LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Felix Heinzl & Gerhard Tutz

Clustering in Additive Mixed Models with Approximate Dirichlet Process Mixtures using the EM Algorithm

Technical Report Number 138, 2013
Department of Statistics
University of Munich

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

Clustering in Additive Mixed Models with Approximate Dirichlet Process Mixtures using the EM Algorithm

Felix Heinzl[1] and **Gerhard Tutz**[1]

[1] Department of Statistics, Ludwig-Maximilians-University Munich, Akademiestr. 1, 80799 Munich, Germany

February 14, 2013

SUMMARY: We consider additive mixed models for longitudinal data with a nonlinear time trend. As random effects distribution an approximate Dirichlet process mixture is proposed that is based on the truncated version of the stick breaking presentation of the Dirichlet process and provides a Gaussian mixture with a data driven choice of the number of mixture components. The main advantage of the specification is its ability to identify clusters of subjects with a similar random effects structure. For the estimation of the trend curve the mixed model representation of penalized splines is used. An Expectation-Maximization algorithm is given that solves the estimation problem and that exhibits advantages over Markov chain Monte Carlo approaches, which are typically used when modeling with Dirichlet processes. The method is evaluated in a simulation study and applied to body mass index profiles of children.

KEY WORDS: *Additive mixed models; Dirichlet process mixture; EM algorithm; penalized splines; stick breaking*

1 Introduction

For the modeling of longitudinal data with a nonlinear time trend, additive mixed models are an useful tool. The model considered in this paper assumes a nonparametric term for the variation over time and parametric terms for the random effects and the fixed effects of other covariates. Due to this combination of nonparametric and parametric terms the model is called *semiparametric mixed model*. More concretely, the conditional distribution of the response y_{ij} observed for subject i at observation time t_{ij} is given by

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

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 random effects $\boldsymbol{b}_i$ and the 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 paper combines an approximate Dirichlet process mixture (DPM) for the random effects with a penalized spline (P-spline) for approximating the trend function $f(\cdot)$ and uses an Expectation-Maximization (EM) algorithm for estimation. In model (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 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). The most innovative aspect of our method is that we introduce an EM algorithm for inference instead of the popular Markov chain Monte Carlo (MCMC) methods, which are used, for example, in Li et al. (2010) and Heinzl et al. (2012) in semiparametric mixed models. 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 Heinzl and Tutz (2013), 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 paper for clustering nonlinear longitudinal data.

The paper is organized as follows: In Section 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 3 compares our approach to the MCMC-method in Heinzl et al. (2012) and to additive mixed models with normally distributed random effects. In Section 4, the theophylline data and body mass index (BMI) profiles of children are analyzed.

2 Additive Mixed Models with Dirichlet Process Mixtures

2.1 Model Hierarchy

Let the time trend in model (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 $\boldsymbol{\gamma}$. 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 2.3 for more details of the decomposition and other properties of 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 (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:

$$\begin{aligned} \boldsymbol{b}_i | \boldsymbol{\theta}_i &\overset{ind.}{\sim} N(\boldsymbol{\theta}_i, \boldsymbol{D}), \qquad i = 1, \ldots, n, \\ \boldsymbol{\theta}_i | G &\overset{i.i.d.}{\sim} G, \qquad i = 1, \ldots, n, \\ G &\sim DP(\alpha, G_0). \end{aligned} \tag{2}$$

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 (2) can be rewritten as

$$\begin{array}{rcll} \boldsymbol{b}_i|\boldsymbol{v} & \overset{i.i.d.}{\sim} & \sum_{h=1}^{N} \pi_h \, N(\boldsymbol{\mu}_h, \boldsymbol{D}), & i = 1, \ldots, n, \\ \pi_h & = & v_h \prod_{l<h}(1 - v_l), & h = 1, \ldots, N, \\ v_h & \overset{i.i.d.}{\sim} & Be(1, \alpha), & h = 1, \ldots, N-1, \end{array} \tag{3}$$

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 a recommendation how to choose N. As customary, in this context two constraints have to hold: $\sum_{h=1}^{N} \pi_h \boldsymbol{\mu}_h = \mathbf{0}$ 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 $v_N = 1$.

2.2 Inference

In what follows, an EM algorithm for the additive mixed model described in Section 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 $\boldsymbol{w}_i$

$$\begin{array}{rcll} \boldsymbol{y}_i|\boldsymbol{w}_i, \boldsymbol{\gamma}_p & \overset{ind.}{\sim} & N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{T}\boldsymbol{\gamma}_0 + \boldsymbol{B}_i\boldsymbol{W}\boldsymbol{\gamma}_p + \boldsymbol{Z}_i\boldsymbol{\mu}_h, \, \boldsymbol{V}_i), & i = 1, \ldots, n, \\ \boldsymbol{w}_i|\boldsymbol{v} & \overset{i.i.d.}{\sim} & M(1, \boldsymbol{\pi}), & i = 1, \ldots, n, \\ v_h & \overset{i.i.d.}{\sim} & Be(1, \alpha), & h = 1, \ldots, N-1, \\ \boldsymbol{\gamma}_p & \sim & N(\mathbf{0}, \tau^2 \boldsymbol{I}_{d-k}), & \end{array} \tag{4}$$

