Indian Buffet Processes with Power-law Behaviour

Bobby He Hector McKimm Natalia Garcia Martin

October 18, 2019

Abstract

In this paper, we review the Indian Buffet Process (IBP) and its generalisation the Stable-beta Indian Buffet Process (SB-IBP), two Bayesian nonparametric methods used to model data with multiple features. As nonparametric models, both are able to model data for which the total number of features in potentially unbounded. The SB-IBP has the additional benefit of being able to better model data displaying power-law behaviour, that is, where particular features dominate in their popularity. To illustrate and compare the strength of these models, this report considers an application to document modelling, where features correspond to the presence or absence of particular words.

1 Introduction

Model selection is often a key question of interest for statistical inference. In particular, for featural models, the number of features used in the model is crucial. Although the context of a given problem can make clear the appropriate number of features, in some scenarios it will be unclear how many features should be used, thus any upper bound on this number will be arbitrary.

Bayesian nonparametric models provide a means of dealing with data for which the total number of features is unknown and potentially infinite; the number of features is allowed to grow as more data is collected. More precisely, the number of features in the model are unbounded, and it is assumed only a finite subset of features are observed via the given data.

This report will focus on two Bayesian nonparametric featural models: the Indian Buffet Process (IBP), which is presented in Section 2, and its generalisation the Stable-beta Indian Buffet Process (SB-IBP), described in Section 4. In particular, we will show in Sections 3 and 4 respectively how both these processes can be represented mathematically. In Section 5, we consider whether the SB-IBP can be represented as the infinite limit of a parametric model. A potentially advantageous property of the SB-IBP is that its additional parameter enables it to better model data exhibiting power-law behaviour, in which certain features are significantly more popular than others. The extent to which the SB-IBP successfully models data exhibiting such properties is tested in Section 6, where we analyse data concerning the presence or absence of words in related documents to compare how the models perform. Finally, we conclude with a discussion in Section 7.

2 The Indian Buffet Process

The IBP is an infinitely exchangeable distribution over binary matrices that allows data points to share multiple features. Objects are represented as binary vectors Z_i with entries indicating the presence or absence of each feature. The IBP assumes that the number of classes or features (represented by the number of columns in the matrix) is unbounded and that the observed objects manifest a finite subset of these features [1]. We can explain this process using a metaphor of n customers sequentially serving themselves from a buffet at an Indian restaurant by choosing a finite number of dishes from infinitely many possibilities:

- The first customer tries Poisson(α) dishes.
- The n^{th} customer tries:

– Each of the current dishes with probability m_k/n, where m_k denotes the number of times dish k has been chosen, and

– Poisson(α/n) new dishes.

The IBP uses a single parameter α. Introducing a second parameter c, we get a generalisation of the process [2], presented below. Note that by simply setting $c = 1$, we recover the one parameter model.

- The first customer tries Poisson(α) dishes.
- The nth customer tries each of the current dishes with probability $m_k/(c+n-1)$ as well as Poisson $(\alpha c/(c+n-1))$ new dishes.

3 Connection between the IBP and the Beta Process

3.1 de Finetti's Theorem

The IBP generates an exchangeable distribution over binary matrices: $P(Z_1, \ldots, Z_n) = P(Z_{\sigma(1)}, \ldots, Z_{\sigma(n)})$ for all permutations σ and all n. de Finetti's Theorem states the distribution of any infinitely exchangeable sequence can be written in the following way [3]:

$$P(Z_1, \ldots, Z_n) = \int \left\{ \prod_{i=1}^{n} P(Z_i \mid B) \right\} dP(B), \tag{1}$$

with $P(B)$ known as the "de Finetti mixing distriubtion", or mixing distriubtion for short. There must exist such a de Finetti mixing distriubtion for the IBP; in this section we find it for the two parameter generalisation.

