hat{\beta}_1} + x_2 \underbrace{\frac{s_y}{s_{x_2}}\tilde{\beta}_2}_{\hat{\beta}_2} +$$

$$+ \underbrace{\tilde{b}_{i0} s_y - \bar{t}\frac{s_y}{s_t}\tilde{b}_{i1} - \bar{t^2}\frac{s_y}{s_{t^2}}\tilde{b}_{i2}}_{\hat{b}_{i0}} + t \underbrace{\frac{s_y}{s_t}\tilde{b}_{i1}}_{\hat{b}_{i1}} + t^2 \underbrace{\frac{s_y}{s_{t^2}}\tilde{b}_{i2}}_{\hat{b}_{i2}}.$$

For the means $\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_N$ of the random effects the same transformations are used as for the random effects itself. The transformations of the variance parameters follow from usual calculation rules.

For the additive mixed model in Chapter 6 we get the same transformations as in (A.4), but instead of the intercept β_0 the spline coefficients have to be updated:

$$\hat{\gamma}_j = \tilde{\gamma}_j s_y + \bar{y} - \bar{x}_1 \frac{s_y}{s_{x_1}} \tilde{\beta}_1 - \ldots - \bar{x}_p \frac{s_y}{s_{x_p}} \tilde{\beta}_p, \quad j = 1, \ldots, d.$$

Again, this is visualized for the special case $p = 2$ and $q = 2$:

$$\begin{aligned}
\left(\frac{y - \bar{y}}{s_y}\right) &= \sum_{j=1}^{d} B_j \left(\frac{t - \bar{t}}{s_t}\right) \tilde{\gamma}_j + \left(\frac{x_1 - \bar{x}_1}{s_{x_1}}\right) \tilde{\beta}_1 + \left(\frac{x_2 - \bar{x}_2}{s_{x_2}}\right) \tilde{\beta}_2 + \\
&\quad + \tilde{b}_{i0} + \left(\frac{t - \bar{t}}{s_t}\right) \tilde{b}_{i1} + \left(\frac{t^2 - \bar{t^2}}{s_{t^2}}\right) \tilde{b}_{i2} \\
y &= \sum_{j=1}^{d} B_j(t) \underbrace{\left(\tilde{\gamma}_j s_y + \bar{y} - \bar{x}_1 \frac{s_y}{s_{x_1}} \tilde{\beta}_1 - \bar{x}_2 \frac{s_y}{s_{x_2}} \tilde{\beta}_2\right)}_{\hat{\gamma}_j} + x_1 \underbrace{\frac{s_y}{s_{x_1}} \tilde{\beta}_1}_{\hat{\beta}_1} + x_2 \underbrace{\frac{s_y}{s_{x_2}} \tilde{\beta}_2}_{\hat{\beta}_2} + \\
&\quad + \underbrace{\tilde{b}_{i0} s_y - \bar{t} \frac{s_y}{s_t} \tilde{b}_{i1} - \bar{t^2} \frac{s_y}{s_{t^2}} \tilde{b}_{i2}}_{\hat{b}_{i0}} + t \underbrace{\frac{s_y}{s_t} \tilde{b}_{i1}}_{\hat{b}_{i1}} + t^2 \underbrace{\frac{s_y}{s_{t^2}} \tilde{b}_{i2}}_{\hat{b}_{i2}}.
\end{aligned}$$

Here, it is used that $B_j\left(\frac{t-\bar{t}}{s_t}\right) = B_j(t)$ is fulfilled if the number and the positioning of knots are the same for the original variable t and the standardized one.

A.6. Predictive Cross-Validation

In what follows, the notation of Section 3.2.2 is used. In analogue to Braun et al. (2012), we consider the joint distribution of $\boldsymbol{z}_{ij}^T \boldsymbol{b}_i$ and $\boldsymbol{y}_{i,-j} = \boldsymbol{X}_{i,-j}\boldsymbol{\beta} + \boldsymbol{Z}_{i,-j}\boldsymbol{b}_i + \boldsymbol{\varepsilon}_{i,-j}$. If it is known that subject i belongs to cluster h, i.e. $w_{ih} = 1$, one obtains $\boldsymbol{b}_i | w_{ih} = 1 \sim N(\boldsymbol{\mu}_h, \boldsymbol{D})$. In addition, a multivariate normal distribution for the error variable is assumed: $\boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{n_i})$. Thus, the joint distribution of $\boldsymbol{z}_{ij}^T \boldsymbol{b}_i$ and $\boldsymbol{y}_{i,-j}$ under the condition $w_{ih} = 1$ is also a multivariate normal distribution, in which the mean is composed by

$$\begin{aligned}
&\mathbb{E}(\boldsymbol{z}_{ij}^T \boldsymbol{b}_i | w_{ih} = 1) = \boldsymbol{z}_{ij}^T \mathbb{E}(\boldsymbol{b}_i | w_{ih} = 1) = \boldsymbol{z}_{ij}^T \boldsymbol{\mu}_h, \\
&\mathbb{E}(\boldsymbol{y}_{i,-j} | w_{ih} = 1) = \mathbb{E}(\boldsymbol{X}_{i,-j}\boldsymbol{\beta} + \boldsymbol{Z}_{i,-j}\boldsymbol{b}_i + \boldsymbol{\varepsilon}_{i,-j} | w_{ih} = 1) = \boldsymbol{X}_{i,-j}\boldsymbol{\beta} + \boldsymbol{Z}_{i,-j}\boldsymbol{\mu}_h.
\end{aligned}$$

Due to the mutual independence of $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_{i,-j}$ the covariances are given by

$$\operatorname{Cov}(\boldsymbol{z}_{ij}^T\boldsymbol{b}_i, \boldsymbol{z}_{ij}^T\boldsymbol{b}_i|w_{ih}=1) = \operatorname{Cov}(\boldsymbol{z}_{ij}^T\boldsymbol{b}_i|w_{ih}=1) = \boldsymbol{z}_{ij}^T\operatorname{Cov}(\boldsymbol{b}_i|w_{ih}=1)\boldsymbol{z}_{ij} = \boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ij},$$

$$\begin{aligned}&\operatorname{Cov}(\boldsymbol{y}_{i,-j}, \boldsymbol{y}_{i,-j}|w_{ih}=1)\\ &\quad= \operatorname{Cov}(\boldsymbol{X}_{i,-j}\boldsymbol{\beta} + \boldsymbol{Z}_{i,-j}\boldsymbol{b}_i + \boldsymbol{\varepsilon}_{i,-j}, \boldsymbol{X}_{i,-j}\boldsymbol{\beta} + \boldsymbol{Z}_{i,-j}\boldsymbol{b}_i + \boldsymbol{\varepsilon}_{i,-j}|w_{ih}=1)\\ &\quad= \operatorname{Cov}(\boldsymbol{\varepsilon}_{i,-j}) + \operatorname{Cov}(\boldsymbol{Z}_{i,-j}\boldsymbol{b}_i|w_{ih}=1) = \sigma^2\boldsymbol{I}_{n_i-1} + \boldsymbol{Z}_{i,-j}\boldsymbol{D}\boldsymbol{Z}_{i,-j}^T,\end{aligned}$$

$$\begin{aligned}\operatorname{Cov}(\boldsymbol{z}_{ij}^T\boldsymbol{b}_i, \boldsymbol{y}_{i,-j}|w_{ih}=1) &= \operatorname{Cov}(\boldsymbol{z}_{ij}^T\boldsymbol{b}_i, \boldsymbol{X}_{i,-j}\boldsymbol{\beta} + \boldsymbol{Z}_{i,-j}\boldsymbol{b}_i + \boldsymbol{\varepsilon}_{i,-j}|w_{ih}=1)\\ &= \operatorname{Cov}(\boldsymbol{z}_{ij}^T\boldsymbol{b}_i, \boldsymbol{Z}_{i,-j}\boldsymbol{b}_i|w_{ih}=1) = \boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{Z}_{i,-j}^T,\end{aligned}$$

$$\begin{aligned}\operatorname{Cov}(\boldsymbol{y}_{i,-j}, \boldsymbol{z}_{ij}^T\boldsymbol{b}_i|w_{ih}=1) &= \operatorname{Cov}(\boldsymbol{X}_{i,-j}\boldsymbol{\beta} + \boldsymbol{Z}_{i,-j}\boldsymbol{b}_i + \boldsymbol{\varepsilon}_{i,-j}, \boldsymbol{z}_{ij}^T\boldsymbol{b}_i|w_{ih}=1)\\ &= \operatorname{Cov}(\boldsymbol{Z}_{i,-j}\boldsymbol{b}_i, \boldsymbol{z}_{ij}^T\boldsymbol{b}_i|w_{ih}=1) = \boldsymbol{Z}_{i,-j}\boldsymbol{D}\boldsymbol{z}_{ij}.\end{aligned}$$