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 (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 distribu-

tions 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 (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 (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. Instead of maximizing the penalized incomplete likelihood function

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

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

$$v_h = \frac{\sum_{i=1}^{n}\pi_{ih}}{\sum_{i=1}^{n}\sum_{l=h}^{N}\pi_{il} + \alpha - 1}, \qquad h = 1, \ldots, N-1, \tag{5}$$

and

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

Without further restrictions it could happen that $v_h \notin [0,1]$ if $\alpha \in (0,1)$. For preventing this we use the following correction approach: Update v_h by (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: $\sum_{l=1}^{h} \pi_l^\circ > 1$. 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. As proposed by Ohlssen et al. (2007) 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

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

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

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

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) 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: 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 Heinzl and Tutz (2013). Note that after convergence all parameters have to be restandardized internally because the algorithm works with standardized variables.

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 2.2. First, note that the decomposition of the basis coefficient vector mentioned in Section 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

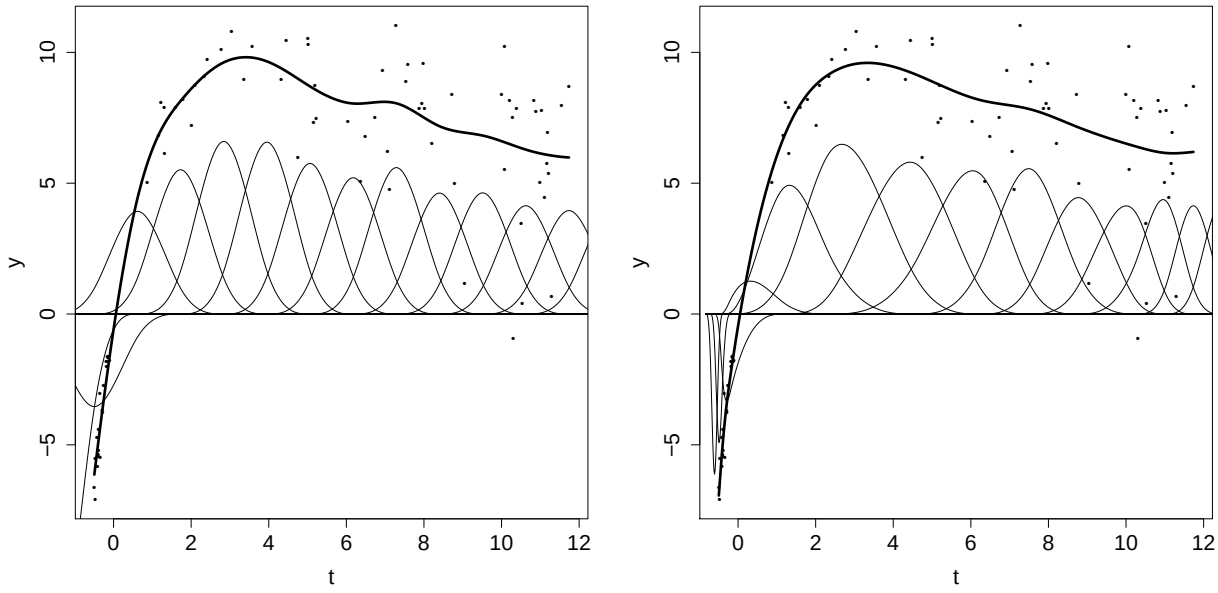
$$\boldsymbol{K} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T = \begin{pmatrix} \boldsymbol{\Gamma}_p & \boldsymbol{\Gamma}_0 \end{pmatrix}\begin{pmatrix} \boldsymbol{\Omega}_p & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}\begin{pmatrix} \boldsymbol{\Gamma}_p^T \\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

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

$$\boldsymbol{T} = \begin{pmatrix} 1 & \varsigma_1 & \dots & \varsigma_1^{k-1} \\ \vdots & & & \vdots \\ 1 & \varsigma_d & \dots & \varsigma_d^{k-1} \end{pmatrix} \quad \text{and} \quad \boldsymbol{W} = \boldsymbol{\Delta}^T(\boldsymbol{\Delta}\boldsymbol{\Delta}^T)^{-1}, \tag{6}$$

is also suitable, where $\varsigma_1, \dots, \varsigma_d$ are equidistant grid points on the relevant range. 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) 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.

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 3.1. The corresponding trace plot is shown later in Figure 7 (top left). In Figure 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



(a) Equidistant knots

(b) Knots based on quantiles

Figure 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 $\hat{\gamma}_s B_s^l(t)$.

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

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 of a competing MCMC approach for fitting additive mixed models with DPMs are compared.

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. 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 2 with a DPM as random effects distribution estimated via EM algorithm (DPM-EM) and compare it to the corresponding MCMC-approach from Heinzl et al. (2012) (DPM-MCMC). In addition, based on the considerations in Section 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

$$PE(s) = \frac{1}{n} \sum_{i=1}^{n} \int_{-0.5}^{12} \left(\hat{f}_{is}(t) - f_{is}(t) \right)^2 dt, \tag{7}$$

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 (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.