3.2 The Beta process

Define a beta process $B \sim \mathrm{BP}(c, B_0)$ as a positive Lévy process whose Lévy measure depends on two parameters, a positive "concentration" function c over a space Θ and a fixed "base" measure B_0 on Θ. In the following we will assume c is a constant and refer to it as the concentration parameter. If B_0 is continuous the Lévy measure is defined on $[0, 1] \times \Theta$ as:

$$\nu(du, d\theta) = cu^{-1}(1-u)^{c-1} du B_0(d\theta). \tag{2}$$

If B_0 is discrete, writing it as $B_0 = \sum_i v_i \delta_{\theta_i}$ with $v_i \in [0, 1]$, then B will have atoms in the same locations: $B = \sum_i u_i \delta_{\theta_i}$, with

$$u_i \sim \mathrm{Beta}(cv_i, c(1-v_i)). \tag{3}$$

If B_0 is a mixture of discrete and continuous, then B will be the sum of two independent parts, one continuous and the other discrete.

3.3 Bernoulli Process

Define a Bernoulli process [3] $Z \sim \mathrm{BernoulliP}(B)$ with "hazard" measure B as the Lévy process with Lévy measure:

$$\zeta(du, d\theta) = \delta_1(du) B(d\theta). \tag{4}$$

If B is continuous, Z is a Poisson process with intensity B and N points where:

$$N \sim \mathrm{Poi}(B(\Theta)). \tag{5}$$

If B is discrete, expressed in form $B = \sum_i u_i \delta_{\theta_i}$ then $Z = \sum_i b_i \delta_{\theta_i}$ with

$$\mathbb{P}(b_i = 1) = u_i \ , \ \mathbb{P}(b_i = 0) = 1 - u_i \tag{6}$$

If B is a mixture, then Z is the sum on two independent contributions arising from a discrete measure and a continuous measure.

3.4 The Beta process as the mixing distriubtion of the IBP

Consider the following model, with B_0 a continuous measure and Z_i drawn independently for $i = 1, \ldots, n$:

$$B \sim \mathrm{BP}(c, B_0) \tag{7}$$

$$Z_i \mid B \sim \mathrm{BernoulliP}(B) \tag{8}$$

The beta process is conjugate to the Bernoulli process [3]; the posterior for B is:

$$B \mid Z_1, \ldots, Z_n \sim \mathrm{BP}\Big(c+n, \frac{c}{c+n}B_0 + \frac{1}{c+n}\sum_{i=1}^{n} Z_i\Big). \tag{9}$$

Z_1 is a Lévy process - the value it takes on an interval is independent of its value on a disjoint interval. It assigns weight 0 or 1 to atoms, so it is a Bernoulli process. Consider its expected value: $\mathbb{E}_{Z_1}[Z_1] = \mathbb{E}_B[\mathbb{E}_{Z_1|B}[Z_1]] = \mathbb{E}_B[B] = B_0$, with the last equality shown to hold by [3]. We have thus found the hazard measure of Z_1:

$$Z_1 \sim \mathrm{BernoulliP}(B_0). \tag{10}$$

Denote the unique atoms in $Z_1, \ldots, Z_n$ as $\theta_1^*, \ldots, \theta_K^*$, with θ_k^* appearing m_k times. Using the expression $p(Z_{n+1} \mid Z_1, \ldots, Z_n) = \mathbb{E}_{B|Z_1,\ldots,Z_n}[p(Z_{n+1} \mid B)]$ in combination with (9) we find:

$$Z_{n+1} \mid Z_1, \ldots, Z_n \sim \mathrm{BernoulliP}\Big(\frac{c}{c+n}B_0 + \sum_j \frac{m_j}{c+n}\delta_{\theta_j^*}\Big). \tag{11}$$

The equations (10) and (11) show that the model defined by (7) and (8) is the two parameter generalisation of the IBP. Since B_0 is a continuous measure, using property (5) of the Bernoulli process, the first customer 1 samples a number of dishes according to a Poisson process with rate parameter $\alpha := B(\Theta)$. Proceeding customers then sample dishes according to a Bernoulli process whose measure is a mixture of continuous and discrete. The discrete part results in customer $n+1$ trying previously tasted dishes with probability $m_j/(c+n)$. The continuous part results in customer $n+1$ trying $N \sim \mathrm{Poisson}(\frac{\alpha c}{c+n})$ new dishes.

