Truncated Inverse-Lévy Measure Representation of the Beta Process

Junyi Zhang[1,2] **Angelos Dassios**[3] **Chong Zhong**[2] **Qiufei Yao**[1]
[1]Bocconi University
[2]The Hong Kong Polytechnic University
[3]London School of Economics

Abstract

The beta process is a widely used nonparametric prior in Bayesian machine learning. While various inference schemes have been developed for the beta process and related models, the current state-of-the-art method relies heavily on the stick-breaking representation with decreasing atom weights, which is available only for a special hyperparameter. In this paper, we introduce the truncated inverse-Lévy measure representation (TILe-Rep) that extends the decreasing atom weights representation of the beta process to general hyperparameters. The TILe-Rep fills the gap between the two stick-breaking representations in Teh et al. (2007) and Paisley et al. (2010). Moreover, it has a lower truncation error compared to other sequential representations of the beta process and potentially leads to the posterior consistency property of the Bayesian factor models. We demonstrate the usage of the TILe-Rep in the celebrated beta process factor analysis model and beta process sparse factor model.

1 INTRODUCTION

The beta process is a widely used nonparametric prior. It was introduced by Hjort (1990) to study the hazard rate in the life history data models. Later on, Kim (1999) considered the usage of the beta process in the multiplicative counting process model. In recent years, the beta process became popular in Bayesian machine learning research. The popularity is attributed to Thibaux and Jordan (2007), who showed that the beta process is the de Finetti mixing distribution underlying the Indian buffet process (IBP, Ghahramani and Griffiths 2005). Reminiscent of the stick-breaking construction of the Dirichlet process (Sethuraman, 1994), Paisley et al. (2010) developed a stick-breaking representation for the beta process. Paisley et al. (2012) showed that the stick-breaking representation could be obtained from the characterisation of the beta process as a Poisson process. Teh et al. (2007) provided a different stick-breaking representation for the beta process based on a decreasing sequence of atom weights. Teh and Görür (2009) introduced the stable-beta process, a generalisation of the beta process, and provided two construction methods using the size-biased representation and the inverse-Lévy measure method, respectively. Broderick et al. (2012) derived a stick-breaking representation for the stable-beta process. A recent review of the beta process and its properties can be found in Phadia (2016).

Due to the infinite-dimensional support, an explicit representation of the beta process is not available in practice, as we cannot simulate or store infinite random variables. To facilitate the posterior inference of the beta process, the existing literature has proposed various methods. Ghahramani and Griffiths (2005) truncated the support of the number of new features generated by the marginal distribution of the beta process. Teh et al. (2007) developed a slice sampler for the exact sampling from the stick-breaking representation of the beta process. Doshi-Velez et al. (2009) proposed a variational inference scheme for the stick-breaking representation of the beta process proposed by Teh et al. (2007). Paisley et al. (2010) derived an inference procedure that uses the Monte Carlo integration to reduce the number of parameters to be inferred. Paisley et al. (2011) designed a variational inference scheme for the stick-breaking representation of the beta process proposed by Paisley et al. (2010). In addition, Teh and Görür (2009) used the results of Kim (1999) to derive the marginal distribution of the stable-beta process. Their result is known as the stable-beta Indian buffet process.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

Alongside these methods, it is also possible to consider the truncated inverse-Lévy measure representation (TILe-Rep) of the beta process. In fact, both Teh and Görür (2009) and Al Labadi and Zarepour (2018) have discussed this approach, while the approximation error and posterior inference scheme are yet to be developed. The main barrier is the intractable tail distribution of the Lévy measure associated with the beta process. However, since the tail distribution can be expressed in terms of the hypergeometric function, it is still possible to compute it efficiently. See, for example, Luke (1969). On the other hand, there are some appealing properties for us to use the TILe-Rep: (i) It enables us to focus on the atoms of the beta process with the highest weights, which are more likely to lead to active features in the observations; (ii) The approximation error of both the prior and posterior admit explicit decompositions and bounds; (iii) It provides a unified inference scheme which is applicable to the generalisations of the beta process.

In this paper, we construct the TILe-Rep of the beta process. It fills the gap between the two stick-breaking representations in Teh et al. (2007) and Paisley et al. (2010). We provide a decomposition in distribution for the truncation error of the TILe-Rep. Then, we develop two posterior inference schemes for the TILe-Rep using the Markov chain Monte Carlo (MCMC) method and the variational inference (VI) scheme, respectively. We illustrate the usage of the TILe-Rep in the binary latent feature model (Ghahramani and Griffiths, 2005), the beta process factor analysis model (Paisley and Carin, 2009) and the beta process sparse factor model (Ohn and Kim, 2022).

The rest of the paper is organised as follows. Section 2 reviews the construction and basic properties of the beta process. Section 3 introduces the TILe-Rep of the beta process and investigates its truncation error. Section 4 develops the posterior inference schemes for the TILe-Rep and related models. Section 5 discusses our findings and gives some final remarks.

2 BETA PROCESS AND RELATED MODELS

Let G_0 be a continuous measure on the space $(\mathcal{S}, \mathcal{B})$ and let $G_0(\mathcal{S}) = \gamma < \infty$. Also, let $\alpha \in (0, \infty)$ be a positive scalar. Define a random measure $\tilde{H}_K$ as

$$\begin{aligned} \tilde{H}_K &= \sum_{k=1}^{K} \pi_k \delta_{\psi_k}, \\ \pi_k \overset{iid}{\sim} \text{Beta}\left(\frac{\alpha\gamma}{K}, \alpha\left(1-\frac{\gamma}{K}\right)\right), \;\; \psi_k &\overset{iid}{\sim} \frac{1}{\gamma} G_0. \end{aligned} \tag{1}$$

Then, the beta process H, abbreviated by $H \sim \text{BP}(\alpha, G_0)$, is defined as the limiting distribution of $\tilde{H}_K$ as $K \to \infty$.

To facilitate the simulation and posterior inference schemes, Paisley et al. (2010) developed a stick-breaking representation for the beta process in terms of

$$\begin{aligned} H &= \sum_{i=1}^{\infty} \sum_{j=1}^{N_i} V_{ij}^{(i)} \prod_{l=1}^{i-1} (1 - V_{ij}^{(l)}) \delta_{\psi_{ij}}, \\ N_i \overset{iid}{\sim} \text{Pois}(\gamma), \;\; V_{ij}^{(l)} &\overset{iid}{\sim} \text{Beta}(1, \alpha), \;\; \psi_{ij} \overset{iid}{\sim} \frac{1}{\gamma} G_0. \end{aligned} \tag{2}$$

The equivalence between the limiting distribution of (1) and the distribution of (2) can be found in Paisley et al. (2010).

The beta process can also be obtained from the characterisation of a Poisson point process. Consider a Poisson process τ with the mean measure

$$v(dw, d\psi) = \alpha w^{-1}(1-w)^{\alpha-1} \mathbb{1}_{\{0<w<1\}} dw G_0(d\psi). \tag{3}$$

We refer to $\rho(dw) = \alpha w^{-1}(1-w)^{\alpha-1}\mathbb{1}_{\{0<w<1\}}dw$ as the Lévy measure of the Poisson process and G_0 as its base measure. Since $\int_0^1 \rho(dw) = \infty$ and $\int_0^1 w\rho(dw) < \infty$, the Poisson process will generate infinite atoms with an almost surely finite total mass. We denote by $\{\tilde{J}_k\}_{k\geq 1}$ the atom weights of the Poisson process, where the tilde notation emphasises that the sequence of atom weights is presented according to the time of appearance without any reordering. Then, the beta process has the sequential representation

$$H = \sum_{k=1}^{\infty} \tilde{J}_k \delta_{\psi_k}. \tag{4}$$

The equivalence between (2) and (4) has been shown in Theorem 2 of Paisley et al. (2012). They viewed each round of the stick-breaking process (2) as a Poisson process and showed that the sum of the mean measures from all the rounds has a closed-form solution, which is exactly the mean measure (3). In the supplementary material, we provide an alternative proof using the ϵ-perturbation of the Lévy measure. Our proof aims to show that both representations have the same Laplace transform.

In Bayesian nonparametric models, the beta process is usually equipped with a Bernoulli likelihood process $\text{BeP}(\Theta)$, meaning that each atom weight $\tilde{J}_k$ denotes the prior probability of the component ψ_k. In this case, we obtain the celebrated beta-Bernoulli process (Thibaux and Jordan, 2007). Next, we introduce three models based on the beta-Bernoulli process prior. We

start with the binary latent feature model:

$$X_n \sim \mathcal{MN}(Z_n\mathbf{\Phi}, \sigma_X^2 \boldsymbol{I}_D),\ \ \Phi_k \sim \mathcal{MN}(0, \sigma_\phi^2 \boldsymbol{I}_D),$$
$$Z_n \sim \mathrm{BeP}(H),\ \ H \sim \mathrm{BP}(\alpha, G_0), \tag{5}$$

where the D-dimensional observation vector X_n has the Gaussian likelihood with the mean vector $Z_n\mathbf{\Phi}$ and the covariance matrix $\sigma_X^2 \boldsymbol{I}_D$, the binary feature allocation vector Z_n is a sample of the beta-Bernoulli process, and the independent latent features $\Phi_1, \Phi_2, \dots$ have the Gaussian distribution prior. If the sample X_n contains the feature Φ_k, $z_{nk} = 1$; Otherwise, $z_{nk} = 0$. Given a set of observations $\mathbf{X} := (X_1; \dots; X_N)$, the inferential target is to estimate the feature allocation matrix $\mathbf{Z} := (Z_1; \dots; Z_N)$ and the latent feature matrix $\mathbf{\Phi} := (\Phi_1; \Phi_2; \dots)$.

In many real-world datasets, the observations share common features but have different scores. For example, two images may contain the same objects but show different brightnesses for the objects. The binary latent feature model is not applicable in this case because the score is not represented by any variable in the model. To this end, the beta process factor analysis model introduces the factor value vector W_n for each observation X_n. Its element w_{nk} describes the score of the feature Φ_k. The model has the format

$$X_n \sim \mathcal{MN}((W_n \circ Z_n)\mathbf{\Phi}, \sigma_X^2 \boldsymbol{I}_D),$$
$$Z_n \sim \mathrm{BeP}(H),\ \ H \sim \mathrm{BP}(\alpha, G_0), \tag{6}$$
$$w_{nk} \sim \mathcal{N}(0, \sigma_w^2),\ \ \Phi_k \sim \mathcal{MN}(0, \boldsymbol{I}_D),$$

where the symbol $\circ$ represents the Hadamard, or element-wise multiplication of two vectors. The vector $(W_n \circ Z_n)$ is known as the factor value of the n-th sample. Due to the binary support of the beta-Bernoulli process Z_n, the factor value vector is sparse. Given $z_{nk} = 1$, the feature Φ_k would be more significant in the observation X_n if w_{nk} is larger. The inferential target is to estimate the feature allocation matrix $\mathbf{Z} := (Z_1; \dots; Z_N)$, the factor loading matrix $\mathbf{\Phi} := (\Phi_1; \Phi_2; \dots)$ and the factor value matrix $\mathbf{W} := (W_1; \dots; W_N)$.

Alternatively, it is possible to place the beta-Bernoulli prior on $\mathbf{\Phi}$ to obtain sparse factor loadings. To this end, we recall the beta process sparse factor model (Ohn and Kim 2022) in terms of

$$X_n \sim \mathcal{MN}(W_n(\mathbf{Z} \circ \mathbf{\Phi}), \sigma_X^2 \boldsymbol{I}_D),$$
$$Z_d = (z_{1d}, z_{2d}, \dots) \sim \mathrm{BeP}(H),\ \ H \sim \mathrm{BP}(\alpha, G_0), \tag{7}$$
$$w_{nk} \sim \mathcal{N}(0, \sigma_w^2),\ \ \phi_{kd} \sim \mathrm{Laplace}(0, 1).$$

In this model, the factor loading matrix $(\mathbf{Z} \circ \mathbf{\Phi})$ is sparse, and the elements of the factor value matrix $\mathbf{W}$ are almost surely non-zero. The inferential target is to estimate $\mathbf{Z}$, $\mathbf{\Phi}$ and $\mathbf{W}$ as well.

3 TRUNCATED INVERSE-LÉVY MEASURE REPRESENTATION

In this section, we introduce the TILe-Rep and demonstrate its connection to the existing sequential representations of the beta process. Since the random variables $\{\psi_k\}_{k\geq 1}$ are i.i.d., the sequential representation (4) is identical in distribution to a reordering of its components. In particular, let $J_1 > J_2 > \dots$ be the ranked values of $\{\tilde{J}_k\}_{k\geq 1}$, then the bate process has the following inverse-Lévy measure representation:

$$H = \sum_{k=1}^{\infty} J_k \delta_{\psi_k}. \tag{8}$$

The ranked atom weights $\{J_k\}_{k\geq 1}$ can be derived from the inverse-Lévy measure method (Rosiński, 2001) by setting $J_k := \rho^{\leftarrow}(\Gamma_k)$, where $\rho^{\leftarrow}(w) := \inf\{x \mid \rho(x, 1) \leq w\}$, $\rho(x, 1) := \int_x^1 \rho(dw)$ denotes the tail distribution of the Lévy measure $\rho(dw)$, $\Gamma_k := \sum_{j=1}^k E_j$, and $E_j \sim \mathrm{Exp}(1)$ are i.i.d. exponential random variables with mean 1.

