Research Article

MCMC Sampling Estimation of Poisson-Dirichlet Process Mixture Models

Xiang Qiu , Linlin Yuan , and Xueqin Zhou

College of Sciences, Shanghai Institute of Technology, Shanghai, China

Correspondence should be addressed to Xueqin Zhou; xueqinzhou@163.com

Received 2 December 2020; Revised 6 September 2021; Accepted 15 September 2021; Published 21 October 2021

Academic Editor: Erik Cuevas

Copyright © 2021 Xiang Qiu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In this article, we aim to estimate the parameters of Poisson-Dirichlet mixture model with multigroup data structure by empirical Bayes. The number of mixture components with Bayesian nonparametric process priors is not fixed in advance and it can grow with the increase of data. Empirical Bayes is the useful method to estimate the mixture components without information on them in advance. We give the procedure to construct smooth estimates of base distribution G_0 and estimates of the two parameters (α, θ) . The performances of estimations for parameters under multigroup data are better than those of the single-group data with the same total size of individuals in the perspectives of bias, standard deviations, and mean squared errors by numerical simulation. Also, we applied Poisson-Dirichlet mixture models to well-known real datasets.

1. Introduction

Mixture models are typically used to model data in which each observation belongs to one of a set of distributions, being finite or infinite, with certain probabilities. In many situations, some elements or all of the mixing distributions are unknown. However, accuracy of prior information plays an important role in statistical inferences, and the deviation of prior information inevitably results in biased inference which may cause serious problems in subsequent decisions. A possible solution to the inference on parameters of the prior is the empirical Bayes (EB) methods with the price of using more data sources than pure Bayesian inference. EB is a kind of modern statistical analysis methodologies that blend frequentist inference into Bayesian ideas with the parameters in prior distributions estimated from the observed data (or Bayes decisions estimated from observed data). EB has been widely applied in many fields in its development process of more than half a century. Relevant contributions are, among others, Seidel [1] on estimating partial prior information by EB, George and Foster [2] on EB variable selection in regression analysis, Clyde and George [3] on EB estimation for wavelets, Yang and Wu [4] on EB estimates of the precision parameter and baseline

distribution with monotone missing data subject to Dirichlet process priors, Bi and Davuluri [5] developing a novel nonparametric empirical Bayesian-based approach to model the RNA-seq data, and Zhou et al. [6] presenting careful comparisons among several empirical Bayes estimates to the precision parameter of Dirichlet process prior. As of EB inference for nonparametric problems when priors are modeled by mixture models, one can refer to, for example, the works of Liu [7] and McAuliffe et al. [8].

A Dirichlet process (DP) [9] is a distribution of distributions and, thanks to its characteristics, can act as a prior distribution in nonparametric problems with possibly tied observations. Therefore, a DP mixture model allows for common distributions being shared by subsets of observations and thus forms generalizations of the finite mixture models (FMM). In FMM, the number of unknown distributions is supposed to be known and fixed and every individual is sampled from one of those distributions in different probability, whereas, in DP mixture models, the number of clusters is not needed to be given in advance and inferred based on the data. The number of clusters in a DP mixture model is random and grows in $\log(\text{sample size})$. The research efforts on DP mixture model in Bayesian nonparametric are quite large, which can be found in, for

example, the works of Ferguson [9], MacEachern and Müller [10], Neal [11], Ishwaran and James [12], Quintana et al. [13], and so on.

Being proposed by Pitman and Yor [14], Poisson-Dirichlet process (PDP) is also known as Pitman-Yor process, which can be seen as the extension of DPs. In PDP model, a new observation is either from the previous ones in some cluster with a probability which is an increasing function of the cluster size or a new draw from the base distribution.

As an effective and widely applied density estimation method, mixture modeling covers a great deal of phenomena in real-world practice. But it is usually hard to determine the number of mixture components. There are many methods treating the number of components as an unknown constant which is based on the data beforehand, for example, model selection methods. However, the assumption could be too rigid or wrong for the new individual may be from the previous cluster or new cluster as mentioned above. There are a lot of researches on the DP mixture model, such as the works of Berry and Christensen [15], McAuliffe et al. [8], Mostofi and Kharrati-Kopaei [16], and Hu et al. [17]. But there are few literatures about the PDP mixture model and its parameters' estimation.

In this article, we are interested in the estimation of base distribution under the multigroup data, where the density of base distribution is continuous. There are two parameters and base distribution in PDP mixture model, different from DP mixture model. One difficulty is the initial value of (α, θ) and its impact on the estimation of base distribution, where we suppose that $\alpha \in [0, 1], \theta > -\alpha$ in the context. Another one is that the base distribution could not be directly observed, which is the parameter of infinite dimensional distribution. The individuals of every group follow the random probability distribution in view of the base. We handle this by the nonparametric empirical Bayes; and the estimations of α, θ are referred to in the works, among others, of Carlton [18] and Favaro et al. [19]. The performance of estimations for α, θ under multigroup data is better than that of the single-group data with the same total size of individuals in the perspectives of bias, standard deviations, and mean squared errors.

The remainder of this paper is structured as follows. In Section 2, we review PDP mixture models. In Section 3, under the framework of multigroup data, the procedure of estimation on the base density is described and the algorithm with the nonconjugate Gibbs sampler is given. In Section 4, some simulations with different base distributions and pairs of (α, θ) are shown and we compare the results between single-group and multigroup data. In Section 5, the PDP mixture model is applied to the well-known real-world datasets.

2. Poisson-Dirichlet Process Mixture Model

This section formulates the Poisson-Dirichlet processes mixture models we are going to discuss.

The fundamental setting is a set of independent random couples $(X_1, \eta_1), (X_2, \eta_2), \dots, (X_n, \eta_n)$ such that, for every $i = 1, 2, \dots, n$, $X_i | \eta_i \sim f(x | \eta_i)$ for certain known density function f and $\eta_1, \eta_2, \dots, \eta_n$ are identically distributed as an unknown probability measure G , as discussed in, for example, in the works of Neal [11] and McAuliffe et al. [8]. There are quite a few real-world applications that can be set in this framework. Examples include such general insurance practice where the observations X_i are the claims of an insured at n periods distributed identically as $f(x | \eta_i)$ with η_i reflecting the overall profile of the economic, political, and other situations at period i , such as presence of extreme weather, earthquake, war, and epidemics. In contrast to the most extensively discussed model under which $X_1, X_2, \dots, X_n | \eta \sim \text{iid } f(x | \eta), \eta \sim G$, the parameters η_i are of evolutionary feature from one period to another [20]. In many real scenarios, the values of evolutionary $\eta_i, i = 1, 2, \dots, n$ exhibit clustering feature and the number and the profiles of the clusters are unknown, so that it is appropriate to use stick-breaking distributions to model $\eta_i, i = 1, 2, \dots, n$.

