Technical report from Automatic Control at Linköpings universitet

Bayesian inference for mixed effects models with heterogeneity

Johan Dahlin, Robert Kohn, Thomas B. Schön

Division of Automatic Control

E-mail: johan.dahlin@liu.se, r.kohn@unsw.edu.au,
thomas.schon@it.uu.se

1st April 2016

Report no.: LiTH-ISY-R-3091

Address:
Department of Electrical Engineering
Linköpings universitet
SE-581 83 Linköping, Sweden

WWW: http://www.control.isy.liu.se

AUTOMATIC CONTROL
REGLERTEKNIK
LINKÖPINGS UNIVERSITET

Technical reports from the Automatic Control group in Linköping are available from http://www.control.isy.liu.se/publications.

Abstract

We are interested in Bayesian modelling of panel data using a mixed effects model with heterogeneity in the individual random effects. We compare two different approaches for modelling the heterogeneity using a mixture of Gaussians. In the first model, we assume an infinite mixture model with a Dirichlet process prior, which is a non-parametric Bayesian model. In the second model, we assume an over-parametrised finite mixture model with a sparseness prior. Recent work indicates that the second model can be seen as an approximation of the former. In this paper, we investigate this claim and compare the estimates of the posteriors and the mixing obtained by Gibbs sampling in these two models. The results from using both synthetic and real-world data supports the claim that the estimates of the posterior from both models agree even when the data record is finite.

Keywords: Bayesian inference, mixed effects model, panel/longitudinal data, Dirichlet process mixture, finite mixture, sparseness prior

Avdelning, Institution
Division, Department

Division of Automatic Control
Department of Electrical Engineering

Datum
Date

2016-04-01

Språk
Language

□ Svenska/Swedish
☒ Engelska/English
□ ________________

Rapporttyp
Report category

□ Licentiatavhandling
□ Examensarbete
□ C-uppsats
□ D-uppsats
☒ Övrig rapport
□ ________________

ISBN
—

ISRN
—

Serietitel och serienummer
Title of series, numbering

ISSN
1400-3902

LiTH-ISY-R-3091

URL för elektronisk version

`http://www.control.isy.liu.se`

Titel
Title

Bayesian inference for mixed effects models with heterogeneity

Författare
Author

Johan Dahlin, Robert Kohn, Thomas B. Schön

Sammanfattning
Abstract

We are interested in Bayesian modelling of panel data using a mixed effects model with heterogeneity in the individual random effects. We compare two different approaches for modelling the heterogeneity using a mixture of Gaussians. In the first model, we assume an infinite mixture model with a Dirichlet process prior, which is a non-parametric Bayesian model. In the second model, we assume an over-parametrised finite mixture model with a sparseness prior. Recent work indicates that the second model can be seen as an approximation of the former. In this paper, we investigate this claim and compare the estimates of the posteriors and the mixing obtained by Gibbs sampling in these two models. The results from using both synthetic and real-world data supports the claim that the estimates of the posterior from both models agree even when the data record is finite.

Nyckelord
Keywords

Bayesian inference, mixed effects model, panel/longitudinal data, Dirichlet process mixture, finite mixture, sparseness prior

Bayesian inference for mixed effects models with heterogeneity

Johan Dahlin, Robert Kohn and Thomas B. Schön*

April 1, 2016

Abstract

We are interested in Bayesian modelling of panel data using a mixed effects model with heterogeneity in the individual random effects. We compare two different approaches for modelling the heterogeneity using a mixture of Gaussians. In the first model, we assume an infinite mixture model with a Dirichlet process prior, which is a non-parametric Bayesian model. In the second model, we assume an over-parametrised finite mixture model with a sparseness prior. Recent work indicates that the second model can be seen as an approximation of the former. In this paper, we investigate this claim and compare the estimates of the posteriors and the mixing obtained by Gibbs sampling in these two models. The results from using both synthetic and real-world data supports the claim that the estimates of the posterior from both models agree even when the data record is finite.

Keywords: Bayesian inference, mixed effects model, panel/longitudinal data, Dirichlet process mixture, finite mixture, sparseness prior.

*E-mail to corresponding author: *johan.dahlin@liu.se*. JD is with the Department of Electrical Engineering, Linköping University, Sweden. RK is with the Department of Economics, University of New South Wales Business School, Sydney, Australia. TS is with the Department of Information Technology, Uppsala University, Sweden.

1 Introduction

In many fields, we are interested in modelling the dependence of an observation y_{it} of individual i at time t given some regressors/covariates x_{it}. This type of data is known as panel data in economics [Baltagi, 2008] and longitudinal data in statistics [Verbeke and Molenberghs, 2009]. That is, when we obtain multiple observations T of a number of N individuals over time. A common application is health surveys in which annual questionnaires are sent out to a group of individuals. The focus is then to isolate different factors that are correlated with disease or visits to the doctor. However, the importance of these factors can vary between different sub-groups in the population and this is referred to as heterogeneity. For prediction purposes, it is therefore important to captures these variations and be able to identify to which sub-group a specific individual belongs.

Another popular application is recommendation systems for online retailing or streaming sites such as Amazon, Netflix and Spotify. The underlying algorithms vary between different sites and are often proprietary information unknown to the public. However, academic work in recommendation system by Condliff et al. [1999] and Ansari et al. [2000] have made used of panel data models. The common theme for both applications are that the number of observations T for each individual is typically much smaller than the size N of the population. It is therefore essential to pool information together from similar individuals to construct a model from data.

For this end, we consider a linear mixed effects model [Verbeke and Lesaffre, 1996] with observation $y_{it} \in \mathbb{R}$ for individual $i = 1, \ldots, N$ at time $t = 1, \ldots, T$. This model can be expressed as

$$y_{it} \,|\, \alpha, \beta_i^s, \sigma_e^2, x_{it} \sim \mathcal{N}\Big(y_{it}, \alpha x_{it} + \beta_i^s z_{it}, \sigma_e^2/\omega_i\Big), \tag{1}$$

where $\alpha \in \mathbb{R}^d$ denotes the fixed effects and $\beta_i^s = \{\beta_{ij}^s\}_{j=1}^p \in \mathbb{R}^p$ denotes the random effects for individual i. Here, x_{it} and z_{it} denote the design matrices connected to the fixed and random effects, respectively. Furthermore, $\sigma_e > 0$ denotes the standard deviation of noise affecting the observations and $\omega_i > 0$ denotes the individual scaling of the variance to allow for variance heterogeneity. We denote a Gaussian distribution with mean μ and standard deviation $\sigma > 0$ by $\mathcal{N}(\mu, \sigma^2)$.

In this paper, we assume that β_i^s can be modelled as an infinite mixture of Gaussians. This means that the random effects can for example be distributed according to a multi-modal distribution. Each mode would then potentially correspond to a certain sub-group of the population with similar behaviour. This information could be important in marketing, economics, medicine and other applications as discussed by Allenby et al. [1998], Canova [2004] and Lopes et al. [2003]. A potential benefit of having an infinite mixture is that the data determines the number of components to include in the model.

An infinite mixture of Gaussians can be expressed by

$$\beta_i^s \sim \sum_{k=1}^{\infty} \eta_k \mathcal{N}(\beta_i^s; \beta_k, Q_k), \qquad \text{s.t.} \ \sum_{k=1}^{\infty} \eta_k = 1, \tag{2}$$

for some weights $\eta_k > 0$, mean vector $\beta_k \in \mathbb{R}^p$ and covariance matrix $Q_k \in \mathbb{R}^{p \times p}$. We proceed to model this mixture in a Bayesian setting in Section 2, where a Dirichlet process (DP; Ferguson, 1973, 1974) prior is employed. This is a Bayesian non-parametric method, which can act as a prior for probability distributions such as (2). However, the inference often relies on Markov chain Monte Carlo (MCMC; Robert and Casella, 2004), which can be challenging to implement and sometimes mixes poorly. The latter is discussed by Hastie et al. [2015] and could be a problem when applying this kind of model for big data.