3.2 Results

Clearly separated clusters

In Figure 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 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 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 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

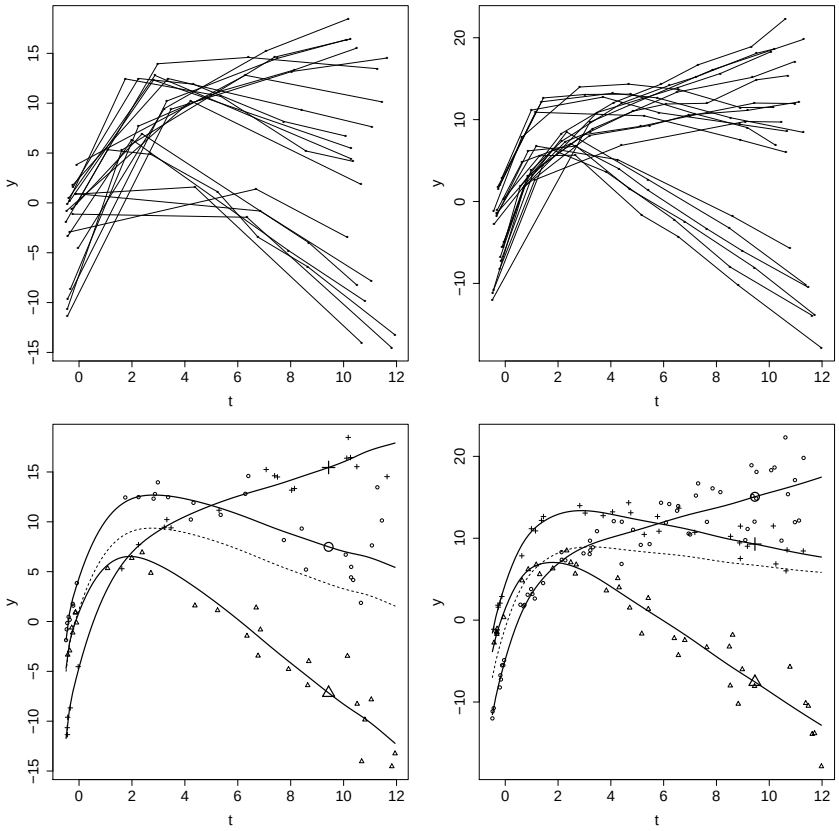
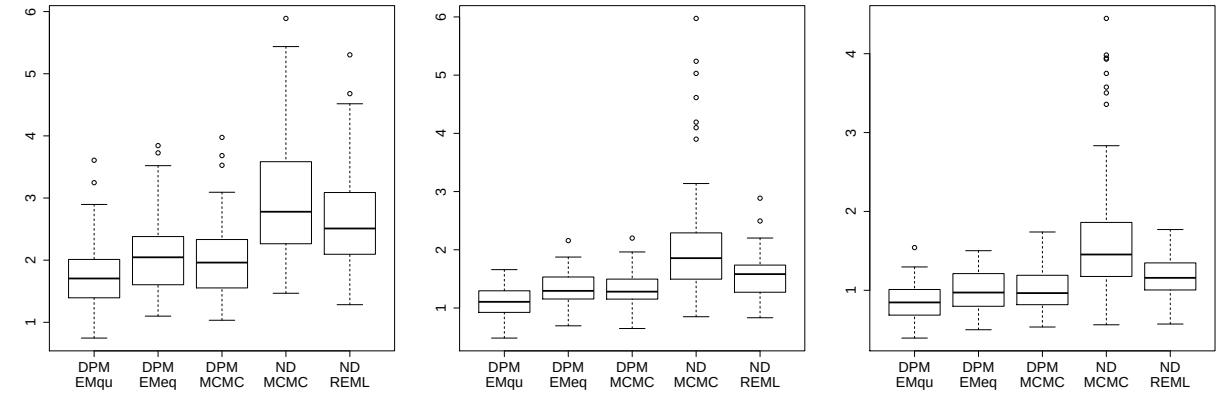


Figure 2: Trace plots (top) and clustering by the DPM-EM approach with knots based on quantiles (below) with clearly separated clusters for few individual observations ($\nu = 1$) (left) and a medium number of individual observations ($\nu = 3$) (right).