Pitman-Yor processes [14, 21] are a class of stick-breaking distributions covering Dirichlet processes as special elements. Let $\alpha \in [0, 1)$ (referred to as the discount parameter), let $\theta > -\alpha$ (referred to as the concentration parameter), and let $(V_n)_{n \geq 1}$ be a sequence of mutually independent random variables with $V_n \sim \text{Beta}(1 - \alpha, \theta + n\alpha)$. Set $\tilde{V}_1 = V_1$ and $\tilde{V}_n = V_n \prod_{i=1}^{n-1} (1 - V_i), n \geq 2$, so that $\sum_{n=1}^{\infty} \tilde{V}_n = 1$. Let $(\xi_n)_{n \geq 1}$ be a sequence of i.i.d. random variables following a common probability measure G_0 , taking values in a measurable space, $(\mathcal{X}, \mathcal{B})$, for example, and independent also of $(V_n)_{n \geq 1}$. Let $\delta(\xi, \cdot)$ stand for the Dirac measure degenerated at point ξ . Then the stick-breaking distribution $G(\cdot) = \sum_{n=1}^{\infty} \tilde{V}_n \delta(\xi_n, \cdot)$ on $(\mathcal{X}, \mathcal{B})$ is referred to as a Poisson-Dirichlet process (PDP) with parameters α, θ and base distribution G_0 , writing $G \sim \text{PD}(G_0; \alpha, \theta)$. Note that parameters α and θ determine the characteristic of power law of a PDP and the case $\alpha = 0$ corresponds to a Dirichlet process.

Let $\eta = \{\eta_1, \eta_2, \dots, \eta_n\}$ be a sequence of i.i.d. observations drawn from G so as to reflect the clustering feature of distributions. Write K for the number of clusters of η , $\eta_1^*, \eta_2^*, \dots, \eta_K^*$ for the distinct observations among $\{\eta_1, \eta_2, \dots, \eta_n\}$ and m_k is the number of elements in cluster k . Then, by Pitman [21], given $\{\eta_1, \eta_2, \dots, \eta_n\}$, the probability distribution of a new future η_{n+1} is

$$\Pr(\eta_{n+1} \in A | \eta_1, \eta_2, \dots, \eta_n) = \frac{K\alpha + \theta}{n + \theta} G_0(A) + \sum_{k=1}^K \frac{m_k - \alpha}{n + \theta} \delta(\eta_k^*, A), \quad (1)$$

that is, taking values η_k^* with probability $m_k - \alpha/n + \theta$ and a new one from G_0 with probability $K\alpha + \theta/n + \theta$. This is the generalized Pólya urn scheme to produce $\eta_i, i = 1, 2, \dots$

A PDP mixture model is defined by

$$\begin{aligned} G &\sim \text{PD}(G_0; \alpha, \theta), \\ \eta_1, \dots, \eta_n | G &\stackrel{\text{i.i.d.}}{\sim} G, \quad i = 1, 2, \dots, n, \\ X_i | \eta_i &\sim f(\cdot | \eta_i), \quad i = 1, 2, \dots, n, \end{aligned} \quad (2)$$

where $f(\cdot | \eta)$ is a known density function characterized by parameter η , $\eta = (\eta_1, \eta_2, \dots, \eta_n)$ consists of the latent parameters, and G_0 is absolutely continuous with respect to certain σ -finite measure with an entirely unknown density function g_0 , as depicted in Figure 1. Thanks to the clustering feature of elements of η from PD processes, under this model, the observations $X_1, X_2, \dots, X_n$ are randomly divided into a random number of clusters so that all the elements of every cluster share a common parameter picked from $\eta_1^*, \eta_2^*, \dots, \eta_K^*$ and a new observation X_{n+1} either is generated from a distribution with a new parameter, meaning that X_{n+1} comes from a new cluster, or falls into one of the previously existing clusters with parameter from $\eta_1^*, \eta_2^*, \dots, \eta_K^*$. Note that $\alpha = 0$ corresponds to the special case of DP mixture models, on which some recent efforts can be found in, for example, the work of Neal [11] who discussed the methods of sampling from the posterior distribution of a DP mixture model and the work of McAuliffe et al. [8] who studied the estimation of the base distribution G_0 in DP mixture model with a single group of data $\{X_1, X_2, \dots, X_n\}$.

In this paper, however, we consider the estimation of the parameters in the distribution G when data is organized in independent and identically distributed groups. To be specific, we consider the following PDP mixture model: the data are divided into I groups and the parameters $\eta_{ij}, j = 1, 2, \dots, n_i$, of each group form an i.i.d. sample from G . A graphical representation of a PDP mixture model is shown in Figure 2, and its hierarchical structure in terms of conditional distributions is clearly expressed by

$$\begin{aligned} G_1, \dots, G_I &\stackrel{\text{i.i.d.}}{\sim} \text{PD}(G_0; \alpha, \theta), \\ \eta_{i1}, \dots, \eta_{in_i} | G_i &\stackrel{\text{i.i.d.}}{\sim} G_i, \quad i = 1, 2, \dots, I, \\ X_{ij} | \eta_{ij} &\sim f(\cdot | \eta_{ij}), \quad j = 1, 2, \dots, n_i, \quad i = 1, 2, \dots, I. \end{aligned} \quad (3)$$

Under this setting, the individuals from different groups are mutually independent and the random probabilities $G_i, i = 1, 2, \dots, I$, are independent and identically distributed as a PD($G_0; \alpha, \theta$) with unknown α, θ , and G_0 to be estimated.

On the one hand, the structure defined above arises quite frequently in real-world applications. In the context of general insurance, it is a typical structure of data extensively investigated by the so-called credibility theory [20]. On the other hand, it has been commonly recognized that, for the majority of EB problems, it is impossible to make efficient estimate of the unknown elements in the prior distributions if one has only a single group of conditionally independent

FIGURE 1: PDP mixture models with a single-group data.

FIGURE 2: PDP mixture models with multigroup data.

and identically distributed observations given the random parameters following the priors. The only exception arises when the prior can be characterized by a so-called stick-breaking distribution such as Dirichlet processes (see [4, 22]). Even for stick-breaking priors, existing research efforts have exhibited the possibly low efficiency in estimating the parameters in prior distributions with a single group of data; see, for example, the works of Zehnwirth [22], Yang and Wu [4], and Zhang et al. [23].

For $i = 1, 2, \dots, I$, denote by K_i the number of distinct values among $\eta_i = \{\eta_{i1}, \eta_{i2}, \dots, \eta_{in_i}\}$, $\{\eta_{i1}^*, \eta_{i2}^*, \dots, \eta_{iK_i}^*\}$ those distinct values, and $m_{ij}, j = 1, 2, \dots, K_i$, the number of elements in cluster j of group i . Then, similar to Theorem 2.5 of Korwar and Hollander [24], given K_i , the distinct values $\eta_{i1}^*, \eta_{i2}^*, \dots, \eta_{iK_i}^* \sim \text{i.i.d.} G_0$ and are also independent of $m_{i1}, m_{i2}, \dots, m_{iK_i}$, so that we can retrieve information as of the base distribution G_0 from the information delivered by the i.i.d. $\{\eta_{i1}^*, \eta_{i2}^*, \dots, \eta_{iK_i}^*, i = 1, 2, \dots, I\}$ and of the parameters θ and α from $\{(K_i, m_{i1}, \dots, m_{iK_i}), i = 1, 2, \dots, I\}$.

