Statistical Modelling 2013; **13**(**1**): 41–67

Clustering in linear mixed models with approximate Dirichlet process mixtures using EM algorithm

Felix Heinzl[1] **and Gerhard Tutz**[1]
[1]Department of Statistics, Ludwig-Maximilians-University Munich, Munich, Germany

Abstract: In linear mixed models, the assumption of normally distributed random effects is often inappropriate and unnecessarily restrictive. The proposed approximate Dirichlet process mixture assumes a hierarchical Gaussian mixture that is based on the truncated version of the stick breaking presentation of the Dirichlet process. In addition to the weakening of distributional assumptions, the specification allows to identify clusters of observations with a similar random effects structure. An Expectation-Maximization algorithm is given that solves the estimation problem and that, in certain respects, may exhibit advantages over Markov chain Monte Carlo approaches when modelling with Dirichlet processes. The method is evaluated in a simulation study and applied to the dynamics of unemployment in Germany as well as lung function growth data.

Key words: approximate Dirichlet process mixture; EM algorithm; likelihood inference; linear mixed models; stick breaking

Received February 2012; **revised** June & October 2012; **accepted** November 2012

1 Introduction

Linear mixed models (LMMs), which were proposed by Laird and Ware (1982), are a common tool for the modelling of longitudinal data. The classical model has the form

$$y_{ij}|\boldsymbol{b}_i \overset{ind.}{\sim} N(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i,\ \sigma^2), \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i, \tag{1.1}$$

where y_{ij} denotes the response observed for subject i at observation times t_{ij} with $t_{i1} < \ldots < t_{ij} < \ldots < t_{in_i}$. Population effects of covariates $\boldsymbol{x}_{ij}$ are collected in the parameter vector $\boldsymbol{\beta}$, whereas individual-specific effects of covariates $\boldsymbol{z}_{ij}$ are represented in the parameter vector $\boldsymbol{b}_i$. The classical assumption in (1.1) is a Gaussian distribution for the random effects, i.e., $\boldsymbol{b}_i$ is i.i.d. $N(0, \boldsymbol{D})$, see e.g., Verbeke and Molenberghs (2000) and Ruppert *et al.* (2003). While this choice is mathematically convenient, in

Address for correspondence: Felix Heinzl, Department of Statistics, Ludwig-Maximilians-University Munich, Munich, Germany. E-mail: felix.heinzl@stat.uni-muenchen.de

 10.1177/1471082X12471372

applications it is often questionable for several reasons. The normal distribution is symmetric, unimodal and has light tails. Since the distributional assumption is made on unobserved quantities, it is typically hard to validate these properties. Possible skewness and multimodality (arising, e.g., from an unconsidered grouping structure in the data) may be masked when checking the normal distribution in terms of estimated random effects. A finite mixture of normal distributions as a random effects distribution as suggested by Verbeke and Lesaffre (1996) is much more flexible. One assumes

$$\boldsymbol{b}_i \sim \sum_{h=1}^{N} \pi_h \, N(\boldsymbol{\mu}_h, \boldsymbol{D}), \tag{1.2}$$

where $\pi_1, \ldots, \pi_N$ are mixture weights. Several extensions and alternatives to this heterogeneity model have been proposed. For example, Gaffney and Smyth (2003) used random effects regression mixtures in the context of curve clustering. Approaches for clustering functional data were proposed by James and Sugar (2003) and Liu and Yang (2009). Celeux *et al.* (2005), Ng *et al.* (2006) and Scharl *et al.* (2010) dealt with mixtures of linear mixed effects models. In these approaches, the mixture weights, the variance parameters and all fixed effects are cluster specific, whereas in equation (1.2) just the mixture weights and the locations corresponding to the time trend depend on the cluster. While Booth *et al.* (2008) extended this concept by proposing a stochastic search algorithm for finding the partition that maximizes an objective function based on the classification likelihood, De la Cruz-Mesía *et al.* (2008) generalized the approach to a mixture of non-linear hierarchical models. Villarroel *et al.* (2009) extended the heterogeneity model to allow for a multivariate response variable. In addition, a heteroscedastic normal mixtures in the random effect distribution for multiple longitudinal markers was considered by Komárek *et al.* (2010) for LMMs and by Komárek and Komárková (2012) for generalized linear mixed models. However, in all these approaches, it is necessary to fix the number of mixture components for estimation even though in most applications the number of mixture components is unknown. Further procedures are typically provided for selecting this number, which are usually based on information criteria. A data-driven choice of this number is desirable and could be achieved by a penalization of the mixture weights π_h. For example, Komárek and Lesaffre (2008) penalized differences between reparameterized weights. In contrast, Magder and Zeger (1996) used component-specific covariance matrices subject to the constraint that their determinants are greater than or equal to some minimum value.

In this paper, we present an alternative penalization approach. The basic concept is to shrink the weights π_h towards zero in order to reduce the number of clusters. Therefore, we consider an approximate Dirichlet process mixture (DPM) for the random effects distribution by using the truncated version of the stick breaking presentation of the Dirichlet process (DP); see Ferguson (1973) for the theory behind the DP and Sethuraman (1994) for the stick breaking presentation of the DP. The main advantage of DPs is the cluster property: by using a DPM for the random effects distribution, we obtain automatically a clustering of individuals.

Under the assumption that the population can be described by few clusters we want to identify and interpret them. Since a DP allows to specify a prior on probability measures, it has been widely used in Bayesian inference. For LMMs, DP priors for random effects were first proposed by Bush and MacEachern (1996). The first application of a DPM of Gaussian distributions to random effects was given by Müller and Rosner (1997).

We aim at establishing the DP as a tool for frequentist modelling. Therefore, instead of using Markov chain Monte Carlo (MCMC) methods, which are usually applied for estimation in random effects models with DPs (compare, e.g., Heinzl *et al.*, 2012), we extend the traditional Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977) used in the heterogeneity model of Verbeke and Lesaffre (1996) and refer to it as DPM-EM model. We will illustrate that the EM algorithm has an essential advantage over MCMC methods, as far as DPs are concerned. In summary, on the one hand, our DPM-EM model provides a regularization approach for the number of mixture components in (1.2). On the other hand, our model is a method to obtain clustering of individuals in longitudinal data.

The paper is organized as follows: In Section 2.1, the model hierarchy as well as the cluster property of DPs are illustrated. In Section 2.2, we present our DPM-EM algorithm in detail. Simulation results can be seen in Section 3 while applications are shown in Section 4. Finally, Section 5 subsumes the main aspects of our approach.

2 LMMs with DPMs

2.1 Model hierarchy

Collecting observations y_{ij}, $j = 1, \ldots, n_i$, for individual i in the vector $\boldsymbol{y}_i$, model (1.1) can be written in matrix notation as

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

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

$$\begin{aligned} \boldsymbol{b}_i | \boldsymbol{\theta}_i &\overset{ind.}{\sim} N(\boldsymbol{\theta}_i, \boldsymbol{D}), \quad i = 1, \ldots, n, \\ \boldsymbol{\theta}_i | G &\overset{i.i.d.}{\sim} G, \quad i = 1, \ldots, n, \\ G &\sim DP(\alpha, G_0). \end{aligned} \tag{2.1}$$