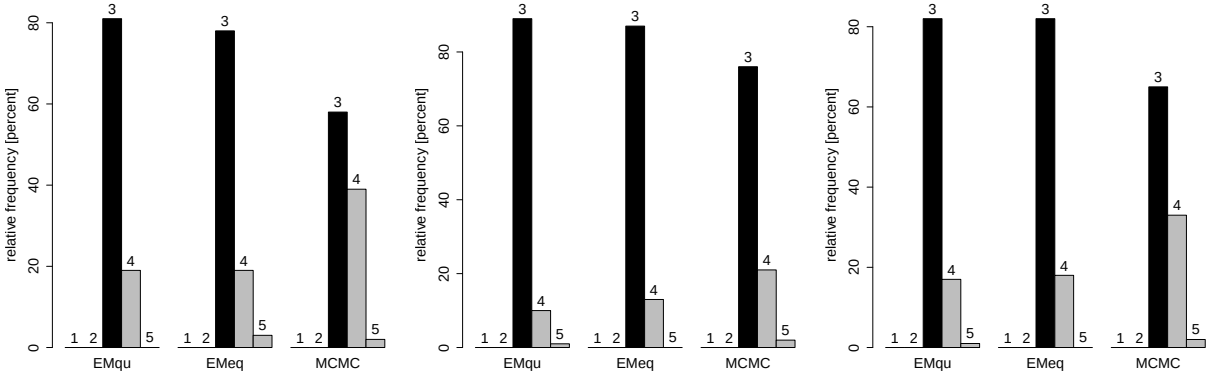
the estimation procedure. Using knots based on quantiles (DPM-EMqu), the prediction accuracy can even be improved.

The clustering related characteristics are shown in Figure 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.



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

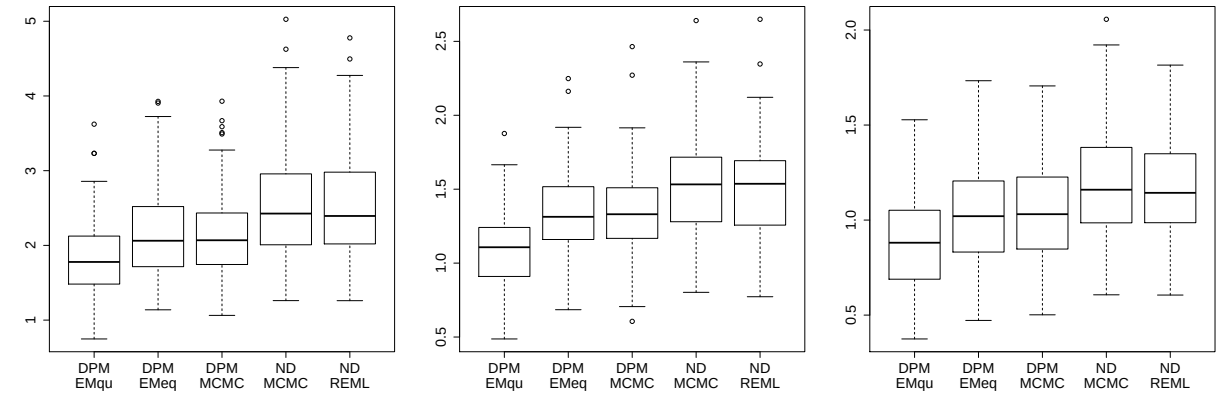
Figure 3: Box plots of PE with clearly separated clusters for few individual observations (left), a medium number of individual observations (middle) and many individual observations (right).



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

Figure 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).

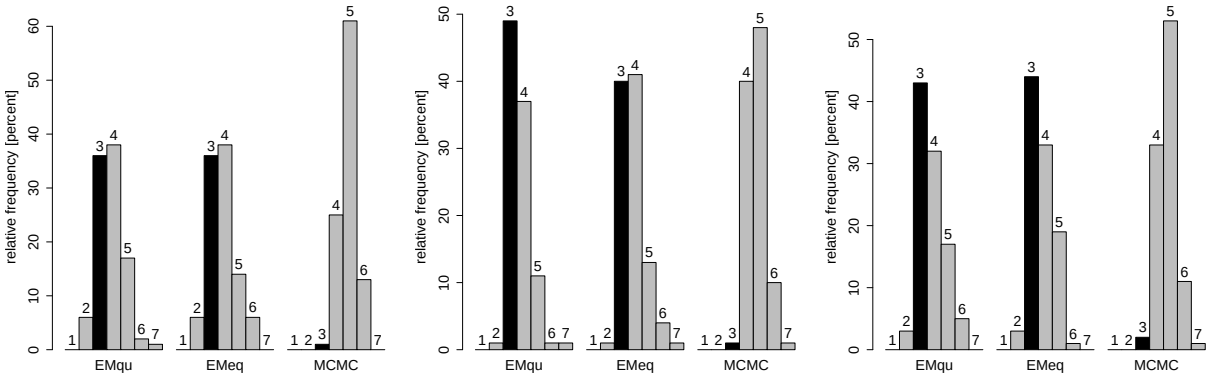
Moderately separated clusters



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

Figure 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).

For a smaller separation of the cluster centers the DPM approaches still outperform the classical mixed models with normally distributed random effects (Figure 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).



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

Figure 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).

