Noname manuscript No.
(will be inserted by the editor)

Semiparametric Bayesian analysis for longitudinal mixed effects models with non-normal AR(1) errors

Junshan Shen · Hanjun Yu · Jin Yang · Chunling Liu

Received: date / Accepted: date

Abstract This paper studies Bayesian inference on longitudinal mixed effects models with non-normal AR(1) errors. We model the nonparametric zero-mean noise in the autoregression residual with a Dirichlet process (DP) mixture model. Applying the empirical likelihood tool, an adjusted sampler based on the Pólya urn representation of DP is proposed to incorporate information of the moment constraints of the mixing distribution. A Gibbs sampling algorithm based on the adjusted sampler is proposed to approximate the posterior distributions under DP priors. The proposed method can easily be extended to address other moment constraints owing to the wide application background of the empirical likelihood. Simulation studies are used to evaluate the performance of the proposed method. Our method is illustrated via the analysis of a longitudinal dataset from a psychiatric study.

Keywords Autocorrelation · Dirichlet process mixture models · Empirical likelihood · Pólya urn representation · Random effects.

Junshan Shen
School of Statistics, Capital University of Economics and Business, Fengtai, Beijing, P. R. China

Hanjun Yu
School of Statistics, Capital University of Economics and Business, Fengtai, Beijing, P. R. China

Jin Yang
Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

Chunling Liu
Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
Tel.: +852 2766 6931
Fax: +852 2362 9045
E-mail: macliu@polyu.edu.hk

1 Introduction

The seminal linear mixed effects model by Laird and Ware (1982) has been popularly accepted and adopted to formulate longitudinal data since it combines the merits of incorporating both individual random effect components and within-individual errors and of accommodating the unbalanced observations and more parsimonious covariance structures. The autoregressive time series process of order p, denoted by AR(p), is widely used to characterize the residual serial dependence correlation structure of repeated measurements considering the similarities between the time series and the hierarchical longitudinal data; refer to Chi and Reinsel (1989), Bartolucci and Bacci (2014), among others for details. There is extensive work under conditions where the residual error sequence follows an uncorrelated normal distribution. However, this assumption is critical and might be violated in practice; refer to Arnau et al (2012) for examples in a wide range of application fields. In this paper, we consider the case where the individual errors are non-normal with zero-mean in the sense that the zero-mean noise term in the AR(1) residual is not normally distributed. We develop a computationally feasible method to address the Bayesian semiparametric model inference in the linear mixed effects model, where the noise follows a Dirichlet process mixture (DPM) model. The strategy is to employ the empirical likelihood (EL) tool to incorporate the moment constraint information into an equivalent representation of the Dirichlet process (DP) to obtain zero-mean posterior distributions and hence to develop an adjusted Gibbs sampler for posterior inference.

Let us explore the literature under the model presented by Laird and Ware (1982) when the within-subject residual autocorrelation is assumed under serial autoregression AR(p). Chi and Reinsel (1989) presented a score test to check the autocorrelation in the AR(1) errors for the random effects model. Goldstein et al (1994) proposed a multilevel model for repeated measured data augmented by AR(1) and AR(2) models together with a seasonal component for the residuals. Fan et al (2014) adopted mixture priors between a point mass and a normal distribution on both effects together with the aid of the MCMC method to select variables and to choose the model. However, within the applied contexts, the residual errors tend to depart from the normal distribution, which is particularly common in psychometric, educational, psychological, and other social and health sciences backgrounds. Goldstein et al (1994) noted that in some cases, especially when measurements are close together in time, the normality assumption is doubtful, and some additional correlation structures should be modeled. Bartolucci and Bacci (2014) studied a longitudinal dataset from the Health and Retirement study by a linear mixed effects model with AR(1) errors following a logistic distribution. Wang and Fan (2011) presented a score test for autocorrelation and a hybrid ECME-scoring procedure to calculate the MLE, where the individual error has an AR(p) dependence structure and follows a t distribution. Damsleth and El-Shaarawi (1989) applied the AR(1) model with double-exponential noise to a series of weekly measurements of sulphate concentrations. Tiku et al (2000) listed dozens of papers that assumed non-normal assumptions for the noise, and they considered AR(q) models in time series with non-normal noise represented by a member of a wide family of symmetric distributions. Furthermore, some work considered nonparametric random effects (Kleinman and Ibrahim, 1998; Li et al, 2011); there is also increasing attention on Bayesian quantile regression for longitudinal and even clustered data (Luo et al, 2012; Reich et al, 2010).

The questions that we seek to address are difficult to consider by frequentist approaches to statistical analysis. In practice, Bayesian methods appear to provide a relatively computationally feasible and effective technique. Bayesian nonparametric methods have become increasingly popular for addressing non-normal issues. Among them, the Dirichlet process proposed by Ferguson (1973) is commonly used as a nonparametric prior for an unknown discrete distribution. Note that samples for Dirichlet processes can easily be drawn from its Pólya urn representation (Blackwell and MacQueen, 1973) or its stick-breaking construction (Sethuraman, 1994). The Dirichlet process mixture models proposed by Antoniak (1974) can be used to model priors for continuous distributions. Further developments of DPM models can be found in Escobar (1994), Escobar and West (1995), and MacEachern and Müller (1998), among others. Neal (2000) presented several Markov chain methods for sampling from the posterior distributions of DPM models.

Encouraged by the effectiveness of DPM models for the aforementioned longitudinal mixed effects model with non-normal AR(1) errors, we provide a semiparametric Bayesian method as a solution. First, we add a DPM model to the non-normal noise term in the autoregression residual. Note that the noise term is subject to the zero-mean constraint. Next, using the property that the DP prior is a conjugate for mixing distributions, we obtain a DP posterior in an equivalent Pólya urn representation. After the above preparation, the key step is to incorporate the moment constraint information via the empirical likelihood so that the posterior of the mixing distribution is zero-mean. As a result, we have an adjusted Gibbs sampler for Bayesian inference. Consequently, the proposed posterior sampling algorithm is an approximate form of the standard Markov chain Monte Carlo (MCMC) algorithm with DP prior.

Returning to the literature, there are several ways to make full use of this type of moment constraints. Brunner and Lo (1989) created DP mixtures of uniform distributions to obtain symmetric densities with the mean and mode at zero. Hoff (2003) proposed a general method to define probability measures in a convex set and obtained measures with a mean constraint. Hoff (2000) noted that his approach could be used to construct probability measures with a mean of zero and a variance of one. Yang et al (2010) proposed a concise centered stick-breaking process (CSBP) to induce mean and variance constraints on an unknown distribution that can apply centered stick-breaking mixtures (CSBM) in parallel for the DPM model under the semiparametric latent factor regression. Griffin (2016) applied the CSBM method in a semiparametric linear mixed model with mean constraints on the adaptive truncated stick-breaking mixtures of both the random effects and regression errors. However, the CSBP and CSBM methods are limited to mean and variance constraints and could not be applied to solve complicated constraints related to the distribution function, for example, a quantile constraint such as a median of zero.

In addition to the aforementioned work, some studies adjusted the prior rather than the posterior to address auxiliary information. Kitamura and Otsu (2011) constructed an exponential tilting projection of a DP process with the Kullback-Leibler divergence for independent and identically distributed data; Shin (2014) completed their method for general moment constrained models for dependent data and drew posterior samples by a sequential Monte Carlo algorithm; Choi (2016) used the same principle to obtain exponentially tilted DP priors and presented a Metropolis-Hastings algorithm for posterior sampling.

The novelty of our proposed method is that we apply EL to the moment constraints for the posterior sampler, which has not previously been investigated in the literature. Lazar (2003) were the first to apply EL to Bayesian analysis, but they treated EL mainly as an alternative of the parametric likelihood. Nevertheless, we apply EL as a tool to cancel the moment constraints. The contribution of our semiparametric Bayesian method is twofold. On one hand, under the longitudinal data setting, the nonparametric instead of normality assumption on the noise term provides robust modeling, which leads to broader applications. On the other hand, we propose a new method to address auxiliary information for DP prior inference. We achieve our goal by adjusting the Gibbs sampler. Gibbs sampling is feasible for any models that considers distributions subject to moment constraints. Meanwhile, the application scope of our methodology is not restricted to the longitudinal data setting, in that the development of the methodology does not rely on the data type.

The remainder of this paper is organized as follows. Section 2 establishes the procedures for how we develop an EL-based modification on posterior distributions to obtain the adjusted Gibbs sampler. Section 3 presents the Gibbs sampling algorithm for the posterior computation. Section 4 details the simulation results, and Section 5 illustrates an application to a real dataset. Section 6 ends the article with a discussion.

