Comput Stat (2013) 28:2007–2027
DOI 10.1007/s00180-012-0392-3

ORIGINAL PAPER

Semi-parametric Bayesian estimation of mixed-effects models using the multivariate skew-normal distribution

Reyhaneh Rikhtehgaran · Iraj Kazemi

Received: 7 October 2011 / Accepted: 18 December 2012 / Published online: 13 January 2013

Abstract In this paper, we develop a semi-parametric Bayesian estimation approach through the Dirichlet process (DP) mixture in fitting linear mixed models. The random-effects distribution is specified by introducing a multivariate skew-normal distribution as base for the Dirichlet process. The proposed approach efficiently deals with modeling issues in a wide range of non-normally distributed random effects. We adopt Gibbs sampling techniques to achieve the parameter estimates. A small simulation study is conducted to show that the proposed DP prior is better at the prediction of random effects. Two real data sets are analyzed and tested by several hypothetical models to illustrate the usefulness of the proposed approach.

Keywords Dirichlet process mixture · Gibbs sampling · Hierarchical models · Identifibiliaty problem · Random effects

1 Introduction

In recent years, semi-parametric Bayesian models have become popular in the analysis of correlated data. These models try to reflect our knowledge about the underlying distributions as well as their associated uncertainty. Among the semi-parametric Bayesian strategies, Dirichlet process (DP) is more familiar by introducing an unknown distribution G over the space of all possible distribution functions (Ferguson 1973). The DP has two parameters, a base distribution, G_0, stating our guess about the true non-parametric shape of G, and a precision parameter, M, reflecting our belief about

R. Rikhtehgaran · I. Kazemi (✉)
Department of Statistics, University of Isfahan, Isfahan, Iran
e-mail: i.kazemi@stat.ui.ac.ir

R. Rikhtehgaran
e-mail: r_rikhtehgaran@stat.ui.ac.ir

how similar G is to G_0. The DP has the restriction of being almost surely discrete which makes it inapplicable in situations where continuous distributions are needed. The Dirichlet process mixture (DPM) model is introduced to relax this restriction by adding a hierarchy level to the model (e.g., Escobar 1994; MacEachern 1994).

A number of authors in the context of linear mixed models have used DPM models with DP as a prior for the random effects (Kyung et al. 2010; Li et al. 2010). It is illustrated in the literature (e.g. Verbeke and Lesaffre 1997) that a misspecified random-effects distribution does not have a large impact on the estimation of fixed effects while the application of correctly specified distributions can lead to efficient estimates of these effects. This misspecification has also considerable effects on the applications where inferences about the subject effects are of interest (Verbeke and Lesaffre 1996).

Lots of efforts are made in the literature to tackle this misspecification by using some flexible parametric models. Widely applying of skewed distributions in recent modeling contexts is a good reason for this idea (e.g., Lachos et al. 2009; Ho and Lin 2010; Ferreira et al. 2011). Several methods are also proposed to make asymmetric a given family of symmetric distributions (e.g., Azzalini and Dalla Valle 1996; Sahu et al. 2003; Arellano-Valle et al. 2007). In most of these studies, some parameters are embedded into the underlying distribution in order to address the skewness.

In this paper, we propose the use of the multivariate skew-normal (SN) distribution in the structure of the DPM model in order to make more adaptable the analysis of longitudinal data. This includes the DPM model with the base of the multivariate normal distribution as its special case. Furthermore, by assuming a large value for M, the DP prior tends to its underlying base distribution and thus we deal with fitting the skew-normal mixed models. These concepts mean that our proposed prior is more flexible than the previous usual approaches. For this motivation, we use the multivariate skew-normal, introduced originally by Sahu et al. (2003) and extended by, for example, Arellano-Valle et al. (2007). We then use the multivariate SN as the base of the DP prior in the estimation process. Then we apply Markov chain Monte Carlo (McMC) methods to estimate the model parameters.

Another justification for our proposed prior relates to the stick-breaking representation of the DP in which we may approximate the related infinite mixture distribution with a truncated mixture. The use of this method helps doing Bayesian computations of the MCMC methods simple, specifically in the freely available software OpenBUGs (Lunn et al. 2009). In this case, for those applications in which the shape of the underlying random-effects distribution is distorted, the finite mixture based on normal distributions may tend to impose additional components to capture the skewness. Increasing the number of components will increase the cost of computation. While, by using a skew-normal distribution in the mixture structure, we may overcome this issue.

The remainder of this paper is organized as follows. In Sect. 2, we briefly review the semi-parametric Bayesian approach and introduce the DPM model. Section 3 considers the specification of normal mixed models. Section 4 briefly introduces a multivariate skew-normal distribution and derives the complete conditional posteriors required for the Gibbs sampling approach. In this section, we assume that the random effects follow the multivariate SN distribution and the error terms are also skew-normally distributed. Section 5 represents the Polya urn and the stick-breaking representations of the proposed DP. In Sect. 6, the identifibility problem of the proposed model is discussed. A small simulation study is conducted in Sect. 7. Section 8 includes applications of the proposed model on two real data sets. The last section includes concluding remarks.

2 Dirichlet process mixture models

This section introduces shortly the Dirichlet process and DPM models. The DP introduced by Ferguson (1973) and developed by others, such as Sethuraman (1994), generates a discrete random probability measure G as $G(\cdot) = \sum_{j=1}^{\infty} \pi_j \delta_{\boldsymbol{\xi}_j}(\cdot)$, where $\pi_j = \gamma_j \prod_{i=1}^{j-1}(1-\gamma_i)$ with $\gamma_j \overset{\text{iid}}{\sim} Beta(1, M)$, i.e. $f\left(\gamma_j|M\right) = M(1-\gamma_j)^{M-1}$ for $0 < \gamma_j < 1,\ M > 0$, and $\delta(\cdot)$ denotes a degenerated distribution of unit mass centered at $\boldsymbol{\xi}_j$ with $\boldsymbol{\xi}_j \overset{\text{iid}}{\sim} G_0$. This representation, so-called stick-breaking, is useful for carrying out the computational Bayesian inference using the MCMC approach. We denote the distribution of G by $DP(M, G_0)$. The variation of G around G_0 is controlled by the dispersion parameter $M(> 0)$ in a stochastic way such that, if M gets large then G becomes close to G_0. The role of this parameter is given below. Being applicable, the above infinite summation is mostly being truncated to a finite integer C. In order to obtain a proper distribution, the restriction $\pi_C = 1 - \sum_{j=1}^{C-1} \pi_j$ is generally considered.

An attractive feature of the DP for the general Bayesian statistical modeling is to specify it as a prior over distributions which provides wide support for the random-effects distribution. Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ be a statistically exchangeable sequence distributed according to a probability density function $f(.|\,\boldsymbol{\xi})$ where $\boldsymbol{\xi} \in \boldsymbol{\Xi}$ and let $\boldsymbol{\xi} \mid G \sim G$, where $G \sim DP(M, G_0)$ with G_0 depends possibly on some additional hyper-parameters. The density estimation problem, expressed by this hierarchical Bayesian model, is often known as DPM. This extension of the DP relaxes the discreteness restriction of using DP by adding a level to the hierarchy and has an interpretation as an infinite mixture model by assuming the stick-breaking representation of the DP. Escobar (1994) show that the DPM may be simplified by the use of the Polya urn representation through developing Bayesian computational techniques. More specifically, let $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N$ be N random samples from G, where $G \sim DP(M, G_0)$. Conditional on the other $\boldsymbol{\xi}$'s, $\boldsymbol{\xi}_i$ $(i = 1, \ldots, N)$ is shown to have the following specific mixture distribution

$$\boldsymbol{\xi}_i|\boldsymbol{y}, \boldsymbol{\xi}_{(-i)} \propto \sum_{j \neq i} q_j \delta_{\boldsymbol{\xi}_j} + M q_0 g_0(\boldsymbol{\xi}_i) f(\boldsymbol{y}_i|\boldsymbol{\xi}_i), \tag{1}$$

where $\boldsymbol{\xi}_{(-i)} = \left(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{i-1}, \boldsymbol{\xi}_{i+1}, \ldots, \boldsymbol{\xi}_N\right)^{'}$, $\delta_{\boldsymbol{\xi}_j}$ is a degenerated distribution with point mass at $\boldsymbol{\xi}_i = \boldsymbol{\xi}_j$, $q_j = f\left(y_i \mid \boldsymbol{\xi}_j\right),\ j = 1, \ldots, i-1, i+1, \ldots, N,\ g_0$ is the corresponding density function of G_0, and

$$q_0 = \int f\left(\boldsymbol{y}_i|\boldsymbol{\xi}_i\right) g_0(\boldsymbol{\xi}_i) d\boldsymbol{\xi}_i. \tag{2}$$

As was noted, the realizations of the DP are discrete, thus several unobservable $\boldsymbol{\xi}_i$'s prone to take similar values, such that the number of distinct values of $\boldsymbol{\xi}_i$, denoted by K, being less than or equal to N. The number of clusters, K, depends on the scale parameter M, which controls the amount of clustering between the center effects. It is shown by Antoniak (1974) that, given a fixed sample size N, $E\left[K \mid M, N\right] = \sum_{i=1}^{N} M/\left(M+i-1\right) \simeq Mlog\left(1+\frac{N}{M}\right)$, which implies K be approximated practically by M. There have been several works on specifying parameters M and K (e.g., Dorazio 2009).

In the application of the stick-breaking representation, it is required to determine the number of mass points, C, for approximating the DP. There are methods for the estimation of this parameter (see e.g., Ishwaran and James 2001, 2002). However, in this paper, depending on N, we consider a large value for C and then evaluate it with the estimation of the number of active clusters, i.e., the number of mixture components which are not empty.

3 Specification of Bayesian mixed models

Let y_{it} denotes the tth measurement taken on the ith individual, $t = 1, \ldots, T$, $i = 1, \ldots, N$. Consider the following linear mixed model

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_i, \tag{3}$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})'$, the $\boldsymbol{X}_i$ and the $\boldsymbol{Z}_i$ are $T \times p$ and $T \times q$ design matrices, $\boldsymbol{\beta}$ and $\boldsymbol{\xi}_i$ are unknown regression parameters corresponding to the fixed and random effects, respectively, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})'$ is a vector of error terms. The usual assumptions are $\boldsymbol{\varepsilon}_i \overset{\text{iid}}{\sim} N_T\left(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I}_T\right)$, where $\boldsymbol{I}_T$ is an identity matrix of order T, and the vector of random effects $\boldsymbol{\xi}_i \overset{\text{iid}}{\sim} N_q\left(\boldsymbol{0}, \boldsymbol{\varXi}\right)$, where $\boldsymbol{\varXi}$ is a $q \times q$ positive definite matrix.