In Figure 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

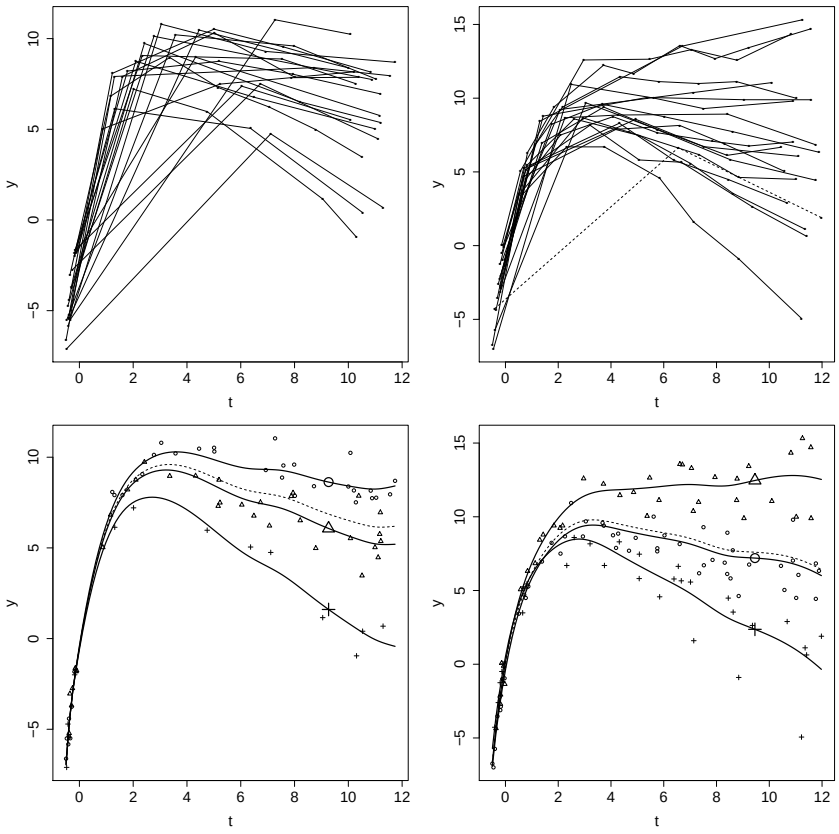
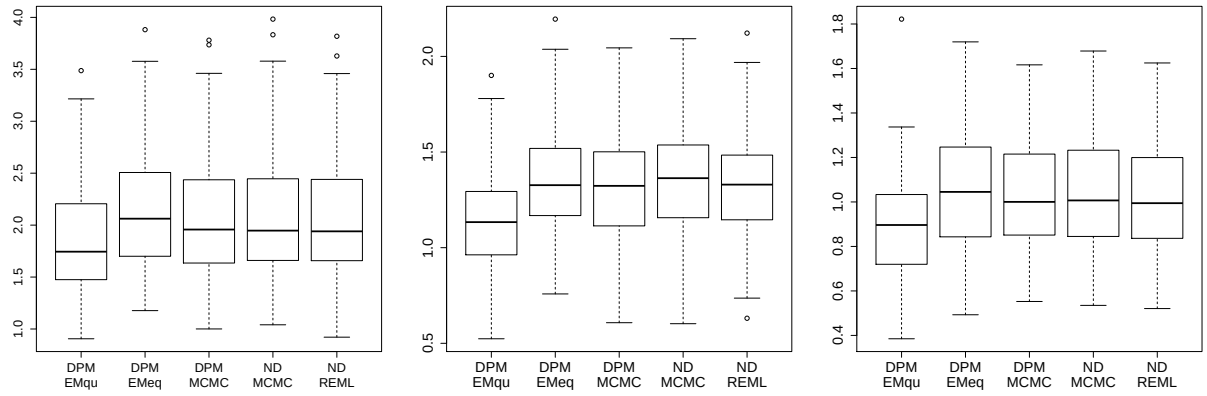


Figure 7: Trace plots (top) and clustering by the DPM-EM approach with knots based on quantiles (below) with substantially overlapping clusters for few individual observations ($\nu = 1$) (left) and a medium number of individual observations ($\nu = 3$) (right).

In the scenario of substantially overlapping clusters we pick up the example of Section 2.3 for few individual observations. See Figure 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 7, top right) three clusters are found by our DPM-EM approach (below right), too. Let regard these plots in more detail. In Figure 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.



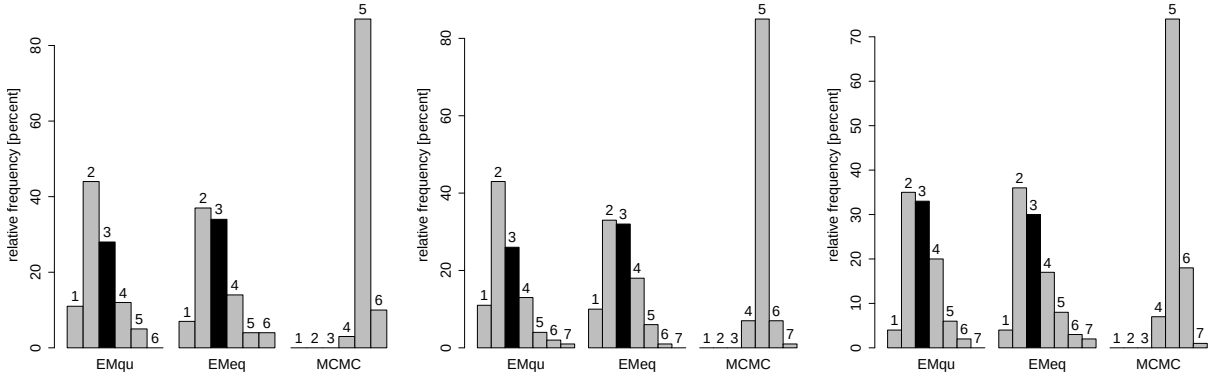
(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

Figure 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 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 1. For the DPM-EM approach with knots based on quantiles, however, the best performance can be observed. See Section 2.3 for a discussion about the choice of knots.