Here, $DP(\alpha, G_0)$ is a distributional assumption for the unknown mixing distribution G. A special feature of the DP is that each realization of G is a discrete probability measure (Blackwell, 1973). So in the DPM specification, choosing a DP for the $\boldsymbol{\theta}_i$, $i = 1, \ldots, n$, creates ties among these and therefore forms clusters of subjects,

whereas each subject still has its own unique random effects value. In general, there are $k \leq n$ clusters and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ can be represented by cluster locations $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ and cluster allocation variables. The strength of clustering and therefore the number of clusters is determined by the parameter α, which controls the confidence in the base distribution G_0. According to the relationship between Bayesian and likelihood inference, we choose a diffuse uniform distribution on $(-\infty, \infty)$ for G_0. So, in principle, no cluster location is preferred over others. Although in theory, an automatic clustering structure is induced by the DP, a severe practical problem arises within the Bayesian framework when using MCMC methods, namely how to obtain a single clustering estimate $\hat{c}$ based on an MCMC sample of clusterings $c^{(1)}, \ldots, c^{(M)}$, where $c^{(m)}$, $m = 1, \ldots, M$, describes the cluster allocation at iteration m and $\hat{c}$ the final cluster allocation. By using MCMC methods in each iteration, ties among the $\boldsymbol{\theta}_i$, $i = 1, \ldots, n$, are created and clusters are formed. But when approximating the posterior means by the means over MCMC samples $\hat{\boldsymbol{\theta}}_i = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\theta}_i^{(m)}$, $i = 1, \ldots, n$, the clustering of subjects gets lost. Fritsch and Ickstadt (2009) gave an overview on operations of how the MCMC sample of clusterings $c^{(1)}, \ldots, c^{(M)}$ can be aggregated to a single clustering $\hat{c}$ but due to the high number of possible clusterings, these methods are typically not feasible in larger problems. By using EM type algorithms all these strategies for rescuing the cluster property of the DP are unneeded. The reason is that the EM algorithm converges to fixed values, whereas MCMC methods converge to distributions. So with EM type algorithms, the cluster property of the DP can be used more directly. While other alternatives to the MCMC methods as the recursive algorithm of Newton and Zhang (1999) or the variational method of Blei and Jordan (2006) are based on approximative posterior distributions, our EM algorithm aims at maximizing the posterior given in Section 2.2 directly.

In practice, inference with DPs can be built on the constructive definition of the DP by Sethuraman (1994). This stick breaking representation implies that $G \sim DP(\alpha, G_0)$ is equivalent to

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

with locations simulated by $\boldsymbol{\mu}_h \overset{i.i.d.}{\sim} G_0$ and weights constructed through the stick breaking procedure

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

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

The recursive definition of weights $\pi_h = v_h \left(1 - \sum_{l<h} \pi_l\right), h \in \mathbb{N}$, which is clarified in Appendix A.1, and is visualized in Figure 1, gives the procedure its name. It works

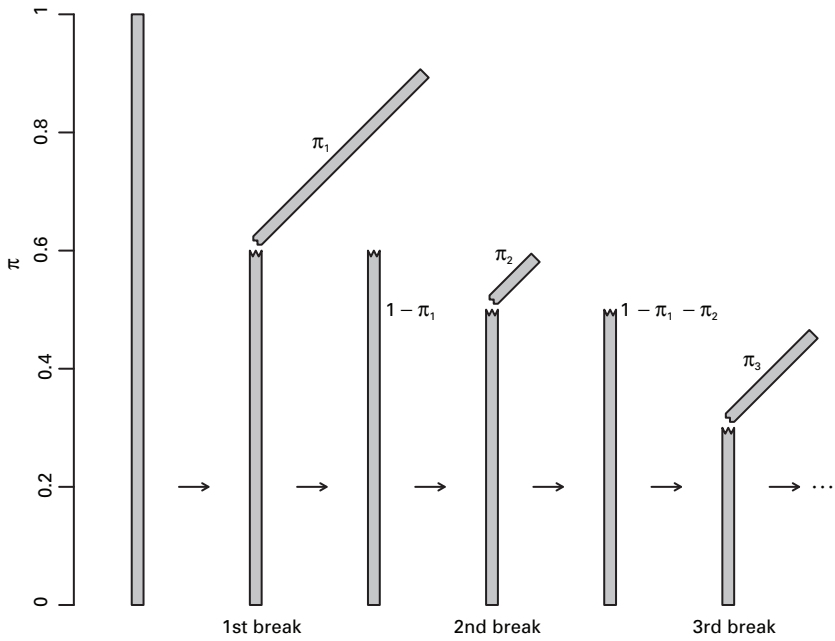

Figure 1 Construction of $\pi_1, \pi_2, \ldots$ by stick breaking

Source: Authors' own.

as follows: first, for getting π_1, a piece is broken away from a stick of length one. Next, from the remainder of the stick, $1 - \pi_1$, breaks a further piece away, called π_2 and so on. So the random weights decrease stochastically as the index h grows. More concretely, $E(\sum_{h=N+1}^{\infty} \pi_h)$ converges to zero exponentially with $N \to \infty$ (see Appendix A.2). So an established concept to make the stick breaking procedure applicable in practice is to approximate the DP by considering

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

with large enough N. Here, all locations $\boldsymbol{\mu}_h$ and all weights v_h and π_h are constructed as before with the exception of $v_N = 1$. In summary, by using the stick breaking procedure, the distribution assumption for the random effects (2.1) can be rewritten as

$$\begin{aligned} \boldsymbol{b}_i | \boldsymbol{v} &\overset{i.i.d.}{\sim} \textstyle\sum_{h=1}^{N} \pi_h \, N(\boldsymbol{\mu}_h, \boldsymbol{D}), \quad i = 1, \ldots, n, \\ \pi_h &= v_h \textstyle\prod_{l<h}(1 - v_l), \qquad h = 1, \ldots, N, \\ v_h &\overset{i.i.d.}{\sim} Be(1, \alpha), \qquad h = 1, \ldots, N-1 \end{aligned} \tag{2.2}$$

with $\boldsymbol{v} = (v_1, \ldots, v_{N-1})^T$. Therefore, for the random effects distribution, we get a finite mixture of normal distributions as in equation (1.2) in which the number of

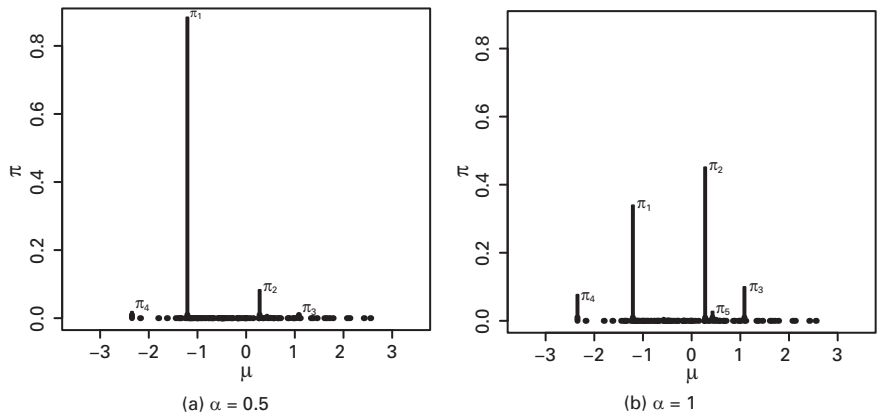

Figure 2 Realizations of $G \sim DP(\alpha, G_0)$ with $G_0 = N(0, 1)$
Source: Authors' own.

mixture components with $\pi_h \neq 0$ is penalized. It should be noted that a generalization to a heteroscedastic normal mixture with different covariance matrices over components is also possible—following, e.g., the approach of Yao and Holmes (2011). Nevertheless, the assumption (2.2) seems to be sufficiently flexible and avoids numerical problems, which arise in the case of a heteroscedastic normal mixture (Verbeke and Molenberghs, 2000).