In summary, the joint distribution of $\boldsymbol{z}_{ij}^T\boldsymbol{b}_i$ and $\boldsymbol{y}_{i,-j}$ conditional on $w_{ih}=1$ is given by

$$\left.\begin{pmatrix}\boldsymbol{z}_{ij}^T\boldsymbol{b}_i\\ \boldsymbol{y}_{i,-j}\end{pmatrix}\right| w_{ih}=1 \sim$$

$$\sim N\left(\begin{pmatrix}\boldsymbol{z}_{ij}^T\boldsymbol{\mu}_h\\ \boldsymbol{X}_{i,-j}\boldsymbol{\beta} + \boldsymbol{Z}_{i,-j}\boldsymbol{\mu}_h\end{pmatrix}, \begin{pmatrix}\boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ij} & \boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{Z}_{i,-j}^T\\ \boldsymbol{Z}_{i,-j}\boldsymbol{D}\boldsymbol{z}_{ij} & \sigma^2\boldsymbol{I}_{n_i-1} + \boldsymbol{Z}_{i,-j}\boldsymbol{D}\boldsymbol{Z}_{i,-j}^T\end{pmatrix}\right).$$

Using standard properties of the multivariate normal distribution, the conditional distribution of $\boldsymbol{z}_{ij}^T\boldsymbol{b}_i|\boldsymbol{y}_{i,-j}, w_{ih}=1$ is a normal distribution with the moments

$$\begin{aligned}\mathbb{E}(\boldsymbol{z}_{ij}^T\boldsymbol{b}_i|\boldsymbol{y}_{i,-j}, w_{ih}=1) &= \boldsymbol{z}_{ij}^T\boldsymbol{\mu}_h + \boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{Z}_{i,-j}^T\left(\sigma^2\boldsymbol{I}_{n_i-1} + \boldsymbol{Z}_{i,-j}\boldsymbol{D}\boldsymbol{Z}_{i,-j}^T\right)^{-1}\\ &\quad\cdot(\boldsymbol{y}_{i,-j} - \boldsymbol{X}_{i,-j}\boldsymbol{\beta} - \boldsymbol{Z}_{i,-j}\boldsymbol{\mu}_h),\\ \operatorname{Var}(\boldsymbol{z}_{ij}^T\boldsymbol{b}_i|\boldsymbol{y}_{i,-j}, w_{ih}=1) &= \boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ij} - \boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{Z}_{i,-j}^T\left(\sigma^2\boldsymbol{I}_{n_i-1} + \boldsymbol{Z}_{i,-j}\boldsymbol{D}\boldsymbol{Z}_{i,-j}^T\right)^{-1}\boldsymbol{Z}_{i,-j}\boldsymbol{D}\boldsymbol{z}_{ij}.\end{aligned}$$

Thus, the moments of the predictive distribution $y_{ij}|\boldsymbol{y}_{i,-j}, w_{ih}=1$ with $y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i + \boldsymbol{\varepsilon}_{ij}$ are given by

$$\begin{aligned}\mathbb{E}(y_{ij}|\boldsymbol{y}_{i,-j}, w_{ih}=1) &= \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{\mu}_h + \boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{Z}_{i,-j}^T\left(\sigma^2\boldsymbol{I}_{n_i-1} + \boldsymbol{Z}_{i,-j}\boldsymbol{D}\boldsymbol{Z}_{i,-j}^T\right)^{-1}\\ &\quad\cdot(\boldsymbol{y}_{i,-j} - \boldsymbol{X}_{i,-j}\boldsymbol{\beta} - \boldsymbol{Z}_{i,-j}\boldsymbol{\mu}_h),\\ \operatorname{Var}(y_{ij}|\boldsymbol{y}_{i,-j}, w_{ih}=1) &= \boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ij} - \boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{Z}_{i,-j}^T\left(\sigma^2\boldsymbol{I}_{n_i-1} + \boldsymbol{Z}_{i,-j}\boldsymbol{D}\boldsymbol{Z}_{i,-j}^T\right)^{-1}\boldsymbol{Z}_{i,-j}\boldsymbol{D}\boldsymbol{z}_{ij} + \sigma^2.\end{aligned}$$

A.7. Proof of Full Conditionals

In the following all full conditionals for the block Gibbs sampler in Section 5.2.3 are derived. Large parts of these derivations can also be found in Heinzl (2009). In general, in the framework of a block Gibbs sampler an unknown parameter vector $\boldsymbol{\xi}$ is partitioned into $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_S)^T$. For updating $\boldsymbol{\xi}_s$, $s = 1, \ldots, S$, in the MCMC algorithm as proposal density the full conditional density is used − conditional on all the other parameters $\boldsymbol{\xi}_{-s} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{s-1}, \boldsymbol{\xi}_{s+1}, \ldots, \boldsymbol{\xi}_S)^T$. Each full conditional is proportional to the posterior distribution:

$$p(\boldsymbol{\xi}_s|\boldsymbol{\xi}_{-s}, \boldsymbol{y}) = \frac{p(\boldsymbol{\xi}|\boldsymbol{y})}{p(\boldsymbol{\xi}_{-s}|\boldsymbol{y})} \propto p(\boldsymbol{\xi}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\xi})\, p(\boldsymbol{\xi}).$$

However, for updating $\boldsymbol{\xi}_s$ only the $\boldsymbol{\xi}_s$ including terms in the product of likelihood and prior are necessary. For the additive mixed model with DPM prior in Chapter 5 the likelihood is given by

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}, \sigma^2) &= \prod_{i=1}^{n} f(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}_i, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{0.5n_i}} \exp\left(-\frac{1}{2\sigma^2}\tilde{\boldsymbol{y}}_i^T\tilde{\boldsymbol{y}}_i\right) \\ &\propto \frac{1}{(\sigma^2)^{0.5n_d}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\tilde{\boldsymbol{y}}_i^T\tilde{\boldsymbol{y}}_i\right), \end{aligned}$$

with $\tilde{\boldsymbol{y}}_i := \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{\gamma} + \boldsymbol{Z}_i\boldsymbol{b}_i$. Here, the number of all measurements is given by n_d. For a clearer notation in the following $\tilde{\boldsymbol{y}}_i$ denotes the relative working response as given in Section 5.2.3.