According to Figure 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 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).

4 Applications

4.1 Theophylline

In the following, the approach introduced in Section 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 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). See Section 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

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.

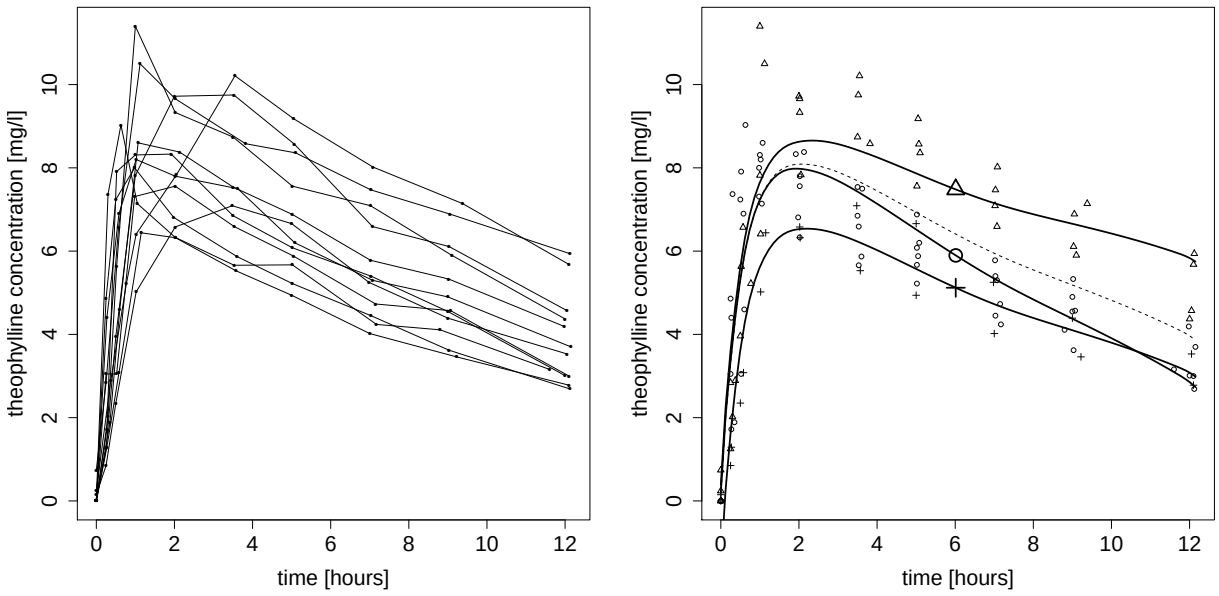


Figure 10: Theophylline concentration in the sample across time: raw data (left) and clustering by the DPM-EM approach (right). On the right observations belonging to the same cluster are marked with the same symbol. The dashed line represents the population effect, the solid lines symbolize the cluster 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 1: Estimation results for the fixed effects and variance parameters by the DPM-EM approach for the theophylline data.

Table 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 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 10, right).

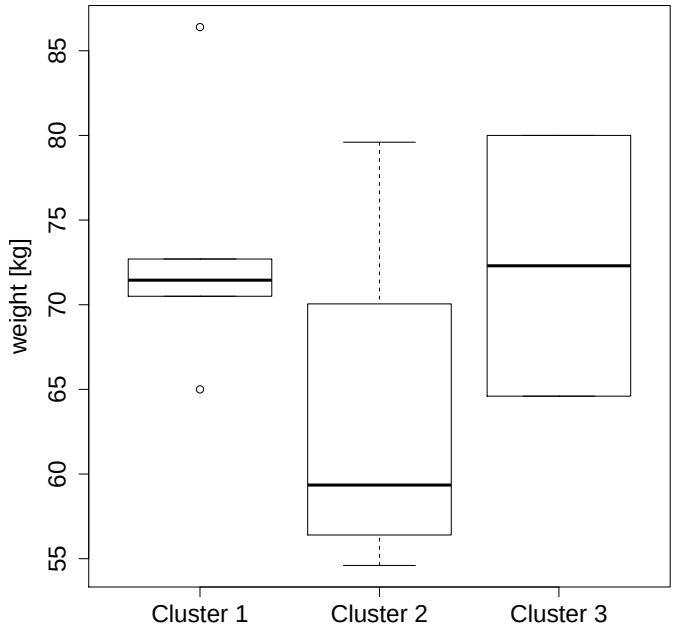


Figure 11: Distribution of the variable `weight` in the three clusters corresponding to the clustering by the DPM-EM approach.