It follows from Eq. (3) that, conditional on the random effects $\boldsymbol{\xi}_i$, the vector $\boldsymbol{y}_i$ is normally distributed with the mean vector $\boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\xi}_i$ and the covariance matrix $\sigma_\varepsilon^2 \boldsymbol{I}_T$. Let the joint prior density be $\pi\left(\boldsymbol{\beta}, \sigma_\varepsilon^2, \boldsymbol{\varXi}\right)$. The joint posterior density of all unobservables is then given by

$$\pi\left(\boldsymbol{\xi}, \boldsymbol{\beta}, \sigma_\varepsilon^2, \boldsymbol{\varXi} | \boldsymbol{y}\right) \propto \prod_{i=1}^{N} \varphi_T\left(\boldsymbol{y}_i | \boldsymbol{\mu}_i, \sigma_\varepsilon^2 \boldsymbol{I}_T\right) \varphi_q\left(\boldsymbol{\xi}_i | \boldsymbol{0}, \boldsymbol{\varXi}\right) \times \pi\left(\boldsymbol{\beta}, \sigma_\varepsilon^2, \boldsymbol{\varXi}\right), \tag{4}$$

where $\varphi_m\left(\boldsymbol{x}_i | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right)$ is the probability density function of $N_m\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right)$ evaluated at $\boldsymbol{x}_i$. The Gibbs sampling implementation has been appeared in many mixed-effects contexts and thus we do not present it in this paper.

4 The skew-normal mixed models

In traditional mixed models, the random effects $\boldsymbol{\xi}_i$'s are conventionally assumed to be randomly drawn from the multivariate normal distribution while the multivariate SN distribution may fit better the real data in many practical applications (e.g., Bandyopadhyay et al. 2010). Thus, we use this distribution for the random effects to improve fitting of Eq. (3).

4.1 The multivariate skew-normal distribution

Let $\boldsymbol{y}_i$ be a T-dimensional random vector. A T-variate SN distribution with the location vector $\boldsymbol{\mu}_i$, the scale matrix $\boldsymbol{\Sigma}$ ($T \times T$ positive definite matrix) and a $T \times T$ skewness matrix $\boldsymbol{\Delta}$, denoted by $SN_T\left(\boldsymbol{y}_i | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}; \boldsymbol{\Delta}\right)$, is given by

$$f\left(\boldsymbol{y}_i | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}; \boldsymbol{\Delta}\right) = 2^T \varphi_T\left(\boldsymbol{y}_i | \boldsymbol{\mu}_i, \boldsymbol{\Sigma} + \boldsymbol{\Delta}\boldsymbol{\Delta}'\right) \times \Phi_T\left(\boldsymbol{\Delta}'\left(\boldsymbol{\Sigma} + \boldsymbol{\Delta}\boldsymbol{\Delta}'\right)^{-1}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_i\right) | \mathbf{0}, \left(\boldsymbol{I} + \boldsymbol{\Delta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}\right)^{-1}\right), \tag{5}$$

where $\Phi_T\left(\boldsymbol{x}_i | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right)$ is the cumulative distribution function of $N_T\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right)$ evaluated at $\boldsymbol{x}_i$. Here, we assume that $\boldsymbol{\Delta} = diag\{\boldsymbol{\delta}\}$, where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_T)'$. For $\boldsymbol{\delta} = \mathbf{0}$, the original multivariate normal distribution is retrieved; for $\boldsymbol{\delta} > \mathbf{0}$, positively skewed and for $\boldsymbol{\delta} < \mathbf{0}$, negatively skewed distributions are obtained. The expectation and the covariance matrix are $\boldsymbol{E}\left(\boldsymbol{y}_i\right) = \boldsymbol{\mu}_i + \sqrt{2/\pi}\boldsymbol{\delta}$ and $\boldsymbol{Cov}\left(\boldsymbol{y}_i\right) = \boldsymbol{\Sigma} + (1 - 2/\pi)\boldsymbol{\Delta}^2$. Assume that the latent variable $\boldsymbol{w}_i$ follows the truncated-multivariate normal distribution. The SN distribution is then shown to be in a hierarchical form

$$\begin{aligned} \boldsymbol{y}_i | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}, \boldsymbol{\Delta}, \boldsymbol{w}_i &\overset{\text{ind}}{\sim} N_T\left(\boldsymbol{\mu}_i + \boldsymbol{\Delta}\boldsymbol{w}_i, \boldsymbol{\Sigma}\right), \\ \boldsymbol{w}_i &\overset{\text{ind}}{\sim} TN_T\left(\mathbf{0}, \boldsymbol{I}_T\right)\boldsymbol{I}\{\boldsymbol{w}_i > \mathbf{0}\}, \end{aligned} \tag{6}$$

where the notation $\boldsymbol{I}\{\boldsymbol{w}_i > \mathbf{0}\}$ means that each element of $\boldsymbol{w}_i$ is greater than zero.

4.2 Bayesian skew-normal mixed models

We now extend the normal mixed model (4) based on the multivariate SN distribution for both the error terms and the random effects to accommodate asymmetry in the model fitting process. More specifically, we consider the hierarchical model

$$\begin{aligned} \boldsymbol{y}_i | \boldsymbol{\xi}_i, \boldsymbol{\beta}, \sigma_\varepsilon^2, \boldsymbol{\Delta}_\varepsilon &\overset{\text{ind}}{\sim} SN_T\left(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\xi}_i, \sigma_\varepsilon^2\boldsymbol{I}_T; \boldsymbol{\Delta}_\varepsilon\right), \\ \boldsymbol{\xi}_i | \boldsymbol{\Xi}, \boldsymbol{\Delta}_\xi &\overset{\text{iid}}{\sim} SN_q\left(\mathbf{0}, \boldsymbol{\Xi}; \boldsymbol{\Delta}_\xi\right), \end{aligned} \tag{7}$$

for $i = 1, \ldots, N$. Depending on specific applications, various assumptions may be made for the skewness matrices $\boldsymbol{\Delta}_\varepsilon$ and $\boldsymbol{\Delta}_\xi$. In this paper, for the residual skewness matrix, we assume each depends only on one skewness parameter and let $\boldsymbol{\Delta}_\varepsilon = \delta_\varepsilon\boldsymbol{I}_T$. For the subjects skewness matrix we let $\boldsymbol{\Delta}_\xi = diag\left(\boldsymbol{\delta}_\xi\right)$, where $q \times 1$ vector $\boldsymbol{\delta}_\xi$

includes q skewness parameters. By the use of hierarchical form (6), we rewrite (7) as

$$\begin{aligned} \boldsymbol{y}_i|\boldsymbol{w}_{\varepsilon_i}, \boldsymbol{\xi}_i, \boldsymbol{\beta}, \sigma_\varepsilon^2, \delta_\varepsilon &\overset{\text{ind}}{\sim} N_T\left(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\xi}_i + \delta_\varepsilon \boldsymbol{w}_{\varepsilon_i}, \sigma_\varepsilon^2 \boldsymbol{I}_T\right), \\ \boldsymbol{w}_{\varepsilon_i} &\overset{\text{ind}}{\sim} TN_T\left(\boldsymbol{0}, \boldsymbol{I}_T\right) \boldsymbol{I}\{\boldsymbol{w}_{\varepsilon_i} > \boldsymbol{0}\}, \end{aligned} \tag{8}$$

$$\begin{aligned} \boldsymbol{\xi}_i|\boldsymbol{w}_{\boldsymbol{\xi}_i}, \boldsymbol{\Xi}, \delta_{\boldsymbol{\xi}} &\overset{\text{ind}}{\sim} N_q\left(\boldsymbol{\Delta}_{\boldsymbol{\xi}} \boldsymbol{w}_{\boldsymbol{\xi}_i}, \boldsymbol{\Xi}\right), \\ \boldsymbol{w}_{\boldsymbol{\xi}_i} &\overset{\text{ind}}{\sim} TN_q\left(\boldsymbol{0}, \boldsymbol{I}_q\right) I\{\boldsymbol{w}_{\boldsymbol{\xi}_i} > \boldsymbol{0}\}. \end{aligned} \tag{9}$$

Assuming that the unobserved parameters are independent, the following prior distributions are usually adopted; the inverse-Gamma with the hyper-parameters τ_1 and τ_2 with the mean $\tau_2/(\tau_1 - 1)$ for σ_ε^2, the inverse-Wishart with the hyper-parameters ν and $\boldsymbol{\Gamma}$ for $\boldsymbol{\Xi}$ and the multivariate normal with the mean vector $\boldsymbol{\beta}_0$ and the covariance matrix $\boldsymbol{\Lambda}$ for $\boldsymbol{\beta}$, the normal with zero mean and variance $\boldsymbol{\sigma}_{\delta_\varepsilon}^2$ for δ_ε, and the multivariate normal with the mean vector zero and the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\delta}_{\boldsymbol{\xi}}} = diag(\boldsymbol{\sigma}_{\boldsymbol{\delta}_{\boldsymbol{\xi}}}^2)$, where $q \times 1$ vector $\boldsymbol{\sigma}_{\boldsymbol{\delta}_{\boldsymbol{\xi}}}^2$ includes q variance components, for $\boldsymbol{\delta}_{\boldsymbol{\xi}}$.