Generally, denote $\mathbf{m}_i = (m_{i1}, \dots, m_{iK_i}), i = 1, 2, \dots, I$ and rearrange $X_{i1}, \dots, X_{in_i}$ into K_i groups $(X_{i1}^1, \dots, X_{i1}^{m_{i1}}), \dots, (X_{iK_i}^1, \dots, X_{iK_i}^{m_{iK_i}})$ so that, for every $k = 1, 2, \dots, K_i$, the observations in $\{X_{jk}^i, j = 1, 2, \dots, m_k^i\}$ share the common

random parameter η_{ik}^* of cluster k in group i . Then, by means of marginalization, the likelihood under framework (3) with data $\mathbf{X} = \{X_{ij}, j = 1, 2, \dots, n_i; i = 1, 2, \dots, I\}$ can be expressed as

$$\begin{aligned} L(g_0, \alpha, \theta) &= \prod_{i=1}^I E_{\eta_i} \left(\prod_{j=1}^{n_i} f(X_{ij} | \eta_{ij}) \right) \\ &= \prod_{i=1}^I E_{\eta_i} \left(\prod_{j=1}^{m_{i1}} f(X_{j1}^i | \eta_{i1}^*) \dots \prod_{j=1}^{m_{iK_i}} f(X_{jK_i}^i | \eta_{iK_i}^*) \right) \\ &= \prod_{i=1}^I E_{K_i, \mathbf{m}_i} \left(\int \prod_{j=1}^{m_{i1}} f(X_{j1}^i | u) g_0(u) du \dots \int \prod_{j=1}^{m_{iK_i}} f(X_{jK_i}^i | u) g_0(u) du \right), \end{aligned} \quad (4)$$

where the expectation E_{η_i} is taken with respect to marginal distribution of η_i with X_{ij} taken fixed, $E_{K_i, \mathbf{m}_i}$ is understood in the same way, and g_0 is the unknown density of G_0 . Imaginably, the likelihood function (4) of (g_0, α, θ) is generally expressed as a product of I summations, of which every summand is, in turn, a product of the form expressed inside the parentheses in (4). So, it is too intricate to be maximized; even numerically and certain indirect alternatives to the maximum likelihood estimates of (g_0, α, θ) need to be developed, of which the following is constructed with the help of the ideas from empirical Bayes.

3. Estimation of Poisson-Dirichlet Mixture Models

This section concentrates on developing reasonable estimates of the parameters g_0, α , and θ and elucidates in detail algorithms deriving those estimates. We proceed in two subsections with one for the idea and the other for algorithms to realize the ideas.

3.1. Constructing Estimates. Should one have observed $\eta = \{\eta_{ij}, j = 1, 2, \dots, n_i, i = 1, 2, \dots, I\}$ or, equivalently, $\{K_1, K_2, \dots, K_I\}$ together with $\{(\eta_{ij}^*, m_{ij}), j = 1, 2, \dots, K_i, i = 1, 2, \dots, I\}$. By the fact that $\eta_{ij}^* \sim \text{i.i.d. } g_0$ for all $j = 1, 2, \dots, K_i, i = 1, 2, \dots, I$, a natural choice is to estimate g_0 by certain maturely developed density estimators. We use kernel method to result in

$$\tilde{g}_0(\cdot; \boldsymbol{\eta}) = \frac{1}{\sum K_i} \sum_{i=1}^I \sum_{j=1}^{K_i} \omega_h(\cdot, \eta_{ij}^*), \quad (5)$$

where $\omega_h(x, \eta_{ij}^*) = (1/h)\omega(x - \eta_{ij}^*/h)$ is defined by a certain kernel function $\omega(\cdot)$ together with a bandwidth $h > 0$. For α and θ , similar to Carlton [18] who studied the estimation of PDP priors and Favaro et al. [19] who established the PDP model so as to predict species for the problem of species searching, we can use the maximum likelihood estimate derived in terms of the distribution of $\{(K_i, m_{i1}, \dots, m_{iK_i}), i = 1, 2, \dots, I\}$; that is,

$$(\tilde{\alpha}(\boldsymbol{\eta}), \tilde{\theta}(\boldsymbol{\eta})) = \underset{\alpha, \theta}{\text{Argmax}} \prod_{i=1}^I \left\{ \frac{1}{(\theta + 1)^{n_i-1}} \cdot \prod_{l=1}^{K_i-1} (\theta + l\alpha) \cdot \prod_{j=2}^{n_i} [(1 - \alpha)_{j-1}]^{a_j^i} \right\}, \quad (6)$$

where $(1 + x)_k = (1 + x)(2 + x) \dots (k - 1 + x)$ and a_j^i is the cardinality of $\{m_{ik}: k = 1, 2, \dots, K_i \text{ such that } m_{ik} = j\}$, that is, each of the clusters of which has j items.

Now that the latent variables $\{\eta_{ij}, j = 1, 2, \dots, n_i, i = 1, 2, \dots, I\}$ are unobservable, we construct unbiased pseudoestimates:

$$\begin{cases} \tilde{g}_0(\cdot; g_0, \alpha, \theta) = E(\tilde{g}_0(\cdot; \boldsymbol{\eta}) | \mathbf{X}) = E\left(\frac{1}{\sum_{i=1}^I K_i} \sum_{i=1}^I \sum_{j=1}^{K_i} \omega_h(\cdot, \eta_{ij}^*) | \mathbf{X}\right), \\ (\tilde{\alpha}, \tilde{\theta})(g_0, \alpha, \theta) = E[(\tilde{\alpha}(\boldsymbol{\eta}), \tilde{\theta}(\boldsymbol{\eta})) | \mathbf{X}], \end{cases} \quad (7)$$

where the arguments (g_0, α, θ) indicate that the “estimates” nonetheless depend on the unknown parameters g_0, α , and θ . Note that these pseudoestimates can be computed by means of the posterior distribution for η given the observations $\mathbf{X}$.

