LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Felix Heinzl & Gerhard Tutz

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

Technical Report Number 115, 2011
Department of Statistics
University of Munich

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

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

Felix Heinzl & Gerhard Tutz

Department of Statistics, Ludwig-Maximilians-University Munich,

Akademistr. 1, 80799 Munich, Germany

SUMMARY: In linear mixed models the assumption of normally distributed random effects is often inappropriate and unnecessary restrictive. The proposed Dirichlet process mixture assumes a hierarchical Gaussian mixture. In addition to the weakening of distributions assumptions the specification allows to estimate clusters of observations with a similar random effects structure identified. An Expectation-Maximization algorithm is given that solves the estimation problem and that exhibits advantages over in this framework usually used Markov chain Monte Carlo approaches. The method is evaluated in a simulation study and applied to dynamics of unemployment in Germany as well as lung function growth data.

KEY WORDS: *Dirichlet process mixture; mixed models; likelihood inference; EM algorithm*

1 Introduction

Linear mixed models (LMM) are a common tool for the modeling of longitudinal data. The classical model has the form

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

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) is a Gaussian distribution for the random effects, i.e. $\boldsymbol{b}_i$ i.i.d. $N(\boldsymbol{0}, \boldsymbol{D})$, see for example Verbeke and Molenberghs (2000) and Ruppert et al. (2003). While this choice is mathematically convenient, in 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 based on estimates. Possible skewness and multimodality (arising, for example, from an unconsidered grouping structure in the data) may be masked when checking the normal distribution in terms of estimated random effects. A finite mixture of normal distributions as a random effects distribution suggested, for example, by Verbeke and Lesaffre (1996), Verbeke and Molenberghs (2000), and Grün (2008) is much more flexible. One assumes

$$\boldsymbol{b}_i \sim \sum_{h=1}^{N} \pi_h\, N(\boldsymbol{\mu}_h, \boldsymbol{D}), \tag{2}$$

where $\pi_1, \ldots, \pi_N$ are mixture weights. The number of mixture components is unknown and has to be chosen. A data driven choice of this number is desirable and could be achieved by a penalization of the mixture weights π_h. For example, Komarék and Lesaffre (2008) penalized differences between reparametrized 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 a new 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 a Dirichlet process mixture (DPM) for the random effects distribution and use the stick breaking procedure of the Dirichlet process (see Ferguson, 1973, for the theory behind the Dirichlet process and Sethuraman, 1994, for the stick breaking presentation of the Dirichlet process). The main advantage of Dirichlet processes 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

Dirichlet process allows to specify a prior on probability measures, it has been mainly used in the Bayesian inference for density estimation and random effects models. For linear mixed models, Dirichlet process priors for random effects were first proposed by Kleinman and Ibrahim (1998).

We aim at establishing the Dirichlet process as a tool for frequentist modeling. Therefore, instead of using Markov chain Monte Carlo (MCMC) methods, which are usually applied for estimation in random effects models with Dirichlet processes (compare for example Heinzl et al., 2011), we extend the traditional Expectation-Maximization (EM) algorithm (Dempster et al., 1977) used in the heterogeneity model of Verbeke and Molenberghs (2000) and call it DPM-EM model. We will show that the EM algorithm has an essential advantage over MCMC methods, where Dirichlet processes are concerned. In summary, on the one hand our DPM-EM model is a regularization approach for the number of mixture components in (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 Dirichlet processes are illustrated. In Section 2.2 we present our DPM-EM algorithm in detail. Simulation results can be seen in Chapter 3 while applications are shown in Chapter 4. Finally Chapter 5 subsumes the main aspects of our approach.

2 Linear mixed models with Dirichlet process mixtures

2.1 Model hierarchy

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

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

where $\boldsymbol{I}$ is the identity matrix 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

$$\begin{aligned} \boldsymbol{b}_i|\boldsymbol{\theta}_i &\overset{ind.}{\sim} N(\boldsymbol{\theta}_i, \boldsymbol{D}), && i = 1, \ldots, n, \\ \boldsymbol{\theta}_i &\overset{i.i.d.}{\sim} G, && i = 1, \ldots, n, \\ G &\sim DP(\alpha, G_0). \end{aligned} \tag{3}$$

Here, the Dirichlet process $DP(\alpha, G_0)$ is a distributional assumption for the unknown

mixing distribution G. A special feature of the Dirichlet process is, that each realization of G is a discrete probability measure. So in the DPM specification, choosing a Dirichlet process for the $\boldsymbol{\theta}_i,\ i = 1, \ldots, n$, creates ties among these and therefore forms clusters of subjects whereas each subject still has its own unique random effects value. In general, there are $k \leq n$ clusters and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ can be represented by cluster locations $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ and cluster allocation variables. Although in theory there is an automatic clustering structure induced by the Dirichlet process, some practical problems arise in the Bayesian context from using MCMC methods: One obtains a clustering of subjects within each iteration, but it is unclear how these can merged into an universal clustering. Several operations exist to handle this (see for example Fritsch and Ickstadt, 2009), but due to the high number of possible clusterings, these methods are typically not feasible in larger problems. The advantage of the EM algorithm over MCMC methods is that the EM algorithm converges to fixed values, while MCMC methods converge to distributions. So with EM type algorithms the cluster property of the Dirichlet process can be used directly.