4 Connection between the SB-IBP and the Stable-Beta Process

4.1 The Stable-Beta Indian Buffet Process

Distributions of the form

$$p(x) = Cx^{-\alpha}$$

are said to follow a power law with exponent α. On the log-log scale, this corresponds to a straight line of the form $\log p(x) = -\alpha \log x + c$, where α and $c = \log C$ are constants. Power-law distributions occur in a diverse range of phenomena including city populations, earthquakes sizes, computer files, the frequency of use of words in any human language, the frequency of occurrence of personal names in most cultures and the number of citations received by papers [4].

The SB-IBP is a generalisation of the IBP exhibiting power-law behaviour, with parameters $\alpha > 0$, $c > -\sigma$ and $\sigma \in [0, 1)$. $\sigma = 0$ corresponds to the standard two parameter IBP. It can be described using the following metaphor:

- The first customer tries Poisson(α) dishes.
- The nth customer tries
  - each of the current dishes with probability $(m_k - \sigma)/(c + n - 1)$
  - as well as Poisson$\Big(\alpha\frac{\Gamma(1+c)\Gamma(n-1+c+\sigma)}{\Gamma(n+c)\Gamma(c+\sigma)}\Big)$ new dishes.

The left image in Figure 1 shows through simulations how σ controls the power-law behaviour of the SB-IBP: the cumulative number of dishes tried by customers increases as σ increases. The right plot shows that a few dishes are exceptionally popular, whilst many dishes are tried only a few times.

In the following subsections, a Completely Random Measure (CRM) called the stable-beta process (SBP) will be introduced and shown to be the mixing distribution of the SB-IBP.

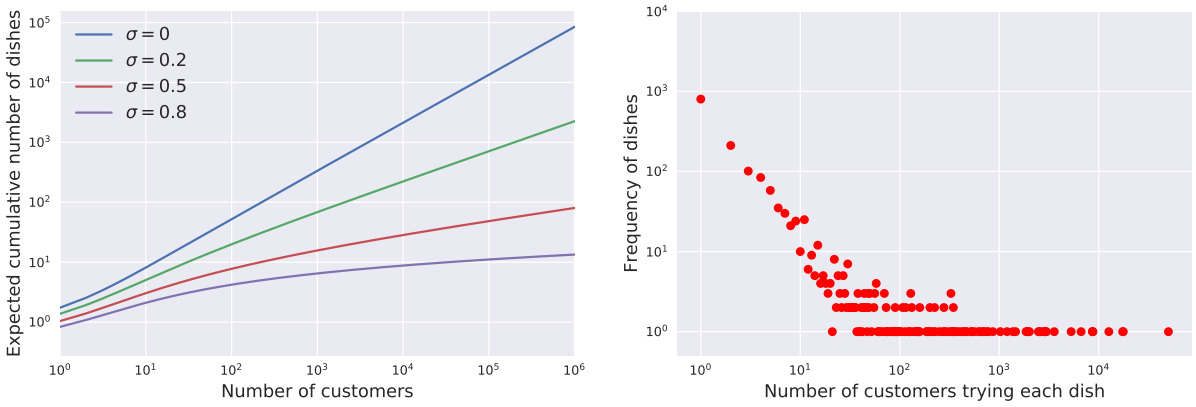

Figure 1: Simulations to demonstrate power-law properties of the SB-IBP. Parameters used: (Left) $\alpha = 1, c = 1$, σ varies; (Right) $\alpha = 1, c = 1, \sigma = 0.5$.

4.2 Completely Random Measures

A CRM μ on (Θ, Σ) is a random measure such that $\mu(X)$ is independent of $\mu(Y)$ for all disjoint measurable sets $X, Y \in \Theta$. A CRM can be written as the sum of three independent parts:

$$\mu = \mu_0 + \sum_{k=1}^{M} u_k \delta_{\phi_k} + \sum_{l=1}^{N} p_l \delta_{\theta_l}, \tag{12}$$

where μ_0 is a non-random measure, u_k and $p_l > 0$ are random masses, ϕ_k are fixed atoms and θ_l are random atoms. Each u_k is independent of everything else and has distribution F_k. The random atoms and their weights $\{p_l, \theta_l\}$ are the points of a 2D Poisson Process over $(0, \infty] \times \Theta$ with some nonatomic underlying weight measure Λ. We write