By these pseudoestimates, an algorithm can be motivated, which, starting from certain appropriately selected initial values of g_0, α , and θ , iteratively computes for $r = 0, 1, \dots$

$$\begin{cases} \hat{g}_0^{(r+1)}(\cdot) = E_{(\hat{g}_0^{(r)}, \hat{\alpha}^{(r)}, \hat{\theta}^{(r)})} [\tilde{g}_0(\cdot; \boldsymbol{\eta}) | \mathbf{X}], \\ (\hat{\alpha}^{(r+1)}, \hat{\theta}^{(r+1)}) = E_{(\hat{g}_0^{(r)}, \hat{\alpha}^{(r)}, \hat{\theta}^{(r)})} [(\tilde{\alpha}(\boldsymbol{\eta}), \tilde{\theta}(\boldsymbol{\eta})) | \mathbf{X}], \end{cases} \quad (8)$$

with the symbol $E_{(\hat{g}_0^{(r)}, \hat{\alpha}^{(r)}, \hat{\theta}^{(r)})}$ indicating that the posterior expectations of $(\tilde{g}_0(\cdot; \boldsymbol{\eta}), \tilde{\alpha}(\boldsymbol{\eta}), \tilde{\theta}(\boldsymbol{\eta}))$ given $\mathbf{X}$ are computed with the unknown parameters (g_0, α, θ) replaced by $(\hat{g}_0^{(r)}, \hat{\alpha}^{(r)}, \hat{\theta}^{(r)})$. Consequently, estimate g_0, α , and θ by the limit

$$(\hat{g}_0, \hat{\alpha}, \hat{\theta}) = \lim_{r \rightarrow \infty} (\hat{g}_0^{(r)}, \hat{\alpha}^{(r)}, \hat{\theta}^{(r)}). \quad (9)$$

In every iteration to compute $(\hat{g}_0^{(r+1)}(\cdot), \hat{\alpha}^{(r+1)}, \hat{\theta}^{(r+1)})$ from $(\hat{g}_0^{(r)}, \hat{\alpha}^{(r)}, \hat{\theta}^{(r)})$, numerical solutions are necessary because it is not an easy task to work out closed forms of the posterior expectations in the equations in (8). A natural way is to employ an MCMC technique that will be described in detail later on.

To sum up, the estimates are approximated with the following workable procedure.

3.2. Algorithms. This subsection details Algorithm 1 in the selection of initial values, the steps to perform the MCMC sampling, and the computation of the estimation proposed.

3.2.1. Initial Values. To fix an initial estimate $\hat{g}_0^{(0)}$, compute first the maximum likelihood estimate $\hat{\eta}_{ij}$ by $f(X_{ij} | \hat{\eta}_{ij}) = \max_u f(X_{ij} | u)$ and then compute a kernel estimate of g_0 by

$$\hat{g}_0^{(0)}(\cdot) = \frac{1}{\sum_{i=1}^I n_i} \sum_{i=1}^I \sum_{j=1}^{n_i} \omega_h(\cdot, \hat{\eta}_{ij}). \quad (10)$$

Similar to the formula in (5), in many important cases, f is a location translation of some standard density with mode

zero (e.g., a standard normal distribution) such that $\hat{\eta}_{ij} = X_{ij}$. In these cases, the initial values of g_0 are given by

$$\hat{g}_0^{(0)}(\cdot) = \frac{1}{\sum_{i=1}^I n_i} \sum_{i=1}^I \sum_{j=1}^{n_i} \omega_h(\cdot, X_{ij}). \quad (11)$$

For parameters α, θ , the initial value of α could be 0.1 or 0.5 and θ could be 1 or 5 in case there is no further information on them. The simulation, which will be reported later on, showed that the initial values of (α, θ) had only quite minor effects on the convergence of the algorithm.

3.2.2. The Gibbs Sampling. We next state the full conditional distribution used in the MCMC procedure, that is, the conditional distribution of a single latent variable η_{ij} given the data $\mathbf{X}_i = \{X_{i1}, X_{i2}, \dots, X_{in_i}\}$, $i = 1, 2, \dots, I$, and the latent variables $(\eta_i \setminus \eta_{ij}) \cup_{k \neq i} \eta_k$. Denote by $\eta_i^{-j} = \eta_i \setminus \eta_{ij}$ the subvector of η_i by taking away η_{ij} and K_i^{-j} correspondingly the numbers of the distinct values in η_i^{-j} , $j = 1, 2, \dots, n_i$, $i = 1, 2, \dots, I$. With an abuse of notation, let $\eta_{i1}^*, \eta_{i2}^*, \dots, \eta_{iK_i^{-j}}^*$ be the K_i^{-j} distinct values in η_i^{-j} and $m_{il} = \#\{r: r \in \{1, 2, \dots, n_i\} \setminus \{j\} \text{ and } \eta_{ir} = \eta_{il}^*\}$ the number of such elements in η_i^{-j} which take the value η_{il}^* .

Thanks to the mutual independence among the random vectors $(G_i, (X_{i1}, \eta_{i1}), \dots, (X_{in_i}, \eta_{in_i}))$, $i = 1, 2, \dots, I$, it suffices to discuss the single-group situation, as stated in Theorem 3.1.

Theorem 1. For each $j = 1, 2, \dots, n_i$, given η_i^{-j} and $\mathbf{X}_i$, the conditional distribution of η_{ij} is

$$(\eta_{ij} | \boldsymbol{\eta}_i^{-j}, \mathbf{X}_i; g_0, \alpha, \theta) \sim q_0 G_\eta(\cdot | X_{ij}) + \sum_{l=1}^{K_i^{-j}} q_l \delta(\eta_{il}^*, \cdot), \quad (12)$$

where $G_\eta(\cdot | X_{ij})$ is the posterior distribution of η_{ij} derived from a single observation X_{ij} with density $g_\eta(u | X_{ij}) = f(X_{ij} | u) g_0(u) / \int f(X_{ij} | u) g_0(u) du$, $\{q_l\}_{l=0}^{K_i^{-j}}$ are such that $q_0 = b(\theta + \alpha K_i^{-j}) \int f(X_{ij} | u) g_0(u) du$, $q_l = b(m_{il} - \alpha) f(X_{ij} | \eta_{il}^*)$, $l = 1, \dots, K_i^{-j}$, and the normalization quantity b is determined by the equality $\sum_{l=0}^{K_i^{-j}} q_l = 1$.

