Preprint of article published in *Statistics and Computing*

Restricted Indian Buffet Processes

Finale Doshi-Velez · Sinead A. Williamson

Received: 20 August 2015 / Accepted: 27 June 2016

Abstract Latent feature models are a powerful tool for modeling data with globally-shared features. Nonparametric distributions over exchangeable sets of features, such as the Indian Buffet Process, offer modeling flexibility by letting the number of latent features be unbounded. However, current models impose implicit distributions over the number of latent features per data point, and these implicit distributions may not match our knowledge about the data. In this work, we demonstrate how the Restricted Indian Buffet Process circumvents this restriction, allowing arbitrary distributions over the number of features in an observation. We discuss several alternative constructions of the model and apply the insights to develop Markov Chain Monte Carlo and variational methods for simulation and posterior inference.

Keywords Bayesian nonparametrics · Latent feature models · Indian Buffet Process

1 Introduction

Generative models are a popular approach for identifying latent structure in data. For example, a musical piece may be naturally modeled as a collection of notes, each with associated frequencies. A patient's health may be naturally modeled as a collection of diseases, each with associated symptoms. The text of a news article may be naturally modeled as a collection of topics, each with associated words. In each of these cases, we posit that there exists a small set of underlying features that are responsible for generating the structure that we observe in the data.

When the number of these underlying features is unknown, Bayesian nonparametric models such as the Indian Buffet Process (IBP) [Griffiths and Ghahramani, 2011] provide an elegant generative modeling approach. The IBP posits that there are an infinite number of potential underlying features, but only a finite number of features underlie any particular observation. The IBP has been the foundation for a variety of modeling applications including choice behavior [Görür et al., 2006], psychiatric comorbitities [Ruiz et al., 2014], network models [Miller et al., 2009], blind source separation [Knowles and Ghahramani, 2007], image modeling [Zhou et al., 2009], and time-series [Fox et al., 2009].

Under the IBP, the prior distribution over the number of features underlying an observation is Poisson(α), where α is the concentration parameter for the IBP. A three-parameter extension of the IBP [Teh and Görür, 2009] retains this strong requirement for Poisson-distributed feature cardinality, as do IBP variants which posit correlations between observations [Gupta et al., 2013, Miller et al., 2008] or features [Doshi-Velez and Ghahramani, 2009]. Other non-parametric latent variable models such as the infinite gamma-Poisson process [Titsias, 2008] and the beta-negative Binomial process [Broderick et al., 2015, Zhou et al., 2012] also exhibit a Poisson distribution over the number of non-zero features.

One exception is [Caron, 2012], which number of features underlying each observation to follow a mixture

F. Doshi-Velez
Harvard Paulson School
29 Oxford Street
Cambridge, MA 02138
E-mail: finale@seas.harvard.edu

S.A. Williamson
McCombs School of Business
University of Texas at Austin
2110 Speedway
Austin, TX 78705

of Poissons. This allows us to capture overdispersed distributions over the number of features. This might be appropriate when modeling word occurrence in natural language, and degree distribution in networks, where we often see power-law behavior.

Overdispersed distributions are just one situation where a Poisson distribution over the number of features may be inappropriate. We may have an upper bound on the number of features expected for an observation, for example if we expect features to correspond to individuals in a conversation or instruments in a musical recording. We may expect our distribution to be zero-deflated: for example when modeling articles, we may wish to preclude the possibility of having no topics represented. A textual label for an image may provide strong clues about the number of objects we expect in the image. The IBP does not provide the flexibility to put an arbitrary prior distribution on the number of latent features in an observation; the mixture of Poissons approach [Caron, 2012] constrains us to overdispersed distributions with full support on the non-negative integers.

In this article, we present and describe the Restricted Indian Buffet Process (R-IBP), a recently developed model that allows an arbitrary prior distribution to be placed over the number of features underlying each observation. Unlike the model of [Caron, 2012], this distribution can have arbitrary support, or even be degenerate on a single value. We expand upon the original exposition of the R-IBP in [Williamson et al., 2013] with several alternative constructions, new insights, and novel efficient inference techniques.

2 Background: Completely Random Measures and the Indian Buffet Process

Many Bayesian nonparametric models, including the IBP, can be expressed in terms of completely random measures (CRMs) [Kingman, 1967]. A completely random measure μ is a random measure consisting of a collection of atoms[1] $\mu = \sum_i \pi_i \delta_{\theta_i}$ on some space $(\Theta, \mathcal{A})$ such that for any disjoint subsets $A_1, A_2 \in \mathcal{A}, A_1 \cap A_2 = \emptyset$, the masses $\mu(A_1), \mu(A_2)$ assigned to those subsets are independent.

The atoms fall into two categories: atoms where the locations θ_i are random, and atoms where the locations θ_i are fixed a priori. The size π_i and locations θ_i of the randomly-located atoms are governed by a Lévy measure $\nu(d\pi, d\theta)$. Different Lévy measures yield different properties. For example, the Lévy measure $\nu(d\pi, d\theta) = c\alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi H(d\theta)$ describes the homogeneous beta process [Hjort, 1990], in which the atom sizes π_i are equal in distribution to the limit as $I \to \infty$ of Beta $\left(\frac{c\alpha}{I}, c\left(1-\frac{\alpha}{I}\right)\right)$ random variables. The Lévy measure $\nu(d\pi, d\theta) = \gamma\pi^{-1}e^{-\lambda\pi}d\pi H(d\theta)$ describes the gamma process, whose atom sizes correspond to the infinitesimal limit of a gamma distribution. If the Lévy measure can be decomposed as $\nu(d\pi, d\theta) := \nu_\pi(d\pi)\nu_\theta(d\theta)$—as is the case for all Lévy measures considered in this paper—then the atoms' locations are independent of their sizes.

The masses of the fixed-location atoms are governed by a finite or countable measure ξ with atoms at the fixed locations. In a Bayesian inference setting, these fixed atoms are normally associated with observed values, and the corresponding fixed-location jumps characterize the effect of those observations on the posterior. A CRM with both fixed and random atom locations can be decomposed as the sum of two CRMs, one with fixed-location atoms and one with randomly located atoms. In this paper, we will write an arbitrary CRM with Lévy measure ν governing atoms with random location and atomic measure ξ governing atoms with fixed location as CRM $(\nu + \xi)$; since ν is continuous and ξ is discrete there should be no difficulty distinguishing the two component measures.

We can use a CRM to construct a distribution over an exchangeable sequence of measures ζ_n on Θ. To do so, we first define a *directing measure* $\mu := \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i} \sim$ CRM (ν) to be a CRM with Lévy measure ν.[2] We then let $\zeta_n := \sum_{n=1}^{\infty} z_{ni}\delta_{\theta_i} \overset{\text{i.i.d.}}{\sim}$ CRM$(g(\mu)), n = 1, 2, \ldots$ be a sequence of CRMs parametrized by some functional $g(\mu)$ of this directing measure μ. If we integrate out μ, the sequence $\zeta_1, \zeta_2, \ldots$ is an infinitely exchangeable sequence of measures on Θ.

For example, the beta-Bernoulli process [Thibaux and Jordan, 2007] is a distribution over exchangeable counting measures with finite support. The directing measure μ is distributed according to a beta process

$$\text{BP}(c, \alpha, H) := \text{CRM}\left(c\alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi H(d\theta)\right),$$

where $c, \alpha > 0$ and H is a probability measure on Θ. Conditioned on μ, the ζ_n are distributed according to a Bernoulli process

$$\text{BeP}(\mu) := \text{CRM}\left(\delta_1(d\pi)\mu(d\theta)\right).$$

The Bernoulli process, parametrized by finite, countable measure μ, assigns mass 1 to each of the members of a finite subset of the singletons in the support of μ;

[1] Technically, a CRM can also include a deterministic, non-atomic component; however we ignore this for simplicity.