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.

In practice, inference with Dirichlet processes can be handled by using the stick breaking representation of the Dirichlet process in its truncated version (see for example Ishwaran and James, 2002)

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

where $\delta_{\boldsymbol{\mu}_h}$ denotes the Dirac measure on $\boldsymbol{\mu}_h$. Hence, the unknown distribution G is represented as a weighted sum of point masses with random weights π_h linked to the locations $\boldsymbol{\mu}_h$. The weights can be constructed through the stick breaking procedure

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

where $Be(\cdot)$ denotes the beta distribution. In the truncated version $v_N = 1$ ensures that the sum of weights π_h is one. Sethuraman (1994) showed that for $\boldsymbol{\mu}_h \overset{i.i.d.}{\sim} G_0$ (in the limit $N \to \infty$) the probability measure of G is given by $DP(\alpha, G_0)$. The truncated version still is a good approximation because the random weights decrease stochastically as the index h grows (Ishwaran and James, 2001). This is obvious by the recursive definition of weights

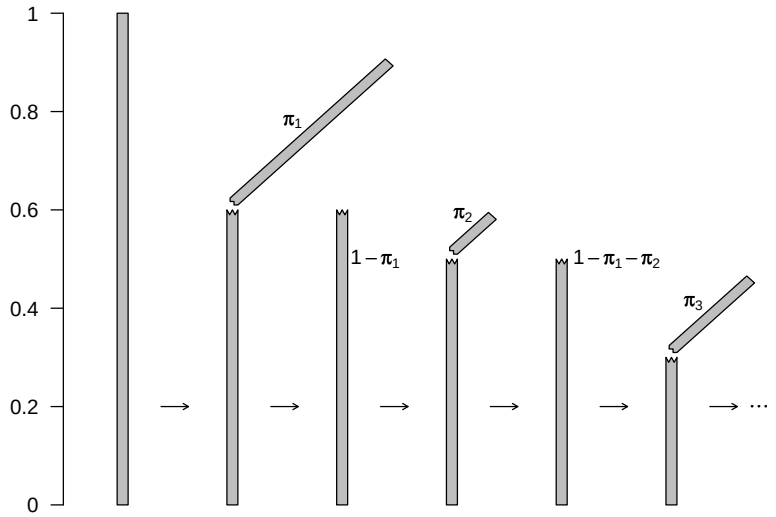


Fig. 1: Construction of $\pi_1, \pi_2, \ldots$ by stick breaking.

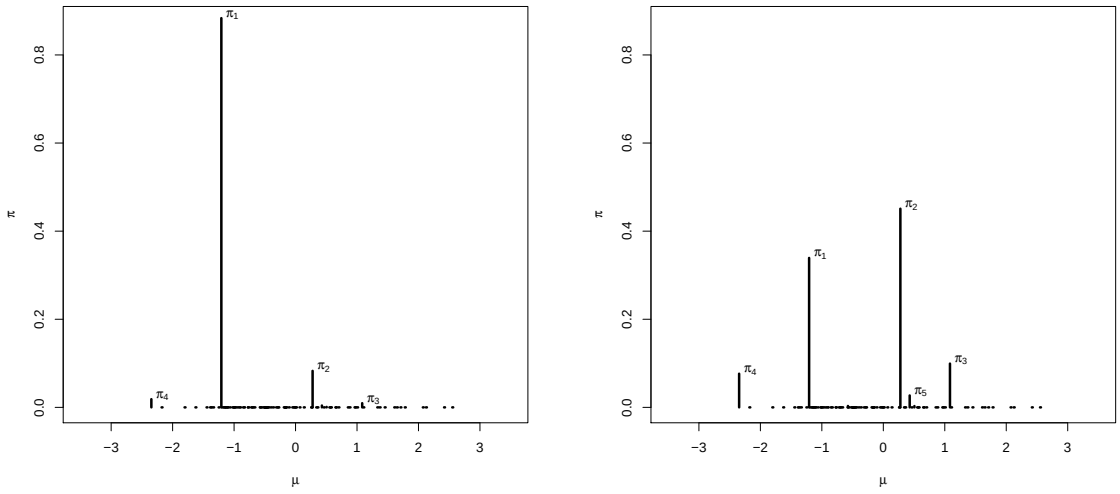
$$\pi_h = v_h \left(1 - \sum_{l<h} \pi_l\right), \quad h = 1, \ldots, N,$$

which gives the procedure its name and is visualized in Figure 1. It works as follows: first, for getting π_1 a piece is broken away from a stick of length one. Next, from the remainder of the stick, $1 - \pi_1$ breaks a further piece away and calls it π_2 and so on. So for large h the weights get very small and can be omitted. More mathematically, $E(\sum_{h=N+1}^{\infty} \pi_h)$ converges to zero exponentially with $N \to \infty$ (Ohlssen et al., 2007). It should be noted that N can also be seen as the maximum number of clusters. So in our simulations and applications, we truncate the stick breaking representation at $N = \min\{n, 100\}$.