Proof. By the generalized Pólya urn scheme (1), the posterior distribution under PD($g_0; \alpha, \theta$) is

$$\Pr(\eta_{ij} \in \cdot | \boldsymbol{\eta}_i^{-j}) = \frac{\theta + \alpha K_i^{-j}}{\theta + n_i - 1} g_0(\cdot) + \frac{1}{\theta + n_i - 1} \sum_{l=1}^{K_i^{-j}} (m_{il} - \alpha) \delta(\eta_{il}^*, \cdot). \quad (13)$$

Therefore, the conditional distribution of η_{ij} given $(\eta_i^{-j}, \mathbf{X}_i)$ can be computed by

- (1) The procedure begins with certain appropriately set initial values $(\hat{g}_0^{(0)}(\cdot), \hat{\alpha}^{(0)}, \hat{\theta}^{(0)})$ at round 0.
- (2) Having computed $(\hat{g}_0^{(r)}(\cdot), \hat{\alpha}^{(r)}, \hat{\theta}^{(r)})$ at round r , the $(r+1)$ -th round proceeds in the following steps:
 - (a) Start from certain initial values η^0 of the latent variable η and produce a sequence of $\eta^b, b = 1, 2, \dots, B$ by certain MCMC method for a sufficiently large positive integer B .
 - (b) Compute $\tilde{g}_0(\cdot; \eta^b)$ by (5) and $(\tilde{\alpha}(\eta^b), \tilde{\theta}(\eta^b))$ by (6) for every $b = 1, 2, \dots, B$.
 - (c) End by assigning a new set of values $(\hat{g}_0^{(r+1)}(\cdot), \hat{\alpha}^{(r+1)}, \hat{\theta}^{(r+1)}) \approx (1/B) \sum_{b=1}^B (\tilde{g}_0(\cdot; \eta^b), \tilde{\alpha}(\eta^b), \tilde{\theta}(\eta^b))$ to the parameters $(g_0(\cdot), \alpha, \theta)$.
- (3) Repeat the algorithm until the sequence $(\hat{g}_0^{(r)}(\cdot), \hat{\alpha}^{(r)}, \hat{\theta}^{(r)})$ is stable.

ALGORITHM 1: Estimation procedure.

$$\begin{aligned}
 dG(\eta_{ij}|\mathbf{\eta}_i^{-j}, \mathbf{X}_i) &= \frac{f(X_{ij}|\eta_{ij})dF(\eta_{ij}|\mathbf{\eta}_i^{-j})}{\int f(X_{ij}|\eta_{ij})dF(\eta_{ij}|\mathbf{\eta}_i^{-j})} \\
 &= \frac{(\theta + \alpha K_i^{-j})f(X_{ij}|\eta_{ij})g_0(\eta_{ij})d\eta_{ij} + \sum_{l=1}^{K_i^{-j}} (m_{il} - \alpha)f(X_{ij}|\eta_{il}^*)\delta(\eta_{il}^*, \eta_{ij})}{(\theta + \alpha K_i^{-j}) \int f(X_{ij}|\eta_{ij})g_0(d\eta_{ij}) + \sum_{l=1}^{K_i^{-j}} (m_{il} - \alpha)f(X_{ij}|\eta_{il}^*)}
 \end{aligned} \tag{14}$$

Set

$$\Lambda(X_{ij}) = \frac{1}{(\theta + \alpha K_i^{-j}) \int f(X_{ij}|\eta_{ij})g_0(d\eta_{ij}) + \sum_{l=1}^{K_i^{-j}} (m_{il} - \alpha)f(X_{ij}|\eta_{il}^*)}, \tag{15}$$

and rewrite the equation above as

$$\begin{aligned}
 dG(\eta_{ij}|\mathbf{\eta}_i^{-j}, \mathbf{X}_i) &= \Lambda(X_{ij})(\theta + \alpha K_i^{-j})f(X_{ij}|\eta_{ij})g_0(\eta_{ij})d\eta_{ij} \cdot \frac{f(X_{ij}|\eta_{ij})g_0(\eta_{ij})}{\int f(X_{ij}|\eta_{ij})g_0(d\eta_{ij})} \\
 &\quad + \Lambda(X_{ij}) \cdot \sum_{l=1}^{K_i^{-j}} (m_{il} - \alpha)f(X_{ij}|\eta_{il}^*)\delta(\eta_{il}^*, \eta_{ij}).
 \end{aligned} \tag{16}$$

That is, the conditional distribution of η_{ij} is

$$(\eta_{ij}|\mathbf{\eta}_i^{-j}, \mathbf{X}_i; g_0, \alpha, \theta) \sim q_0 G_\eta(\eta_{ij}|X_{ij}) + \sum_{l=1}^{K_i^{-j}} q_l \delta(\eta_{il}^*, \eta_{ij}), \tag{17}$$

where $q_0 = \Lambda(X_{ij})(\theta + \alpha K_i^{-j}) \int f(X_{ij}|\eta_{ij})g_0(\eta_{ij})d\eta_{ij}$ and $q_l = \Lambda(X_{ij})(m_{il} - \alpha)f(X_{ij}|\eta_{il}^*)$. Obviously, q_0 and q_l s satisfy $\sum_{l=0}^{K_i^{-j}} q_l = 1$. The proof is completed. $\square$

3.2.3. The Iteration Procedure. With the theorem above and the initial values $(\hat{g}_0^{(0)}, \hat{\alpha}^{(0)}, \hat{\theta}^{(0)})$, we now describe the iteration to compute $(\hat{g}_0^{(r+1)}, \hat{\alpha}^{(r+1)}, \hat{\theta}^{(r+1)})$ from $(\hat{g}_0^{(r)}, \hat{\alpha}^{(r)}, \hat{\theta}^{(r)})$ for all $r = 0, 1, 2, \dots$, which consists of two alternatively running phases: one for sampling from the posterior distribution of the latent variables η with the unknown

parameters replaced by the estimates $(\hat{g}_0^{(r)}, \hat{\alpha}^{(r)}, \hat{\theta}^{(r)})$ and the other for estimation of the parameters using the sampled latent variables η , as depicted in what follows:

(1) *Sampling.* The sampling phase is indeed a two-step procedure applied to every group $i = 1, 2, \dots, I$ separately so as to produce the posterior samples η_i , starting with an arbitrarily selected $\eta_i^{(0)} = \{\eta_{i1}^{(0)}, \eta_{i2}^{(0)}, \dots, \eta_{iK_i}^{(0)}\}$. The first step is the standard Gibbs sampling derived from the full conditionals presented in Theorem 1 and the second performs an acceleration.

As frequently done in the literature (see, e.g., [11]), introduce an auxiliary vector $\mathbf{c}(\eta_i) \triangleq \mathbf{c} = (c_1, c_2, \dots, c_{n_i})$ to stand for the cluster indices of η_{ij} s in $\{\eta_{i1}^*, \eta_{i2}^*, \dots, \eta_{iK_i}^*\}$; that is, $c_j = k \Leftrightarrow \eta_{ij} = \eta_{ik}^*$, so that $\eta_i = (\eta_{ic_1}^*, \eta_{ic_2}^*, \dots, \eta_{ic_{n_i}}^*)$. In this notation, η_i is equivalent to the combination of $\mathbf{c}$ and $\{\eta_{i1}^*, \eta_{i2}^*, \dots, \eta_{iK_i}^*\}$. With the cluster indicators in $\mathbf{c}$, in this

two-step procedure, the first step updates the values of c_j and in passing the values of η_{ij} respecting the full conditionals in Theorem 1. For the case $\sum_{l=1}^{K_i^j} q_l$ is much larger than q_0 , the Markovian chain directly derived from Theorem 1 may converge slowly because the chance to introduce new values for η_{ij} is quite small; see equation (12). Hence, an acceleration method (e.g., [11]) may be helpful, which samples another round of η_i for acceleration. This is realized by sampling η_i conditional on $\mathbf{X}_i$ and the cluster indicators $\mathbf{c}$ just obtained in Step 1 and then sampling $\mathbf{c}$ given $(\mathbf{X}_i, \eta_i)$. In other words, Step 2 is driven by the grouped conditional distribution of η_i given $(\mathbf{X}_i, \mathbf{c})$ and $\mathbf{c}$ given $(\mathbf{X}_i, \eta_i)$. Because of the particular feature that $\mathbf{c}$ is a deterministic function of η_i , Step 2 only needs to draw values from the conditional distribution of η_i given $(\mathbf{X}_i, \mathbf{c})$.

