Non-exchangeable feature allocation models with sublinear growth of the feature sizes

Giuseppe Di Benedetto
University of Oxford

François Caron
University of Oxford

Yee Whye Teh
University of Oxford
DeepMind

Abstract

Feature allocation models are popular models used in different applications such as unsupervised learning or network modeling. In particular, the Indian buffet process is a flexible and simple one-parameter feature allocation model where the number of features grows unboundedly with the number of objects. The Indian buffet process, like most feature allocation models, satisfies a symmetry property of exchangeability: the distribution is invariant under permutation of the objects. While this property is desirable in some cases, it has some strong implications. Importantly, the number of objects sharing a particular feature grows linearly with the number of objects. In this article, we describe a class of non-exchangeable feature allocation models where the number of objects sharing a given feature grows sublinearly, where the rate can be controlled by a tuning parameter. We derive the asymptotic properties of the model, and show that such model provides a better fit and better predictive performances on various datasets.

1 Introduction

Feature allocation models are probabilistic models over multisets (Broderick et al., 2013), which represent the allocation of a set of objects to a (potentially unbounded) set of features. Contrary to models on partitions, where each object is assigned to a single group, these models allow each object to be allocated more than one feature. For example, the objects may be movies, and the features correspond to the actors performing in that movie. Movies have different number of actors, and actors participate to a different number of movies. Informally, feature allocations models can be interpreted as distributions over sparse binary matrices whose rows represent the objects and whose columns represent the features; non-zero entries indicate the allocation of features to objects. Feature allocation models have been used in various applications, including topic modeling (Williamson et al., 2010b), image analysis (Zhou et al., 2011a), network modeling (Palla et al., 2012; Cai et al., 2016) or inference in tumor heterogeneity (Lee et al., 2015; Xu et al., 2015).

A classical assumption is that of exchangeability: the feature allocation model is invariant over permutations of the objects. Taking the interpretation as a binary matrix, the matrix is invariant over permutations of its rows. The most remarkable example of an exchangeable feature allocation is the Indian buffet process (IBP) (Ghahramani and Griffiths, 2006; Thibaux and Jordan, 2007; Griffiths and Ghahramani, 2011) which has a simple and intuitive generative model. The model allows the number of features K_n to grow unboundedly with the number of objects n at a logarithmic rate. The IBP admits a three-parameter generalisation, the stable IBP (Teh and Gorur, 2009), which is also exchangeable. For some values of its parameters, the stable IBP can capture a power-law behavior, where the number of features grows at a rate of n^σ for some $\sigma \in (0, 1)$.

While the exchangeability assumption is reasonable for many applications and has computational advantages, it may not be adequate in some cases. In particular, assuming exchangeability implies that, out of n objects, the number $m_{n,j} \leq n$ of objects having a particular feature j (called feature's size) scales linearly with the number of objects n (see Figure 4(b) for an illustration). Such assumption may be undesirable. For instance, even if she's a prolific actress, one does not expect the filmography of Meryl Streep to scale linearly with the overall number of movies released.

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

The objective of this article is to present a model that can have a sublinear growth of the size of the features, while retaining the properties of the (stable) Indian buffet in terms of overall growth of the number of features and power-law properties. The article is organised as follows. Section 2 provides some background on feature allocation models and the IBP. Our non-exchangeable model is presented in Section 3 and its asymptotic properties given in Section 4. In Section 5 we derive a Gibbs sampler for posterior inference. Experimental results are presented in Section 6. Related approaches are discussed in Section 7.

2 Background

2.1 Feature allocation models and the Indian buffet process

A feature allocation (Broderick et al., 2012) $f_n = \{A_{n,1}, \ldots, A_{n,K_n}\}$ is a multiset of a set of objects $\{1, \ldots, n\}$ such that $A_{n,k}$, $k = 1, \ldots, K_n$ are (possibly overlapping) non-empty subsets of $\{1, \ldots, n\}$; $A_{n,k}$ represents the set of objects having feature k and K_n is the number of different features shared by the n objects. For example, $f_4 = \{\{1,3\},\{1,3\},\{3,4\},\{4\},\{4\}\}$ indicates that object 1 has features 1 and 2, object 2 has no feature, object 3 has features 1, 2, 3 and object 4 has features 3, 4 and 5. Note that the labelling of the features as 1 to 5 is arbitrary.

A feature allocation model is a distribution over a growing family of random feature allocations $(f_n)_{n=1,2,\ldots}$. The most popular feature allocation model is the Indian buffet process, where f_n has distribution proportional to

$$\Pr(f_n) \propto \eta^{K_n} \prod_{k=1}^{K_n} \frac{(\widetilde{m}_{n,k}-1)!(n-\widetilde{m}_{n,k})!}{n!} \qquad (1)$$

where $\widetilde{m}_{n,k}$ is the number of objects having feature k, for $k = 1, \ldots, K_n$ and $\eta > 0$ is a tuning parameter. The IBP admits a three-parameter generalisation, called stable IBP (Teh and Gorur, 2009), where[1]

$$\Pr(f_n) \propto \frac{\eta^{K_n}}{\Gamma(1-\sigma)^{K_n}} \prod_{k=1}^{K_n} \frac{\Gamma(\widetilde{m}_{n,k}-\sigma)\Gamma(n-\widetilde{m}_{n,k}+\zeta)}{\Gamma(n+\zeta-\sigma)} \qquad (2)$$

with $\eta > 0$, $\sigma \in (-\infty, 1)$ and $\zeta > 0$. It reduces to the one-parameter IBP when $\zeta = 1$ and $\sigma = 0$. When $\sigma > 0$, the model exhibits power-law properties.