In the following, the order of $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ is given by the corresponding weights in decreasing order under the restrictions $\sum_{h=1}^{N} \pi_h \boldsymbol{\mu}_h = 0$ and $\sum_{h=1}^{N} \pi_h = 1$. The first restriction ensures $E(\boldsymbol{y}_i) = \boldsymbol{X}_i \boldsymbol{\beta}$. The second constraint is standard and is automatically fulfilled by $v_N = 1$. See Figure 2 for an illustration of two discrete probability measures simulated by DPs with different values of α. Obviously α controls the number of cluster locations $\boldsymbol{\mu}_h$ with weights $\pi_h \neq 0$ and thus the effective number of clusters.

For example, the truncated DP was used by Muliere and Tardella (1998), Ishwaran and James (2002), Kottas and Gelfand (2001), Gelfand and Kottas (2002) and Ohlssen *et al.* (2007), see Section 2.2 for a strategy of choosing N. Even though other methods exist that are based on the stick breaking representation and that avoid the truncation (see, e.g., Walker 2007 and Papaspiliopoulos and Roberts 2008) the truncated version distinguishes oneself by simplicity and theoretical justifications as shown in Muliere and Tardella (1998), Ishwaran and James (2001) as well as Ishwaran and James (2002). In our case, this truncation is still more attractive because our approach is formally similar to the heterogeneity model of Verbeke and Lesaffre (1996) but with 'penalized' weights referred to the stick breaking procedure which induces that only the relevant clusters get comparably high weights. Inference is possible by extending the EM algorithm of the heterogeneity model. Another inference approach within the framework of DPs is based on Pólya urn scheme (Blackwell and MacQueen, 1973) and thus on integrating out the unknown distribution G (compare

Escobar, 1994, MacEachern, 1994, Escobar and West, 1995 as well as MacEachern and Müller, 1998). Nevertheless when using this marginal method instead of the stick breaking procedure, the connection between the DP and the heterogeneity model of Verbeke and Lesaffre (1996) is hidden. This is the main reason why the stick breaking presentation is much more appealing to us and seems to be more user-friendly than the Pólya urn inference scheme, which also has other drawbacks (see, e.g., Ishwaran and James, 2001). In the next section, we will explain how DPs can be embedded in the EM framework. As is seen, elaborate handling of the DP parameters is necessary.

2.2 Inference

In the following, we give an EM algorithm for the LMM described in Section 2.1. The algorithm is based on derivations by McLachlan and Peel (2000) and McLachlan and Krishnan (1997) and is similar to the algorithm used by Verbeke and Lesaffre (1996) but includes a penalty term. The following approach can be parameterized either by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^T$ or by $\boldsymbol{v}$. Since the latter parametrization simplifies calculations, it is used in the following. Nevertheless, only for a compact presentation, we write π_h instead of $v_h \prod_{l<h}(1 - v_l)$. Let $\boldsymbol{\xi} = (\alpha, \boldsymbol{v}, \boldsymbol{\psi})^T$, where $\boldsymbol{\psi}$ is the vector containing all the remaining parameters $\boldsymbol{\beta}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N, \boldsymbol{D}, \sigma^2$. The cluster membership of each individual can be described by the latent variable $z_i := (z_{i1}, \ldots, z_{iN})^T$, where $z_{ih} = 1$ if subject i belongs to cluster h and 0 otherwise. Marginalization over the random effects yields the complete model with observed data $\boldsymbol{y}_i$ as well as unobserved data z_i and $\boldsymbol{v}$:

$$\begin{aligned} \boldsymbol{y}_i | z_i &\overset{ind.}{\sim} N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\mu}_h,\ \boldsymbol{V}_i), && i = 1, \ldots, n, \\ z_i | \boldsymbol{v} &\overset{i.i.d.}{\sim} M(1, \boldsymbol{\pi}), && i = 1, \ldots, n, \\ v_h &\overset{i.i.d.}{\sim} Be(1, \alpha), && h = 1, \ldots, N-1, \end{aligned} \tag{2.3}$$

with $\boldsymbol{V}_i = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T + \sigma^2 \boldsymbol{I}_{n_i}$ and $M(\cdot)$ denoting the multinomial distribution. Equation (2.3) describes the data generating process for the data $(\boldsymbol{y}_i, z_i, \boldsymbol{v})$ given the parameters $(\alpha, \boldsymbol{\psi})$, i.e.,

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

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

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

when assuming a flat prior for α and $\boldsymbol{\psi}$. Here $f_{ih}(\cdot)$ denotes the density function of $N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\mu}_h,\ \boldsymbol{V}_i)$. Note that from a Bayesian point of view $\boldsymbol{\psi}$ and $\boldsymbol{v}$ are parameters,

whereas α is the hyperparameter for the prior on $\boldsymbol{v}$. In an empirical Bayes context, such a hyperparameter would be estimated by maximizing the marginal incomplete likelihood (Maritz and Lwin, 1989). However, in the present case, the marginalization is analytically not feasible. Following the strategy of McAuliffe *et al.* (2006), in the case of a DPM model, such an integration could be avoided by many alternations between an inference phase where the parameters $\boldsymbol{v}$ and $\boldsymbol{\psi}$ are estimated and an estimation phase where the hyperparameter α is estimated. This procedure would be very time-consuming in our case. Thus we prefer to handle α like any other parameter and to estimate α conditionally on the actual state of the other parameters during the algorithm. In general, vague priors like our diffuse prior for α are an alternative to empirical Bayes inference for achieving robustness (McAuliffe *et al.*, 2006).

Finally, as log-posterior one obtains

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

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

E-step

Collecting all observed data by $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$ for the E-step, we get

$$Q(\boldsymbol{\xi}) = E\left(l_P(\boldsymbol{\xi})|\,\boldsymbol{y}, \boldsymbol{\xi}^{(t)}\right)$$

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

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

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

M-step

For clarity, in the following we write $\pi_{ih} := \pi_{ih}(\boldsymbol{\xi}^{(t)})$ but note that for the M-step it is essential that π_{ih} is fixed from the last iteration t because then using that $Q(\boldsymbol{\xi})$ is the

sum of $Q(\alpha, \boldsymbol{v})$ and $Q(\boldsymbol{\psi})$ the optimization problem in the M-step can be separated into two parts: The maximization of

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

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

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

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

$$v_h = \frac{\sum_{i=1}^{n} \pi_{ih}}{\sum_{i=1}^{n} \sum_{l=h}^{N} \pi_{il} + \alpha - 1}, \qquad h = 1, \ldots, N-1, \tag{2.4}$$

and

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

Without further restrictions it could happen that $v_h \notin [0, 1]$. To avoid this we use the following correction approach: Update v_h by (2.4) 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: $\sum_{l=1}^{h} \pi_l^\circ > 1$. Here the idea of the penalization approach becomes evident. First note that for $\alpha = 1$ we get the usual estimators for π_h and no restrictions are needed. Compared to these estimators, for $\alpha \in (0, 1)$, all weights π_h for $h < h^*$ are stretched by the factor $\frac{n}{n+\alpha-1}$, while all weights π_h for $h > h^*$ are set to zero. The amount of stretching is controlled by the parameter α. If $\alpha \approx 0$ a very strong clustering is achieved while for larger values of α only few clusters drop out. In order to avoid log(0) we choose $v_h = 1 - 10^{-300}$ instead of $v_h = 1$ in the algorithm. Then $\pi_h \approx 0$ for $h > h^*$.

In the second part of the M-step, we get the current state for $\boldsymbol{\psi}$ by alternating separate maximization of $Q(\boldsymbol{\psi})$ to $\boldsymbol{\beta}$, to $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ and to the variance parameters

$\boldsymbol{D}$ and σ^2. Conditional on the actual state of the other parameters, the maximization of $\boldsymbol{\beta}$ results in

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

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

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

For the simultaneous maximization of the variance parameters given $\boldsymbol{\beta}$ and $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$, a numerical procedure like the Nelder-Mead method is necessary.