$$\mu \sim \text{CRM}(\Lambda, \{\phi_k, F_k\}_{k=1}^{N}) \tag{13}$$

4.3 The Stable-Beta process

The SBP is a CRM with no non-random measure, no fixed atoms and Lévy measure defined on $(0, 1) \times \Theta$ as:

$$\Lambda_0(dp, d\theta) = \alpha \frac{\Gamma(1 + c)}{\Gamma(1 - \sigma)\Gamma(c + \sigma)} p^{-\sigma - 1}(1 - p)^{c + \sigma - 1} dp H(d\theta). \tag{14}$$

H is a smooth base distribution over the locations of the atoms.

4.4 The Stable-Beta process as the mixing distribution of the SB-IBP

Consider the following model:

$$\mu \sim \text{CRM}(\Lambda_0, \{\}), \tag{15}$$

$$Z_i \mid \mu \sim \text{BernoulliP}(\mu). \tag{16}$$

Note its similarity to hierarchical model previously proposed, with the difference being that the random measure on which $Z_i \mid \mu$ depends (for $i = 1, \ldots, n$), is now a SBP instead of a beta process. The empty curly brackets indicate the SBP has no fixed atoms. In this section we will show that this model is the SB-IBP. Recall the unique atoms in $Z_1, \ldots, Z_n$ are denoted $\theta_1^*, \ldots, \theta_K^*$, with θ_k appearing m_k times. First we must consider the posterior of μ, given by [5]:

$$\mu \mid Z_1, \ldots, Z_n \sim \text{CRM}(\Lambda_n, \{\theta_k^*, F_{nk}\}_{k=1}^{K}). \tag{17}$$

F_{nk} has a beta distribution with parameters $(m_k - \sigma, n - m_k + c + \sigma)$. Λ_n is the Lévy measure of an updated SBP process:

$$\Lambda_n(dp, d\theta) = \alpha \frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} p^{-\sigma-1}(1-p)^{n+c+\sigma-1} dp H(d\theta). \tag{18}$$

The posterior is a CRM, but not a SBP; it is the independent sum of a SBP with updated parameters and of fixed atoms with beta distributed masses [6].

The probability of a new data point having a previously selected feature is derived as follows:

$$p(Z_{n+1}(\theta_k^*) = 1 \mid Z_1, \ldots, Z_n) = \int_M p(Z_{n+1}(\theta_k^*) = 1, \mu \mid Z_1, \ldots, Z_n) d\mu \tag{19}$$

$$= \int_M p(Z_{n+1}(\theta_k^*) = 1 \mid \mu, Z_1, \ldots, Z_n) p(\mu \mid Z_1, \ldots, Z_n) d\mu \tag{20}$$

The first term in the second equation depends on a measure consisting of a continuous part and a discrete part, only the latter of which can contribute to the probability of the random variable Z_{n+1} taking a value at point θ_k^*. The property of the Bernoulli process described by (6) means this probability is given by the weight of the measure at θ_k^*. We therefore have:

$$p(Z_{n+1}(\theta_k^*) = 1 \mid Z_1, \ldots, Z_n) = \int_M \mu(\theta_k^*) p(\mu \mid Z_1, \ldots, Z_n) d\mu, \tag{21}$$

$$= \mathbb{E}[\mu(\theta_k^*) \mid Z_1, \ldots, Z_n], \tag{22}$$

$$= \frac{m_k - \sigma}{n + c}. \tag{23}$$

The expectation is evaluated, using the fact that the posterior of μ at atoms has a beta distribution.

The probability of a new data point having a new feature - the probability of Z_{n+1} having an atom at some new position - can likewise be evaluated as an expectation;

$$p(Z_{n+1}(d\theta) = 1 \mid Z_1, \ldots, Z_n) = \mathbb{E}[\mu(d\theta) \mid Z_1, \ldots, Z_n], \tag{24}$$

$$= \int_0^1 u \Lambda_n(du \times d\theta), \tag{25}$$

$$= \alpha \frac{\Gamma(1+c)\Gamma(n+c+\sigma)}{\Gamma(n+1+c)\Gamma(c+\sigma)} H(d\theta). \tag{26}$$

The last step is found by substituting in the expression for Λ_n and integrating. Continuing with the metaphor, this step can be interpreted as the number of new dishes tasted by a new customer being drawn from a Poisson process with rate measure given above.