A convenient way to encode a feature allocation model is via a collection of atomic random measures $(Z_1, \ldots, Z_n)$ on some space (here $\mathbb{R}_+$) where for $i = 1, \ldots, n$,

$$Z_i = \sum_{j \geq 1} z_{ij} \delta_{\theta_j} \qquad (3)$$

with $z_{ij} = 1$ if object i has feature j and $(\theta_j)_{j\geq 1}$ are continuous random variables on $\mathbb{R}_+$ whose distribution is irrelevant here. Note that in this notation there is no particular ordering of the features, and we now use the index j instead of k to emphasize this difference.

2.2 Completely random measures

A homogeneous completely random measure (CRM) (Kingman, 1967; Lijoi and Prünster, 2010) on $\mathbb{R}_+ = [0, \infty)$ is an almost surely discrete random measure

$$B = \sum_{j \geq 1} \omega_j \delta_{\theta_j} \qquad (4)$$

where $\{(\omega_j, \theta_j)_{j\geq 1}\}$ are points of a Poisson point process on $(0, \infty) \times \mathbb{R}_+$ with mean measure $\nu(d\omega, d\theta) = \rho(\omega)d\omega\alpha(\theta)d\theta$ where ρ and α satisfy

$$\int_0^\infty (1 - e^{-\omega})\rho(\omega)d\omega < \infty \text{ and } \int_A \alpha(\theta)d\theta < \infty$$

for any bounded set $A \subset \mathbb{R}_+$. The above condition ensures that $B(A) < \infty$ almost surely. The CRM is said to be infinite-activity if $\int_0^\infty \rho(\omega)d\omega = \infty$. In this case, the measure B has a countably infinite support on any non-empty interval $A \subset \mathbb{R}_+$.

The (stable) Indian buffet process admits the following hierarchical construction via CRMs (Thibaux and Jordan, 2007; Teh and Gorur, 2009). Let $\widetilde{B} = \sum_j \pi_j \delta_{\theta_j}$ be a homogeneous CRM with

$$\rho(\pi) = \frac{\eta}{\Gamma(1-\sigma)} \pi^{-1-\sigma}(1-\pi)^{\zeta-1}\, 1_{\pi\in(0,1)} \qquad (5)$$

and $\alpha(\theta) = 1_{\theta\leq 1}$. The parameter π_j can be interpreted as the popularity of feature j. For $i = 1, \ldots, n$ define the atomic measure Z_i as in Equation (3) with

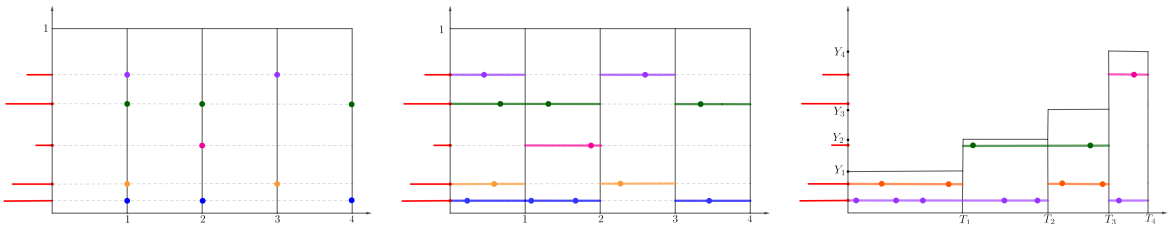
$$z_{ij} \mid \pi_j \sim \text{Ber}(\pi_j)$$

for $j \geq 1$, where $\text{Ber}(\pi)$ denotes the Bernoulli distribution with parameter $\pi \in [0, 1]$.

2.3 An alternative construction for the Indian buffet process

We present here an equivalent construction for the IBP. It relies on the introduction of latent Poisson processes, similar to the construction proposed by Di Benedetto et al. (2017) for non-exchangeable random partitions, adapted to feature models. This approach will give the intuition for the generalisation of the IBP introduced in the next section.

[1] Note that we use a slightly different parameterisation compared to that of Teh and Gorur (2009).



(a) Beta-Bernoulli process (b) Continuous time IBP (c) Non-exchangeable point process

Figure 1: Point process defining the feature models: red sticks represent weights of the CRM, points and bars of the same colour refer to the same feature. (a) Beta-Bernoulli process; (b) Point process construction of the IBP: each rectangle of basis $[T_{i-1}, T_i]$ represents the $i-th$ object whose features are depicted by the coloured bars, which are present if at least one of the corresponding points falls in the rectangle; (c) Non-exchangeable point process.

Using the change of variable $\omega_j = -\log(1-\pi_j) \in (0,\infty)$, the random measure $B = \sum_j \omega_j \delta_{\theta_j}$ is itself a CRM with $\alpha(\theta) = 1_{\theta \leq 1}$ and

$$\rho(\omega) = \frac{\eta}{\Gamma(1-\sigma)} e^{-\zeta\omega}(1-e^{-\omega})^{-1-\sigma}. \tag{6}$$

We call (6) a transformed stable beta (TSB) Lévy measure, and the associated random measure B a TSB process (TSBP). Consider for each j a homogeneous Poisson process $N_j(t)$ on $\mathbb{R}_+$ with rate ω_j. Let $z_{ij} = 1_{N_j(i)-N_j(i-1)>0}$ be a binary variable indicating if there is any event in the time interval $[i-1, i)$. Then

$$\begin{aligned}\Pr(z_{ij} = 1 \mid \omega_j) &= \Pr(N_j(i) - N_j(i-1) > 0 \mid \omega_j) \\ &= 1 - e^{-\omega_j} = \pi_j.\end{aligned}$$

This construction is illustrated in Figure 1.

3 Non-exchangeable feature allocation model

Let $B = \sum_{j\geq 1} \omega_j \delta_{\theta_j}$ be a homogeneous completely random measure on $\mathbb{R}_+$. While the model can be defined for a general CRM, we focus here on the case where $\alpha(\theta) = 1$ and ρ is either the transformed stable beta measure (6), or the generalised gamma (GG) measure (Hougaard, 1986), given by

$$\rho(\omega) = \frac{\eta}{\Gamma(1-\sigma)} \omega^{-1-\sigma} e^{-\zeta\omega} 1_{\omega>0} \tag{7}$$

where $\eta > 0$, $\sigma \in (-\infty, 1)$, $\zeta > 0$. As we will show in Section 4, both models lead to the same asymptotic behavior. The GG measure has however a conjugate form that makes it more amenable to posterior inference, as detailed in Section 5. A CRM with mean measure (7) will be called a Generalised Gamma Process (GGP).

