LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Felix Heinzl, Thomas Kneib & Ludwig Fahrmeir

Additive mixed models with Dirichlet process mixture and P-spline priors

Technical Report Number 068, 2009
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Additive mixed models with Dirichlet process mixture and P-spline priors

Felix Heinzl

Department of Statistics

Ludwig-Maximilians-University Munich

Thomas Kneib

Department of Mathematics

Carl von Ossietzky University Oldenburg

Ludwig Fahrmeir

Department of Statistics

Ludwig-Maximilians-University Munich

Abstract

Longitudinal data often require a combination of flexible trends and individual-specific random effects. In this paper, we propose a fully Bayesian approach based on Markov chain Monte Carlo simulation techniques that allows for the semiparametric specification of both the trend function and the random effects distribution. Bayesian penalized splines are considered for the former, while a Dirichlet process mixture (DPM) specification allows for an adaptive amount of deviations from normality for the latter. We investigate the advantages of DPM prior structures for random effects in terms of a simulation study and present a challenging application that requires semiparametric mixed modeling.

Keywords: *Dirichlet process mixture, mixed models, penalized splines, nonparametric Bayes inference*

1 Introduction

Complex longitudinal data often require a combination of nonlinear effects such as flexible trends and individual-specific effects. For example, in the application on childhood obesity that motivated our research, we investigate the temporal development of the body mass index (BMI) in children. The data are obtained from a prospective birth cohort study on roughly 3,000 healthy neonates, where the BMI as well as a number of further covariates are collected at up to nine mandatory medical examinations starting at birth and ending at an age of 60 months. The development of the BMI within this time period is well-known to be highly nonlinear but the precise form of the temporal trend differs widely across children.

Such behavior can be represented in a longitudinal semiparametric regression model

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

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 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 $\boldsymbol{b}_i$. Note that we will not use a centered parameterization of the random effects $\boldsymbol{b}_i$ and therefore the vectors $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are required to be disjoint to avoid nonidentifiability of the model. In particular, $\boldsymbol{x}_{ij}$ does not comprise an intercept. The nonlinear time trend will be approximated using Bayesian versions of penalized splines (Brezger & Lang, 2006; Jullion & Lambert, 2007).

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

$$\boldsymbol{z}_{ij}'\boldsymbol{b}_i = b_{0i} + t_{ij}b_{1i}$$

for individual-specific linear deviations or

$$\boldsymbol{z}_{ij}'\boldsymbol{b}_i = b_{0i} + t_{ij}b_{1i} + h(t_{ij})b_{2i} \tag{2}$$

with a known nonlinear transformation $h(t)$ to gain additional flexibility. In our application, we will primarily make use of the latter possibility to adapt individual-specific deviations to the structure of the trend observed in the obesity data. The classical assumption in (1) and more general additive mixed models is a Gaussian (prior) distribution for the random effects, i.e. $\boldsymbol{b}_i$ i.i.d. $\mathrm{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_b)$, see for example Lin & Zhang (1999), Fahrmeir & Lang (2001), Ruppert, Wand & Carroll (2003), Fahrmeir, Kneib & Lang (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.

Therefore, we will consider a Dirichlet process mixture (DPM) prior for the random effects that allows to specify a hyperprior on their prior distribution (see Ferguson (1973) for theory of Dirichlet process). For linear mixed models, Dirichlet process (DP) priors for random effects were first proposed by Kleinman & Ibrahim (1998), and the DP Package in R (Jara, 2007) has options for DP 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.

Note that the model considered in this paper combines two different notions of nonparametric inference: On the one hand, the trend function is modeled nonparametrically to obtain a flexible trend reconstruction that avoids possible pitfalls of parametric specifications. On the other hand, the random effects distribution is specified nonparametrically as detailed in the following paragraph.

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 DP prior $G \sim \text{DP}$ 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 have

$$\begin{aligned} G &\sim \text{DP}(\alpha_0, G_0), \\ \theta_i|G &\overset{i.i.d.}{\sim} G, & i = 1, \ldots, n, \\ \boldsymbol{b}_i|\theta_i &\overset{ind}{\sim} p(\boldsymbol{b}_i|\theta_i), & i = 1, \ldots, n. \end{aligned}$$

As a consequence, the random effects distribution is a mixture of distributions with a Dirichlet process for the mixing distribution:

$$p(\boldsymbol{b}_i) = \int p(\boldsymbol{b}_i|\theta_i) dG(\theta_i).$$

In this DPM specification, each subject still has its own unique random effects value whereas choosing a DP for the $\theta_i, i = 1, \ldots, n$, 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 $\phi_1, \ldots, \phi_k$ 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 $\theta_i = \phi_{c_i}$. The strength of clustering is determined by the concentration parameter α_0 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, Lin & Müller (2009) consider a model that is comparable to (1), but assume a Bayesian smoothing spline for the time trend $f(t)$ in combination with a DP prior for the random effects distribution. In contrast, we use a low-rank Bayesian P-spline for the nonlinear time effect and a DPM prior for the random effects distribution. Therefore, inference for the nonparametric trend function is considerably facilitated, since the P-spline specification dramatically reduces the number of regression coefficients. The DPM prior allows to overcome the restriction to discrete random effects distributions imposed by the DP prior. While Li, Lin & Müller (2009) consider a Pólya urn scheme for implementing their model, our Markov chain Monte Carlo (MCMC) simulation algorithm is based on a truncated stick breaking representation of the Dirichlet process according to Sethuraman (1994). In combination with conjugate priors we obtain a blocked Gibbs sampler. All computations are implemented in C++, allowing for an efficient treatment of loop-intensive calculations, and are made easily accessible by providing an R wrapper function.

The rest of this paper is organized as follows: Section 2 considers the additive mixed model (AMM) with DPM priors for the random effects in more detail. Subsection 2.1 deals with the model hierarchy of the AMM and details prior specifications for all model parameters as well as associated hyperparameter choices. In subsection 2.2 the blocked Gibbs sampler we use for inference is described at full length. The impact of deviations from a Gaussian random effects distribution is investigated in a simulation study in section 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 4 describes the application of the AMM for the childhood obesity data that served as a motivation at the beginning of the introduction. Here we utilize the cluster property of the DP to detect specific patterns in the data. Section 5 concludes with a short summary of our main findings.

2 Additive mixed models with DPM priors

2.1 Model hierarchy

AMMs for longitudinal data extend linear mixed models (LMMs) 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 and consider the model (1). Extensions to AMMs with further nonlinear effects of continuous covariates are conceptually straightforward. In the following, we describe the model hierarchy and prior settings we use in this paper in more detail. In general, the prior choices, in particular choices for hyperparameters are to be understood as default values that worked well in our experience but may have to be adapted to specific data situations. In most cases, the prior settings are motivated by the aim to achieve noninformative priors but for some parameters we will also use specific prior settings to enforce desirable properties of the posterior estimates.

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 $\sigma^2_{\beta_r} \sim IG(a_\beta, b_\beta)$ 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 grows large. We will therefore also use this restriction in the specification of the random effects prior later on. 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 effect distribution $p(\boldsymbol{b}_i|\boldsymbol{\theta}_i)$, we assume a hierarchical Gaussian mixture prior

$$\begin{aligned} \boldsymbol{b}_i|\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_b &\overset{ind}{\sim} N(\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_b), & i = 1, \ldots, n, \\ \boldsymbol{\theta}_i|G &\overset{i.i.d.}{\sim} G, & i = 1, \ldots, n, \\ G &\sim \text{DP}(\alpha_0, G_0). \end{aligned}$$

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 & James, 2002) where

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

and $\delta_{\boldsymbol{\phi}_h}$ denotes the Dirac measure in $\boldsymbol{\phi}_h$. Hence, the unknown distribution G is represented as a weighted sum of point masses with random weights π_h which are

independent of locations $\boldsymbol{\phi}_h$. The locations are i.i.d. random variables from the base distribution G_0, i.e.

$$\boldsymbol{\phi}_h \overset{i.i.d.}{\sim} G_0, \quad h = 1, \ldots, N,$$

while weights are constructed through the stick breaking procedure

$$\begin{array}{rcll} \pi_h & := & V_h \prod_{l<h}(1 - V_l), & h = 1, \ldots, N, \\ V_h & \overset{i.i.d.}{\sim} & Be(1, \alpha_0), & h = 1, \ldots, N-1, \quad V_N = 1. \end{array}$$

Sethuraman (1994) showed that (in the limit $N \to \infty$) the probability measure of G is given by $\mathrm{DP}(\alpha_0, G_0)$. The truncated version still is a good approximation for G because the random weights decrease stochastically as the index h grows. Furthermore $E(\sum_{h=N+1}^{\infty} \pi_h)$ converges to zero exponentially with $N \to \infty$. In our simulations and applications, we truncate the stick breaking representation at $N = 100$.