Choice of N

By truncation of the DP, the originally infinite constraints $\sum_{h=1}^{\infty} \pi_h \boldsymbol{\mu}_h = \boldsymbol{0}$ and $\sum_{h=1}^{\infty} \pi_h = 1$ are converted into finite ones, which can be handled easily. Concretely, the second constraint is fulfilled by $v_N = 1$. On the one hand, this idea avoids reparameterizations as in Jara *et al.* (2009) or post-processing strategies as in Li *et al.* (2011). On the other hand, $v_N = 1$ actually means that the last weight π_N absorbs all the remaining probabilities $\pi_N, \ldots, \pi_\infty$ of the untruncated DP, which can be seen as zero in the truncated version. But this is only correct if N is large enough, so that

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

with $\varepsilon > 0$. So it is crucial to choose N correctly. This is still more challenging because the choice of N depends on α, which itself is estimated. As proposed by Ohlssen *et al.* (2007) the postulation (2.5) can be transformed into

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

which can be seen from Appendix A.2. Thus for a given range on α, a lower bound for N can be determined. For inducing a very strong clustering and according to the previous considerations within this section, we allow only the range $\alpha \in (0, 1)$ which is automatically fulfilled by a very low starting value for α. This means that even for $N \geq 15$ a good approximation can be achieved ($\varepsilon = 0.0001$). So in the majority of cases, $N = \min\{n, 100\}$ is a satisfying choice.

Start and stop of the algorithm

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

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

The EM algorithm stops if the penalized incomplete log-likelihood is not ascending any more. After convergence we get the cluster membership by the matrix of estimated π_{ih}. 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 a few relevant clusters k. Based on the weights of all clusters the random effects are predicted by using the mean of the posterior $\boldsymbol{b}_i | \boldsymbol{y}_i$, which is given by

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

where q denotes the dimension of random effects. A proof of this formula is given in Appendix A.3.

Implementation

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

(and so invertible, too) if $\boldsymbol{L}$ is a matrix with exclusively nonzero diagonal entries (Lindstrom and Bates, 1988).

3 Simulation study

3.1 Setting

In the following simulation study the performance of the DPM-EM is evaluated. The study aims at clarifying in which data situations our approach improves estimation compared to the LMM with a normal distribution or a finite mixture of normal distributions as random effects distribution. Note that for prediction accuracy of random effects, there is a trade-off with regard to the assumed number of clusters: On the one hand, for prediction of $\boldsymbol{b}_i$ it makes sense to borrow information from other similar subjects. On the other hand, it is not reasonable to incorporate individuals which show a basically different behaviour. For examining this trade-off, we compare the commonly used LMM with Gaussian random effects distribution (one cluster model) as well as the three, five and ten cluster model to our DPM-EM model with a data-driven choice for the number of clusters. Moreover, in the simulation study, we investigate the impact of the number of observations within clusters and the separation between clusters. We generated datasets assuming a simple linear trend model

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

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

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

imitating a population consisting of three clusters of overlapping subpopulations.

Throughout the simulations, we set $n = 20$ and

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

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

For given n_i, observation times are generated from

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

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

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

corresponding to *clearly separated clusters*,

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

corresponding to *moderately separated clusters*,

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

corresponding to *only one cluster*.

Combining these different settings for observation times and clusters results in nine different scenarios. For each of them, we compare the estimation results from the DPM-EM algorithm with results based on Gaussian random effects using the R-function lmer() from the lme4 package by Bates *et al.* (2012) and with results of models using an unpenalized ($\alpha = 1$) finite normal mixture as random effects distribution. In each simulation run s, we calculate the average prediction error

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

for uncentered random intercepts $b^*_{i0} = \beta_0 + b_{i0}$ and random slopes $b^*_{i1} = \beta_1 + b_{i1}$. In addition, the estimation accuracy of the fixed effects is investigated by the relative bias $RB_k = 100 \cdot (\hat{\beta}_k - \beta_k)/\beta_k$, $k = 0, 1$.

3.2 Results

In the following, we summarize results of the nine combinations. For some scenarios, the empirical distribution of $PE_k(s)$ values obtained from simulation run $s = 1, \ldots, 100$ is represented through box plots.

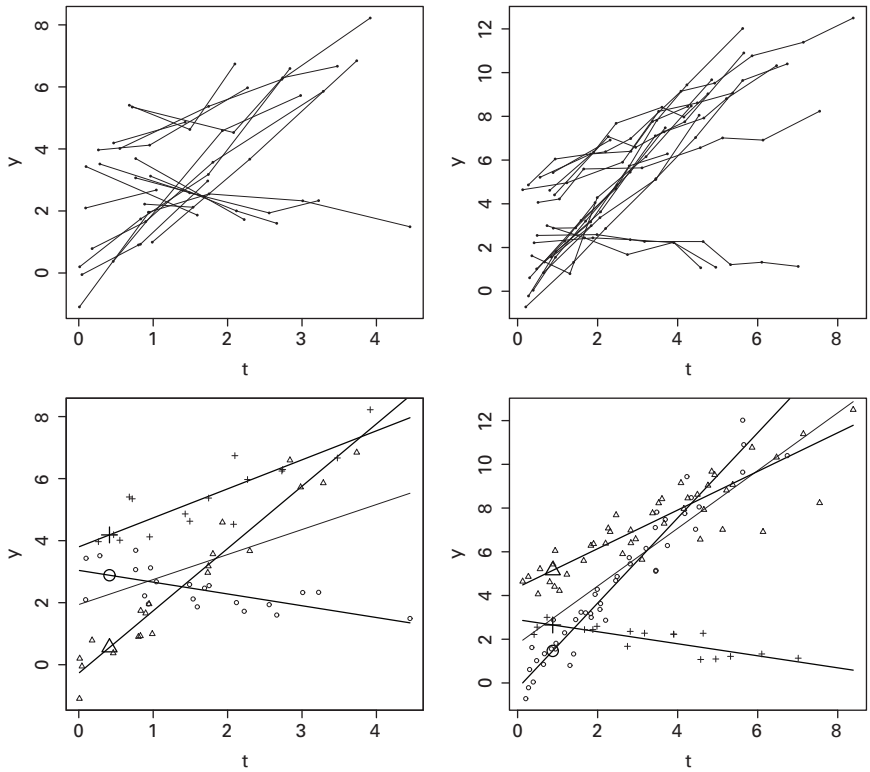

Figure 3 Trace plots (top) and clustering by DPM-EM model (bottom) with clearly separated clusters for few individual observations ($\nu = 1$) (left) and a moderate number of observations on individuals ($\nu = 3$) (right)
Source: Authors' own.

Clearly separated clusters

Figure 3 (top) displays trace plots of typical longitudinal data generated in the setting of clearly separated clusters that shows that cluster effects can easily be detected visually. On the left, there are only a few observations for each subject while on the right the mean of the number of repeated measurements is 5 corresponding to several observations. Not surprisingly the DPM-EM model detects three clusters in both cases (Figure 3 (bottom)). The thin line shows the overall effect and the thick lines visualize the means of the resulting clusters. Which observation is assigned to which cluster is marked by the same symbol.