2 Empirical-likelihood-based Dirichlet process mixture

Let $\mathbf{y}_i$ be an $n_i \times 1$ vector of responses whose element y_{ij} is the jth response of the ith subject for $i = 1,\dots,m$. The longitudinal mixed effects model with AR(1) errors in our study can be characterized by

$$\begin{aligned}&\mathbf{y}_i = \mathbf{x}_i\boldsymbol{\beta} + \mathbf{z}_i\mathbf{b}_i + \mathbf{w}_i,\; i = 1,\dots,m;\\&\mathbf{w}_i = (w_{i1},\dots,w_{in_i})^T;\; w_{ij} = \rho w_{i,j-1} + \varepsilon_{ij},\; j = 2,\dots,n_i,\end{aligned} \tag{1}$$

where $\boldsymbol{\beta}$ is a $p\times 1$ vector of fixed effect parameters, $\mathbf{b}_i$ a $q\times 1$ Gaussian random vector representing the subject-specific random effects, $\mathbf{x}_i = (\mathbf{x}_{i1},\dots,\mathbf{x}_{in_i})^T$ and $\mathbf{z}_i = (\mathbf{z}_{i1},\dots,\mathbf{z}_{in_i})^T$ are $n_i \times p$ and $n_i \times q$ design matrices linking $\boldsymbol{\beta}$ and $\mathbf{b}_i$ to $\mathbf{y}_i$, respectively, and $\mathbf{w}_i = (w_{i1},\dots,w_{in_i})^T$ is an $n_i \times 1$ vector of model errors. In the AR(1) serial correlation, ρ is the autoregressive coefficient and the $\varepsilon_{ij}' s$ are i.i.d. noises. Denote $N = n_1 + \dots + n_m$. When the noise sequence $\{\varepsilon_{ij}\}$ is non-normal, we assume a mixture model on the noise

$$f_G(\varepsilon|\sigma^2) = \int \varphi(\varepsilon|u,\sigma^2)dG(u), \tag{2}$$

where $\varphi(\cdot|u,\sigma^2)$ is the normal probability density function with mean u and variance σ^2, and G is an unknown probability distribution of u satisfying $\int u dG(u) = 0$, which ensures that ε comes from a zero-mean mixture distribution.

If there were no constraint $\int u dG(u) = 0$, the Dirichlet process (DP) proposed by Ferguson (1973) is often used as a prior of the mixing distribution G. We hence *first* recall the concept of a Dirichlet process. A DP is parameterized in terms of a positive precision parameter α and a base distribution G_0 defined on a probability space Θ. Suppose that $A_1,\dots,A_k$ is a finite measurable partition of Θ. A Dirichlet process, denoted as DP(α, G_0), is a random probability measure G such that the joint distribution of the vector $(G(A_1),\dots,G(A_k))$ is a Dirichlet distribution with parameters $(\alpha G_0(A_1),\dots,\alpha G_0(A_k))$.

Note that a DP can be equivalently characterized by a Pólya urn representation, e.g., refer to Blackwell and MacQueen (1973). That is, for $G\sim DP(\alpha,G_0)$, a random sample $u_1,\dots,u_n$ drawn from G can be represented as

$$u_n|u_1,\dots,u_{n-1} \sim \sum_{l=1}^{n-1}\frac{1}{n-1+\alpha}\delta_{u_l} + \frac{\alpha}{n-1+\alpha}G_0,$$

where δ_u is the probability mass at point u. Denote $u_{-i} = \{u_1,\dots,u_n\} - \{u_i\}$. The exchangeability property of the Pólya urn representation shows that

$$u_i|u_{-i} \sim \frac{n-1}{n-1+\alpha}G_{-i}(\cdot) + \frac{\alpha}{n-1+\alpha}G_0(\cdot), \tag{3}$$

where $G_{-i}(\cdot) = \frac{1}{n-1}\sum_{j\neq i}\delta_{u_j}(\cdot)$. The Gibbs sampler Eq. (3) can be seen as a posterior sampler based on the prior $DP(\alpha,G_0)$ and dataset u_{-i} due to the conjugate property of a DP. On the basis of the right side of Eq. (3), a zero-mean base distribution G_0 is a natural selection due to the constraint $\int u dG(u) = 0$. However, $G_{-i}(\cdot)$ may not be zero-mean.

Next, we develop an empirical-likelihood-based zero-mean approximation to replace $G_{-i}(\cdot)$ on the right side of Eq. (3). We broaden our perspective to $G_{-i}^{(\mathbf{p})}(\cdot) = \sum_{j\neq i} p_{j,-i}\delta_{\mu_j}(\cdot)$, the weighted empirical form of $G_{-i}(\cdot)$, where each coordinate of the vector $\mathbf{p} = (p_{1,-i},\dots,p_{i-1,-i},p_{i+1,-i},\dots,p_{n,-i})^T$ is a probability weight summing to one.

By choosing the weights appropriately, we construct a zero-mean measure $G_{-i}^{(\mathbf{p})}(\cdot)$ close to $G_{-i}(\cdot)$. Recall that in Chapter 2 of Owen (2001), the empirical likelihood is defined as the maximum of the likelihood ratio

$$R(\mathbf{p}) = \prod_{j\neq i}(n-1)p_{j,-i} \tag{4}$$

subject to the constraints

$$\sum_{j\neq i} p_{j,-i} = 1,\; \sum_{j\neq i} p_{j,-i}u_j = 0,\; p_{j,-i} \geq 0, j = 1,\dots,n, j\neq i.$$

Note that in Chapter 3 of Owen (2001), the equation (4) can be seen as a negative Cressie-Read power divergence between $G_{-i}^{(\mathbf{p})}(\cdot)$ and $G_{-i}(\cdot)$ for $\gamma \to 0$ denoted by

$$CR_\gamma(\mathbf{p}) = \frac{2}{\gamma(\gamma+1)}\sum_{j\neq i}\left[\left((n-1)p_{j,-i}\right)^{-\gamma} - 1\right].$$

$R(\mathbf{p})$ can be maximized at points

$$p_{j,-i}^{*} = \frac{1}{n-1}\frac{1}{1+\lambda u_j}, \tag{5}$$

where λ is the solution of the equation

$$\sum_{j\neq i}\frac{u_j}{1+\lambda u_j} = 0.$$

The typical solution of λ via the Newton-Raphson algorithm can be found in Owen (2001). In our numerical studies, we use the R package `emplik` to solve the empirical likelihood function; refer to http://www.ms.uky.edu/∼mai/EmpLik.html.

The resultant zero-mean adjustment of Eq. (3) with the involvement of EL becomes

$$\begin{aligned} u_i|u_{-i} \sim & \frac{n-1}{n-1+\alpha} G_{-i}^{(\mathbf{p}^*)} + \frac{\alpha}{n-1+\alpha} G_0 \\ = & \frac{n-1}{n-1+\alpha} \sum_{j\neq i} p^*_{j,-i} \delta_{\mu_j} + \frac{\alpha}{n-1+\alpha} G_0. \end{aligned} \tag{6}$$

Eq. (6) can be seen as a synthesized posterior sampler based on dataset u_{-i}, prior $DP(\alpha, G_0)$, and constraint information $\int u dG(u) = 0$. We call it an EL-DP sampler hereafter, and can be used to generate a Markov chain of $\{u_1, \ldots, u_n\}$.

Then, returning to the mixture model (2), we can construct a so-called EL-DPM Gibbs sampler. That is, suppose that ε_i's, $i = 1, \ldots, n$, meet both the mixture model (2) and $G \sim \mathrm{DP}(\alpha, G_0)$. Given the latent variable $u_i \sim G$, we have $\varepsilon_i|u_i \sim \varphi(\varepsilon_i|u_i, \sigma^2)$. Based on Eq. (6) and a similar discussion as in Section 3 of Neal (2000), we obtain the Gibbs sampler

$$u_i|u_{-i}, \varepsilon_i \sim \sum_{j\neq i} q_{i,j} \delta_{u_j} + r_i H_i(u_i), \tag{7}$$

where

$$\begin{aligned} q_{i,j} &= k \cdot (n-1) \cdot p^*_{j,-i} \varphi(\varepsilon_i|u_j, \sigma^2), \\ r_i &= k \cdot \alpha \int \varphi(\varepsilon_i|u, \sigma^2) dG_0(u), \\ H_i(u_i) &= \frac{\alpha \varphi(\varepsilon_i|u_i, \sigma^2) dG_0(u_i)}{\alpha \int \varphi(\varepsilon_i|u_i, \sigma^2) dG_0(u_i)} \\ &\propto \varphi(\varepsilon_i|u_i, \sigma^2) dG_0(u_i), \end{aligned} \tag{8}$$

and k is a parameter such that $\sum_{j\neq i} q_{i,j} + r_i = 1$.

Finally, it is safe to return to the model (1). Once $\{u_1, \ldots, u_n\}$ or, equivalently, the unknown G can be drawn by the proposed EL-DPM Gibbs sampler, we can apply the standard Gibbs sampling algorithm to draw posterior samples for the remaining unknown parameter components β, σ^2, ρ, and $\mathbf{b}_i$, $i = 1, \ldots, m$.