We define the TILe-Rep as the finite-dimensional approximation of the inverse-Lévy measure representation (8):

$$H_K := \sum_{k=1}^{K} J_k \delta_{\psi_k}, \tag{9}$$

abbreviated by $H_K \sim \text{K-BP}(\alpha, G_0)$. The TILe-Rep involves the K largest atom weights of the beta process. Their joint distribution follows from the basic properties of Poisson random measure (Kyprianou, 2014):

$$\mathbb{P}(J_1 \in dx_1, \dots, J_K \in dx_K)$$
$$= \exp(-\gamma\rho(x_K, 1))\gamma^K \prod_{k=1}^{K} \rho(dx_k),$$

for $1 > x_1 > \dots > x_K > 0$. In practice, however, it is more convenient to use the ratio between the ranked jumps. To this end, we denote by $R_k := J_{k+1}/J_k$ the ratio between the $(k+1)$-th and k-th largest atom weights. Then, the joint distribution of $J_1, R_1, \dots, R_{K-1}$ follows from a change of variable:

$$\begin{aligned}&\mathbb{P}(J_1 \in dx_1, R_1 \in dr_1, \dots, R_{K-1} \in dr_{K-1})\\ &= \exp(-\gamma\rho(x_1 r_1 \dots r_{K-1}, 1))\\ &\quad \times \gamma^K \alpha^K (x_1 r_1 \dots r_{K-1})^{-1}\\ &\quad \times \prod_{k=0}^{K-1} \left(1 - x_1 \prod_{i=1}^{k} r_i\right)^{\alpha-1} dx_1 dr_1 \dots dr_{K-1},\end{aligned} \tag{10}$$

for $x_1 \in (0, 1)$ and $r_k \in (0, 1)$.

Recall that Teh et al. (2007) proposed a stick-breaking representation for the beta process based on the decreasing sequence of atom weights. We can easily

revert to their representation by setting $\alpha = 1$ and renaming γ as α. In this case, $\rho(x_1 r_1 \dots r_{K-1}, 1) = -\ln(x_1 r_1 \dots r_{K-1})$, and the joint distribution (10) becomes

$$\begin{aligned}&\mathbb{P}(J_1 \in dx_1, R_1 \in dr_1, \dots, R_{K-1} \in dr_{K-1}) \\ &= \alpha x_1^{\alpha-1} \alpha r_1^{\alpha-1} \dots \alpha r_{K-1}^{\alpha-1} dx_1 dr_1 \dots dr_{K-1}.\end{aligned} \tag{11}$$

Thus, we can answer the question in the final comments of Paisley et al. (2010), which sets out a future work to show that the stick-breaking representation (2) is equivalent to the one in Teh et al. (2007). Since (4) is equivalent to the former, and (8) is equivalent to the latter, the equivalence between the two stick-breaking representations follows immediately.

Next, we investigate the truncation error of the TILe-Rep. It is clear that the truncation error is revealed by the sum of the smaller atom weights, namely $\tau_K := \sum_{k=K+1}^{\infty} J_k$. We call τ_K a K-trimmed beta process as it is derived from the Poisson process τ by removing the K largest atom weights. Given the K-th largest atom weight J_K, τ_K is a Poisson process with the mean measure $v(dw, d\psi)\mathbb{1}_{\{0<w<J_K\}}$. It follows that τ_K has the density

$$f_{\rho,K}(z) = \sum_{i=0}^{n-1} \frac{(-\alpha)^i}{i!} L_i(z), \tag{12}$$

where $L_i(t)$ is defined recursively as follows:

$$L_0(z) = f_\rho(z) \exp(\gamma \rho(J_K, 1)), \;\; z \in (0, J_K),$$

$$L_{i+1}(z) = \int_{J_K}^{z-iJ_K} L_i(z-s)\gamma\rho(ds), \;\; z > (i+1)J_K,$$

and $f_\rho(z)$ denotes the density of the beta process τ. The detailed derivation of the density can be found in the supplementary material.

Using the joint distribution (10), it is possible to integrate out J_K and derive the unconditional distribution of τ_K. However, the convolution of $L_i(t)$ makes it extremely hard to derive an explicit expression. To better understand the truncation error, we provide a decomposition in distribution for τ_K.

Conditioning on J_K, τ_K has the following decomposition in distribution:

$$\frac{\tau_K}{J_K} \overset{d}{=} \begin{cases} \sigma_1 + \sum_{i=1}^{\text{Pois}(\gamma C_1)} X_i, & 0 < \alpha < 1, \\ \sigma_1, & \alpha = 1, \\ \sigma_2 + \sum_{i=1}^{\text{Pois}(\gamma C_2)} Y_i + \sum_{i=1}^{\text{Pois}(\gamma C_3)} Z_i, & \alpha > 1, \end{cases}$$

where σ_1 is a truncated Dickman process with the Lévy measure $\alpha w^{-1} \mathbb{1}_{\{0<w<1\}} dw$ at time γ, σ_2 is a truncated gamma process with the Lévy measure $\alpha w^{-1} e^{-K(\alpha-1)w} \mathbb{1}_{\{0<w<w^*\}} dw$ at time γ, the constants C_1, C_2, C_3 are defined such that

$$\begin{aligned} f_X(x) &= \frac{(1 - J_K x)^{\alpha-1} - 1}{C_1 x/\alpha} \mathbb{1}_{\{0<x<1\}}, \\ f_Y(y) &= \frac{(1 - J_K y)^{\alpha-1} - e^{-K(\alpha-1)y}}{C_2 y/\alpha} \mathbb{1}_{\{0<y<w^*\}}, \\ f_Z(z) &= \frac{(1 - J_K z)^{\alpha-1}}{C_3 z/\alpha} \mathbb{1}_{\{w^*<z<1\}}, \end{aligned}$$

are valid probability density functions, where w^* is the solution of the equation $1 - J_K w^* = e^{-K w^*}$, and X_i, Y_i, Z_i are i.i.d. random variables with the density functions $f_X(x), f_Y(y), f_Z(z)$, respectively. See supplementary material for the detailed derivation.

The usage of the decomposition is illustrated as follows. From the decomposition, we derive the conditional expectation of τ_K as

$$\begin{aligned}&\mathbb{E}(\tau_K \mid J_K) = J_K \gamma \times \\ &\begin{cases} \alpha + C_1 \mathbb{E}(X), & 0 < \alpha < 1, \\ \alpha, & \alpha = 1, \\ \alpha \frac{1 - e^{-K(\alpha-1)w^*}}{K(\alpha-1)} + C_2 \mathbb{E}(Y) + C_3 \mathbb{E}(Z), & \alpha > 1. \end{cases}\end{aligned}$$

The conditional expectation, when $\alpha = 1$, has been used in the existing literature (Ohn and Kim, 2022) to prove the posterior consistency of the beta process sparse factor model. Besides, the decomposition of τ_K also leads to a simulation algorithm for the truncation error. Recall that Dassios et al. (2019) developed the exact simulation algorithms for the truncated Dickman process σ_1 and the truncated gamma process σ_2. Also, we can simulate the random variables X_i, Y_i, Z_i via rejection sampling. Then the simulation algorithm of τ_K follows immediately.

We present some numerical results for the simulation algorithm. In the experiment, we use the inverse-Lévy measure method to sample from $J_1, \dots, J_K$, then simulate τ_K via the decomposition. The sample averages of τ_K with different hyperparameters are presented in Figure 1. The numerical results are based on 1000 random samples. From the figure we can see that the sample average of τ_K decreases fast. With a relatively low truncation level $K = 10$, it approaches zero for all the hyperparameters considered in the experiment.

Next, we investigate the posterior approximation error arising from the TILe-Rep. Consider the Bayesian nonparametric model of the format

$$\begin{aligned} X_n \mid Z_n &\sim f(. \mid Z_n), \\ Z_n \sim \text{BeP}(H), \;\; H &\sim \text{BP}(\alpha, G_0), \end{aligned} \tag{13}$$

for $n = 1, \dots, N$. We replace the beta process by the TILe-Rep and consider the following finite approxima-

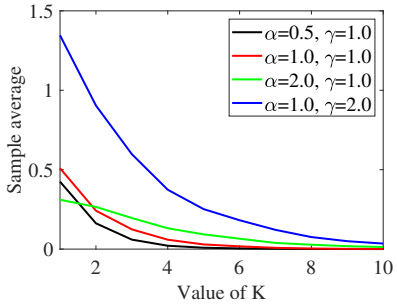


Figure 1: Sample Average of τ_K

tion of the model,

$$\begin{aligned} X_n \mid Z_n &\sim f(. \mid Z_n), \\ Z_n \sim \text{BeP}(H_K),\ \ H_K &\sim \text{K-BP}(\alpha, G_0). \end{aligned} \tag{14}$$

The approximation only uses the K largest atom weights of the beta process. Thus, the Bernoulli likelihood process reduces to a sequence of Bernoulli random variables in terms of $Z_n = (\text{Ber}(J_1), \ldots, \text{Ber}(J_K))$.

Denote by $p_{N,\infty}$ and $p_{N,K}$ the marginal densities of the observations $X_1, \ldots, X_N$ in model (13) and (14), respectively. The posterior approximation error is defined as $\|p_{N,\infty} - p_{N,K}\|_1$. From Theorem 4.2 of Campbell et al. (2019), we know

$$0 \leq 0.5\|p_{N,\infty} - p_{N,K}\|_1 \leq 1 - e^{-B_{N,K}} \leq 1, \tag{15}$$

where $B_{N,K}$ is a function with the upper bound

$$B_{N,K} \leq N \int_0^1 F_K(\gamma\rho(x,1))x\gamma\rho(dx),$$

and $F_K(.)$ denotes the cumulative distribution function of the gamma random variable $\text{Ga}(K,1)$. To derive a more explicit upper bound, a common method is to use the inequality $F_K(t) \leq (3t/K)^K$. However, it leads to the power of the tail distribution of the Lévy measure, which turns out to be hard to evaluate. Instead, we use a very rough upper bound, namely $F_K(t) \leq t(K-1)^{K-1}e^{-(K-1)}/\Gamma(K)$. In this case, the posterior approximation error has the upper bound (15) with

$$B_{N,K} \leq N\gamma^2\alpha^2(K-1)^{K-1}e^{-(K-1)}\zeta(2,\alpha)/\Gamma(K),$$

where $\zeta(s,\alpha)$ is the Hurwitz zeta function. The detailed derivation of the upper bound can be found in the supplementary material.

As $K \to \infty$, $(K-1)^{K-1}e^{-(K-1)}/\Gamma(K) \to 0$ decreasingly. Thus, it suffices to conclude that as the truncation level grows, the posterior approximation error decreases, and the marginal densities $p_{N,\infty}$ and $p_{N,K}$ become identical.

4 POSTERIOR INFERENCE SCHEME

In this section, we develop the MCMC and VI posterior inference schemes for the TILe-Rep. We will demonstrate the usage of these algorithms in the three models introduced in Section 2.

4.1 MCMC Algorithm

Consider a Bernoulli likelihood process with the TILe-Rep prior, namely $Z_n \sim \text{BeP}(H_K)$, $H_K \sim \text{K-BP}(\alpha, G_0)$. Let $\mathbf{Z} := (Z_1; \ldots; Z_N)$ be the $N \times K$ observation matrix. Denote by $m_{1,k} := \sum_{n=1}^N z_{nk}$ and $m_{0,k} := N - m_{1,k}$ the number of 1's and 0's in the k-th column of $\mathbf{Z}$, respectively. The posterior of the TILe-Rep has the expression

$$\begin{aligned} &\mathbb{P}(x_1, r_1, \ldots r_{K-1} \mid \mathbf{Z}) \\ &\propto \prod_{k=1}^K (x_1 r_1 \cdots r_{k-1})^{m_{1,k}} (1 - x_1 r_1 \cdots r_{k-1})^{m_{0,k}} \\ &\quad \times \mathbb{P}(J_1 \in dx_1, R_1 \in dr_1, \ldots, R_{K-1} \in dr_{K-1}). \end{aligned} \tag{16}$$

We use the Hamiltonian Monte Carlo (HMC, Neal 2011) algorithm to sample from the posterior. The algorithm is based on the gradient of the posterior and, thus, can discover the posterior more efficiently than the Metropolis-Hastings algorithm with a predefined transition kernel. The implementation of the HMC algorithm is straightforward, we provide the details in the supplementary material.