[2] The directing measure may also have a fixed-location part, however we ignore this in our analysis.

there is no random-location component. Each singleton θ in the support of μ is included in the support of $\zeta \sim \text{BeP}(\mu)$ with probability $\mu\{\theta\}$.

We can interpret the beta-Bernoulli process as a *feature allocation* model – each counting measure ζ_n selects a subset of possible locations, or features, in the space Θ of all possible features. Since the beta process and the Bernoulli process form a conjugate pair, we can integrate out the directing beta process measure and work directly with the exchangeable sequence of features. The distribution over this sequence is known as the Indian buffet process (IBP).[3] Let each location $\theta \in \Theta$ parametrize a "dish," or feature, and then consider a buffet of all these dishes. The first customer selects a Poisson(α) number of dishes from the buffet, selected from some distribution H on Θ. When the nth customer arrives at the buffet, there are a finite number of previously sampled dishes and infinite unsampled dishes. He selects a each previously sampled dish with probability $\frac{m_i}{c+n}$, where m_i is the number of previous customers that have chosen that dish. He then selects Poisson $\left(\frac{c\alpha}{c+n}\right)$ new dishes, chosen according to H.

The features selected by each "customer," or data point, can be used as the basis of a latent feature model, by combining the IBP prior with an appropriate likelihood. A priori, both the number of latent features exhibited by a given data point, and the total number of latent features, are unknown.

Different choices of CRM may be substituted for the beta-Bernoulli pair, each yielding different properties. The three-parameter Indian Buffet Process replaces the beta process directing measure with a stable-beta process; the resulting random sequence exhibits power-law behavior in the total number of features underlying N observations [Teh and Görür, 2009]. A gamma process directing measure with a sequence of Poisson processes results in the infinite gamma-Poisson process [Titsias, 2008], a distribution over integer-valued sequences that can be interpreted as feature selection with repeats. Other exchangeable sequences constructed in this manner include the beta-negative binomial process [Zhou et al., 2012, Broderick et al., 2015] and the gamma-exponential process [Saeedi and Bouchard-Côté, 2011].

[3] In its original formulation [Griffiths and Ghahramani, 2011], the IBP imposes an ordering on the features which breaks the exchangeability in the more abstract feature-allocation representation. Here, we slightly modify the construction to refer to the more flexible feature allocation representation.

3 Exchangeable Sequences of Counting Measures with Arbitrary Marginals: The Restricted Indian Buffet Process

The exchangeable measures described in Section 2 offer significant modeling flexibility, with different choices of CRM yielding different properties. One property, however, cannot be avoided by judicious choice of CRM: either each measure ζ_n is associated with a countably infinite number of atoms with non-zero weight, or the distribution over the number of atoms with non-zero weight is marginally Poisson. This property is a direct consequence of the complete randomness of the underlying random measures μ and ζ_n: because the directing measure μ assigns independent finite masses $\pi_1, \pi_2, \ldots$ to a countably infinite subset $\Omega := \{\theta_1, \theta_2, \ldots\} \subset \Theta$, there will be a countably infinite number of locations where ζ_n might assign positive mass. If we marginalize out the π_k, these locations have equal, independent probability, resulting in a binomial distribution over the number of non-zero atoms (or selected features). With infinitely many potential features, the binomial distribution converges to a Poisson distribution.

Imposing an arbitrary distribution over the number of non-zero atoms in ζ_n must break the complete randomness. Suppose that we know that each measure ζ_n has exactly J atoms with positive mass. Next, suppose that we observe ζ_n has exactly J atoms with positive mass in some subset $A \subset \Theta$. We know that $\zeta(A^c) = 0$, so the probabilities of the entries in A and A^c are no longer independent.

The Restricted Indian Buffet Process (R-IBP), introduced in [Williamson et al., 2013], is a distribution over exchangeable subsets of features, with an arbitrary distribution over the number of features per observation. In the following sections, we describe several equivalent formulations for the R-IBP. We focus on restricted versions of the Indian Buffet Process, and in Section 4 describe how the ideas in this section can be applied to create other distributions with arbitrary marginals on the number of features.

3.1 Construction of the R-IBP using Restricted Bernoulli Processes

The R-IBP was originally constructed (in [Williamson et al., 2013]) by manipulating the beta-Bernoulli process construction of the IBP. The IBP can be represented as an i.i.d. collection of Bernoulli processes, directed by a beta process:

$$\begin{aligned} \mu &:= \textstyle\sum_i \pi_i \delta_{\theta_i} \sim \text{BP}(c, \alpha, H) \\ \zeta_n &:= \textstyle\sum_i z_{ni} \delta_{\theta_i} \overset{\text{i.i.d.}}{\sim} \text{BeP}(\mu) \end{aligned} \tag{1}$$

We can modify this construction to give a restricted model where the total mass of each measure ζ_n is constrained to be some integer J, by replacing the Bernoulli process in Equation 1 with a *restricted* Bernoulli process with restricting function δ_J, which we will write as R-BeP$(\mu, f = \delta_J)$ and which has law

$$P_{\text{R-BeP}}(\zeta_n; \mu, f = \delta_J) \propto \begin{cases} P_{\text{BeP}}(\zeta_n; \mu) & \text{if } \zeta(\Theta) = J \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

where $P_{\text{BeP}}(\zeta_n; \mu)$ describes the law of a Bernoulli process parametrized by μ. The distribution over atom sizes $Z_n = (z_{n1}, z_{n2}, \dots)$ is

$$\begin{aligned} &P_{\text{R-BeP}}\left(Z_n; \mu, f = \delta_J\right) \\ =&\frac{\prod_{i=1}^{\infty} \pi_i^{z_{ni}}(1-\pi_i)^{1-z_{ni}} \mathbb{I}(\sum_i z_{ni} = J)}{\sum_{z' \in \mathcal{Z}} \prod_i \pi_i^{z'_i}(1-\pi_i)^{1-z'_i} \mathbb{I}(\sum_i z'_i = J)}, \end{aligned} \quad (3)$$

where $\mathcal{Z}$ is the set of all collections of atom weights in the support of BeP(μ). The associated atom locations $\theta_i \overset{\text{i.i.d.}}{\sim} H$, as before. This restricted Bernoulli process can be seen as a nonparametric extension of the conditional Bernoulli distribution [Chen, 2000]. It is no longer a completely random measure: disjoint subsets of ζ_n depend on each other via the total mass.

More generally, we may wish to have some arbitrary distribution f on the number of atoms of ζ_n. To do so, we create an f-mixture of the distributions described by Equation 3, so that the probability of a sequence of weights Z_n associated with a measure ζ_n is given by

$$P_{\text{R-BeP}}(Z_n; \mu, f) \overset{d}{=} f\left(\textstyle\sum_i z_{ni}\right) P_{\text{R-BeP}}\left(Z_n; \mu, \delta_{\sum_i z_{ni}}\right). \quad (4)$$

Substituting restricted Bernoulli processes (Equations 3, 4) for the Bernoulli processes in Equation 1, we get the Restricted Indian Buffet Process:

$$\begin{aligned} \mu &\sim \text{BP}(c, \alpha, H) \\ \zeta_n &\overset{\text{i.i.d.}}{\sim} \text{R-BeP}(\mu, f). \end{aligned} \quad (5)$$

Since the ζ_n are i.i.d. given μ, marginalizing out μ results in an exchangeable sequence of counting measures ζ_n.

Importantly, setting $f(z) = \text{Poisson}(z; \alpha)$ does not recover the IBP. The IBP has Poisson(α) marginals over the total mass of each counting measure; however, if we condition on some previously observed measures $\zeta_1, \dots, \zeta_n$, the total mass of ζ_{n+1} is distributed according to a Poisson-binomial distribution.