Lastly, the likelihood of the data with μ marginalised out can be shown to be [6]:

$$p(Z_1, \ldots, Z_n) = \exp\left(-\alpha \sum_{i=1}^{n} \frac{\Gamma(1+c)\Gamma(i-1+c+\sigma)}{\Gamma(i+c)\Gamma(c+\sigma)} \right) \prod_{k=1}^{K} \frac{\Gamma(m_k - \sigma)\Gamma(n - m_k + c + \sigma)\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)\Gamma(n+c)} \alpha h(\theta_k^*), \tag{27}$$

where h is the density of H.

5 IBPs as limits of parametric models

Consider the following hierarchical model for the $n \times p$ binary matrix Z.

$$\pi_j \sim \text{Beta}\left(\frac{\alpha}{p}, 1\right) \qquad \forall j \leq p$$

$$z_{i,j} | \pi_j \overset{\text{ind}}{\sim} \text{Ber}(\pi_j) \qquad \forall i \leq n,\, j \leq p$$

One can integrate out $\{\pi_j\}_j$ and use standard properties of the Beta function in relation to the Gamma function in order to obtain the marginal for Z:

$$\mathbb{P}[Z] = \prod_{j=1}^{p} \frac{\frac{\alpha}{p}\Gamma\left(m_j + \frac{\alpha}{p}\right)\Gamma\left(n - m_j + 1\right)}{\Gamma\left(n + 1 + \frac{\alpha}{p}\right)}$$

where $m_j = \sum_{i=1}^{n} z_{i,j}$. Now let's define K_n to be the number of columns of Z with at least one non-zero entry and $\kappa_0 = p - K_n$. Then, if f_n is the multiset corresponding to Z, we can use column exchangeability to group together all matrices corresponding to f_n to give:

$$\mathbb{P}[f_n] = \frac{p!}{\kappa_0! \prod_{h=1}^{\tilde{K}_n} \kappa_h!} \left(\frac{\frac{\alpha}{p}\Gamma\left(\frac{\alpha}{p}\right)\Gamma(n+1)}{\Gamma\left(n+1+\frac{\alpha}{p}\right)}\right)^{\kappa_0} \prod_{j=1}^{K_n} \frac{\frac{\alpha}{p}\Gamma\left(m_j + \frac{\alpha}{p}\right)\Gamma\left(n - m_j + 1\right)}{\Gamma\left(n+1+\frac{\alpha}{p}\right)} \tag{28}$$

$$= \frac{\alpha^{K_n}}{\prod_{h=1}^{\tilde{K}_h} \kappa_h!} \frac{p!}{\kappa_0! p^{K_n}} \cdot \left(\frac{n!\Gamma\left(\frac{\alpha}{p}+1\right)}{\Gamma\left(n+1+\frac{\alpha}{p}\right)}\right)^{p} \cdot \prod_{j=1}^{K_n} \frac{\Gamma\left(m_j + \frac{\alpha}{p}\right)(n-m_j)!}{\Gamma\left(\frac{\alpha}{p}+1\right)n!} \tag{29}$$

$$\xrightarrow{p\to\infty} \frac{\alpha^{K_n}}{\prod_{h=1}^{\tilde{K}_h} \kappa_h!} \cdot e^{-\alpha H_n} \cdot \prod_{j=1}^{K_n} \frac{(m_j - 1)!(n - m_j)!}{n!} \tag{30}$$

where κ_l denotes the multiplicity of the l^{th} unique value in f_n, H_n denotes the truncated Harmonic series, and we suppose that the columns are reordered in such a way that allows the non-zero columns be on the left of Z.