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 $\boldsymbol{\phi} = (\boldsymbol{\phi}_1', \ldots, \boldsymbol{\phi}_N')'$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)'$ and $\boldsymbol{c} = (c_1, \ldots, c_n)'$. In addition, one obtains estimates for $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ via $\boldsymbol{\theta}_i = \boldsymbol{\phi}_{c_i}$. Contrary to a Pólya urn Gibbs sampling scheme, the stick breaking representation offers the possibility to estimate G itself. See Ishwaran & 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 $\boldsymbol{\Sigma}_0 = \mathrm{diag}(\sigma_{0_0}^2, \ldots, \sigma_{0_q}^2)$ with $\sigma_{0_r}^2 \sim IG(a_0, b_0)$ for $r = 0, \ldots, q$. In analogy to the specifications for the fixed effects hyperpriors, we suggest the following default values for hyperparameters: $\boldsymbol{m}_0 = \mathbf{0}_{q+1}$, $\boldsymbol{S}_0 = 100\boldsymbol{I}_{q+1}$, $a_0 = b_0 = 0.5$. For the prior covariance of the random effects, we also assume a diagonal structure, leading to $\boldsymbol{\Sigma}_b = \mathrm{diag}(\sigma_{b_0}^2, \ldots, \sigma_{b_q}^2)$ with $\sigma_{b_r}^2 \sim IG(a_b, b_b)$ for $r = 0, \ldots, q$ 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. This prior structure yields a high power for detecting clusters in the data.

For the concentration parameter α_0, we consider a discrete prior

$$\alpha_0 \sim \sum_{\omega \in \Omega} P(\alpha_0 = \omega)\, \delta_\omega$$

with support $\Omega = \{0.5, 0.6, \ldots, 100\}$ and probabilities which resemble a gamma distribution. This specification avoids difficulties with too small values for α_0 that would naturally appear in a blocked Gibbs sampler with a gamma prior. For illustration, assume that $\alpha_0 \sim Ga(a_\alpha, b_\alpha)$. In this case, the full conditionals for V_h, $h = 1, \ldots, N-1$ are given by

$$\begin{array}{rcl} V_h | c_1, \ldots, c_n, \alpha_0 & \sim & Be(1 + n_h,\ \alpha_0 + \sum_{l=h+1}^{N} n_l), \\ \alpha_0 | V_1, \ldots, V_{N-1} & \sim & Ga(N - 1 + a_\alpha,\ b_\alpha - \sum_{h=1}^{N-1} \log(1 - V_h)), \end{array}$$

where n_h denotes the number of subjects in cluster h. Updating V_h in stick breaking representation for a small value of α_0 near zero could lead to $V_h = 1$, where h represents an empty cluster followed by further empty clusters. This results in $\alpha_0 = 0$ 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_0 allows us to avoid such problems.

The prior for the concentration parameter of course influences the resulting number of clusters. To express our prior preference for few clusters, we chose standardized values of the $Ga(2, 2)$ density for the discrete prior of α_0.

The nonlinear time trend $f(t)$ is modeled through a Bayesian P-spline. That is we assume that $f(t)$ can be represented through

$$f(t) = \sum_{k=1}^{d} \gamma_k B_k^l(t),$$

where $B_k^l(t)$ 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 (1) can be written in matrix notation as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{\gamma} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n, \tag{3}$$

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 x_{ij} and 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 & Brezger, 2004), a Gaussian smoothness prior

$$p(\boldsymbol{\gamma}|\tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}\right)$$

is assumed for the vector of basis coefficients. The precision matrix acts as a penalty matrix to enforce smoothness and is defined through $\boldsymbol{K} = \boldsymbol{D}'\boldsymbol{D}$, where $\boldsymbol{D}$ 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 log-penalty $\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}$ corresponds exactly to the penalty term introduced by Eilers & Marx (1996) in a frequentist penalized likelihood setting. For the variance parameter, we assume an inverse gamma prior

$$\tau^2 \sim IG(a_\gamma, b_\gamma).$$

In 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 $a_\varepsilon = b_\varepsilon = 0.005$.

Note that the AMM (3) can be written compactly by merging all individual vectors and matrices thereunder, e.g. $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_n')'$, except $\boldsymbol{Z} = \text{diag}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$, yielding

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

This matrix notation of the AMM is advantageous when presenting the blocked Gibbs sampler in Section 2.2.

2.2 Inference

In the following, we describe the resulting block Gibbs sampler for the AMM described in 2.1. The derivation of all full conditionals is given in Heinzl (2009).

In general, inference in AMM has to deal with an identifiability problem. Consider, as a typical example, the AMM (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 $b_{0i} + b_{1i}t$ 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 about zero, modeling only individual linear deviations from $f(t)$. Li, Lin & Müller (2009) deal with this problem in a post-processing step while Dunson, Yang & Baird (2007) introduce a centered DP prior. We suggest centering random effects about zero in the MCMC algorithm as a simple but effective device. To specify $f(t)$ as a non-linear deviation from a linear population trend would require additional, comparably complicated linear constraints for B-splines with a second order random walk prior or for cubic smoothing splines, see Panagiotelis & Smith (2008). For linear regression splines represented through a TP-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 appropriate working responses for updating parameters referring to the P-spline, the fixed effects and the random effects to take into account remaining parameters in full conditionals. 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.

Taking account of these specifics, the blocked Gibbs sampler can be summarized in the following way:

Blocked Gibbs algorithm for AMM:

Let the state of the Markov chain consist of $\boldsymbol{\gamma}$, τ^2, $\boldsymbol{\beta}$, $\boldsymbol{\mu}_\beta$, $\boldsymbol{\Sigma}_\beta$, $\boldsymbol{b}$, $\boldsymbol{\Sigma}_b$, $\boldsymbol{\phi}$, $\boldsymbol{c}$, $\boldsymbol{\pi}$, α_0, $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$ and σ^2.

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:
     - $\boldsymbol{\gamma}|\tau^2, \boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{y}, \sigma^2 \sim N(\boldsymbol{\mu}_\gamma^*, \boldsymbol{\Sigma}_\gamma^*)$,
       $\boldsymbol{\mu}_\gamma^* = (\frac{1}{\tau^2}\boldsymbol{K} + \frac{1}{\sigma^2}\boldsymbol{B}'\boldsymbol{B})^{-1}\frac{1}{\sigma^2}\boldsymbol{B}'\tilde{\boldsymbol{y}}$,
       $\boldsymbol{\Sigma}_\gamma^* = (\frac{1}{\tau^2}\boldsymbol{K} + \frac{1}{\sigma^2}\boldsymbol{B}'\boldsymbol{B})^{-1}$,
     - $\tau^2|\boldsymbol{\gamma} \sim IG\left(a_\gamma + 0.5\, rg(\boldsymbol{K}), b_\gamma + 0.5\, \boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}\right)$.