3.2 Construction via Subsets of an Exchangeable Sequence

In Section 3.1, we saw how the R-IBP can be represented a sequence of restricted Bernoulli processes parametrized by a beta process directing measure. We can also construct the R-IBP directly from a sequence of IBP-distributed measures (or sets of features).

We can generate an IBP-distributed sequence $\boldsymbol{\zeta}^* = (\zeta_1^*, \zeta_2^*, \dots)$ of measures using the buffet-based predictive distribution from Section 2. Since this sequence is infinitely exchangeable, its law is invariant to shuffling the order of any finite subset [Aldous, 1983]. As a direct consequence, any infinite sub-sequence $\boldsymbol{\zeta}$ of $\boldsymbol{\zeta}^*$ is infinitely exchangeable. Thus, we can construct an R-IBP-distributed sequence $\boldsymbol{\zeta}$ by sampling a sequence $\boldsymbol{\zeta}^*$ according to an IBP, and including each proposed measure ζ_n^* into our sequence $\boldsymbol{\zeta}$ with probability $f(\zeta_n^*(\Theta))$.

This construction is equivalent to the restricted Bernoulli process method described in Section 3.1: if we integrate out the directing measure, a sequence of Bernoulli process-distributed measures is described by the IBP. However, unlike the Bernoulli process-based procedure, the IBP-based procedure requires sufficient statistics of the entire sequence $\boldsymbol{\zeta}^*$ to generate the next candidate for $\boldsymbol{\zeta}$. If we generate our proposed distributions based on the sum of the measures in $\boldsymbol{\zeta}$ rather than $\boldsymbol{\zeta}^*$, the resulting sequence of counting measures will not have the desired law and in general will not even be exchangeable (see [Williamson et al., 2013] for an example of such non-exchangeability).

3.3 Construction via Tilting the Bernoulli Process

A tilted CRM μ^* is a random measure obtained by scaling the law P_μ of a CRM μ on $(\Theta, \mathcal{A})$ by its total mass, according to some function $h(\Theta)$ [Lau, 2013], so that

$$P_{\mu^*}(A) := \frac{1}{\mathbb{E}[h(\mu(\Theta))]} \int_A h(\nu(\Theta)) P_\mu(d\nu). \quad (6)$$

For example, if $h(x) = e^{-\gamma x}$, then μ^* is said to be exponentially tilted. Exponentially tilting, or Esscher transforming [Gerber and Shiu, 1993], a CRM yields a different CRM [Lau, 2013]; for example an exponentially tilted α-stable process is equal (in distribution) to a generalized gamma process [Brix, 1999]. In general, however, a tilted CRM will not be a completely random measure. For example, if $h(x) = x^{-q}$ for some $q > 0$, then μ^* is said to be polynomially tilted and is no longer a CRM. Random measures constructed via polynomial tilting include the Pitman-Yor process [Pitman

and Yor, 1997] (obtained by polynomially tilting an α-stable process) and the beta-gamma process [James, 2005] (obtained by polynomially tilting a gamma process).

In Equation 4, the probability of a measure ζ_n under the restricted Bernoulli process is given by its probability under the Bernoulli process, scaled by a function f of its total mass. Thus the restricted Bernoulli process can be described as a tilted Bernoulli process[4] with tilting function $h(x) = f(x)$.

3.4 Construction via the Beta Prime Process and Invariance with respect to the Directing Measure

As shown in Equation 1, the IBP can be written as a sequence of Bernoulli processes with a beta process directing measure μ. If only a finite number N of measures ζ_n have been observed, our uncertainty about μ is described by a beta process with parameters $c + N$, $\frac{c\alpha}{c+N} H + \frac{1}{c+N} \sum_{n=1}^{N} \zeta_n$. As N tends to infinity, this posterior will tend towards the uniquely defined directing measure μ.

In contrast, the beta process directing measure μ for the R-IBP can *never* be uniquely determined, even with infinitely many observations. To show this, we can reconstruct the R-IBP in terms of a beta-prime process [Broderick et al., 2014]. A beta-prime process-distributed CRM $\tau := \sum_i w_i \delta_{\theta_i}$ is obtained by transforming the atoms π_i of a beta process-distributed CRM $\mu := \sum_i \pi_i \delta_{\theta_i}$ according to

$$w_i := \frac{\pi_i}{1 - \pi_i}.$$

We can therefore reformulate the R-IBP as

$$\begin{aligned} \tau &\sim \text{Beta-prime}(c, \alpha, H) \\ J_n &\sim f \\ P_{\text{R-BeP}}(Z_n; W, f) &= \frac{\prod_i w_i^{z_{ni}} \mathbb{I}(\sum_i z_{ni} = J)}{\sum_{z' \in \mathcal{Z}} \prod_i w_i^{z'_i} \mathbb{I}(\sum_i z_{ni} = J)} \\ \theta_i &\overset{\text{i.i.d.}}{\sim} H. \end{aligned} \tag{7}$$

where we use the notation $Z_n = (z_{ni})_{i=1}^{\infty}$ and $W = (w_i)_{i=1}^{\infty}$ to describe the atom sizes associated with ζ_n and τ respectively. The law $P_{\text{R-BeP}}(Z_n; W, f)$ in Equation 7 is invariant to rescaling the w_i by any e^{β}:

$$P_{\text{R-BeP}}(Z; W = \{w_i\}, J) \overset{d}{=} P_{\text{R-BeP}}(Z; \widetilde{W} = \{e^{\beta} w_i\}, J)$$

for any $\beta \in \mathbb{R}$. Intuitively, this scale invariance occurs because the R-IBP first chooses the *number* of atoms $J_n \sim f$ and then selects *where* in the support of τ (or equivalently μ) to place these atoms. Conditioned on J_n, the absolute scale of the weights w_i no longer matters; only their relative sizes are important.

The connection between the restricted IBP and the beta-prime process makes it possible to remove extra degree of freedom present in the previous constructions by fixing the scale through a normalized beta-prime process. While this is theoretically appealing – it leads to a unique directing measure for each infinite sequence – it offers little practical advantage, due to the lack of a tractable representation for such a process.

This invariance with respect to scaling the beta-prime process can equivalently be interpreted to an invariance with respect to exponentially tilting the Bernoulli process. Rescaling the beta-prime process weights w_i by e^{β} is equivalent to rescaling the atoms π_i of the corresponding beta process according to the nonlinear function

$$\pi'_i = \frac{\pi_i e^{\beta}}{\pi_i e^{\beta} + 1 - \pi_i}. \tag{8}$$

Equation 8 describes the Esscher transform of a Bernoulli random variable, and this operation therefore corresponds to exponentially tilting the (unrestricted) Bernoulli process, as described in Section 3.3. We will make use of this formulation of the invariance in Section 5.2 to improve simulation efficiency.

4 Extensions and Variations

In Section 3, we focused on exchangeable models based on the IBP. However, the same ideas apply to exchangeable models based on other completely random measures. One can also relax the exchangeability assumption to allow partial exchangeability, leading to models for data with observation-specific covariates.

4.1 Restricted Partially-Exchangeable Counting Measures

In Section 3, we assumed that our data (or more concretely, the underlying counting measures ζ_n) were exchangeable. We can easily modify the R-IBP to yield a partially exchangeable model conditioned on observation-specific covariates. For example, if each observation has an associated label indicating a group membership $m \in \{1, \ldots, M\}$, then each group could

[4] Arguably, the tilted Bernoulli process nomenclature is perhaps a better fit for the R-IBP, since for arbitrary f the "restricted Bernoulli process" is in fact a mixture of restricted distributions. However, the tilting interpretation was not apparent when the models described in this paper were first introduced in [Williamson et al., 2013], so we continue to use original term "restricted" for consistency.

have group-specific restricting distribution f_m, so that

$$\begin{aligned}\mu \sim& \text{BP}(c, \alpha, H)\\ \zeta_n \overset{\text{i.i.d.}}{\sim}& \text{R-BeP}(\mu, f_{m(n)}).\end{aligned}$$

where $m(n)$ describes the group for observation n. The resulting sequence would be partially exchangeable in that distribution over features is invariant to permuting measures in the same group.