With these specifications, the data analysis is conducted by implementing the Gibbs sampling technique which simulates iteratively from the complete conditional posterior distribution of each unknown stochastic parameter, given the current values of all other model parameters and the observations. After some algebra is done, it can readily be shown that these posteriors are as follows. For the regression coefficients, $\boldsymbol{\beta}$, we obtain

$$\boldsymbol{\beta}|\, \sigma_\varepsilon^2, \boldsymbol{\Xi}, \boldsymbol{\xi}, \boldsymbol{y} \sim N_p\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right), \tag{10}$$

with the mean $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\left(\boldsymbol{\Lambda}^{-1}\boldsymbol{\beta}_0 + \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^N \boldsymbol{X}_i'\boldsymbol{d}_i\right)$, where $\boldsymbol{d}_i = (\boldsymbol{y}_i - \boldsymbol{Z}_i\boldsymbol{\xi}_i - \delta_\varepsilon \boldsymbol{w}_{\varepsilon_i})$ and the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left(\boldsymbol{\Lambda}^{-1} + \frac{1}{\sigma_\varepsilon^2}\boldsymbol{T}_{XX}\right)^{-1}$, where $\boldsymbol{T}_{XX} = \sum_i \boldsymbol{X}_i^{'}\boldsymbol{X}_i$ represents the total-variation matrix. For the residual variance, σ_ε^2, we have

$$\sigma_\varepsilon^2|\, \boldsymbol{\beta}, \boldsymbol{w}_{\varepsilon_i}, \delta_\varepsilon, \boldsymbol{\xi}, \boldsymbol{y} \sim InvGamma\left(\tau_1^*, \tau_2^*\right), \tag{11}$$

where $\tau_1^* = \tau_1 + \frac{NT}{2}$ and $\tau_2^* = \tau_2 + \frac{1}{2}\sum_i(\boldsymbol{d}_i - \boldsymbol{X}_i\boldsymbol{\beta})^{'}(\boldsymbol{d}_i - \boldsymbol{X}_i\boldsymbol{\beta})$. For the random-effect variance, $\boldsymbol{\Xi}$, we derive

$$\boldsymbol{\Xi}|\boldsymbol{w}_{\xi_i}, \delta_{\xi}, \boldsymbol{\xi}, \boldsymbol{y} \sim InvWish\left(\nu^*, \boldsymbol{\Gamma}^*\right), \tag{12}$$

where $\nu^* = \nu + N$ and $\boldsymbol{\Gamma}^* = \boldsymbol{\Gamma} + \frac{1}{2}\sum_{i=1}^N(\boldsymbol{\xi}_i - \boldsymbol{\Delta}_{\boldsymbol{\xi}}\boldsymbol{w}_{\boldsymbol{\xi}_i})(\boldsymbol{\xi}_i - \boldsymbol{\Delta}_{\boldsymbol{\xi}}\boldsymbol{w}_{\boldsymbol{\xi}_i})'$. For the skewness parameter δ_ε we have

$$\delta_\varepsilon|\boldsymbol{\beta}, \sigma_\varepsilon^2, \boldsymbol{\xi}_i, \boldsymbol{w}_{\varepsilon_i}, \boldsymbol{y} \sim N\left(\mu_{\delta_\varepsilon}, \sigma_{\delta_\varepsilon}^{*2}\right), \tag{13}$$

where $\mu_{\delta_\varepsilon} = \sigma^{*2}_{\delta_\varepsilon} \sum_{i=1}^{N} \left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{\xi}_i\right)' \boldsymbol{w}_{\varepsilon i}/\sigma^2_\varepsilon$ and $\sigma^{*2}_{\delta_\varepsilon} = \left(\frac{1}{\sigma^2_{\delta_\varepsilon}} + \frac{1}{\sigma^2_\varepsilon}\boldsymbol{T}_{\boldsymbol{w}_\varepsilon \boldsymbol{w}_\varepsilon}\right)^{-1}$ with $\boldsymbol{T}_{\boldsymbol{w}_\varepsilon \boldsymbol{w}_\varepsilon} = \sum_i \boldsymbol{w}'_{\varepsilon_i}\boldsymbol{w}_{\varepsilon_i}$ representing the total variation of the latent variables $\boldsymbol{w}_{\varepsilon_i}$. For the skewness parameter $\boldsymbol{\delta}_{\boldsymbol{\xi}}$ we derive

$$\boldsymbol{\delta}_{\boldsymbol{\xi}} | \boldsymbol{\Sigma}_{\boldsymbol{\delta}_{\boldsymbol{\xi}}}, \boldsymbol{\Xi}, \boldsymbol{\xi}_i, \boldsymbol{w}_{\boldsymbol{\xi}_i}, \boldsymbol{y} \sim N_q\left(\boldsymbol{\mu}_{\boldsymbol{\delta}_{\boldsymbol{\xi}}}, \boldsymbol{\Sigma}^*_{\boldsymbol{\delta}_{\boldsymbol{\xi}}}\right), \tag{14}$$

where $\boldsymbol{\mu}_{\boldsymbol{\delta}_{\boldsymbol{\xi}}} = \boldsymbol{\Sigma}^*_{\boldsymbol{\delta}_{\boldsymbol{\xi}}} \sum_{i=1}^{N} \boldsymbol{\Xi}^{-1} diag(\boldsymbol{w}_{\boldsymbol{\xi}_i})\boldsymbol{\xi}_i$ and $\boldsymbol{\Sigma}^*_{\boldsymbol{\delta}_{\boldsymbol{\xi}}} = \left(\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\delta}_{\boldsymbol{\xi}}} + \boldsymbol{T}_{\boldsymbol{w}_\xi \boldsymbol{w}_\xi}\right)^{-1}$ with $\boldsymbol{T}_{\boldsymbol{w}_\xi \boldsymbol{w}_\xi} = \sum_{i=1}^{N} diag(\boldsymbol{w}_{\boldsymbol{\xi}_i})\boldsymbol{\Xi}^{-1}diag(\boldsymbol{w}_{\boldsymbol{\xi}_i})$. For each $\boldsymbol{w}_{\boldsymbol{\varepsilon}_i}$ we obtain

$$\boldsymbol{w}_{\varepsilon_i} | \boldsymbol{\beta}, \sigma^2_\varepsilon, \boldsymbol{\xi}_i, \delta_\varepsilon, \boldsymbol{y} \sim TN_T\left(\boldsymbol{\mu}_{\boldsymbol{w}_\varepsilon}, \boldsymbol{\Sigma}^2_{\boldsymbol{w}_\varepsilon}\right) \boldsymbol{I}\left(\boldsymbol{w}_{\boldsymbol{\varepsilon}_i} > \boldsymbol{0}\right), \tag{15}$$

where $\boldsymbol{\mu}_{\boldsymbol{w}_\varepsilon} = \boldsymbol{\Sigma}^2_{\boldsymbol{w}_\varepsilon}\delta_\varepsilon\sigma_\varepsilon^{2^{-1}}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{\xi}_i\right)$ and $\boldsymbol{\Sigma}^2_{\boldsymbol{w}_\varepsilon} = \left(1 + \delta^2_\varepsilon\sigma_\varepsilon^{2^{-1}}\right)^{-1}\boldsymbol{I}_T$. Similarly, for each $\boldsymbol{w}_{\boldsymbol{\xi}_i}$ we derive

$$\boldsymbol{w}_{\boldsymbol{\xi}_i} | \boldsymbol{\Xi}, \boldsymbol{\xi}_i, \boldsymbol{\delta}_{\boldsymbol{\xi}}, \boldsymbol{y} \sim TN_q\left(\boldsymbol{\mu}_{\boldsymbol{w}_\xi}, \boldsymbol{\Sigma}_{\boldsymbol{w}_\xi}\right) \boldsymbol{I}\left(\boldsymbol{w}_{\boldsymbol{\xi}_i} > \boldsymbol{0}\right), \tag{16}$$

where $\boldsymbol{\mu}_{\boldsymbol{w}_\xi} = \boldsymbol{\Sigma}_{\boldsymbol{w}_\xi}\boldsymbol{\Delta}_\xi\boldsymbol{\Xi}^{-1}\boldsymbol{\xi}_i$ and $\boldsymbol{\Sigma}_{\boldsymbol{w}_\xi} = \left(\boldsymbol{I}_q + \boldsymbol{\Delta}_\xi\boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_\xi\right)^{-1}$. For each random effect $\boldsymbol{\xi}_i$ we derive

$$\boldsymbol{\xi}_i | \boldsymbol{\beta}, \sigma^2_\varepsilon, \boldsymbol{w}_\xi, \boldsymbol{\delta}_{\boldsymbol{\xi}}, \boldsymbol{\Xi}, \boldsymbol{y} \sim N_q\left(\boldsymbol{\mu}^*_i, \boldsymbol{\Xi}^*\right), \tag{17}$$

where $\boldsymbol{\mu}^*_i = \boldsymbol{\Xi}^*\left(\boldsymbol{Z}'_i\widetilde{\boldsymbol{r}}_i/\sigma^2_\varepsilon + \boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_\xi\boldsymbol{w}_{\boldsymbol{\xi}_i}\right)$, $\boldsymbol{\Xi}^* = \left(\boldsymbol{\Xi}^{-1} + \boldsymbol{Z}'_i\boldsymbol{Z}_i\,/\sigma^2_\varepsilon\right)^{-1}$ and the $\widetilde{\boldsymbol{r}}_{\boldsymbol{i}} = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \delta_\varepsilon\boldsymbol{w}_{\varepsilon_i}$ denote the adjusted residuals.

Now, the Gibbs sampling implementation is straightforward using of all complete conditional posteriors (10)–(17), by noting that all are in closed forms belonging to standard families of distributions. Then, the sampler simulates iteratively from these posteriors by running a sufficient burn-in period until converges to the target stationary distributions. Gibbs samples can then be used in the computation of any feature of either marginal posterior distribution.

5 Application of the DP prior with SN base in Bayesian mixed models

While the DPM model with base normal has received specific attention in many practical applications, observations with having asymmetric structures of mixture components may not be modeled appropriately. Thus, we propose here a new DPM model with a multivariate skew-normal as the base distribution of the DP prior. Both Polya urn and stick-breaking representations of the proposed DP are illustrated.

5.1 The Polya urn representation of the proposed DP

By using the multivariate SN as the base distribution of the DP, the proposed mixed model, for $i = 1, \ldots, N$, can be written in the following hierarchical form

$$\begin{aligned} \boldsymbol{y}_i|\boldsymbol{\xi}_i, \boldsymbol{\beta}, \sigma_\varepsilon^2, \delta_\varepsilon &\overset{\text{ind}}{\sim} SN_T\left(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\xi}_i, \sigma_\varepsilon^2\boldsymbol{I}_T, \delta_\varepsilon\boldsymbol{I}_T\right), \\ \boldsymbol{\xi}_i|G &\overset{\text{iid}}{\sim} G, \\ G|M \sim DP\left(M, G_0\right), G_0 &\equiv SN_q\left(\boldsymbol{0}, \boldsymbol{\Xi}, \boldsymbol{\delta}_{\boldsymbol{\xi}}^{\prime}\boldsymbol{I}_q\right). \end{aligned} \tag{18}$$

Assuming that the priors are similar to those given in the previous section, we derive the complete conditional posterior distributions in order to implement MCMC methods.