The main contribution of this paper is to compare two different models for the heterogenity described by (2). In the first approach, we make use of a DP model for (2). In the second approach, we truncate the infinite mixture to obtain an overparametrised finite mixture (FM) for which we make use of a sparseness prior. The latter model is often simpler to implement and can also enjoy better mixing properties. Ishwaran and Zarepour [2002] and Rousseau and Mengersen [2011] have analysed the properties of this approximation. In short, the results are that the approximation is asymptotically consistent (in NT) and converges to the solution obtained when using a DP. The FM is over-parametrised but the sparseness prior empties the superfluous components. Our aim is to compare these two approaches for mixed effects models in terms of the estimate of the posterior and the mixing in the Markov chain constructed by MCMC algorithms.

The comparisons are made on both synthetic and real-world data. The results indicate that the posterior estimates are similar in most cases and that the similarity increases as N and T tend to infinity. The mixing is compared using a Monte Carlo simulation with synthetic data. The resulting mixing seems to be similar for the finite and infinite mixture models. Therefore, there is nothing lost or gained by using either model compared with the other. More work is needed to see if this result holds for even larger sets of data and alternative sampling schemes.

2 Bayesian mixture modelling

In this section, we discuss the details of the two approaches for modelling the mixture of the random effects (2). The first approach uses the DP prior of the infinite mixture of Gaussians and the second approach approximates the infinite mixture by a FM. We return to the problem of sampling from the posterior of the parameters in the mixed effects model and the mixture for the random effects in Section 3.

2.1 Infinite mixture model using a Dirichlet process

In the first approach, we model the random effects β_i^s using the infinite mixture model in (2) with the DP [Ferguson, 1973, 1974] as a prior. The mixture model can be seen as a hierarchical Bayesian model, which we develop step by step in this section.

The DP is an example of a Bayesian non-parametric model, where the number of parameters grow with the number of observations and can be viewed as infinite. A realisation G from a DP is a random discrete probability distribution in the form of an empirical distribution. Hence, we can express G [Gelman et al., 2013] by

$$G = \sum_{k=1}^{\infty} \eta_k \delta_{\vartheta_k}, \tag{3}$$

where the weights $\{\eta_k\}_{k=1}^{\infty}$ and locations $\{\vartheta_k\}_{k=1}^{\infty}$ are random variables. Here, $\delta_{\vartheta'}$ denotes a Dirac measure placed at ϑ', where ϑ denotes the parameters of the mixture component. Furthermore, we have that $\sum_{k=1}^{\infty} \eta_k = 1$ with probability 1, which means that G can be interpreted as a probability measure.

Let $\mathcal{DP}(\eta_0, G_0)$ denote a DP with the *concentration parameter* $\eta_0 > 0$ and the *base measure* G_0. We say that G is distributed according to a DP if all of its marginal distributions are Dirichlet distributed. Let $\mathcal{D}(\eta)$ denote the Dirichlet distribution with concentration parameter $\eta = \{\eta_1, \ldots, \eta_R\}$. Hence, if G_0 is a probability measure on the space $(\Omega, \mathcal{F})$, we have that

$$\Big(G(A_1), G(A_2), \ldots, G(A_N)\Big) \sim \mathcal{D}\Big(\eta_0 G_0(A_1), \eta_0 G_0(A_2), \ldots, \eta_0 G_0(A_N)\Big), \tag{4}$$

for any finite (measurable) partition $A_{1:N}$ of Ω. Note that the expected value of G is the base measure and therefore G has the same support as G_0. Moreover, G is discrete with probability one even if the base measure is continuous, which is useful in mixture models as discussed below.

Assume that we obtain some data generated from the model given by

$$G \sim \mathcal{DP}(\eta_0, G_0), \qquad \vartheta_i \,|\, G \sim G, \qquad i = 1, 2, \ldots.$$

In many applications, we would like to compute the predictive distribution for some new parameter $\vartheta_\star$ given the observations $\vartheta_{1:N}$. The predictive distribution can be computed as a marginalisation given by

$$p(\vartheta_\star \,|\, \vartheta_{1:N}) = \int G(\vartheta_\star) p(G|\vartheta_{1:N}) \mathrm{d}G,$$

which is possible to carry out in closed-form. The result is a so-called Pólya urn scheme discussed by

Blackwell and MacQueen [1973]. This scheme can be expressed mathematically

$$\vartheta_\star \,|\, \vartheta_{1:N} \sim \frac{\eta_0}{\eta_0 + N} G_0 + \frac{1}{\eta_0 + N} \sum_{i=1}^{N} n_i \delta_{\vartheta_i}. \tag{5}$$

Here, n_i denotes the number of parameters that are identical to ϑ_i, i.e.,

$$n_i = \sum_{j=1}^{N} \mathbb{I}\,[\vartheta_j = \vartheta_i],$$

where $\mathbb{I}\,[A]$ denotes the indicator function which assumes the value one if A is true and zero otherwise.

From (5), we note that the DP is a discrete process with a non-zero probability of ties, i.e., that two or more realisations are identical to each other. From the Pólya urn scheme, we either draw a new parameter from the base measure or an existing parameter from the Dirac mixture. The probability of sampling a new parameter from the base measure is determined by concentration parameter. We are more likely to sample from the base measure if $\eta_0 \gg N$, which means that the predictive posterior concentrates to the base measure. If the concentration parameter is small, we often sample from the existing parameters, which gives many ties and a strong clustering behaviour.

Hence, we can make use of this clustering effect to reformulate (2) as a Dirichlet process mixture (DPM; Antoniak, 1974) given by

$$G \sim \mathcal{DP}(\alpha, G_0), \qquad \beta, Q \,|\, G \sim G, \qquad \beta_i^s \sim \mathcal{N}(\beta, Q), \tag{6}$$

where we choose $\vartheta = \{\beta, Q\}$ for this particular model. The presence of ties means that several β_i^s will be drawn from a Gaussian distribution with the same parameters. Hence, this is an alternative presentation of (2).

We have now introduced the DPM and discussed the properties of the underlying DP as well as how to compute the predictive posterior distribution. What remains is to discuss how to simulate from a DPM and how to compute the prior-posterior update given some data. In this paper, we make use of the stick-breaking by Sethuraman [1994] to simulate from the mixture. The procedure iterates

$$\beta_k, Q_k \sim G_0, \qquad \eta_k = V_k \prod_{l<k} (1 - V_l), \qquad V_k \sim \mathcal{B}(1, \eta_0), \tag{7}$$

for $k = 1, \ldots, R$, where R is an upper bound on the number of components in the mixture selected by the user. Here, $\mathcal{B}(a, b)$ denotes a Beta distribution with shape $a > 0$ and scale $b > 0$. Moreover, $\prod_{l<k}(1 - V_l)$ denotes the length remaining of a unit stick after breaking off $k - 1$ pieces, where each V_k denotes the fraction of the remaining stick to break off.

Note that the truncation implied by the stick-breaking corresponds a mixture given by

$$\beta_i^s \,|\, \eta_{1:R}, \beta_{1:R}, Q_{1:R} \sim \sum_{r=1}^{R} \eta_r \mathcal{N}(\beta_i^s; \beta_r, Q_r).$$

In many cases, we can choose R to be larger than the number of individuals N, so this is not a problem when n is rather small. To see the connection between the sick-breaking and (5), we note that the expected value of the Beta distribution sampled from in (7) is $(1+\eta_0)^{-1}$. Hence, the expected part of the stick to break off tends to zero when $\eta_0 \to \infty$ and tends to one when $\eta_0 \to 0$. In the latter case, we usually obtain a few components, which is in agreement with the previous discussion of the concentration parameter. We return to computing the posterior of a DPM in Section 3 using Gibbs sampling.