Such a model, which we evaluate experimentally in Section 7, would be appropriate where we have observation-specific information about the number of non-zero features. For example, we might wish to construct a topic model with different distributions over the number of topics depending on the type of document (feature stories may focus on a few topics, summaries may cover many topics). An image model may be informed by textual descriptions indicating how many objects are present.

4.2 Restricted Exchangeable Counting Measures based on Different Completely Random Measures

In Section 3.1 we constructed the Restricted IBP from the beta-Bernoulli process representation of the IBP, and replacing the Bernoulli process with a restricted Bernoulli process. In general, we could pick any conjugate pair of CRMs to generate an exchangeable sequence $(\zeta_n)_{n=1}^N$, [Orbanz, 2009]:

$$\begin{aligned}\mu \;\sim\; & \text{CRM}\left(\nu(d\pi, d\theta)\right)\\ \zeta_n \overset{\text{i.i.d.}}{\sim}& \text{CRM}(g(\mu)).\end{aligned} \tag{9}$$

If the support of the random measures ζ_n in Equation 9 consists almost surely of measures with a finite number of non-zero atoms, the resulting exchangeable sequence of measures can be interpreted as selecting (and possibly weighting) a finite number of features from a countably infinite number of potential features. Examples of such feature selection models include the beta-negative binomial process [Zhou et al., 2012, Broderick et al., 2015] and the gamma-Poisson process [Titsias, 2008], both of which weight selected features using a positive integer which can be interpreted as the number of instances of that feature. All such models exhibit the property that the total number of selected features—excluding repeated features—is (marginally) Poisson-distributed, a direct consequence of the complete randomness of the underlying random measures (as described in Section 3).

For any such exchangeable model, we can restrict the support of the ζ_n to generate an exchangeable sequence of measures with restrictions on the total mass or the number of non-zero atoms. If $\zeta_n \sim \text{BeP}(\mu)$, the total mass and the number of atoms are the same. Other choices of CRM present a wider range of possible restrictions, as the support of the unrestricted CRM will not be limited to binary counting measures. We give three examples.

Restricting the number of unique features If ζ_n is distributed according to a Bernoulli process, imposing a distribution over the total mass $\zeta_n(\Theta)$ is equivalent to imposing a distribution over the number of atoms with positive mass. For more general CRMs, these two cases are not the same. We first consider imposing a function $f\left(\sum_k \mathbb{I}(z_k > 0)\right)$ on the number of atoms with non-zero mass. This yields a restricted CRM with law

$$\begin{aligned}&\text{R-CRM}^{(1)}\left(\zeta := \sum_k z_k \delta_{\theta_k}; g(\mu), f\left(\textstyle\sum_k \mathbb{I}(z_k > 0)\right)\right)\\ \propto& f\left(\textstyle\sum_k \mathbb{I}(z_k > 0)\right) \text{CRM}(\zeta; g(\mu)).\end{aligned} \tag{10}$$

where $CRM(g(\mu))$ is the law of the corresponding unrestricted CRM. This is equivalent to restricting the number of unique features present in a feature selection model, ignoring weights or repetitions.

Restricting the mass of each measure We can also impose a function $f(\sum_k z_k)$ on the total mass $\zeta_n(\Theta)$ of each measure, yielding

$$\begin{aligned}&\text{R-CRM}^{(2)}\left(\zeta := \sum_k z_k \delta_{\theta_k}; g(\mu), f\left(\textstyle\sum_k z_k\right)\right)\\ \propto& f\left(\textstyle\sum_k z_k\right)) \text{CRM}(\zeta; g(\mu)).\end{aligned} \tag{11}$$

If the ζ_n has integer-valued weights, we can interpret z_{ni} as the number of repetitions of the ith feature in the nth measure. This restriction corresponds to restricting the total number of features, including repetitions.

A special case of the construction in Equation 11 is when the directing measure μ is distributed according to a gamma process with parameter αH for some probability measure H, and the ζ_n are distributed according to a Poisson process with mean measure μ [Titsias, 2008]. In this case, if we restrict the total mass of each measure ζ_n following Equation 11, the distribution over the ζ_n is equivalent to the following Dirichlet process-multinomial model:

$$\begin{aligned}\rho \sim& \text{DP}(\alpha, H)\\ J_n \sim& f\\ \zeta_n \sim& \text{Mult}(\rho, J_n)\end{aligned}$$

Restricting the total mass and the number of unique features We can combine the previous examples to specify more complex restrictions. For example, we could generate an exchangeable sequence of binary counting measures with total mass J by letting the ζ_n be a CRM with integer-valued atoms, and restricting both the number of non-zero elements to be J and the values of the non-zero elements to be one. If the directing measure μ is distributed according to a gamma process, and the ζ_n are distributed according to a Poisson process, the features associated with each ζ_n corresponds to sampling J features using conditional Poisson sampling from a Dirichlet-distributed random measure.

5 Simulation from the R-IBP

In Section 3, we presented several constructions for the R-IBP. Since the atom locations θ_i are independent of the atom locations z_{ni}, this problem can be reduced to simulating the feature indicators Z_n. In this section, we present four methods for sampling from the R-IBP, which we compare empirically in Section 5.5:

1. (Section 5.1)An approximate method where atoms of ζ_n are sequentially added, conditioned on an approximation to μ.
2. (Section 5.2) An approximate method based on rejection sampling entire measures ζ_n, conditioned on an approximation μ.
3. (Section 5.3) An exact retrospective method based on rejection sampling entire measures ζ_n, using an adaptive truncation of μ.
4. (Section 5.4) An exact method based on rejection sampling in a collapsed representation.

The first three methods decompose the R-IBP into a beta (or beta-prime) process-distributed directing measure and a sequence of restricted Bernoulli process. Of course, we cannot represent the entire infinite-dimensional beta process random measure μ (or beta-prime process measure τ). Instead, we use a finite-dimensional approximation $\tilde{\pi} = (\tilde{\pi_1}, \dots \tilde{\pi}_I)$ to the full sequence $\pi_1, \pi_2, \dots$ of atom sizes, obtained using one of two approximation methods:

– *Weak Limit* One approach is to use a finite vector of beta random variables that converges (in a weak limit sense) to the (unordered) atom sizes of a beta process-distributed measure [Zhou et al., 2009]. We approximate the atom sizes of μ with a vector $\tilde{\pi} = (\tilde{\pi}_1, \dots, \tilde{\pi}_I)$

$$\tilde{\pi}_i \overset{i.i.d.}{\sim} \text{Beta}\left(\frac{c\alpha}{I}, c - \frac{c\alpha}{I}\right). \tag{12}$$

– *Size-Ordered Stick-breaking Representation* Another approach is to transform the arrival times of a unit-rate Poisson process based on the beta process Lévy measure [Rosiski, 2001, Ferguson and Klass, 1972]. This approach gives exact samples from the size-ordered atoms of the beta process. In the special case where $c = 1$, this approach yields a simple stick-breaking construction [Teh et al., 2007]:

$$u_i \sim \text{Beta}(\alpha, 1) \qquad \pi_i = \prod_{j=1}^{i} u_i. \tag{13}$$

We can obtain a truncated approximation $\tilde{\pi}$ by taking the first I strictly-ordered atom sizes generated according to this procedure.

In the remainder of the paper, we restrict our analysis to the case where $c = 1$, since it allows us to use the representation in Equation 13 and is the most commonly used setting in the literature. We note that for the approximate algorithms described in Sections 5.1 and 5.2, and for the posterior inference methods in Section 6, it is straightforward to substitute the more general weak limit representation in Equation 12.