Error Variance

Prior:

$$p(\sigma^2) \propto \frac{1}{(\sigma^2)^{a_\varepsilon+1}} \exp\left(-\frac{b_\varepsilon}{\sigma^2}\right).$$

Full conditional:

$$\begin{aligned} p(\sigma^2|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}, \boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}, \sigma^2)\; p(\sigma^2) \\ &\propto \frac{1}{(\sigma^2)^{0.5n_d}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\tilde{\boldsymbol{y}}_i^T\tilde{\boldsymbol{y}}_i\right) \frac{1}{(\sigma^2)^{a_\varepsilon+1}} \exp\left(-\frac{b_\varepsilon}{\sigma^2}\right) \\ &= \frac{1}{(\sigma^2)^{a_\varepsilon+0.5n_d+1}} \exp\left(-\frac{1}{\sigma^2}\left(b_\varepsilon + \frac{1}{2}\sum_{i=1}^{n}\tilde{\boldsymbol{y}}_i^T\tilde{\boldsymbol{y}}_i\right)\right). \end{aligned}$$

Parameters of the P-spline

Priors:

$$\begin{aligned} p(\boldsymbol{\gamma}|\tau^2) &\propto \frac{1}{(\tau^2)^{0.5\,(d-k)}} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}^T\boldsymbol{K}\boldsymbol{\gamma}\right), \\ p(\tau^2) &\propto \frac{1}{(\tau^2)^{a_\gamma+1}} \exp\left(-\frac{b_\gamma}{\tau^2}\right). \end{aligned}$$

Full conditionals:

$$\begin{aligned} p(\boldsymbol{\gamma}|\tau^2,\boldsymbol{\beta},\boldsymbol{b},\boldsymbol{y},\sigma^2) &\propto p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{b},\sigma^2)\; p(\boldsymbol{\gamma}|\tau^2) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\tilde{\boldsymbol{y}}_i-\boldsymbol{B}_i\boldsymbol{\gamma})^T(\tilde{\boldsymbol{y}}_i-\boldsymbol{B}_i\boldsymbol{\gamma})\right)\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}^T\boldsymbol{K}\boldsymbol{\gamma}\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\boldsymbol{\gamma}^T\boldsymbol{B}_i^T\boldsymbol{B}_i\boldsymbol{\gamma}-2\boldsymbol{\gamma}^T\boldsymbol{B}_i^T\tilde{\boldsymbol{y}}_i)\right)\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}^T\boldsymbol{K}\boldsymbol{\gamma}\right) \\ &= \exp\left(-\frac{1}{2}\left(\boldsymbol{\gamma}^T\left(\frac{1}{\tau^2}\boldsymbol{K}+\frac{1}{\sigma^2}\sum_{i=1}^{n}\boldsymbol{B}_i^T\boldsymbol{B}_i\right)\boldsymbol{\gamma}-2\boldsymbol{\gamma}^T\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}\boldsymbol{B}_i^T\tilde{\boldsymbol{y}}_i\right)\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\boldsymbol{\gamma}^T\underbrace{\left(\frac{1}{\tau^2}\boldsymbol{K}+\frac{1}{\sigma^2}\boldsymbol{B}^T\boldsymbol{B}\right)}_{=:\boldsymbol{\Sigma}_\gamma^{*\,-1}}\boldsymbol{\gamma}-2\boldsymbol{\gamma}^T\left(\frac{1}{\sigma^2}\boldsymbol{B}^T\tilde{\boldsymbol{y}}\right)\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\boldsymbol{\gamma}^T\boldsymbol{\Sigma}_\gamma^{*\,-1}\boldsymbol{\gamma}-2\boldsymbol{\gamma}^T\boldsymbol{\Sigma}_\gamma^{*\,-1}\underbrace{\boldsymbol{\Sigma}_\gamma^*\left(\frac{1}{\sigma^2}\boldsymbol{B}^T\tilde{\boldsymbol{y}}\right)}_{=:\boldsymbol{\mu}_\gamma^*}\right)\right). \end{aligned}$$

$$\begin{aligned} p(\tau^2|\boldsymbol{\gamma}) &\propto p(\boldsymbol{\gamma}|\tau^2)\; p(\tau^2) \\ &\propto \frac{1}{(\tau^2)^{0.5\,(d-k)}}\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}^T\boldsymbol{K}\boldsymbol{\gamma}\right)\frac{1}{(\tau^2)^{a_\gamma+1}}\exp\left(-\frac{b_\gamma}{\tau^2}\right) \\ &= \frac{1}{(\tau^2)^{a_\gamma+0.5\,(d-k)+1}}\exp\left(-\frac{1}{\tau^2}(b_\gamma+0.5\boldsymbol{\gamma}^T\boldsymbol{K}\boldsymbol{\gamma})\right). \end{aligned}$$

Parameters of the Fixed Effects Part

Priors:

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) &\propto |\boldsymbol{\Sigma}_\beta|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)^T \boldsymbol{\Sigma}_\beta^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)\right), \\ p(\sigma^2_{\beta_r}) &\propto \frac{1}{(\sigma^2_{\beta_r})^{a_\beta+1}} \exp\left(-\frac{b_\beta}{\sigma^2_{\beta_r}}\right), \\ p(\mu_{\beta r}) &\propto \exp\left(-\frac{1}{2s^2_{\beta_r}}(\mu_{\beta r}-m_{\beta r})^2\right). \end{aligned}$$

Full conditionals:

$$\begin{aligned} &p(\boldsymbol{\beta}|\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta, \boldsymbol{\gamma}, \boldsymbol{b}, \boldsymbol{y}, \sigma^2) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}, \sigma^2)\; p(\boldsymbol{\beta}|\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \\ &\quad\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\tilde{\boldsymbol{y}}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T(\tilde{\boldsymbol{y}}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right) \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)^T\boldsymbol{\Sigma}_\beta^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)\right) \\ &\quad\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\boldsymbol{\beta}^T\boldsymbol{X}_i^T\boldsymbol{X}_i\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{X}_i^T\tilde{\boldsymbol{y}}_i)\right) \exp\left(-\frac{1}{2}(\boldsymbol{\beta}^T\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta)\right) \\ &\quad= \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}^T\left(\boldsymbol{\Sigma}_\beta^{-1} + \frac{1}{\sigma^2}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{X}_i\right)\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \frac{1}{\sigma^2}\sum_{i=1}^{n}\boldsymbol{X}_i^T\tilde{\boldsymbol{y}}_i\right)\right)\right) \\ &\quad= \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}^T\underbrace{\left(\boldsymbol{\Sigma}_\beta^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}\right)}_{=:\boldsymbol{\Sigma}_\beta^{*\,-1}}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \frac{1}{\sigma^2}\boldsymbol{X}^T\tilde{\boldsymbol{y}}\right)\right)\right) \\ &\quad= \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}^T\boldsymbol{\Sigma}_\beta^{*\,-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{\Sigma}_\beta^{*\,-1}\underbrace{\boldsymbol{\Sigma}_\beta^*\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \frac{1}{\sigma^2}\boldsymbol{X}^T\tilde{\boldsymbol{y}}\right)}_{=:\boldsymbol{\mu}_\beta^*}\right)\right). \end{aligned}$$

$$\begin{aligned} p(\sigma^2_{\beta_r}|\mu_{\beta r}, \beta_r) &\propto p(\beta_r|\mu_{\beta r}, \sigma^2_{\beta_r})\; p(\sigma^2_{\beta_r}) \\ &\propto \frac{1}{(\sigma^2_{\beta_r})^{0.5}} \exp\left(-\frac{1}{2\sigma^2_{\beta_r}}(\beta_r - \mu_{\beta r})^2\right) \frac{1}{(\sigma^2_{\beta_r})^{a_\beta+1}} \exp\left(-\frac{b_\beta}{\sigma^2_{\beta_r}}\right) \\ &= \frac{1}{(\sigma^2_{\beta_r})^{a_\beta+0.5+1}} \exp\left(-\frac{1}{\sigma^2_{\beta_r}}\left(b_\beta + \frac{1}{2}(\beta_r - \mu_{\beta r})^2\right)\right). \end{aligned}$$