2.2 Finite mixture model

The second approach that we consider is the FM, which can be recovered as a special case of the DPM model (6) given by

$$\beta_i^s \,|\, S_i, \beta_{1:R}, Q_{1:R} \sim \mathcal{N}(\beta_{S_i}, Q_{S_i}), \qquad S_i \,|\, \eta \sim \mathcal{M}(1, \eta_1, \ldots, \eta_R), \tag{8a}$$

$$\beta_k, Q_k \sim G_0, \qquad \eta \sim \mathcal{D}(\eta_{0,1}, \ldots, \eta_{0,R}), \tag{8b}$$

where the latent variable S_i denotes which component that individual i belongs to. Here, $\mathcal{M}(1,p)$ denotes the multinomial distribution with one try and event probabilities given by p. The parameters of the mixture components are still realisations from the base measure. However, the DP is now a Dirichlet distribution, which follows from the property of the DP discussed in connection with (4). The Dirichlet distribution determines the probability of a certain cluster membership S_i. Hence, we get a similar clustering effect as for the DPM if we select η_0 to be small. This is essentially the same result as we had for the Pólya urn scheme in (5).

An interesting property discussed by Ishwaran and Zarepour [2002] and Rousseau and Mengersen [2011] is that the FM can approximate the DPM. This means that the sparsity will empty unnecessary components and that the estimate of the posterior tend to the true one. However, we need to be careful when setting the hyperparameters in the Dirichlet prior distribution for this approximation to work. On choice advocated by Ishwaran and Zarepour [2002] is to use the concentration parameter $\eta_{0,1} = \ldots = \eta_{0,R} = \eta_0/R$ with $\eta_0 = 1$. Here, the number of components R can be selected as n if the amount of data is small. Here, the small value of η_0 results in that a few samples from the Dirichlet distribution are allocated most of the probability mass. Therefore, we get a sparsity effect as only a few components are occupied a priori when $\eta_0 \leq 1$.

We discuss how to sample from the posterior of the FM (8) in Section 3

3 Sampling from the posterior

In this section, we make use of MCMC sampling to estimate the posterior of the parameters in the mixed effects model (1) and the mixture modelling the random effects (2) given the data $\{y, x, z\} = \{\{y_{it}, x_{it}, z_{it}\}_{i=1}^{n}\}_{t=1}^{T}$. The parameter vector is given by $\theta = \{\alpha, \beta_{1:R}, Q_{1:R}, \beta^{s}_{1:N}, \omega_{1:N}, \sigma^2_e, S_{1:N}, \eta_{1:R}\}$, which includes the parameters of the DPM (6) and the FM (8). To decrease the notational burden, we have used the same notation for the parameters in both mixture models.

We make use of conjugate priors to obtain closed-form conditional posteriors, which allows for Gibbs sampling as discussed by Gelman et al. [2013], Frühwirth-Schnatter [2006] and Neal [2000]. Note that there are many other interesting alternatives to Gibbs sampling for estimating the posterior of a DPM. Sequential Monte Carlo algorithms [Del Moral et al., 2006] discussed by Fearnhead [2004] and Ulker et al. [2010] do not have problems with mixing but can be challenging to implement.

Slice samplers are also an interesting alternative discussed by Walker [2007], which are easy to implement and give moderate or good mixing. Finally, split-merge algorithms can give good mixing as discussed by Jain and Neal [2004] and Bouchard-Côté et al. [2015] but can be challenging to implement. For the FM, a simple Gibbs sampling approach often gives good mixing and is easy to implement.

3.1 Prior distributions

To compute the posterior, we need to assign prior distributions for each of the elements in θ. Here, we make use of the prior distributions from Frühwirth-Schnatter [2006, p. 265] for all the parameters except for the mixture weights $\eta_{1:R}$. For the fixed effects and the noise variance with heterogeneity, we choose

$$\alpha \sim \mathcal{N}(\alpha; a_0, A_0), \qquad \sigma^2_e \sim \mathcal{G}^{-1}(\sigma^2_e; c^e_0, C^e_0), \qquad \omega_i \sim \mathcal{G}\Big(\frac{\nu}{2}, \frac{\nu}{2}\Big), \tag{9}$$

for individuals $i = 1, \ldots, n$. Here, we introduce the notation $\mathcal{G}(a, b)$ for the Gamma distribution with shape $a > 0$ and rate $b > 0$, which means that the expected value is ab^{-1}. The inverse Gamma distribution is denoted by $\mathcal{G}^{-1}(a, b)$ with the expected value $b(a-1)^{-1}$. Hence, we have the hyperparameters $\{a_0, A_0, c^e_0, C^e_0, \nu\}$ for the user to choose.

We also need to choose a number of priors for the mixture model (2) describing the distribution of the random effects. Here, we make use of the priors

$$\beta_k \sim \mathcal{N}(\beta_k; b_0, B_0), \qquad Q^{-1}_k \sim \mathcal{W}(Q^{-1}_k; c^Q_0, C^Q_0), \tag{10}$$

where $\mathcal{W}(n, V)$ denotes the Wishart distribution with $n > 0$ degrees of freedom and scale matrix $V > 0$, respectively. Hence, we have $\{b_0, B_0, c^Q_0, C^Q_0\}$ as additional hyperparameters for the user to choose.

Algorithm 1 Gibbs sampling for mixture models

```
INPUTS: ${\{y_{it}, x_{it}, z_{it}\}_{i=1}^N\}_{t=1}^T$ (data), $\theta_0$ (hyperparameters).
OUTPUTS: $\theta'$ (approximate sample from the parameter posterior).

During one iteration of the Gibbs sampler:
1: Update parameters given the cluster allocation.
    (i: FM) Sample $\eta'_r | S_{1:N}$ using (12) for $r = 1, 2, \ldots, R$.
    (i: DPM) Sample $\eta'_r | S_{1:N}$ using (13) for $r = 1, 2, \ldots, R$.
    (ii) Sample $\alpha^{\star,\prime} | y, x, z, Q_{1:R}, \omega_{1:N}, S_{1:N}$ using (14) with $\alpha^\star = \{\alpha, \beta_{1:R}\}$.
    (iii) Sample $Q'_r | \beta^s_{1:N}, \beta'_{1:R}, S_{1:N}$ using (15) for $r = 1, 2, \ldots, R$.
    (iv) Sample $\sigma_e^{2,\prime} | y, x, z, \alpha', \beta^s_{1:N}, \omega_{1:N}$ using (16).
2: Sample allocations $S'_i | y, x, z, \alpha', \beta_i^{s,\prime}, \beta'_{1:R}, Q'_{1:R}, \omega_{1:N}$ using (17) for $i = 1, 2, \ldots, N$.
3: Sample the random effects $\beta_i^{s,\prime} | y_i, x_i, z_i, \alpha', \beta'_{S'_i}, Q'_{S'_i}, \omega_i$ using (18) for $i = 1, 2, \ldots, N$.
4: Sample the variance heterogeneity $\omega'_i | y, x, z, \alpha', \beta_i^{s,\prime}, \sigma_e^{2,\prime}$ using (19) for $i = 1, 2, \ldots, N$.
5: [DPM] Sample the concentration parameter $\eta'_0$ using (20).
```

Furthermore, we assign a prior for the concentration parameter in the DPM given by

$$\eta_0 \sim \mathcal{G}(c_0^\eta, C_0^\eta), \tag{11}$$

for some hyperparameters c_0^η and C_0^η.

3.2 Gibbs sampling

To sample from the posterior, we make use of blocked Gibbs sampling algorithms. For the DPM model, we make use of the algorithm proposed by Gelman et al. [2013, p. 552], which is a truncated approximation using at most R clusters. The stick-breaking procedure is used to sample from the DPM in this formulation.