Having taken the limit $p \to \infty$, we assume that there are infinitely many features (or dishes in the culinary analogy). For $j \in \mathbb{N}$, feature j has some location θ_j in some feature space Θ, regardless of whether it is observed in our finite n sample, and these locations are i.i.d. drawn from some smooth density g. If we instead view the rows of matrix Z to be corresponding to a point process with binary weights $Z_i = \sum_{j=1}^{\infty} z_{i,j}\delta_{\theta_j}$, we obtain the density of our observed n sample:

$$\begin{aligned} p[Z_1, \ldots, Z_n] =& \mathbb{P}[f_n(Z_1, \ldots, Z_n)] \cdot \prod_{h=1}^{\tilde{K}_h} \kappa_h! \cdot \prod_{j=1}^{K_n} g_0(\theta_j^*) \\ =& \alpha^{K_n} e^{-\alpha H_n} \prod_{j=1}^{K_n} \frac{(m_j - 1)!(n - m_j)!}{n!} \prod_{j=1}^{K_n} g(\theta_j^*), \end{aligned}$$

where recall $\{\theta_j^*\}_{j=1}^{K_n}$ are the set of observed features. This completes the derivation of the IBP as a limit of parametric models and is the derivation the authors in [1] used.

A similar derivation can be obtained for the two-parameter IBP, as found in [7] where the following parametric model is taken to its infinite limit in p:

$$\begin{aligned} \pi_j &\sim \text{Beta}\left(\frac{\alpha c}{p}, c\right) \qquad \forall j \leq p \\ z_{i,j}|\pi_j &\overset{\text{ind}}{\sim} \text{Ber}(\pi_j) \qquad \forall i \leq n,\, j \leq p \end{aligned}$$

The authors in [6] were interested in finding a derivation of the Stable-beta IBP corresponding to the infinite limit of a parametric model. However, recall that the Pitman-Yor Process can be viewed as a Power-law extension of the Dirichlet Process, but whereas the Dirichlet Process can be viewed as an infinite limit of finite-dimensional Dirichlet distributions, the Pitman-Yor does not have such a property. This suggests that the Stable-beta Process, which is the Power-law extension of the Beta process, may not be able to be viewed as an infinite limit either.

Nevertheless, we attempt to investigate this question. Let $f_p(\cdot;\theta)$ be some parametric density supported on $(0,1)$, with parameters $\theta = (\alpha, c, \sigma)$. Suppose that:

$$\begin{aligned} \pi_j &\sim f_p(\cdot;\theta) \qquad \forall j \leq p \\ z_{i,j}|\pi_j &\overset{\text{ind}}{\sim} \text{Ber}(\pi_j) \qquad \forall i \leq n,\, j \leq p \end{aligned}$$

We are interested in finding properties that f_p must satisfy, in order for the Stable-beta process to be the limit for Z as $p \to \infty$. For example, we need to have $f_p(\cdot;\theta) = \text{Beta}\big(\frac{\alpha c}{p}, c\big)$ when $\sigma = 0$.
Integrating out $\{\pi_j\}_j$ as before, we obtain:

$$\begin{aligned}\mathbb{P}[Z] &= \prod_{j=1}^{p} \int_0^1 \pi_j^{m_j} \big(1-\pi_j\big)^{n-m_j} f_p\big(\pi_j;\theta\big) d\pi_j \\ &= \prod_{j=1}^{p} b_p(m_j, n, \theta)\end{aligned}$$

for some function b_p using the same definition $m_j = \sum_i z_{i,j}$ as before. When f_p is simply a beta density we obtain a closed form for b_p. Notice that the red parts of equation (28) correspond to the contributions of the unobserved features, when $m_j = 0$, to the likelihood, where some minor algebraic manipulations are used to gain exponent p instead of κ_0. Moreover, this is the term that gives the exponential term in the limit.

What happens when we try to match these exponent terms in the Stable-beta case? From equation (10) of paper [6], we see that we want:

$$\Big(b_p\big(0,n,\theta\big)\Big)^p \xrightarrow{p\to\infty} \exp\Big(-\alpha \sum_{i=1}^{n} \frac{(c+\sigma)_{(i-1)}}{(1+c)_{(i-1)}}\Big),$$

where $a_{(n)} = \frac{\Gamma(a+n)}{\Gamma(a)}$ is the Pochhammer notation. Using the identity $(1+\frac{x}{n})^n \xrightarrow{n} e^x$, this suggests, in a highly non-rigorous way, that a potential candidate for b_p that will definitely satisfy the exponential terms in the limit is:

$$b_p(0,n,\theta) = \prod_{i=1}^{n} \frac{(1+c)_{(i-1)}}{(1+c)_{(i-1)} + \frac{\alpha}{p}(\sigma+c)_{(i-1)}} \qquad \forall n \in \mathbb{N}. \tag{31}$$