$$
\begin{aligned}
p(\mu_{\beta r}|\sigma^2_{\beta_r}, \beta_r) &\propto p(\beta_r|\mu_{\beta r}, \sigma^2_{\beta_r})\; p(\mu_{\beta r}) \\
&\propto \exp\left(-\frac{1}{2\sigma^2_{\beta_r}}(\beta_r - \mu_{\beta r})^2\right) \exp\left(-\frac{1}{2s^2_{\beta_r}}(\mu_{\beta r} - m_{\beta r})^2\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2_{\beta_r}}\beta_r^2 - 2\frac{1}{\sigma^2_{\beta_r}}\beta_r\mu_{\beta r} + \frac{1}{\sigma^2_{\beta_r}}\mu^2_{\beta r} + \frac{1}{s^2_{\beta_r}}\mu^2_{\beta r} - 2\frac{1}{s^2_{\beta_r}}\mu_{\beta r}m_{\beta r} + \frac{1}{s^2_{\beta_r}}m^2_{\beta r}\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\mu^2_{\beta r}\left(\frac{1}{\sigma^2_{\beta_r}} + \frac{1}{s^2_{\beta_r}}\right) - 2\mu_{\beta r}\left(\frac{\beta_r}{\sigma^2_{\beta_r}} + \frac{m_{\beta r}}{s^2_{\beta_r}}\right)\right)\right) \\
&= \exp\left(-\frac{1}{2\left(\frac{1}{\sigma^2_{\beta_r}} + \frac{1}{s^2_{\beta_r}}\right)^{-1}}\left(\mu^2_{\beta r} - 2\mu_{\beta r}\left(\frac{1}{\sigma^2_{\beta_r}} + \frac{1}{s^2_{\beta_r}}\right)^{-1}\left(\frac{\beta_r}{\sigma^2_{\beta_r}} + \frac{m_{\beta r}}{s^2_{\beta_r}}\right)\right)\right).
\end{aligned}
$$

Parameters of the Random Effects Part

Priors:

$$
\begin{aligned}
p(\boldsymbol{b}_i|\boldsymbol{\theta}_i, \boldsymbol{D}) &\propto |\boldsymbol{D}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{b}_i - \boldsymbol{\theta}_i)^T\boldsymbol{D}^{-1}(\boldsymbol{b}_i - \boldsymbol{\theta}_i)\right), \\
p(\sigma^2_{b_r}) &\propto \frac{1}{(\sigma^2_{b_r})^{a_b+1}} \exp\left(-\frac{b_b}{\sigma^2_{b_r}}\right), \\
p(\boldsymbol{\theta}_i|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) &\propto |\boldsymbol{\Sigma}_0|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\mu}_0)\right), \\
p(\boldsymbol{\mu}_h|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) &\propto |\boldsymbol{\Sigma}_0|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\mu}_h - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_h - \boldsymbol{\mu}_0)\right), \\
p(\mu_{0r}) &\propto \exp\left(-\frac{1}{2s^2_{0_r}}(\mu_{0r} - m_{0r})^2\right), \\
p(\sigma^2_{0_r}) &\propto \frac{1}{(\sigma^2_{0_r})^{a_0+1}} \exp\left(-\frac{b_0}{\sigma^2_{0_r}}\right), \\
p(\alpha) &\propto \alpha^{a_\alpha - 1}\exp(-b_\alpha\alpha) \qquad \text{(gamma prior version)}, \\
\alpha &\sim \sum_{\omega\in\Omega} P(\alpha = \omega)\delta_\omega \qquad \text{(discrete prior version)}.
\end{aligned}
$$

Further distributions:

$$
\begin{aligned}
p(\boldsymbol{c}|\boldsymbol{v}) = p(n_1,\dots,n_N|\boldsymbol{v}) &\propto \prod_{h=1}^{N} \pi_h^{n_h} = \prod_{h=1}^{N}\left(v_h \prod_{l=1}^{h-1}(1-v_l)\right)^{n_h} \\
&= \prod_{h=1}^{N} v_h^{n_h} \cdot \prod_{h=1}^{N}\prod_{l=1}^{h-1}(1-v_l)^{n_h} = \prod_{h=1}^{N-1} v_h^{n_h} \cdot \prod_{h=1}^{N-1}\prod_{l=h+1}^{N}(1-v_h)^{n_l} \\
&= \prod_{h=1}^{N-1} v_h^{n_h} \cdot \prod_{h=1}^{N-1}(1-v_h)^{\sum_{l=h+1}^{N} n_l}, \\
p(\boldsymbol{v}|\alpha) = \prod_{h=1}^{N-1} p(v_h|\alpha) &= \prod_{h=1}^{N-1} \frac{1}{B(1,\alpha)}(1-v_h)^{\alpha-1} = \prod_{h=1}^{N-1} \frac{\Gamma(1+\alpha)}{\Gamma(1)\Gamma(\alpha)}(1-v_h)^{\alpha-1} \\
&= \prod_{h=1}^{N-1} \alpha(1-v_h)^{\alpha-1} = \alpha^{N-1}\prod_{h=1}^{N-1}(1-v_h)^{\alpha-1} \\
&= \exp\left(\log\left(\alpha^{N-1}\prod_{h=1}^{N-1}(1-v_h)^{\alpha-1}\right)\right) \\
&= \exp\left(\log\alpha^{N-1} + \sum_{h=1}^{N-1}\log(1-v_h)^{\alpha-1}\right) \\
&= \exp\left((N-1)\cdot\log\alpha + (\alpha-1)\cdot\sum_{h=1}^{N-1}\log(1-v_h)\right).
\end{aligned}
$$

Full conditionals:

$$
\begin{aligned}
p(\boldsymbol{b}_i|\boldsymbol{\theta}_i,\boldsymbol{D},\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{y},\sigma^2) &\propto p(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{b}_i,\sigma^2)\; p(\boldsymbol{b}_i|\boldsymbol{\theta}_i,\boldsymbol{D}) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{y}}_i - \boldsymbol{Z}_i\boldsymbol{b}_i)^T(\tilde{\boldsymbol{y}}_i - \boldsymbol{Z}_i\boldsymbol{b}_i)\right)\exp\left(-\frac{1}{2}(\boldsymbol{b}_i-\boldsymbol{\theta}_i)^T\boldsymbol{D}^{-1}(\boldsymbol{b}_i-\boldsymbol{\theta}_i)\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{b}_i^T\boldsymbol{Z}_i^T\boldsymbol{Z}_i\boldsymbol{b}_i - 2\boldsymbol{b}_i^T\boldsymbol{Z}_i^T\tilde{\boldsymbol{y}}_i)\right)\exp\left(-\frac{1}{2}(\boldsymbol{b}_i^T\boldsymbol{D}^{-1}\boldsymbol{b}_i - 2\boldsymbol{b}_i^T\boldsymbol{D}^{-1}\boldsymbol{\theta}_i)\right) \\
&= \exp\left(-\frac{1}{2}\left(\boldsymbol{b}_i^T\underbrace{\left(\boldsymbol{D}^{-1}+\frac{1}{\sigma^2}\boldsymbol{Z}_i^T\boldsymbol{Z}_i\right)}_{=:\boldsymbol{D}_i^{*-1}}\boldsymbol{b}_i - 2\boldsymbol{b}_i^T\left(\boldsymbol{D}^{-1}\boldsymbol{\theta}_i + \frac{1}{\sigma^2}\boldsymbol{Z}_i^T\tilde{\boldsymbol{y}}_i\right)\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\boldsymbol{b}_i^T\boldsymbol{D}_i^{*-1}\boldsymbol{b}_i - 2\boldsymbol{b}_i^T\boldsymbol{D}_i^{*-1}\underbrace{\boldsymbol{D}_i^*\left(\boldsymbol{D}^{-1}\boldsymbol{\theta}_i + \frac{1}{\sigma^2}\boldsymbol{Z}_i^T\tilde{\boldsymbol{y}}_i\right)}_{=:\boldsymbol{\theta}_i^*}\right)\right).
\end{aligned}
$$

$$
\begin{aligned}
p(\sigma^2_{b_r}|\boldsymbol{\theta},\boldsymbol{b}) &\propto \left(\prod_{i=1}^{n} p(b_{ir}|\theta_{ir},\sigma^2_{b_r})\right)\ p(\sigma^2_{b_r}) \\
&\propto \frac{1}{(\sigma^2_{b_r})^{0.5n}} \exp\left(-\frac{1}{2\sigma^2_{b_r}}\sum_{i=1}^{n}(b_{ir}-\theta_{ir})^2\right) \frac{1}{(\sigma^2_{b_r})^{a_b+1}} \exp\left(-\frac{b_b}{\sigma^2_{b_r}}\right) \\
&= \frac{1}{(\sigma^2_{b_r})^{a_b+0.5n+1}} \exp\left(-\frac{1}{\sigma^2_{b_r}}\left(b_b+\frac{1}{2}\sum_{i=1}^{n}(b_{ir}-\theta_{ir})^2\right)\right).
\end{aligned}
$$