LMMs with DPM penalty substantially improve upon results based on a misspecified Gaussian random effects assumption, especially in the case of several and many observations (see Table 1 and, e.g., Figure 4). In general, models with a finite mixture as random effects distribution yield better predictions for random effects than the classical LMM with normally distributed random effects. Of course, the best prediction can be observed for the model with fixed $N = 3$ clusters because this

Table 1 Medians of PE_k and RB_k with $k = 0, 1$ for clearly separated clusters

	$\nu = 1$				$\nu = 3$				$\nu = 5$			
	PE_0	PE_1	RB_0	RB_1	PE_0	PE_1	RB_0	RB_1	PE_0	PE_1	RB_0	RB_1
Normal	0.373	0.185	−4.091	2.068	0.222	0.054	−1.048	4.710	0.148	0.015	−2.127	0.957
DPM-EM	0.135	0.063	−6.818	4.697	0.060	0.012	−5.212	6.935	0.048	0.006	−1.377	0.887
$N = 3$	0.111	0.058	−3.698	4.313	0.054	0.011	−2.914	5.197	0.045	0.005	−0.457	1.741
$N = 5$	0.145	0.062	−2.906	4.802	0.072	0.015	−2.760	4.387	0.050	0.006	−0.243	2.026
$N = 10$	0.222	0.112	−3.331	2.062	0.101	0.020	−2.188	6.324	0.080	0.008	−0.240	1.514

Source: Authors' own.

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

Moderately separated clusters

In the following the differences between clusters get smaller. See Figure 5 for two typical trace plots in the case of few, respectively, many individual observations.

Still the DPM-EM model outperforms both the homogeneity model (LMM with normal random effect distribution) and the unpenalized heterogeneity model with $N = 5$ and $N = 10$ clusters (Figure 6). Only the 'true' model with $N = 3$ clusters is able to feature a lower error in predicting the random effects (Table 2). Note that the superiority of the DPM-EM model over the classical LMM with normal random effects distribution is even higher in the case of many individual observations.

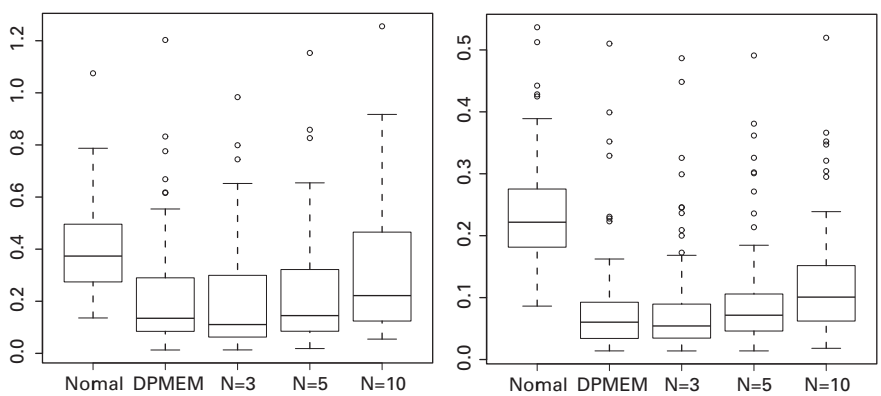

Figure 4 Box plots of PE_0 with clearly separated clusters for few individual observations ($\nu = 1$) (left) and a moderate number of observations on individuals ($\nu = 3$) (right)

Source: Authors' own.

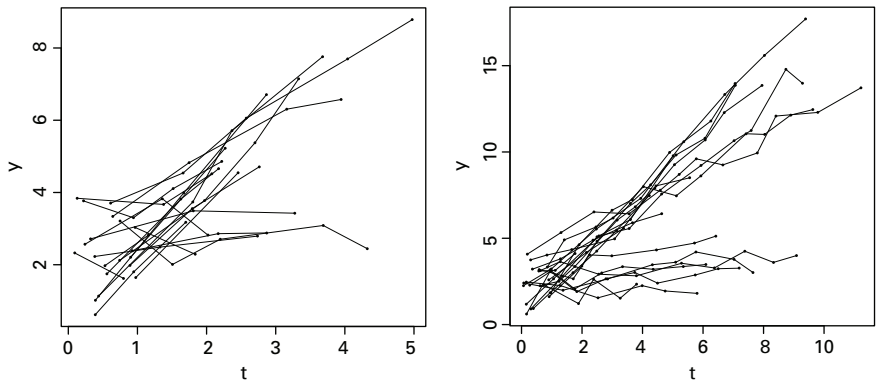

Figure 5 Trace plots with moderately separated clusters for few individual observations ($\nu = 1$) (left), respectively, many individual observations ($\nu = 5$) (right)

Source: Authors' own.

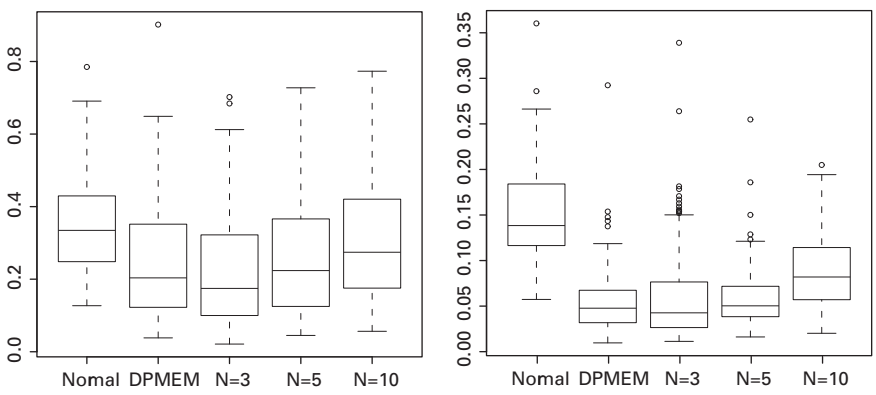

Figure 6 Box plots of PE_0 with moderately separated clusters for few individual observations ($\nu = 1$) (left), respectively, many individual observations ($\nu = 5$) (right)

Source: Authors' own.

Table 2 Medians of PE_k and RB_k with $k = 0, 1$ for moderately separated clusters

	$\nu = 1$				$\nu = 3$				$\nu = 5$			
	PE_0	PE_1	RB_0	RB_1	PE_0	PE_1	RB_0	RB_1	PE_0	PE_1	RB_0	RB_1
Normal	0.335	0.164	−2.112	1.912	0.207	0.046	−0.751	2.204	0.138	0.015	−1.122	0.750
DPM-EM	0.204	0.114	−6.088	4.673	0.082	0.018	−3.104	2.335	0.048	0.005	−0.920	1.117
$N = 3$	0.175	0.097	−3.799	2.111	0.063	0.014	−0.108	3.193	0.043	0.005	−1.275	0.945
$N = 5$	0.224	0.122	−3.091	2.028	0.082	0.018	−0.108	3.089	0.050	0.006	−1.226	0.693
$N = 10$	0.274	0.140	−2.987	1.381	0.126	0.025	−0.344	3.114	0.082	0.008	−1.304	1.469

Source: Authors' own.

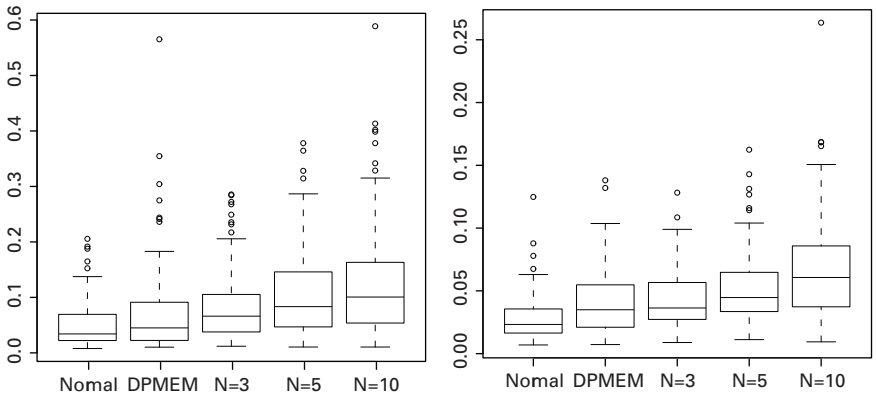

Figure 7 Box plots of PE_0 with only one cluster for few individual observations (ν = 1) (left), respectively, many individual observations (ν = 5) (right)

Source: Authors' own.