Let $\boldsymbol{\xi}_{-i}$ be the vector of $\boldsymbol{\xi}$'s, after removing $\boldsymbol{\xi}_i$. In order to find each complete conditional posterior $\boldsymbol{\xi}_i|\boldsymbol{\beta}, \sigma_\varepsilon^2, \boldsymbol{\Xi}, \boldsymbol{\xi}_{-i}, \boldsymbol{y}$, we use relation (1). We first find an expression for the probability q_0, given in (2), by noting from Eq. (5) that

$$\begin{aligned} g_0\left(\boldsymbol{\xi}_i\right) f\left(\boldsymbol{y}_i \mid \boldsymbol{\xi}_i\right) = 2^{T+q}\varphi_q\left(\boldsymbol{\xi}_i|\boldsymbol{0}, \boldsymbol{\Xi} + \boldsymbol{\Delta}_\xi\boldsymbol{\Delta}_\xi\right) \times \varphi_T\left(\boldsymbol{y}_i|\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\xi}_i, \left(\sigma_\varepsilon^2 + \delta_\varepsilon^2\right)\boldsymbol{I}_T\right) \\ \times \Phi_q\left(\boldsymbol{\Delta}_\xi\left(\boldsymbol{\Xi} + \boldsymbol{\Delta}_\xi\boldsymbol{\Delta}_\xi\right)^{-1}\boldsymbol{\xi}_i|\boldsymbol{0}, \boldsymbol{I}_q - \boldsymbol{\Delta}_\xi\left(\boldsymbol{\Xi} + \boldsymbol{\Delta}_\xi\boldsymbol{\Delta}_\xi\right)^{-1}\boldsymbol{\Delta}_\xi\right) \\ \times \Phi_T\left(\frac{\delta_\varepsilon}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{\xi}_i\right)|\boldsymbol{0}, \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}\boldsymbol{I}_T\right). \end{aligned} \tag{19}$$

By introducing notations $A = \varphi_q\left(.\right) \times \varphi_T\left(.\right)$ and $B = \Phi_q\left(.\right) \times \Phi_T\left(.\right)$ in the above equations, we first easily show that

$$A = \varphi_q\left(\boldsymbol{\xi}_i \mid \boldsymbol{\mu}_i^{**}, \boldsymbol{\Xi}^{**}\right) \varphi_T\left(\boldsymbol{y}_i\,|\boldsymbol{X}_i\boldsymbol{\beta}, \left(\sigma_\varepsilon^2 + \delta_\varepsilon^2\right)\boldsymbol{\Omega}_i{}^{-1}\right), \tag{20}$$

where $\boldsymbol{\mu}_i^{**} = \boldsymbol{\Xi}^{**}\boldsymbol{Z}_i^{\prime}\boldsymbol{r}_i / \left(\sigma_\varepsilon^2 + \delta_\varepsilon^2\right)$, $\boldsymbol{\Xi}^{**} = \left(\left(\boldsymbol{\Xi} + \boldsymbol{\Delta}_\xi\boldsymbol{\Delta}_\xi\right)^{-1} + \boldsymbol{Z}_i^{\prime}\boldsymbol{Z}_i \,/\left(\sigma_\varepsilon^2 + \delta_\varepsilon^2\right)\right)^{-1}$ where $\boldsymbol{r}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}$ denote residuals and $\boldsymbol{\Omega}_i = \boldsymbol{I}_T - \left(\sigma_\varepsilon^2 + \delta_\varepsilon^2\right)^{-1}\boldsymbol{Z}_i\boldsymbol{\Xi}^{**}\boldsymbol{Z}_i^{\prime}$. Also, we obtain

$$B = \Pr\left(\boldsymbol{V}_i < \boldsymbol{C}\boldsymbol{\xi}_i\right) = \Phi_{T+q}\left(\boldsymbol{C}\boldsymbol{\xi}_i \mid \boldsymbol{\mu}_{v_i}, \boldsymbol{\Psi}\right), \tag{21}$$

where $\boldsymbol{V}_i \sim N_{T+q}\left(\boldsymbol{\mu}_{v_i}, \boldsymbol{\Psi}\right)$, $\boldsymbol{C} = \left(\left(\boldsymbol{\Xi} + \boldsymbol{\Delta}_\xi\boldsymbol{\Delta}_\xi\right)^{-1}\boldsymbol{\Delta}_\xi, -\frac{\delta_\varepsilon}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}\boldsymbol{Z}_i^{\prime}\right)^{\prime}$,

$$\boldsymbol{\mu}_{v_i} = \left(\boldsymbol{0}^{\prime}, \; -\frac{\delta_\varepsilon}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right)^{\prime}\right)^{\prime}, \tag{22}$$

and $\boldsymbol{\Psi}$ is a positive-definite blocked-diagonal matrix of the form

$$\boldsymbol{\Psi} = bldiag\left(\boldsymbol{I}_q - \boldsymbol{\Delta}_\xi\left(\boldsymbol{\Xi} + \boldsymbol{\Delta}_\xi\boldsymbol{\Delta}_\xi\right)^{-1}\boldsymbol{\Delta}_\xi, \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}\boldsymbol{I}_T\right). \tag{23}$$

Now, Eq. (19) is simplified by the multiplication of 2^{T+q} in A and B as

$$g_0\left(\boldsymbol{\xi}_i\right) f\left(\boldsymbol{y}_i \mid \boldsymbol{\xi}_i\right) = 2^{T+q}\varphi_T\left(\boldsymbol{y}_i \mid \boldsymbol{X}_i\boldsymbol{\beta}, \left(\sigma_\varepsilon^2+\delta_\varepsilon^2\right)\boldsymbol{\Omega}_i{}^{-1}\right) \tag{24}$$

$$\times\Phi_{T+q}\left(\boldsymbol{C}\boldsymbol{\xi}_i \mid \boldsymbol{\mu}_{v_i}, \boldsymbol{\Psi}\right)\varphi_q\left(\boldsymbol{\xi}_i \mid \boldsymbol{\mu}_i^{**}, \boldsymbol{\Xi}^{**}\right). \tag{25}$$

The probability q_0 is then achieved by integrating out the $\boldsymbol{\xi}_i$ from the above relation and by illustrating that the first term (24) is constant in terms of the $\boldsymbol{\xi}_i$ and the second term (25) is the expectation of $\Phi_{T+q}\left(\boldsymbol{C}\boldsymbol{\xi}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\Psi}\right)$ with respect to the normal density $\varphi_q\left(\boldsymbol{\xi}_i \mid \boldsymbol{\mu}_i^{**}, \boldsymbol{\Xi}^{**}\right)$. Therefore,

$$q_0 = 2^{T+q}\varphi_T\left(\boldsymbol{y}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \left(\sigma_\varepsilon^2+\delta_\varepsilon^2\right)\boldsymbol{\Omega}_i^{-1}\right) E\left(\Phi_{T+q}\left(\boldsymbol{C}\boldsymbol{\xi}_i|\boldsymbol{\mu}_{v_i}, \boldsymbol{\Psi}\right)\right). \tag{26}$$

Furthermore, according to Lemma 1 given in the “Appendix”, we have

$$q_0 = 2^{T+q}\varphi_T\left(\boldsymbol{y}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \left(\sigma_\varepsilon^2+\delta_\varepsilon^2\right)\boldsymbol{\Omega}_i^{-1}\right)\Phi_{T+q}\left(\boldsymbol{0}|\boldsymbol{\mu}_{v_i}-\boldsymbol{C}\boldsymbol{\mu}_i^{**}, \boldsymbol{\Psi}+\boldsymbol{C}\boldsymbol{\Xi}^{**}\boldsymbol{C}'\right). \tag{27}$$

Therefore, by using the mixture distribution (1), the simulation of $\boldsymbol{\xi}_i$'s is done based on the following scheme. With the probability proportional to $q_j = f\left(\boldsymbol{y}_i|\boldsymbol{\xi}_j\right)$, i.e. the first term in Eq. (18) by considering $\boldsymbol{\xi}_j$ instead of $\boldsymbol{\xi}_i$, we draw $\boldsymbol{\xi}_i$ according to $\boldsymbol{\delta}_{\boldsymbol{\xi}_j}$. Also, with the probability proportional to Mq_0, we draw $\boldsymbol{\xi}_i$ according to $\varphi_q\left(\boldsymbol{\xi}_i|\boldsymbol{\mu}_i^{**}, \boldsymbol{\Xi}^{**}\right)\Phi_{T+q}\left(\boldsymbol{C}\boldsymbol{\xi}_i|\boldsymbol{\mu}_{v_i}, \boldsymbol{\Psi}\right)$ achieved in Eq. (25). The complete conditional posteriors of $\boldsymbol{\beta}$, σ_ε^2 and δ_ε are to be found in a similar way of Sect. 4. For $\boldsymbol{\Xi}$ and $\boldsymbol{\delta}_{\boldsymbol{\xi}}$, we follow the approach introduced by West et al. (1994). As was mentioned, the realizations of the Dirichlet process are discrete, thus some of the $\boldsymbol{\xi}_i$'s may appear the same in the process. This particularly happens when the distribution of latent parameters is thought to be clustered. Thus, the similar $\boldsymbol{\xi}_i$'s discriminate in the same clusters. Consequently, there will be, say, K unique subject effects, denoted by $\boldsymbol{\eta}_l$, $l = 1, \ldots, K$. Conditional on the $\boldsymbol{\xi}_i$'s in finding the complete posterior of $\boldsymbol{\Xi}$, the $\boldsymbol{\xi}_i$'s act like known quantities. Therefore, $\boldsymbol{\xi}_l$'s, K and cluster memberships are also known. Consequently, we assume that the K independent variables $\boldsymbol{\eta}_l$'s, conditional on other parameters and the data, are distributed as $SN_q\left(\boldsymbol{0}, \boldsymbol{\Xi}; \boldsymbol{\Delta}_{\boldsymbol{\xi}}\right)$, where $\boldsymbol{\Delta}_{\boldsymbol{\xi}} = diag\left\{\boldsymbol{\delta}_{\boldsymbol{\xi}}\right\}$. Then, the complete conditional posterior of $\boldsymbol{\Xi}$ is as follows

$$\boldsymbol{\Xi}|\boldsymbol{\Delta}_{\boldsymbol{\xi}}, \boldsymbol{\xi}, \boldsymbol{y} \sim InvWish\left(\nu, \boldsymbol{\Gamma}\right) \times |\boldsymbol{\Upsilon}|^{-K/2} exp\left\{-\frac{1}{2}\boldsymbol{\Upsilon}^{-1} tr\left(\sum_{l=1}^{K}\boldsymbol{\eta}_l\boldsymbol{\eta}_l'\right)\right\}$$

$$\times\Phi_{Kq}\left(bldiag\left\{\boldsymbol{\Delta}_{\boldsymbol{\xi}}\boldsymbol{\Upsilon}^{-1}\right\}\boldsymbol{\eta}|\boldsymbol{0}, bldiag\left\{\boldsymbol{I}_q-\boldsymbol{\Delta}_{\boldsymbol{\xi}}\boldsymbol{\Upsilon}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\xi}}\right\}\right), \tag{28}$$

where $\boldsymbol{\Upsilon} = \boldsymbol{\Xi} + \boldsymbol{\Delta}_{\boldsymbol{\xi}}\boldsymbol{\Delta}_{\boldsymbol{\xi}}$, $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_K')'$ and $bldiag\{\boldsymbol{D}\}$ denotes a $Kq \times Kq$ blocked-diagonal matrix with elements $\boldsymbol{D}$. The complete conditional distribution of $\boldsymbol{\delta}_{\boldsymbol{\xi}}$ is obtained similarly by replacing $InvWish\left(\nu, \boldsymbol{\Gamma}\right)$ in Eq. (35) with $N_q\left(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\delta}_{\boldsymbol{\xi}}}\right)$.