$p(\mu_{hr}|\sigma^2_{b_r},\mu_{0r},\sigma^2_{0_r},\boldsymbol{b},\boldsymbol{c}) \propto \left(\prod_{i:c_i=h} p(b_{ir}|\mu_{hr},\sigma^2_{b_r})\right)\ p(\mu_{hr})$

$$
\begin{aligned}
&\propto \exp\left(-\frac{1}{2\sigma^2_{b_r}}\sum_{i:c_i=h}(b_{ir}-\mu_{hr})^2\right)\exp\left(-\frac{1}{2\sigma^2_{0_r}}(\mu_{hr}-\mu_{0r})^2\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2_{b_r}}n_h(\bar{b}_{r,h}-\mu_{hr})^2\right)\exp\left(-\frac{1}{2\sigma^2_{0_r}}(\mu_{hr}-\mu_{0r})^2\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{n_h}{\sigma^2_{b_r}}\bar{b}^2_{r,h}-2\frac{n_h}{\sigma^2_{b_r}}\bar{b}_{r,h}\mu_{hr}+\frac{n_h}{\sigma^2_{b_r}}\mu^2_{hr}+\frac{1}{\sigma^2_{0_r}}\mu^2_{hr}-2\frac{1}{\sigma^2_{0_r}}\mu_{hr}\mu_{0r}+\frac{1}{\sigma^2_{0_r}}\mu^2_{0r}\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\mu^2_{hr}\left(\frac{n_h}{\sigma^2_{b_r}}+\frac{1}{\sigma^2_{0_r}}\right)-2\mu_{hr}\left(\frac{n_h}{\sigma^2_{b_r}}\bar{b}_{r,h}+\frac{\mu_{0r}}{\sigma^2_{0_r}}\right)\right)\right) \\
&= \exp\left(-\frac{1}{2\left(\frac{n_h}{\sigma^2_{b_r}}+\frac{1}{\sigma^2_{0_r}}\right)^{-1}}\left(\mu^2_{hr}-2\mu_{hr}\left(\frac{n_h}{\sigma^2_{b_r}}+\frac{1}{\sigma^2_{0_r}}\right)^{-1}\left(\frac{n_h}{\sigma^2_{b_r}}\bar{b}_{r,h}+\frac{\mu_{0r}}{\sigma^2_{0_r}}\right)\right)\right).
\end{aligned}
$$

$p(\mu_{0r}|\sigma^2_{0_r},\boldsymbol{\theta}) \propto \left(\prod_{i=1}^{n} p(\theta_{ir}|\mu_{0r},\sigma^2_{0_r})\right)\ p(\mu_{0r})$

$$
\begin{aligned}
&\propto \exp\left(-\frac{1}{2\sigma^2_{0_r}}\sum_{i=1}^{n}(\theta_{ir}-\mu_{0r})^2\right)\exp\left(-\frac{1}{2s^2_{0_r}}(\mu_{0r}-m_{0r})^2\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2_{0_r}}n(\bar{\theta}_r-\mu_{0r})^2\right)\exp\left(-\frac{1}{2s^2_{0_r}}(\mu_{0r}-m_{0r})^2\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{n}{\sigma^2_{0_r}}\bar{\theta}^2_r-2\frac{n}{\sigma^2_{0_r}}\bar{\theta}_r\mu_{0r}+\frac{n}{\sigma^2_{0_r}}\mu^2_{0r}+\frac{1}{s^2_{0_r}}\mu^2_{0r}-2\frac{1}{s^2_{0_r}}\mu_{0r}m_{0r}+\frac{1}{s^2_{0_r}}m^2_{0r}\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\mu^2_{0r}\left(\frac{n}{\sigma^2_{0_r}}+\frac{1}{s^2_{0_r}}\right)-2\mu_{0r}\left(\frac{n}{\sigma^2_{0_r}}\bar{\theta}_r+\frac{m_{0r}}{s^2_{0_r}}\right)\right)\right) \\
&= \exp\left(-\frac{1}{2\left(\frac{n}{\sigma^2_{0_r}}+\frac{1}{s^2_{0_r}}\right)^{-1}}\left(\mu^2_{0r}-2\mu_{0r}\left(\frac{n}{\sigma^2_{0_r}}+\frac{1}{s^2_{0_r}}\right)^{-1}\left(\frac{n}{\sigma^2_{0_r}}\bar{\theta}_r+\frac{m_{0r}}{s^2_{0_r}}\right)\right)\right).
\end{aligned}
$$

$$
\begin{aligned}
p(\sigma_{0_r}^2|\mu_{0r}, \boldsymbol{\theta}) &\propto \left(\prod_{i=1}^{n} p(\theta_{ir}|\mu_{0r}, \sigma_{0_r}^2)\right)\ p(\sigma_{0_r}^2) \\
&\propto \frac{1}{(\sigma_{0_r}^2)^{0.5n}} \exp\left(-\frac{1}{2\sigma_{0_r}^2}\sum_{i=1}^{n}(\theta_{ir}-\mu_{0r})^2\right) \frac{1}{(\sigma_{0_r}^2)^{a_0+1}} \exp\left(-\frac{b_0}{\sigma_{0_r}^2}\right) \\
&= \frac{1}{(\sigma_{0_r}^2)^{a_0+0.5n+1}} \exp\left(-\frac{1}{\sigma_{0_r}^2}\left(b_0 + \frac{1}{2}\sum_{i=1}^{n}(\theta_{ir}-\mu_{0r})^2\right)\right).
\end{aligned}
$$

$$
\begin{aligned}
p(\alpha|\boldsymbol{v}) &\propto p(\boldsymbol{v}|\alpha)\ p(\alpha) \qquad \text{(gamma prior version)} \\
&\propto \prod_{h=1}^{N-1} \alpha(1-v_h)^{\alpha-1}\alpha^{a_\alpha-1}\exp(-b_\alpha\alpha) \\
&\propto \alpha^{N-1+a_\alpha-1}\exp\left(\log\left(\prod_{h=1}^{N-1}(1-v_h)^{\alpha}\right)\right)\exp\left(-b_\alpha\alpha\right) \\
&= \alpha^{N-1+a_\alpha-1}\exp\left(\sum_{h=1}^{N-1}\alpha\log(1-v_h) - b_\alpha\alpha\right) \\
&= \alpha^{N-1+a_\alpha-1}\exp\left(-\alpha\left(b_\alpha - \sum_{h=1}^{N-1}\log(1-v_h)\right)\right).
\end{aligned}
$$

$$
\begin{aligned}
P(\alpha=\omega|\boldsymbol{v}) &\propto p(\boldsymbol{v}|\alpha)\cdot P(\alpha=\omega) \qquad \text{(discrete prior version)} \\
&= \exp\left((N-1)\cdot\log\omega + (\omega-1)\cdot\sum_{h=1}^{N-1}\log(1-v_h)\right)\cdot P(\alpha=\omega).
\end{aligned}
$$

$$
P(c_i=h|\boldsymbol{v},\boldsymbol{\mu},\boldsymbol{b}_i,\boldsymbol{D}) \propto p(\boldsymbol{b}_i|\boldsymbol{\mu}_h,\boldsymbol{D})\pi_h \propto |\boldsymbol{D}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{b}_i-\boldsymbol{\mu}_h)^T\boldsymbol{D}^{-1}(\boldsymbol{b}_i-\boldsymbol{\mu}_h)\right)\pi_h.
$$

$$
\begin{aligned}
p(\boldsymbol{v}|\boldsymbol{c},\alpha) &\propto p(\boldsymbol{c}|\boldsymbol{v})\ p(\boldsymbol{v}|\alpha) \\
&\propto \prod_{h=1}^{N-1} v_h^{n_h}(1-v_h)^{\sum_{l=h+1}^{N} n_l}\prod_{h=1}^{N-1}(1-v_h)^{\alpha-1} \\
&\propto \prod_{h=1}^{N-1} v_h^{n_h}(1-v_h)^{\alpha+\sum_{l=h+1}^{N} n_l-1}.
\end{aligned}
$$