Only one cluster

When regarding Figure 7 and Table 3 for only one cluster, we can conclude the following: Only the LMM with normal random effects distribution which is the 'true' model in this setting is better than the DPM-EM model. The background for this feature is that the DPM-EM model detects sometimes more than one cluster in the data. Different patterns in the data are taken seriously. Nevertheless the DPM-EM model exhibits lower prediction errors than all unpenalized heterogeneity models because in the majority of cases less clusters than three are observed by the DPM-EM model.

In summary, we draw the following conclusion: The DPM-EM models yield the better estimates for random effects—in terms of prediction errors—the clearer the clusters differ and the more observations are in the data. It makes a good job both for normally distributed random effects and for random effects following a mixture of three normal distributions and is only a little bit inferior to the corresponding correctly specified model. Thus the DPM-EM model turns out to be very flexible

Table 3 Medians of PE_k and RB_k with $k = 0, 1$ for only one cluster

	$\nu = 1$				$\nu = 3$				$\nu = 5$			
	PE_0	PE_1	RB_0	RB_1	PE_0	PE_1	RB_0	RB_1	PE_0	PE_1	RB_0	RB_1
Normal	0.034	0.020	−0.277	−1.081	0.029	0.007	0.605	−0.911	0.023	0.004	−0.163	−0.261
DPM-EM	0.045	0.022	0.004	−1.465	0.040	0.009	0.437	−0.003	0.035	0.005	−0.091	−0.205
$N = 3$	0.066	0.027	0.372	−1.242	0.045	0.010	0.916	−0.848	0.036	0.005	−0.077	−0.421
$N = 5$	0.083	0.034	0.277	−1.218	0.053	0.012	0.493	−1.035	0.045	0.006	−0.782	−0.299
$N = 10$	0.101	0.038	0.582	−1.804	0.062	0.012	0.499	−1.417	0.061	0.006	−0.166	−0.384

Source: Authors' own.

without risk of misspecifying the model like it can happen for the homogeneity model and the unpenalized heterogeneity model.

4 Applications

4.1 Unemployment

The practical use of the proposed method is investigated in two data examples. First, the variation of the unemployment over the federal states of Germany across time is considered (Weise *et al.*, 2011). We examine the unemployment rate of each federal state from 2005 to 2010 in order to identify differences between states. Figure 8 shows different levels of the unemployment rates and a negative time trend which can be regarded as approximately linear. Therefore we consider a random slope model for the annual average of the unemployment rate y_{ij} of state i and measurement j

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

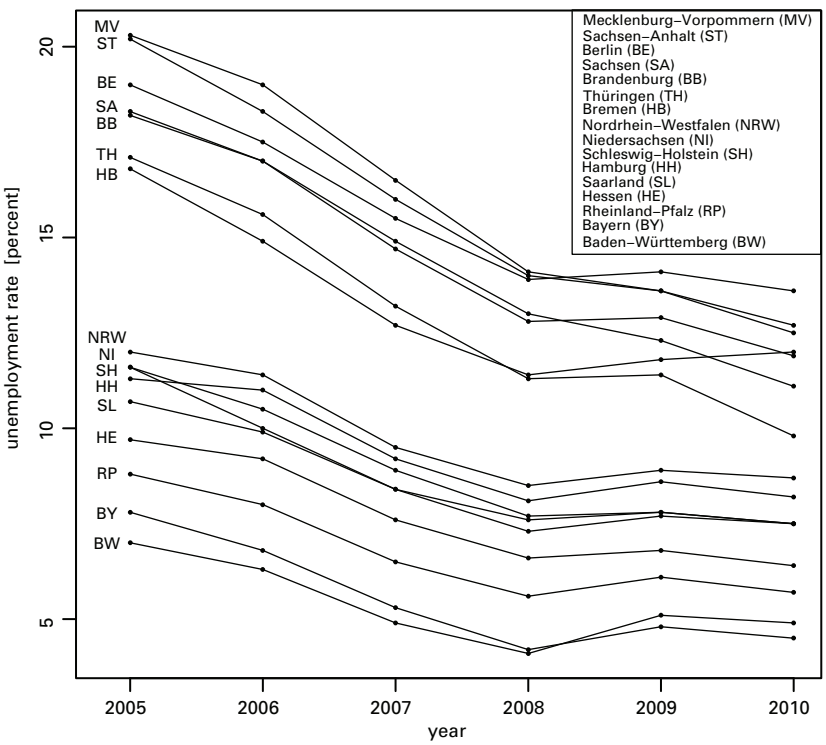

Figure 8 Unemployment rates of the federal states of Germany across time
Source: Authors' own.

Table 4 Estimation results for the fixed effects and variance parameters by DPM-EM model for the unemployment data

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

Source: Authors' own.

Since there is no symmetric unimodal variation of the individual intercepts about the overall mean it would not be appropriate to assume a Gaussian random effect distribution. Instead, the centered i.i.d. random effects $\boldsymbol{b}_i = (b_{i0}, b_{i1})^T$ follow a mixture distribution of Gaussian components with penalized mixture weights (2.2).

We are looking for clustering the federal states in order to expose which states show similar behaviour. Only for a better interpretability we change the zero point of the time variable to 2005. Thus, during calculations, the time variable is labelled by $0, 1, \ldots, 5$ for the years $2005, 2006, \ldots 2010$.

First, Table 4 shows the estimated fixed effects and variance parameters. The standard errors and confidence intervals for the fixed effects have been estimated by the nonparametric bootstrap method proposed by Efron (1979) using the Monte Carlo approximation with 1000 replications. Obviously the time variable has a significant effect on the unemployment rate on the 5% level.

Our DPM-EM model detects three clusters with estimated weights $\hat{\pi}_1 = 0.467$, $\hat{\pi}_2 = 0.425$ and $\hat{\pi}_3 = 0.108$. Figure 9 shows the population effect (thin line) as well as the cluster effects (thick lines). Observations belonging to the same cluster are marked with the same symbol. For identification, this symbol is also added to the corresponding thick line. The southern federal states Bayern, Baden-Württemberg and Rheinland-Pfalz are assigned to cluster 3 (+) which features the lowest unemployment rate and the weakest decrease over time. As Table 5 shows, here, the base level in 2005 is −6.469 lower compared to the overall unemployment rate 13.718. In the south also, the decrease of the unemployment rate is less distinct than in the other states. A similar effect can be observed in cluster 2 (△). Here, the gap to the global

Table 5 Estimates of the cluster locations by DPM-EM model for the unemployment data

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

Source: Authors' own.

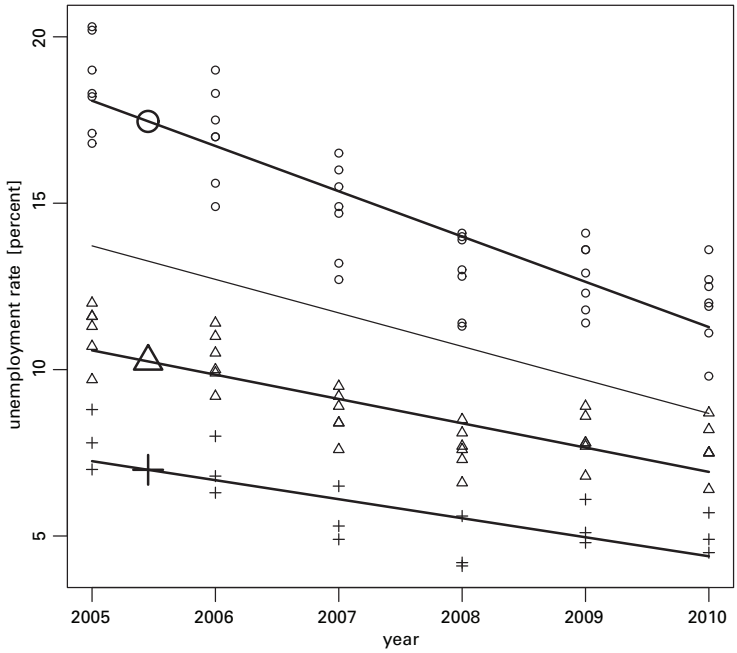

Figure 9 Clustering of unemployment data by DPM-EM model. Observations belonging to the same cluster are marked with the same symbol. The thin line represents the population effect and the thick lines symbolize the cluster effects

Source: Authors' own.