Inspired by the work on random partition models by Di Benedetto et al. (2017), let $(Y_n)_{n\geq 1}$ and $(T_n)_{n\geq 1}$ be two increasing sequences of positive reals defined as

$$T_n = n^{\frac{1}{\xi+1}}, \qquad Y_n = n^{\frac{\xi}{\xi+1}} \tag{8}$$

where $\xi \geq 0$ is a tuning parameter. Define the sequence $(\Delta_n)_{n\geq 1}$ by $\Delta_n := T_n - T_{n-1}$. The feature allocation of an object i is represented by a random measure Z_i on $\mathbb{R}_+$ as in Equation (3) where

$$z_{ij} \mid \omega_j, \theta_j \sim \text{Ber}\left(1 - e^{-\omega_j \Delta_i 1_{\theta_j \leq Y_i}}\right). \tag{9}$$

Note that $z_{ij} = 0$ a.s. if $\theta_j > Y_i$.

The model admits the following construction using a latent Poisson process. For each j, let $(N_j(t))_{t>0}$ be a homogeneous Poisson process with rate ω_j. Then the binary variable $z_{ij} = 1_{N_j(T_i)-N_j(T_{i-1})>0} 1_{\theta_i \leq Y_i}$ has distribution (9). See the illustration in Figure 1(c).

Note that if one sets $\xi = 0$, we have $\Delta_i = Y_i = 1$. The distribution in the right-handside of Equation (9) does not depend on i and the associated feature allocation model is therefore exchangeable. If additionally we use the mean measure ρ as in Equation (6), we recover the three-parameter IBP as a special case.

The model is parameterised by the four parameters η, σ, ζ and ξ. In the next section we show how these parameters tune the asymptotic properties of the model. The critical parameter is the parameter ξ and we show that for $\xi > 0$, the features' sizes grow sublinearly with n, at a rate controlled by this parameter.

4 Asymptotic Properties

4.1 Notations

In this section we use the following notations for asymptotics. $a_n \overset{n\to\infty}{\sim} b_n$ means that $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$ and $a_n \overset{n\to\infty}{\asymp} b_n$ means $\limsup_{n\to\infty} \frac{a_n}{b_n} < \infty$ and $\limsup_{n\to\infty} \frac{b_n}{a_n} < \infty$ (that is, a_n is of the same order as b_n). Let

$$Z_i(\mathbb{R}_+) = \sum_{j\geq 1} z_{ij}$$

be the number of features for object i. For each feature $j = 1, 2, \ldots$, let us define its size as

$$m_{n,j} = \sum_{i=1}^{n} z_{ij}$$

i.e. the number of objects having that feature, and let

$$m_n = \sum_{j\geq 1} m_{n,j} = \sum_{i=1}^{n}\sum_{j\geq 1} z_{ij}$$

be the total number of features allocated to the n objects. Denote

$$K_n = \sum_j 1_{m_{n,j}>0}$$

the number of unique features, and

$$K_{n,r} = \sum_j 1_{m_{n,j}=r}$$

the number of unique features allocated to r objects, where $r \geq 1$. Note that $\sum_r K_{n,r} = K_n$.

We will use the following notation for the Laplace exponent and the tilted moments

$$\psi(t) = \int_0^\infty \left\{1 - e^{-\omega t}\right\} \rho(\omega) d\omega$$
$$\kappa(m, u) = \int_0^\infty \omega^m e^{-u\omega} \rho(\omega) d\omega.$$

for any integer $m \geq 1$ and any $u > 0$. $\kappa(m, 0)$ correspond to the m-th moment of the measure ρ. For the GG measure, we have $\psi(t) = \frac{\eta}{\sigma}((t+\zeta)^\sigma - \zeta^\sigma)$ and $\kappa(m, u) = \eta\frac{\Gamma(m-\sigma)}{\Gamma(1-\sigma)}(u+\zeta)^{\sigma-m}$. For the TSB, we have $\psi(1) = \frac{\eta\Gamma(\zeta)}{\Gamma(\zeta+1-\sigma)}$, but there is no analytical expression for $\psi(t)$ and $\kappa(m, u)$.

4.2 Asymptotic properties when $\xi = 0$

We first recall here, for comparison, the asymptotic properties of our model for $\xi = 0$ when B is a GGP or a TSBP. As mentioned in Section 3, the model is exchangeable in that case and if B is the TSBP, it reduces to the three parameter IBP, whose asymptotic properties are well known (Ghahramani and Griffiths, 2006; Thibaux and Jordan, 2007; Teh and Gorur, 2009; Broderick et al., 2012). Asymptotics when B is the GGP model are similar. First, for the number of features of object i,

$$Z_i(\mathbb{R}_+) \sim \text{Poisson}\left(\psi(1)\right) \tag{10}$$

where $\psi(1) = \frac{\eta\Gamma(\zeta)}{\Gamma(\zeta+1-\sigma)}$ for the TSBP and $\psi(1) = \frac{\eta}{\sigma}((\zeta+1)^\sigma - \zeta^\sigma)$ for the GGP.