A.8. Reparametrization of the P-spline

The following section is based on Tutz (2012). Using the notation of Section 6.2.1 the basis coefficients vector $\boldsymbol{\gamma}$ could be separated into an unpenalized vector $\boldsymbol{\gamma}_0$ and a penalized vector $\boldsymbol{\gamma}_p$ by considering $\boldsymbol{\gamma} = \boldsymbol{T}\boldsymbol{\gamma}_0 + \boldsymbol{W}\boldsymbol{\gamma}_p$ if the conditions

(i) $\boldsymbol{T}^T\boldsymbol{K} = \boldsymbol{0}$,

(ii) $\boldsymbol{W}^T\boldsymbol{K}\boldsymbol{W} = \boldsymbol{I}_{d-k}$,

are fulfilled. Then the penalty term can be transformed into one that is known from mixed models with normally distributed random effects:

$$\begin{aligned}\boldsymbol{\gamma}^T\boldsymbol{K}\boldsymbol{\gamma} &= (\boldsymbol{T}\boldsymbol{\gamma}_0 + \boldsymbol{W}\boldsymbol{\gamma}_p)^T\boldsymbol{K}(\boldsymbol{T}\boldsymbol{\gamma}_0 + \boldsymbol{W}\boldsymbol{\gamma}_p) \\ &= \boldsymbol{\gamma}_0^T \underbrace{\boldsymbol{T}^T\boldsymbol{K}}_{\boldsymbol{0}}\boldsymbol{T}\boldsymbol{\gamma}_0 + 2\boldsymbol{\gamma}_0^T \underbrace{\boldsymbol{T}^T\boldsymbol{K}}_{\boldsymbol{0}}\boldsymbol{W}\boldsymbol{\gamma}_p + \boldsymbol{\gamma}_p^T \underbrace{\boldsymbol{W}^T\boldsymbol{K}\boldsymbol{W}}_{\boldsymbol{I}_{d-k}}\boldsymbol{\gamma}_p = \boldsymbol{\gamma}_p^T\boldsymbol{\gamma}_p.\end{aligned}$$

The conditions (i) and (ii) are fulfilled for $\boldsymbol{T} = \boldsymbol{\Gamma}_0$ and $\boldsymbol{W} = \boldsymbol{\Gamma}_p\boldsymbol{\Omega}_p^{-\frac{1}{2}}$:

$$\begin{aligned}\boldsymbol{T}^T\boldsymbol{K} &= \boldsymbol{\Gamma}_0^T\boldsymbol{K} = \underbrace{\boldsymbol{\Gamma}_0^T\boldsymbol{\Gamma}_p}_{\boldsymbol{0}}\boldsymbol{\Omega}_p\boldsymbol{\Gamma}_p^T = \boldsymbol{0}, \\ \boldsymbol{W}^T\boldsymbol{K}\boldsymbol{W} &= \boldsymbol{\Omega}_p^{-\frac{T}{2}} \underbrace{\boldsymbol{\Gamma}_p^T\boldsymbol{\Gamma}_p}_{\boldsymbol{I}_{d-k}}\boldsymbol{\Omega}_p \underbrace{\boldsymbol{\Gamma}_p^T\boldsymbol{\Gamma}_p}_{\boldsymbol{I}_{d-k}}\boldsymbol{\Omega}_p^{-\frac{1}{2}} = \underbrace{\boldsymbol{\Omega}_p^{-\frac{T}{2}}\boldsymbol{\Omega}_p^{\frac{T}{2}}}_{\boldsymbol{I}_{d-k}} \underbrace{\boldsymbol{\Omega}_p^{\frac{1}{2}}\boldsymbol{\Omega}_p^{-\frac{1}{2}}}_{\boldsymbol{I}_{d-k}} = \boldsymbol{I}_{d-k}.\end{aligned}$$

If the penalty matrix is given by $\boldsymbol{K} = \boldsymbol{\Delta}^T\boldsymbol{\Delta}$ the conditions (i) and (ii) are also fulfilled for the choice given in 6.6. Then one becomes:

$$\begin{aligned}\boldsymbol{T}^T\boldsymbol{K} &= \underbrace{\boldsymbol{T}^T\boldsymbol{\Delta}^T}_{\boldsymbol{0}}\boldsymbol{\Delta} = \boldsymbol{0}, \\ \boldsymbol{W}^T\boldsymbol{K}\boldsymbol{W} &= \underbrace{((\boldsymbol{\Delta}\boldsymbol{\Delta}^T)^{-1})^T\boldsymbol{\Delta}\boldsymbol{\Delta}^T}_{\boldsymbol{I}_{d-k}} \underbrace{\boldsymbol{\Delta}\boldsymbol{\Delta}^T(\boldsymbol{\Delta}\boldsymbol{\Delta}^T)^{-1}}_{\boldsymbol{I}_{d-k}} = \boldsymbol{I}_{d-k}.\end{aligned}$$

A.9. Simulation of Observation Times

In the following the concept for the simulation of observation times in Section 6.3 is visualized. As an example we assume $n_i = 6$. See Figure A.1 for an illustration of the underlying intervals.

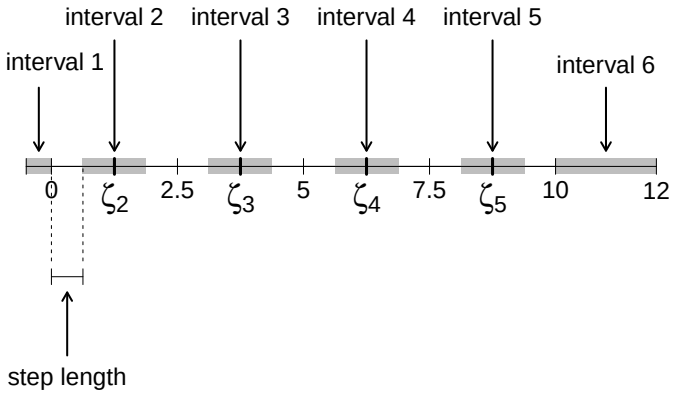


Figure A.1.: The observation times $t_{i1}, \ldots, t_{i6}$ are drawn from uniform distributions on the grey intervals.

In this case the step length is given by

$$\frac{\frac{10}{n_i-2}}{4} = \frac{2.5}{n_i - 2} = 0.625.$$

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics 2*, 1152–1174.

Arenz, S., R. Rückerl, B. Koletzko, and R. von Kries (2004). Breast-feeding and childhood obesity - a systematic review. *International Journal of Obesity and Related Metabolic Disorders 28*, 1247–1256.

Basford, K. E., D. R. Greenway, G. J. McLachlan, and D. Peel (1997). Standard errors of fitted means under normal mixture models. *Computational Statistics 12*, 1–17.

Bates, D. M. and M. Maechler (2012). *Matrix: Sparse and dense matrix classes and methods*. R package version 1.0-10.

Bates, D. M., M. Maechler, and B. Bolker (2012). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999999-0.

Bates, D. M. and W. Venables (2011). *splines: Regression spline functions and classes*. R package version 2.13.0.