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:
     - $\boldsymbol{\beta}|\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta, \boldsymbol{\gamma}, \boldsymbol{b}, \boldsymbol{y}, \sigma^2 \sim N(\boldsymbol{\mu}_\beta^*, \boldsymbol{\Sigma}_\beta^*)$,
       $\boldsymbol{\mu}_\beta^* = (\boldsymbol{\Sigma}_\beta^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \frac{1}{\sigma^2}\boldsymbol{X}'\tilde{\boldsymbol{y}})$,
       $\boldsymbol{\Sigma}_\beta^* = (\boldsymbol{\Sigma}_\beta^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X})^{-1}$,
     - 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:

     $$\begin{aligned}\boldsymbol{b}_i|\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_b, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{y}_i, \sigma^2 &\sim N(\boldsymbol{\mu}_b^*, \boldsymbol{\Sigma}_b^*),\\ \boldsymbol{\mu}_b^* &= (\boldsymbol{\Sigma}_b^{-1} + \tfrac{1}{\sigma^2}\boldsymbol{Z}_i'\boldsymbol{Z}_i)^{-1}(\boldsymbol{\Sigma}_b^{-1}\boldsymbol{\theta}_i + \tfrac{1}{\sigma^2}\boldsymbol{Z}_i'\tilde{\boldsymbol{y}}_i),\\ \boldsymbol{\Sigma}_b^* &= (\boldsymbol{\Sigma}_b^{-1} + \tfrac{1}{\sigma^2}\boldsymbol{Z}_i'\boldsymbol{Z}_i)^{-1}.\end{aligned}$$

   - Centering (if there is a P-spline in the model):
     - 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 $\boldsymbol{\phi}_h$ using:
     - If $\nexists\, i : c_i = h$:

       $$\boldsymbol{\phi}_h|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0 \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).$$

     - If $\exists\, i : c_i = h$: For $r = 1, \ldots, q$:

       $$\begin{aligned}\phi_{h_r}|\sigma_{b_r}^2, \mu_{0_r}, \sigma_{0_r}^2, \boldsymbol{b}, \boldsymbol{c} &\sim N(\mu_{0_r}^*, \sigma_{0_r}^{2*}),\\ \mu_{0_r}^* &= (\tfrac{n_h}{\sigma_{b_r}^2} + \tfrac{1}{\sigma_{0_r}^2})^{-1}(\tfrac{n_h}{\sigma_{b_r}^2}\bar{b}_{r,h} + \tfrac{\mu_{0_r}}{\sigma_{0_r}^2}),\\ \sigma_{0_r}^{2*} &= (\tfrac{n_h}{\sigma_{b_r}^2} + \tfrac{1}{\sigma_{0_r}^2})^{-1}.\end{aligned}$$

   - For $i = 1, \ldots, n$:
     - Draw a new value for c_i using:

       $$c_i|\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{b}_i, \boldsymbol{\Sigma}_b \sim \textstyle\sum_{h=1}^{N} c^\star f(\boldsymbol{b}_i|\boldsymbol{\phi}_h, \boldsymbol{\Sigma}_b)\pi_h\delta_h,$$

       $f \mathrel{\widehat{=}}$ multivariate normal density,

       $c^\star \mathrel{\widehat{=}}$ constant so that the sum of probabilities is 1,
     - Set $\boldsymbol{\theta}_i = \boldsymbol{\phi}_{c_i}$.
   - For $h = 1, \ldots, N$:
     - Draw a new value for V_h (except for $h = N$) using:

       $$V_h|\boldsymbol{c} \sim Be(1 + n_h, \alpha_0 + \textstyle\sum_{l=h+1}^{N} n_l),$$

     - Create π_h using:

       $$\pi_h = V_h \textstyle\prod_{l<h}(1 - V_l).$$

   - Draw a new value $\omega \in \Omega = \{0.5, 0.6, \ldots, 100\}$ for α_0 using:
     - $\alpha_0|\boldsymbol{\pi} \sim \sum_{\omega\in\Omega} \exp\left((N-1)\log(\omega) + (\omega-1)\sum_{h=1}^{N-1}\log(1-V_h)\right) \cdot$
       $\cdot\, P(\alpha_0 = \omega)\,\delta_\omega$.

- For $r = 1, \ldots, p$: Draw new values for μ_{0_r}, $\sigma^2_{0_r}$ and $\sigma^2_{b_r}$ using:
  - $\mu_{0_r}|\sigma^2_{0_r}, \boldsymbol{\theta} \sim N\left((\frac{n}{\sigma^2_{0_r}} + \frac{1}{s^2_{0_r}})^{-1}(\frac{n}{\sigma^2_{0_r}}\bar{\theta}_r + \frac{m_{0_r}}{s^2_{0_r}}), (\frac{n}{\sigma^2_{0_r}} + \frac{1}{s^2_{0_r}})^{-1}\right)$,
  - $\sigma^2_{0_r}|\mu_{0_r}, \boldsymbol{\theta} \sim IG\left(a_0 + 0.5\, n,\ b_0 + 0.5 \sum_{i=1}^{n}(\theta_{i_r} - \mu_{0_r})^2\right)$,
  - $\sigma^2_{b_r}|\boldsymbol{\theta}, \boldsymbol{b} \sim IG\left(a_b + 0.5\, n,\ b_b + 0.5 \sum_{i=1}^{n}(b_{i_r} - \theta_{i_r})^2\right)$.

4. Update the error variance: Draw a new value for σ^2 using:

$$\sigma^2|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}, \boldsymbol{y} \sim IG\left(a_\varepsilon + 0.5\,(\textstyle\sum_{i=1}^{n} n_i), b_\varepsilon + 0.5 \sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{ij} - \mu_{ij})^2\right),$$
$$\mu_{ij} = (\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{\gamma} + \boldsymbol{Z}_i\bar{\boldsymbol{b}} + \boldsymbol{Z}_i\boldsymbol{b}_i)_j.$$

Note that 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 1000 for inference.

3 Simulation study

3.1 Setting

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 actually quite close to the empirical distribution obtained with DP or DPM priors. See for example in Figure 1 the kernel density estimate of the estimated random intercepts in a random slope model for simulated data where the true random effects distribution is a Gaussian mixture. The generation of these data will be explained in this section later on. Note that even in the case of a Gaussian random effects prior the kernel density estimate has a bimodal form. The reason for this 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 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, the number of clusters and the separation between clusters.

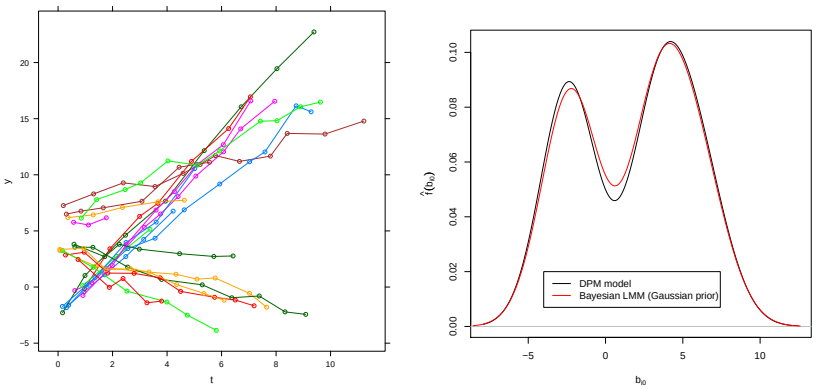


Fig. 1: Trace plot (left) and kernel density estimate of $\hat{b}_{i0}$ with Gauss kernel and bandwidth = 1.793 (optimal in the DPM model) (right) for many individual observations ($\lambda = 5$) with clearly separated clusters.