For the FM, we make use of the Gibbs sampler proposed by Frühwirth-Schnatter [2006, p. 368], which samples from the mixture using the conjugacy between the Dirichlet prior distribution and the multinomial data. The remaining steps of the Gibbs sampler are identical and the full procedure performed during one iteration of the sampler is presented in Algorithm 1. Note that the differences between the two alternatives for modelling appears in Steps 1(i) and 5.

For the FM, we sample the mixture weights from the posterior given by the conjugacy property as previously discussed. This results in the sampling scheme

$$\eta'_{1:R} \sim \mathcal{D}(\eta_0/R + n_1, \eta_0/R + n_2, \ldots, \eta_0/R + n_R), \tag{12}$$

where n_r denotes the number of individuals in component r. The details are discussed by Gelman et al. [2013, p. 534]. For the DPM model, we make use of the stick-breaking construction in (7) to generate the weights $\eta'_{1:R}$ for the truncated model, i.e.,

$$\eta'_r = V_r \prod_{l<r}(1-V_l), \qquad V_r \sim \mathcal{B}\left(1+n_r, \eta_0 + N - \sum_{j=1}^{r} n_j\right), \tag{13}$$

for $r = 1, \ldots, R$ and starting with a stick of unit length. The remaining parameters are sampled from the conditional posteriors derived in the subsequent section.

3.3 Conditional posteriors

In this section, we outline the details for sampling from each of the conditional posteriors in Algorithm 1. In Step 1(ii), we are required to sample from the conditional of $\alpha^\star \triangleq \{\alpha, \beta_1, \ldots, \beta_R\}$, which are the fixed effects and the mean of each component in the mixture (2). This essentially requires us to solve a regression problem where the regressors are given by $z_i = (x_i, z_i D_{i1}, \ldots, z_i D_{iK})$ with $D_{ik} = 1$ if $S_i = k$ and zero otherwise. Hence, we rewrite the regression model to obtain

$$y_i = z_i \alpha^\star + \widetilde{e}_i,$$

where $\widetilde{e}_i \sim \mathcal{N}(0, V_i)$ with $V_i = z_i Q_{S_i}(z_i)^\top + \sigma_e^2/\omega_i \mathbf{I}_T$. We can therefore compute the posterior using a standard Bayesian linear regression with known covariance and Gaussian prior for the regression coefficients. The conditional posterior is given by

$$\alpha^\star \,|\, y_{1:N}, Q_{1:R}, \sigma_e^2, \omega_{1:N}, S_{1:N} \sim \mathcal{N}_{d+Kp}(\alpha^\star; a_N^\star, A_N^\star), \tag{14}$$

where the mean and the covariance are given by

$$(A_N^\star)^{-1} = \sum_{i=1}^{N} (z_i^\star)^\top V_i^{-1} z_i^\star + (A_0^\star)^{-1}, \qquad a_N^\star = A_N^\star \left(\sum_{i=1}^{N} (z_i^\star)^\top V_i^{-1} y_i + (A_0^\star)^{-1} a_0^\star\right).$$

In Step 1(iii), we sample the covariance of the mixture components. The posterior is given by the conjugacy of the Wishart prior for the covariance matrix and a Gaussian likelihood. We can compute the conditional posterior for $r = 1, \ldots, R$ by

$$Q_r^{-1} \,|\, \beta_{1:R}, \beta_{1:N}^s, S_{1:N} \sim \mathcal{W}\left(Q_r^{-1}; c_r^Q, C_r^Q\right), \tag{15}$$

where the sufficient statistics (assuming independence between B_0 and C_r^Q) are given by

$$c_r^Q = c_0^Q + \frac{n_r}{2}, \qquad C_r^Q = C_0^Q + \frac{1}{2} \sum_{\{i: S_i = r\}} (\beta_i^s - \beta_r)(\beta_i^s - \beta_r)^\top,$$

where again n_r denotes the number of individuals in component r, i.e.,

$$n_r = \sum_{i=1}^{N} \mathbb{I}(S_i = r).$$

In Step 1(iv), we sample the variance of the noise σ_e^2, which we assume to have an inverse-Gamma prior distribution. As the likelihood is Gaussian, we obtain the conjugate posterior given by

$$\sigma_e^2 \,|\, y_{1:N}, \alpha, \beta_{1:N}^s, \omega_{1:N} \sim \mathcal{G}^{-1}\left(\sigma_e^2; c_k^e, C_k^e\right), \tag{16}$$

where the sufficient statistics are given by

$$c_k^e = c_0^e + \frac{NT}{2}, \qquad C_k^e = C_0^e + \frac{1}{2}\sum_{i=1}^{N} \omega_i \big(y_i - \alpha x_i - \beta_i^s z_i\big)^2.$$

Note that the data in this update is given by the scaled residuals for each of the individuals.

In Step 2, we update the allocation of each individual to a component by sampling S_i for $i = 1, \ldots, N$. The latent variable S_i is sampled from a multinomial distribution, where the probabilities reflect the likelihood that individual i belongs to a certain component. That is, we sample

$$S_i \,|\, \alpha, \beta_{1:R}, Q_{1:R}, \omega_i, \sigma_e^2, y_i \sim \mathcal{M}(1, p_{1:R}), \tag{17}$$

where the probabilities are given by

$$p_r = \frac{\eta_r \mathcal{N}\big(y_i; \alpha x_i + \beta_k z_i, W_{ik}\big)}{\sum_{l=1}^{R} \eta_l \mathcal{N}\big(y_i; \alpha x_i + \beta_l z_i, W_{il}\big)}, \qquad W_{il} = z_i Q_l (z_i)^\top + \frac{\sigma_e^2}{\omega_i}\mathbf{I}_T.$$

Note that the probabilities follow from the mixture model (2) directly. The intuition is that we are more likely to select the component that best explains the data in terms of its contribution to the likelihood.

In Step 3, we sample the random effects for each individual from the component to which the individual is assigned. The prior for each individual is given by $p(\beta_i) = \mathcal{N}(\beta_{S_i'}, Q_{S_i'})$ and the likelihood is Gaussian and given by

$$y_i - \alpha x_i = z_i \beta_i^s + \frac{\sigma_e}{\sqrt{w_i}} e_i,$$

where e_i is a standard Gaussian random variable. Note that this again is a Bayesian linear regression with a Gaussian prior for the regression coefficient β_i^s, where the observations are given by $y_i - \alpha x_i$ and the regressors are x_i. In this case, the variance is known and the conditional posterior for $i = 1, \ldots, N$ is

$$\beta_i^s \,|\, y_i, \alpha, \beta_{S_i}, Q_{S_i}, \omega_i \sim \mathcal{N}(\beta_i^s; b_i^s, B_i^s), \tag{18}$$

where the sufficient statistics are given by

$$B_i^s = \left(Q_{S_i}^{-1} + (z_i)^\top z_i \frac{\omega_i}{\sigma_e^2} \right)^{-1}, \qquad b_i^s = B_i^s \left(Q_{S_i}^{-1} \beta_{S_i} + (z_i)^\top (y_i - \alpha x_i) \frac{\omega_i}{\sigma_e^2} \right),$$

In Step 4, we sample the individual scaling of the noise of the observations. The derivation is similar to for Step 1(iv), but here we have a Gamma distributed prior for ω_i and a Gaussian likelihood. The conditional posterior is given by

$$\omega_i \,|\, y_{1:N}, \alpha, \beta_{1:N}^s, \sigma_e^2 \sim \mathcal{G}\left(\omega_i; c_k^\omega, C_k^\omega\right), \tag{19}$$

where the sufficient statistics are given by

$$c_k^\omega = \frac{\nu}{2} + \frac{T}{2}, \qquad C_k^\omega = \frac{\nu}{2} + \frac{1}{2\sigma_e^2}\big(y_i - \alpha x_i - \beta_i^s z_i\big)^2.$$

In Step 5, we sample the concentration parameter η_0 for the DPM model using the Gamma prior in (11). In this case, the conditional posterior [Gelman et al., 2013, p. 553] is given by

$$\eta_0' \sim \mathcal{G}\left(c_0^\eta + N - 1, C_0^\eta - \sum_{r=1}^{R-1} \log(1 - V_r) \right), \tag{20}$$

where V_r is the fraction broken off during the stick-breaking in Step 1(i) of the procedure.