Note that the HMC algorithm involves evaluating the tail distribution of the Lévy measure $\rho(dw)$. It can be expressed in terms of the hypergeometric function:

$$\rho(x,1) = (1-x)^\alpha F_{2,1}(1, \alpha; \alpha+1; 1-x).$$

The efficient evaluation of the hypergeometric function has been well-studied. See, for example, Luke (1969). The package is available in various programming languages.

The posterior (16) uses fixed hyperparameters. It is also possible to put priors on the hyperparameters and estimate them in the posterior inference scheme. To this end, let the prior be $\pi(\alpha, \gamma)$. Then the posterior of the hyperparameters is given by

$$\begin{aligned} &\mathbb{P}(\alpha, \gamma \mid x_1, r_1, \ldots r_{K-1}) \\ &\propto \exp\Big(-\gamma \textstyle\int_{x_1 r_1 \ldots r_{K-1}}^1 \alpha w^{-1}(1-w)^{\alpha-1} dw\Big) \\ &\quad \times e^{-(-\ln(C_1))\alpha} \pi(\alpha, \gamma), \end{aligned}$$

where $C_1 := (1-x_1)(1-x_1r_1)\dots(1-x_1r_1\dots r_{K-1})$, so that $0 < C_1 < 1$. Various algorithms can be used for sampling from the posterior. For example, we may adopt the Metropolis-Hastings algorithm with an adaptive transition kernel. See, Haario et al. (2001) and Griffin and Stephens (2013) for the details.

We remark that although one can put priors on the hyperparameters and sample from them, it is worth investigating the sensitivity of the posterior with respect to the hyperparameters and derive a more explicit rule for selecting them. For example, an extension of the methods in Lijoi et al. (2007) and Giordano et al. (2023) will give us a principled approach for selecting the hyperparameters.

4.2 VI Scheme

Next, we discuss the VI scheme for the TILe-Rep. As shown in Section 3, the TILe-Rep generalises the stick-breaking representation of beta process proposed by Teh et al. (2007), and the VI scheme for the latter has been established in Doshi-Velez et al. (2009). Thus, we expect our VI scheme to be an extension of the method in Doshi-Velez et al. (2009). To ease the comparison, we rename the variables $(x_1, r_1, \dots, r_{K-1})$ in (10) by $\boldsymbol{v} = (v_1, \dots, v_K)$ as in the existing literature.

We approximate the posterior $\mathbb{P}(\boldsymbol{v} \mid \mathbf{Z})$ using the mean-filed variational inference (Wainwright and Jordan, 2008) method. Consider the variational distribution $q_{\boldsymbol{\tau}}(\boldsymbol{v}) = \prod_{k=1}^{K} \mathrm{Beta}(v_k; \tau_{k1}, \tau_{k2})$. The evidence lower bound (ELBO) of this approximation is given by

$$\begin{aligned}\mathcal{L}(q) :&= \mathcal{H}(q) + \mathbb{E}_q(\log \mathbb{P}(\boldsymbol{v}, \mathbf{Z})) \\ &= \mathcal{H}(q) + \mathbb{E}_q(\log \mathbb{P}(\boldsymbol{v})) \\ &\quad + \textstyle\sum_{k=1}^{K}\sum_{n=1}^{N} \mathbb{E}_q(\log \mathbb{P}(z_{nk} \mid \boldsymbol{v})),\end{aligned} \tag{17}$$

where $\mathcal{H}(q)$ denotes the entropy of the variational distribution $q_{\boldsymbol{\tau}}(\boldsymbol{v})$. The inference procedure aims to maximise the ELBO by optimising the parameters τ_{k1} and τ_{k2}, i.e., $\arg\max_{\boldsymbol{\tau}}(\mathcal{L}(q))$. The optimisation relies on the evaluation of the intractable component $\mathbb{E}_q(\log \mathbb{P}(\boldsymbol{v}))$, and we consider its lower bound instead. When $\alpha > 1$, we have the inequality

$$\begin{aligned}&\mathbb{E}_q\left(\log \mathbb{P}(\boldsymbol{v})\right) \\ &> (\gamma\alpha - 1)\mathbb{E}_q\left(\log(v_1 \dots v_K)\right) \\ &\quad + K\log(\gamma) + K\log(\alpha) \\ &\quad + (\alpha-1)\textstyle\sum_{k=1}^{K} \mathbb{E}_q(\log(1 - v_1 \dots v_k)),\end{aligned} \tag{18}$$

and it remains to evaluate the intractable expectation $\mathbb{E}_q(\log(1 - v_1 \dots v_k))$. To this end, we use the multinomial lower bound (Doshi-Velez et al., 2009). The details can be found in the supplementary material.

Finally, we find the optimal parameters of the variational distribution $q_{\boldsymbol{\tau}}(\boldsymbol{v})$:

$$\begin{aligned}\tau_{k1} =&\gamma\alpha + \textstyle\sum_{i=k}^{K}\sum_{n=1}^{N} \nu_{ni} \\ &+ (\alpha-1)\textstyle\sum_{i=k+1}^{K}\sum_{y=k+1}^{i} q_i(y) \\ &+ \textstyle\sum_{i=k+1}^{K}\sum_{n=1}^{N}(1-\nu_{ni})\sum_{y=k+1}^{i} q_i(y), \\ \tau_{k2} =&1 + (\alpha-1)\textstyle\sum_{i=k}^{K} q_i(k) \\ &+ \textstyle\sum_{i=k}^{K}\sum_{n=1}^{N}(1-\nu_{ni}) q_i(k),\end{aligned} \tag{19}$$

for $k = 1, \dots, K$, where $q_k(y)$ is a valid probability mass function for $y = 1, \dots, k$, such that

$$q_k(y) \propto e^{\psi(\tau_{y2}) + \sum_{m=1}^{y-1}\psi(\tau_{m1}) - \sum_{m=1}^{y}\psi(\tau_{m1}+\tau_{m2})}. \tag{20}$$

By setting $\alpha = 1$ and renaming γ as α, the results in equation (19) revert to the VI scheme proposed by Doshi-Velez et al. (2009) as we expect.

4.3 Binary Latent Feature Model

The TILe-Rep can be used to approximate the binary latent feature model (5). We replace the beta process prior by the TILe-Rep and obtain the finite approximation in terms of

$$\begin{gathered}X_n \sim \mathcal{MN}(Z_n\boldsymbol{\Phi}, \sigma_X^2\boldsymbol{I}_D), \;\; \Phi_k \sim \mathcal{MN}(0, \sigma_\phi^2\boldsymbol{I}_D), \\ Z_n \sim \mathrm{BeP}(H_K), \;\; H_K \sim \text{K-BP}(\alpha, G_0).\end{gathered} \tag{21}$$

To simplify the posterior inference scheme, we follow the collapsed approach (Ghahramani and Griffiths, 2005) and integrate out the latent feature matrix $\boldsymbol{\Phi}$. The likelihood of $\mathbf{X}$ can then be expressed as

$$\mathbb{P}(\mathbf{X} \mid \mathbf{Z}) = \frac{\exp\left(-T/(2\sigma_X^2)\right)}{(2\pi)^{\frac{ND}{2}}\sigma_X^{(N-K)D}\sigma_A^{KD}|\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_X^2}{\sigma_A^2}\mathbf{I}_K|^{\frac{D}{2}}},$$

where

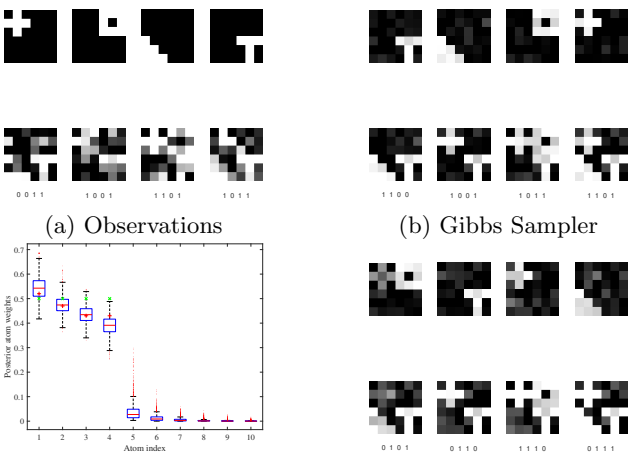
$$T := \mathrm{tr}\Big(\mathbf{X}^T(\mathbf{I} - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z} + (\sigma_X^2/\sigma_A^2)\mathbf{I})^{-1}\mathbf{Z}^T)\mathbf{X}\Big).$$

Note that it is also possible to retain $\boldsymbol{\Phi}$ in the likelihood and adopt the accelerated sampler (Doshi-Velez and Ghahramani, 2009). For simplicity, however, we focus on the collapsed sampler in the current paper.

The model posterior has the format $\mathbb{P}(H_K, \mathbf{Z} \mid \mathbf{X})$. We first develop a blocked Gibbs sampler to sample from H_K and $\mathbf{Z}$ iteratively. Given $\mathbf{Z}$, the posterior of H_K is given by (16), and we sample from it using the HMC algorithm. Next, we update $z_{nk} \in \{0, 1\}$ according to

$$\mathbb{P}(z_{nk} \mid X_n, \mathbf{Z}_{-nk}, H_K) \propto \mathbb{P}(z_{nk} \mid H_K)\mathbb{P}(X_n \mid Z_n),$$

for $n = 1, \dots, N$ and $k = 1, \dots, K$, and the blocked Gibbs sampler is completed.



(a) Observations (b) Gibbs Sampler

(c) Atom Weights (d) VI Scheme

Figure 2: Numerical Implementation of the Binary Latent Feature Model. (a) The true underlying features and four randomly selected observations. (b) Posterior inference results of the blocked Gibbs sampler. (c) Posterior atom weights of the TILe-Rep. (d) Posterior inference results of the VI scheme

Then, we develop the VI posterior inference scheme. Consider a tractable variational distribution $q = q_{\boldsymbol{\tau}}(\boldsymbol{v})q_{\boldsymbol{\phi}}(\boldsymbol{\Phi})q_{\boldsymbol{\nu}}(\mathbf{Z})$, where

$$\begin{aligned} q_{\boldsymbol{\tau}}(\boldsymbol{v}) &= \textstyle\prod_{k=1}^{K} \text{Beta}(v_k; \tau_{k1}, \tau_{k2}), \\ q_{\boldsymbol{\phi}}(\Phi_k) &= \textstyle\prod_{k=1}^{K} \text{Gaussian}(\Phi_k; \eta_k, \xi_k), \\ q_{\nu_{nk}}(z_{nk}) &= \textstyle\prod_{n=1}^{N}\prod_{k=1}^{K} \text{Bernoulli}(z_{nk}; \nu_{nk}). \end{aligned}$$

The ELBO of the variational distribution is given by

$$\begin{aligned} \mathcal{L}(q) &= \mathcal{H}(q) + \mathbb{E}_q(\log \mathbb{P}(\mathbf{X}, \boldsymbol{v}, \mathbf{Z}, \boldsymbol{\Phi})) \\ &= \mathcal{H}(q) + \mathbb{E}_q(\log \mathbb{P}(\boldsymbol{v})) \\ &\quad + \textstyle\sum_{k=1}^{K} \mathbb{E}_q(\log \mathbb{P}(\Phi_k)) \\ &\quad + \textstyle\sum_{k=1}^{K}\sum_{n=1}^{N} \mathbb{E}_q(\log \mathbb{P}(z_{nk} \mid \boldsymbol{\nu})) \\ &\quad + \textstyle\sum_{n=1}^{N} \mathbb{E}_q(\log \mathbb{P}(X_n \mid Z_n, \boldsymbol{\Phi})). \end{aligned} \tag{22}$$

We maximise the ELBO by optimising the parameters τ_{k1}, τ_{k2}, η_k, ξ_k and ν_{nk}. The parameter update for $q_{\boldsymbol{\tau}}(\boldsymbol{v})$ has been derived in (19), and the update of the other parameters follows the standard procedure. We provide the details in the supplementary material.

Next, we provide some numerical results for the blocked Gibbs sampler and the VI scheme. Consider four underlying features $\boldsymbol{\Phi} := (\Phi_1, \Phi_2, \Phi_3, \Phi_4)$ presented in the first row of Figure 2(a). Assume that each feature is included in the observation with a probability of 0.5 independently. We generate 100 samples based on the Gaussian likelihood and present four randomly selected samples in the second row of Figure 2(a). The underlying feature allocations of the observations are recorded below the images. We first analyse the observations using the blocked Gibbs sampler. We run the algorithm for 2000 iterations and present the posterior feature estimations in the first row of Figure 2(b). The experiment costs 386 seconds on MATLAB 2024a on a 64-bit Windows desktop with an Intel i9-12900 processor and 64GB RAM. Using the posterior values of $\mathbf{Z}$ and $\boldsymbol{\Phi}$, we reconstruct the four observations in the second row of Figure 2(b). Also, we record the posterior atom weights of the TILe-Rep (in the box plot), the prior probabilities of the features (in the green cross) and the actual proportions of the features in the observations (in the red plus) in Figure 2(c). Then, we analyse the observations using the VI scheme. We run the algorithm for 2000 iterations and present the results in Figure 2(d). The experiment costs 182 seconds. From the posterior inference results, we find that both algorithms recover the underlying features and the allocations correctly. However, the numerical results of the blocked Gibbs sampler include fewer noise.