In summary, by using the stick breaking procedure the distribution assumption for the random effects (3) can be rewritten as

$$\begin{array}{rcll} \boldsymbol{b}_i|\boldsymbol{v} & \overset{i.i.d.}{\sim} & \sum_{h=1}^{N} \pi_h \, N(\boldsymbol{\mu}_h, \boldsymbol{D}), & i = 1, \ldots, n, \\ \pi_h & = & v_h \prod_{l<h}(1 - v_l), & h = 1, \ldots, N, \\ v_h & \overset{i.i.d.}{\sim} & Be(1, \alpha), & h = 1, \ldots, N-1, \end{array} \tag{4}$$

where $\boldsymbol{v} = (v_1, \ldots, v_{N-1})^T$ symbolizes reparameterized weights. Therefore for the



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

Fig. 2: Realizations of $G \sim DP(\alpha, G_0)$ with $G_0 = N(0, 1)$

random effects distribution we get a finite mixture of normal distributions in which the number of mixture components with $\pi_h \neq 0$ is penalized. See Figure 2 for an illustration of two discrete probability measures simulated by Dirichlet processes 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. In the following the order of $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ is given by the corresponding weights in decreasing order under the restrictions $\sum_{h=1}^{N} \pi_h \boldsymbol{\mu}_h = \mathbf{0}$ and $\sum_{h=1}^{N} \pi_h = 1$. The first restriction ensures $E(\boldsymbol{y}_i) = \boldsymbol{X}_i \boldsymbol{\beta}$. The second constraint is standard and is automatically fulfilled by $v_N = 1$.

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 Molenberghs (2000) but includes a penalty term. Let $\boldsymbol{\xi} = (\alpha, \boldsymbol{v}, \boldsymbol{\gamma})^T$, where $\boldsymbol{\gamma}$ is the vector containing all the remaining parameters $\boldsymbol{\beta}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N, \boldsymbol{D}, \sigma^2$. The cluster membership of each individual can be described by the latent variable $\boldsymbol{z}_i := (z_{i1}, \ldots, z_{iN})^T$ 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 $\boldsymbol{z}_i$:

$$\begin{aligned} \boldsymbol{y}_i|\boldsymbol{z}_i &\overset{ind.}{\sim} N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\mu}_h,\, \boldsymbol{V}_i), && i = 1, \ldots, n, \\ \boldsymbol{z}_i|\boldsymbol{v} &\overset{i.i.d.}{\sim} M(1, \boldsymbol{\pi}), && i = 1, \ldots, n, \\ v_h &\overset{i.i.d.}{\sim} Be(1, \alpha), && h = 1, \ldots, N-1, \end{aligned} \tag{5}$$

with $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \sigma^2\boldsymbol{I}$ and $M(\cdot)$ symbolizing the multinomial distribution. This model can either be parameterized by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^T$ or by $\boldsymbol{v}$. Since the latter parametrization simplifies calculations it is used in the following. Nevertheless, only for a compact presentation, we write π_h instead of $v_h \prod_{l<h}(1 - v_l)$. The likelihood function corresponding to (5) is given by

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

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

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

This function can either be seen as log-posterior in the Bayesian context or as penalized log-likelihood whose penalization term results from the stick breaking procedure of the Dirichlet process. Obviously for $\alpha = 1$ the penalization term drops out. According to the general EM algorithm procedure we alternate between taking the expectation of $l_P(\boldsymbol{\xi})$ over all unobserved 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

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

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

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

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

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

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

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

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

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

$$\hat{v}_h = \frac{\sum_{i=1}^{n} \pi_{ih}}{\sum_{i=1}^{n}\sum_{l=h}^{N} \pi_{il} + \alpha - 1}, \qquad h = 1, \ldots, N-1. \tag{6}$$

and

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

Without further restrictions it could happen that $\hat{v}_h \notin [0, 1]$. To avoid this we use the following correction approach: Update $\hat{v}_h$ by (6) for increasing h. If $\hat{v}_{h^*} > 1$ set $\hat{v}_h$ to 1 for $h = h^*, \ldots, N-1$. This constraint for $\hat{\boldsymbol{v}}$ is equivalent to the following restriction on $\hat{\boldsymbol{\pi}}$ by using the stick breaking procedure:

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

where h^* is the lowest index h for which $\sum_{l=1}^{h} \hat{\pi}_l > 1$ is fulfilled. Here the idea of the penalization approach becomes evident. First note that for $\alpha = 1$ we get the usual estimates for $\hat{\pi}_h$ and no restrictions are needed. Compared to these estimates, for $\alpha \in (0, 1)$, all weights $\hat{\pi}_h$ for $h < h^*$ are stretched by the factor $\frac{n}{n+\alpha-1}$, while all weights $\hat{\pi}_h$ 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 general, the algorithm starts with $N = n$ clusters and successively merges clusters until there is no further ascent of the penalized incomplete log-likelihood. 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. In order to avoid $\log(0)$ we choose $\hat{v}_h = 1 - 10^{-300}$ instead of $\hat{v}_h = 1$ in the algorithm. Then $\hat{\pi}_h \approx 0$ for $h > h^*$.

In the second part of the M-step we get the estimate for $\boldsymbol{\gamma}$ by alternating separate maximization of $Q(\boldsymbol{\gamma}|\boldsymbol{\xi}^{(t)})$ to $\boldsymbol{\beta}$, $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ 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

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

Setting the derivative of $Q(\boldsymbol{\gamma}|\boldsymbol{\xi}^{(t)})$ with respect to $\boldsymbol{\mu}_h$, $h = 1, \ldots, N$, given $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{D}}$ and $\hat{\sigma}^2$ yields

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

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

Stop criterion

The EM algorithm stops if the penalized incomplete log-likelihood is not ascending any more. Then 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 k. Based on the weights of all clusters the random effects are predicted by

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

This result can be shown by using derivations from Lindley and Smith (1972).

Choice of starting values

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.

Implementation

All computations are implemented in C++, allowing for an efficient treatment of loop-intensive calculations and with regard to slow convergence of the EM algorithm, and are made easily accessible by providing an R wrapper function. All variables are standardized internally for calculations. For updating variance parameters we use an implementation of the Nelder-Mead algorithm in C++ (library ASA047). 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 behavior. 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 data sets assuming a simple linear

trend model

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

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

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

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

$$\sigma^2 = 0.25, \qquad \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \qquad \boldsymbol{D} = \begin{pmatrix} \sigma^2_{b_0} & \sigma_{b_{01}} \\ \sigma_{b_{01}} & \sigma^2_{b_1} \end{pmatrix} = \begin{pmatrix} 0.02 & 0.01 \\ 0.01 & 0.02 \end{pmatrix}.$$

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 λ. Setting $\lambda = 1$ corresponds to longitudinal data with only few (3 on average) repeated observations per unit, $\lambda = 3$ to a moderate number and $\lambda = 5$ to (comparably) large numbers of repeated observations.

For given n_i, observation times are generated from

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

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 observations 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 and with results of models using a unpenalized ($\alpha = 1$) finite normal mixture as random effects distribution. In each simulation run s, we calculate the average prediction error

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

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.

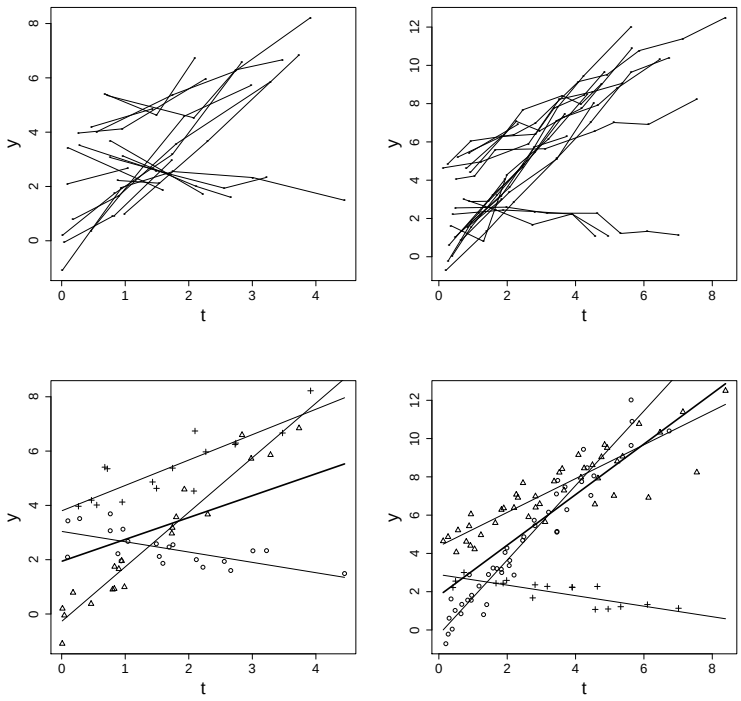


Fig. 3: Trace plots (top) and clustering by DPM-EM model (bottom) with clearly separated clusters for few individual observations ($\lambda = 1$) (left) and a moderate number of observations on individuals ($\lambda = 3$) (right).