The scale parameter of the DP, M, can be assumed fixed or be estimated by the MCMC outputs. To achieve the Bayes estimate of this parameter, a Gamma prior with hyper-parameters of κ_1 and κ_2 with the mean κ_1/κ_2 is assumed. Then, according to the method proposed by Escobar and West (1995) the complete conditional posterior of M is given by

$$\begin{aligned} M|u, K \ &\sim \pi_u Gamma\left(\kappa_1 + K, \kappa_2 - log(u)\right) \\ &\quad +(1-\pi_u) Gamma\left(\kappa_1 + K - 1, \kappa_2 - log(u)\right), \end{aligned} \tag{29}$$

where u is sampled from a Beta distribution, $u|M, K \sim Beta(M + 1, K)$, and π_u is defined by $\pi_u/(1 - \pi_u) = (\kappa_1 + K - 1)/\left(K(\kappa_2 - log(u))\right)$.

The Gibbs sampler now proceeds by simulating a sequence of samples from the complete conditional posteriors by illustrating that for the conditionals with admitting no closed forms, the Metropolis-Hastings algorithm is implemented within the Gibbs sampler.

Note that the DP with the multivariate normal as the base distribution is a special case of the DP with SN base and the complete conditional posteriors can be achieved in a similar way by putting skewness parameters equal to zero.

5.2 The stick-breaking representation of the proposed DP

Another representation of the DP with SN base is the stick-breaking. As the definition of the DP is not changed in the proposed distribution, all the properties of the DP are preserved. Therefore, the stick-breaking representation of the proposed DP can simply be achieved. Being applicable in the OpenBUGs software, we use an equivalent structure introduced by Ishwaran and James (2001) which considers the following DPM model. Let $\boldsymbol{y}_i$'s have the probability density functions $f\left(.|\boldsymbol{\xi}_{\lambda_i}\right)$ and $G\left(\lambda_i\right) = \sum_{j=1}^{C} \pi_j \delta_j\left(\lambda_i\right)$, where $\pi_j = \gamma_j \prod_{i=1}^{j-1}\left(1 - \gamma_i\right)$ with $\gamma_j \overset{\text{iid}}{\sim} Beta\left(1, M\right)$ and $\boldsymbol{\xi}_j \overset{\text{iid}}{\sim} G_0$ for $j = 1, \ldots, C$. According to this representation, the mixed-effects model is written as

$$\begin{aligned} &\boldsymbol{y}_i|\boldsymbol{\xi}_{\lambda_i}, \boldsymbol{w}_{\varepsilon_i}, \boldsymbol{\beta}, \sigma_\varepsilon^2, \delta_\varepsilon \overset{\text{ind}}{\sim} N\left(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\xi}_{\lambda_i} + \delta_\varepsilon \boldsymbol{w}_{\varepsilon_i}, \sigma_\varepsilon^2\right), \\ &\quad \boldsymbol{w}_{\varepsilon_i} \overset{\text{ind}}{\sim} N_T\left(\boldsymbol{0}, \boldsymbol{I}_T\right) \boldsymbol{I}\{\boldsymbol{w}_{\varepsilon_i} > \boldsymbol{0}\}, \\ &\quad \lambda_i \overset{\text{ind}}{\sim} G\left(\lambda_i\right) = \sum_{j=1}^{C} \pi_j \delta_j\left(\lambda_i\right), \\ &\quad \boldsymbol{\xi}_j|\boldsymbol{w}_{\xi_j} \overset{\text{ind}}{\sim} N_q\left(\boldsymbol{\Delta}_\xi \boldsymbol{w}_{\xi_j}, \boldsymbol{\Xi}\right), \\ &\quad \boldsymbol{w}_{\xi_j} \overset{\text{ind}}{\sim} TN_q\left(\boldsymbol{0}, \boldsymbol{I}_q\right) \boldsymbol{I}\{\boldsymbol{w}_{\xi_j} > \boldsymbol{0}\}, \end{aligned} \tag{30}$$

with $j = 1, \ldots, C$. Now, by considering the priors assumed in the previous sections, we readily find the complete conditional posteriors. These posteriors are simply obtained by using the above hierarchical form in a similar way of Sect. 4.2.

6 The identifibiliaty problem

As is mentioned in Sect. 4.1, the SN distribution has a non-zero mean. Therefore, the use of this as the distribution of random effects, or residuals, in Eq. (7) causes an identifibility issue and consequently makes difficulties in the interpretation of fixed parameters related to these random effects, or residuals. A solution suggested by Jara et al. (2008) is to subtract the expectations $\sqrt{2/\pi}\boldsymbol{\Delta}_\varepsilon$ and $\sqrt{2/\pi}\boldsymbol{\Delta}_{\boldsymbol{\xi}}$ from the location parameters of SN distributions in Eq. (7).

Also, a derived distribution, G, from the DP prior, has a non-zero mean. Therefore, this would cause another identifibility issue when we use the DP as the prior of the random-effects distribution in linear mixed models. This is discussed by Li et al. (2011). Based on their notations, let $\boldsymbol{\mu}_G = E(\boldsymbol{\xi}_i|G) = \int \boldsymbol{\xi}_i dG(\boldsymbol{\xi}_i)$ be the random mean of G and $\boldsymbol{\beta}^R$ be the subvector of $\boldsymbol{\beta}$ which is paired with those random effects that have the DP prior. They mention that the components of $\boldsymbol{\beta}_{pair} = \boldsymbol{\beta}^R + \boldsymbol{\mu}_G$ are not identifiable and then present a solution to solve the problem when the base distribution of the DP is normal. They assume the $\boldsymbol{\beta}^R$ equals zero which leads to $\boldsymbol{\beta}_{pair} = \boldsymbol{\mu}_G$ and then consider G_0 as $N(\boldsymbol{\beta}_\xi, \boldsymbol{\Xi})$. Following their approach, by the use of Theorems 3 and 4 of Ferguson (1973), the Bayes estimate of $\boldsymbol{\beta}_{pair}$ for the DP with base G_0 is given by

$$E\left(\boldsymbol{\mu}_G|\boldsymbol{y}\right) = E\left(\boldsymbol{\mu}_{G_\star}|\boldsymbol{y}\right) = \frac{M}{N+M}E\left(\boldsymbol{\mu}_{G_0}|\boldsymbol{y}\right) + \frac{1}{N+M}\sum_{i=1}^{N} E\left(\boldsymbol{\xi}_i|\boldsymbol{y}\right), \tag{31}$$

and its variance-covariance matrix is

$$Cov\left(\boldsymbol{\mu}_G|\boldsymbol{y}\right) = \frac{1}{N+M+1}E\left(\mathbf{Cov}_{G_\star} \mid \boldsymbol{y}\right) + Cov\left(\boldsymbol{\mu}_{G_\star}|\boldsymbol{y}\right), \tag{32}$$

where

$$\boldsymbol{\mu}_{G_\star} = \frac{M}{N+M}\boldsymbol{\mu}_{G_0} + \frac{1}{N+M}\sum_{i=1}^{N}\boldsymbol{\xi}_i,$$

and

$$\mathbf{Cov}_{G_\star} = \frac{1}{N+M}\left\{M(\boldsymbol{\mu}_{G_0}\boldsymbol{\mu}'_{G_0} + \boldsymbol{Cov}_{G_0}) + \sum_{i=1}^{N}\boldsymbol{\xi}_i\boldsymbol{\xi}'_i\right\} - \boldsymbol{\mu}_{G_\star}\boldsymbol{\mu}'_{G_\star},$$

are respectively the expectation and the variance of $G_\star = \frac{1}{N+M}\{M.G_0 + \sum_{i=1}^{N}\delta_{\boldsymbol{\xi}_i}\}$. The $G_\star$ is the base distribution of the predictive distribution of random-effects. By obtaining the posterior samples of model parameters and random effects based on MCMC outputs, Eqs. (31) and (32) are computed.

When the base distribution is normal, a non-zero mean is assumed for the base distribution, i.e., $N(\boldsymbol{\beta}_{\boldsymbol{\xi}}, \boldsymbol{\Xi})$. Thus, $\boldsymbol{\mu}_{G_0} = \boldsymbol{\beta}_{\boldsymbol{\xi}}$ and $\boldsymbol{Cov}_{G_0} = \boldsymbol{\Xi}$. However, when the base distribution has a non-zero mean, such as SN distribution, assuming a non-zero location parameter for the base distribution causes the identifibility problem. Thus, we consider $SN(\mathbf{0}, \boldsymbol{\Xi}; \boldsymbol{\delta}_{\boldsymbol{\xi}})$ for the base distribution and therefore, $\boldsymbol{\mu}_{G_0} = \sqrt{2/\pi}\boldsymbol{\delta}_{\boldsymbol{\xi}}$ and $\boldsymbol{Cov}_{G_0} = \boldsymbol{\Xi} + (1 - 2/\pi)\boldsymbol{\Delta}^2_{\boldsymbol{\xi}}$.