But notice that:

$$b_p(0,n,\theta) = \int_0^1 (1-\pi_j)^n f_p(\pi_j,\theta) d\pi_j \qquad \forall n \in \mathbb{N}. \tag{32}$$

And thus, matching together equations (31) and (32) we may iteratively obtain all moments of the candidate density $f_p(\cdot,\theta)$, which we know exist as f_p has support $[0,1]$. But now, theoretically, we can use these moments to obtain the characteristic function of f_p, which determines it. Once we have obtained the density f_p we may then verify that our candidate f_p does indeed give the Stable-beta IBP as its limit. Of course, this is all hypothetical as upon inspection, the moments given by our equations hold no nice form for $\sigma > 0$.

6 Experiments and Results

6.1 Recreating results from paper [6]

We compared the goodness of fit of the two-parameter IBP and the SB-IBP for the 20newsgroups dataset as in the first experiment of [6]. For each of the 20 news groups we found the Maximum Likelihood parameters using equation (27), for α, c and σ. For the SB-IBP we obtained: $\hat{\alpha} = 179.2 \pm 38.0$, $\hat{c} = 3.25 \pm 0.738$ and $\hat{\sigma} = 0.48 \pm 0.044$, where the first number denotes the mean over the 20 newsgroups and the second denotes the standard deviation. For the IBP, we obtained $\hat{\alpha} = 182.7 \pm 40.3$ and $\hat{c} = 23.5 \pm 5.49$. These values are similar to those obtained in the original paper and also depict that the IBP needs much larger estimates for c in order to compensate for the fact that a large number of the words occur in only a few documents. Figure 2 depicts the power law properties of the 20newsgroups dataset. We can see from both plots that the SB-IBP offers a better fit to the real data due to its ability to capture the power-law behaviour.

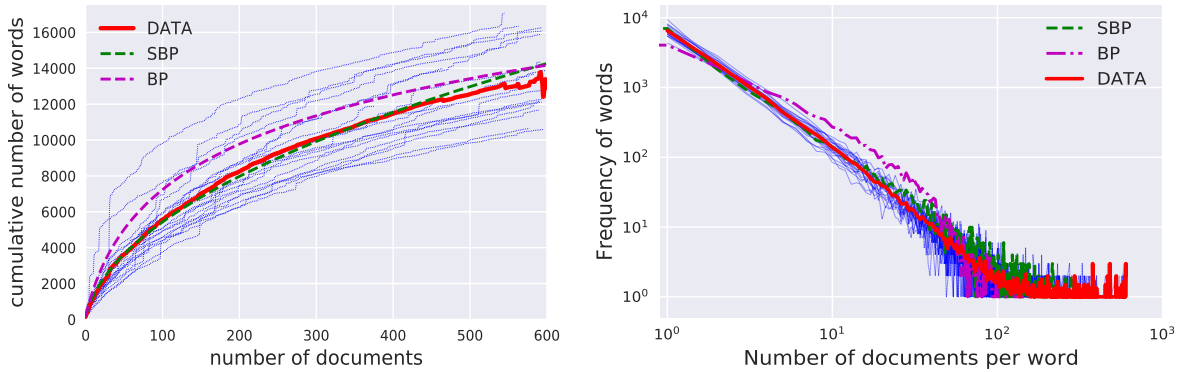

Figure 2: Power-law properties of the 20newsgroups dataset. The left plot depicts how the cumulative number of words increases as more documents are observed. The faint dashed lines correspond to each of the 20 newsgroups, and the solid red curve is the mean of these. The SB-IBP (green) and BP-IBP (magenta) use the expected number of words from the mean of the 20 ML parameters. The right plot depicts the distribution of the number of different documents a word appears in, on a log-log scale