Since we do not expect the presence or absence of a flexible trend function to have significant impact on this question, we generated data sets assuming a simple linear

trend model

$$y_{ij} = b_0 + \tilde{b}_{i0} + (b_1 + \tilde{b}_{i1})t_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i$$

with i.i.d. errors $\varepsilon_{ij} \sim N(0, \sigma^2)$ and excluding any fixed effects apart from the expectations of the random effects. The i.i.d. random effects $\tilde{\boldsymbol{b}}_i = (\tilde{b}_{i0}, \tilde{b}_{i1})'$ follow a mixture distribution with three Gaussian components:

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

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

$$\sigma^2 = 0.25, \qquad \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_b = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}.$$

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 $X_i \sim Po(\lambda)$. Setting $\lambda = 0.5$ corresponds to longitudinal data with only few (2.5 on average) repeated observations per unit, $\lambda = 2.5$ to a moderate number and $\lambda = 5$ to (comparably) large numbers of repeated 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 $\tilde{\boldsymbol{b}}_i^{(s)}$ are drawn from the Gaussian mixture distribution in each simulation run. For the cluster locations, we chose

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

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*.

Combining these different settings for observations times and clusters results in nine different scenarios. For each of them, we compare mean square errors of estimated random effects obtained from full Bayesian inference based on DPM priors with estimation results based on Gaussian random effects priors, using full Bayesian (MCMC) or empirical Bayesian (REML) inference as implemented in BayesX (Brezger, Kneib & Lang, 2005). In each simulation run s, we compare true parameters with corresponding estimates through squared differences. For random effects, we sum up over the $n = 20$ individual parameters and obtain a sum of squares,

$$SSQ_k(s) = \sum_{i=1}^{n} \left(\hat{b}_{ik}(s) - b_{ik} \right)^2, \quad k = 0, 1$$

for (uncentered) random intercepts and slopes. The empirical distribution of $SSQ_k(s)$ values obtained from simulation run $s = 1, \ldots, 100$ is then represented through box plots. In addition, we also compare estimation of the "fixed effects" b_0 and b_1 (after centering when using DPM priors), estimates $\hat{\sigma}^2$ for the observation variance σ^2 as well as estimated variances $\hat{\sigma}^2_{b_0}$, $\hat{\sigma}^2_{b_1}$ obtained from a Gaussian prior assumption with estimates of variances $\hat{\sigma}^2_{0_0}$, $\hat{\sigma}^2_{0_1}$ of the Gaussian base distribution and of the Gaussian prior $\boldsymbol{b}_i | \boldsymbol{\theta}_i, \boldsymbol{\Sigma}_b \sim N(\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_b)$ for DPM priors through MSE box plots.

In the following, we will summarize results for some scenarios selected from the nine combinations.

3.2 Results

Few individual observations with *clearly* and *moderately separated clusters*

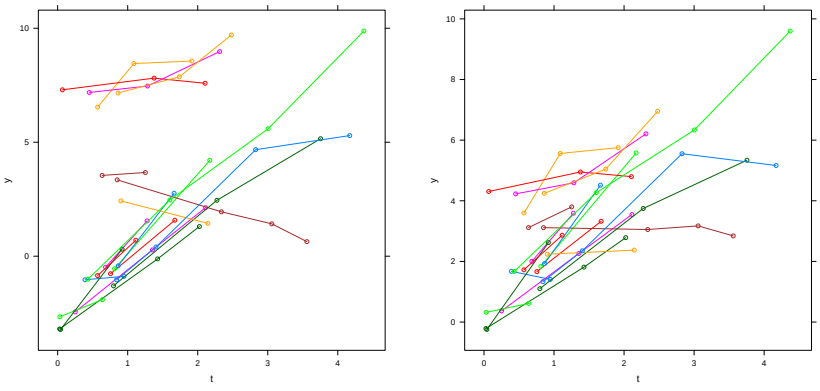
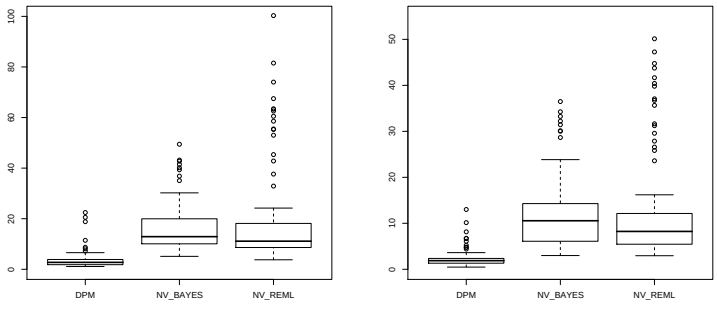
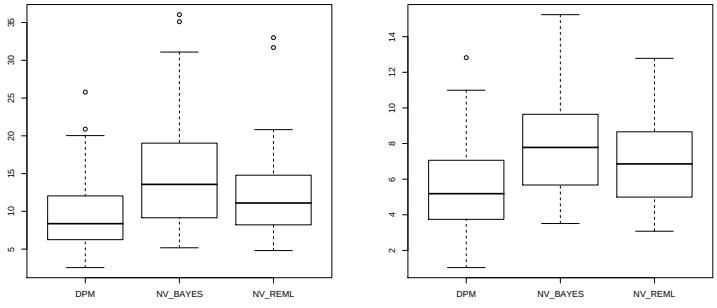


Fig. 2: Trace plots for few individual observations ($\lambda = 0.5$) with clearly (left) and moderately separated clusters (right).

Figure 2 (left) displays a trace plot of typical longitudinal data generated in the setting of clearly separated clusters, showing that cluster effects can easily be detected visually. As we would expect, LMMs with DPM priors substantially improve upon results based on a clearly misspecified Gaussian random effects assumption (Figure 3).



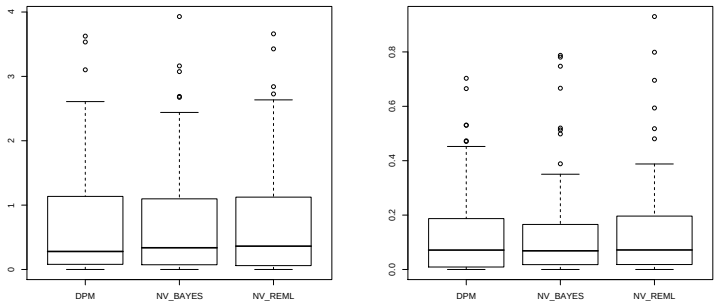
(a) Box plot for $SSQ_0(s)$ (b) Box plot for $SSQ_1(s)$



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

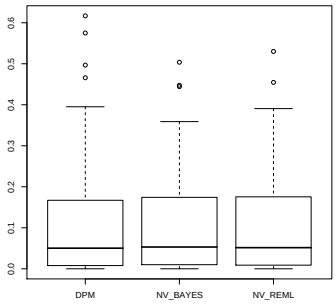
Fig. 3: Box plots for uncentered random intercepts (left) respectively random slopes (right) with clearly (top) and moderately separated clusters (bottom).

As can be seen from the trace plot (Figure 2, right) of typical longitudinal data with moderately separated clusters, it is not immediately evident to recognize that the random curves come from three clusters. So it may be much more tempting to apply a mixed model with Gaussian random effects to analyze such data. Still, the SSQ box plots in Figure 3 demonstrate that DPM priors are superior to Gaussian priors.