5.1 Approximate Sampling using Inclusion Probabilities

Given an approximation $\tilde{\pi}$ to the directing measure μ for the IBP, the presence or absence of each feature can be sampled independently according to $z_{ni} \sim \text{Bernoulli}(\tilde{\pi}_i)$. In the R-IBP, the atoms can no longer be sampled independently; however given the total number of features J_n, we can construct a sequential procedure for selecting atoms given $\tilde{\pi}$ by computing the inclusion probabilities $P(z_{ni}|\tilde{\pi}, k_n)$ of each feature [Aires, 1999].[5]

If I feature indicators $z_1, \dots, z_I$ are sampled independently with unequal probabilities $\tilde{\pi_1}, \dots, \tilde{\pi}_I$ then, conditioned on the total sum $\sum_i z_i = J$, the probability $\eta_{k;J} = P(z_k = 1 | \sum_i z_i = J)$ that the kth feature is included is given by

$$\eta_{k;J} = \tilde{\pi}_k \frac{S_{J-1}^{I-1}(\tilde{\pi}_1, ..., \tilde{\pi}_{k-1}, \tilde{\pi}_{k+1}, ..., \tilde{\pi}_I)}{S_J^I(\tilde{\pi}_1, ..., \tilde{\pi}_I)}, \tag{14}$$

where S_J^I corresponds to the unconditional probability $P(\sum_i z_i = J)$ of selecting exactly J features:

$$S_J^I = \textstyle\sum_{s \in A_J(I)} \prod_{k \in s} \tilde{\pi}_k \prod_{j \ni s} (1 - \tilde{\pi}_j), \tag{15}$$

[5] More generally, [Hanif and Brewer, 1983] lists over 50 ways to sample without replacement with unequal weights in the finite case.

where $A_J(I)$ is the set of all subsets of size J that can be drawn from the the I features. The $\{S_J^I\}$ can be computed in $O(I^2)$-time using the following recursion:

$$S_J^I = \tilde{\pi}_I S_{J-1}^{I-1}(\tilde{\pi}_1, ..., \tilde{\pi}_{I-1}) + (1-\tilde{\pi}_I) S_J^{I-1}(\tilde{\pi}_1, ..., \tilde{\pi}_{I-1}). \tag{16}$$

Given the marginal inclusion probabilities η_{ik}, we have a draw-by-draw algorithm for sampling a set of feature indicators Z_n given a vector of probabilities $(\pi_1, \ldots, \pi_I)$. We first select the total number J_n of features for the nth observation by sampling $J_n \sim f$. Given $J_n > 0$, we use Equation 14 calculate and apply the inclusion probabilities for each feature to select a feature. We then recurse on the remaining $I-1$ features to select the second feature, and so on. With appropriate caching, we can use Equation 16 to compute (and cache) all of the elements $\eta_{k,J}$ in $O(I^3)$ time.

The error due to using an I-dimensional approximation depends I and the distribution over the cardinalities f; in Appendix A.1, we derive bounds for the inclusion probabilities.

5.2 Approximate Sampling from using Rejection Sampling with Tilted Bernoulli Process Proposal

While the inclusion probabilities in Equation 14 are available analytically, they are expensive to compute. Conversely, we can very cheaply sample feature indicators $Z^* \sim \text{BeP}(\tilde{\pi})$, by sampling each $z_i^* \sim \text{Bernoulli}(\tilde{\pi}_i), i = 1, \ldots, I$. We can use these samples from the Bernoulli process as proposals in a rejection sampler for the restricted Bernoulli process: If f is the desired distribution over the total number of features, we accept a proposal Z^* with probability $f(\sum_i z_i^*)$. However, this approach will give low acceptance rates—and therefore high computational cost—if the target distribution f differs significantly from the Poisson(α) distribution implied by the IBP.

We can improve the acceptance rate—and hence ameliorate the computational costs—by exponentially tilting the Bernoulli process likelihood. As we saw in Section 3.4, exponentially tilting the underlying Bernoulli process does not change the law of a restricted Bernoulli process, i.e.

$$\begin{aligned}&\text{R-BeP}((\tilde{\pi}_1, \ldots, \tilde{\pi}_I)) \\ &\stackrel{d}{=} \text{R-BeP}\left(\left(\frac{e^\beta \tilde{\pi}_1,}{e^\beta \tilde{\pi}_1 + 1 - \tilde{\pi}_1} \ldots, \frac{e^\beta \tilde{\pi}_1,}{e^\beta \tilde{\pi}_I + 1 - \tilde{\pi}_I}\right)\right).\end{aligned}$$

Exponentially tilting the Bernoulli process *does* however change the distribution over the number of features selected in the unrestricted setting [Brostrom and Nilsson, 2000]. Thus, manipulating the tilting parameter β changes the proportion of proposals that are accepted. To maximize the likelihood of selecting exactly J features, we set β to be the unique solution to

$$J = \sum_{i=1}^{I} \frac{e^\beta \tilde{\pi}_i}{e^\beta \tilde{\pi}_i + 1 - \tilde{\pi}_i}.$$

Thus, we can first sample the number of features J_n for the nth observation from f, Esscher transform the weights $\tilde{\pi}_i$ to maximize the chance of getting exactly J_n features, and then sample Z_n using the transformed weights. For computational efficiency, the transformed weights can be cached for each value of J_n.

Since this approach is based on a finite approximation to the true directing measure, it will only yield approximate samples from the R-IBP; we discuss the errors introduced in Appendix A.2. It can however be adapted to give exact samples from the R-IBP, as we will show in Section 5.3

5.3 Exact Sampling from using Rejection Sampling with Tilted Bernoulli Process Proposal

Since the rejection sampler described in Section 5.2 is based on a finite-dimensional approximation to the beta process-distributed μ, it will not generate exact samples from the R-IBP. However, if we use a truncated stick-breaking representation to approximate μ, we can dynamically adapt the truncation level, to obtain exact samples from the R-IBP via retrospective sampling [Papaspiliopoulos and Roberts, 2008].

When working with a fixed truncation, there are two forms of error that might occur. Recall that a given proposal ζ^* is a truncated version of a "correct" proposal measure, and we make our accept/reject decision based on I atoms of this this "correct" proposal (corresponding to the I largest atoms in μ). Consider the special case where $f = \delta_J$, i.e. proposals should be accepted if and only if the total number of features is J. Let Ω be the support of the full, untruncated measure μ, and let $A \subset \Omega$ contain the I largest atoms in μ. If the "correct" proposal ζ^* has $\zeta^*(A^c) > 0$ and $\zeta^*(A) = J-k$, the truncated version of this proposal would be erroneously rejected (since in the truncated representation, we only see $\zeta^*(A) < J$ features) . Conversely, if $\zeta^*(A) = J$ but $\zeta^*(A^c) > 0$, the proposal would be erroneously accepted because the truncated approximation doesn't "see" the features in A^c.

In Appendix A.2, we show that a proposal with $\sum_{i=1}^I z_i^* < J$ should have been accepted with probability Poisson$(J - \sum_{i=1}^I z_i^*; \pi_I \alpha)$, where π_I is the smallest atom in the truncated representation, and a proposal where $\sum_{i=1}^I z_i^* = J$ should have been rejected

with probability $1 - \exp(-\pi_I \alpha)$. We can use these probabilities to probabilistically reject proposals where $\sum_{i=1}^{I} z_i^* = J$, and propose decreasing our truncation level when $\sum_{i=1}^{I} z_i^* < J$:

- Sample an initial truncated vector of feature probabilities $\tilde{\pi} = (\pi_1 \dots \pi_I)$ according to the size-ordered stick breaking representation of the Beta process (Equation 13).
- For $n = 1, \dots, N$, repeat the following proposal step until we have accepted a sequence of feature indicators Z_n:
  - Sample $z_1^* \dots z_I^* \sim \pi_1 \dots \pi_I$, and compute the sum $K^* = \sum_{i=1}^{I} z_i^*$.
    - If $K^* > J$, reject Z^*.
    - If $K^* = J$, accept with probability $\exp(-\pi_I \alpha)$.
    - If $K^* < J$,
      - Reject with probability $1 - \text{Poisson}(J - \sum_i z_{ni}; \pi_I \alpha)$.
      - Otherwise, expand the representation by sampling new π_i, z_i^* for $i = I+1, I+2, \dots$ according to the stick breaking representation, until $\sum z_n^* = J$. Accept the resulting Z^*, and update I and $\tilde{\pi}$ to incorporate the new atoms.