Remark 1 In practice, constraints other than zero-mean may be added to distribution G. Assume that G satisfies the moment constraints

$$\int g(\nu, \mu) dG(\mu) = 0,$$

for a known function $g(\nu, \cdot)$ with a nuisance parameter ν. With our method, the empirical likelihood can be obtained by

$$\max \left\{ \prod_{j\neq i} (n-1) p_{j,-i} | p_{j,-i} \geq 0, \sum_{j\neq i} p_{j,-i} = 1, \sum_{j\neq i} p_{j,-i} g(\nu, \mu_j) = 0 \right\}.$$

In practice, the base distribution G_0 should be selected to satisfy the constraint

$$\int g(\nu, \mu) dG_0(\mu) = 0,$$

for all ν. For example, in a quantile regression model, the median of the mixing distribution G may be assumed to be zero, which means $\int_{-\infty}^{\infty} \Phi(-u/\sigma) dG(u) = \frac{1}{2}$, where $\Phi(\cdot)$ is the normal cumulative distribution and σ is a nuisance parameter. Accordingly, G_0 can be selected to satisfy the equation $\int_{-\infty}^{\infty} \Phi(-u/\sigma) dG_0(u) = \frac{1}{2}$, for example, a normal distribution with zero-mean.

In simple cases of constraints on the mean and median, the questions appear to be solvable. However, it is non-trivial to find an appropriate G_0 when it is subject to complex moment restrictions. If such a G_0 does not exist, G_0 can be selected to satisfy the conditions

$$\int \int g(\nu, \mu) \pi(\nu) dG_0(\mu) = 0,$$

where $\pi(\nu)$ is the prior for parameter ν. Practically, an ad hoc way to select G_0 is to take a normal distribution with hyperparameters including location μ_0 and variance τ_0^2, both of which can be obtained by solving the equation above. This process is feasible in most situations.

Remark 2 As a special case of the Cressie-Read power divergence family $CR_\gamma(\mathbf{p})$ for $\gamma \to 0$, the empirical likelihood employed in our paper acts as a criterion to measure the distance between two discrete distributions. The Cressie-Read divergence $CR_\gamma(\mathbf{p})$ proposed abundant distances for different needs when γ varies; refer to Chapter 3 of Owen (2001). For example, as $\gamma \to -1$, the limitation $CR_{-1}(\mathbf{p})$ is the Kullback-Leibler divergence (empirical entropy) between $G_{-i}^{(\mathbf{p})}$ and G_{-i}, and for $\gamma = -2$, $CR_{-2}(\mathbf{p})$ is the Euclidean likelihood distance between $G_{-i}^{(\mathbf{p})}$ and G_{-i}. One obvious advantage of the empirical likelihood compared to the Euclidean likelihood is that the latter may produce negative probabilities $p_{j,-i}$ while the empirical likelihood does not. Therefore, we use the empirical likelihood as a measure to search out a mean-zero approximate distribution such that it is the nearest to the distribution generated by the DP.

For studies involving a Dirichlet process mixture subject to constraints, under various modeling, the Cressie-Read power divergence is the most common tool. Kitamura and Otsu (2011), Shin (2014), and Choi (2016) all applied the Kullback-Leibler divergence, or relative entropy, to address the restriction information, whereas we apply the empirical likelihood. They modified the prior by incorporating the auxiliary information at the beginning and conducted the analysis in the usual Bayesian framework. Our insight based on the empirical likelihood is different: we modify the posterior sampler by incorporating the constraint information during the Gibbs sampling by calculating the projection $G_{-i}^{(p*)}$ for each G_{-i}.

Remark 3 In practice, latent samples $u_1,\dots,u_n$ with zero-mean can be simulated with the Gibbs sampling algorithm by Eq. (7). The precondition to use the algorithm is that the exchangeability of $u_1,\dots,u_n$ can be verified based on our method. We admit that using the adjusted Pólya urn scheme to draw $u_1,\dots,u_n$ may destroy the exchangeability for $u_1,\dots,u_n$. *Fortunately*, the adjusted Pólya urn scheme that we propose can be seen as a projection of the standard Pólya urn scheme of a DP. The exchangeability of $u_1,\dots,u_n$ drawn from the latter was verified in Escobar and West (1995). We illustrate the geometric interpretation in Figure 1. Denote $\mathscr{H}$ as the space of all distributions with form $G_{-i}^{(\mathbf{p})}$, and $\mathscr{H}_0$ is a subset of $\mathscr{H}$ with zero-mean distributions. Then, $G_{-i}^{(\mathbf{p}*)}$ is the projection of $G_{-i}^{(\mathbf{p})}$ in $\mathscr{H}_0$. In fact, for any point $G_{-i}^{(p)} \in \mathscr{H}$, its projection in $\mathscr{H}_0$ is defined as the nearest point in $\mathscr{H}_0$.

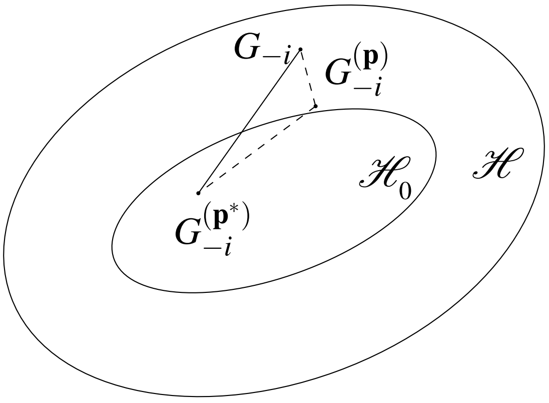


Fig. 1 The adjusted Pólya urn scheme can be looked as a projection of a standard Pólya urn scheme of DP

3 Gibbs Sampling Algorithm

In this section, we sketch the framework of an approximate MCMC sampling based on model (1), the mixture distribution (2), posterior sampler (6), and other parametric priors for $\alpha,\sigma^2,\rho,\boldsymbol{\beta}$, and $\mathbf{b}_i$, $i=1,\dots,m$.

The observations y_{ij}s can be produced from the following hierarchical model:

$$\begin{aligned}
&\sigma^2,\rho,\boldsymbol{\beta},\mathbf{b}_i \sim \pi(\sigma^2,\rho,\boldsymbol{\beta},\mathbf{b}_i), i=1,\dots,m,\\
&\varepsilon_{ij} \sim \mathrm{N}(u_{ij},\sigma^2), j=2,\dots,n_i, i=1,\dots,m,\\
&u_{ij}|G \sim G, G \sim \mathrm{DP}(\alpha,G_0), \qquad (9)\\
&w_{ij} = \rho w_{i,j-1} + \varepsilon_{ij},\\
&y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i + w_{ij}.
\end{aligned}$$

In the analysis below, the prior specifications for the parameters are:

$$\begin{aligned}
&G_0 = \mathrm{N}(0,\sigma_0^2), \alpha \sim \mathrm{Ga}(a_0,b_0),\\
&\sigma^2 \sim \mathrm{IG}(c_0,d_0), \rho \sim \mathrm{U}(-1,1),\\
&\boldsymbol{\beta} \sim \mathbf{N}_p(\mathbf{0},\Sigma_0), \mathbf{b}_i \sim \mathbf{N}_q(\mathbf{0},D(\boldsymbol{\eta})), \boldsymbol{\eta} \sim g(\boldsymbol{\eta}).
\end{aligned}$$

Here, $\mathrm{Ga}(a_0,b_0)$ stands for the gamma distribution with shape parameter a_0 and rate parameter b_0, $\mathrm{IG}(c_0,d_0)$ stands for the inverse-gamma distribution with shape parameter c_0 and scale parameter d_0, $D(\boldsymbol{\eta})$ is the covariance matrix characterized by the hyperparameter vector $\boldsymbol{\eta}$, and g is a known probability distribution.