4 Numerical illustrations

In this section we present three numerical experiments to compare the difference of modelling heterogeneity in the random effects using the FM and DPM model. The aim is to compare the posterior estimates and the mixing in the Markov chain created by the Gibbs sampler. The latter is important as it determines the inefficiency of the sampling and the asymptotic variance of the posterior estimates.

4.1 Mixture of Gaussians

We begin with a simple mixture of Gaussians model [Escobar and West, 1995] to validate our implementation and present the setup that we use for the comparisons in this section. In this model, we simplify the mixed effects model (1) such that $y_i = \beta_i^s$, where the heterogeneity of the random effects is modelled by (2). We generate $T = 1$ observations for $N = 100$ individuals using $R = 3$ components and $\eta_{1:3} = \{0.3, 0.5, 0.2\}$, $\beta_{1:3} = \{-1, 0, 2\}$ and $Q_{1:3} = \{0.2^2, 1, 0.05^2\}$. We carry out the inference using Algorithm 1 with the settings presented in Appendix A.

The results are presented in Figure 1. In the upper part, we present the estimates of the posterior

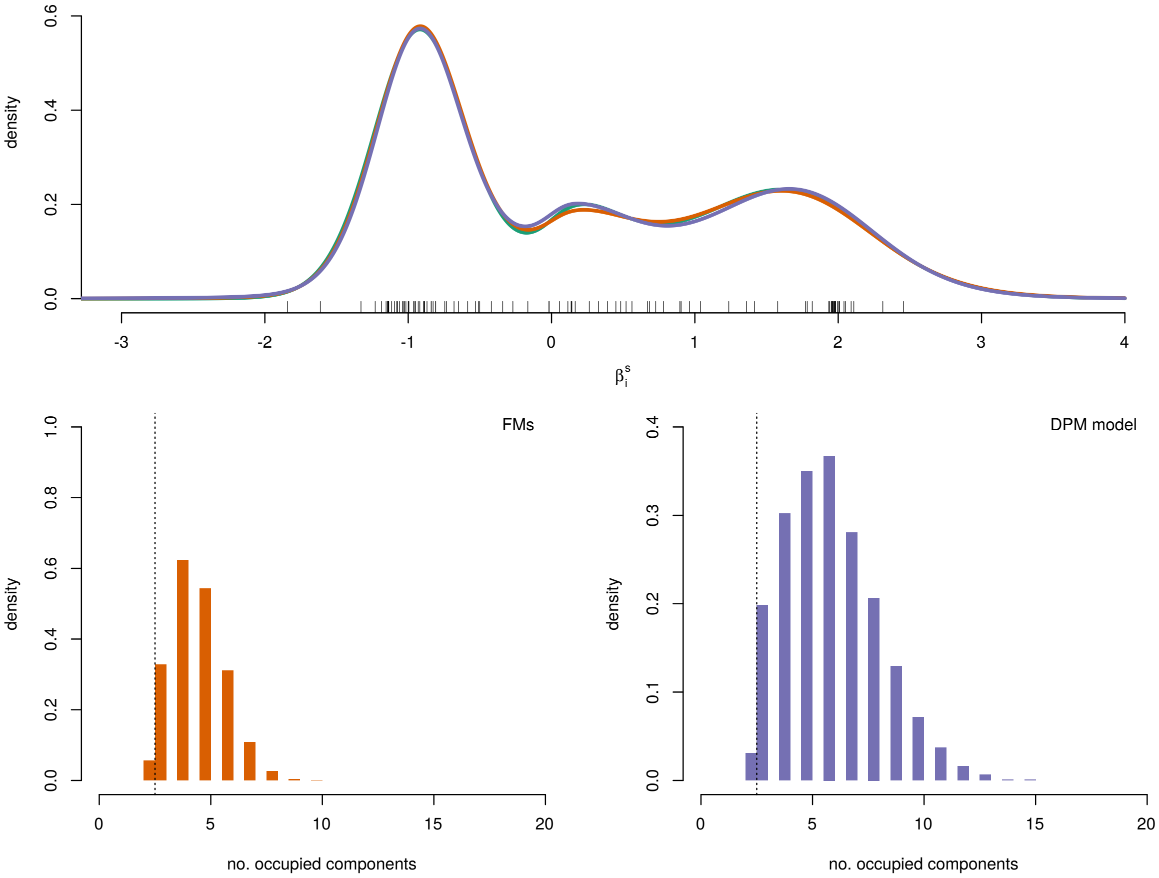

Figure 1: Upper: the posterior estimate from a finite mixture with 3 components (green), the FM (orange) and DPM model (purple). The *rug* indicate the observed data. Lower: the number of active clusters, where vertical dotted lines indicate the true number of clusters.

of β_i^s using three different models: (i) a FM with the true number of components $R = 3$, (ii) a FM with $R = 20$ components and (iii) a DPM model. We note that the three models give almost the same estimate of the posterior, which indicates that the sparsity effect of the FM works satisfactory and empties the extra components. This can be seen in the lower part of the same figure as less than 20 components are active at any time in the FM. Furthermore, the DPM model use on average more components than the FM. However, both models often make use of more components than in the model from which we generated the data.

4.2 Mixed effects model with synthetic data

We now consider the complete mixed effects model (1) with heterogenity in the individual random effects (2). We simulate 40 independent data sets using $T = 100$ observations in each while varying the number of individuals N in $\{10, 20, 50, 100, 200, 500\}$. We make use of a mixture for the random effects with $R = 3$ components and parameters

$$\eta_{1:3} = \{0.4, 0.3, 0.3\}, \qquad \beta_{1:3} = \{2, 3, -2\}, \qquad Q_{1:3} = \{1, 0.2, 0.2\}.$$

Moreover, we make use of $\sigma_e^2 = 1$, $\alpha = 2$ and $\omega_{1:N} \sim |\mathcal{N}(0, 1)|$ for the remaining parameters of the model. We carry out the inference using Algorithm 1 with the settings presented in Appendix A.

We proceed by comparing the difference in the posterior estimates of the random effects. This is done by computing the mean squared error (MSE) between the kernel density estimates (KDEs) of the sought posterior using the FM and DPM model. Hence for each data set, we compute the KDEs of the posterior estimates and compute the squared distance between them. In Table 1, we present the results which indicate that the MSE (after an initial increase) tends to decrease when N grows. We therefore conclude that the posterior estimates tend to become similar as the amount of data increases. Furthermore, we present some graphical comparisons in Figure 2 for four Monte Carlo runs. We see that the posterior estimates are similar most of the time, which is promising and validates the conclusion from the MSE comparison.