7 A small simulation study

Now, we conduct a simulation study to highlight the improvement created by the proposed DP prior for the random effects and to compare it with the normal, SN and, the DP prior with base SN. The data set is sampled by

$$y_{it} = \beta_0 + \beta_1 x_{it} + \xi_i + \varepsilon_{it}, \quad i = 1, \ldots, 300; \; t = 1, \ldots, 6, \tag{33}$$

where $x_{it} \overset{\text{iid}}{\sim} N(0, 9)$, $\varepsilon_{it} \overset{\text{iid}}{\sim} N(0, 16)$, and the random effects are generated by a mixture of skew-normal distributions. The components of this distribution are selected such that being highly overlapped and not easily distinguishable. Specifically, we generate 300 subjects according to $0.2SN(-4, 0.25; 2.5) + 0.2SN(-3, 0.25; 2.5) + 0.2SN(-1, 0.25; 2.5) + 0.2SN(2, 0.38; 4.5) + 0.2SN(6, 0.38; 3.5)$. Therefore, the theoretical mean and variance of random effects are respectively 2.474 and 18.311. Also, the empirical mean and variance of generated random effects are, 2.574 and 19.152, respectively and the empirical mean and variance of generated residuals are -0.018 and 16.525, respectively. Figure 2, panel left, shows the histogram of generated random effects. We also set $\beta_0 = 0$ and $\beta_1 = 1.5$. It should be mentioned that the theoretical intercept of the model is $\lambda_0 = \beta_0 + E(\xi_i) + E(\varepsilon_{it}) = 2.474$ and the empirical mean of $\beta_0 + \xi_i + \varepsilon_{it}$ across all (i,t) is 2.556(=0 + 2.574 − 0.018).

Then, we fit several models specified by Eq. (33) by assuming $\beta_0 = 0$ and with $\varepsilon_{it} \overset{\text{iid}}{\sim} N(0, \sigma^2_\varepsilon)$, while assuming $\xi_i \overset{\text{iid}}{\sim} N(\beta_\xi, \sigma^2_\xi)$ for model M1, $\xi_i \overset{\text{iid}}{\sim} SN(0, \sigma^2_\xi; \delta_\xi)$ for model M2, and in other models the random-effects distributions are DP priors with $N(\beta_\xi, \sigma^2_\xi)$ and $SN(0, \sigma^2_\xi; \sigma^2_\xi)$ as the base distributions.

We know that when the scale parameter M gets large then the DP becomes close to its base distribution. Here, to show the performance of the DP with base SN in comparison to the DP with base normal, we consider two cases: First, we set $M = 1$ which is a small value for the dispersion parameter of the DP and consider model M3 with base normal and M4 with base SN; Then, in the second case, we set $M = 100$ which is a large value and consider model M5 with base normal and M6 with base SN. As is already mentioned, there is a relationship between M and the number of clusters in the DP such that for a large value of M, we have a large value for K. Therefore, for the first case, the stick-breaking representations of DPs are approximated by $C = 20$ components and for the second case $C = 100$ components are considered to allow for a large K. To show that these values suffice to approximate infinite mixtures, the number of active clusters, K, are also estimated.

To implement the Gibbs sampler, the following independent priors are adopted: $N(0, 100)$ for the regression coefficient and for the skewness parameter δ_ξ, and the $InvGamma(0.01, 0.01)$ for σ^2_ε and σ^2_ξ. We use the OpenBUGs software by setting

Table 1 Bayesian estimation results of models M1–M6 for the simulated data set

Parm.	M1	M2	M3	M4	M5	M6
Model parameters						
λ_0	2.554	2.647	2.557	2.565	2.566	2.581
	(0.257)	(0.271)	(0.243)	(0.243)	(0.273)	(0.275)
β_1	1.486	1.486	1.484	1.484	1.484	1.484
	(0.035)	(0.035)	(0.034)	(0.035)	(0.035)	(0.035)
σ_ε^2	16.800	16.810	16.980	16.970	16.750	16.750
	(0.624)	(0.616)	(0.628)	(0.631)	(0.604)	(0.595)
Hyper-parameters of random-effects distributions						
δ_ξ	–	3.318	–	5.374	–	3.597
		(0.339)		(2.575)		(0.880)
σ_ξ^2	17.790	13.550	14.830	7.954	16.050	10.660
	(1.593)	(1.833)	(7.103)	(6.381)	(3.242)	(3.622)
β_ξ	–	–	3.225	–	2.596	–
			(1.807)		(0.651)	
K	–	–	8.618	8.287	72.510	55.590
			(2.090)	(1.916)	(3.882)	(3.014)

Bayesian standard deviations are given in parentheses

1,000 burn-in and then running 30,000 and 8,000 samples, respectively, for the first and the second cases. Results, after the convergence is achieved, are reported in Table 1.

The posterior probability density functions of parameter K in models M3–M6 are shown in Fig. 1. These graphs indicate that we achieved good approximations of DPs by the use of truncated stick-breaking representations. Based on the results of Table 1, a comparison of the various models shows that in all models biases and precisions in the estimation of β_1 and σ_ε^2 are closed. These showes that the distribution of random effects does not considerably influence the estimation of longitudinal effects. This fact was already illustrated by, e.g., Verbeke and Lesaffre (1997). The estimation of hyper-parameters σ_ξ^2 and δ_ξ are also given in Table 1. These values should not be used for the comparison of models M1–M6, since in models M2–M6 these values are not estimates of variance and skewness of random effects. For more details and solutions, see Jara et al. (2008) and Li et al. (2011).

The main advantage of using flexible distributions for the subject random-effects relates to their predicted values. These values are shown to be important in many applications where the prediction of response for each subject is of interest (e.g., Van Der Merwe and Pretorius 2003). We use the sum of squared of differences between the generated random effects and the predicted values to compare these predictions for all fitted models. Finding these measures as 718, 699, 684, 680, 659 and 640 for M1–M6, respectively, reveals that flexible models M2–M6 are better than M1. Also, models M4 and M6, respectively, in comparison to M3 and M5 are better fitted to the random effects. It is also seen that smaller number of components are needed to approximate the stick-breaking representations of DPs in models with SN bases. These are also

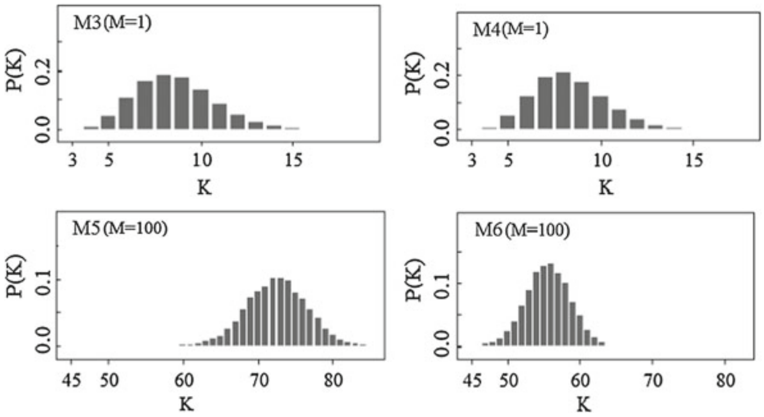

Fig. 1 Posterior probability density functions of active number of clusters, K, for models M3–M6

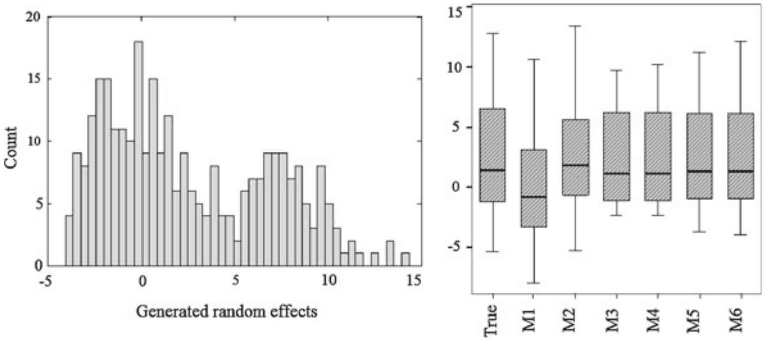

Fig. 2 Histogram of the generated random effects is depicted in the *left panel*. The right panel shows *Box-plots* of generated random effects (True) and the predicted ones for models M1–M6.

confirmed by the Box-plots shown in the right panel of Fig. 2. It is seen that some quantities, such as the quartiles and specially the median of the random effects, are biased in models M1 and M2, while they are correctly estimated in models M3–M6. Furthermore, the M4 and M6 have better covered the right tail of the generated random effects in comparison to M3 and M5, respectively.

It should be mentioned that because of the approximation of DPs with $C = 100$ components and due to the computationally intensive nature of MCMC methods, the inferences have been based on one generated data set. More comprehensive researches are required to evaluate the proposed model by considering larger values for N, M and C based on a large number of data sets.

8 Two illustrative empirical applications

Tax liability study: In this example, we fit a linear mixed model including the subject-specific effects. The data are taken from the Statistics of Income (SOI) panel of Individual Returns. These data are previously analyzed by Frischmann and Frees (1999);

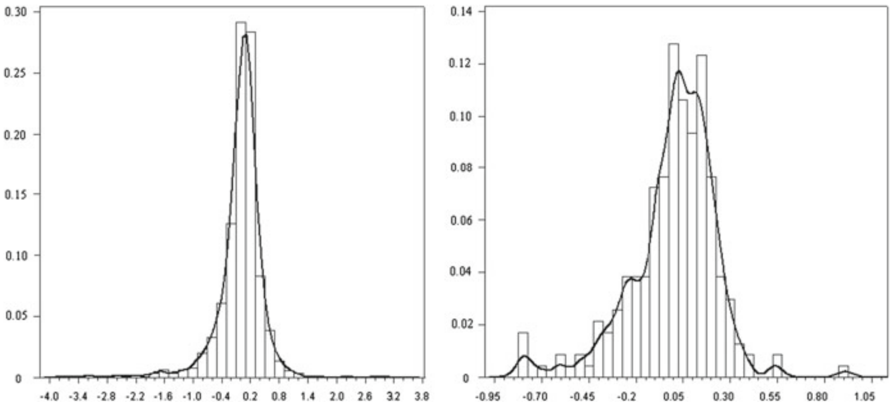

Fig. 3 Histograms of the fitted residuals and the predicted random effects based on M1 are shown in the *left* and the *right panels*, respectively. *Fitted curves* based on a kernel density estimate of the fitted residuals and the predicted random effects are also shown