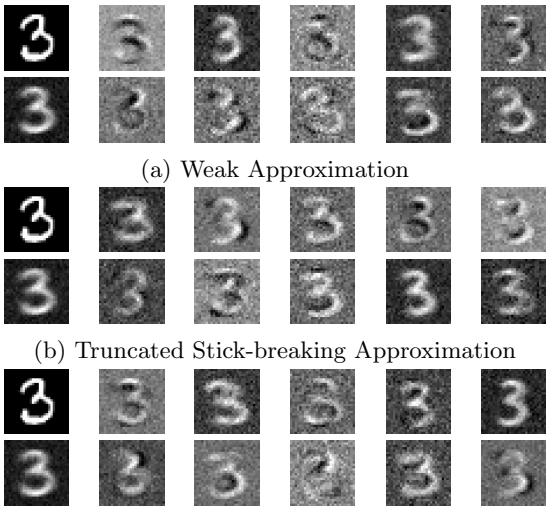
4.4 Beta Process Factor Analysis Model

Next, we use the TILe-Rep to approximate the beta process factor analysis model (6). We replace the beta process prior by the TILe-Rep and consider the approximation in terms of

$$\begin{aligned} X_n &\sim \mathcal{MN}((W_n \circ Z_n)\boldsymbol{\Phi}, \sigma_X^2 \boldsymbol{I}_D), \\ Z_n &\sim \text{BeP}(H_K), \;\; H_K \sim \text{K-BP}(\alpha, G_0), \\ w_{nk} &\sim \mathcal{N}(0, \sigma_w^2), \;\; \Phi_k \sim \mathcal{MN}(0, \boldsymbol{I}_D). \end{aligned} \tag{23}$$

The posterior of model (23) can be expressed as $\mathbb{P}(H_K, \mathbf{W}, \mathbf{Z}, \boldsymbol{\Phi} \mid \mathbf{X})$. We could use both the blocked Gibbs sampler and the VI scheme for posterior inference. For the blocked Gibbs sampler, we need to sample from H_K, $\mathbf{W}$, $\mathbf{Z}$, $\boldsymbol{\Phi}$ iteratively. While for the VI scheme, we approximate the posterior via the mean-filed variational distributions. Since our algorithms differ from the existing literature only in the prior distribution H_K, we refer the readers to Paisley and Carin (2009) for the details.

We illustrate some numerical results for the beta process factor analysis model based on the MNIST handwritten digital dataset (LeCun et al., 1998). We use the weak representation (1), the truncated stick-breaking representation (11) and the TILe-Rep to approximate the beta process. Note that for the weak representation, the beta process reduces to a sequence of beta random variables, which are conjugate to the Bernoulli likelihood. Thus, the posterior inference

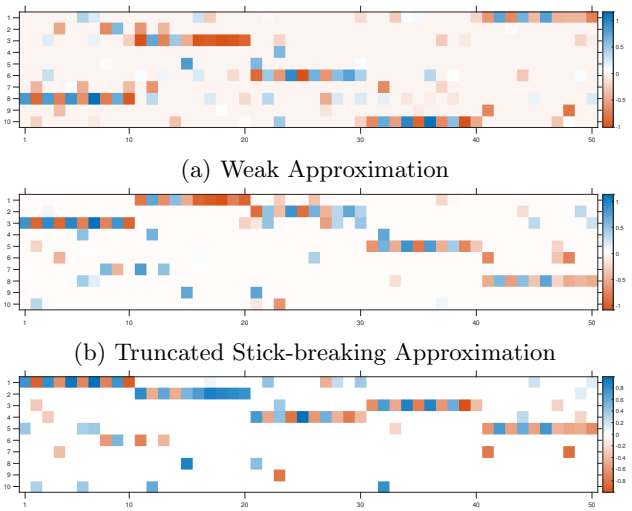


(a) Weak Approximation

(b) Truncated Stick-breaking Approximation

(c) TILe-Rep ($\alpha = 2.0$)

Figure 3: Posterior Estimation of the Factor Loadings and Image Reconstructions via the Beta Process Factor Analysis Model



(a) Weak Approximation

(b) Truncated Stick-breaking Approximation

(c) TILe-Rep ($\alpha = 2.0$)

Figure 4: Posterior Estimation of the Factor Loadings via the Beta Process Sparse Factor Model

scheme is straightforward, namely

$$\pi_k \mid \mathbf{Z} \sim \text{Beta}\left(\frac{\alpha\gamma}{K} + \sum_{n=1}^{N} z_{nk}, \alpha\left(1 - \frac{\gamma}{K}\right) + N - \sum_{n=1}^{N} z_{nk}\right).$$

We run the blocked Gibbs sampler to analyse 100 images containing the digital 3 and present the numerical results in Figure 3. In these figures, the upper figure in the first column is the original observation, and the lower figure in the first column is the image reconstruction result. The reconstruction is based on the posterior factor loadings presented in the second to the sixth columns. The blocked Gibbs sampler costs 506, 587 and 667 seconds for 25000 iterations based on the three approximation methods, respectively.

4.5 Beta Process Sparse Factor Model

Finally, we use the TILe-Rep to approximate the beta process sparse factor model (7). The approximation has the format

$$\begin{aligned} X_n &\sim \mathcal{MN}(W_n(\mathbf{Z} \circ \mathbf{\Phi}), \sigma_X^2 \boldsymbol{I}_D), \\ Z_d &\sim \text{BeP}(H), \quad H \sim \text{K-BP}(\alpha, G_0), \\ w_{nk} &\sim \mathcal{N}(0, \sigma_w^2), \quad \phi_{kd} \sim \text{Laplace}(0, 1). \end{aligned} \tag{24}$$

The posterior of model (24) can be expressed as $\mathbb{P}(H_K, \mathbf{W}, \mathbf{Z}, \mathbf{\Phi} \mid \mathbf{X})$, and both the blocked Gibbs sampler and the VI scheme can be used for posterior inference.

We illustrate some numerical results for the beta process sparse factor model based on the big five personality traits dataset (Goldberg, 2013). The dataset contains answers of 50 questions about personality with the five-level scale. The 50 questions can be grouped into five clusters which represent the five personality traits of extraversion, agreeableness, openness, conscientiousness, and neuroticism. We use the weak representation (1), the truncated stick-breaking representation (11) and the TILe-Rep to approximate the beta process. We run the blocked Gibbs sampler to analyse 200 randomly selected samples and present the posterior estimations of the factor loadings in Figure 4. From the figures we can see that all the methods induce five active factor loadings, which match the five personality traits in the dataset. However, the TILe-Rep method produce fewer noise for the factor loadings than the other two methods.

5 DISCUSSION

In this paper, we develop the TILe-Rep of the beta process and illustrate its applications in various models. We devise two posterior inference schemes for the TILe-Rep based on the blocked Gibbs sampler and the variational inference scheme.

It is straightforward to extend the TILe-Rep to the

generalisations of the beta process. Consider, for example, the stable-beta process (three-parameter beta process, Teh and Görür 2009; Broderick et al. 2012) induced by the Lévy measure

$$\rho(dw) = C_{\alpha,\theta} w^{-1-\alpha}(1-w)^{\theta+\alpha-1}\mathbb{1}_{\{0<w<1\}}dw, \quad (25)$$

where $C_{\alpha,\theta} := \Gamma(1+\theta)\Gamma(1-\alpha)^{-1}\Gamma(\theta+\alpha)^{-1}$, for $\alpha \in (0,1)$ and $\theta > -\alpha$. We denote by $J_1 > J_2 > \dots$ the ranked atom weights of the stable-beta process and $R_k := J_{k+1}/J_k$ the ratio between the ranked jumps. Then, the joint density of $(J_1, R_1, \dots, R_{K-1})$ is

$$\begin{aligned}
&\mathbb{P}(J_1 \in dx_1, R_1 \in dr_1 \dots, R_{K-1} \in dr_{K-1}) \\
&= \exp(-\gamma\rho(x_1 r_1 \dots r_{K-1}, 1))\gamma^K C_{\alpha,\theta}^K \\
&\quad \times x_1^{-K\alpha-1} r_1^{-(K-1)\alpha-1} \cdots r_{K-1}^{-\alpha-1} \\
&\quad \times \prod_{k=0}^{K-1} \left(1 - x_1 \prod_{i=1}^{k} r_i\right)^{\theta+\alpha-1} dx_1 dr_1 \dots dr_{K-1},
\end{aligned} \quad (26)$$

where $x_1 \in (0,1)$, $r_k \in (0,1)$, $k = 1, \dots, K-1$, and $\rho(x_1 r_1 \dots r_{K-1}, 1)$ denotes the tail distribution of the Lévy measure (25). For the Bernoulli likelihood process, the posterior of the TILe-Rep has the same format as (16), but the prior is replaced by (26). We can use the HMC algorithm to sample from the posterior. To this end, we provide the gradient of the posterior in the supplementary material.

The TILe-Rep method belongs to the general class of the truncated finite approximation (TFA, Nguyen et al. 2020, 2024) of completely random measures (CRMs). It also belongs to the deterministic arrival times construction (Lee et al. 2023) of the series representations of the CRMs. Therefore, the existing results from these literature are also applicable to the TILe-Rep. We have used Theorem 4.2 of Campbell et al. (2019) to derive the upper bound of the posterior approximation error in Section 3. From the discussion of Campbell et al. (2019), we know that the TILe-Rep has the lowest posterior approximation error within the TFA family. Therefore, we can choose another TFA method whose posterior approximation error is more interpretable compared to the upper bound of $B_{N,K}$ in Section 3 and use it as the upper bound for the error of the TILe-Rep. For example, we can choose the Bondesson representation whose posterior approximation error has the upper bound (Nguyen et al. 2020, 2024)

$$\begin{aligned}
||p_{N,\infty} - p_{N,K}|| &\leq ||p_{N,\infty} - p_{N,K}^{\text{Bondesson}}|| \\
&\leq \frac{C' + C''\ln^2 N + C'''\ln N \ln K}{K}.
\end{aligned}$$

The first inequality follows from the fact that the posterior approximation error of the Bondesson representation is higher than that of the TILe-Rep. From the second inequality, we can see that the error grows as $O(\ln^2 N)$ with fixed K and decreases as $O((\ln K)/K)$ for fixed N. For a fixed K, the error increases as N increases. In particular, as the sample size N increases, we would expect increasingly smaller components represented in the sample. To capture these components, we require finite approximation of increasingly larger sizes. For fixed N, the error goes to zero at least as fast as $O((\ln K)/K)$.

The TILe-Rep focuses on the K largest atom weights of the prior and ignores the infinite number of smaller atom weights. Alternatively, it is possible to truncate the support of the Lévy measure, such that only a finite number of atoms will be generated. For example, we could approximate the beta process with the truncated Lévy measure $\rho_\epsilon(dw) := \alpha w^{-1}(1-w)^{\alpha-1}\mathbb{1}_{\{\epsilon<w<1\}}dw$. A similar idea was considered by Argiento et al. (2016) in the context of Bayesian nonparametric mixture model. This approach has the advantage that all the atom weights larger than ϵ are included in the approximation. However, this means the total number of jumps follows a Poisson distribution, and the posterior inference scheme must take into account the randomness of the number of atoms. Also, the choice of the truncation level ϵ needs careful consideration.

The VI scheme introduced in this paper has two limitations. First, it is applicable only for $\alpha > 1$. Second, the lower bound used in (18) is relatively rough. Both points could be improved via a careful discussion of the properties of the hypergeometric function. They will be considered in future work.

Acknowledgements

We are grateful to the five anonymous reviewers for giving us detailed and constructive comments which have helped us a lot in improving the manuscript. Junyi Zhang's research is supported in part by the European Union – Next Generation EU, PRIN-PNRR 2022 (P2022H5WZ9) and the Research Grant Council of Hong Kong (15303524).

References

Al Labadi, L. and Zarepour, M. (2018). On approximations of the beta process in latent feature models: Point processes approach. *Sankhya A*, 80:59–79.

Argiento, R., Bianchini, I., and Guglielmi, A. (2016). Posterior sampling from ε-approximation of normalized completely random measure mixtures. *Electronic Journal of Statistics*, 10(2):3516–3547.

Broderick, T., Jordan, M. I., and Pitman, J. (2012). Beta processes, stick-breaking and power laws. *Bayesian Analysis*, 7(2):439–475.