Given the CRM B, we have almost surely

$$m_{n,j} \overset{n\to\infty}{\sim} n(1 - e^{-\omega_j}) \tag{11}$$
$$m_n \overset{n\to\infty}{\sim} n\sum_{j\geq 1}(1 - e^{-\omega_j}). \tag{12}$$

Hence both the features' sizes and the total number of features grow linearly with n, as a consequence of the exchangeability assumptions. For the number of unique features, we have almost surely

$$K_n \overset{n\to\infty}{\sim} \begin{cases} \eta\log(n) & \sigma = 0 \\ \frac{\eta}{\sigma} n^\sigma & \sigma \in (0,1) \end{cases}. \tag{13}$$

Finally we have, for the proportions of features allocated to r objects for $\sigma \in (0, 1)$

$$\frac{K_{n,r}}{K_n} \to \frac{\sigma\Gamma(r-\sigma)}{r!\Gamma(1-\sigma)} \tag{14}$$

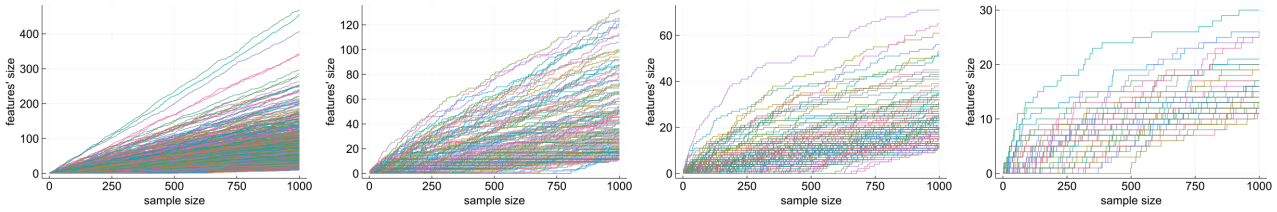
almost surely as n tends to infinity, and $K_{n,r}/K_n \to 0$ almost surely otherwise for all $r \geq 1$. Equation (14) corresponds to a power-law behaviour as $\Gamma(r-\sigma)/r! \simeq r^{-(1+\sigma)}$ for large r.

4.3 Asymptotic properties when $\xi > 0$

When $\xi > 0$, the model is non-exchangeable. In this case, most of the properties of the exchangeable case are retained, such as the Poisson number of features per object (see Proposition 1 below), the linear growth of the total number of features (Proposition 2) and the power-law behaviour, solely controlled by the parameter σ (Proposition 5). The number of unique features K_n still grows sublinearly, but at a rate now controlled by both σ and ξ (Proposition 4). The key difference is that, as shown in Proposition 3, the features' sizes grow at a rate $n^{1/(1+\xi)}$, which is sublinear for $\xi > 0$.

Propositions 1, 2 and 3 are valid for any choice of Lévy measure ρ. Propositions 4 and 5 hold for any Lévy measure ρ such that

$$\psi(t) \overset{t\to\infty}{\sim} \begin{cases} \eta\log(t) & \sigma = 0 \\ \frac{\eta}{\sigma} t^\sigma & \sigma \in (0,1) \end{cases}. \tag{15}$$



(a) IBP (b) $\xi = 0.25$ (c) $\xi = 1$ (d) $\xi = 2$

Figure 2: Evolution of the features' sizes with the sample size for different synthetic datasets: (a) IBP; (b-d) Non-exchangeable model for $\xi = 0.25, 1, 2$. The feature size growth is (a) n, (b) $n^{0.8}$, (c) $n^{1/2}$ and (d) $n^{1/3}$.

Equation (15) holds in particular for the GGP and the TSBP. The proofs are given in the Supplementary Material.

The first proposition shows that the number of features per object is Poisson distributed, with a rate converging to a constant.

Proposition 1 (number of features per object)**.** *Consider the model of Section 3 with a generic ρ and $\xi > 0$. For each object n, we have*

$$Z_n(\mathbb{R}_+) \sim \text{Poisson}\left(\mathbb{E}\left[Z_n(\mathbb{R}_+)\right]\right) \tag{16}$$

where

$$\mathbb{E}\left[Z_n(\mathbb{R}_+)\right] = Y_n \psi(\Delta_n) \to (1+\xi)^{-1}\kappa(1,0),$$

with $\kappa(1,0) = \eta\zeta^{\sigma-1}$ for the GGP.

The next result is on the asymptotic behaviour of the total number of features observed in the first n objects.

Proposition 2 (total number of features)**.** *Consider the model of Section 3 with a generic ρ and $\xi > 0$. The total number of features observed in the first n objects satisfies*

$$m_n \overset{n\to\infty}{\sim} (1+\xi)^{-1}\kappa(1,0)\, n.$$

The following proposition describes the growth of the features' sizes $m_{n,j}$ with respect to n.

Proposition 3 (number of objects per feature)**.** *Consider the model of Section 3 with a generic Lévy measure ρ and $\xi \geq 0$. For each $j \geq 1$ and conditionally on the CRM B, we have that, almost surely,*

$$m_{n,j} \overset{n\to\infty}{\sim} \omega_j T_n = \omega_j n^{1/(1+\xi)}.$$

The features' sizes therefore grow linearly if $\xi = 0$, and sublinearly when $\xi > 0$, with a rate decreasing at ξ increases. This is illustrated in Figure 2.

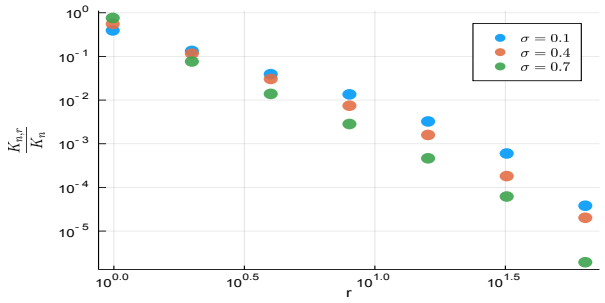


Figure 3: Log-log plot of the proportions of features shared by a fixed number of objects. Simulated examples for $\sigma \in \{0.1, 0.4, 0.7\}$, with $\xi = 1$ and $\eta = 30$.