We can adapt this procedure to arbitrary restricting function f, by first sampling a feature count $J_n \sim f$ for each observation. The growth of the truncation level I will depend on f; if $J_n \sim f$ is large then we may have to expand to very large truncation levels I. Specifically, starting with a truncation level too small may result in many, many rejections before the truncation level is sufficiently expanded. However, the samples that we do accept will be from the correct R-IBP prior.

5.4 Sub-sampling from an Exchangeable Model

In Section 3.2, we showed that the R-IBP can be constructed by subset selection of an IBP-distributed sequence of binary vectors. This directly suggests an exact scheme for generating a sequence $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \dots) \sim \text{R-IBP}(\alpha, c, H)$. We generate a sequence $(\zeta_1^*, \zeta_2^*, \dots)$ according to the IBP predictive distribution, and include each ζ_n^* in our sequence $\boldsymbol{\zeta}$(with probability $P(\zeta_n^* \in \boldsymbol{\zeta}) = f(\zeta_n^*(\Theta))$.

5.5 Empirical Comparison of Simulation Methods

We empirically compared the simulation approaches described in this Section by measuring the number of rejections and CPU time required to generate samples from the R-IBP with concentration parameter $\alpha = 5$ and restricting function $f = \delta_J$ for $J = \{2, 5, 8\}$. We generated 25 samples of 100 observations from each of five approaches: approximate inclusion sampling (Section 5.1); approximate uncollapsed rejection sampling, using both an untilted ("Approx Uncollapsed") and a tilted ("Approx Tilted") Bernoulli process (Section 5.2); exact uncollapsed rejection sampling using a tilted Bernoulli process (Section 5.3); and exact collapsed subset selection (Section 5.4).

Rejections per 100 observations are shown in figure 1. As expected, rejection rates are lowest for $J = 5$ because $\alpha = 5$. Inclusion sampling, a draw-by-draw procedure, has no rejections, and tilting significantly reduces the number of rejections—and the variance in the number of rejections—regardless of J. The other procedures all have large rejection rates varying over several orders of magnitude. Figure 2 shows CPU time on a standard laptop. Again, the time to 100 acceptances is shortest when J is equal to the expected value of features α. The approximate methods are faster than the exact methods, and the approximate tilted rejection sampler is the fastest, closely followed by the approximate sampler that uses inclusion probabilities.[6]

Figures 3 and 4 show the mean of the empirical feature probabilities, sorted in descending order, for various truncation levels for $J = 5$ and $J = 8$. When $J = 5$, the exact samplers instantiate between 30-40 hidden features. The mean probabilities of the approximate methods follow the exact probabilities closely even with truncations of $I = 10$ or $I = 20$, with only slight overestimation to account for the fewer features. When $J = 8$, the exact methods tend to instantiate 35-45 features. The approximate methods have a noticeable overestimation of feature probabilities when the truncation I is too small (e.g. $I = 10$). However, as the truncation is increased, the mean probabilities from the approximate methods again closely match the exact methods. Interestingly, there do not seem to be large differences between the different approximate methods.

These explorations suggest that the approximate methods can be accurate, computationally-efficient alternatives when the truncation is set to a reasonable value, justifying the use of such approximations for posterior inference in Section 6.

[6] The wall-clock time difference between the draw-by-draw procedure using inclusion probabilities and the approximate rejection samplers may be due in part due to Matlab vectorization; a draw-by-draw procedure requires a loop to sequentially compute whether a feature is present while the rejection sampler can sample all elements of Z_n together.