Campbell, T., Huggins, J. H., How, J. P., and Broderick, T. (2019). Truncated random measures. *Bernoulli*, 25(2):1256–1288.

Dassios, A., Qu, Y., and Lim, J. W. (2019). Exact simulation of generalised Vervaat perpetuities. *Journal of Applied Probability*, 56(1):57–75.

Doshi-Velez, F. and Ghahramani, Z. (2009). Accelerated sampling for the Indian buffet process. In *Proceedings of the 26th International Conference on Machine Learning*, pages 273–280.

Doshi-Velez, F., Miller, K., Van Gael, J., and Teh, Y. W. (2009). Variational inference for the Indian buffet process. In *Artificial Intelligence and Statistics*, pages 137–144.

Ghahramani, Z. and Griffiths, T. (2005). Infinite latent feature models and the Indian buffet process. *Advances in Neural Information Processing Systems*, 18.

Giordano, R., Liu, R., Jordan, M. I., and Broderick, T. (2023). Evaluating sensitivity to the stick-breaking prior in Bayesian nonparametrics. *Bayesian Analysis*, 18(1):287–366.

Goldberg, L. R. (2013). An alternative "description of personality": The big-five factor structure. In *Personality and Personality Disorders*, pages 34–47. Routledge.

Griffin, J. E. and Stephens, D. A. (2013). Advances in Markov chain Monte Carlo. In *Bayesian theory and applications*, pages 104–142. Oxford Univ. Press, Oxford.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, 18(3):1259–1294.

Kim, Y. (1999). Nonparametric Bayesian estimators for counting processes. *The Annals of Statistics*, 27(2):562–588.

Kyprianou, A. E. (2014). *Fluctuations of Lévy processes with applications: Introductory Lectures*. Springer Berlin, Heidelberg.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, J., Miscouridou, X., and Caron, F. (2023). A unified construction for series representations and finite approximations of completely random measures. *Bernoulli*, 29(3):2142–2166.

Lijoi, A., Mena, R. H., and Prünster, I. (2007). Controlling the reinforcement in Bayesian nonparametric mixture models. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 69(4):715–740.

Luke, Y. L. (1969). *The special functions and their approximations, Vol. I*, volume Vol. 53 of *Mathematics in Science and Engineering*. Academic Press, New York-London.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 113–162. CRC Press, Boca Raton, FL.

Nguyen, T. D., Huggins, J., Masoero, L., Mackey, L., and Broderick, T. (2024). Independent finite approximations for bayesian nonparametric inference. *Bayesian Analysis*, 19(4):1187–1224.

Nguyen, T. D., Huggins, J. H., Masoero, L., Mackey, L., and Broderick, T. (2020). Independent versus truncated finite approximations for bayesian nonparametric inference. In *"I Can't Believe It's Not Better!" NeurIPS 2020 workshop*.

Ohn, I. and Kim, Y. (2022). Posterior consistency of factor dimensionality in high-dimensional sparse factor models. *Bayesian Analysis*, 17(2):491–514.

Paisley, J., Blei, D., and Jordan, M. (2012). Stick-breaking beta processes and the Poisson process. In *Artificial Intelligence and Statistics*, pages 850–858.

Paisley, J. and Carin, L. (2009). Nonparametric factor analysis with beta process priors. In *Proceedings of the 26th International Conference on Machine Learning*, pages 777–784.

Paisley, J., Carin, L., and Blei, D. (2011). Variational inference for stick-breaking beta process priors. In *Proceedings of the 28th International Conference on Machine Learning*.

Paisley, J. W., Zaas, A. K., Woods, C. W., Ginsburg, G. S., and Carin, L. (2010). A stick-breaking construction of the beta process. In *Proceedings of the 27th International Conference on Machine Learning*, pages 847–854.

Phadia, E. G. (2016). *Prior processes and their applications*. Springer Series in Statistics. Springer, second edition.

Rosiński, J. (2001). Series representations of Lévy processes from the perspective of point processes. In *Lévy processes*, pages 401–415. Birkhäuser Boston, Boston, MA.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650.

Teh, Y. W. and Görür, D. (2009). Indian buffet processes with power-law behavior. *Advances in Neural Information Processing Systems*, 22.

Teh, Y. W., Görür, D., and Ghahramani, Z. (2007). Stick-breaking construction for the Indian buffet process. In *Artificial Intelligence and Statistics*, pages 556–563.

Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In *Artificial Intelligence and Statistics*, pages 564–571.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
   Yes. The detailed mathematical settings and assumptions of the algorithms and models are clearly described in Section 2.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
   Yes. The CPU times of the different algorithms are included in Section 4.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
   Yes. The MATLAB code is included.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results.
   Yes. The statements of the full set of assumptions of all theoretical results are included in Section 3.
   (b) Complete proofs of all theoretical results.
   Yes. The detailed proofs are provided in the supplementary material.
   (c) Clear explanations of any assumptions.
   Yes. All the assumptions are clearly explained.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
   Yes. The data and instructions needed to reproduce the main experimental results are included.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
   Yes. The hyperparameters used to generate the figures and tables are included.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
   Not Applicable. The numerical experiments in this paper aim to illustrate the usefulness of the proposed methods.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
   Yes. A description of the computing infrastructure used is included.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets.
   Yes. This paper uses the MNIST handwritten digital dataset and the big five personality traits dataset. The citations of the creators are included.
   (b) The license information of the assets, if applicable. Not Applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.
   (d) Information about consent from data providers/curators. Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

A MISSING PROOFS

In this section, we provided detailed proofs of the results that are missing in the main paper.

A.1 Derivation of the Equivalence between (2) and (4)

Let $X \sim \text{Beta}(1, \alpha)$ and $Y_i \sim \text{Beta}(\alpha, 1)$ be independent beta random variables. Denote by $Z_m := XY_1 \dots Y_m$. Let $Z_m^{(1)}, Z_m^{(2)}, \dots$ be independent copies of Z_m, and let $N \sim \text{Pois}(\gamma)$ be an independent Poisson random variable. We first calculate the Laplace transform of the compound Poisson process $S_m := \sum_{i=1}^{N} Z_m^{(i)} \delta_{\psi_i}$.

Recall that the Laplace transform of the $\text{Beta}(1, \alpha)$ random variable is

$$\begin{aligned}\mathbb{E}(e^{-\beta X}) &= 1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{1+r}{1+\alpha+r} \right) \frac{(-\beta)^k}{k!} \\ &= 1 + \frac{1}{1+\alpha} \sum_{k=1}^{\infty} \left(\prod_{r=1}^{k-1} \frac{1}{1+\alpha+r} \right) (-\beta)^k,\end{aligned}$$

and the k-th moment of the $\text{Beta}(\alpha, 1)$ random variable is

$$\mathbb{E}(Y_i^k) = \int_0^1 y^k \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\Gamma(1)} y^{\alpha-1} dy = \frac{\alpha}{k+\alpha}.$$

It follows that the Laplace transform of Z_m is

$$\begin{aligned}\mathbb{E}(e^{-\beta Z_m}) &= \int_0^1 \cdots \int_0^1 \int_0^1 e^{-\beta x y_1 \dots y_m} f(x) f(y_1) \dots f(y_m) dx dy_1 \dots dy_m \\ &= \int_0^1 \cdots \int_0^1 \left(1 + \frac{1}{1+\alpha} \sum_{k=1}^{\infty} \left(\prod_{r=1}^{k-1} \frac{1}{1+\alpha+r} \right) (-\beta y_1 \dots y_m)^k \right) f(y_1) \dots f(y_m) dy_1 \dots dy_m \\ &= 1 + \frac{1}{1+\alpha} \sum_{k=1}^{\infty} \left(\prod_{r=1}^{k-1} \frac{1}{1+\alpha+r} \right) (-\beta)^k \left(\frac{\alpha}{k+\alpha} \right)^m,\end{aligned}$$

where $f(x)$ and $f(y_i)$ denotes the density of X and Y_i, respectively. Then, S_m has the Laplace transform

$$\begin{aligned}\mathbb{E}(e^{-\beta \sum_{i=1}^{N} Z_m^{(i)}}) &= \sum_{n=0}^{\infty} \frac{\gamma^n}{n!} e^{-\gamma} \left(1 + \frac{1}{1+\alpha} \sum_{k=1}^{\infty} \left(\prod_{r=1}^{k-1} \frac{1}{1+\alpha+r} \right) (-\beta)^k \left(\frac{\alpha}{k+\alpha} \right)^m \right)^n \\ &= \exp(-\gamma) \exp\left(\gamma \left(1 + \frac{1}{1+\alpha} \sum_{k=1}^{\infty} \left(\prod_{r=1}^{k-1} \frac{1}{1+\alpha+r} \right) (-\beta)^k \left(\frac{\alpha}{k+\alpha} \right)^m \right) \right) \\ &= \exp\left(\gamma \left(\frac{1}{1+\alpha} \sum_{k=1}^{\infty} \left(\prod_{r=1}^{k-1} \frac{1}{1+\alpha+r} \right) (-\beta)^k \left(\frac{\alpha}{k+\alpha} \right)^m \right) \right) \\ &= \exp\left(\frac{\gamma}{1+\alpha} \sum_{k=1}^{\infty} \left(\prod_{r=1}^{k-1} \frac{1}{1+\alpha+r} \right) (-\beta)^k \frac{\alpha^m}{(k+\alpha)^m} \right).\end{aligned}$$

Next, we show that the representations (2) and (4) have the identical Laplace transform. From the construction (2) we know $\mathbb{E}(e^{-\beta H}) = \mathbb{E}(e^{-\beta \sum_{m=0}^{\infty} S_m})$. Since S_m, $m = 0, 1, \ldots$, are independent, it follows that

$$\begin{aligned}\mathbb{E}(e^{-\beta H}) &= \prod_{m=0}^{\infty} \exp\left(\frac{\gamma}{1+\alpha}\sum_{k=1}^{\infty}\left(\prod_{r=1}^{k-1}\frac{1}{1+\alpha+r}\right)(-\beta)^k\left(\frac{\alpha}{k+\alpha}\right)^m\right)\\ &= \exp\left(\frac{\gamma}{1+\alpha}\sum_{k=1}^{\infty}\left(\prod_{r=1}^{k-1}\frac{1}{1+\alpha+r}\right)(-\beta)^k\frac{k+\alpha}{k}\right)\\ &= \exp\left(\gamma\sum_{k=1}^{\infty}\left(\prod_{r=1}^{k-1}\frac{1}{\alpha+r}\right)\frac{(-\beta)^k}{k}\right).\end{aligned} \tag{S.1}$$

Thus, we have obtained the Laplace transform the stick-breaking representation (2).

On the other hand, let $\lambda := \Gamma(\epsilon)\Gamma(\alpha)/\Gamma(\epsilon+\alpha)$, for $\epsilon > 0$, and consider the function

$$L(\epsilon) := \exp\left(-\gamma\alpha\lambda\int_0^1 (1-e^{-\beta w})\frac{\Gamma(\epsilon+\alpha)}{\Gamma(\epsilon)\Gamma(\alpha)}w^{\epsilon-1}(1-w)^{\alpha-1}dw\right).$$

The integral inside the exponent part of $L(\epsilon)$ can be evaluated using the Laplace transform of a Beta(ϵ, α) random variable. We rewrite the function $L(\epsilon)$ as

$$\begin{aligned}L(\epsilon) &= \exp\left(-\gamma\alpha\lambda\left(1-\int_0^1 e^{-\beta w}\frac{\Gamma(\epsilon+\alpha)}{\Gamma(\epsilon)\Gamma(\alpha)}w^{\epsilon-1}(1-w)^{\alpha-1}dw\right)\right)\\ &= \exp(-\gamma\alpha\lambda)\exp\left(\gamma\alpha\lambda\left[1+\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\epsilon+r}{\epsilon+\alpha+r}\right)\frac{(-\beta)^k}{k!}\right]\right)\\ &= \exp\left(\gamma\alpha\lambda\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\epsilon+r}{\epsilon+\alpha+r}\right)\frac{(-\beta)^k}{k!}\right).\end{aligned}$$

The summation inside the exponent term can be written as

$$\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\epsilon+r}{\epsilon+\alpha+r}\right)\frac{(-\beta)^k}{k!} = \frac{\epsilon}{\epsilon+\alpha}\sum_{k=1}^{\infty}\left(\prod_{r=1}^{k-1}\frac{\epsilon+r}{\epsilon+\alpha+r}\right)\frac{(-\beta)^k}{k!}.$$

Recall that $\lambda := \Gamma(\epsilon)\Gamma(\alpha)/\Gamma(\epsilon+\alpha)$ and

$$\lim_{\epsilon\to 0}\gamma\alpha\lambda\frac{\epsilon}{\epsilon+\alpha} = \lim_{\epsilon\to 0}\gamma\alpha\frac{\Gamma(\epsilon)\Gamma(\alpha)}{\Gamma(\epsilon+\alpha)}\frac{\epsilon}{\epsilon+\alpha} = \gamma\alpha\frac{\Gamma(\alpha)}{\Gamma(\alpha+1)} = \gamma.$$

Thus,

$$\lim_{\epsilon\to 0}\gamma\alpha\lambda\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\epsilon+r}{\epsilon+\alpha+r}\right)\frac{(-\beta)^k}{k!} = \gamma\sum_{k=1}^{\infty}\left(\prod_{r=1}^{k-1}\frac{r}{\alpha+r}\right)\frac{(-\beta)^k}{k!},$$

and $L(\epsilon)$ converges to the right hand side of the equation (S.1) as $\epsilon \to 0$. It follows that the limit of $L(\epsilon)$ is exactly the Laplace transform of (2).