We compare the mixing in the two models using the integrated autocorrelation time (IACT) also known as the inefficiency factor for the Gibbs sampler applied to the two models. The IACT is estimated by

$$\widehat{\mathsf{IACT}}\big(\varphi(\theta_{K_b:K})\big) = 1 + 2 \sum_{l=1}^{L} \widehat{\rho}_l\big(\varphi(\theta_{K_b:K})\big), \tag{21}$$

where K_b denotes the number of iterations in the *burn-in* and K denotes the number of iterations of the Gibbs sampler. Here $\widehat{\rho}_l(\varphi(\theta_{K_b:K}))$ denotes the empirical autocorrelation at lag l of some test function φ

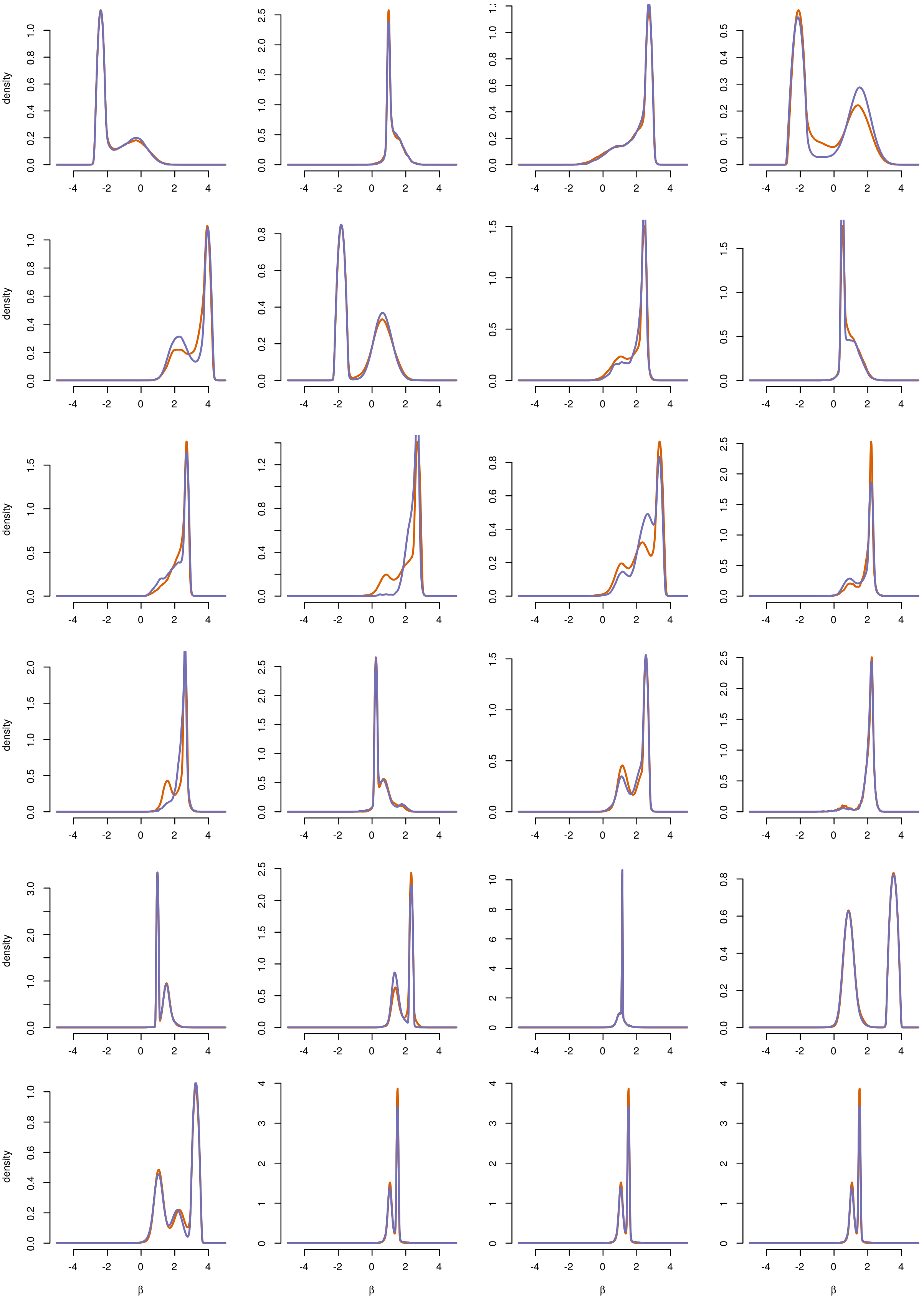

Figure 2: Estimates of $p(\beta_i^s|y)$ from a FM (orange) and DPM model (purple) for $n = \{10, 20, 50, 100, 200, 500\}$ (rows) and 4 independent Monte Carlo runs (columns).

		$N10$	$N=20$	$N=50$	$N=100$	$N=200$	$N=500$
Log-MSE		-6.1	-5.5	-5.1	-3.7	-4.1	-4.9
$\widehat{\mathsf{IACT}}(\psi_1)$	FM	16 (27)	22 (48)	36 (75)	19 (61)	2 (5)	1 (1)
	DPM	11 (25)	24 (34)	29 (45)	10 (46)	2 (4)	1 (1)
$\widehat{\mathsf{IACT}}(\psi_2)$	FM	4 (1)	7 (3)	8 (4)	8 (3)	7 (1)	11 (5)
	DPM	12 (10)	12 (10)	13 (9)	11 (10)	8 (5)	8 (4)

Table 1: The logarithm of the MSE of the estimates and the IACT of the log-likelihood and the number of occupied components obtained in the FM and in a DPM model. The IACT values are the median with the interquartile range in parenthesise from 40 Monte Carlo simulations.

that depends on θ. Here, we make use of the log-likelihood and the number of occupied components to compute the IACT. The log-likelihood for the mixed effects model (1) is given by

$$\psi_1(\theta) = \sum_{i=1}^{N} \sum_{t=1}^{T} \log \mathcal{N}\left(y_{it}; \alpha x_{it} - \beta_i^s z_{it}, \frac{\sigma_e^2}{\omega_i} \right).$$

The number of occupied clusters is given by

$$\psi_2(\theta) = \sum_{k=1}^{R} \mathbb{I}(n_r > 0),$$

where n_r denotes the number of individuals in component r. A small value of the IACT for both quantities indicates that we obtain many uncorrelated samples from the sought posterior. This implies that the chain is mixing well and that the asymptotic variance of the parameter estimates is rather small. We choose $L = \max\{3, \lfloor 0.5(M - M_b) \rfloor\}$ and make use of the `IAT` command in the R-package `LaplacesDemon` [Hall, 2012] to compute the estimate of the IACT.

Table 1 also presents the median IACT with the interquartile range (the length between the first and third quartiles) from 40 Monte Carlo runs over different data sets. We compare the IACT for the FM and the DPM model for both the log-likelihood and the number of occupied components. We note that there are no significant differences between the two models. Also, the IACT in the log-likelihood seems to decrease with increasing N and the opposite holds for the mixing in the number of occupied components.

4.3 Mixed effects model with sleep deprivation data

Finally, we consider the data set presented by Belenky et al. [2003] of the reaction times in a sleep deprivation trial of $N = 18$ individuals. We present the data in Figure 3 from the test during $T = 10$ consecutive days during which the subjects are only allowed to sleep three hours every night. Moreover, we present the linear regression for each individual indicated by the green lines. It seems that there are two different types of individuals in the data. In the first sub-group, the reaction time significantly increases for each day of sleep deprivation, e.g., individuals 337, 308 and 350. In the second sub-group, the lack of sleep does not seem to have a large impact on the reaction times, e.g., individual 351, 335 and 309. Hence, we can assume that the distribution of the random effects is possibly multi-modal, where the modes represent these sub-groups.

We make use of the model discussed by Bates et al. [2015] to try to capture this effect. Let $\{\{y_{it}\}_{i=1}^{N}\}_{t=1}^{T}$ denote the reaction time in millisecond for individual i on day t. We assume that the observations can be modelled by a mixed effects model given by

$$y_{it}\,|\,\alpha, \beta_i^s, \sigma_e^2 \sim \mathcal{N}\left(y_{it}; (\alpha_0 + \beta_{i0}^s) + (\alpha_1 + \beta_{i1}^s)(t-1), \frac{\sigma_e^2}{\omega_i}\right). \tag{22}$$

Furthermore, we assume that the individual random effects can be modelled using a mixture of Gaussians (2). The model parameters are $\theta = \{\alpha_0, \alpha_1, \beta_{1:R,0}, \beta_{1:R,1}, \sigma_e^2, \omega_{1:N}\}$. The interpretation of this model is that the mean reaction time and change in reaction time for every day of sleep deprivation are given by α_0 and α_1. The individual variations of these two quantities are captured by β_{i0}^s and β_{i1}^s, respectively.