Table 6 Estimates of $\hat{\pi}_{ij}$ for unemployment data by DPM-EM model

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

Source: Authors' own.

intercept is considerably smaller. Furthermore, there is one cluster (○) with a much more higher base level and a stronger decrease of the unemployment rates. It is remarkable that these states are all in Eastern Germany or city states. Only the city state Hamburg makes an exception to that feature and belongs to cluster 2.

Table 6 shows the estimated probabilities $\hat{\pi}_{ij}$. Here, it can be seen that for most of the states, the assignment to a specific cluster is very distinct. Only for Rheinland-Pfalz, the probability for cluster 3 and cluster 2 is very similar. The parameter α which controls the number of clusters is estimated by $\hat{\alpha} = 0.00155$. It is a typical feature that estimated αs are very small. This means that the strongest clustering as allowed by the data is the best one.

4.2 Lung function growth

In the second application, the lung function growth of girls in Topeka (USA) is examined by our DPM-EM model. These data are a subsample from the six cities' study of air pollution and health in Dockery *et al.* (1983). The response variable is

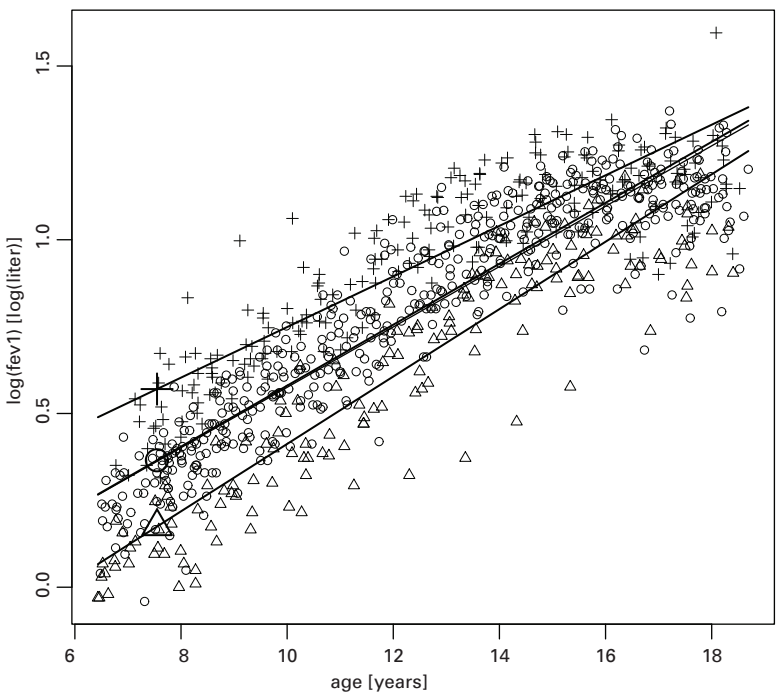

Figure 10 Clustering of lung function growth data by DPM-EM model: Observations belonging to the same cluster are marked with the same symbol. The thin line represents the population effect and the thick lines symbolize the cluster effects

Source: Authors' own.

the logarithmic forced expiratory volume in one second (fev1). Our sample consists of 100 girls, with a minimum of two and a maximum of 12 observations over time. Again, we use a LMM with random intercepts and random slopes

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

and a DPM as random effects distribution (2.2). While the plot of all measurements over time (Figure 10) is not very informative because of the large number of measurements, the clustering effect of the DPM-EM model can be seen much easily from Figure 11. Here the axes represent the intercepts and slopes, respectively. The square at coordinates (0,0) marks the population effect. All other icons are interpreted as deviations from the population effect. The thick big ones symbolize the cluster locations $\boldsymbol{\mu}_h$ and the thin small ones the random effects $\boldsymbol{b}_i$. Girls assigned to the same cluster are marked with the same symbol and are arranged around the three cluster locations in the form of ellipses.

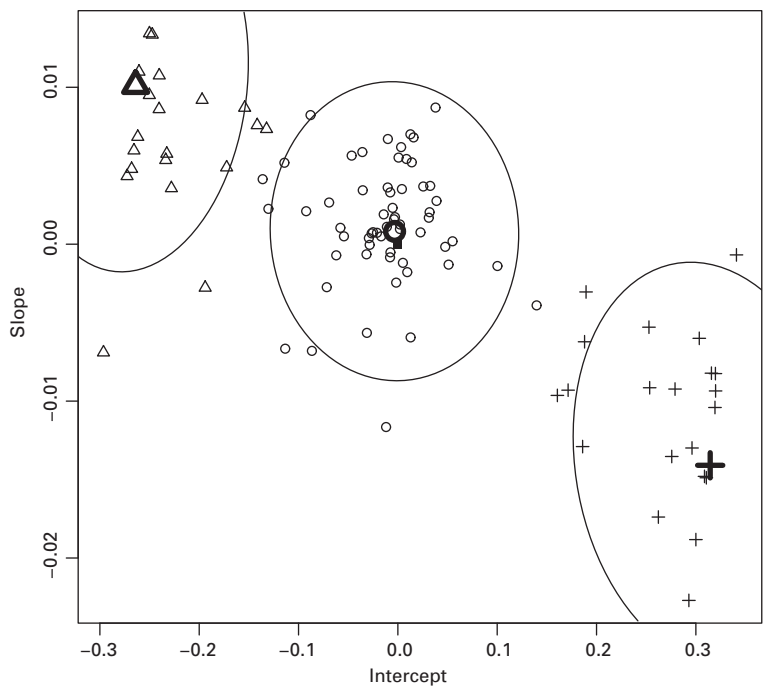

Figure 11 Cluster locations and corresponding random effects of DPM-EM model for lung function growth data: The thick big icons symbolize the cluster locations $\hat{\boldsymbol{\mu}}_h$ and the thin small ones the random effects $\hat{\boldsymbol{b}}_i$. The square at coordinates (0,0) marks the population effect. Ellipses with level 0.95 visualize the estimated conditional distribution of random effects in the clusters

Source: Authors' own.

5 Conclusion

We introduced LMMs with a DPM for the random effects distribution in order to penalize the number of clusters in the finite mixture of normal distribution. While models with DPs are typically fitted by Bayesian methods like MCMC, we used the EM algorithm because then the cluster property of the DP can be used directly. So our method can be called an agglomerative clustering approach of individuals for longitudinal data. The DPM-EM algorithm itself was presented in detail. Furthermore, we showed in a simulation study that our approach outperforms the classical LMM in the case of an underlying grouping structure. Applications of this DPM-EM algorithm were demonstrated by considering unemployment data and lung function growth data. Extensions of this DPM-EM algorithm to additive mixed models are planned.