Frees (2004), and Rabe-Hesketh and Skrondal (2005) based on a panel from 1982 to 1984 and 1986 to 1987 taxpayers included in the SOI panel. The aim of the analysis is to determine whether tax preparers significantly affect tax liability. The response of interest is the tax liability ($lntax_{it}$) as stated on the return in 1983 dollars, in logarithmic unit. The basic taxpayer characteristics that may affect tax liability include married ($married_i$), head of household ($head_{it}$), at least 65 years of age (age_{it}), self-employed ($employ_{it}$), paid preparer ($prep_{it}$) as binary variables, the number of dependents ($depend_{it}$), the marginal tax rate measure ($margtax_{it}$), and the logarithm of total positive income in 1983 dollars ($lntpinc_{it}$). We have removed the subjects with less than 3 follow-ups and those who have not reported their tax. We consider the following linear mixed-effects model

$$\begin{aligned} lntax_{it} &= \beta_0 + \beta_1 lntpinc_{it} + \beta_2 margtax_{it} + \beta_3 married_i + \beta_4 head_{it} + \beta_5 age_{it} \\ &\quad + \beta_6 employ_{it} + \beta_7 prep_{it} + \beta_8 depend_{it} + \xi_i + \varepsilon_{it}, \end{aligned} \tag{34}$$

for $i = 1, \ldots, 235$ and $t = 1, \ldots, T_i$, where T_i is the total number of follow-ups, the $\boldsymbol{\xi}_i$ are subject-specific effects and the ε_{it} are true residual terms, independent of $\boldsymbol{\xi}_i$ for all i, t. We fit the following candidate models. The residual terms ε_{it} and the random effects ξ_i are assumed to be normally distributed with zero means, variances σ_ε^2 and σ_ξ^2, respectively (M_1). We first fit this model in a frequentist perspective. The fitted residuals and the predicted random intercepts, shown in Fig. 3, illustrate that the underlying distributions are asymmetric. The p-values of Shapiro-Wilk tests are also obtained closed to 0, confirming that the normality assumptions are inappropriate. Thus, we assume that the residuals ε_{it} and the random effects ξ_i follow skew-normal distributions, centering at zero, having scale parameters σ_ε^2 and σ_ξ^2, and skewness parameters δ_ε and δ_ξ, respectively (M_2). In order to assess the ability of the proposed model with DP of base SN, we first fit a model similar to that specified in Sect. 5 but by assuming normality of residuals and the normal base distribution for the DP prior

Table 2 Bayesian estimation results of models M1–M4 for the tax liability data

Parm.	M1	M2	M3	M4
Model parameters				
β_0	0.049 (0.397)	2.229 (0.364)	−0.388 (0.034)	0.790 (0.045)
β_1	0.644 (0.047)	0.408 (0.041)	0.689 (0.047)	0.564 (0.043)
β_2	0.051 (0.004)	0.054 (0.003)	0.048 (0.003)	0.047 (0.003)
β_3	0.249 (0.073)	0.324 (0.070)	0.213 (0.066)	0.228 (0.058)
β_4	0.082 (0.098)	0.047 (0.095)	0.130 (0.084)	0.095 (0.069)
β_5	−0.167 (0.077)	−0.115 (0.079)	−0.121 (0.069)	−0.045 (0.068)
β_6	−0.213 (0.068)	−0.125 (0.068)	−0.128 (0.066)	−0.039 (0.062)
β_7	−0.031 (0.047)	0.032 (0.046)	0.009 (0.041)	0.026 (0.035)
β_8	−0.098 (0.023)	−0.090 (0.022)	−0.083 (0.021)	−0.074 (0.019)
σ_ε^2	0.306 (0.015)	0.095 (0.022)	0.300 (0.015)	0.035 (0.017)
δ_ε	–	−0.734 (0.047)	–	−0.875 (0.041)
Hyper-parameters of random-effects distributions				
σ_ξ^2	0.097 (0.017)	0.076 (0.028)	1.422 (2.399)	1.400 (2.662)
δ_ξ	–	0.186 (0.277)	–	2.216 (0.993)
M	–	–	1.971 (1.283)	2.130 (1.412)
K	–	–	8.831 (3.688)	9.150 (3.614)

Bayesian standard deviations are given in parentheses

(M_3). Then, we fit a model with the DP prior with base SN for the random effects with the specification of Sect. 5 (M_4).

To achieve the parameter estimates, independent priors are assumed for model parameters with the following distributions: the inverse-Gamma distribution with hyper-parameters both equal to 0.01 for σ_ε^2 and σ_ξ^2, and the normal distribution with zero mean and precision 0.01 for each regression coefficients and for the skewness parameters δ_ε and δ_ξ. For parameter M, we assume the Gamma distribution with 0.01 for both of its hyper-parameters, which implies a non-informative prior for the dispersion parameter of the DP. The truncated stick-breaking representation is assumed for the DP priors. We set C=30 mixture components for the approximation of the DP priors and then estimate the active number of clusters, K.

The OpenBUGs software is used to obtain the Bayes estimates of model parameters. With thin equals to 30, we set 1,000 burn-in to ensure that the convergence is achieved. Then, 20,000 samples are taken afterwards to get the MCMC outputs. Bayesian estimation results are given in Table 2. Skewness parameters in M2 and M4 are significant, meaning that the skewed distributions are suitable for both residuals and random effects. Because of appearing difficulties in working with marginal likelihoods, prediction-based criteria are considered for the model comparison issues. We use the predicted values, after the burn-in, to compute the relative absolute deviation (RAD). This criterion is simply achieved by computing the sum of ratios of differences between observed responses and their predicted values divided by the observed values. The values of RAD for models M1–M4 are 92.5, 51.3, 91.2 and 30.3, respectively.

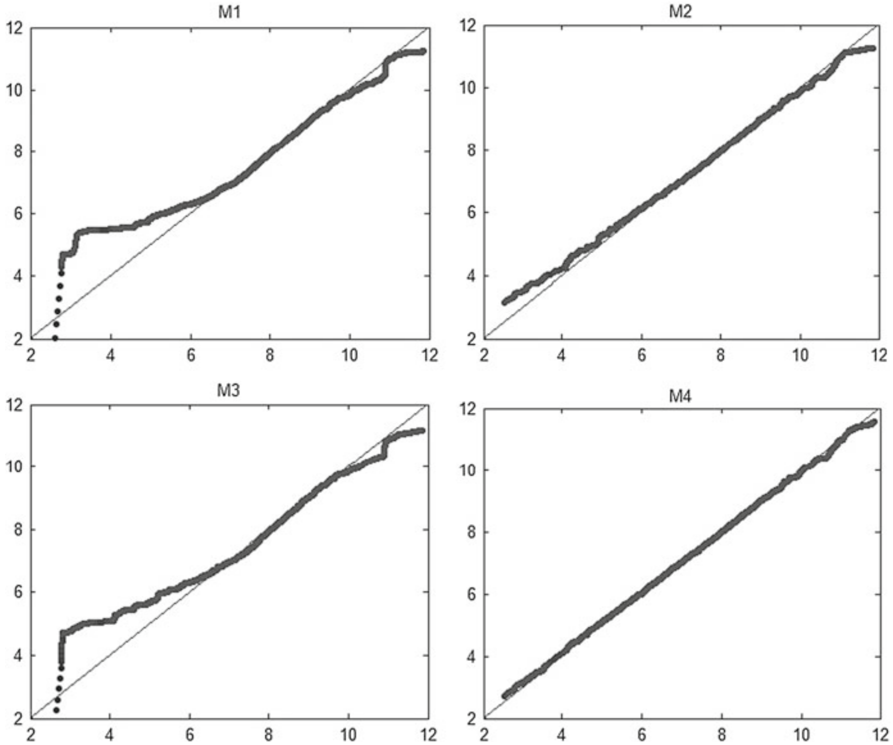

Fig. 4 Q–Q plots of predicted values (*vertical axis*) against observed values (*horizontal axis*) of models M1–M4 for the tax liability data

These values with the related Q–Q plots depicted in Fig. 4, show that the proposed model M4 which implements a more flexible structure for the random-effect distribution, is better fitted to the data.

Framingham cholesterol study: We analyze a subset of 133 randomly selected subjects from Framingham cholesterol data. The response variable includes subjects cholesterol levels measured at the beginning of the study and then every 2 years for 10 years. The aim of the study is to qualify changes in cholesterol levels based on time, baseline age, gender and intra-subject variation.

Zhang and Davidian (2001) analyzed the data by fitting a semi-parametric linear mixed model from a frequentist perspective. Ghidey et al. (2004) apply a smooth random-effect distribution and estimate the parameters by maximizing a penalized Gaussian mixed model. Arellano-Valle et al. (2005) provide additional inferences by implementing an EM type algorithm with the skew-normal distribution for the random effects and the residuals. Arellano-Valle et al. (2007) and Jara et al. (2008) refit the model with skewed distributions by proposing a Bayesian approach and the Gibbs sampler implementation.

We adopt the same linear mixed model used by these authors with baseline effects of age and sex, in the familiar form

$$y_{it} = \beta_0 + \beta_1 sex_i + \beta_2 age_i + \xi_{i1} + \xi_{i2} timec_{it} + \varepsilon_{it}, \tag{35}$$

Table 3 Bayesian estimation results of models M1–M4 for cholesterol data

Parm.	M1	M2	M3	M4
Model parameters				
β_0	1.502 (0.181)	1.656 (0.173)	1.689 (0.133)	1.447 (0.082)
β_1	−0.037 (0.065)	−0.003 (0.052)	−0.034 (0.053)	0.011 (0.048)
β_2	0.018 (0.004)	0.016 (0.004)	0.016 (0.004)	0.021 (0.004)
σ_ε^2	0.044 (0.003)	0.044 (0.003)	0.044 (0.003)	0.028 (0.006)
δ_ε	–	0.152 (0.115)	–	0.213 (0.057)
Hyper-parameters of random-effects distributions				
$\sigma_{\xi_1}^2$	0.132 (0.019)	0.081 (0.029)	0.161 (0.052)	0.147 (0.202)
$\sigma_{\xi_2}^2$	0.115 (0.022)	0.073 (0.031)	0.053 (0.025)	0.085 (0.064)
$\sigma_{\xi_1\xi_2}$	0.039 (0.020)	0.052 (0.017)	0.030 (0.025)	0.024 (0.091)
δ_{ξ_1}	–	0.375 (0.160)	–	1.954 (0.598)
δ_{ξ_2}	–	−0.420 (0.200)	–	0.341 (0.160)
M	–	–	51.530 (49.070)	3.664 (1.947)
K	–	–	34.150 (49.070)	13.400 (3.661)

Bayesian standard deviations are given in parentheses

for $i = 1, \ldots, 133$ and $t = 1, \ldots, 6$, where the y_{it} are cholesterol levels divided by 100 at time t for the ith subject, the $timec_{it}$ are $(time_{it} - 5)/10$, with time measured in years from baseline, sex is the gender indicator (0 = female, 1 = male), the ξ_{i1} and the ξ_{i2} are random intercepts and slopes, respectively, and the ε_{it} are residual terms.