6.2 Neurips Dataset

We also used the SB-IBP to model word occurence data from the Neurips Conference from 1987 to 2015. Again, using equation (27) we calculated the ML parameters for the SB-IBP and the IBP. For the SB-IBP we obtained values $\hat{\alpha} = 721$, $\hat{c} = -0.19$ and $\hat{\sigma} = 0.71$. For the IBP, we obtained values $\hat{\alpha} = 1168$ and $\hat{c} = 60.3$. These are in line with prior beliefs, as α should be approximately the average number of words in a given document, which was calculated to be 900, and σ can be show to be approximately the proportion of observed words that are contained in exactly one document, which was calculated to be 0.71. Again, we observe that $\hat{c}$ is much larger in the IBP case in order to compensate for high proportion of words occuring in only a few documents. We recreated the same plots as for the 20newsgroups dataset and these are shown in figure 3

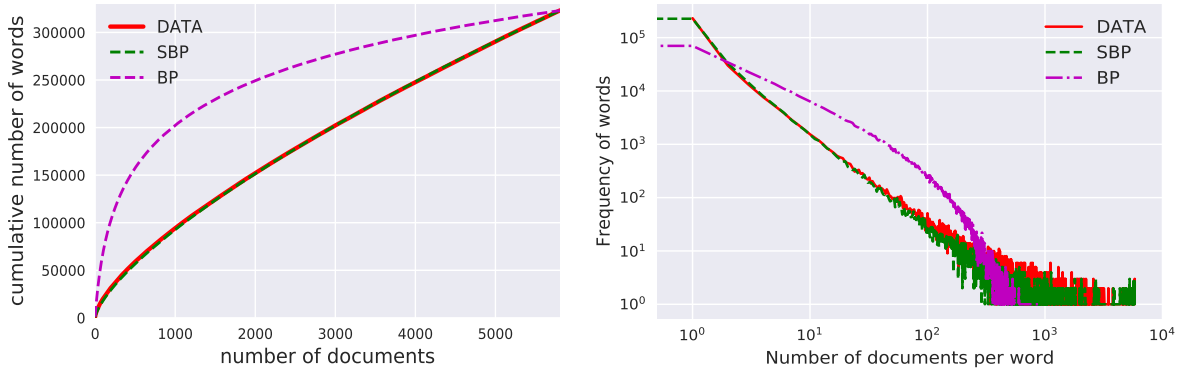

Figure 3: Power-law properties of the Neurips dataset

Again we see that the SB-IBP is far superior to the standard IBP in capturing the power-law behaviour of the Neurips word count data, and this difference is even more pronounced as $\hat{\sigma} = 0.71$ is much higher than for the 20newsgroups dataset.

7 Discussion

Both the IBP and SB-IBP are powerful Bayesian nonparametric models capable of modelling data with an unbounded number of features. In particular, we have demonstrated via an application to document modelling how the SB-IBP is better suited to phenomenon displaying power-law behaviour. A fascinating further use of both processes, not dealt with in this short report, is in classification. The basic idea for

this task is, given a new object with certain features, to use labelled training data to calculate the likelihood of the new object belonging to each of the possible classes. The new object is then classified based on which category they are most likely to belong to.

References

[1] Z. Ghahramani and T. L. Griffiths, "Infinite latent feature models and the indian buffet process," pp. 475–482, 2006.

[2] E. Xing, "Probabilistic graphical models lecture notes: The indian buffet process," 2014.

[3] R. Thibaux and M. I. Jordan, "Hierarchical beta processes and the indian buffet process," in *Artificial Intelligence and Statistics*, pp. 564–571, 2007.

[4] M. E. Newman, "Power laws, pareto distributions and zipf's law," *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.

[5] Y. Kim, "Nonparametric bayesian estimators for counting processes," *Annals of Statistics*, pp. 562–588, 1999.

[6] Y. W. Teh and D. Görür, "Indian buffet processes with power-law behavior," pp. 1838–1846, 2009.

[7] Z. Ghahramani, T. L. Griffiths, and P. Sollich, "Bayesian nonparametric latent feature models," 2007.

[8] J. Pitman *et al.*, "Combinatorial stochastic processes," tech. rep., Technical Report 621, Dept. Statistics, UC Berkeley, 2002.