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 thick line shows the overall effect and the thin lines visualize the means of the resulting clusters. Which observation is assigned to which cluster is marked by the same symbol.

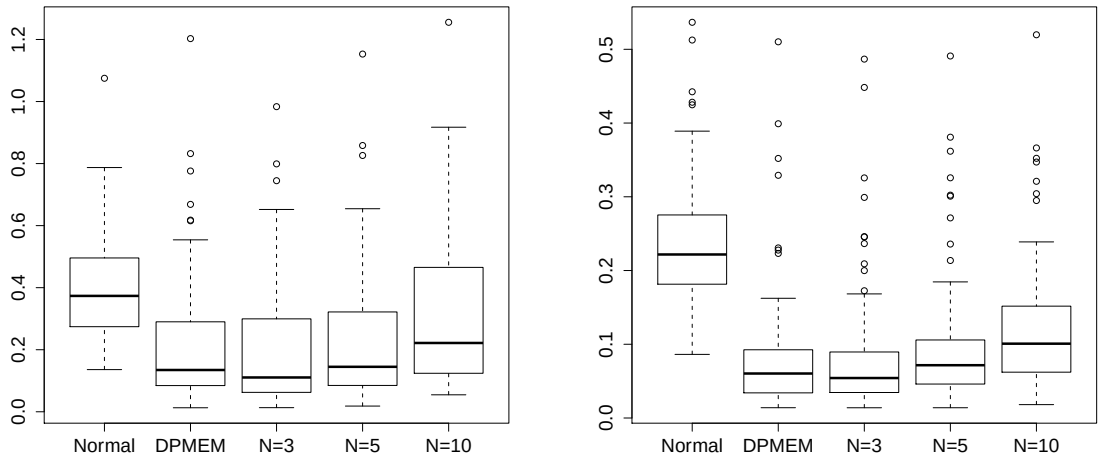


Fig. 4: Box plots of PE_0 with clearly separated clusters for few individual observations ($\lambda = 1$) (left) and a moderate number of observations on individuals ($\lambda = 3$) (right).

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, for example, 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 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.

	$\lambda = 1$				$\lambda = 3$				$\lambda = 5$			
	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

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

Moderately separated clusters

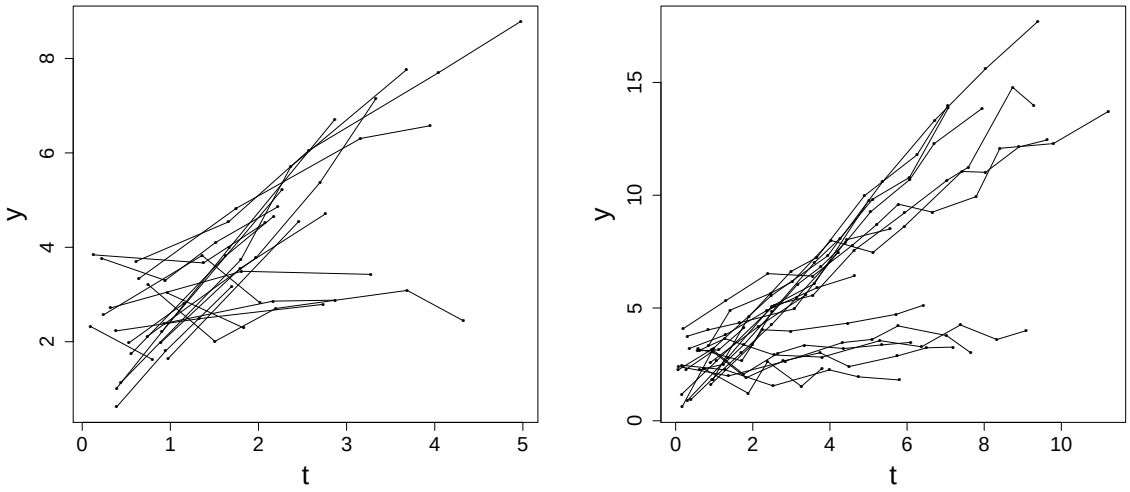


Fig. 5: Trace plots with moderate separated clusters for few individual observations ($\lambda = 1$) (left) respectively many individual observations ($\lambda = 5$) (right).

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. Note that the superiority of the DPM-EM model over the classical linear mixed model with normal random effects distribution is even higher in the case of many individual observations.