In the meanwhile, recall that the Lévy-Khintchine representation of the beta process is

$$\mathbb{E}(e^{-\beta H}) = \exp\left(-\gamma\int_0^1 (1-e^{-\beta w})\alpha w^{-1}(1-w)^{\alpha-1}dw\right).$$

It is clear that $L(\epsilon)$ converges to the Lévy-Khintchine representation:

$$\lim_{\epsilon\to 0} L(\epsilon) = \lim_{\epsilon\to 0}\exp\left(-\gamma\alpha\int_0^1 (1-e^{-\beta w})w^{\epsilon-1}(1-w)^{\alpha-1}dw\right) = \mathbb{E}(e^{-\beta H}).$$

Thus, the limit of $L(\epsilon)$ is also the Laplace transform of (4).

We have shown that the Laplace transform of both (2) and (4) are the limit of $L(\epsilon)$. Thus, (2) and (4) are identical in distribution.

A.2 Density of the K-trimmed Beta Process τ_K

The K-trimmed beta process has the conditional Lévy-Khintchine representation

$$\mathbb{E}\left(\exp(-\beta\tau_K) \mid J_K\right) = \exp\left(-\gamma \int_0^{J_K} (1-e^{-\beta w})\alpha w^{-1}(1-w)^{\alpha-1}dw\right).$$

The density of τ_K can be derived via the inverse Laplace transform as follows,

$$\begin{aligned}
&f_{\rho,J_K}(z)\\
&= \mathcal{L}^{-1}\left\{\mathbb{E}\left(\exp(-\beta\tau_K) \mid J_K\right)\right\}\\
&= \mathcal{L}^{-1}\left\{\exp\left(-\gamma\int_0^{J_K}(1-e^{-\beta w})\alpha w^{-1}(1-w)^{\alpha-1}dw\right)\right\}\\
&= \mathcal{L}^{-1}\left\{\exp\left(-\gamma\int_0^{1}(1-e^{-\beta w})\alpha w^{-1}(1-w)^{\alpha-1}dw\right)\exp\left(\gamma\int_{J_K}^{1}(1-e^{-\beta w})\alpha w^{-1}(1-w)^{\alpha-1}dw\right)\right\}\\
&= \mathcal{L}^{-1}\left\{\mathbb{E}(e^{-\beta\tau})\exp\left(\gamma\int_{J_K}^{1}\alpha w^{-1}(1-w)^{\alpha-1}dw\right)\exp\left(-\gamma\int_{J_K}^{1}e^{-\beta w}\alpha w^{-1}(1-w)^{\alpha-1}dw\right)\right\}\\
&= \mathcal{L}^{-1}\left\{\mathbb{E}(e^{-\beta\tau})\exp(\gamma\rho(J_K,1))\sum_{k=0}^{\infty}\frac{(-1)^k}{k!}\left(\gamma\int_{J_K}^{1}e^{-\beta w}\alpha w^{-1}(1-w)^{\alpha-1}dw\right)^k\right\}.
\end{aligned}$$

Denote by $f_\rho(z)$ the density of τ, then the density of τ_K can be derived by convolution.

A.3 Decomposition of the K-trimmed Beta Process τ_K

The conditional Lévy-Khintchine representation of the K-trimmed beta process can be written as

$$\mathbb{E}\left(\exp(-\beta\tau_K) \mid J_K\right) = \exp\left(-\gamma\int_0^1(1-e^{-\beta J_K w})\alpha w^{-1}(1-J_K w)^{\alpha-1}dw\right).$$

When $0<\alpha\leq 1$, we rewrite the Lévy measure of τ_K as

$$\alpha w^{-1}(1-J_K w)^{\alpha-1}\mathbb{1}_{\{0<w<1\}} = \alpha w^{-1}\mathbb{1}_{\{0<w<1\}} + \alpha\frac{(1-J_K w)^{\alpha-1}-1}{w}\mathbb{1}_{\{0<w<1\}}.$$

The first term on the right hand side of the equation is the Lévy measure of a truncated Dickman process. While for the second term, since

$$\lim_{w\to 0}\frac{(1-J_K w)^{\alpha-1}-1}{w} = (1-\alpha)J_K < \infty,$$

$$\lim_{w\to 1}\frac{(1-J_K w)^{\alpha-1}-1}{w} = (1-J_K)^{\alpha-1}-1 < \infty,$$

and $((1-J_K w)^{\alpha-1}-1)/w$ is continuous in $w\in(0,1)$, we have that

$$C_1 = \int_0^1 \alpha\frac{(1-J_K w)^{\alpha-1}-1}{w}dw < \infty.$$

Thus, we can define a random variable X whose probability density function is $f_X(x)$.

When $1<\alpha<\infty$, we define w^* as the solution to the equation $1-J_K w^* = e^{-Kw^*}$, then we rewrite the Lévy measure of τ_K as

$$\begin{aligned}
&\alpha w^{-1}(1-J_K w)^{\alpha-1}\mathbb{1}_{\{0<w<1\}}\\
&= \alpha w^{-1}e^{-K(\alpha-1)w}\mathbb{1}_{\{0<w<w^*\}} + \alpha\frac{(1-J_K w)^{\alpha-1}-e^{-K(\alpha-1)w}}{w}\mathbb{1}_{\{0<w<w^*\}} + \alpha\frac{(1-J_K w)^{\alpha-1}}{w}\mathbb{1}_{\{w^*<w<1\}}.
\end{aligned}$$

Since

$$\lim_{w\to 0}\frac{(1-J_K w)^{\alpha-1}-e^{-K(\alpha-1)w}}{w}=(\alpha-1)(K-J_K)<\infty,$$

$$\lim_{w\to w^*}\frac{(1-J_K w)^{\alpha-1}-e^{-K(\alpha-1)w}}{w}=0,$$

and $((1-J_K w)^{\alpha-1}-e^{-K(\alpha-1)w})/w$ is continuous in $w\in(0,w^*)$, we have that

$$C_2=\int_0^{w^*}\alpha\frac{(1-J_K w)^{\alpha-1}-e^{-K(\alpha-1)w}}{w}dw<\infty.$$

Thus, we can define a random variable Y whose probability density function is $f_Y(y)$. Also, since

$$\lim_{w\to w^*}\frac{(1-J_K w)^{\alpha-1}}{w}=\frac{(1-J_K w^*)^{\alpha-1}}{w^*}<\infty,$$

$$\lim_{w\to 1}\frac{(1-J_K w)^{\alpha-1}}{w}=(1-J_K)^{\alpha-1}<\infty,$$

and $((1-J_K w)^{\alpha-1})/w$ is continuous in $w\in(w^*,1)$, we have that

$$C_3=\int_{w^*}^1\alpha\frac{(1-J_K w)^{\alpha-1}}{w}dw<\infty.$$

Thus, we can define a random variable Z whose probability density function is $f_Z(z)$.

A.4 Posterior Approximation Error of TILe-Rep

From Theorem 4.2 of Campbell et al. (2019), we know

$$B_{N,K}\le N\int_0^1 F_K\left(\gamma\rho(x,1)\right)x\gamma\rho(dx),$$

where $F_K(.)$ denotes the cumulative distribution function of the gamma random variable $\mathrm{Ga}(K,1)$. The cdf has the upper bound $F_K(t)\le t(K-1)^{K-1}e^{-(K-1)}/\Gamma(K)$, for $K>1$. Also, the tail distribution of the Lévy measure $\rho(dw)$ can be written as

$$\rho(x,1)=\int_x^1\alpha w^{-1}(1-w)^{\alpha-1}dw=(1-x)^{\alpha}F_{2,1}(1,\alpha;\alpha+1;1-x).$$

Then we can rewrite the upper bound of $B_{N,K}$ as

$$\begin{aligned}
B_{N,K}&\le N\int_0^1\gamma(1-x)^{\alpha}F_{2,1}(1,\alpha;\alpha+1;1-x)\frac{(K-1)^{K-1}e^{-(K-1)}}{\Gamma(K)}x\gamma\alpha x^{-1}(1-x)^{\alpha-1}dx\\
&=N\gamma^2\alpha\frac{(K-1)^{K-1}e^{-(K-1)}}{\Gamma(K)}\int_0^1 x^{2\alpha-1}F_{2,1}(1,\alpha;\alpha+1;x)dx\\
&\le N\gamma^2\alpha\frac{(K-1)^{K-1}e^{-(K-1)}}{\Gamma(K)}\int_0^1 x^{\alpha-1}F_{2,1}(1,\alpha;\alpha+1;x)dx\\
&=N\gamma^2\alpha\frac{(K-1)^{K-1}e^{-(K-1)}}{\Gamma(K)}\alpha\zeta(2,\alpha),
\end{aligned}$$

where $\zeta(s,\alpha)$ is the Hurwitz zeta function.

A.5 Derivation of the Gradient for HMC Algorithm

For the TILe-Rep of the stable beta process, the posterior is given by

$$\begin{aligned}\mathbb{P}(x_1, r_1, \dots r_{K-1} \mid \mathbf{Z}) \propto \exp(-\gamma\rho(x_1 r_1 \dots r_{K-1}, 1))(1-x_1)^{m_{0,1}+\theta+\alpha-1} \\ \times (1-x_1 r_1)^{m_{0,2}+\theta+\alpha-1} \cdots (1-x_1 r_1 \cdots r_{K-1})^{m_{0,K}+\theta+\alpha-1} \\ \times x_1^{(m_{1,1}+\cdots+m_{1,K})-K\alpha-1} r_1^{(m_{1,2}+\cdots+m_{1,K})-(K-1)\alpha-1} \cdots r_{K-1}^{m_{1,K}-\alpha-1}.\end{aligned}$$

Note that we can revert to the TILe-Rep of the beta process by setting $\alpha = 0$ and rename θ as α in the derivation. To facilitate the HMC algorithm, we first make the change of variables $\mathcal{X}_1 := \tan(\pi(x_1 - 0.5))$ and $\mathcal{R}_k := \tan(\pi(r_k - 0.5))$ to obtain unconstrained variables. It follows that

$$x_1 = \frac{1}{2} + \frac{1}{\pi}\operatorname{atan}(\mathcal{X}_1) \quad \text{and} \quad r_k = \frac{1}{2} + \frac{1}{\pi}\operatorname{atan}(\mathcal{R}_k). \tag{S.2}$$

To simplify the expression, we use both the constrained variables $(x_1, r_1, \dots, r_{K-1})$ and the unconstrained variables $(\mathcal{X}_1, \mathcal{R}_1, \dots, \mathcal{R}_{K-1})$ in the following derivations. But we will keep in mind that all the constrained variables are functions of the unconstrained variables in terms of (S.2). Then we have

$$\begin{aligned}\mathbb{P}(\mathcal{X}_1, \mathcal{R}_1, \dots \mathcal{R}_{K-1} \mid \mathbf{Z}) \propto \exp(-\gamma\rho(x_1 r_1 \dots r_{K-1}, 1))(1-x_1)^{m_{0,1}+\theta+\alpha-1} \\ \times (1-x_1 r_1)^{m_{0,2}+\theta+\alpha-1} \cdots (1-x_1 r_1 \cdots r_{K-1})^{m_{0,K}+\theta+\alpha-1} \\ \times x_1^{(m_{1,1}+\cdots+m_{1,K})-K\alpha-1} r_1^{(m_{1,2}+\cdots+m_{1,K})-(K-1)\alpha-1} \cdots r_{K-1}^{m_{1,K}-\alpha-1} \\ \times \pi^{-1}(1+\mathcal{X}_1^2)^{-1} \pi^{-1}(1+\mathcal{R}_1^2)^{-1} \cdots \pi^{-1}(1+\mathcal{R}_{K-1}^2)^{-1}.\end{aligned}$$