Proposition 4 (number of unique features)**.** *Consider the model of Section 3 with $\xi \geq 0$ and a Lévy measure ρ whose Laplace exponent ψ satisfies Equation* (15)*. The number of unique features K_n is such that*

$$K_n \overset{n\to\infty}{\sim} \begin{cases} \eta\, n^{\frac{\xi}{\xi+1}} \log(n) & \sigma = 0 \\ \eta\, \frac{\Gamma(\xi+1)\Gamma(\sigma)}{\Gamma(\sigma+\xi+1)}\, n^{\frac{\xi+\sigma}{\xi+1}} & \sigma \in (0,1) \end{cases}. \tag{17}$$

Analogously to the three-parameter IBP model, the proposed non-exchangeable feature allocation model exhibits the power-law property for the number of features shared by a given number of objects (see Figure 3 for an illustration of this property).

Proposition 5 (power-law properties)**.** *Consider the model of Section 3 with $\xi > 1 - \sigma$ and a Lévy measure ρ whose Laplace exponent ψ satisfies Equation* (15)*. We have, almost surely as n tends to infinity*

$$\frac{K_{n,r}}{K_n} \to \frac{\sigma\Gamma(r-\sigma)}{r!\Gamma(1-\sigma)}.$$

We conjecture that the condition $\xi > 1-\sigma$ introduced in the above proposition, which is needed for the proof, is not a necessary condition, and that the above proposition actually holds for any $\xi \geq 0$.

5 Inference

5.1 Data augmentation and conditional distribution

For posterior inference, we introduce a latent process, similar to that of Caron (2012). For $i = 1, \dots, n$, let $U_i = \sum_{j \geq 1} u_{ij} \delta_{\theta_j}$ where $u_{ij} = 1$ if $z_{ij} = 0$, and

$$u_{ij} \mid z_{ij} = 1, \omega_j \sim \text{tExp}(\Delta_i \omega_j, 1) \tag{18}$$

$\text{tExp}(\lambda, T)$ denotes the right-truncated exponential distribution with rate $\lambda > 0$ and truncation $T > 0$ with probability density function $\lambda e^{-\lambda x}(1 - e^{-T\lambda})^{-1} 1_{x<T}$. Note that by construction, $z_{ij} = 1_{u_{ij}<1}$ is a deterministic function of u_{ij}.

Denote $\widetilde{\theta}_1, \dots, \widetilde{\theta}_{K_n}$ the set of θ_j's such that $m_{n,j} > 0$, $\widetilde{\omega}_1, \dots, \widetilde{\omega}_{K_n}$ the corresponding weights, $(\widetilde{u}_{i,k})_{i=1,\dots,n;k=1,\dots,K_n}$ the corresponding latent variables, $(\widetilde{z}_{i,k})_{i=1,\dots,n;k=1,\dots,K_n}$ the corresponding binary variables and $\widetilde{m}_{n,1}, \dots, \widetilde{m}_{n,K_n}$ the associated features' sizes. For $k = 1, \dots, K_n$, let $Y_k^* = \inf_i\{Y_i : \widetilde{z}_{ik} = 1\}$. With analogous calculations to Caron (2012) it is possible to write down the conditional distributions

$$\begin{aligned}
&\mathbb{P}(U_1, \dots, U_n \,|\, B) \\
&= e^{-\sum_{i=1}^n \Delta_i \left(\bar{B}(Y_i) - \sum_{k=1}^{K_n} \widetilde{\omega}_k 1_{\widetilde{\theta}_k < Y_i}\right)} \left(\prod_{i=1}^n \Delta_i^{\sum_{k=1}^{K_n} \widetilde{z}_{ik}}\right) \\
&\left(\prod_{k=1}^{K_n} \widetilde{\omega}_k^{\widetilde{m}_{n,k}} e^{-\widetilde{\omega}_k \sum_{i=1}^n \Delta_i \widetilde{u}_{ik} 1_{\widetilde{\theta}_k < Y_i}} 1_{\widetilde{\theta}_k < Y_k^*}\right)
\end{aligned}$$

where $\bar{B}(Y_i) = B((0, Y_i)) = \sum_j \omega_j 1_{\theta_j < Y_i}$. Using the results on Poisson partition calculus (James, 2002), we can marginalize out the CRM

$$\begin{aligned}
\mathbb{P}(U_1, \dots, U_n) = e^{-\int_0^{Y_n} \psi(f_n(\theta)) d\theta} \left(\prod_{i=1}^n \Delta_i^{\sum_{k=1}^{K_n} \widetilde{z}_{ik}}\right) \\
\times \prod_{k=1}^{K_n} \kappa\left(\widetilde{m}_{n,k}, \sum_{i=1}^n \Delta_i \widetilde{u}_{ik} 1_{\widetilde{\theta}_k < Y_i}\right) 1_{\widetilde{\theta}_k < Y_k^*}
\end{aligned} \tag{19}$$

with $f_n(\theta) := \sum_{i=1}^n \Delta_i 1_{\theta \leq Y_i}$. The conditional distribution is

$$B \,|\, (\widetilde{u}_{ik}) \stackrel{d}{=} B^* + \sum_{k=1}^{K_n} \widetilde{\omega}_k \delta_{\widetilde{\theta}_k} \tag{20}$$

where B^* is an inhomogeneous CRM with Lévy measure $\nu^*(d\omega, d\theta) = e^{-\omega \sum_{i=1}^n \Delta_i 1_{\theta < Y_i}} \rho(\omega) d\omega d\theta$, and it is independent of $(\widetilde{\omega}_k, \widetilde{\theta}_k)_{k=1,\dots,K_n}$ which are independent (in k), with joint posterior probability density

$$p(\widetilde{\omega}_k, \widetilde{\theta}_k | \, (\widetilde{u}_{ik})) \propto \widetilde{\omega}_k^{\widetilde{m}_{n,k}} e^{-\widetilde{\omega}_k \sum_{i=1}^n \Delta_i \widetilde{u}_{ik} 1_{\widetilde{\theta}_k < Y_i}} \rho(\widetilde{\omega}_k) 1_{\widetilde{\theta}_k < Y_k^*}.$$