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 3097 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 12 (left) shows the development of the BMI for

twelve randomly selected children, while in Figure 12 (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.

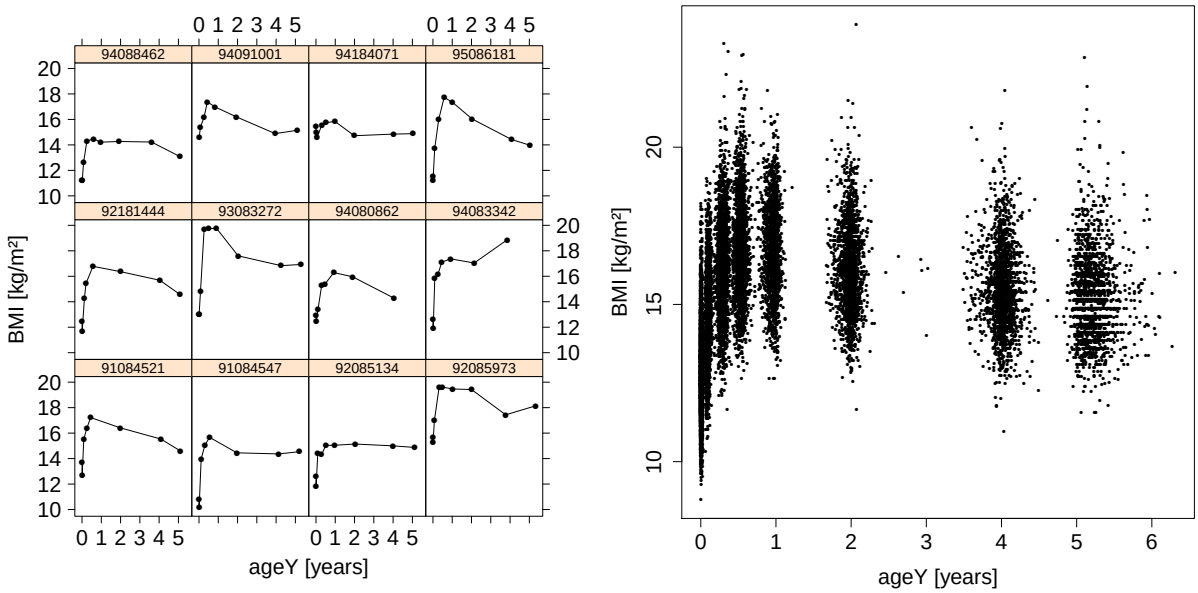


Figure 12: BMI against age: trace plots (left) for twelve randomly selected children and a scatter plot for all children (right) of the LISA data (Heinzl et al., 2012).

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 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 2.2, we truncate the Dirichlet process at $N = 11$. See Table 2 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

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

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 2: Description of the used categorial and continuous covariates of the LISA data with 2043 children (Heinzl et al., 2012).

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 3) we obtain the same significant predictors as in Heinzl et al. (2012) 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 $\hat{\beta}_r / \widehat{sd}(\hat{\beta}_r)$, 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. 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.

Figure 13 shows that five clusters are detected by the DPM-EM model, which was not obvious when looking at the raw data in Figure 12. Note that the concentration

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

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. 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 14, 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.

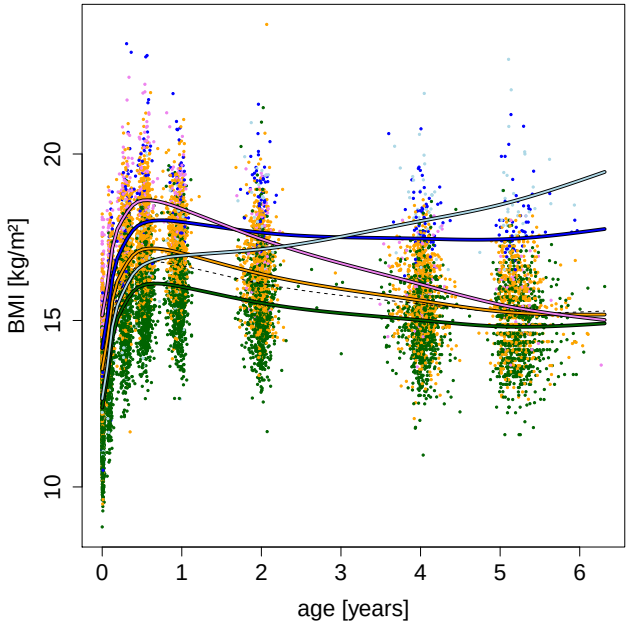


Figure 13: Clustering of the LISA data by the DPM-EM model. Observations belonging to the same cluster are marked with the same color. The dashed black line represents the population effect, the solid colored lines symbolize the cluster effects.