Appendix

A.1 Recursive definition of weights

$$\prod_{h=1}^{N}(1-v_h) = (1-v_N)\prod_{h=1}^{N-1}(1-v_h) = \prod_{h=1}^{N-1}(1-v_h) - \pi_N = \ldots =$$
$$= (1-v_1) - \sum_{h=2}^{N}\pi_h = 1 - \sum_{h=1}^{N}\pi_h$$

A.2 Convergence of weights

$$E\left(\sum_{h=N+1}^{\infty}\pi_h\right) = E\left(1-\sum_{h=1}^{N}\pi_h\right) \overset{A.1}{=} E\left(\prod_{h=1}^{N}(1-v_h)\right) = \prod_{h=1}^{N}E(1-v_h) =$$
$$= \prod_{h=1}^{N}(1-E(v_h)) = \prod_{h=1}^{N}\left(1-\frac{1}{\alpha+1}\right) = \left(\frac{\alpha}{\alpha+1}\right)^N \overset{N\to\infty}{\longrightarrow} 0.$$

A.3 Prediction of random effects

Proposition:

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

Proof:

According to (4) − (8) in Lindley and Smith (1972), it follows from

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

that

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

holds. By defining

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

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

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

Note that in $(*)$, the matrix lemma (10) in Lindley and Smith (1972) with $\boldsymbol{A}_1 := \boldsymbol{Z}_i^T$, $\boldsymbol{C}_1 := \hat{\boldsymbol{D}}^{-1}$ and $\boldsymbol{C}_2 := \hat{\boldsymbol{\Sigma}}_i^{-1}$ is used.

Thus without knowing the cluster membership one obtains

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

References

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

Blackwell D (1973) Discreteness of Ferguson selections. *The Annals of Statistics*, **1**, 356–58.

Blackwell D and MacQueen JB (1973) Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, **1**, 353–55.

Blei DM and Jordan MI (2006) Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, **1**, 121–44.

Booth JG, Casella G and Hobert JP (2008) Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society, Series B*, **70**, 119–39.

Burkhardt J (2008) *ASA047: Nelder-Mead minimization algorithm*. C++ library.

Bush CA and MacEachern SN (1996) A semiparametric Bayesian model for randomised block designs. *Biometrika*, **83**, 275–85.

Celeux G, Martin O and Lavergne C (2005) Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling*, **5**, 243–67.

De la Cruz-Mesía R, Quintana FA and Marshall G (2008) Model-based clustering for longitudinal data. *Computational Statistics & Data Analysis*, **52**, 1441–57.

Dempster AP, Laired NM and Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Dockery DW, Berkey CS, Ware JH, Speizer FE and Ferris BG (1983) Distribution of fvc and fev1 in children 6 to 11 years old. *American Review of Respiratory Disease*, **128**, 405–12.

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

Escobar MD (1994) Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, **89**, 268–77.

Escobar MD and West M (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577–88.

Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–30.

Fritsch A and Ickstadt K (2009) Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*, **4**, 367–92.

Gaffney SJ and Smyth P (2003) Curve clustering with random effects regression mixtures. In CM Bishop and BJ Frey (eds) *Proceedings of the ninth international workshop on artificial intelligence and statistics*. Key West, FL.

Gelfand AE and Kottas A (2002) A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **11**, 289–305.

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

Heinzl F, Fahrmeir L and Kneib T (2012) Additive mixed models with Dirichlet process mixture and P-spline priors. *Advances in Statistical Analysis*, **96**, 47–68.

Ishwaran H and James LF (2001) Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161–73.

Ishwaran H and James LF (2002) Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics*, **11**, 508–32.

James GM and Sugar CA (2003) Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, **98**, 397–408.

Jara A, Hanson TE and Lesaffre E (2009) Robustifying generalized linear mixed models using a new class of mixtures of multivariate Polya trees. *Journal of Computational and Graphical Statistics*, **18**, 838–60.

Komárek A, Hansen BE, Kuiper EMM, van Buuren HR and Lesaffre E (2010) Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Statistics in Medicine*, **29**, 3267–83.

Komárek A and Komárková L (2013) Clustering for multivariate continuous and discrete longitudinal data. *The Annals of Applied Statistics*. To appear.

Komárek A and Lesaffre E (2008) Generalized linear mixed model with a penalized Gaussian mixture as a random effects distribution. *Computational Statistics and Data Analysis*, **52**, 3441–58.

Kottas A and Gelfand AE (2001) Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, **96**, 1458–68.

Laird NM and Ware JH (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–74.

Li Y, Müller P and Lin X (2011) Center-adjusted inference for a nonparametric Bayesian random effect distribution. *Statistica Sinica*, **21**, 1201–23.

Lindley DV and Smith AFM (1972) Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, **34**, 1–41.

Lindstrom MJ and Bates DM (1988) Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data. *Journal of the American Statistical Association*, **83**, 1014–22.

Liu X and Yang MCK (2009) Simultaneous curve registration and clustering for functional data. *Computational Statistics & Data Analysis*, **53**, 1361–76.

MacEachern SN (1994) Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics - Simulation and Computation*, **23**, 727–41.

MacEachern SN and Müller P (1998) Estimating mixtures of Dirichlet process models. *Journal of Computational and Graphical Statistics*, **7**, 223–38.

Magder LS and Zeger SL (1996) A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *Journal of the American Statistical Association*, **91**, 1141–51.

Maritz JS and Lwin T (1989) *Empirical Bayes methods. Monographs on statistics and applied probability*. London: Chapman & Hall.

McAuliffe JD, Blei DM and Jordan MI (2006) Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistics and Computing*, **16**, 5–14.

McLachlan GJ and Krishnan T (1997) *The EM algorithm and extensions*. New York: Wiley.

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

Muliere P and Tardella L (1998) Approximating distributions of random functionals of Ferguson-Dirichlet priors. *The Canadian Journal of Statistics*, **26**, 283–97.

Müller P and Rosner GL (1997) A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association*, **92**, 1279–92.

Nelder JA and Mead R (1965) A simplex method for function minimization. *Computer Journal*, **7**, 308–13.

Newton MA and Zhang Y (1999) A recursive algorithm for nonparametric analysis with missing data. *Biometrika*, **86**, 15–26.

Ng SK, McLachlan GJ, Wang K, Ben-Tovim Jones L and Ng SW (2006) A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics*, **22**, 1745–52.

Ohlssen DI, Sharples LD and Spiegelhalter DJ (2007) Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons. *Statistics in Medicine*, **26**, 2088–112.

O'Neill R (1971) Algorithms AS 47: function minimization using a simplex procedure. *Journal of the Royal Statistical Society, Series C*, **20**, 338–45.

Papageorgiou G and Hinde J (2012) Multivariate generalized linear mixed models with semi-nonparametric and smooth nonparametric random effects densities. *Statistics and Computing*, **22**, 79–92.

Papaspiliopoulos O and Roberts GO (2008) Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, **95**, 169–86.

R Development Core Team (2012) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ruppert D, Wand MP and Carroll RJ (2003) *Semiparametric regression*. Cambridge: Cambridge University Press.

Scharl T, Grün B and Leisch F (2010) Mixtures of regression models for time course gene expression data: evaluation of initialization and random effects. *Bioinformatics*, **26**, 370–77.

Sethuraman J (1994) A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–50.

Stroustrup B (1997) *The C++ programming language* (3rd ed). Amsterdam: Addison-Wesley.

Verbeke G and Lesaffre E (1996) A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, **91**, 217–21.

Verbeke G and Molenberghs G (2000) *Linear mixed models for longitudinal data*. New York: Springer.

Villarroel L, Marshall G and Barón AE (2009) Cluster analysis using multivariate mixed effects models. *Statistics in Medicine*, **28**, 2552–65.

Walker SG (2007) Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, **36**, 45–54.

Weise FJ, Alt H and Becker R (eds) (2011) *Arbeitsmarkt in Zahlen*. Nürnberg: Statistik der Bundesagentur für Arbeit.

Yao C and Holmes C (2011) Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination. *Bayesian Analysis*, **6**, 329–51.