Belitz, C., A. Brezger, T. Kneib, S. Lang, and N. Umlauf (2012). *BayesX - Software for Bayesian inference in structured additive regression models.*

Beyerlein, A., L. Fahrmeir, U. Mansmann, and A. Toschke (2008). Alternative regression models to assess increase in childhood BMI. *BMC Medical Research Methodology 8*, (59).

Beyerlein, A., A. Toschke, and R. von Kries (2008). Breastfeeding and childhood obesity: Shift of the entire BMI distribution or only the upper parts? *Obesity 16*, 2730–2733.

Blackwell, D. (1973). Discreteness of Ferguson selections. *The Annals of Statistics 1*, 356–358.

Blackwell, D. and J. B. MacQueen (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics 1*, 353–355.

Blei, D. M. and M. I. Jordan (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis 1*, 121–144.

Boeckmann, A. J., L. B. Sheiner, and S. L. Beal (1994). *NONNEM users guide: part V*. San Francisco: University of California.

Booth, J. G., G. Casella, and J. P. Hobert (2008). Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society B 70*, 119–139.

Braun, J., L. Held, and B. Ledergerber (2012). Predictive cross-validation for the choice of linear mixed-effects models with application to data from the swiss HIV cohort study. *Biometrics 68*, 53–61.

Brezger, A., T. Kneib, and S. Lang (2005). BayesX: Analysing Bayesian structured additive regression models. *Journal of Statistical Software 14*, (11).

Brezger, A. and S. Lang (2006). Generalized additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis 50*, 967–991.

Bühlmann, P. and T. Hothorn (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science 22*, 477–505.

Burkhardt, J. (2008). *ASA047: Nelder-Mead Minimization Algorithm*. C++ library.

Bush, C. A. and S. N. MacEachern (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika 83*, 275–285.

Celeux, G., O. Martin, and C. Lavergne (2005). Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling 5*, 243–267.

Chen, C.-M., P. Rzehak, A. Zutavern, B. Fahlbusch, W. Bischof, O. Herbarth, M. Borte, I. Lehmann, H. Behrendt, U. Krämer, H.-E. Wichmann, and J. Heinrich (2007). Longitudinal study on cat allergen exposure and the development of allergy in young children. *The Journal of Allergy and Clinical Immunology 119*, 1148–1155.

Davidian, M. and D. M. Giltinan (1995). *Nonlinear models for repeated measurement data*. London: Chapman & Hall.

Davies, R. B. (2008). *Newmat C++ matrix library*. C++ library version 11.

De Boor, C. (1978). *A practical guide to splines*. New York: Springer-Verlag.

De la Cruz-Mesía, R., F. A. Quintana, and G. Marshall (2008). Model-based clustering for longitudinal data. *Computational Statistics & Data Analysis 52*, 1441–1457.

Dempster, A. P., N. M. Laired, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B 39*, 1–38.

Diggle, P. J., P. Heagerly, K. Y. Liang, and S. L. Zeger (2002). *Analysis of longitudinal data* (2nd ed.). Oxford: Oxford University Press.

Dockery, D. W., C. S. Berkey, J. H. Ware, F. E. Speizer, and B. G. Ferris (1983). Distribution of fvc and fev1 in children 6 to 11 years old. *American Review of Respiratory Disease 128*, 405–412.

Dunson, D., M. Yang, and D. Baird (2007). Semiparametric Bayes hierarchical models with mean and variance constraints. *Computational Statistics & Data Analysis 54*, 2172–2186.

Durban, M., J. Harezlak, M. P. Wand, and R. J. Carroll (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine 24*, 1153–1167.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*, 1–26.

Efron, B. and R. Tibshirani (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*, 89–121.

Eilers, P. H. C. and B. D. Marx (2010). Splines, knots and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics 2*, 637–653.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association 89*, 268–277.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*, 577–588.

Fahrmeir, L. and T. Kneib (2011). *Bayesian smoothing and regression for longitudinal, spatial and event history data*. Oxford: Oxford University Press.

Fahrmeir, L., T. Kneib, and S. Lang (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica 14*, 731–761.

Fahrmeir, L., T. Kneib, and S. Lang (2007). *Regression - Modelle, Methoden und Anwendungen*. Berlin: Springer.

Fahrmeir, L. and S. Lang (2001). Bayesian semiparametric regression analysis of multicategorical time-space data. *Annals of the Institute of Statistical Mathematics 53*, 11–30.

Fan, J. and R. Li (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association 99*, 710–723.

Fenske, N., L. Fahrmeir, P. Rzehak, and M. Höhle (2008). Detection of risk factors for obesity in early childhood with quantile regression methods for longitudinal data. Technical Report 38, Ludwig-Maximilians-University Munich.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*, 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics 2*, 615–629.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh 52*, 399–433.

Fitzmaurice, G. M., N. M. Laird, and J. H. Ware (2004). *Applied longitudinal analysis* (2nd ed.). Wiley Series in Probability and Statistics. New Jersey: Wiley.

Freund, Y. and R. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences 55*, 119–139.

Frigyik, B. A., A. Kapila, and M. R. Gupta (2010). Introduction to the Dirichlet distribution and related processes. Technical Report 6, University of Washington.

Fritsch, A. and K. Ickstadt (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis 4*, 367–392.

Gaffney, S. J. and P. Smyth (2003). Curve clustering with random effects regression mixtures. In C. M. Bishop and B. J. Frey (Eds.), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Key West, FL.

Gelfand, A. E. and A. Kottas (2002). A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *Journal of Computational and Graphical Statistics 11*, 289–305.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis 1*, 515–553.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review 55*, 245–259.

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

Groll, A. and G. Tutz (2012). Regularization for generalized additive mixed models by likelihood-based boosting. *Methods of Information in Medicine 51*, 168–177.

Groll, A. and G. Tutz (2013). Variable selection for generalized linear mixed models by L1-penalized estimation. *Statistics and Computing*. (to appear).

Harder, T., R. Bergmann, G. Kallischnigg, and A. Plagemann (2005). Duration of breastfeeding and risk of overweight: a meta-analysis. *American Journal of Epidemiology 162*, 397–403.

Heinzl, F. (2009). Nonparametrische Bayes-Inferenz in additiven gemischten Modellen. Diploma Thesis, Ludwig-Maximilians-University Munich.

Heinzl, F. (2012). *clustmixed: Clustering in linear and additive mixed models*. R package version 0.1.

Heinzl, F., L. Fahrmeir, and T. Kneib (2012). Additive mixed models with Dirichlet process mixture and P-spline priors. *Advances in Statistical Analysis 96*, 47–68.

Heinzl, F. and G. Tutz (2012). Clustering in linear mixed models with a group fused lasso penalty. Technical Report 123, Ludwig-Maximilians-University Munich.

Heinzl, F. and G. Tutz (2013). Clustering in linear mixed models with approximate Dirichlet process mixtures using EM algorithm. *Statistical Modelling 13*, 41–67.

Hjort, N. L., C. Holmes, P. Müller, and S. G. Walker (2010). *Bayesian nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*, 161–173.

Ishwaran, H. and L. F. James (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics 11*, 508–532.

James, G. M. and C. A. Sugar (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association 98*, 397–408.

Jara, A. (2007). Applied Bayesian non- and semiparametric inference using DPpackage. *R News 3*, 17–26.

Jara, A., T. E. Hanson, and E. Lesaffre (2009). Robustifying generalized linear mixed models using a new class of mixtures of multivariate Pólya trees. *Journal of Computational and Graphical Statistics 18*, 838–860.

Jullion, A. and P. Lambert (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. *Computational Statistics & Data Analysis 51*, 2542–2558.

Kleinman, K. and J. Ibrahim (1998). A semiparametric Bayesian approach to the random effects model. *Biometrics 54*, 921–938.

Koenker, R. (2005). *Quantile regression*. Economic Society Monographs. Cambridge: Cambridge University Press.

Komárek, A., B. E. Hansen, E. M. M. Kuiper, H. R. van Buuren, and E. Lesaffre (2010). Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Statistics in Medicine 29*, 3267–3283.