Next, we establish the posterior sampling of G through the EL-DPM sampler (7), and then perform Gibbs sampling for the remaining parameters. Denote $\mathbf{u} = \{u_{ij}, j=2,\dots,n_i, i=1,\dots,m\}$ and $\mathbf{b} = \{\mathbf{b}_i, i=1,\dots,m\}$. Based on the likelihood function

$$\prod_{i=1}^{m}\prod_{j=2}^{n_i} f_G(\varepsilon_{ij}|\sigma^2) = \prod_{i=1}^{m}\prod_{j=2}^{n_i}\int \varphi(\varepsilon_{ij}|u,\sigma^2)dG(u), \qquad (10)$$

where $\varepsilon_{ij} = (y_{ij} - \mathbf{x}_{ij}^T\boldsymbol{\beta} - \mathbf{z}_{ij}^T\mathbf{b}_i) - \rho(y_{i,j-1} - \mathbf{x}_{i,j-1}^T\boldsymbol{\beta} - \mathbf{z}_{i,j-1}^T\mathbf{b}_i)$, we develop the Gibbs sampler for sampling $u_{ij},\alpha,\sigma^2,\rho,\boldsymbol{\beta}$, $\mathbf{b}$ and $\boldsymbol{\eta}$ through the following steps:

1. Drawing $u_{ij}, j=2,\dots,n_i, i=1,\dots,m$: Denote $\mathbf{u}_{-(i,j)} = \{u_{ls},(l,s)\neq(i,j)\}$. The conditional density of u_{ij} given $\mathbf{u}_{-(i,j)},\alpha,\sigma^2,\rho,\boldsymbol{\beta},\mathbf{b}$ and the data is

$$\begin{aligned}
&f(u_{ij}|\mathbf{u}_{-(i,j)},\alpha,\sigma^2,\rho,\boldsymbol{\beta},\mathbf{b},\text{data})\\
&= \frac{\alpha\varphi(\varepsilon_{ij}|u_{ij},\sigma^2)dG_0(u_{ij}) + \sum_{(l,s)\neq(i,j)} q_{ij,ls}\delta(u_{ls})}{r_{ij} + \sum_{(l,s)\neq(i,j)} q_{ij,ls}}
\end{aligned}$$

where

$$\begin{aligned}
&q_{ij,ls} = (N-m-1)\cdot p^*_{ls,-ij}\varphi(\varepsilon_{ij}|u_{ls},\sigma^2),\\
&r_{ij} = \alpha\int\varphi(\varepsilon_{ij}|u,\sigma^2)dG_0(u),
\end{aligned}$$

and $p^*_{ls,-ij}$s are empirical likelihood weights calculated with equation (5) based on $\mathbf{u}_{-(i,j)}$. Specifically, let $s_{ij} = \sum_{(l,s)\neq(i,j)} q_{ij,ls}$,

1a. With probability $r_{ij}/\{r_{ij}+s_{ij}\}$, generate u_{ij} from the density that is proportional to $\alpha\varphi(\varepsilon_{ij}|u_{ij},\sigma^2)dG_0(u_{ij})$;

1b. With probability $s_{ij}/\{r_{ij}+s_{ij}\}$, generate u_{ij} from the set $\{u_{ls},(l,s)\neq(i,j)\}$ using probabilities proportional to $q_{ij,ls}$.

2. Drawing α: Let d be the number of distinct values of $\mathbf{u}$. Then, α is updated by the following steps.

2a. Based on the current value of α, draw ν from $\mathrm{Beta}(\alpha+1,N-m)$;

2b. Given ν, draw a new value for α from a mixed gamma distribution:

$$\pi_\nu \mathrm{Ga}(a_0+d, b_0-\log(\nu)) + (1-\pi_\nu)\mathrm{Ga}(a_0+d-1, b_0-\log(\nu)),$$

where $\pi_\nu/(1-\pi_\nu) = (a_0+d-1)/[(N-m)(b_0-\log(\nu))]$.

3. Drawing σ^2: The prior distribution for σ^2 is $\mathrm{IG}(c_0,d_0)$. Based on the likelihood function (10), we draw σ^2 from the posterior distribution

$$[\sigma^2|\mathbf{u},\rho,\boldsymbol{\beta},\mathbf{b},\text{data}] \sim \mathrm{IG}(c,d),$$

where

$$c = c_0 + l/2,$$
$$l = (n_1-1)+\cdots+(n_m-1),$$
$$d = d_0 + \sum_{i=1}^{m}\sum_{j=2}^{n_i}(w_{ij}-\rho w_{i,j-1}-u_{ij})^2/2.$$

4. Drawing ρ: By combining the uniform prior with the likelihood function, we derive the posterior distribution for ρ:

$$[\rho|\mathbf{u},\sigma^2,\boldsymbol{\beta},\mathbf{b},\text{data}] \sim \mathrm{N}\left(\hat{\rho}, \left(\sum_{i=1}^{m}\sum_{j=2}^{n_i}\sigma^{-2}w_{i,j-1}^2\right)^{-1}\right)\mathbf{I}_{[-1,1]},$$

where

$$\hat{\rho} = \sum_{i=1}^{m}\sum_{j=2}^{n_i} w_{i,j-1}(w_{ij}-u_{ij}) \Big/ \sum_{i=1}^{m}\sum_{j=2}^{n_i} w_{i,j-1}^2.$$

5. Drawing $\boldsymbol{\beta}$: Based on the multivariate normal prior $\mathbf{N}_p(\mathbf{0},\Sigma_0)$ and the likelihood function, the posterior distribution for $\boldsymbol{\beta}$ is a multivariate normal distribution

$$[\boldsymbol{\beta}|\mathbf{u},\sigma^2,\rho,\mathbf{b},\text{data}] \sim \mathbf{N}_p(\mu_\beta,\Sigma_\beta),$$

where

$$\Sigma_\beta = \left(\Sigma_0^{-1} + \sigma^{-2}\sum_{i=1}^{m}\sum_{j=2}^{n_i}(\mathbf{x}_{ij}-\rho\mathbf{x}_{i,j-1})(\mathbf{x}_{ij}-\rho\mathbf{x}_{i,j-1})^T\right)^{-1},$$

$$\mu_\beta = \sigma^{-2}\sum_{i=1}^{m}\sum_{j=2}^{n_i}(A_{ij}-u_{ij})\Sigma_\beta(\mathbf{x}_{ij}-\rho\mathbf{x}_{i,j-1}),$$

$$A_{ij} = (y_{ij}-\mathbf{z}_{ij}^T\mathbf{b}_i) - \rho(y_{i,j-1}-\mathbf{z}_{i,j-1}^T\mathbf{b}_i).$$

6. Drawing $\mathbf{b}_i, i=1,\ldots,m$: By combining the multivariate normal prior $\mathbf{N}_q(\mathbf{0},D(\boldsymbol{\eta}))$ for $\mathbf{b}_i$ and the likelihood function, the posterior distribution is

$$[\mathbf{b}_i|\mathbf{u}_i,\sigma^2,\rho,\boldsymbol{\beta},\boldsymbol{\eta},\text{data}] \sim \mathbf{N}(\mu_{b_i},\Sigma_{b_i}),$$

where

$$\Sigma_{b_i} = \left(D(\boldsymbol{\eta})^{-1} + \sigma^{-2}\sum_{j=2}^{n_i}(\mathbf{z}_{ij}-\rho\mathbf{z}_{i,j-1})(\mathbf{z}_{ij}-\rho\mathbf{z}_{i,j-1})^T\right)^{-1},$$

$$\mu_{b_i} = \sigma^{-2}\sum_{j=2}^{n_i}(B_{ij}-u_{ij})\Sigma_{b_i}(\mathbf{z}_{ij}-\rho\mathbf{z}_{i,j-1}),$$

$$B_{ij} = (y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}) - \rho(y_{i,j-1}-\mathbf{x}_{i,j-1}^T\boldsymbol{\beta}), j=2,\ldots,n_i.$$

7. Drawing $\boldsymbol{\eta}$: Based on the prior distribution $g(\boldsymbol{\eta})$, the posterior distribution density for $\boldsymbol{\eta}$ is proportional to

$$f(\boldsymbol{\eta}|\mathbf{b}) \propto g(\boldsymbol{\eta})\prod_{i=1}^{m}\varphi_q(\boldsymbol{b}_i|\mathbf{0},D(\boldsymbol{\eta})),$$

where $\varphi_q(\cdot|\mathbf{0},D(\boldsymbol{\eta}))$ stands for the q-dimensional multivariate normal density with mean $\mathbf{0}$ and covariance matrix $D(\boldsymbol{\eta})$.