Moreover, for every $j = 1, \dots, n_i$, the integration $\int f(X_{ij}|\eta_{ij})\hat{g}_0^{(r)}(\eta_{ij})d\eta_{ij}$ involved in computing $G_{\eta}(\cdot|X_{ij})$ and $\{q_l, l = 0, 1, \dots, K_i^j\}$ is approximated by an average

$$h_{rj} = \frac{1}{d} \sum_{l=1}^d f(X_{ij}|u_l), \quad (18)$$

where d is a positive integer and $u_1, \dots, u_d \sim \text{iid} \hat{g}_0^{(r)}$. Because $u_1, \dots, u_d \sim \text{iid} \hat{g}_0^{(r)}$ depend only on the iteration indicator r , this approximation can be prepared at the beginning of each iteration.

As a final point, we would like to note that, due to the presence of the acceleration Step 2, the update of the values of η_{ij} in Step 1 essentially takes effect only on the computation of the probabilities q_l in equation (12) for every $j = 1, 2, \dots, n_i$.

The details of this two-step procedure of generating $\eta_i^{(b)}$ are presented in Algorithm 2. For each group $i = 1, 2, \dots, I$, this procedure is repeated to generate $\eta_i^{(b)} = (\eta_{i1}^{(b)}, \dots, \eta_{in_i}^{(b)})$, $b = 1, 2, \dots, B$, for a sufficiently large B .

(2) *Estimation.* In phase estimation, with $\eta^{(b)}$, $b = 1, 2, \dots, B$, in hand, the estimates of α , θ , and g_0 are obtained according to (10).

By these newly obtained estimates, repeat the above two phases for the next iteration until the sequence of estimates is stable. The above computation process on the estimation of α , θ , g_0 under the framework in (3) can be briefly arranged as in Algorithm 3.

We now conclude this section by the following remark on one of Neal's [11] algorithms.

Remark 1. Algorithm 2 is similar to, but with two minor differences, to Algorithm 8 of Neal [11] at Step 1. First, we only draw a single set of d i.i.d. values from $\hat{g}_0^{(r)}$ in each iteration, while Neal [11] suggested drawing d i.i.d. values for each $j = 1, \dots, n_i$, hence needing $n_i - 1$ times more than the sampling computation here. The algorithm here hence may cause a significant reduction in the total computation load. Second, we draw a value for η_{ic_j} from $G_{\eta}(\cdot|X_{ij})$ if a new cluster for c_j is produced, while Neal [11] simply drew a

value from $g_0^{(r)}$. According to Theorem 1, we think Algorithm 2 draws a correct value.

4. Simulation

In this section, some simulations are reported to show the performance of estimates provided above. The data structure of the PDP mixture model is according to equation (3) and the simulations were conducted under the two settings.

- (i) We considered two settings: (1) $I = 100$ with 20 individuals in every group: $n_1 = n_2 = \dots = n_{100} = 20$ and (2) $I = 1$ with 200 individuals in a single group.
- (ii) The amount of Gibbs sampling is $B = 20$ and total iteration of estimation is 200.
- (iii) The combinations of (α, θ, G_0) : (a) standard normal distribution $(0.3, 10, N(0, 1))$, (b) t distribution with 3 degrees of freedom, and (c) the distribution of $-\log \mathcal{X}_1^2$. α took values of 0.3, 0.5, 0.8, respectively. $\theta = 10$. The initial values $(\hat{\alpha}^{(0)}, \hat{\theta}^{(0)})$ were (0.5, 5) and (0.1, 1).
- (iv) The conditional density function f was normal with mean η_{ij} and standard deviation was $0.1 \cdot \text{SD}(G_0)$, where $\text{SD}(G_0)$ is the standard deviation of G_0 .

In the three base probabilities we chose, the first two are symmetric, t_3 has heavy tails, and the last has a positive skew. We tried a few distinct initial values of α , θ in all cases; the simulations showed that the estimations were insensitive to the initial values. For every combination of the values, α , θ , I , and g_0 , we produced 1000 replicates of the datasets under the setting described above. The simulation results are concisely discussed below.

- (1) The estimates of the base measures g_0 . In the simulation, g_0 was estimated more accurately under the setting $I = 100$ than under the setting $I = 1$. The results from some replicates out of the 1000 are presented in Figures 3–5 under some parameters α , θ , and g_0 , in terms of the iterations $r = 0, 5$, and 200 under $I = 1$ and $I = 200$, in which the dashed lines represent the curves of estimates and the solids are the true density curves. The figures show the two following clear features: the dashed lines are more close to the solids curves with the increase of iteration times in both data structures and apparently the performance of estimated curves under multi-group data is better than that of single-group data. The center of dashed curves with single data is a little deviation at round 200 in Figure 3, the estimated curve is more concentrated than real density with heavy tails in Figure 4, and the curves of $\hat{g}_0$ are more fluctuating and skewed in Figure 5 under the single-group data. Overall, in both the single-group and multigroup environments, our estimates are still near the true density curves. The histograms of L_1 distance between $\hat{g}_0$ and true g_0 with 1000 replicates are shown in Figure 6 under single group and multiple groups. Figure 7 delineates the total

- (1) Draw iid values $u_1, u_2, \dots, u_d$ from $\hat{g}_0^{(r)}$ and compute $\{h_{r,j}, j = 1, \dots, n_i\}$ by (13).
- (2) Initiate the algorithm with an arbitrarily selected $\eta_i^{(0)}$. Correspondingly, set $\mathbf{c} = \mathbf{c}(\eta_i^{(0)})$. Let $\eta^* = (\eta_{i1}^*, \eta_{i2}^*, \dots, \eta_{iK_i}^*, 0, \dots, 0)$ be the vector of dimension n_i in which the first K_i elements store the distinct value of η_i^0 and $\mathbf{m} = (m_{i1}, m_{i2}, \dots, m_{iK_i}, 0, \dots, 0)$ be an n_i -vector to represent the numbers of η_{ij}^* occurring in $\eta_i^{(0)}$.
- (3) Now assume that we have got $\eta_i^{(b-1)}$ and the corresponding auxiliary vectors $\mathbf{c}, \mathbf{m}$ and η^* . The sampled values $\eta_i^{(b)}$ are generated by the following two-step procedure:

Step 1. For $j = 1, 2, \dots, n_i$,

- (1) set $m_{ic_j} = m_{ic_j} - 1$, $K_i^{-j} = \begin{cases} K_i - 1 & \text{if } m_{ic_j} = 0 \\ K_i & \text{otherwise} \end{cases}$;

- (2) draw a new value for c_j from $\{1, 2, \dots, n_i\}$ according to the distribution law

$$q_c = \Pr(c_j = c | \mathbf{X}_i, \mathbf{c} \setminus c_j, \eta_{i1}^*, \dots, \eta_{iK_i}^*)$$

$$= \begin{cases} \mathcal{B}(m_{ic} - \hat{\alpha}^{(r)}) f(X_{ij} | \eta_{ic}^*) & \text{for } 1 \leq c \leq n_i, m_{ic} \neq 0 \\ b/n_i - K_i^{-j} (\hat{\theta}^{(r)} + \hat{\alpha}^{(r)} K_i^{-j}) h_{r,j} & \text{for } 1 \leq c \leq n_i, m_{ic} = 0 \end{cases}$$

and then, in case $m_{ic_j} = 0$, draw a value for η_{ic_j} from $G_{\eta}(\cdot | X_{ij})$ with g_0 replaced by $\hat{g}_0^{(r)}$, where $\mathcal{B}$ is the normalization constant subject to $\sum_{c=0}^{n_i} q_c = 1$; and then

- (3) set $m_{ic_j} = m_{ic_j} + 1$.

Step 2. For all $c \in \{c_1, c_2, \dots, c_{n_i}\}$ which are obtained by above Step 1, draw $\eta^{(b)}$ from the conditional density

$$f(\eta_{ic} | \mathbf{c}, \mathbf{X}_i) \propto \hat{g}_0^{(r)}(\eta_{ic}) \prod_{\{j: c_j=c\}} f(X_{ij} | \eta_{ic}).$$

ALGORITHM 2: Gibbs sampling under single group.

Step 1. Initialize $\hat{\alpha}^{(0)}, \hat{\theta}^{(0)}$ and $\hat{g}_0^{(0)}$;
 Step 2. At round $r + 1, r = 1, 2, \dots$, for all $i = 1, 2, \dots, I$,
 (i) arbitrarily selection initial values for $\eta_i^{(0)}$ and
 (ii) update η_i with Algorithm 2 to produce $\eta_i^{(b)}, b = 1, 2, \dots, B$;
 Step 3. Obtain $\hat{\alpha}^{(r+1)}, \hat{\theta}^{(r+1)}$ and $\hat{g}_0^{(r+1)}$ according to (10).

ALGORITHM 3: Gibbs sampling under multigroup.

FIGURE 3: $(\alpha, \theta) = (0.3, 10)$ and $g_0 = N(0, 1)$ with the upper panels for $I=1$ and the lower for $I=100$.

variation distance (TVD) between G_0 and $\hat{G}_0$. The TVD of two probability distributions P and Q is $\sup_A |P(A) - Q(A)|$, with the supremum taken over measurable sets A . Each panel plots the TVD between $\hat{G}_0$ and G_0 as the rounds progress. As can be seen, the value of TVD gradually reaches a steady

value after about 15 iterations with the bases G_0 of $N(0, 1)$ and t_3 , and the other needs to go through about 25 iterations.

- (2) The estimates of (α, θ) with the 1000 replicates are summarized in Table 1 in terms of the mean and standard deviation and their histograms are

FIGURE 4: $(\alpha, \theta) = (0.5, 10)$ and $g_0 = t_3$ with the upper panels for $I = 1$ and the lower for $I = 100$.

FIGURE 5: $(\alpha, \theta) = (0.8, 10)$ and $g_0 = -\log \chi^2_1$ with the upper panels for $I = 1$ and the lower for $I = 100$.

presented in Figure 8. The comparisons of estimates are listed with means, SD, and 95% credible intervals and the estimates of multigroup structure are better than those of the single-group structure no matter what kind of bases in three parametric combinations. It is more stable on the estimations of (α, θ) in the standard deviations criterion when $I = 100$.

5. Applications to Real Data

In this section, the PDP mixture model is employed to analyze the three following well-known real-world datasets.

The first is the galaxies data [8, 25] consisting of velocities of 82 galaxies from six well-separated conic sections in Corona Borealis region. The model with $I = 1$ was applied and the algorithm provided in Section 3 was run 200 rounds, giving rise to the estimate $(\hat{\alpha}, \hat{\theta}) = (0.429, 3.988)$ and $\hat{g}_0$ that is depicted in Figure 9, simply showing that the base distribution has multimodality corresponding to identifying different clumped galaxies.

The second dataset consists of records of stamp thickness [26] in millimeters of 485 unwatermarked Hidalgo stamps dating from 1872 to 1874. Figure 10 presents estimates $\hat{g}_0$ of stamp thickness at round 200 of the procedure mentioned above on the stamp data, which exhibits the close bimodality of the density and one is more steep than another. The curve of $\hat{g}_0$ reflects the high probability around the bimodality of η_i^*s corresponding to the number of types of paper used. The stamp data depicts the concentration of shared variables on the modes. After 200 rounds' iterations, we obtain the estimates $(\hat{\alpha}, \hat{\theta}) = (0.425, 6.810)$.

The last is on Danish fire losses [27] which are heterogeneous as reflected in multimodality, skewness, and heavy tail distributions, collected by Copenhagen Reinsurance and consist of 2492 records over the period of 1980–1990. The losses are in millions of Danish Krone and not adjusted for inflation over time. The data is divided into multiple groups based on time and the observations between groups are independent as the conventional assumption in the industry of nonlife insurance. The estimated density $\hat{g}_0$ is

FIGURE 6: Histograms of L1 distance with different bases and groups. The left plots are for $I=1$ and the bases from top to bottom are $N(0,1)$, t_3 and $-\log \mathcal{X}_1^2$.

FIGURE 7: Progression of $\text{TVD}(G_0, \hat{G}_0)$ with the base G_0 s of $N(0,1)$, t_3 and $-\log \mathcal{X}_1^2$.

FIGURE 8: Histograms of $\hat{\alpha}$ and $\hat{\theta}$ with the plots in rows 1 and 3 for $I=1$ and the others for $I=100$. The bases from left to right are $N(0, 1)$, t_3 and $-\log \mathcal{X}_1^2$.

TABLE 1: Performance of estimates for (α, θ) .