Komárek, A. and L. Komárková (2013). Clustering for multivariate continuous and discrete longitudinal data. *The Annals of Applied Statistics 7*, 177–200.

Komárek, A. and E. Lesaffre (2008). Generalized linear mixed model with a penalized Gaussian mixture as a random effects distribution. *Computational Statistics and Data Analysis 52*, 3441–3458.

Kottas, A. and A. E. Gelfand (2001). Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association 96*, 1458–1468.

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

Lang, S. and A. Brezger (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics 13*, 183–212.

Li, Y., X. Lin, and P. Müller (2010). Bayesian inference in semiparametric mixed models for longitudinal data. *Biometrics 66*, 70–78.

Li, Y., P. Müller, and X. Lin (2011). Center-adjusted inference for a nonparametric Bayesian random effect distribution. *Statistica Sinica 21*, 1201–1223.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*, 13–22.

Lin, X. and D. Zhang (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society B61*, 381–400.

Lindley, D. V. and A. F. M. Smith (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society B 34*, 1–41.

Lindstrom, M. J. and D. M. Bates (1988). Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data. *Journal of the American Statistical Association 83*, 1014–1022.

Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *The Annals of Statistics 24*, 911–930.

Liu, X. and M. C. K. Yang (2009). Simultaneous curve registration and clustering for functional data. *Computational Statistics & Data Analysis 53*, 1361–1376.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics - Simulation and Computation 23*, 727–741.

MacEachern, S. N. and P. Müller (1998). Estimating mixtures of Dirichlet process models. *Journal of Computational and Graphical Statistics 7*, 223–238.

Magder, L. S. and S. L. Zeger (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *Journal of the American Statistical Association 91*, 1141–1151.

Maritz, J. S. and T. Lwin (1989). *Empirical Bayes methods. Monographs on statistics and applied probability*. London: Chapman & Hall.

Marshall, E. C. and D. J. Spiegelhalter (2003). Approximate crossvalidatory predictive checks in disease mapping models. *Statistics in Medicine 22*, 1649–1660.

Mayr, A., T. Hothorn, and N. Fenske (2012). Prediction intervals for future BMI values of individual children - a non-parametric approach by quantile boosting. *BMC Medical Research Methodology 12*, (6).

McAuliffe, J. D., D. M. Blei, and M. I. Jordan (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistics and Computing 16*, 5–14.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models* (2nd ed.). New York: Chapman & Hall.

McLachlan, G. J. and T. Krishnan (1997). *The EM algorithm and extensions*. New York: Wiley.

McLachlan, G. J., T. Krishnan, and S. K. Ng (2004). *The EM algorithm*. Humboldt-University Berlin: Center for Applied Statistics and Economics.

McLachlan, G. J. and D. Peel (2000). *Finite mixture models*. New York: Wiley.

Muliere, P. and L. Tardella (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *The Canadian Journal of Statistics 26*, 283–297.

Müller, P. and G. L. Rosner (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association 92*, 1279–1292.

Murdoch, D. and E. D. Chow (2012). *ellipse: Functions for drawing ellipses and ellipse-like confidence regions*. R package version 0.3-7.

Nelder, J. A. and R. Mead (1965). A simplex method for function minimization. *Computer Journal 7*, 308–313.

Newton, M. A. and Y. Zhang (1999). A recursive algorithm for nonparametric analysis with missing data. *Biometrika 86*, 15–26.

Ng, S. K., G. J. McLachlan, K. Wang, L. Ben-Tovim Jones, and S.-W. Ng (2006). A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics 22*, 1745–1752.

Ohlssen, D. I., L. D. Sharples, and D. J. Spiegelhalter (2007). Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons. *Statistics in Medicine 26*, 2088–2112.

O'Neill, R. (1971). Algorithms AS 47: Function minimization using a simplex procedure. *Journal of the Royal Statistical Society C 20*, 338–345.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science 1*, 505–527.

Panagiotelis, A. and M. Smith (2008). Bayesian identification, selection and estimation of semiparametric functions in high-dimensional additive models. *Journal of Econometrics 143*, 291–316.

Papageorgiou, G. and J. Hinde (2012). Multivariate generalized linear mixed models with semi-nonparametric and smooth nonparametric random effects densities. *Statistics and Computing 22*, 79–92.

Papaspiliopoulos, O. and G. O. Roberts (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika 95*, 169–186.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-effects models in S and S-Plus*. New York: Springer.

Plummer, M., N. Best, K. Cowles, K. Vines, D. Sarkar, and R. Almond (2012). *coda: Output analysis and diagnostics for MCMC*. R package version 0.16-1.

R Development Core Team (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Reinsch, C. (1967). Smoothing by spline functions. *Numerische Mathematik 10*, 177–183.

Rigby, R. and D. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Applied Statistics 54*, 507–554.

Ruppert, D. and R. J. Carroll (2000). Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics 42*, 205–223.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric regression*. Cambridge: Cambridge University Press.

Rzehak, P., S. Sausenthaler, S. Koletzko, C. P. Bauer, B. Schaaf, A. von Berg, D. Berdel, M. Borte, O. Herbarth, U. Krämer, N. Fenske, H.-E. Wichmann, and J. Heinrich (2009). Period-specific growth, overweight and modification by breastfeeding in the GINI and LISA birth cohorts up to age 6 years. *European Journal of Epidemiology 24*, 449–467.

Scharl, T., B. Grün, and F. Leisch (2010). Mixtures of regression models for time course gene expression data: evaluation of initialization and random effects. *Bioinformatics 26*, 370–377.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica 4*, 639–650.

Stroustrup, B. (1997). *The C++ programming language* (3rd ed.). Amsterdam: Addison-Wesley.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B 58*, 267–288.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Kneight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society B 67*, 91–108.

Tutz, G. (2012). *Regression for categorical data*. Cambridge: Cambridge University Press.

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

Verbeke, G. and E. Lesaffre (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association 91*, 217–221.

Verbeke, G. and E. Lesaffre (1999). The effect of drop-out on the efficiency of longitudinal experiments. *Applied Statistics 48*, 363–375.

Verbeke, G. and G. Molenberghs (2000). *Linear mixed models for longitudinal data*. Springer Series in Statistics. New York: Springer.

Verbyla, A. P., B. R. Cullis, M. G. Kenward, and S. J. Welham (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society C 48*, 269–300.

Verdonck, A., L. de Ridder, G. Verbeke, J. P. Bourguignon, C. Carels, E. R. Kuhn, V. Darras, and F. de Zegher (1998). Comparative effects of neonatal and prepurbetal castration on craniofacial growth in rats. *Archives of Oral Biology 43*, 861–871.

Villarroel, L., G. Marshall, and A. E. Barón (2009). Cluster analysis using multivariate mixed effects models. *Statistics in Medicine 28*, 2552–2565.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation 36*, 45–54.

Wang, N., R. J. Carroll, and X. Lin (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association 100*, 147–157.

Weise, F.-J., H. Alt, and R. Becker (Eds.) (2011). *Arbeitsmarkt in Zahlen*. Nürnberg: Statistik der Bundesagentur für Arbeit.

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. London: Chapman & Hall.

Yao, C. and C. Holmes (2011). Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination. *Bayesian Analysis 6*, 329–351.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B 68*, 49–67.

Zeger, S. L. and P. J. Diggle (1994). Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics 50*, 689–699.

Zhang, D., X. Lin, J. Raz, and M. F. Sowers (1998). Semi-parametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association 93*, 710–719.

Zutavern, A., P. Rzehak, I. Brockow, B. Schaaf, C. Bollrath, A. von Berg, E. Link, U. Krämer, M. Borte, O. Herbarth, H.-E. Wichmann, and J. Heinrich (2007). Day care in relation to respiratory-tract and gastrointestinal infections in a German birth cohort study. *Acta Paediatrica 96*, 1494–1499.

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 31. Januar 2013 Felix Heinzl