Thus, the log-posterior is

$$\begin{aligned}&\log(\mathbb{P}(\mathcal{X}_1, \mathcal{R}_1, \dots \mathcal{R}_{K-1} \mid \mathbf{Z})) \\ &= C - \gamma\rho(x_1 r_1 \dots r_{K-1}, 1) + (m_{0,1} + \theta + \alpha - 1)\log(1-x_1) \\ &\quad + (m_{0,2} + \theta + \alpha - 1)\log(1-x_1 r_1) + \cdots + (m_{0,K} + \theta + \alpha - 1)\log(1 - x_1 r_1 \dots r_{K-2} r_{K-1}) \\ &\quad + ((m_{1,1} + \cdots + m_{1,K}) - K\alpha - 1)\log(x_1) \\ &\quad + ((m_{1,2} + \cdots + m_{1,K}) - (K-1)\alpha - 1)\log(r_1) + \cdots + (m_{1,K} - \alpha - 1)\log(r_{K-1}) \\ &\quad - \log(\pi) - \log\big(1 + \mathcal{X}_1^2\big) - \log(\pi) - \log\big(1 + \mathcal{R}_1^2\big) - \cdots - \log(\pi) - \log\big(1 + \mathcal{R}_{K-1}^2\big).\end{aligned}$$

The derivative of the log-posterior with respect to $\mathcal{X}_1$ is

$$\frac{d}{d\mathcal{X}_1}\log(\mathbb{P}(\mathcal{X}_1, \mathcal{R}_1, \dots \mathcal{R}_{K-1} \mid \mathbf{Z})) = \frac{1}{\pi}\frac{1}{1+\mathcal{X}_1^2}\frac{d}{dx_1}(\cdots) - \frac{2\mathcal{X}_1}{1+\mathcal{X}_1^2},$$

where

$$\begin{aligned}\frac{d}{dx_1}(\cdots) :&= \gamma C_{\alpha,\theta}(x_1 r_1 \cdots r_{K-1})^{-1-\alpha}(1 - x_1 r_1 \cdots r_{K-1})^{\theta+\alpha-1} r_1 \cdots r_{K-1} \\ &\quad + (m_{0,1} + \theta + \alpha - 1)\frac{-1}{1-x_1} + (m_{0,2} + \theta + \alpha - 1)\frac{-r_1}{1-x_1 r_1} + \cdots \\ &\quad + (m_{0,K} + \theta + \alpha - 1)\frac{-r_1 \dots r_{K-2} r_{K-1}}{1 - x_1 r_1 \dots r_{K-2} r_{K-1}} + ((m_{1,1} + \cdots + m_{1,K}) - K\alpha - 1)x_1^{-1}.\end{aligned}$$

The derivative of the log-posterior with respect to $\mathcal{R}_k$ is

$$\frac{d}{d\mathcal{R}_k}\log(\mathbb{P}(\mathcal{X}_1, \mathcal{R}_1, \dots \mathcal{R}_{K-1} \mid \mathbf{Z})) = \frac{1}{\pi}\frac{1}{1+\mathcal{R}_k^2}\frac{d}{dr_k}(\cdots) - \frac{2\mathcal{R}_k}{1+\mathcal{R}_k^2},$$

where

$$\begin{aligned}\frac{d}{dr_k}(\cdots) :&= \gamma C_{\alpha,\theta}(x_1 r_1 \cdots r_{K-1})^{-1-\alpha}(1 - x_1 r_1 \cdots r_{K-1})^{\theta+\alpha-1}\frac{x_1 r_1 \cdots r_{K-1}}{r_k} \\ &\quad + (m_{0,k+1} + \theta + \alpha - 1)\frac{-x_1 r_1 \dots r_k / r_k}{1 - x_1 r_1 \dots r_k} + \cdots + (m_{0,K} + \theta + \alpha - 1)\frac{-x_1 r_1 \dots r_{K-2} r_{K-1}/r_k}{1 - x_1 r_1 \dots r_{K-2} r_{K-1}} \\ &\quad + (m_{1,k+1} + \cdots + m_{1,K} - (K-k)\alpha - 1)r_k^{-1}.\end{aligned}$$

Finally, the gradient of the posterior is denoted by

$$\mathcal{D}(\mathcal{X}_1, \mathcal{R}_1, \dots \mathcal{R}_{K-1}) := \left(\frac{d}{d\mathcal{X}_1}, \frac{d}{d\mathcal{R}_1}, \cdots, \frac{d}{d\mathcal{R}_{K-1}} \right) \log(\mathbb{P}(\mathcal{X}_1, \mathcal{R}_1, \dots \mathcal{R}_{K-1} \mid \mathbf{Z})).$$

The full steps of the Hamiltonian Monte Carlo method is given in Algorithm 1.

Algorithm 1 Hamiltonian Monte Carlo Algorithm for the TILe-Rep

Require: leapfrog steps $L \geq 1$, step size $\epsilon > 0$, current values $\boldsymbol{v} \leftarrow (x_1, r_1, \dots, r_{K-1})$
Set $W^{(0)} \leftarrow \tan(\pi(\boldsymbol{v} - 0.5))$
Sample $p := (p_1, \dots, p_K) \leftarrow \mathcal{N}(0, I_K)$
Set $\tilde{p}^{(0)} \leftarrow p + (\epsilon/2)\mathcal{D}(W^{(0)})$
for $l = 1$ **to** $L - 1$ **do**
    Set $W^{(l)} \leftarrow W^{(l-1)} + \epsilon\tilde{p}^{(l-1)}$
    Set $\tilde{p}^{(l)} \leftarrow \tilde{p}^{(l-1)} + \epsilon\mathcal{D}(W^{(l)})$
end for
Set $\tilde{W} \leftarrow W^{(L-1)} + \epsilon\tilde{p}^{(L-1)}$
Set $\tilde{p} := (\tilde{p}_1, \dots, \tilde{p}_K) \leftarrow -\{\tilde{p}^{(L-1)} + (\epsilon/2)\mathcal{D}(\tilde{W})\}$
Set $\tilde{\boldsymbol{v}} \leftarrow 0.5 + (1/\pi)\text{atan}(\tilde{W})$
Sample $U \leftarrow U(0, 1)$
Set $S \leftarrow \frac{\mathbb{P}(\tilde{\boldsymbol{v}}|\mathbf{Z},\alpha,\gamma)}{\mathbb{P}(\boldsymbol{v}|\mathbf{Z},\alpha,\gamma)} \exp\left(-\frac{1}{2}\sum_{k=1}^{K}(\tilde{p}_k^2 - p_k^2)\right)$
if $U \leq S$ **then**
    Output $\tilde{\boldsymbol{v}}$
else
    Output $\boldsymbol{v}$
end if

A.6 Derivation of the VI Scheme

When $\alpha > 1$ and $w \in (0, 1)$, we know $(1 - w)^{\alpha-1} < 1$, and

$$\int_x^1 w^{-1}(1 - w)^{\alpha-1} dw < \int_x^1 w^{-1} dw = -\log(x).$$

It follows that

$$\mathbb{E}_{\boldsymbol{v}}\left(-\gamma \int_{v_1 \dots v_K}^1 \alpha w^{-1}(1 - w)^{\alpha-1} dw\right) > \gamma\alpha\mathbb{E}_{\boldsymbol{v}}\left(\log(v_1 \dots v_K)\right).$$

Thus, the expectation $\mathbb{E}_q\left(\log \mathbb{P}(v_1, \dots, v_K)\right)$ has the lower bound

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{v}}\left(\log \mathbb{P}(v_1, \dots, v_K)\right) \\
&\geq (\gamma\alpha - 1)\mathbb{E}_{\boldsymbol{v}}\left(\log(v_1 \dots v_K)\right) + K\log(\gamma) + K\log(\alpha) + (\alpha - 1)\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{v}}(\log(1 - v_1 \dots v_k)).
\end{aligned}$$

Using this inequality, we rewrite the ELBO as

$$\begin{aligned}
\mathcal{L}(q) \geq \mathcal{H}(q) + K\log(\gamma) + K\log(\alpha) + (\gamma\alpha - 1)\mathbb{E}_{\boldsymbol{v}}\left(\log(v_1 \dots v_K)\right) + \sum_{i=1}^{K}\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{v}}\left(z_{ni}\log(v_1 \dots v_i)\right) \\
+ (\alpha - 1)\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{v}}(\log(1 - v_1 \dots v_k)) + \sum_{i=1}^{K}\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{v}}\left((1 - z_{ni})\log(1 - v_1 \dots v_i)\right).
\end{aligned}$$

Denote by $M(\tau_{k1}, \tau_{k2})$ the terms containing τ_{k1} and τ_{k2} on the right hand side of the inequality, we get

$$\begin{aligned}
M(\tau_{k1}, \tau_{k2}) = &(\gamma\alpha - 1)(\psi(\tau_{k1}) - \psi(\tau_{k1} + \tau_{k2})) + \sum_{i=k}^{K}\sum_{n=1}^{N} \nu_{ni}(\psi(\tau_{k1}) - \psi(\tau_{k1} + \tau_{k2})) \\
&+ (\alpha - 1)\sum_{i=k}^{K} q_i(k)\psi(\tau_{k2}) + (\alpha - 1)\sum_{i=k+1}^{K}\sum_{y=k+1}^{i} q_i(y)\psi(\tau_{k1}) \\
&- (\alpha - 1)\sum_{i=k}^{K}\sum_{y=k}^{i} q_i(y)\psi(\tau_{k1} + \tau_{k2}) + \sum_{i=k}^{K}\sum_{n=1}^{N}(1 - \nu_{ni})q_i(k)\psi(\tau_{k2}) \\
&+ \sum_{i=k+1}^{K}\sum_{n=1}^{N}(1 - \nu_{ni})\sum_{y=k+1}^{i} q_i(y)\psi(\tau_{k1}) - \sum_{i=k}^{K}\sum_{n=1}^{N}(1 - \nu_{ni})\sum_{y=k}^{i} q_i(y)\psi(\tau_{k1} + \tau_{k2}) \\
&- \log\left(\frac{\Gamma(\tau_{k1} + \tau_{k2})}{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}\right) - (\tau_{k1} - 1)(\psi(\tau_{k1}) - \psi(\tau_{k1} + \tau_{k2})) - (\tau_{k2} - 1)(\psi(\tau_{k2}) - \psi(\tau_{k1} + \tau_{k2})).
\end{aligned}$$

We rearrange the terms into

$$\begin{aligned}
M(\tau_{k1}, \tau_{k2}) = &\left[(\gamma\alpha - 1) + \sum_{i=k}^{K}\sum_{n=1}^{N}\nu_{ni} + (\alpha - 1)\sum_{i=k+1}^{K}\sum_{y=k+1}^{i} q_i(y) + \sum_{i=k+1}^{K}\sum_{n=1}^{N}(1 - \nu_{ni})\sum_{y=k+1}^{i} q_i(y) - (\tau_{k1} - 1)\right] \\
&\times (\psi(\tau_{k1}) - \psi(\tau_{k1} + \tau_{k2})) \\
&+ \left[(\alpha - 1)\sum_{i=k}^{K} q_i(k) + \sum_{i=k}^{K}\sum_{n=1}^{N}(1 - \nu_{ni})q_i(k) - (\tau_{k2} - 1)\right] \times (\psi(\tau_{k2}) - \psi(\tau_{k1} + \tau_{k2})) \\
&- \log\left(\frac{\Gamma(\tau_{k1} + \tau_{k2})}{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}\right).
\end{aligned}$$

Then, the function $M(\tau_{k1}, \tau_{k2})$ is maximised by setting τ_{k1} and τ_{k2} to be (19), and the variational distribution q is obtained. The full steps of the variational inference scheme is given in Algorithm 2.

Algorithm 2 Variational Inference Scheme for the TILe-Rep

Require: hyperparameter α, γ, initial τ_{k1}, τ_{k2}

1: **for** $k = 1$ **to** K **do**

2: Update $q_k(y)$ according to

$$q_k(y) \propto \exp\left(\psi(\tau_{y2}) + \sum_{m=1}^{y-1}\psi(\tau_{m1}) - \sum_{m=1}^{y}\psi(\tau_{m1} + \tau_{m2})\right).$$

3: Update τ_{k1}, τ_{k2} according to

$$\begin{aligned}
\tau_{k1} =&\gamma\alpha + \sum_{i=k}^{K}\sum_{n=1}^{N}\nu_{ni} + (\alpha - 1)\sum_{i=k+1}^{K}\sum_{y=k+1}^{i} q_i(y) + \sum_{i=k+1}^{K}\sum_{n=1}^{N}(1 - \nu_{ni})\sum_{y=k+1}^{i} q_i(y), \\
\tau_{k2} =&1 + (\alpha - 1)\sum_{i=k}^{K} q_i(k) + \sum_{i=k}^{K}\sum_{n=1}^{N}(1 - \nu_{ni})q_i(k).
\end{aligned}$$

4: **end for**

A.7 Blocked Gibbs Sampler for Binary Latent Feature Model

The full steps of the blocked Gibbs sampler is given in Algorithm 3.