5.2 Gibbs sampler

Assume that we have observed the binary feature allocations $(\widetilde{z}_{i,k})_{i=1,\dots,n,k=1,\dots,K_n}$. We take an empirical Bayes approach, and wish to approximate the posterior distribution

$$p((\widetilde{u}_{i,k}), (\widetilde{\omega}_k), (\widetilde{\theta}_k) \mid (\widetilde{z}_{i,k}), \widehat{\phi}) \tag{21}$$

where $\widehat{\phi}$ is a point estimate of the hyperparameters $\phi = (\xi, \sigma, \eta, \zeta)$. We explain in the next section how to obtain consistent estimators of the hyperparameters using our asymptotic results. For the GGP, a Gibbs sampler with target distribution (21) can be derived as follows

- For $i = 1, \dots, n$, $k = 1, \dots, K_n$, such that $\widetilde{z}_{i,k} = 1$ sample $\widetilde{u}_{ik} \mid \text{rest} \sim \text{tExp}(\Delta_i \widetilde{\omega}_k, 1)$.
- For $k = 1, \dots, K_n$, sample

$$\widetilde{\omega}_k \mid \text{rest} \sim \text{Gamma}\left(\widetilde{m}_{n,k} - \sigma, \zeta + \sum_{i=1}^n \Delta_i \widetilde{u}_{ik} 1_{\widetilde{\theta}_k < Y_i}\right).$$

- For $k = 1, \dots, K_n$, sample

$$\widetilde{\theta}_k \mid \text{rest} \sim e^{-\widetilde{\omega}_k \sum_{i=1}^n \Delta_i \widetilde{u}_{ik} 1_{\widetilde{\theta}_k < Y_i}} 1_{\widetilde{\theta}_k < Y_k^*}.$$

The last distribution is piecewise constant, and one can therefore straightforwardly sample from it.

5.3 Estimation of the Hyperparameters

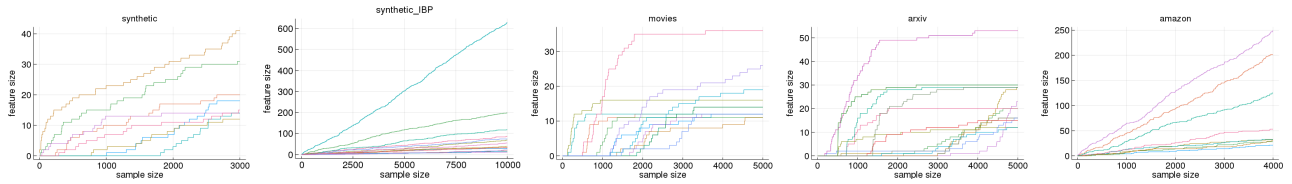
We use here the asymptotic results of Section 4 to derive consistent estimators of the hyperparameters. The parameter ξ controls the features' growth and is consistently estimated using Proposition 3, while Proposition 5 can be used to consistently estimate the parameter σ from the proportion of features of size one:

$$\hat{\xi} = \frac{\log n}{\log \max_{k=1,\dots,K_n} \widetilde{m}_{n,k}} - 1, \qquad \hat{\sigma} = \frac{K_{n,1}}{K_n}.$$

Using Propositions 2 and 4 we have the following consistent estimators for the parameters η and ζ (GGP):

$$\hat{\eta} = \frac{\Gamma(\hat{\sigma} + \hat{\xi} + 1)}{\Gamma(\hat{\sigma})\Gamma(\hat{\xi} + 1)} \frac{K_n}{n^{\frac{\hat{\xi}+\hat{\sigma}}{1+\hat{\sigma}}}}, \qquad \hat{\zeta} = \left(\frac{(\hat{\xi} + 1)\, m_n}{\hat{\eta}\, n}\right)^{\frac{1}{\hat{\sigma}+1}}.$$

Although these estimators are consistent, it should be noted that $\hat{\xi}$ depends logarithmically on the sample size, leading potentially to a slow convergence. However, as shown in the simulated experiments (see section 6 and Table 1), the estimated values of the hyperparamenters ξ and σ are close to the true values.



(a) Synthetic - NE (b) Synthetic - IBP (c) Movies (d) ArXiv (e) Amazon

Figure 4: Evolution of some of features' sizes with respect to the sample size for different datasets.

	Non-exchangeable			Indian Buffet process		
	$\hat{\sigma}_{\text{NE}}$	L2 error	90% CI	$\hat{\sigma}_{\text{IBP}}$	L2 error	90% CI
Synthetic	0.61	15.41	$[15.32, 15.50]$	0.51	19.93	$[19.72, 20.13]$
Synthetic IBP	0.41	30.47	$[30.37, 30.54]$	0.39	31.19	$[31.11, 31.28]$
IMDb movies	0.31	21.27	$[21.20, 21.35]$	0.07	36.78	$[36.59, 37.04]$
Arxiv	0.48	7.96	$[7.89, 7.96]$	0.31	13.95	$[13.81, 14.08]$
Amazon	0.57	142.06	$[141.74, 142.38]$	0.61	142.62	$[142.25, 143.02]$

Table 1: Estimated values for the parameter σ obtained by the Indian Buffet process ($\hat{\sigma}_{\text{IBP}}$) and our non-exchangeable model ($\hat{\sigma}_{\text{NE}}$). Mean and quantiles of the L_2 predictive error.