Figure 4 presents the estimates of the posterior of the fixed and random effects using the FM (orange) and the DPM model (purple). From the upper part, we see the estimates of the mean reaction time at the first day when the participants are well-rested together with the average increase for each day. The estimates of the posterior means $\{\widehat{\alpha}_0, \widehat{\alpha}_1\}$ are $\{237, 1.15\}$ and $\{237, 1.07\}$ for the FM and the DPM model, respectively. This corresponds to that the mean reaction time the first day of the test is 237 milliseconds, which then increases by 1.15 or 1.07 milliseconds for every day of sleep deprivation.

From the lower part, we note that posterior estimate for β_{i0}^s is uni-modal with some skewness to the right. This indicates that most individuals have the same or an increased reaction time on the first day compared with the corresponding fixed effect α_0. However, from the posterior regarding β_{i1}^s, we note that there seems to be three different sub-groups with: a small decrease, small increase and large increase in reaction times for each day of sleep deprivation. This validates some of our initial findings that some individuals have a small (or even negative) change in reaction time when sleep deprived. Finally, the estimates of the random effects posterior means $\{\widehat{\beta^s}_{i0}, \widehat{\beta^s}_{i1}\}$ are $\{9.50, 8.94\}$ and $\{9.50, 9.09\}$ for the FM and the DPM model, respectively.

5 Conclusions

We have compared two different models for describing heterogeneity of the random effects in a mixed effects model. The numerical illustrations show that the two approaches give similar estimates of the posteriors using both synthetic and real-world data. This provides us with some promising indications that the results from Rousseau and Mengersen [2011] holds for this type of models. However, more

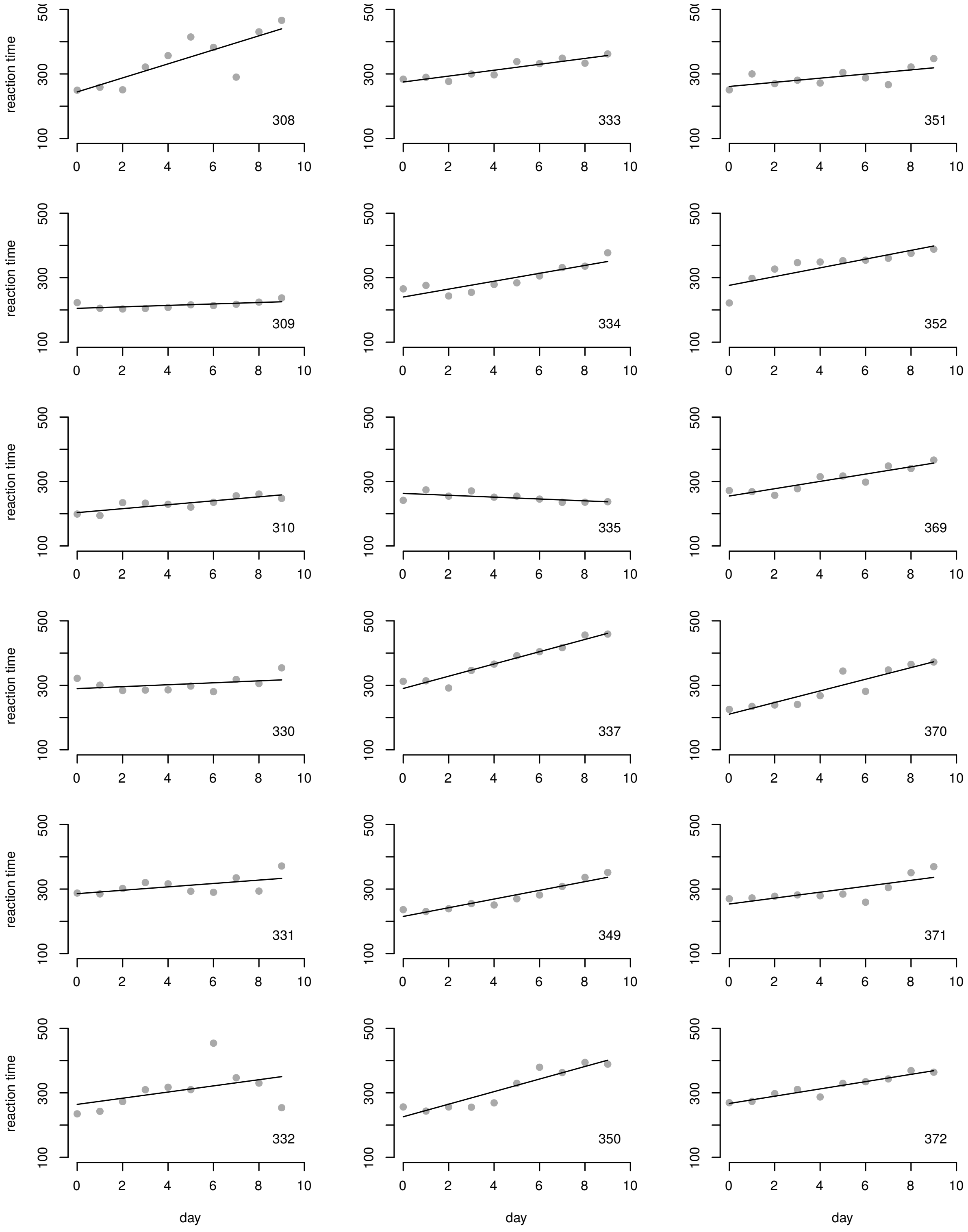

Figure 3: The reaction time in milliseconds for $N = 18$ individuals during $T = 10$ consecutive days with only three hours of sleep per night. The solid lines indicate the best linear fit for each individual and the dots indicate the data points.