Remark 4 In particular, we can extend our algorithm to the nonparametric random effects model situation. We assume that the random effects $\mathbf{b}_i$s follow an unknown distribution F and that F follows $\mathrm{DP}(\gamma,F_0)$, where $F_0 = \mathbf{N}_q(\mathbf{0},D(\boldsymbol{\eta}))$. Note that F is also zero-mean. Then, Steps 6 and 7 above should be modified by the EL-DPM Gibbs sampler (7) as follows:

6' Drawing $\mathbf{b}_i, i=1,\ldots,m$: Denote $\mathbf{b}_{-i} = \{\mathbf{b}_s, s\neq i\}$ and $\varepsilon_{ij}^b = (y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}) - \rho(y_{i,j-1}-\mathbf{x}_{i,j-1}^T\boldsymbol{\beta}) - u_{ij}$. The conditional density of $\mathbf{b}_i$ given $\mathbf{b}_{-i},\gamma,\sigma^2,\rho,\boldsymbol{\beta},\mathbf{u}_i$ and the data is

$$f(\mathbf{b}_i|\mathbf{b}_{-i},\gamma,\sigma^2,\rho,\boldsymbol{\beta},\mathbf{u}_i,\text{data})$$
$$= \frac{\gamma\prod_{j=2}^{n_i}\varphi(\varepsilon_{ij}^b|(\mathbf{z}_{ij}-\rho\mathbf{z}_{i,j-1})^T\mathbf{b}_i,\sigma^2)dF_0(\mathbf{b}_i) + \sum_{s\neq i}q_{i,s}^b\delta(\mathbf{b}_s)}{r_i^b + \sum_{s\neq i}q_{i,s}^b}$$

where

$$q_{i,s}^b = (m-1)\cdot p_{s,-i}^*\prod_{j=2}^{n_i}\varphi(\varepsilon_{ij}^b|(\mathbf{z}_{ij}-\rho\mathbf{z}_{i,j-1})^T\mathbf{b}_s,\sigma^2),$$

$$r_i^b = \gamma\int\prod_{j=2}^{n_i}\varphi(\varepsilon_{ij}^b|(\mathbf{z}_{ij}-\rho\mathbf{z}_{i,j-1})^T\mathbf{b},\sigma^2)dF_0(\mathbf{b}),$$

and $p_{s,-i}^*$s are empirical likelihood weights obtained by solving the following empirical likelihood

$$\max\left\{\prod_{s\neq i}(m-1)p_{s,-i}|p_{s,-i}\geq 0, \sum_{s\neq i}p_{s,-i}=1, \sum_{s\neq i}p_{s,-i}\mathbf{b}_s=\mathbf{0}\right\}.$$

Specifically, let $s_i^b = \sum_{s\neq i}q_{i,s}^b$,

6'a. With probability $r_i^b/\{r_i^b+s_i^b\}$, generate $\mathbf{b}_i$ from the density that is proportional to $\gamma\prod_{j=2}^{n_i}\varphi(\varepsilon_{ij}^b|(\mathbf{z}_{ij}-\rho\mathbf{z}_{i,j-1})^T\mathbf{b}_i,\sigma^2)dF_0(\mathbf{b}_i)$;

6'b. With probability $s_i^b/\{r_i^b+s_i^b\}$, generate $\mathbf{b}_i$ from the set $\{\mathbf{b}_s, s\neq i\}$ using probabilities proportional to $q_{i,s}^b$.

7' Drawing $\boldsymbol{\eta}$: Based on the prior distribution $g(\boldsymbol{\eta})$, the posterior distribution density for $\boldsymbol{\eta}$ is proportional to

$$f(\boldsymbol{\eta}|\mathbf{b}) \propto g(\boldsymbol{\eta})\prod_{i=1}^{s}\varphi_q(\phi_i|\mathbf{0},D(\boldsymbol{\eta})),$$

where $\phi_1,\ldots,\phi_s$ are the distinct values of $\mathbf{b}_i, i=1,\ldots,m$.

Updating the hyperparameter γ is similar to Step 2 of updating α.

4 Simulation Studies

We conduct a series of simulation studies to demonstrate the performance of the proposed method (EL-DPM) and compare it with several alternative methods. We consider both mean regression and median regression models under the mixed effects model setting. In the mean regression scenarios, we consider the two cases where the true random effects are normally distributed and non-normally distributed. In the median regression scenarios, we consider the full model with AR(1) errors and a sub-model without AR(1) errors.

4.1 Normal random effects

In this scenario, 500 datasets are generated from the linear mixed effects model

$$y_{ij} = \beta_0 + \beta_1 x_{ij,1} + \beta_2 x_{ij,2} + b_i + w_{ij},\ j = 1,\ldots,n_i, i = 1,\ldots,m, \tag{11}$$

with $m = 30$ and n_i generated from $\text{Binom}(2,0.5)+4$, i.e., the value of n_i is 4, 5 or 6 with equal probability. We set $\beta_0 = \beta_1 = \beta_2 = 1$ and $b_i \sim \text{N}(0,\sigma_b^2)$ with $\sigma_b^2 = 1$; $x_{ij,1}$ and $x_{ij,2}$ independent with $x_{ij,1} \sim \text{Binom}(1,0.5)$ and $x_{ij,2} \sim \text{N}(0,3)$; and $w_{ij} = \rho w_{i,j-1} + \varepsilon_{ij}$ with $\rho = 0.5$ and $\varepsilon_{ij} \sim 0.5(\text{N}(2,0.5)$ $+\text{LN}(0,0.35) - \exp(0.175)) + 0.5(\text{N}(-2,0.5) + \text{LN}(0,0.35)$ $-\exp(0.175))$, where $\text{LN}(0,0.35)$ is the lognormal distribution with mean $\exp(0.175)$ and variance $\exp(0.35) \times (\exp(0.35)-1)$. The term $\exp(0.175)$ is subtracted to keep the mean of ε_{ij} zero. This is a case in which the noise comes from a nonsymmetric bimodal distribution with a slightly heavy tail for each component.

We compare the EL-DPM model with the centered stick-breaking mixture (CSBM) model proposed by Yang et al (2010), the DP mixture model (DPM) without zero-mean adjustment, the Bayesian parametric model with normal likelihood (Normal) and the Bayesian parametric model with t-likelihood (t-dist). For all the five methods, we assume a relatively flat prior $\text{N}(0, 100 \times \mathbf{I}_{3\times3})$ for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$. An inverse-gamma prior $\text{IG}(1,1)$ is set for both σ_b^2 and σ^2. This vague prior is proper and has no mean. We use the uniform prior $\text{U}(-1,1)$ for ρ. For the three methods related to DP, the hyperprior for the precision parameter α is the gamma distribution $\text{Ga}(1,0.5)$, with mean 2 and variance 4. We set the hyperparameter σ_0^2 of the base distribution $G_0 = \text{N}(0,\sigma_0^2)$ to be the maximum likelihood estimate of the variance of residuals when the noise distribution is normal.

For each replication, after a burn-in of 5,000 iterations, 5,000 samples are obtained, with every 10th saved for posterior inference due to the autocorrelation between the iterations. The convergence of the Markov chains is evaluated by Geweke's method (Geweke, 1992) using the R package coda. The Geweke statistics is a standard Z-score that indicates plausible convergence if its absolute value is less than 1.96. We obtain realizations of the Geweke statistics of each parameter for each replication. For each parameter, the percentage of observations of the Geweke statistic that fall into (-1.96, 1.96) is approximately 85%, which indicates that convergence is good for 10,000 iterations. We also conduct a few simulations with more MCMC iterations and find that the percentage increases. Taking the time issue into account, for simulation studies, 10,000 MCMC iterations is sufficient for convergence.

Table 1 Bias, MSE, SD, SE and CP from different methods with normal random effects

	β_0				
Method	Bias	MSE	SD	SE	CP
EL-DPM	0.0636	0.2128	0.4574	0.5152	0.9660
CSBM	0.0317	0.2292	0.4781	0.5289	0.9400
DPM	0.1406	0.2574	0.4880	0.9782	0.9980
Normal	0.0317	0.2486	0.4981	0.5950	0.9640
t-dist	0.0209	**0.1371**	0.3701	0.4020	0.9580
	β_1				
EL-DPM	-0.0108	**0.0499**	0.2233	0.2222	0.9540
CSBM	-0.0079	0.0540	0.2326	0.2464	0.9660
DPM	-0.0095	0.0623	0.2496	0.2859	0.9800
Normal	-0.0113	0.1474	0.3842	0.3707	0.9400
t-dist	-0.0037	0.1344	0.3669	0.3476	0.9380
	β_2				
EL-DPM	0.0041	**0.0043**	0.0657	0.0648	0.9460
CSBM	0.0037	0.0048	0.0690	0.0716	0.9620
DPM	0.0019	0.0050	0.0711	0.0836	0.9780
Normal	-0.0022	0.0115	0.1075	0.1074	0.9480
t-dist	0.0005	0.0097	0.0983	0.1012	0.9400
	ρ				
EL-DPM	0.0324	**0.0043**	0.0571	0.0587	0.9000
CSBM	0.0729	0.0089	0.0602	0.0718	0.8040
DPM	0.0535	0.0065	0.0608	0.0728	0.8940
Normal	0.0180	0.0098	0.0974	0.1053	0.9520
t-dist	-0.0784	0.0148	0.0933	0.1061	0.9300
	σ_b^2				
EL-DPM	0.0874	0.1692	0.4023	0.6611	0.9980
CSBM	0.1872	0.1723	0.3708	0.8836	1.0000
DPM	-0.1197	**0.0560**	0.2042	0.6421	1.0000
Normal	0.4920	0.4853	0.4937	1.2033	1.0000
t-dist	0.4358	0.4831	0.5419	0.9015	0.9940

Table 1 reports the estimated bias (Bias) given by the mean of the estimates minus the true values, the mean squared error of the estimates (MSE), the standard deviations of the posterior means (SD), the averaged posterior standard deviations of the estimates (SE), and the empirical 95% coverage probabilities (CP). For each parameter, the smallest MSE is set in bold type.