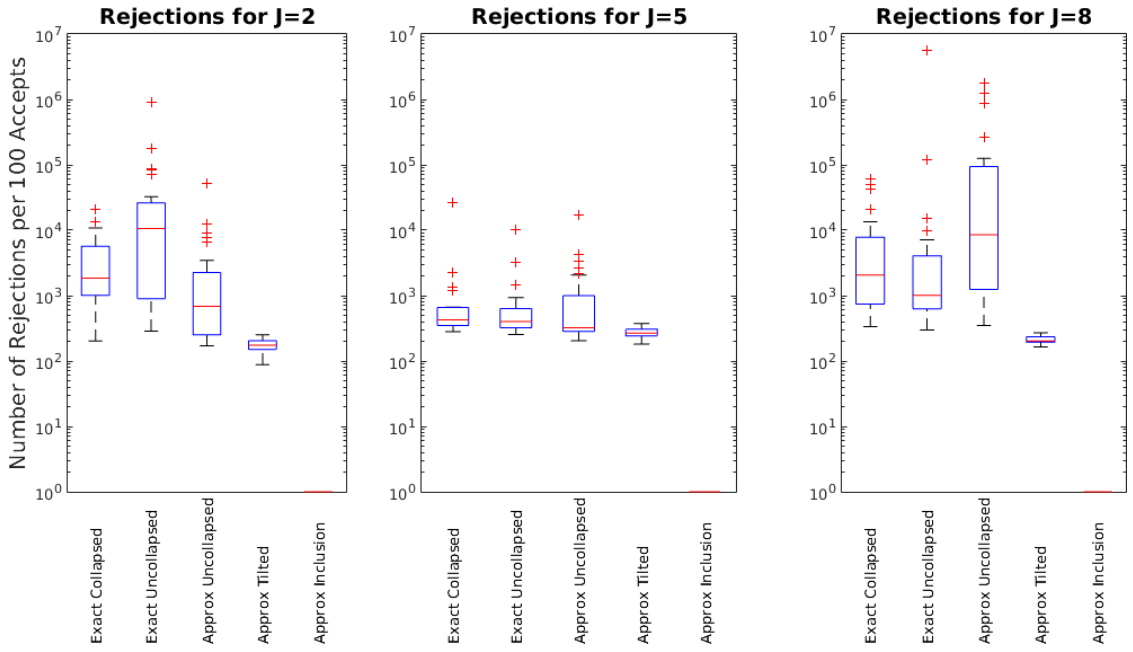


Fig. 1: Rejections per 100 Acceptances

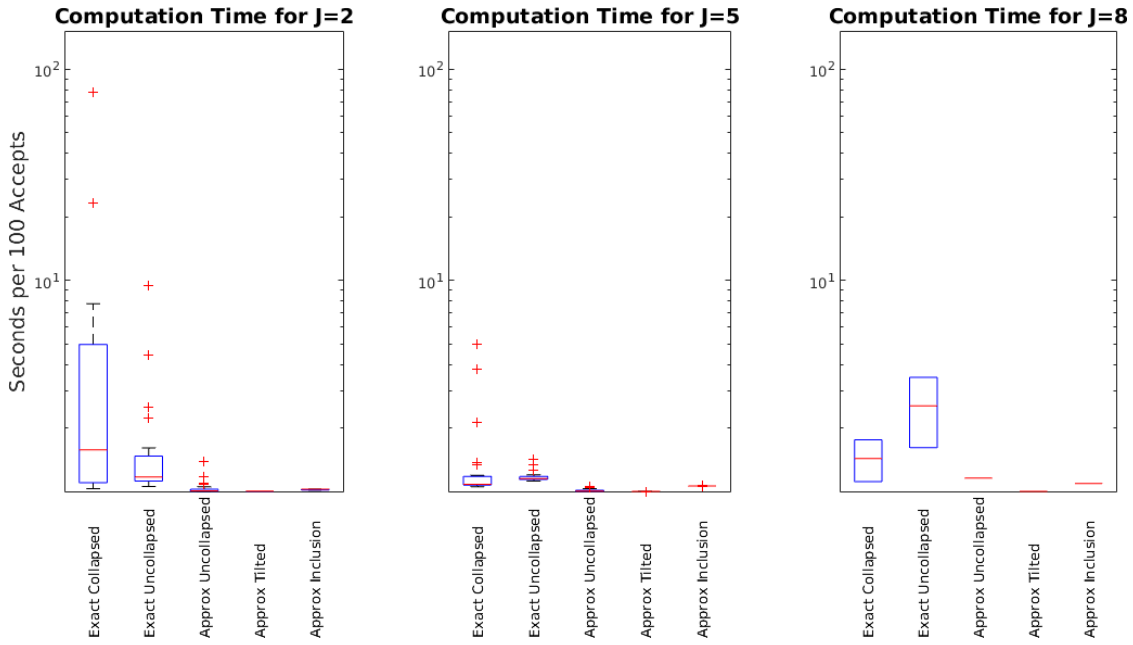


Fig. 2: Time required for 100 Acceptances on a standard laptop

6 Posterior Inference in the R-IBP

In this section, we present two approaches for posterior inference in the R-IBP. In Section 6.1, we present a MCMC-based approach related to the simulation techniques described in Section 5, and in Section 6.2 we present a computationally faster hybrid variational/MCMC approach for posterior inference.

6.1 MCMC-based Posterior Inference in the R-IBP

Mirroring the methods for prior simulation in Sections 5.2 and 5.1, we propose a Gibbs sampler that operates in an uncollapsed representation using a finite-dimensional approximation $\tilde{\pi} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_I)$, obtained either via a weak limit approximation or via truncation in a stick-breaking process. Our Gibbs sampler alternates between sampling from the conditional distribu-

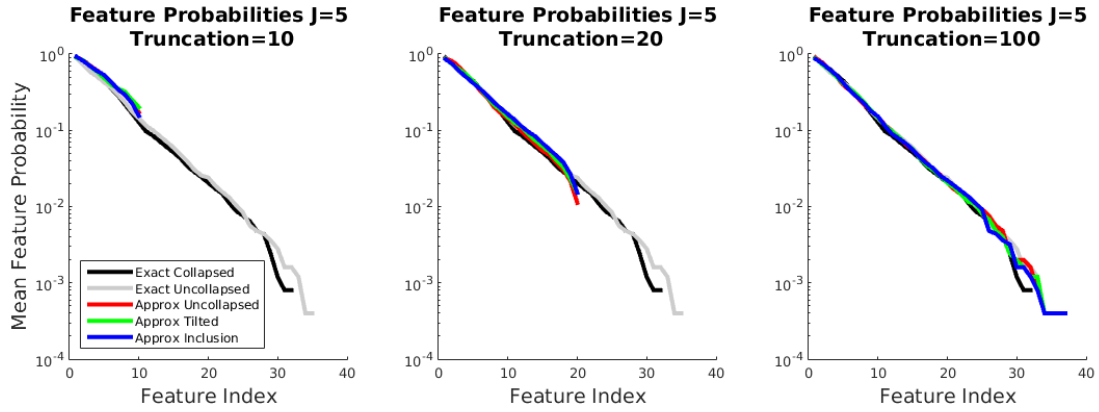


Fig. 3: Mean of empirical feature probabilities, sorted in descending order for varying truncation levels and $J = 5$

.

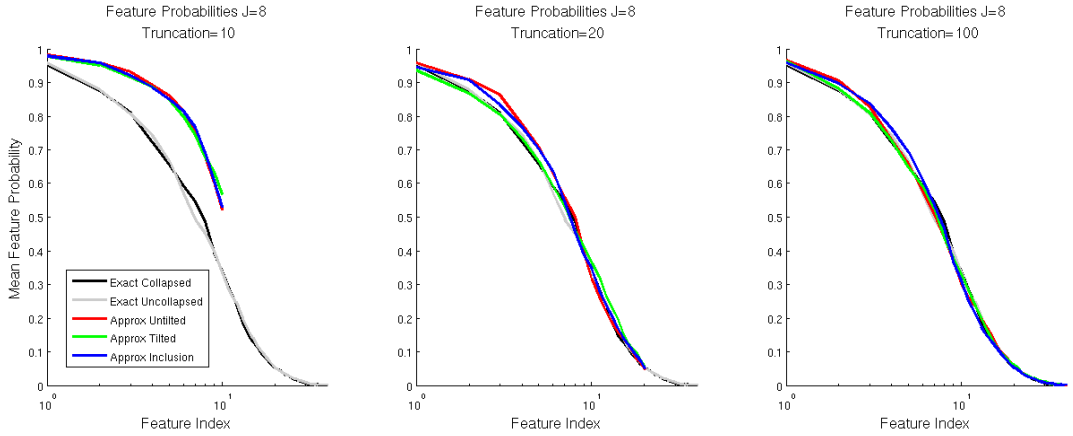


Fig. 4: Mean of empirical feature probabilities, sorted in descending order for varying truncation levels and $J = 8$

.

tion of Z given $\tilde{\pi}$ and the data X and, sampling from the conditional distribution of $\tilde{\pi}$ given Z and X.

Sampling $\mathbf{Z}|\tilde{\pi}$ If the distribution f is not degenerate on a single point, we can use the inclusion probabilities described in Section 5.1, combined with f and the data likelihood $P(X|Z,\Psi)$, to Gibbs sample the presence or absence of a single feature in the nth observation, using the conditional probabilities

$$\begin{aligned} &P(z_{ni}=1|\{\tilde{\pi}_1,\ldots,\tilde{\pi}_I\},\{Z_n\}\setminus z_{ni},X,\Psi) \\ &\propto \tilde{\pi}_i \frac{S_0^{I-k_{n,-i}-1}(\{\tilde{\pi}_k : z_{nk}=0, k\neq i\})}{S_1^{I-k_{n,-i}}(\{\tilde{\pi}_k : z_{nk}=0 \text{ or } k=i\})} \\ &\quad \cdot f(k_{n,-i}+1)P(X|z_{ni}=1,\{Z_n\}\setminus z_{ni},\Psi) \\ &P(z_{ni}=0|\{\tilde{\pi}_1,\ldots,\tilde{\pi}_I\},\{Z_n\}\setminus z_{ni},X,\Psi) \\ &\propto f(k_{n,-i})P(X|z_{ni}=0,\{Z_n\}\setminus z_{ni},\Psi), \end{aligned} \tag{17}$$

where $k_{n,-k}=\sum_{j\neq k} z_{nj}$.

If the distribution f *is* degenerate on a single value J, we cannot construct a Gibbs sampler that sequentially turns elements on or off; doing so would change the number of features. Instead, we use the inclusion probabilities to sample the location of each of the active features for an observation (i.e. each non-zero atom in ζ_n), conditioned on the locations of the other $J-1$. Let ℓ_{nk} be the location of the jth feature entry. Then

$$\begin{aligned} &P(\ell_{nj}=i|\{\tilde{\pi}_1,\ldots,\tilde{\pi}_I\},\ell_{-nj},X,\Psi) \\ &\propto \tilde{\pi}_i \frac{S_0^{I-k_{n,-i}-1}(\{\tilde{\pi}_k : z_{nk}=0, k\neq i\})}{S_1^{I-k_{n,-i}}(\{\tilde{\pi}_k : z_{nk}=0 \text{ or } k=i\})} \\ &\quad \cdot f(k_{n,-i}+1)P(X|\ell_{nj}=i,\ell_{-nj},\Psi) \end{aligned} \tag{18}$$

The Gibbs sampling steps described in Equations 17 and 18 only change a single element of a single feature indicator sequence Z_n at a time. This can lead to slow mixing. We augment these Gibbs sampling steps with Metropolis Hastings proposals generated from the prior, using either the rejection sampling approach of

Section 5.2 or the inclusion probability approach of Section 5.1 to propose an entirely new row Z_n.