6 Experiments

In this section we present some experiments to compare the proposed model with GG measure against the three-parameter IBP. We estimate the parameters on a training set, and compute the empirical normalized L_2 error between the true and predicted features' sizes on a test set

$$\text{Err} = \frac{1}{n_{\text{test}}} \sum_{k=1}^{K_{n_{\text{train}}}} \sum_{i=n_{\text{train}}+1}^{n_{\text{train}}+n_{\text{test}}} (\widetilde{m}_{i,k}^{\text{true}} - \widetilde{m}_{i,k})^2$$

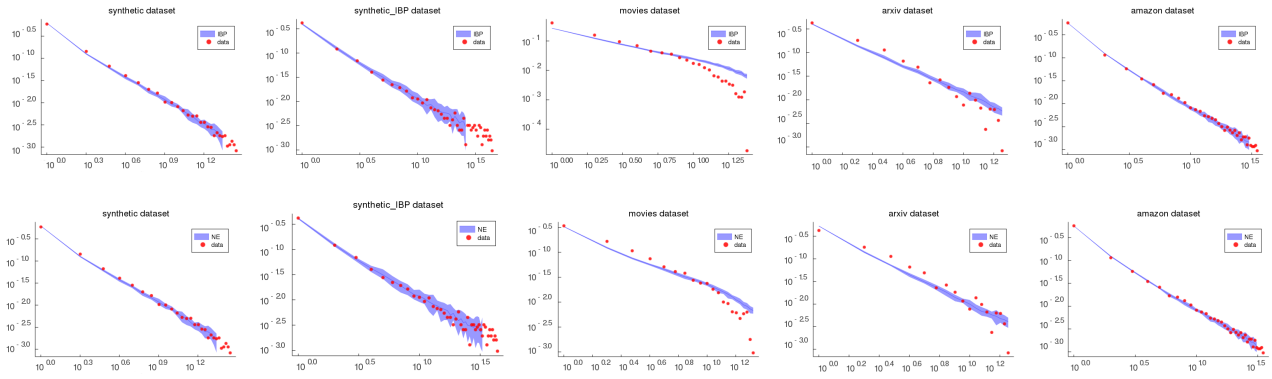
where $\widetilde{m}_{i,k}^{\text{true}}$ is the observed size of feature k, n_{train} and n_{test} are the numbers of objects in the training set and test set respectively and $K_{n_{\text{train}}}$ is the number of features observed in the training set. We aim at reporting the posterior mean L_2 error $\mathbb{E}[\text{Err} \mid (\widetilde{z}_{ik})_{i=1,\ldots,n_{\text{train}},k=1,\ldots,K_{n_{\text{train}}}}]$ and the 90% credible interval of the posterior distribution of the L_2 error. Details on the posterior predictive under our non-exchangeable model are given in Section A of the Supplementary Material. In what follows we denote the total number of objects in the dataset as n. Given the closed form expression (2), the hyperparameters (η, σ, ζ) of the three-parameter IBP are estimated by maximum likelihood.

Synthetic data. The models are first tested on synthetic datasets. We generate a binary matrix with $n = 3000$ rows from our non-exchangeable model with parameters $(\eta, \sigma, \xi, \zeta) = (30, 0.5, 1, 1)$, with a training set of size $n_{\text{train}} = 1000$. Figure 4 shows that the feature' sizes are clearly sublinear, and as expected the prediction error is higher under the IBP model, as shown in Table 1. The second synthetic dataset of size $n = 10000$ is generated from the IBP model with parameters $(\eta, \sigma, \zeta) = (20, 0.4, 2.4)$, with $n_{\text{train}} = 3000$. The estimated value for the parameter ξ in the non-exchangeable model is close to zero ($\hat{\xi} = 0.04$). This shows that our model is able to capture the linearity in the feature size, typical of the exchangeable datasets. Both models are able to recover the parameter σ correctly and the prediction errors are about the same (see Table 1 and Figure 4).

Amazon data. This real dataset contains time-ordered reviews of movies from Amazon[2]. We consider the first $n = 4000$ reviews and use $n_{\text{train}} = 1000$ reviews for training. In this context the binary matrix represents the presence of words in the reviews. Figure 4 shows that the presence of words in the reviews increases linearly with the number of reviews, which is expected to be observed in most of text data. As a consequence, the estimated parameter $\hat{\xi} = 0.01$ is very close to zero and the prediction errors of the two models are the same (see Table 1 and Figure 4).

ArXiv data. The ArXiv dataset[3] contains time-ordered articles uploaded on the Arxiv website. Each article is represented as a binary array that encodes its authors. Features size then represents the number of

[2]https://snap.stanford.edu
[3]https://www.kaggle.com/neelshah18/arxivdataset



(a) Synthetic - NE (b) Synthetic - IBP (c) Movies (d) ArXiv (e) Amazon

Figure 5: Log-log plots of the empirical proportions $(K_{n,r}/K_n)_{r\geq 1}$ (red dots) and 95% posterior predictive intervals (blue). Top: Indian Buffet Process (IBP); bottom: non-exchangeable model (NE).

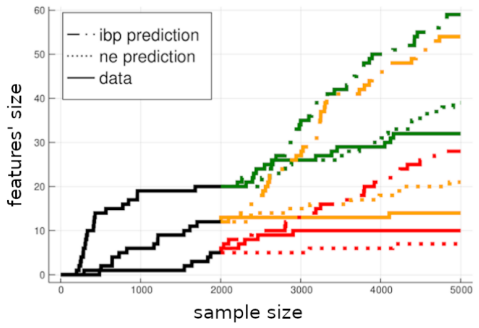


Figure 6: Movies dataset: sizes of some of the features (solid lines, training set in black), predicted features' sizes by the non-exchangeable model (dot lines), predictions by the IBP (dash-dot lines).

articles published by an author as a function of all the articles published on ArXiv. It is therefore expected a sublinear trend of the features' sizes, which is shown in Figure 4. We considered the first $n = 5000$ movies and used the first $n_{train} = 1000$ for training. The non-exchangeable model captures this asymptotic property ($\hat{\xi} = 0.96$) resulting in a better predictive performance compared to the IBP (see Table 1 and Figure 4).