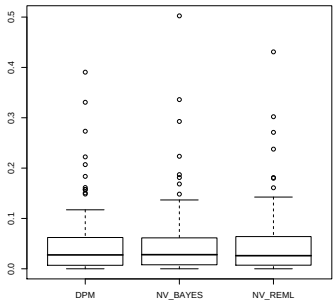


(a) MSE for b_0

(b) MSE for b_1



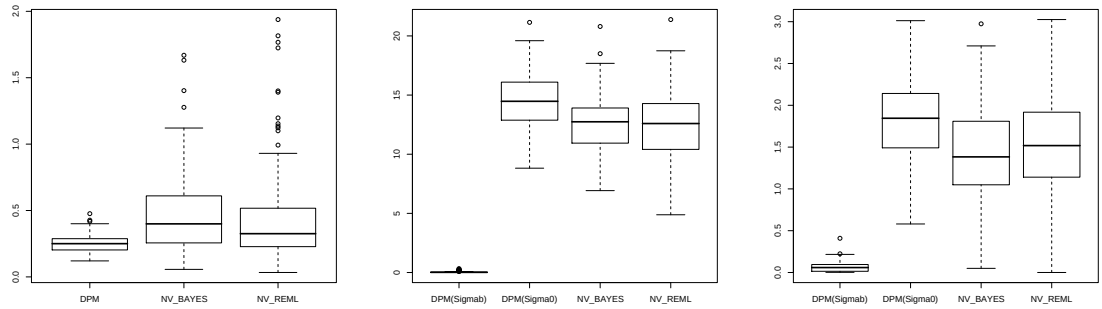
(c) MSE for b_0



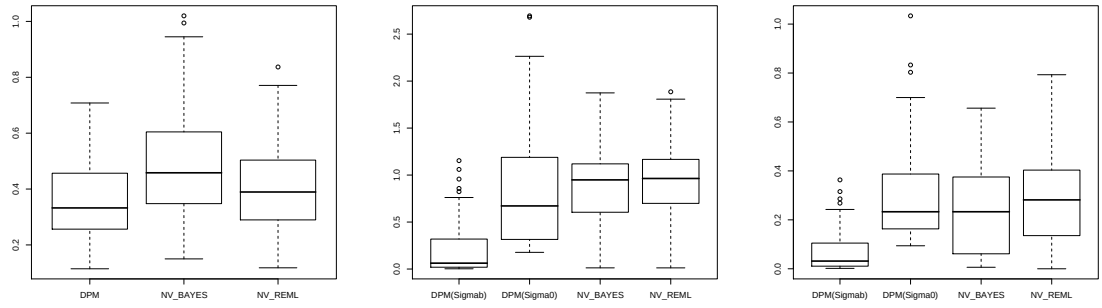
(d) MSE for b_1

Fig. 4: Box plots for fixed effects (intercept (left) and slope (right)) with clearly (top) and moderately separated clusters (bottom).

Figure 4 shows MSE box plots for fixed effects b_0, b_1, confirming that misspecification of the random effects has almost no effect on estimation of fixed effects. Figure 5 displays estimated variances. We can conclude the following: First, estimates for the observation variances are much more precise when using a DPM prior. Second, estimates for the base variances $\sigma^2_{0_0}$, $\sigma^2_{0_1}$ are comparably large for clearly separated clusters, reflecting the fact that the DP parameters $\boldsymbol{\theta}_i$ have high variability. For moderately separated clusters, these estimates have lower MSEs. On the other hand, variation of the random effects $\boldsymbol{b}_i$ around their mean $\boldsymbol{\theta}_i$ is always quite small.



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



(d) $\hat{\sigma}^2$ (e) $\hat{\sigma}^2_{b_0}$ (f) $\hat{\sigma}^2_{b_1}$

Fig. 5: Box plots for the estimated base variance (left), the estimated variances of the random intercept (middle) respectively of the random slope (right) with clearly (top) and moderately separated clusters (bottom).

Moderate number of individual observations with *moderately separated* and *substantially overlapping clusters*

As might be expected from visual inspection of the trace plot (Figure 6), DPM priors still clearly improve upon Gaussian priors for moderately separated clusters, see Figure 7. The superiority of DPM priors is almost lost − in terms of SSQs −, however, for substantially overlapping clusters, where it is already almost impossible to detect heterogeneity visually from the trace plot.

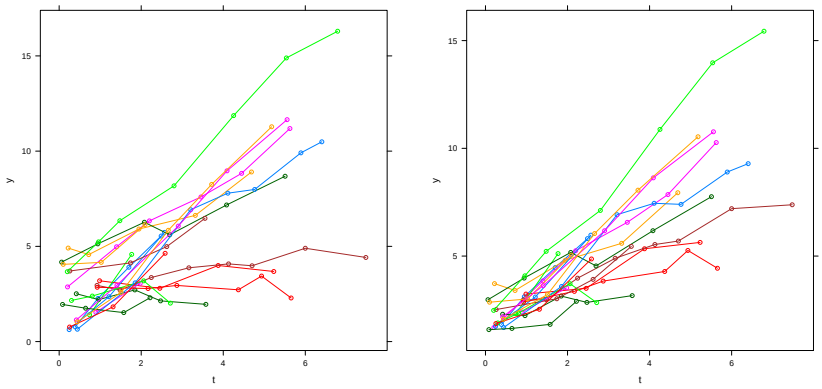
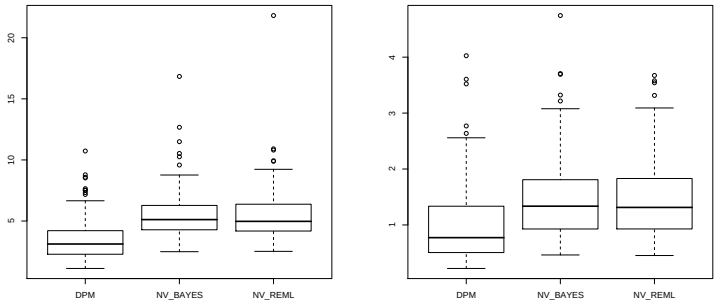
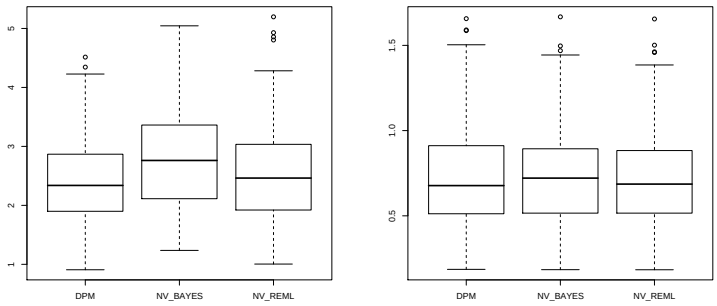


Fig. 6: Trace plots for moderate number of individual observations ($\lambda = 2.5$) with moderately separated (left) and substantially overlapping clusters (right).



(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 $SSQ_1(s)$

Fig. 7: Box plots for uncentered random intercepts (left) respectively random slopes (right) with moderately separated (top) and substantially overlapping clusters (bottom).

Many individual observations with *moderately separated* and *substantially overlapping clusters*

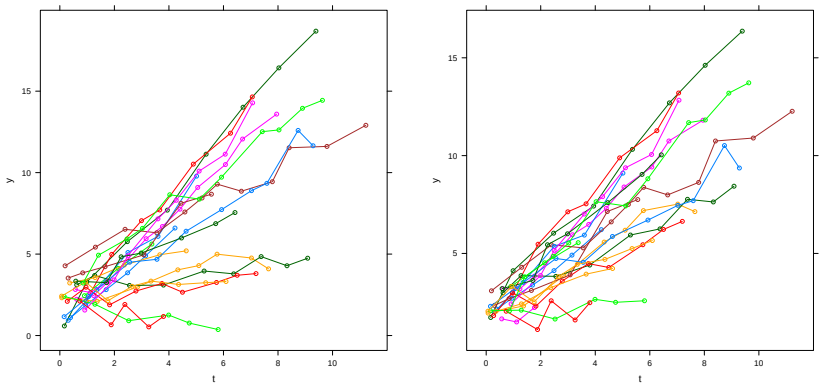
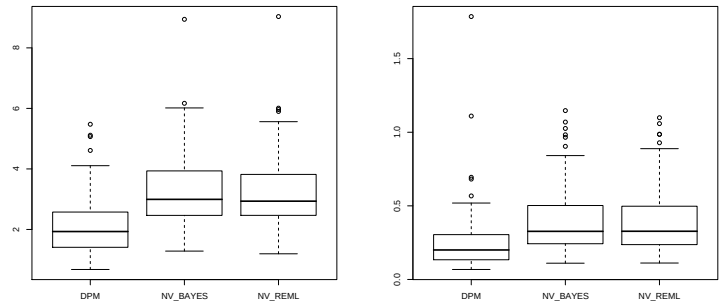
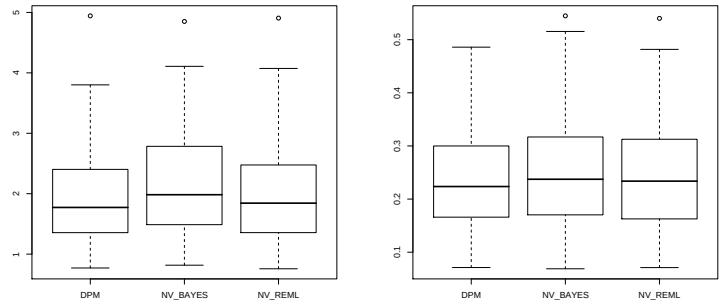


Fig. 8: Trace plots for many individual observations ($\lambda = 5$) with moderately separated (left) and substantially overlapping clusters (right).