From Table 1 we know that the bias of the intercept β_0 estimated by the DPM method is the largest, and its 95% coverage probability is far from the nominal value. For the EL-DPM method, there might be some inflation of the estimation of the intercept β_0 because the constraint is over the mean function, but it performs comparably with the CSBM method and better than the DPM and the Normal model. For the estimation of β_1, β_2 and ρ, our EL-DPM method uniformly beats the other four methods, especially the Normal model and the t-model, with the consistently smallest MSE values of the estimates. The EL-DPM method yields stable estimates of the 95% coverage probabilities, which are close to the nominal value, whereas the DPM model has much higher coverage probabilities. The EL-DPM model also presents the smallest bias of σ_b^2. We conclude that our method works well in this situation.

4.2 Non-normal random effects

We further consider the case where the true random effects are non-normally distributed. The sampling model follows equation (11), except that the random effects are $b_i \sim 0.5\mathrm{N}(\sqrt{1.5}, 0.5) + 0.5\mathrm{N}(-\sqrt{1.5}, 0.5)$, a two-component mixture of normal distributions. All the other simulation settings and the priors of the parameters follow those in Section 4.1.

We compare the performance of the EL-DPM model with that of the other four methods in Section 4.1. For the two Bayesian parametric models, the priors of the random effects are both normal distributions. For the DPM model, we assume a DP prior on the distribution of the random effects. For the CSBM model, we assume the CSBP (Yang et al, 2010) prior on the distribution of the random effects. Table 2 reports the posterior inference with the five models. Since σ_b^2 is estimated with the nonparametric priors for the EL-DPM, DPM and the CSBM methods, we report only the estimations of the parameters that are estimated with parametric priors. We can see that our EL-DPM model is consistently better than the CSBM, DPM and Normal model and yields much smaller MSEs for β_1, β_2 and ρ than those of the t-model.

4.3 Median regression

We illustrate another application of the EL-DPM method with a different constraint through a median regression model for longitudinal data. For median regression, the only change in the algorithm is to replace the zero-mean constraint with that in *Remark 1*. We consider two models in the following:

- **Model 1**:

$$\mathbf{y}_i = \mathbf{x}_i\boldsymbol{\beta} + \mathbf{z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i;$$
$$\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i}),\ i = 1, \dots, m.$$

- **Model 2**:

$$\mathbf{y}_i = \mathbf{x}_i\boldsymbol{\beta} + \mathbf{z}_i\mathbf{b}_i + \mathbf{w}_i,\ i = 1, \dots, m;$$
$$\mathbf{w}_i = (w_{i1}, \dots, w_{in_i});\ w_{ij} = \rho w_{i,j-1} + \varepsilon_{ij}, j = 2, \dots, n_i.$$

Table 2 Bias, MSE, SD, SE and CP from different methods with non-normal random effects

Method	Bias	MSE	SD	SE	CP
			β_0		
EL-DPM	0.0369	0.1688	0.4096	0.5890	0.9900
CSBM	-0.0071	0.2036	0.4516	0.7714	0.9940
DPM	0.0920	0.2360	0.4774	1.1785	1.0000
Normal	-0.0088	0.2122	0.4611	0.6940	0.9920
t-dist	0.0146	**0.1037**	0.3220	0.4334	0.9860
			β_1		
EL-DPM	-0.0053	**0.0546**	0.2338	0.2403	0.9540
CSBM	-0.0066	0.0629	0.2509	0.2735	0.9620
DPM	-0.0053	0.0711	0.2669	0.3056	0.9740
Normal	-0.0068	0.1486	0.3858	0.3706	0.9360
t-dist	-0.0122	0.1322	0.3637	0.3528	0.9280
			β_2		
EL-DPM	0.0024	**0.0053**	0.0728	0.0699	0.9440
CSBM	0.0008	0.0062	0.0785	0.0789	0.9540
DPM	0.0015	0.0068	0.0826	0.0885	0.9660
Normal	0.0016	0.0128	0.1134	0.1073	0.9400
t-dist	0.0011	0.0111	0.1053	0.1021	0.9340
			ρ		
EL-DPM	0.0746	**0.0099**	0.0658	0.0636	0.7360
CSBM	0.1337	0.0214	0.0597	0.0712	0.4760
DPM	0.1012	0.0144	0.0645	0.0731	0.6780
Normal	0.0664	0.0147	0.1017	0.1045	0.8700
t-dist	-0.0634	0.0146	0.1030	0.1141	0.9520

Clearly, Model 1 is a sub-model as well as a special case of Model 2. For both models, we assume that the distribution of ε_{ij} follows the mixture distribution $f_G(\varepsilon) = \int \varphi(\varepsilon|u, \sigma^2)\, dG(u)$ and that the median of $f_G(\varepsilon)$ is zero. In this case, the constraint in *Remark 1* should be added to G.

For each model, 500 datasets are generated. We assume that ε_{ij}s are generated from a three-component mixture of normals, $p_1\varphi(\cdot|\mu_1, \sigma_1^2) + p_2\varphi(\cdot|\mu_2, \sigma_2^2) + (1 - p_1 - p_2)\varphi(\cdot|\mu_3, \sigma_3^2)$, with $p_1 = 0.435$, $p_2 = 0.43, \mu_1 = -0.4, \sigma_1 = 1, \mu_2 = 0, \sigma_2 = 1.5, \mu_3 = 5$ and $\sigma_3 = 2$. The median of the distribution is equal to 0 up to two decimal points. All the other simulation settings follow from Section 4.1, except for Model 2, where the true value of ρ is set to 0.2.

We compare the performance of the EL-DPM method for Model 1 with that of two existing approaches. The first is the fully Bayesian quantile regression model for longitudinal data proposed by Luo et al (2012). They assumed that the error term is subject to an asymmetric Laplace distribution and established a hierarchical Bayesian model. They developed both a Metropolis-Hastings algorithm and Gibbs sampling algorithm for posterior inference. We compare our results with those of their Gibbs sampling approach (Para). The other approach is a flexible Bayesian quantile regression model proposed by Reich et al (2010). They assumed that the error distribution is an infinite mixture of Gaussian densities subject to a stochastic constraint. They extended the proposed approach to analyze conditional and marginal models for clustered data. We compare our results with those

Table 3 Bias, MSE, SD, SE and CP from different methods for Model 1

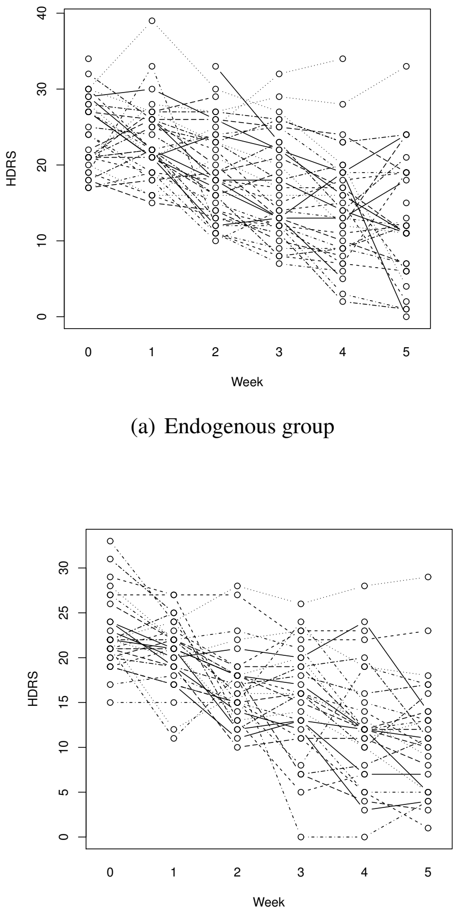
			β_0		
Method	Bias	MSE	SD	SE	CP
EL-DPM	0.0955	0.0782	0.2632	0.2789	0.9460
Para	0.0992	0.0837	0.2720	0.2893	0.9480
FBQR	0.0193	**0.0716**	0.2672	0.2769	0.9460
			β_1		
EL-DPM	0.0064	**0.0836**	0.2894	0.2806	0.9460
Para	0.0059	0.0960	0.3100	0.3165	0.9560
FBQR	0.0056	0.0856	0.2928	0.2792	0.9420
			β_2		
EL-DPM	-0.0074	**0.0067**	0.0819	0.0814	0.9480
Para	-0.0084	0.0083	0.0907	0.0924	0.9480
FBQR	-0.0065	0.0069	0.0829	0.0815	0.9320
			σ_b^2		
EL-DPM	0.0664	0.1407	0.3695	0.4143	0.9540
Para	0.0626	0.1599	0.3954	0.4461	0.9700
FBQR	0.0528	**0.1344**	0.3632	0.4120	0.9560

Table 4 Bias, MSE, SD, SE and CP from different methods for Model 2

			β_0		
Method	Bias	MSE	SD	SE	CP
EL-DPM	0.1020	0.0961	0.2930	0.3420	0.9640
Para	0.0606	**0.0931**	0.2993	0.3575	0.9700
			β_1		
EL-DPM	0.0067	**0.0926**	0.3045	0.3042	0.9440
Para	0.0024	0.1116	0.3343	0.3407	0.9540
			β_2		
EL-DPM	-0.0045	**0.0076**	0.0874	0.0880	0.9420
Para	-0.0022	0.0087	0.0932	0.0981	0.9500
			ρ		
EL-DPM	0.0391	**0.0095**	0.0894	0.0813	0.8900
Para	0.0546	0.0120	0.0952	0.0937	0.8920
			σ_b^2		
EL-DPM	0.0230	0.7928	0.3402	0.5277	0.9960
Para	0.0041	**0.7735**	0.3566	0.5687	0.9940

of their conditional model with homogeneous regression errors (FBQR). For Model 2, we compare the modified Gibbs sampling algorithm of Luo et al (2012) for the AR(1) model setting.