	Bases (α, θ)	$N(0, 1)$ (0.3, 10)	t_3 (0.5, 10)	$-\log \mathcal{X}_1^2$ (0.8, 10)
$I=1$	$(\hat{\alpha}, \hat{\theta})$	(0.355, 10.057)	(0.533, 9.585)	(0.793, 8.605)
	S.D.	(0.076, 2.026)	(0.058, 2.264)	(0.020, 2.444)
	C.I. ($\hat{\alpha}$)	(0.231, 0.512)	(0.426, 0.645)	(0.755, 0.831)
	C.I. ($\hat{\theta}$)	(5.716, 17.408)	(5.377, 13.992)	(3.609, 12.970)
$I=100$	$(\hat{\alpha}, \hat{\theta})$	(0.283, 9.978)	(0.521, 9.662)	(0.811, 9.547)
	S.D.	(0.017, 0.430)	(0.021, 0.752)	(0.010, 1.351)
	C.I. ($\hat{\alpha}$)	(0.225, 0.338)	(0.475, 0.563)	(0.784, 0.835)
	C.I. ($\hat{\theta}$)	(8.942, 10.975)	(8.256, 11.226)	(6.597, 12.014)

FIGURE 9: Galaxy $\hat{g}_0$: km/s (thousands). FIGURE 10: Thickness of stamp $\hat{g}_0$: mm (hundreds). FIGURE 11: Danish fire loss data $\hat{g}_0$: DKK (millions).

clearly described with two intervals of abscissa in Figure 11 at round 200 under the Poisson-Dirichlet process mixture models, having one concentrated modality and heavy tail. We get $(\hat{\alpha}, \hat{\theta}) = (0.481, 0.793)$.

6. Conclusions

In this article, we have shown the posterior distribution in the PDP mixture models and the estimates cannot be directly obtained because of unobserved latent variables η_{ij} s. The unbiased “pseudoeestimates” are constructed and the corresponding algorithm is depicted in detail based on nonparametric empirical Bayes. Some simulations are presented with both data structures, single-group and multi-group, on the pseudoeestimates of base distribution G_0 , discount parameter α , and concentration parameter θ . By the numerical simulation, the performances of estimations for parameters under multigroup data are better than those

of the single-group data with the same total size of individuals. Also, we applied the model to the well-known real-world datasets.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The research was sponsored by Shuguang Program supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission in China

(Project 18SG53) and the National Natural Science Foundation of China (Project 9195210212032016).

References

- [1] W. Seidel, "A note on estimating parameters of partial prior information," *Statistical Papers*, vol. 34, pp. 363–368, 1993.
- [2] E. George and D. P. Foster, "Calibration and empirical Bayes variable selection," *Biometrika*, vol. 87, no. 4, pp. 731–747, 2000.
- [3] M. Clyde and E. I. George, "Flexible empirical Bayes estimation for wavelets," *Journal of the Royal Statistical Society: Series B*, vol. 62, no. 4, pp. 681–698, 2000.
- [4] L. Yang and X. Wu, "Estimation of dirichlet process priors with monotone missing data," *Journal of Nonparametric Statistics*, vol. 25, no. 4, pp. 787–807, 2013.
- [5] Y. Bi and R. V. Davuluri, "NPEBseq: nonparametric empirical Bayesian-based procedure for differential expression analysis of RNA-seq data," *BMC Bioinformatics*, vol. 14, no. 1, p. 262, 2013.
- [6] X. Zhou, L. Yang, and X. Wu, "On several estimates to the precision parameter of Dirichlet process prior," *Communications in Statistics-Simulation and Computation*, vol. 46, no. 4, pp. 3187–3200, 2017.
- [7] J. S. Liu, "Nonparametric hierarchical Bayes via sequential imputation," *Annals of Statistics*, vol. 24, pp. 911–930, 1996.
- [8] J. D. McAuliffe, D. M. Blei, and M. I. Jordan, "Nonparametric empirical Bayes for the Dirichlet process mixture model," *Statistics and Computing*, vol. 16, pp. 5–14, 2006.
- [9] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Annals of Statistics*, vol. 1, pp. 209–230, 1973.
- [10] S. N. MacEachern and P. Müller, "Estimating mixture of Dirichlet process models," *Journal of Computational and Graphical Statistics*, vol. 7, pp. 223–238, 1998.
- [11] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 9, pp. 249–265, 2000.
- [12] H. Ishwaran and L. F. James, "Approximate dirichlet process computing in finite normal mixtures," *Journal of Computational and Graphical Statistics*, vol. 11, no. 3, pp. 508–532, 2002.
- [13] F. A. Quintana, W. O. Johnson, L. E. Waetjen, and E. B. Gold, "Bayesian nonparametric longitudinal data analysis," *Journal of the American Statistical Association*, vol. 515, pp. 1168–1181, 2016.
- [14] J. Pitman and M. Yor, "The Poisson-Dirichlet distribution derived from a stable subordinator," *Annals of Probability*, vol. 25, pp. 855–900, 1997.
- [15] D. A. Berry and R. Christensen, "Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes," *Annals of Statistics*, vol. 7, pp. 558–568, 1979.
- [16] A. G. Mostofi and M. Kharrati-Kopaei, "Bayesian nonparametric inference for unimodal skew-symmetric distributions," *Statistical Papers*, vol. 53, no. 4, pp. 821–832, 2012.
- [17] J. Hu, J. P. Reiter, and Q. Wang, "Dirichlet process mixture model for modeling and generating synthetic versions of nested categorical data," *Bayesian Analysis*, vol. 13, no. 1, pp. 183–200, 2017.
- [18] M. A. Carlton, "Applications of the poisson-dirichlet distribution," Dissertation, University of California, Los Angeles, CA, USA, 1999.
- [19] S. Favaro, A. Lijoi, R. H. Mena, and I. Prünster, "Bayesian non-parametric inference for species variety with a two-parameter Poisson-Dirichlet process prior," *Journal of the Royal Statistical Society: Series B*, vol. 71, no. 5, pp. 993–1008, 2009.
- [20] H. Buhlmann and A. Gisler, *A Course in Credibility Theory and its Applications*, Springer Science & Business Media, Berlin, Germany, 2006.
- [21] J. Pitman, "Exchangeable and partially exchangeable random partitions," *Probability Theory and Related Fields*, vol. 102, pp. 145–158, 1995.
- [22] B. Zehnwirth, "Credibility and the dirichlet process," *Scandinavian Actuarial Journal*, vol. 1, pp. 13–23, 1979.
- [23] J. Zhang, L. Yang, and X. Wu, "Polya tree priors and their estimation with multi-group data," *Statistical Papers*, 2016.
- [24] R. M. Korwar and M. Hollander, "Contributions to the theory of Dirichlet processes," *Annals of Probability*, vol. 1, no. 4, pp. 705–711, 1973.
- [25] K. Roeder, "Density estimation with confidence sets exemplified by superclusters and voids in the galaxies," *Journal of the American Statistical Association*, vol. 102, pp. 145–158, 1990.
- [26] I. G. Wilson, "Add a new dimension to your philately," *The American Philatelist*, vol. 97, pp. 342–349, 1983.
- [27] T. Miljkovic and B. Grün, "Modeling loss data using mixtures of distributions," *Insurance: Mathematics and Economics*, vol. 70, pp. 387–396, 2016.