	$\lambda = 1$				$\lambda = 3$				$\lambda = 5$			
	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

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

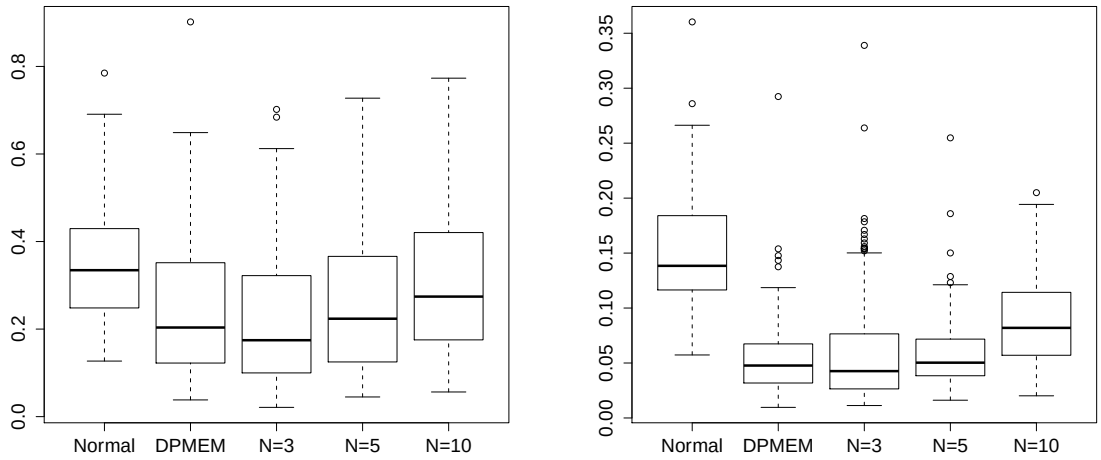


Fig. 6: Box plots of PE_0 with moderate separated clusters for few individual observations ($\lambda = 1$) (left) respectively many individual observations ($\lambda = 5$) (right).

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 effect 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.

	$\lambda = 1$				$\lambda = 3$				$\lambda = 5$			
	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

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

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 without risk

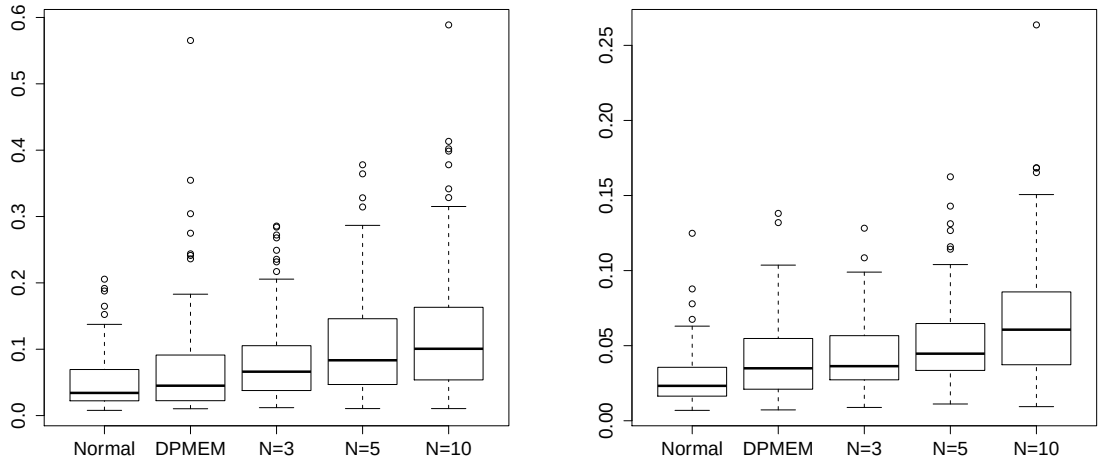


Fig. 7: Box plots of PE_0 with only one cluster for few individual observations ($\lambda = 1$) (left) respectively many individual observations ($\lambda = 5$) (right).

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

$$y_{ij}|\boldsymbol{b}_i \overset{ind.}{\sim} N(\beta_0 + b_{i0} + (\beta_1 + b_{i1})year_{ij}, \sigma^2), \quad i = 1, \ldots, 16, \quad j = 1, \ldots, 6.$$

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 (4).

We are looking for clustering the federal states in order to expose which states show similar behavior. Only for a better interpretability we change the zero point of the