Movies data. This dataset[4] contains time-ordered movies listed in Wikipedia and IMDB. Here the features are represented by the actors who have performed in the movie. The number of movies in which a specific actor has performed is obviously sublinear. We consider a subset of $n = 5000$ movies with $n_{\text{train}} = 2000$. The sublinear trend of the features' sizes is correctly modelled by our model ($\hat{\xi} = 0.45$) which outperforms the IBP in the prediction error (see Table 1 and Figure 4 and 6).

[4]http://udbms.cs.helsinki.fi/?datasets/film_dataset

Finally, Figure 5 shows the posterior predictive of the proportions of features shared by a given number of objects. Both models have the same asymptotic power-law property for this statistic and their fit is similar across all the datasets as expected.

7 Discussion

Various feature allocation models have been proposed in the literature that relax the exchangeability assumption. In particular dependent Indian buffet processes (Caron and Doucet, 2009; Williamson et al., 2010a; Zhou et al., 2011b; Ren et al., 2011; Miller et al., 2012; Gershman et al., 2014; Perrone et al., 2017) include models with covariates (time, space, graph, etc.), so that objects with similar covariates have similar feature weights; see (Foti and Williamson, 2013) for a review. Contrary to this line of work, the aim here is to derive feature allocation models with provable sublinear growth of the features' sizes, retaining the good asymptotic properties of the exchangeable models, namely the power-law behaviour and the control on the number of unique features. The proposed class of models allows to control these quantities by interpretable and tunable parameters, and inference is carried out by a simple Gibbs algorithm. Finally, the problem addressed in this paper is closely related to the problem of microclustering (Miller et al., 2015), which aims at finding models for random partitions where the size of the clusters grows sublinearly.

Acknowledgments. FC acknowledges support from EPSRC under grant EP/P026753/1 and from the Alan Turing Institute under EPSRC grant EP/N510129/1. GDB is funded by EPSRC under grant EP/L016710/1.

References

Broderick, T., Jordan, M. I., and Pitman, J. (2012). Beta processes, stick-breaking and power laws. *Bayesian Analysis*, 7(2):439–476.

Broderick, T., Pitman, J., and Jordan, M. I. (2013). Feature allocations, probability functions, and paintboxes. *Bayesian Analysis*, 8(4):801–836.

Cai, D., Campbell, T., and Broderick, T. (2016). Edge-exchangeable graphs and sparsity. In *Advances in Neural Information Processing Systems*, pages 4249–4257.

Caron, F. (2012). Bayesian nonparametric models for bipartite graphs. In *Advances in Neural Information Processing Systems*, pages 2051–2059.

Caron, F. and Doucet, A. (2009). Bayesian nonparametric models on decomposable graphs. In *Advances in Neural Information Processing Systems*, pages 225–233.

Di Benedetto, G., Caron, F., and Teh, Y. W. (2017). Non-exchangeable random partition models for microclustering. *arXiv preprint arXiv:1711.07287*.

Foti, N. J. and Williamson, S. A. (2013). A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):359–371.

Gershman, S. J., Frazier, P. I., and Blei, D. M. (2014). Distance dependent infinite latent feature models. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):334–345.

Ghahramani, Z. and Griffiths, T. L. (2006). Infinite latent feature models and the indian buffet process. In *Advances in neural information processing systems*, pages 475–482.

Griffiths, T. L. and Ghahramani, Z. (2011). The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73(2):387–396.

James, L. F. (2002). Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. *arXiv preprint math/0205093*.

Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78.

Lee, J., Müller, P., Gulukota, K., and Ji, Y. (2015). A Bayesian feature allocation model for tumor heterogeneity. *The Annals of Applied Statistics*, 9(2):621–639.

Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G., editors, *Bayesian Nonparametrics*. Cambridge University Press.

Miller, J., Betancourt, B., Zaidi, A., Wallach, H., and Steorts, R. (2015). Microclustering: When the cluster sizes grow sublinearly with the size of the data set. *arXiv:1512.00792v1*.

Miller, K. T., Griffiths, T., and Jordan, M. I. (2012). The phylogenetic Indian buffet process: A non-exchangeable nonparametric prior for latent features. *arXiv preprint arXiv:1206.3279*.

Palla, K., Knowles, D. A., and Ghahramani, Z. (2012). An infinite latent attribute model for network data. In *In Proceedings of the International Conference on Machine Learning (ICML)*. Citeseer.

Perrone, V., Jenkins, P. A., Spano, D., and Teh, Y. W. (2017). Poisson random fields for dynamic feature models. *The Journal of Machine Learning Research*, 18(1):4626–4670.

Ren, L., Wang, Y., Carin, L., and Dunson, D. B. (2011). The kernel beta process. In *Advances in Neural Information Processing Systems*, pages 963–971.

Teh, Y. W. and Gorur, D. (2009). Indian buffet processes with power-law behavior. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 1838–1846. Curran Associates, Inc.

Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In *AISTATS*.

Williamson, S., Orbanz, P., and Ghahramani, Z. (2010a). Dependent Indian buffet processes. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 924–931.

Williamson, S., Wang, C., Heller, K. A., and Blei, D. M. (2010b). The IBP compound Dirichlet process and its application to focused topic modeling. In *Proceedings of the 27th international conference on machine learning (ICML)*, pages 1151–1158.

Xu, Y., Müller, P., Yuan, Y., Gulukota, K., and Ji, Y. (2015). MAD Bayes for tumor heterogeneity—feature allocation with exponential family sampling. *Journal of the American Statistical Association*, 110(510):503–514.

Zhou, M., Chen, H., Paisley, J., Ren, L., Li, L., Xing, Z., Dunson, D., Sapiro, G., and Carin, L. (2011a). Nonparametric Bayesian dictionary learning for analysis of noisy and incomplete images. *IEEE Transactions on Image Processing*, 21(1):130–144.

Zhou, M., Yang, H., Sapiro, G., Dunson, D., and Carin, L. (2011b). Dependent hierarchical beta process for image interpolation and denoising. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 883–891.