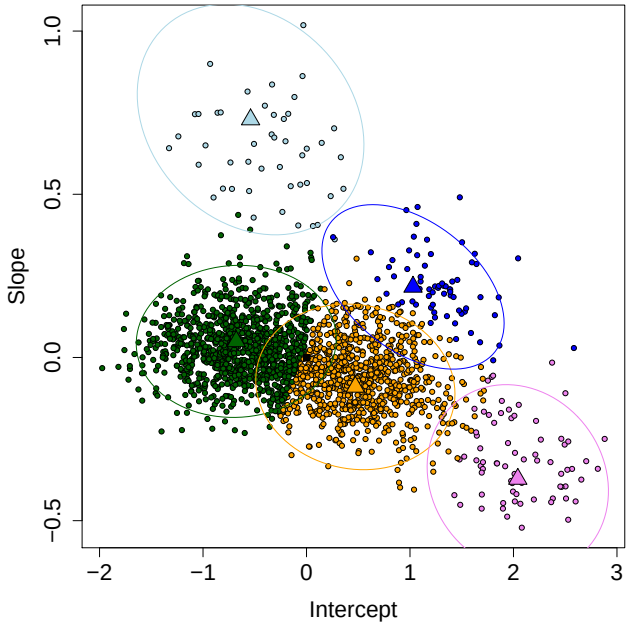
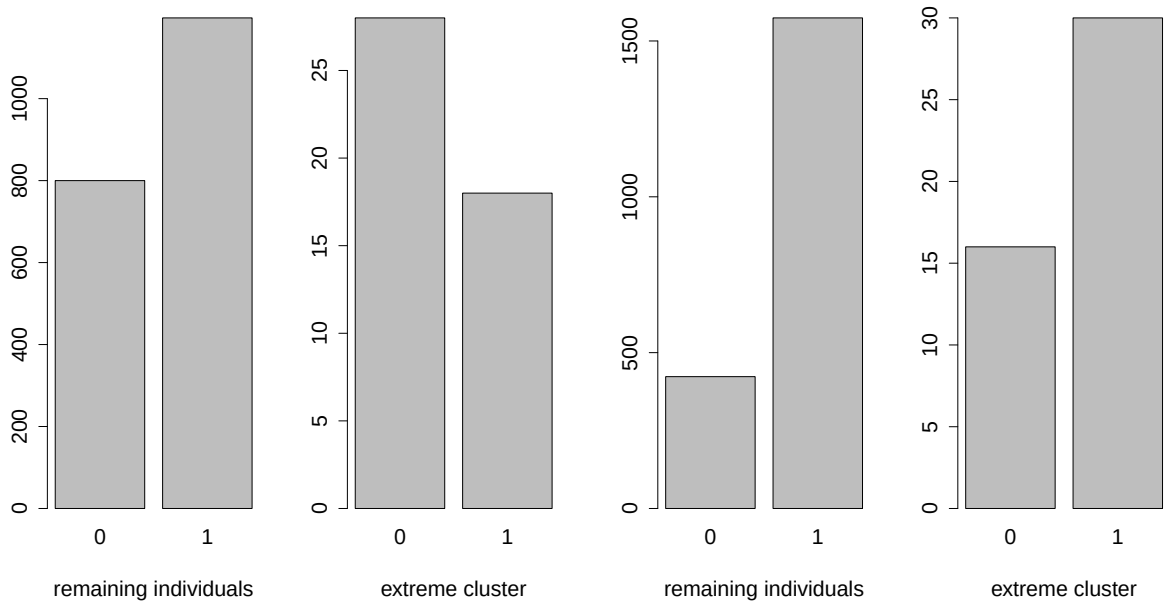


Figure 14: Cluster locations and random effects of DPM-EM model for the LISA data: The big triangles symbolize the cluster locations $\hat{\boldsymbol{\mu}}_h$, the small points the random effects $\hat{\boldsymbol{b}}_i$. Subjects belonging to the same cluster are marked with the same color. The black 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 15: 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 paper as well as in Heinzl et al. (2012). 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 15 (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 13 and Figure 14) 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. 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 15, right).

5 Summary and Discussion

In this paper, 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 a MCMC approach 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.

Bibliography

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.

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.

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

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

Burkhardt, J. (2008). *ASA047: Nelder-Mead Minimization Algorithm*. C++ library.

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.

De Boor, C. (1978). *A practical guide to splines*. New York: Springer-Verlag.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*, 1–26.

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.

Fahrmeir, L., T. Kneib, and S. Lang (2007). *Regression - Modelle, Methoden und Anwendungen*. Berlin: Springer.

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.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review 55*, 245–259.

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. (2012). *clustmixed: Clustering in linear and additive mixed models*. R package version 1.0.

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 (2013). Clustering in linear mixed models with approximate Dirichlet process mixtures using EM algorithm. *Statistical Modelling*. (to appear).

Koenker, R. (2005). *Quantile regression*. Economic Society Monographs. Cambridge: Cambridge University Press.

Li, Y., X. Lin, and P. Müller (2010). Bayesian inference in semiparametric mixed models for longitudinal data. *Biometrics 66*, 70–78.

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.

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).

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. and D. Peel (2000). *Finite mixture models*. New York: Wiley.

Nelder, J. A. and R. Mead (1965). A simplex method for function minimization. *Computer Journal 7*, 308–313.

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.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-effects models in S and S-Plus*. New York: Springer.

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.

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

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 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.

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.

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. London: Chapman & Hall.

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.