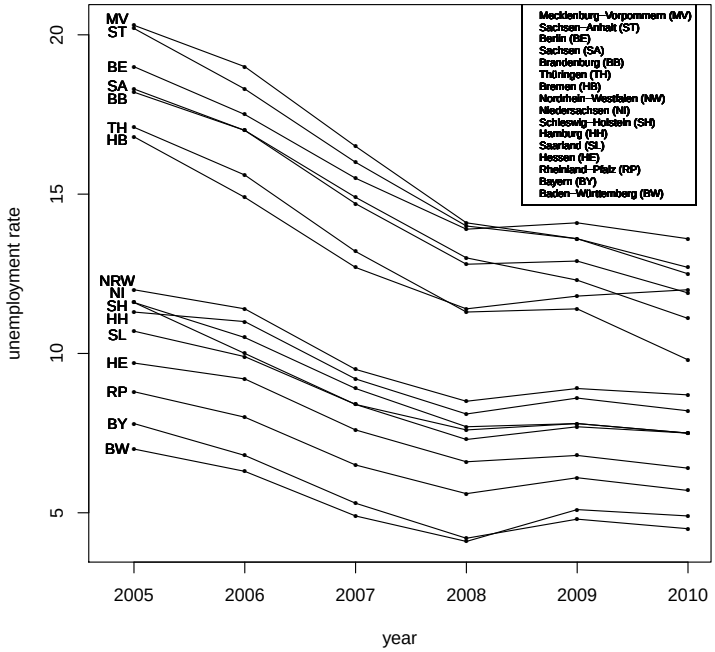


Fig. 8: Unemployment in the federal states of Germany across time

time variable to 2005. Thus, during calculations the time variable is labeled by 0, 1, ..., 5 for the years 2005, 2006, ... 2010.

Figure 9 shows the population effect (thick line) as well as the cluster effects (thin lines). Observations belonging to the same cluster are marked with the same symbol. Our DPM-EM model detects three clusters with estimated weights $\pi_1 = 0.467$, $\pi_2 = 0.425$ and $\pi_3 = 0.108$: 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.

$\boldsymbol{\beta}$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\mu}_3$
13.719	4.361	-3.139	-6.468
-1.007	-0.353	0.277	0.436

Table 4: Estimators for the fixed effects and the cluster locations.

Table 4 shows that here the base level in 2005 is -6.468 lower compared to the overall unemployment rate 13.719. In the south also the decrease of the unemployment rate is less distinct than in the other states. A similar effect can be observed in cluster 2. Here, the gap to the global intercept is considerably smaller. Furthermore, there is one cluster with a much more higher base level and a stronger decrease of the unemployment rates. It is remarkable that these states are all in Eastern Germany

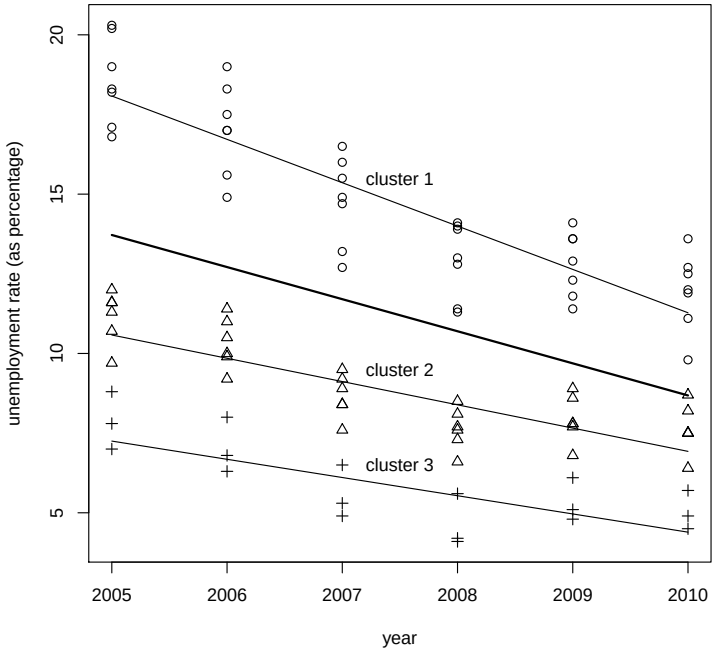


Fig. 9: Clustering of federal states by DPM-EM model

or city states. Only the city state Hamburg makes an exception to that feature and belongs to cluster 2.

			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	0.941	0.059
	7	Rheinland-Pfalz	0	0.421	0.579
state i	8	Baden-Württemberg	0	0.007	0.993
	9	Bayern	0	0.012	0.988
	10	Saarland	0	0.997	0.003
	11	Berlin	1	0	0
	12	Brandenburg	1	0	0
	13	Mecklenburg-Vorpommern	1	0	0
	14	Sachsen	1	0	0
	15	Sachsen-Anhalt	1	0	0
	16	Thüringen	1	0	0

Table 5: Matrix of $\hat{\pi}_{ij}$.