For data with comparably many individual observations, improvement of DPM priors relative to Gaussian priors tends to become smaller. For clearly or moderately separated clusters, SSQ plots (only shown for moderately separated clusters in Figure 9 (top)) still indicate better estimation properties. For substantially overlapping clusters (see Figure 8 (right) for a typical trace plot) it becomes quite difficult to detect heterogeneity caused by clusters through visual inspection.



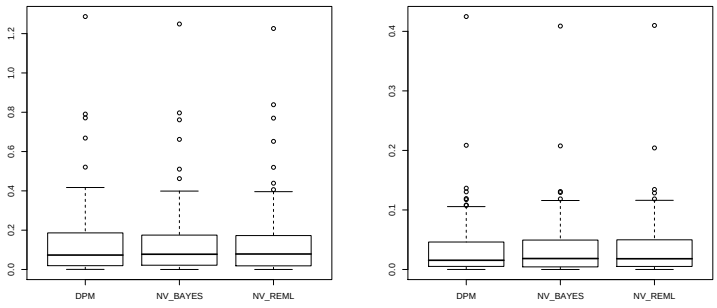
(a) Box plot for $SSQ_0(s)$ (b) Box plot for $SSQ_1(s)$



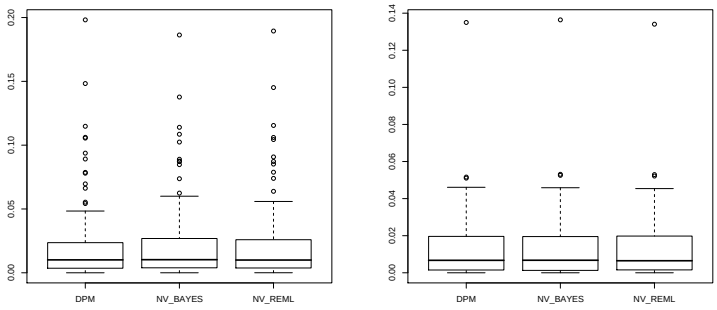
(c) Box plot for $SSQ_0(s)$ (d) Box plot for $SSQ_1(s)$

Fig. 9: Box plots for uncentered random intercepts (left) respectively random slopes (right) with moderately separated (top) and substantially overlapping clusters (bottom).

The SSQ box plots in Figure 9 (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.



(a) MSE for b_0 (b) MSE for b_1

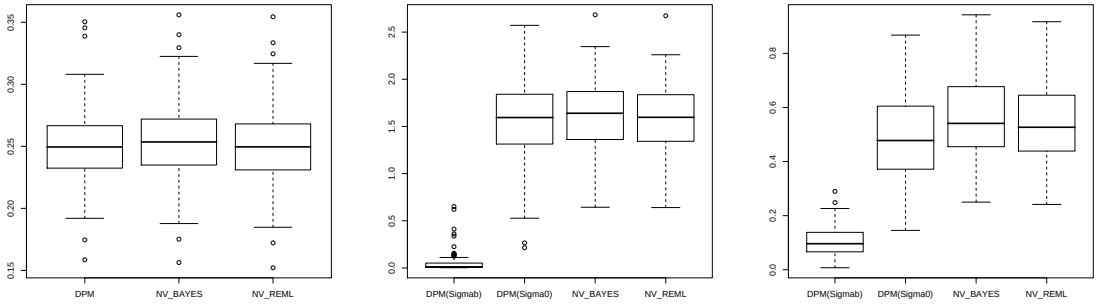


(c) MSE for b_0 (d) MSE for b_1

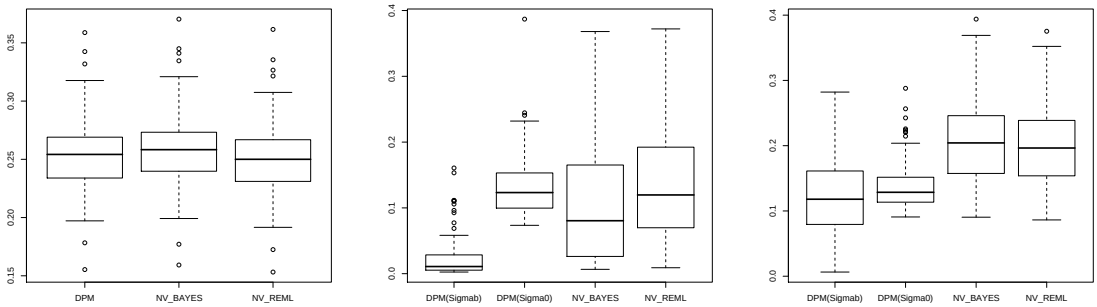
Fig. 10: Box plots for fixed effects (intercept (left) and slope (right)) with moderately separated (top) and substantially overlapping clusters (bottom).

MSE box plots for estimated fixed effects confirm again robustness with regard to the random effects prior (Figure 10). Even estimation of the observation variance σ^2 is now quite robust against misspecification (Figure 11).

In summary, we draw the following conclusions: The DPM priors yield the better estimates for random effects − in terms of SSQs − the clearer the clusters differ and the less observations are in the data. For the estimates of the fixed intercept and slope there is no improvement of the DPM priors over Gaussian random effects in any case. For few observations the error variance is lower in the DPM model than in the models with Gaussian random effects prior.



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



(d) $\hat{\sigma}^2$ (e) $\hat{\sigma}^2_{b_0}$ (f) $\hat{\sigma}^2_{b_1}$

Fig. 11: Box plots for the estimated base variance (left), the estimated variances of the random intercept (middle) respectively of the random slope (right) with moderately separated (top) and substantially overlapping clusters (bottom).

4 Application: Childhood Obesity

Obesity, and more specifically obesity among children, has become a major public health issue in industrialized countries. In the following, we investigate longitudinal data, where the BMI of children is the response variable, measured repeatedly over time. Particular interest is on the effect of age on BMI, adjusted for covariates. We apply the AMM (1) to data from the LISA (Influences of **L**ife-style factors on the development of the **I**mmune **S**ystem and **A**llergies in East and West Germany) study with intent to detect clusters in the data. The LISA study is a prospective birth cohort study in four cities in Germany (Bad Honnef, Leipzig, Munich, Wesel), including 3097 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.

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-feed and/or breast-feeding	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

Table 1: Description of the categorial covariates (related to 2043 children).

Tables 1 and 2 give 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 autocorrelation in the samples.

The effect of age on the BMI is of particular interest to us. Figure 12 (left) shows individual BMI patterns by age for twelve randomly selected children. Here as well as in Figure 12 (right) a nonlinear trend of age is obvious.

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 2: Description of the continuous covariates (related to 2043 children).