Algorithm 3 Blocked Gibbs Sampler for the Binary Latent Feature Model

Require: Initial $H_K, \mathbf{Z}, \sigma_X, \sigma_\phi$
1: Update H_K: Run the HMC Algorithm
2: **for** $n = 1$ **to** N **and** $k = 1$ **to** K **do**
3: Update z_{nk}: Draw $z_{nk} \sim \mathbb{P}(z_{nk} \mid \mathbf{X}, \mathbf{Z}_{-nk}, H_K, \sigma_X, \sigma_\phi)$
4: **end for**
5: Update σ_X, σ_ϕ: Draw $(\sigma_X, \sigma_\phi) \sim \mathbb{P}(\sigma_X, \sigma_\phi \mid \mathbf{X}, \mathbf{Z})$

A.8 VI Scheme for Binary Latent Feature Model

The full steps of the variational inference scheme is given in Algorithm 4.

Algorithm 4 Variational Inference Scheme for the Binary Latent Feature Model

Require: Initial $\tau_{k1}, \tau_{k2}, \eta_k, \xi_k, \nu_{nk}$
1: **for** $k = 1$ **to** K **do**
2: Update $q_k(y)$ according to

$$q_k(y) \propto \exp\left(\psi(\tau_{y2}) + \sum_{m=1}^{y-1}\psi(\tau_{m1}) - \sum_{m=1}^{y}\psi(\tau_{m1}+\tau_{m2})\right)$$

3: Update τ_{k1}, τ_{k2} according to

$$\begin{aligned}\tau_{k1} =& \gamma\alpha + \sum_{i=k}^{K}\sum_{n=1}^{N}\nu_{ni} + (\alpha-1)\sum_{i=k+1}^{K}\sum_{y=k+1}^{i}q_i(y) + \sum_{i=k+1}^{K}\sum_{n=1}^{N}(1-\nu_{ni})\sum_{y=k+1}^{i}q_i(y),\\ \tau_{k2} =& 1 + (\alpha-1)\sum_{i=k}^{K}q_i(k) + \sum_{i=k}^{K}\sum_{n=1}^{N}(1-\nu_{ni})q_i(k),\end{aligned}$$

4: Update η_k, ξ_k according to $\eta_{kd} = 1/(-2p_{kd})$ and $\xi_{kd} = \eta_{kd}m_{kd}$, where

$$m_{kd} = \frac{\sum_{n=1}^{N}\nu_{nk}(X_{nd} - \sum_{s=\{1,\dots,K\}/k}\nu_{ns}\phi_{sd})}{\sigma_X^2} \text{ and } p_{kd} = -\frac{1}{2}\frac{\sigma_X^2 + \sigma_\phi^2\sum_{n=1}^{N}\nu_{nk}}{\sigma_\phi^2\sigma_X^2}.$$

5: **for** $n = 1$ **to** N **do**
6: Update ν_{nk} according to $\nu_{nk} = 1/(1+\exp(-C))$, where

$$\begin{aligned}C =& \sum_{l=1}^{k}(\psi(\tau_{l1}) - \psi(\tau_{l1}+\tau_{l2})) - \frac{1}{2\sigma_X^2}(\mathrm{tr}(\eta_k) + \xi_k\xi_k^T) + \frac{1}{\sigma_X^2}\xi_k\left(X_n^T - \sum_{i=\{1,\dots,K\}/k}\nu_{ni}\xi_i^T\right)\\ &- \sum_{y=1}^{k}q_k(y)\left(\psi(\tau_{y2}) + \sum_{m=1}^{y-1}\psi(\tau_{m1}) - \sum_{m=1}^{y}\psi(\tau_{m1}+\tau_{m2}) - \log(q_k(y))\right).\end{aligned}$$

7: **end for**
8: **end for**

B ADDITIONAL EXPERIMENTS

In this section, we provide the additional experimental results.

In the first numerical experiment, we use the inverse-Lévy measure method to sample from the 10 largest atom weights of the beta process and simulate 100 independent Bernoulli likelihood processes based on these atom weights. Then, we run the HMC algorithm and the VI scheme to estimate the atom weights.

We present the numerical results of the HMC algorithm in Figure 5. In the figures, the green cross represents the true atom weights, the red plus stands for the proportions of the Bernoulli random variables that equal to 1, i.e., $m_{1,k}/(m_{1,k}+m_{0,k})$, and the box plot denotes the posterior atom weights. The results are based on 1000 iterations of the HMC algorithm following an initial 1000 iterations burn-in. We find that the HMC algorithm can recover the atom weights correctly.

The numerical results of the VI scheme are presented in Figure 6. In the figures, the green cross and red plus have the same meanings as before, and the blue interval denotes the $[0.25, 0.75]$ quantile of the variational distribution. The figures suggest that the VI scheme can approximate the posterior accurately.

In the second numerical illustration, we input an empty observation ($m_{1,k} = m_{0,k} = 0$, for $k = 1, \ldots, K$) to the posterior. In this case, the posterior reduces to the joint distribution (10), and the HMC algorithm samples directly from the K largest atom weights of the beta process. We present the sample averages of the atom weights in Table 1. We also use the Monte Caro method to estimate the expectations of the atom weights. The numerical results show that the HMC algorithm can draw from the beta process accurately.

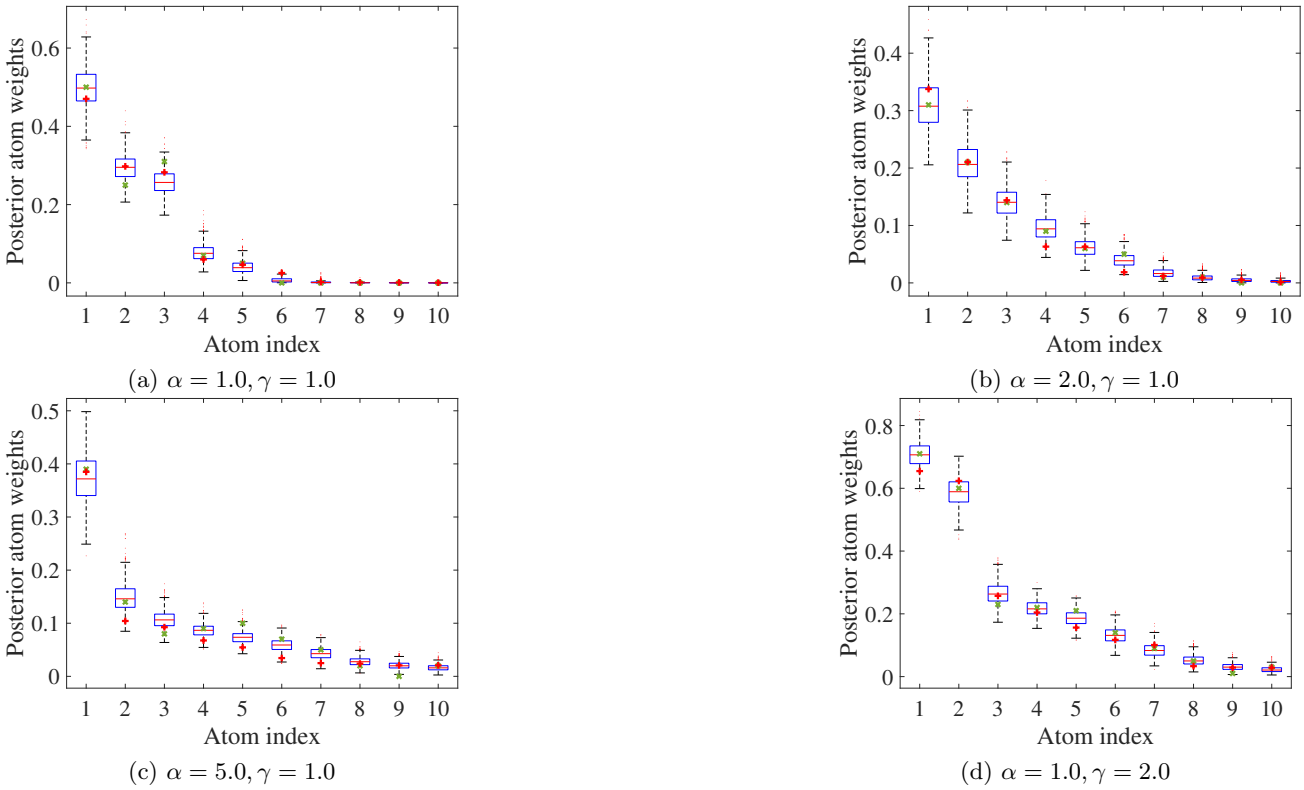
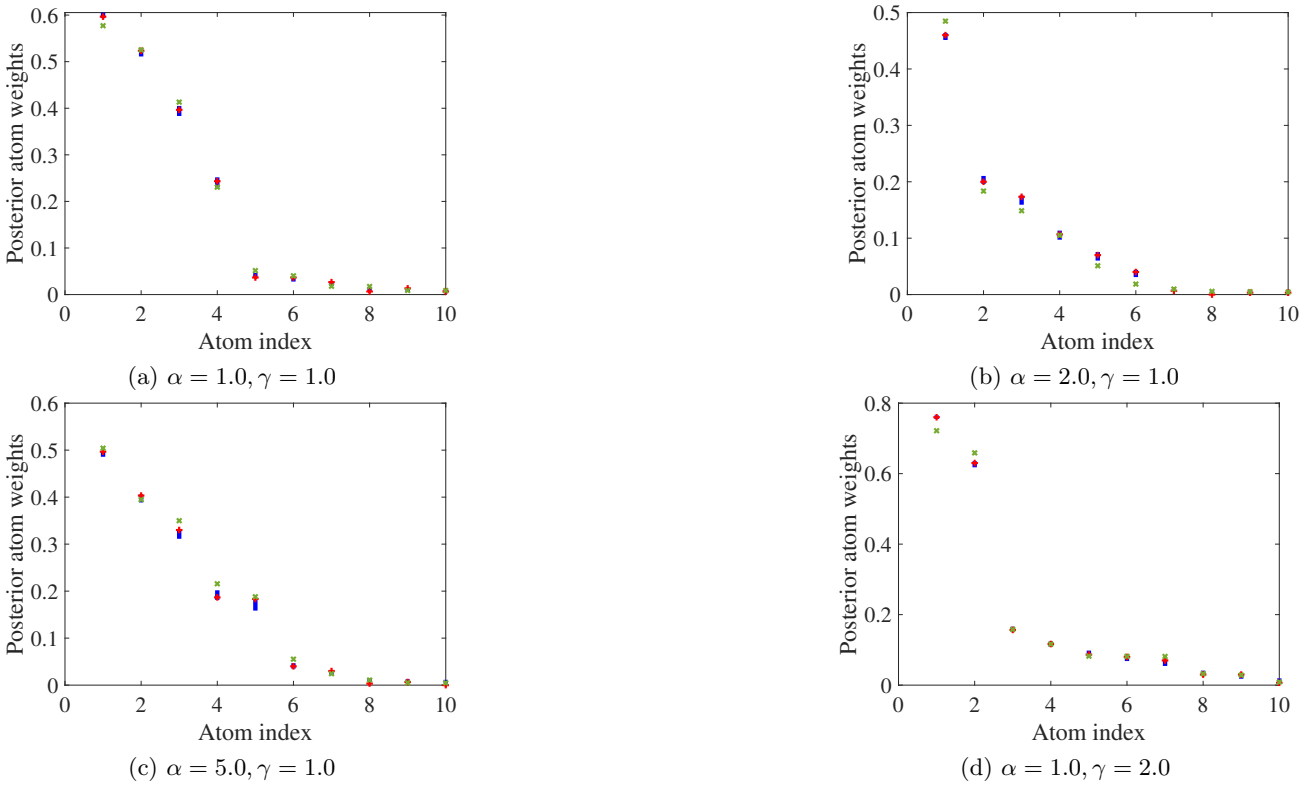


Figure 5: Posterior Atom Weights of the HMC Algorithm



(a) $\alpha = 1.0, \gamma = 1.0$ (b) $\alpha = 2.0, \gamma = 1.0$

(c) $\alpha = 5.0, \gamma = 1.0$ (d) $\alpha = 1.0, \gamma = 2.0$

Figure 6: Posterior Atom Weights of the VI Scheme

Table 1: Sample Averages of the Atom Weights

	Algm.	J_1	J_2	J_3	J_4	J_5
$\alpha = 1.0$	MC	0.50	0.25	0.13	0.06	0.03
$\gamma = 1.0$	HMC	0.48	0.27	0.12	0.06	0.03
$\alpha = 2.0$	MC	0.40	0.23	0.14	0.09	0.05
$\gamma = 1.0$	HMC	0.41	0.23	0.14	0.09	0.06
$\alpha = 5.0$	MC	0.27	0.17	0.12	0.09	0.07
$\gamma = 1.0$	HMC	0.28	0.17	0.12	0.09	0.07
$\alpha = 1.0$	MC	0.67	0.45	0.30	0.19	0.13
$\gamma = 2.0$	HMC	0.66	0.46	0.29	0.19	0.12