Table 3 and Table 4 report the posterior inference for Model 1 and Model 2 from different methods, respectively. Table 3 shows that our EL-DPM method performs comparably with the FBQR method and is uniformly better than the fully parametric Bayesian method, with smaller MSE, SD and SE for all the parameters. The 95% coverage probabilities are also reasonable. From Table 4, we know that the EL-DPM method yields smaller MSE, SD and SE of β_1, β_2 and ρ than those of the fully parametric Bayesian method.



(a) Endogenous group

(b) Nonendogenous group

Fig. 2 Spaghetti plot of observed data

5 Application

We apply our method to analyze the data from a psychiatric study described in Reisby et al (1977). This dataset was also considered in Hedeker and Gibbons (2006). The sample consists of $m = 66$ depressed inpatients, with 29 patients classified as nonendogenous (NE) and 37 as endogenous (E). The dependent variable measured across time is the Hamilton Depression Rating Scale (HDRS). Subjects were rated twice during the baseline placebo week (at the start and end of the week) as well as at the end of each of the four treatment weeks of the study. The number of longitudinal measurements varies from 4 to 6 across patients. For more details of the data, see Hedeker and Gibbons (2006). Figure 2 shows the HDRS scores for all patients in the two groups (E and NE).

We consider the following model

$$y_{ij} = b_{0i} + b_{1i}t_{ij} + w_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, m, \qquad (12)$$

where

$$b_{0i} = \beta_0 + \beta_2 DX_i + v_{0i},\ b_{1i} = \beta_1 + \beta_3 DX_i + v_{1i}. \qquad (13)$$

Here, $DX = 0$ or 1 indicates NE or E, respectively, and t_{ij} is the week (starting at 0). β_0 represents the intercept coefficient of NE patients, and β_1 is the coefficient of time for NE patients. Likewise, β_2 represents the intercept coefficient difference for E patients relative to NE patients, and β_3 is the difference in the coefficient of time for E patients relative to NE patients. Moreover, v_{0i} and v_{1i} are random effects. Equations (12)–(13) correspond to equation (1) by letting $\mathbf{x}_{ij} = (1, t_{ij}, DX_i, t_{ij}DX_i)^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$, $\mathbf{z}_{ij} = (1, t_{ij})^T$ and $\mathbf{b}_i = (v_{0i}, v_{1i})^T$. We assume that

$$w_{ij} = \rho w_{i,j-1} + \varepsilon_{ij}, j = 2, \ldots, n_i, \qquad (14)$$

where ε_{ij} is zero-mean error.

The prior distribution for the fixed effects (β_0, β_1) is chosen as $\mathrm{N}(0, 10^4\mathbf{I}_{2\times 2})$. For the random effects (v_{0i}, v_{1i}), we assume a $\mathrm{N}(0, \mathbf{R})$ prior, and $\mathbf{R}$ is also a 2×2 covariance matrix with an inverse Wishart prior distribution $\mathrm{IW}(2, 2 \times \mathbf{S})$. As in Lee and Niu (1999), $\mathbf{S}$ is set as diagonal with diagonal elements being the sample variances of the corresponding regression coefficients when each subject is regressed on the corresponding design matrix with normal residuals. For the other nuisance parameters, we assume that σ^2 follows $\mathrm{IG}(1, 1)$ and that ρ follows $\mathrm{U}(-1, 1)$. We assume $\mathrm{Ga}(2.5, 0.5)$ for the precision parameter α.

We first consider a special case of equations (12)–(14), where ρ is fixed at 0. In this case, G_0 is chosen as $\mathrm{N}(0, 12.22)$, with the variance of 12.22 being the variance of the residuals under the normality assumption. After a burn-in of 5,000 iterations, 10,000 samples are obtained, with every 10th saved for posterior inference. The trace plots and the Geweke's statistics suggest convergence of the Markov chains. We report the posterior means (Post.Mean), posterior standard deviations (Post.SD) and 95% confidence intervals (95% CI) of this scenario in Table 5. These results are consistent with those in Hedeker and Gibbons (2006).

Table 5 Results of the first scenario

Parameter	Post.Mean	Post.SD	95% CI
β_0	22.5075	0.7772	$(21.0195, 24.0813)$
β_1	-2.3248	0.3185	$(-2.9640, -1.6644)$
β_2	2.0364	1.0555	$(-0.1656, 4.0013)$
β_3	-0.0965	0.4281	$(-0.9388, 0.7351)$
$\sigma^2_{v_0}$	11.7838	3.3556	$(6.4497, 19.3617)$
$\sigma_{v_0 v_1}$	-1.3141	1.0041	$(-3.4488, 0.3115)$
$\sigma^2_{v_1}$	2.2272	0.5431	$(1.3797, 3.5417)$

We then consider fitting the whole model equations (12)–(14). First, we fit the model under the normality assumption of the noise. The normal Q-Q plot of the noise in Figure 3 indicates that the distribution of the noise is heavy tailed. Additionally, the p-values of both the Shapiro-Wilk normality test statistic and the one-sample Kolmogrov-Smirnov test statistic are below 0.05, which suggests that the normality assumption is improper. This result supports the use of the DPM model for the noise.

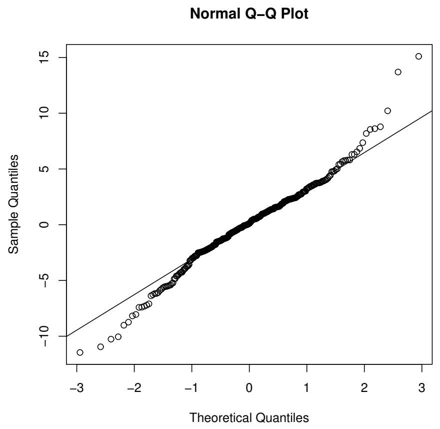


Fig. 3 Q-Q plot of the noises under normality assumption

The difference in the implementation of Gibbs sampling between fitting the whole model and the first scenario ($\rho = 0$) is that the baseline distribution G_0 is chosen as $\mathrm{N}(0, 17.97)$, with the variance 17.97 being the variance of the residuals under the normality assumption. The estimation results in this scenario are listed in Table 6.

The two tables show that the autoregressive coefficient ρ is significant, which indicates that the second model is reasonable. The effect of time (β_1) is significant, and we can conclude that the rate of improvement in the HDRS scores of the patients is significantly different from zero in this study. Additionally, there is no strong evidence that the two diagnostic groups differ in terms of HDRS scores across time.

For comparison, we also report the posterior inference from the CSBM and DPM models in Table 6. We can see that, generally, all three methods yield similar estimation patterns, and the EL-DPM model provides smaller posterior standard deviations and 95% confidence interval lengths than those of the other two.

Finally, we perform a sensitivity analysis of the proposed method with respect to α by considering another gamma prior $\mathrm{Ga}(1, 0.5)$ for α for both scenarios. The results (not shown) follow the same patterns as those in the above two tables. We also conduct multivariate normality tests for the posterior means of $\mathbf{b}_i = (v_{0i}, v_{1i})^T$. The p-values of the Henze-Zirkler's, Mardia's and Royston's multivariate normality tests are all much larger than 0.05, which suggests that the posterior means of these random effects are multivariate-normal distributed. Therefore, it is not necessary to assume a DP prior for the random effects in this data example.