Table 5 shows the estimated probabilities π_{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 estimates of α 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 the logarithmic forced expiratory volume in one second (fev1). Our sample consists of 100 girls, with a minimum of two and a maximum of twelve observations over time. Again, we use a linear mixed model with random intercepts and random slopes

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

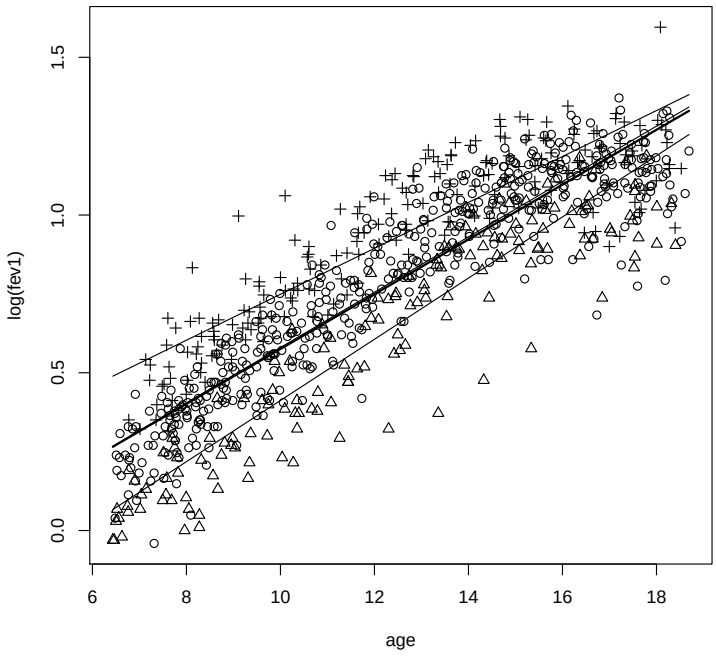


Fig. 10: Clustering of lung function growth data by DPM-EM model

and a DPM as random effects distribution (4). 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 $\boldsymbol{\mu}_h$, the thin small ones the random effects $\boldsymbol{b}_i$. Girls which assigned to the same cluster are marked with the same symbol and are arranged around the three cluster locations in the form of ellipses.

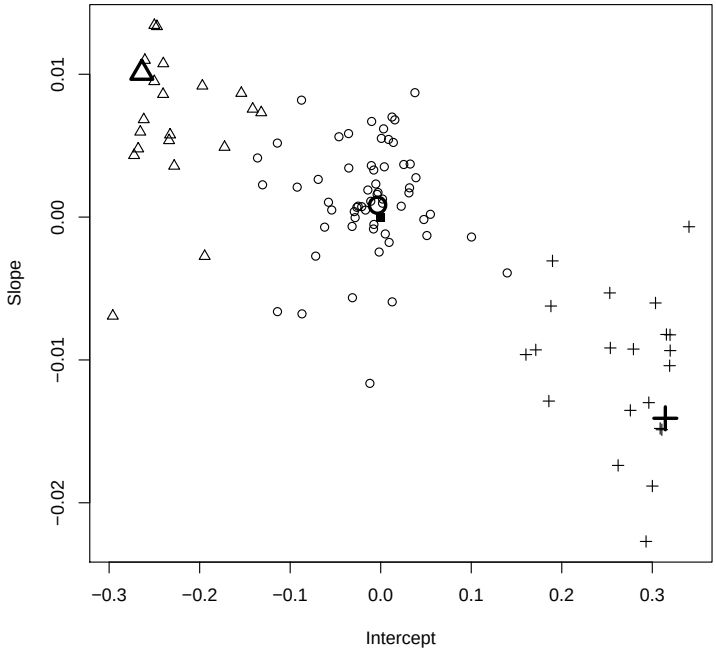


Fig. 11: Cluster locations and corresponding random effects for lung function growth data

5 Conclusion

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

References

Dempster, A. P., N. M. Laired, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B 39*, 1–38.

Dockery, D. W., C. S. Berkey, J. H. Ware, F. E. Speizer, and B. G. Ferris (1983). Distribution of fvc and fev1 in children 6 to 11 years old. *American Review of Respiratory Disease 128*, 405–412.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*, 209–230.

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

Grün, B. (2008). Fitting finite mixtures of linear mixed models with the EM algorithm. In P. Brito (Ed.), *Compstat 2008—Proceedings in Computational Statistics*, Volume II, Heidelberg, pp. 165–173. Physica Verlag.

Heinzl, F., L. Fahrmeir, and T. Kneib (2011). Additive mixed models with Dirichlet process mixture and P-spline priors. *Advances in Statistical Analysis*. Accepted for publication on 2011-05-02.

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

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

Kleinman, K. and J. Ibrahim (1998). A semiparametric Bayesian approach to the random effects model. *Biometrics 54*, 921–938.

Komarék, A. and E. Lesaffre (2008). Generalized linear mixed model with a penalized Gaussian mixture as a random effects distribution. *Computational Statistics and Data Analysis 52*, 3441–3458.

Lindley, D. V. and A. F. M. Smith (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society B 34*, 1–41.

Lindstrom, M. J. and D. M. Bates (1988). Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data. *Journal of the American Statistical Association 83*, 1014–1022.

Magder, L. S. and S. L. Zeger (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *Journal of the American Statistical Association 91*, 1141–1151.

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.

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

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*. New York: Springer.

Weise, F.-J., H. Alt, and R. Becker (Eds.) (2011). *Arbeitsmarkt in Zahlen*. Nürnberg: Statistik der Bundesagentur für Arbeit.