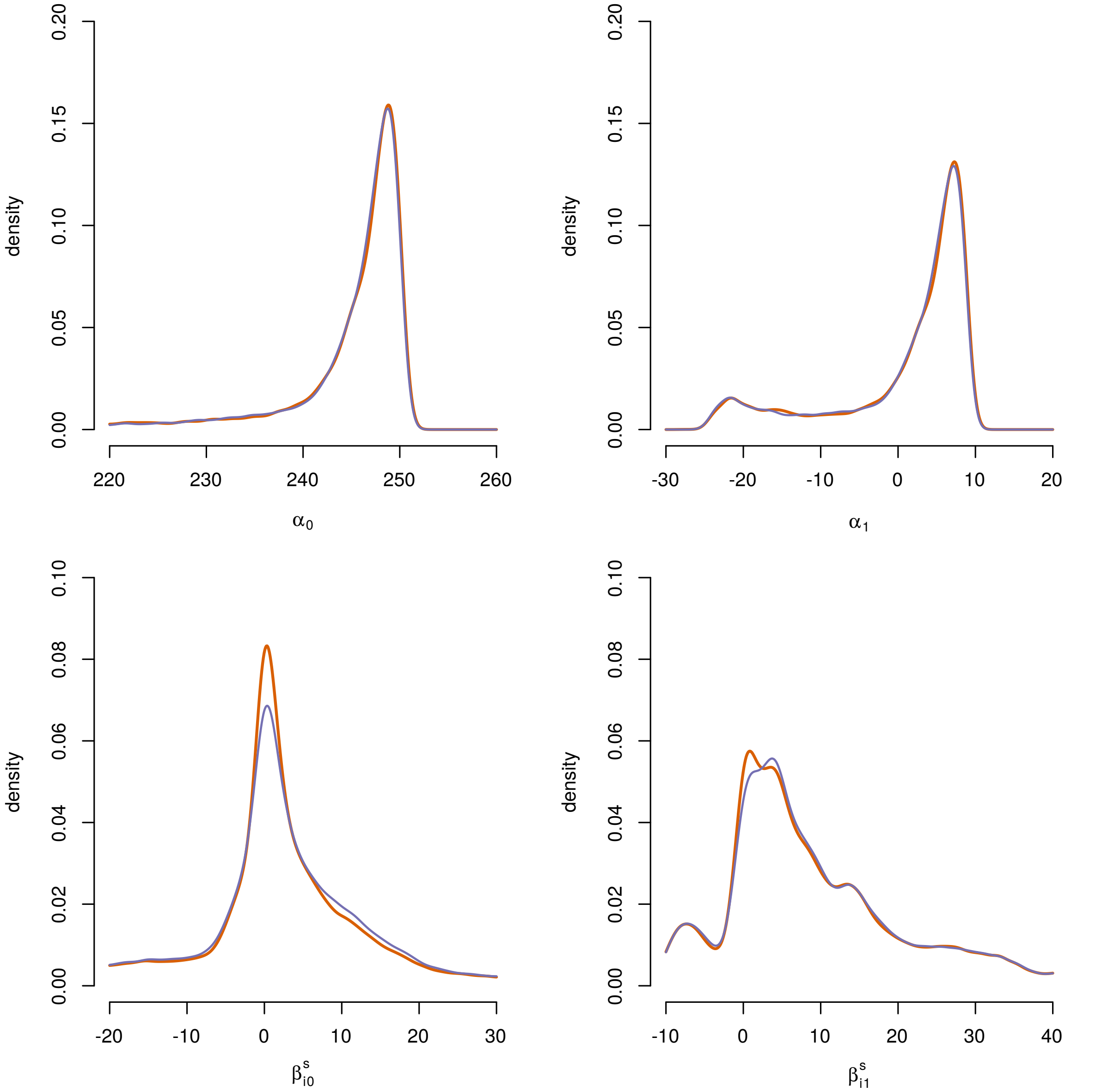

Figure 4: Posterior estimates of α_0 (upper left), α (upper right), β^s_{0i} (lower left) and β^s_{1i} (lower right) for the FM (orange) and the DPM model (purple).

theoretical work is required to generalise the convergence results to this type of model and priors.

Another important future work is to apply this approach for more realistic real-world models as discussed by e.g., Burda and Harding [2013]. It would also be useful to conduct more extensive simulations to compare the mixing in the Markov chain to see if the approximate FM can help mitigating the problems with poor mixing as discussed by Hastie et al. [2015].

The source code and data for the numerical illustrations are available from `https://github.com/compops/panel-dpm2016/`.

Acknowledgements

Financial support for this paper came from the projects *Probabilistic modeling of dynamical systems* (Contract number: 621-2013-5524) and CADICS, a Linnaeus Center, both funded by the Swedish Research Council. The simulations were performed on resources provided by the Swedish National Infrastructure for Computing SNIC at Linköping University, Sweden.

References

G. M. Allenby, N. Arora, and J. L. Ginter. On the heterogeneity of demand. *Journal of Marketing Research*, 35(3):384–389, 1998.

A. Ansari, S. Essegaier, and R. Kohli. Internet recommendation systems. *Journal of Marketing research*, 37(3):363–375, 2000.

C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

B. H. Baltagi. *Econometric analysis of panel data.* John Wiley & Sons, 2008.

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

G. Belenky, N. J. Wesensten, D. R. Thorne, M. L. Thomas, H. C. Sing, D. P. Redmond, M. B. Russo, and T. J. Balkin. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. *Journal of sleep research*, 12(1):1–12, 2003.

D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

A. Bouchard-Côté, A. Doucet, and A. Roth. Particle Gibbs split-merge sampling for Bayesian inference in mixture models. *Pre-print*, 2015. arXiv:1508.02663v1.

M. Burda and M. Harding. Panel probit with flexible correlated effects: quantifying technology spillovers in the presence of latent heterogeneity. *Journal of Applied Econometrics*, 28(6):956–981, 2013.

F. Canova. Testing for convergence clubs in income per capita: a predictive density approach. *International Economic Review*, 45(1):49–77, 2004.

M. K. Condliff, D. D. Lewis, D. Madigan, and C. Posse. Bayesian mixed-effects models for recommender systems. In *Proceedings of ACM SIGIR'99 Workshop on Recommender Systems*, Berkeley, USA, August 1999.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

P. Fearnhead. Particle filters for mixture models with an unknown number of components. *Statistics and Computing*, 14(1):11–21, 2004.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2): 209–230, 1973.

T. S. Ferguson. Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2(4): 615–629, 1974.

S. Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer Verlag, 2006.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. Chapman & Hall/CRC, 3 edition, 2013.

B. Hall. *LaplacesDemon: software for Bayesian inference*, 2012. URL `http://cran.r-project.org/web/packages/LaplacesDemon/index.html`. R package version 12.05.07.

D. I. Hastie, S. Liverani, and S. Richardson. Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Statistics and Computing*, 25(5):1023–1037, 2015.

H. Ishwaran and M. Zarepour. Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, 12(3): 941–963, 2002.

S. Jain and R. M. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182, 2004.

H. F. Lopes, P. Müller, and G. L. Rosner. Bayesian meta-analysis for longitudinal data models using multivariate mixture priors. *Biometrics*, 59(1):66–75, 2003.

R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Verlag, 2 edition, 2004.

J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710, 2011.

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

Y. Ulker, B. Gunsel, and T. A. Cemgil. Sequential Monte Carlo samplers for Dirichlet process mixtures. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 876–883, Sardinia, Italy, May 2010.

F. Verbeke and E. Lesaffre. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):217–221, 1996.

G. Verbeke and G. Molenberghs. *Linear mixed models for longitudinal data*. Springer Verlag, 2009.

S. G. Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36(1):45–54, 2007.

A Implementation details

In all illustrations, we use $K = 10,000$ iterations in the Gibbs sampler and discard the first $K_b = 2,500$ iterations as burn-in.

A.1 Mixture of Gaussians

We use $R = 20$ components in the FM with $\eta_{0,r} = 1/R$ for $r = 1, \ldots, R$ as the concentration parameter to promote sparsity as suggested by Ishwaran and Zarepour [2002]. Furthermore, we use $\eta = 1$ for the finite mixture with the true number of components $R = 3$ so that all components are occupied. For the prior of η_0 in the DPM model, we use $c_o^\eta = 1$ and $C_o^\eta = 0.5$ as the hyperparameters. The remaining hyperparameters for the priors are selected as $b_0 = 0$, $B_0 = 0.2^2$, $c_0^Q = 1$ and $c_0^Q = 1$.

A.2 Mixed effects model with synthetic data

We use $R = 20$ components in the FM with $\eta_{0,r} = 1/R$ for $r = 1, \ldots, R$ to promote sparsity as in the previous example. For the prior of η_0 in the DPM model, we use the hyperparameters $c_0^\eta = 0.01$ and $C_0^\eta = 0.01$. We make use of the mean estimate from the linear regressions of each individual to select $a_0^\star$. Hence, the first element corresponds to the mean intercept and the remaining R elements correspond to the mean slope from the N linear regression estimates. We make use of $A_0^\star = \mathcal{I}_{d+Kp}$ as the prior covariance of $\alpha^\star$. Furthermore, we select $c_0^Q = 10$, $C_0^Q = 1$, $c_0^e = 0$, $C_0^e = 0$ and $\nu = 5$.

A.3 Mixed effects model with sleep deprivation data

We use $R = 20$ components in the FM with $\eta_{0,r} = 1/R$ for $r = 1, \ldots, R$ to promote sparsity as in the previous example. For the prior of η_0 in the DPM model, we use the hyperparameters $c_0^\eta = 0.01$ and $C_0^\eta = 0.01$. We make use of the same approach as in the previous illustration to choose $a_0^\star = \{251.4, 10.5, 0, \ldots, 0\}$ and select $A_0^\star = \mathsf{diag}\{0.01, 0.005, 0.02, \ldots, 0.02\}$. Furthermore, we select $c_0^Q = 10$, $C_0^Q = 0.05\mathbf{I}_2$, $c_0^e = 1$, $C_0^e = 2$ and $\nu = 5$.