Table 6 Results of the second scenario

	NE intercept β_0		
Method	Post.Mean	Post.SD	95% CI
EL-DPM	21.3427	2.1491	$(17.1466, 25.4609)$
CSBM	21.5512	2.7489	$(15.7885, 26.7857)$
DPM	20.9347	3.3847	$(13.7177, 27.2725)$
	NE slope β_1		
EL-DPM	-2.0664	0.4938	$(-3.0108, -1.0678)$
CSBM	-2.0794	0.5820	$(-3.1322, -0.8895)$
DPM	-1.9985	0.6511	$(-3.2267, -0.6461)$
	E intercept difference β_2		
EL-DPM	4.3099	3.2799	$(-1.7065, 10.7073)$
CSBM	4.2358	4.2909	$(-2.8734, 13.3420)$
DPM	5.3382	4.5618	$(-2.2576, 15.1822)$
	E slope difference β_3		
EL-DPM	-0.6008	0.7355	$(-2.0852, 0.7454)$
CSBM	-0.5688	0.8443	$(-2.1753, 1.0005)$
DPM	-0.7437	0.9007	$(-2.6204, 1.0037)$
	ρ		
EL-DPM	0.4837	0.0976	$(0.2873, 0.6690)$
CSBM	0.5494	0.1031	$(0.3501, 0.7562)$
DPM	0.5753	0.0921	$(0.3964, 0.7430)$
	$\sigma^2_{v_0}$		
EL-DPM	15.7999	9.8262	$(4.8476, 40.8242)$
CSBM	16.1515	12.5371	$(4.2617, 47.7944)$
DPM	17.4419	15.7057	$(4.3455, 53.8513)$
	$\sigma^2_{v_0 v_1}$		
EL-DPM	-1.8723	2.4444	$(-7.7006, 1.1521)$
CSBM	-1.8017	2.6177	$(-8.3923, 1.2613)$
DPM	-2.3215	3.2227	$(-10.2081, 0.9836)$
	$\sigma^2_{v_1}$		
EL-DPM	2.0412	0.8403	$(0.8786, 3.9470)$
CSBM	1.9241	0.8610	$(0.7680, 4.0403)$
DPM	1.9168	0.8962	$(0.7565, 4.1008)$

6 Discussion

In this paper, we present a Bayesian nonparametric model for mixing distributions subject to moment constraints within a longitudinal data setting. A useful fact that acts as the stepping stone of our methodology is that we obtain a posterior sampler in the form of a Pólya Urn representation by the conjugation of a DP prior when there are no restrictions.

In the presence of auxiliary information in the form of moment constraints, we employ the empirical likelihood technique to adjust the posterior sampler of the mixing distribution. In the establishment of our method, the EL plays the role of a distance measure to search for zero-mean approximate distributions that are the nearest to the posterior samples of the mixing distribution. Our EL-based adjustment method is actually a projection technique. Other projection methods include Choi (2016), Kitamura and Otsu (2011) and Shin (2014). However, in the presence of auxiliary information, our strategy of Bayesian inference is different from existing methods. For instance, Choi (2016) modified the prior to incorporate expert information via the Kullback-Leibler distance so that the analysis could be performed under the usual Bayesian framework. We approximate the posterior sampler by adjusting the Gibbs sampling algorithm to incorporate the moment constraints through the EL distance.

Our EL-DPM Gibbs sampler procedure is an approximate Markov Chain Monte Carlo (MCMC) method in the sense that we actually modify the standard MCMC by the proposed posterior sampler. Numerically, in our simulation section, we report that the resultant Markov Chain converges according to the criterion of Geweke's method; refer to Geweke (1992). However, theoretically, it is challenging to guarantee convergence for two reasons: the infinite dimensionality of the prior and the fact that the EL measurement and the Cressie-Read divergence are not metrics. To the best of our knowledge, the convergence results of perturbed chains are usually described by metric distances such as the Wasserstein distance or the total variation distance. For example, Roberts et al (1998) and Alquier et al (2016) studied the convergence properties of perturbed and unperturbed chains with metric distances and within a finite-dimensional metric space.

On the other hand, the empirical likelihood can be seen as the likelihood of the multinomial distribution where the support of the distribution is given by the empirical observations. Hence, the empirical likelihood is not the real likelihood since those observations are randomly distributed. In this way, some properties of the Gibbs sampling method based on the real likelihood may not be inherited by our method. Therefore, theoretical proof of the proposed method requires further exploration.

Finally, note that our proposed posterior sampling may act as a tool to facilitate Bayesian analysis that involves unknown distributions subject to moment constraints, regardless of the data setting or models.

Acknowledgments

The authors appreciate the Associate Editor and three referees for their constructive comments and guidance to the current version. The authors thank Prof. Yuichi Kitamura for providing their slide of the manuscript and Prof. Min-Hui Chen for his suggestions during the revision. The first author was partially supported by the National Natural Science Foundation of China (11471024); the research of the first two authors was supported by the Scientific Research Level Improvement Quota Project of Capital University of Economics and Business; the research of the third author was supported by the postgraduate studentship of Hong Kong Polytechnic University; and the research of the fourth author was partially supported by the National Natural Science Foundation of China 11401502, the Hong Kong Polytechnic

University fund G-UB01, and the General Research Fund of Hong Kong 15327216.

References

Alquier P, Friel N, Everitt R, Boland A (2016) Noisy monte carlo: Convergence of markov chains with approximate transition kernels. Statistics and Computing 26(1-2):29–47

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

Arnau J, Bono R, Blanca MJ, Bendayan R (2012) Using the linear mixed model to analyze nonnormal data distributions in longitudinal designs. Behavior Research Methods 44:1224–1238

Bartolucci F, Bacci S (2014) Longitudinal analysis of self-reported health status by mixture latent auto-regressive models. Journal of the Royal Statistical Society: Series C 63:267–288

Blackwell D, MacQueen JB (1973) Ferguson distributions via pólya urn schemes. The Annals of Statistics 1:353–355

Brunner LJ, Lo AY (1989) Bayes methods for a symmetric unimodal density and its mode. The Annals of Statistics 17:1550–1566

Chi EM, Reinsel GC (1989) Models for longitudinal data with random effects and AR(1) errors. Journal of the American Statistical Association 84:452–459

Choi H (2016) Expert information and nonparametric Bayesian inference of rare events. Bayesian Analysis 11(2):421–445

Damsleth E, El-Shaarawi A (1989) Arma models with double-exponentially distributed noise. Journal of the Royal Statistical Society Series B (Methodological) 51:61–69

Escobar MD (1994) Estimating normal means with a Dirichlet process prior. Journal of the American Statistical Association 89:268–277

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

Fan TH, Wang YF, Zhang YC (2014) Baysian model selection in linear mixed effects models with autoregressive (p) errors using mixture priors. Journals of Applied Statistics 41:1814–1829

Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1:209–230

Geweke J (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bayesian Statistics, University Press, pp 169–193

Goldstein H, Healy MJ, Rasbash J (1994) Multilevel time series models with applications to repeated measures data. Statistics in Medicine 13:1643–1655

Griffin JE (2016) An adaptive truncation method for inference in Bayesian nonparametric models. Statistics and Computing 26:423–441

Hedeker D, Gibbons RD (2006) Longitudinal Data Analysis. John Wiley & Sons

Hoff PD (2000) Constrained nonparametric estimation via mixtures. PhD thesis, Department of Statistics, University of Wisconsin

Hoff PD (2003) Nonparametric estimation of convex models via mixtures. The Annals of Statistics 31:174–200

Kitamura Y, Otsu T (2011) Bayesian analysis of moment condition models using nonparametric priors. Tech. rep., Yale University

Kleinman KP, Ibrahim JG (1998) A semiparametric Bayesian approach to the random effects model. Biometrics 54:921–938

Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 38:963–974

Lazar NA (2003) Bayesian empirical likelihood. Biometrika 90(2):319–326

Lee JC, Niu WF (1999) On an unbalanced growth curve model with random effects and AR(1) errors from a Bayesian and the ML points of view. Journal of statistical planning and inference 76:41–55

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

Luo Y, Lian H, Tian M (2012) Bayesian quantile regression for longitudinal data models. Journal of Statistical Computation and Simulation 82:1635–1649

MacEachern SN, Müller P (1998) Estimating mixture of Dirichlet process models. Journal of Computational and Graphical Statistics 7:223–238

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

Owen AB (2001) Empirical Likelihood. Chapman & Hall/CRC

Reich BJ, Bondell HD, Wang HJ (2010) Flexible Bayesian quantile regression for independent and clustered data. Biostatistics 11:337–352

Reisby N, Gram LF, Bech P, Nagy A, Petersen GO, Ortmann J, Ibsen I, Dencker SJ, Jacobsen O, Krautwald O (1977) Imipramine: clinical effects and pharmacokinetic variability. Psychopharmacology 54:263–272

Roberts G, Rosenthal J, Schwartz P (1998) Convergence properties of perturbed markov chains. Journal of Applied Probability 35(1):1–11

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

Shin M (2014) Bayesian generalized method of moments. Tech. rep., University of Illinois

Tiku ML, Wong WK, Vaughan DC, Bian G (2000) Time series models in non-normal situations: symmetric innovations. Journal of Time Series Analysis 21:571–596

Wang WL, Fan TH (2011) Estimation in multivariate *t* linear mixed models for multiple longitudinal data. Statistica Sinica 21:1857–1880

Yang M, Dunson DB, Baird D (2010) Semiparametric Bayes hierarchical models with mean and variance constraints. Computational Statistics and Data Analysis 54:2172–2186