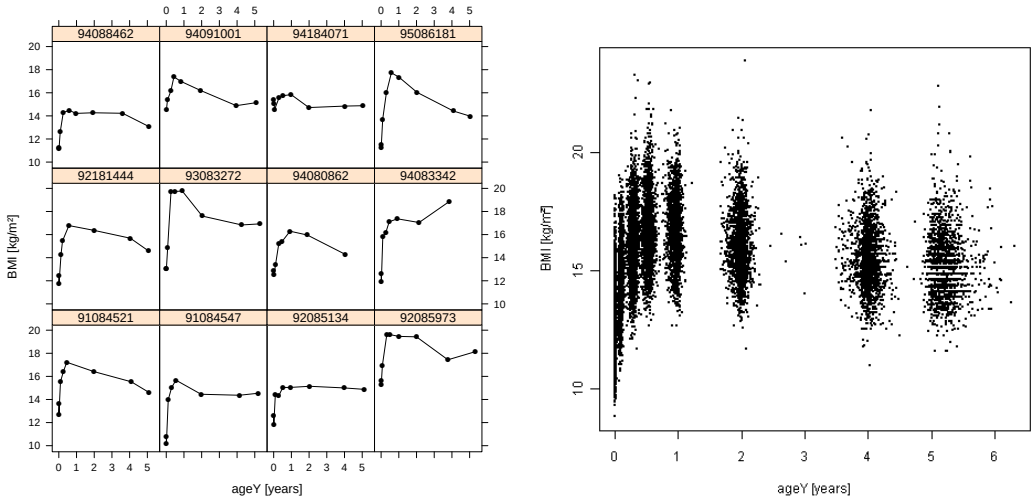


Fig. 12: BMI against time: trace plots (left) for twelve randomly selected children and a scatter plot for all children (right).

To fit a smooth age trend we use a P-spline to penalize the roughness of the fit. Apart from the general effect of the covariate ageY, we are interested in individual deviations from this trend. For this purpose we assume the AMM (1) with $\boldsymbol{z}_{ij}'\boldsymbol{b}_i = b_{0i} + t_{ij}b_{1i}$. Figure 13 visualizes the fit of this model. Here one can see the general time trend (red color) as well as individual fits for four selected subjects. The measurements of these subjects show different peculiar patterns. A person (id = 92189214, orange color) features very low values of BMI where for an another subject (id = 95089461, purple color) there is a nontypical gain of BMI after the age of one year. We find out that the model responds to this features sufficiently. The individual with id = 94182011 (light blue color) shows an extremely high value in the age of two years that probably would be an error measure. However, the model doesn't really react to that fact. Furthermore, another subject (id = 92185191, green color) has a distinct apex in the age of six months. We recognize that this apex is not adequately detected by the model. So we extend the model by an additional random effect as in (2) with

$$h(t) = \frac{\log(t+1)}{(t+1)^2} \tag{4}$$

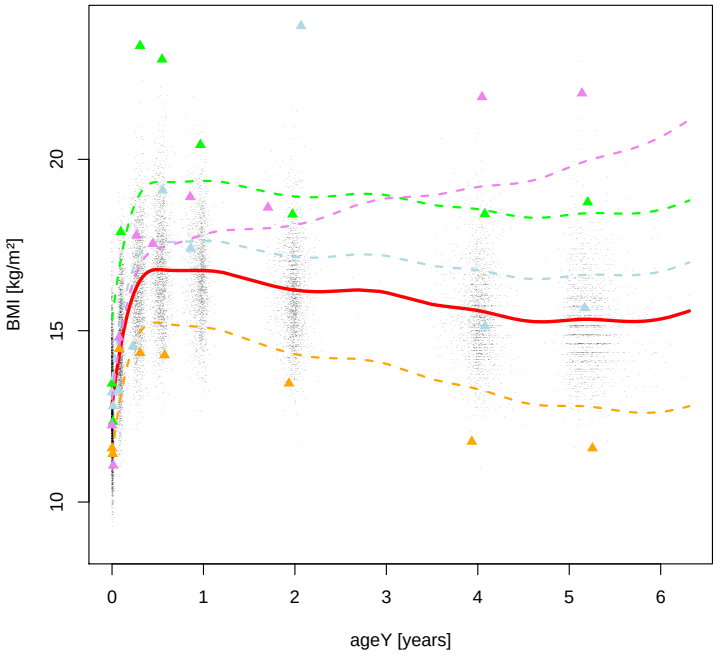


Fig. 13: Fit of AMM (1) with $\boldsymbol{z}_{ij}'\boldsymbol{b}_i = b_{0i} + t_{ij}b_{1i}$. The red line shows the general effect of age while the dashed lines show individual effects: orange color (id = 92189214), purple color (id = 95089461), light blue color (id = 94182011) and green color (id = 92185191).

that is able to fit the special pattern of the LISA data in a convincing way. Sabanés Bové (2009) showed that the goodness of fit can be greatly improved by (4). Indeed, the fit looks very much better (see Figure 14).

Table 3 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 especially for the duration of breastfeeding since some experts suppose that exclusive breastfeeding causes lower BMI values. However, Table 3 does not confirm this general negative effect of breastfeeding on the BMI. In the following, we can get more information about this relation and can see behind the curtain by looking for clusters in the data.

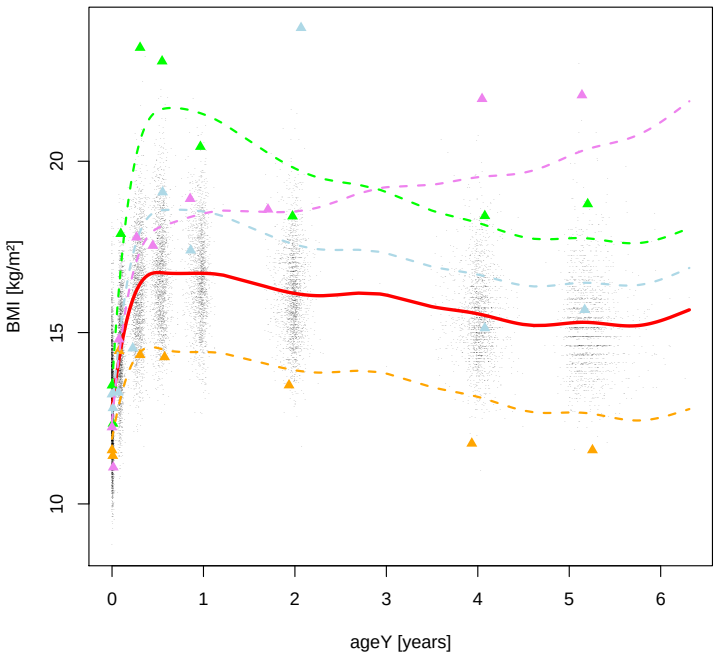


Fig. 14: Fit of AMM (1) with $\boldsymbol{z}_{ij}'\boldsymbol{b}_i = b_{0i} + t_{ij}b_{1i} + \frac{\log(t_{ij}+1)}{(t_{ij}+1)^2}b_{2i}$. The red line shows the general effect of age while the dashed lines show individual effects: orange color (id = 92189214), purple color (id = 95089461), light blue color (id = 94182011) and green color (id = 92185191).

	Mean	**Median**	**Sd**	**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 3: Estimators for the error variance and the fixed effects.

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 we can combine these to an universal clustering? Diverse operations exist to handle this (see for example Fritsch & Ickstadt (2009)), but concerning the high number of subjects and hence the high number of possible clusterings, these methods are either not feasible or not fully convincing. So we pursue an alternative strategy: First, we get an estimated number of clusters using the median for all numbers of cluster in the MCMC iterations. Second, we allocate the $\hat{\boldsymbol{\theta}}_i$ for $i = 1, \ldots, n$ to that estimated number of clusters by k-means clustering. In addition, we do this for $\hat{\boldsymbol{b}}_i$ with $i = 1, \ldots, n$, too, but primarily we are interested in the $\hat{\boldsymbol{\theta}}_i$ clustering, because this is the level the Dirichlet process clustering originally happens on.