As is commonly illustrated in the literature of the cholesterol data analysis, the random effects ξ_{i1} and ξ_{i2} may not be normally distributed. To show this formally, we fit the mixed model in a frequentist perspective with the usual assumptions and test the normality hypotheses. For the predicted random effects, the Mardia's multivariate skewness (Mardia 1970) criterion was estimated 18.53 with the corresponding p-value 0.001. These illustrate that a bivariate normal distribution for the random effects is inappropriate. Furthermore, the Shapiro-Wilk test returns the value 0.99 with the p-value closed to 0, showing the assumption of within-subject normality is violated.

Based on this information we fit four different linear mixed models: the Bivariate normal distribution for the effects $\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2})'$ and the normal distribution for ε_{it} (M1); the Bivariate skew-normal distribution for $\boldsymbol{\xi}_i$ and the skew-normal distribution for ε_{it} (M2); the DP prior with the base bivariate normal distribution for $\boldsymbol{\xi}_i$ and the normal distribution for ε_{it} (M3); and the DP prior with the base bivariate SN for $\boldsymbol{\xi}_i$ and the skew-normal distribution for ε_{it} (M4).

In fitting four Bayesian models with the Gibbs implementation, the priors are assumed the same as those of the previous example except that the prior of $\boldsymbol{\Xi}$ is assumed inverse-Wishart distribution with hyper-parameters ν equals to 2 and $\boldsymbol{\Gamma}$ equals to a diagonal matrix having elements 0.1. We also set $C = 50$.

The OpenBUGs software is used to do the Bayesian inference. We set 1,000 burn-in to ensure that the convergence is achieved and then 40,000 samples are taken afterwards to get the MCMC outputs. Estimation results for models M1–M4 are reported in

Table 3. Findings are different from each of four models. Also, the RAD values for models M1–M4 are reported as 82.1, 67.8, 82.2 and 65.4, respectively. Based on these values, it is obvious that among the four fitted models M1–M4, models M2 and M4 that use the skewed distributions for the random-effects are preferred. In another comparison between models M1 and M3 and also models M2 and M4, it is seen that, in order, models M3 and M4, which exploit the flexible structure of the Dirichlet process, are better fitted to the data. These results are also confirmed based on Q–Q plots of observed values against the predicted values depicted for models M1–M4 (not shown here). In general, we conclude that the DP prior with base SN (M4) can be better fitted to the data because it is more flexible than each of the skew-normal distribution and the DP prior with base normal.

9 Concluding remarks

In the present paper, we combined both the multivariate skew-normal distribution and the DP prior by considering the multivariate SN as the base distribution of the DP in order to achieve a more flexible distribution for the random effects. The proposed DP considers the multivariate normal and the skew-normal distributions as special cases. In comparison to the DP with the normal base distribution, the proposed DP is more adjustable and can be less sensitive to the precision parameter of the DP. This can be achieved by assuming a large value for M. As the role of the precision parameter M in the DP is to specify the amount of clustering between the center effects and to make the distribution flexible, the statistical inference can be sensitive to the choice of M, or alternatively, to the choice of prior assumed for M. There have been several published papers to achieve an estimate for this parameter and also to introduce proper priors (e.g., Escobar and West 1995; Dorazio 2009). In the proposed DP prior, the flexibility of the distribution can be achieved by utilizing the base distribution. Therefore, by assuming a large value for M, the DP prior will be closed to the skew-normal distribution, while our proposed prior as being a DP prior may still be more flexible than the skew-normal distribution.

The proposed DP is applied to a small simulation study and also to two data sets. It is shown that the proposed DP prior is better fitted to the data sets in comparison to the skew-normal distribution and the DP prior with base normal. One may offer the possibility of using more forms of SN distributions in the DP priors by applying various techniques of creating skewed distributions. There are several methods for making skewed distributions based on symmetric distributions. The differences between these methods are the main distinctions between induced DPs. We distinguish here two possible techniques. The first has no effect on the definition of the DP and tries to put the skewness on the DP by utilizing a skewed distribution as the base distribution. The second technique uses a symmetric distribution as the base distribution and tries to impose skewness on the DP by applying some modified methods of creating skewed distributions directly on the DP structure. Operationalizing the first technique is straightforward while the second needs doing more attempt. The only papers we found, related to the second approach, were Iglesias et al. (2009) and Quintana (2010). They propose a class of skewed DPs by assuming the symmetric distribution for the base distribution of the DP and introduce the skewed DP by modifying the definition of the DP familiarized by Ferguson (1973). One interesting subject is to carry on working on these papers. The semi-parametric Bayesian approach we introduced for fitting of mixed models follows the first technique. Other versions of the skew-normal distributions can also be applied in this framework. For the future work we suggest applying other flexible parametric distributions as the base distribution of the DP.

Acknowledgments Authors are grateful to the office of Graduate Studies of the University of Isfahan for their support. We also gratefully acknowledge referees for their invaluable comments.

Appendix

Lemma 1 *(Arellano-Valle and Genton 2005) Let* $\boldsymbol{y} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Then for any fixed* k*-dimensional vector* $\boldsymbol{a}$ *and* $k \times m$ *matrix* $\boldsymbol{D}$,

$$E\left(\Phi_k\left(\boldsymbol{a} + \boldsymbol{D}\boldsymbol{y} | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)\right) = \Phi_k\left(\boldsymbol{a} \mid \boldsymbol{\mu}_0 - \boldsymbol{D}\boldsymbol{\mu}, \boldsymbol{\Sigma}_0 + \boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}'\right).$$

References

Antoniak CE (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann Stat 2:1152–1174

Arellano-Valle RB, Genton MG (2005) Fundamental skew distributions. J Multivar Anal 96:93–116

Arellano-Valle RB, Bolfarine H, Lachos VH (2005) Skew-normal linear mixed models. J Data Sci 3: 415–438

Arellano-Valle RB, Bolfarine H, Lachos VH (2007) Bayesian inference for skew-normal linear mixed models. J Appl Stat 34(6):663–682

Azzalini A, Dalla Valle A (1996) The multivariate skew-normal distribution. Biometrika 83:715–726

Bandyopadhyay D, Lachos VH, Abanto-Valle CA, Ghosh P (2010) Linear mixed models for skew-normal/independent bivariate responses with an application to periodontal disease. Stat Med 29(25):2643–2655

Dorazio RM (2009) On selecting a prior for the precision parameter of Dirichlet process mixture models. J Stat Plan Inf 139:3384–3390

Escobar MD (1994) Estimating normal means with a Dirichlet process prior. J Am Stat Assoc 89(425): 268–277

Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. J Am Stat Assoc 90:577–588

Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. Ann Stat 1:209–230

Ferreira CDS, Bolfarine H, Lachos VH (2011) Skew scale mixtures of normal distributions: properties and estimation. Stat Methodol 8:154–171

Frees EW (2004) Longitudinal and panel data: analysis and applications for the social sciences. Cambridge University Press, Cambridge pp: 65–67 and 81–86

Frischmann PJ, Frees EW (1999) Demand for services: determinants of tax preparation fees. J Am Tax Assoc 21(Supplement):1–23

Ghidey W, Lesaffre E, Eilers P (2004) Smooth random effects distribution in a linear mixed model. Biometrics 60:945–953

Ho HJ, Lin TI (2010) Robust linear mixed models using the skew t distribution with application to schizophrenia data. Biometr J 52(4):449–469

Iglesias PL, Orellana Y, Quintana FA (2009) Nonparametric Bayesian modeling using skewed Dirichlet processes. J Stat Plan Inf 139:1203–1214

Ishwaran H, James LF (2001) Gibbs sampling methods for stick-breaking priors. J Am Stat Assoc 96(453):161–173

Ishwaran H, James LF (2002) Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. J Comput Graph Stat 11:508–532

Jara A, Quintana F, SanMartin E (2008) Linear mixed models with skew-elliptical distributions: a Bayesian approach. Comput Stat Data Anal 52:5033–5045

Kyung M, Gill J, Casella G (2010) Estimation in Dirichlet random effects models. Ann Stat 38:919–1009

Lachos VH, Dey DK, Cancho VG (2009) Robust linear mixed models with skew-normal independent distributions from a Bayesian perspective. J Stat Plan Inf 139:4098–4110

Li Y, Lin X, Müller P (2010) Bayesian inference in Semiparametric mixed models for longitudinal data. Biometrics 66(1):70–78

Li Y, Müller P, Lin X (2011) Center-adjusted inference for a nonparametric Bayesian random effect distribution. Stat Sin 21(3):1201–1223

Lunn D, Spiegelhalter D, Thomas A, Best N (2009) The BUGS project: evolution, critique and future directions (with discussion). Stat Med 28:3049–3082

MacEachern SN (1994) Estimating normal means with a conjugate style Dirichlet process prior. Commun Stat B 23:727–741

Mardia KV (1970) Measures of multivariate skewness and kurtosis with applications. Biometrika 57: 519–530

Van Der Merwe AJ, Pretorius AL (2003) Bayesian estimation in animal breeding using Dirichlet process prior for correlated random effects. Genet Sel Evol 35:137–158

Quintana FA (2010) Linear regression with a dependent skewed Dirichlet process. Chil J Stat 1(2):35–49

Rabe-Hesketh S, Skrondal A (2005) Multilevel and Longitudinal Modeling Using Stata. Stata Press,

Sahu SK, Dey DK, Branco MD (2003) A new class of multivariate skew distributions with applications to Bayesian regression models. Can J Stat 31:129–150

Sethuraman J (1994) A constructive definition of Dirichlet priors. Stat Sin 4:639–650

Verbeke G, Lesaffre E (1996) A linear mixed-effects model with heterogeneity in the random-effects population. J Am Stat Assoc 91:217–221

Verbeke G, Lesaffre E (1997) The effect of misspecifying the random effects distribution in linear mixed models for longitudinal data. Comput Stat Data Anal 23:541–556

West M, Müller P, Escobar MD (1994) Hierarchical priors and mixture models, with applications in regression and density estimation. In: Freeman PR, Smith AFM (eds) Aspects of uncertainty: a tribute to D.V. Lindley. Wiley, New York, pp 363–386

Zhang D, Davidian M (2001) Liner mixed models with flexible distributions of random effects for longitudinal data. Biometrics 57:795–802