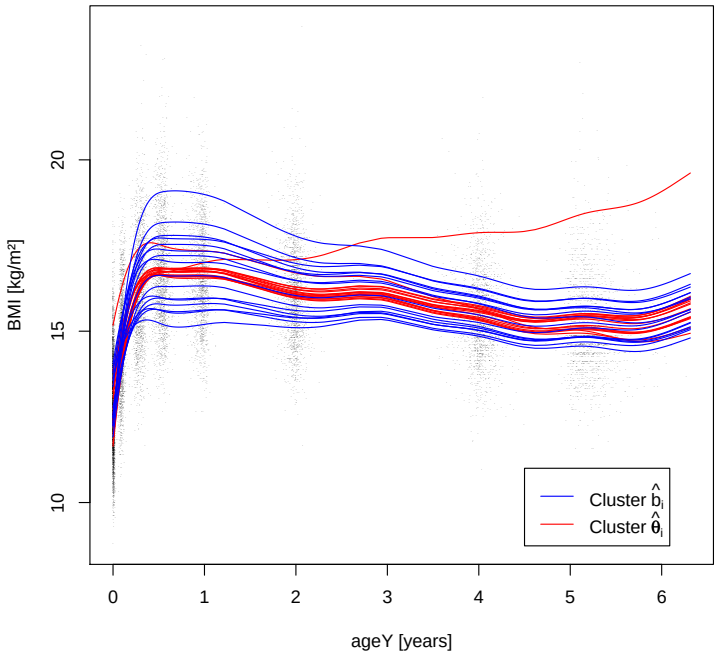
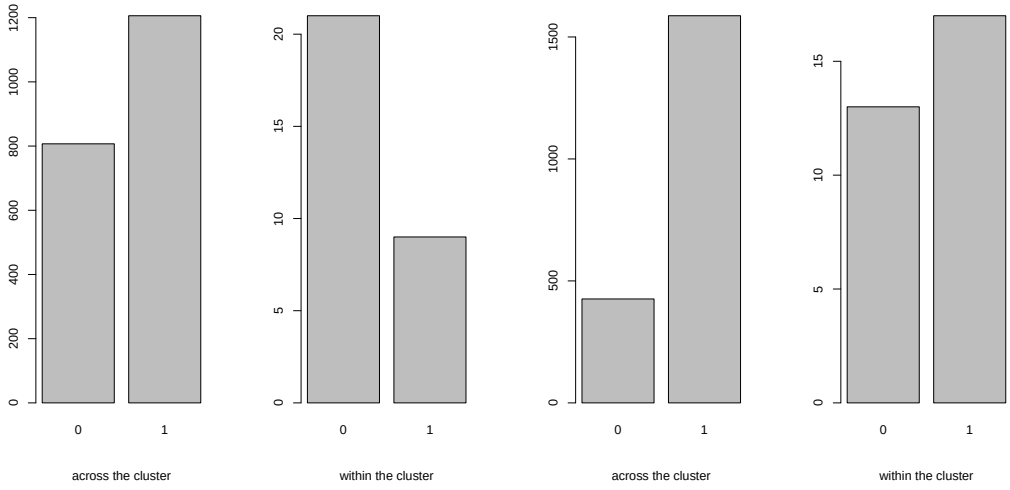


Fig. 15: Clustering of $\hat{\boldsymbol{\theta}}_i$ respectively $\hat{\boldsymbol{b}}_i$ for AMM (1) with $\boldsymbol{z}_{ij}'\boldsymbol{b}_i = b_{0i} + t_{ij}b_{1i} + \frac{\log(t_{ij}+1)}{(t_{ij}+1)^2}b_{2i}$

In Figure 15, one sees 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 attracts attention that differs from all other clusters: For most of the children the BMI steeply increases until the age of about six months and then decreases slowly. In contrast, there is a group for which the development of BMI is very similar to the others only during the first year while a permanent ascent of the BMI is observed afterwards. It is remarkable that this cluster could not be detected for $\hat{\boldsymbol{b}}_i$. More specifically, there are thirty persons which belong

to that special cluster, including the individual with id = 95089461 (purple color). It is of particular interest if there are meaningful differences in the values of covariates between these subjects and the others. Indeed, especially the covariates breast and area show noticeable differences (see Figure 16).



(a) breast (b) area

Fig. 16: 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).

Obviously, most of the children in the extreme cluster were bottlefed and/or breastfed until the age of four months. By contrast, across that cluster the majority of children were breastfed only. So in this sense breastfeeding only is essentially an indicator for a normal and lower development of the BMI like some experts expect, although there is no general negative effect of the covariate breast. Furthermore, one recognizes that the ratio of children living in an urban area to a rural area is nearly balanced in the extreme cluster where most of the other children live in cities.

In summary, using DPM priors for random effects 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 the data. In general, this knowledge can be used to detect indicators for this pattern.

5 Summary

The semiparametric mixed model considered in this paper combines the advantages of Bayesian smoothing of nonlinear time trends as well as of other nonlinear covariate effects, with the flexibility of DPM priors 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, DP and DPM priors allow to detect hidden clusters in the data.

Development of Gibbs samplers for Gaussian AMMs with more complicated structured additive predictors as in Fahrmeir, Kneib & Lang (2004) is conceptually straightforward due to the modular hierarchical structure. Extensions to models with non-Gaussian responses require additional computational effort, involving hybrid Metropolis Hastings algorithms.

Acknowledgment: 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 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 (4) for modeling BMI profiles.

References

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

Brezger, A. & Lang, S. (2006). Generalized additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis*, **50**, 967−991.

Dunson, D., Yang, M. & Baird, D. (2007). Semiparametric Bayes hierarchical models with mean and variance constraints. Technical Report, Dep. of Statistical Science, Duke University.

Durban, M., Harezlak, J., Wand, M. P., and Carroll, R. J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, **24**, 1153−1167.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89−121.

Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, **14**, 731−761.

Fahrmeir, L. & Lang, S. (2001). Bayesian semiparametric regression analysis of multicategorical time-space data. *Annals of the Institute of Statistical Mathematics*, **53**, 11−30.

Fenske, N., Fahrmeir, L., Rzehak, P. & Höhle, M. (2008). *Detection of risk factors for obesity in early childhood with quantile regression methods for longitudinal data*. Technical Report, **38**, Department of Statistics, 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.

Fritsch, A. & Ickstadt, K. (2009). Improved criteria for clustering based on the posterior similarity matrix. *International Society for Bayesian Analysis*, **4**, 367−392.

Heinzl, F. (2009). *Nonparametrische Bayes-Inferenz in additiven gemischten Modellen*, Diploma Thesis, Department of Statistics, Ludwig-Maximilians-University Munich.

Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161−173.

Ishwaran, H. & James, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics*, **11**, 508−532.

Jara, A. (2007). Applied Bayesian Non- and Semiparametric Inference using DPpackage. *R News*, **3**, 17−26.

Jullion, A. & Lambert, P. (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 Ibrahim, J. (1998). A Semiparametric Bayesian Approach to the Random Effects Model. *Biometrics*, **54**, 921−938.

Lang, S. & Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**, 183−212.

Li, Y., Lin, X. & Müller, P. (2009). Bayesian inference in semiparametric mixed models for longitudinal data. *Biometrics*, to appear.

Lin, X. & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society B*, **61**, 381−400.

Panagiotelis, A. & Smith, M. (2008). Bayesian identification, selection and estimation of semiparametric functions in high-dimensional additive models. *Journal of Econometrics*, **143**, 291−316.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge University Press, Cambridge.

Sabanés Bové, D. (2009). *Predictive Assessment of Bayesian Hierarchical Models*, Master Thesis, Department of Statistics, Ludwig-Maximilians-University Munich.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639−650.