Sampling the latent measure Once we have sampled our feature allocations $\{Z_n\}$, we must resample our latent feature weights $\tilde{\pi}$. Unfortunately, since the beta process is not conjugate to the restricted Bernoulli process, we cannot directly Gibbs sample $\tilde{\pi}$ given $\{Z_n\}$. Instead, we use Metropolis-Hastings steps. Since the posterior distribution over $\tilde{\pi}$ given $\{Z_n\}$ is likely to be similar to the poster distribution in the unrestricted IBP, we use the posterior distribution from the unrestricted IBP as a proposal distribution. The acceptance probability depends on the R-IBP likelihood (Equations 3 and 4).

6.2 Hybrid Variational Inference in the R-IBP

The standard variational approach for the IBP [Doshi et al., 2009] uses a mean-field approximation which places independent distributions $q(z_{ni})$ over each feature assignment z_{ni}. Using such a factored distribution is straightforward because each assignment z_{ni} is drawn independently given the weight π_i. However, variational inference in the R-IBP is challenging because fixing the number of active features J_n introduces dependence between the z_{ni}, and because the implied prior distributions over the marginal inclusion probabilities η_{ik} are complex. Further, the invariance of the likelihood to scaling the directing measure, as described in Section 3.4, can lead to inefficiencies in exploring the state space and computational difficulties due to very small atom sizes that may occur at certain scales.

We propose a hybrid variational for inference in the R-IBP that combines variational distributions over the feature assignments and model parameters with MCMC inference over the directing measure. As in Section 6.1, we work with a finite dimensional approximation $\tilde{\pi}$ to the directing measure, and alternate between resampling the $\tilde{\pi}_i$ and updating the variational posterior on the other variables. We demonstrate this approach using a linear-Gaussian likelihood, where the data X are assumed to be generated by $X_n = \sum_i z_{ni}\theta_i + \epsilon$, where each θ_i is an $I \times D$ feature matrix sampled from the beta process base measure $H := \text{Normal}(0, \sigma_A^2)$, and ϵ is a $N \times D$ matrix of independent noise drawn from $\text{Normal}(0, \sigma_n^2)$. We note that the inference of the latent features θ_i is the same as in the standard IBP, and other likelihood models developed for the IBP can be substituted.

The variables in the variational update are the feature assignments Z, the feature values θ_k, and the count of active features per observation J_n. We consider the following mean field approximation for the variational inference:

- $q_{\phi_i}(\theta_i)$ independent Gaussian distributions with mean ϕ_i, variance Φ_i on the posterior of the feature value vector θ_i.
- $q_{\nu_{ni}}(z_{ni})$ independent Bernoulli distributions, where ν_{ni} is the probability that z_{ni} is active.
- $q_{\gamma_{nk}}(J_n)$ multinomial distributions over the number of features in observation n, where γ_{nk} is the probability that observation n has k active features.

Let $W = \{\phi, \Phi, \nu, \gamma\}$ be the set of variational parameters, and let $V = \{\{\theta_i\}, \{Z_n\}, \{J_n\}\}$ be the set of variables. Because the actual and variational distributions belong to the exponential family, coordinate ascent on the variational parameters corresponds to setting the variational distribution $\log(q_{W_i}) = E_{W_{-i}}[\log(P(W, V|X, \Psi))]$, where Ψ denotes the set of hyper-parameters $\{\sigma_n^2, \sigma_a^2, \alpha, f\}$ [Wainwright and Jordan, 2008].

We focus on providing the variational updates for the parameters associated with the Z_n and J_n, as the updates for the parameters associated with the θ_i (i.e., ϕ_k, Φ_k) are exactly the same as in [Doshi et al., 2009]. The update for γ_{nk} is:

$$\begin{aligned}\log(q_{\gamma_n}(J_n)) =& E_Z[\log P(J_n) + \log P(Z_n|\tilde{\pi}, J_n)] \\ =& \textstyle\sum_{k=1}^K I(J_n = k)[\log(f_{nk}) + \nu_{ni}\log(\eta_{ik}) \\ &+ (1-\nu_{ni})\log(1-\eta_{ik})]\end{aligned}$$

where f_{nk} is the prior probability that observation n has k elements. Exponentiating and normalizing, we recover the posterior parameters γ_{nk}.

The update for variational parameters ν_{ni} for the assignments Z are also straightforward given the inclusion probabilities η_{ik}:

$$\begin{aligned}\log(q_{\nu_{ni}}(z_{ni})) =& E_{J_n, Z_{-ni}, A}[\log(P(z_{ni}|Z_{-ni}, \tilde{\pi}, J_n)) \\ &+ \log(P(X_n|Z_n, \{\theta_i\}, \sigma_n^2))]\end{aligned} \tag{19}$$

where the second term is again exactly the same as in [Doshi et al., 2009]. For the first term, we can write

$$\begin{aligned}& E_{J_n, Z_{-ni}}[\log(P(z_{ni}|\{Z_n\} \setminus z_{ni}, \tilde{\pi}, J_n))] \\ =& E_{J_n, Z_{-ni}}[z_{ni} I(J_n = k)\log(\eta_{ik}) \\ &+ (1-z_{ni}) I(J_n = k)\log(1-\eta_{ik})] \\ =& z_{ni}\textstyle\sum_k \gamma_{nk}\log(\frac{\eta_{ik}}{1-\eta_{ik}}) + c\,.\end{aligned} \tag{20}$$

Substituting Equation 20 into Equation 19, we derive the update

$$\xi = \sum_k \gamma_{nk} \log(\tfrac{\eta_{ik}}{1-\eta_{ik}}) - \tfrac{1}{2\sigma_n^2}\Big(-2\phi_i X_n^T + Tr(\Phi_i) + \phi_i\phi_i^T + 2\phi_i(\sum_{j\neq i}\nu_{nj}\phi_j^T)\Big)$$
$$\nu_{ni} = \tfrac{1}{1+\exp(-\xi)}\,. \quad (21)$$

Our inference process iterates through the following steps:

1. Compute the partial variational posterior on Z_n, θ_i, and J_n, as above.
2. Sample values of Z_n, θ_i, and J_n from the variational posterior.
3. Sample new values of $\tilde{\pi}$ given the sampled Z_n, using the Metropolis-Hastings technique described in Section 6.1.

7 Evaluation

We show a variety of evaluations to demonstrate the value of using the R-IBP on real and synthetic data when we have some knowledge about the marginals on the number of non-zero entries.

7.1 Exploration with Synthetic Data

To explore the ability of the R-IBP to recover latent structure, we trained the IBP, the R-IBP with an appropriate restricting distribution, and the partially-exchangeable R-IBP with labeling information described in Section 4.1 on synthetic datasets using a linear Gaussian model.

7.1.1 Knowledge about the Number of Latent Features Assists with Parameter Recovery.

One reason for using the R-IBP is when we have strong ideas of what a "feature" corresponds to, coupled with strong information about the number of such features. While an IBP may be able to model the data using a collection of features, these may not correspond to our preconceived notions of features – for example, the IBP might use multiple features where we expect a single feature.

To explore this, we generated a toy dataset with a total of 15 latent features. We generated 400 observations with 14 of the 15 latent features, and 100 observations with a single latent feature. We assumed a user-defined, observation-specific distribution over the number of features (corresponding to the partially exchangeable model described in Section 4.1). Specifically, if an observation X_n contains k_n features, we used a restricting distribution f_n that is uniform over $k_n \pm 1$.

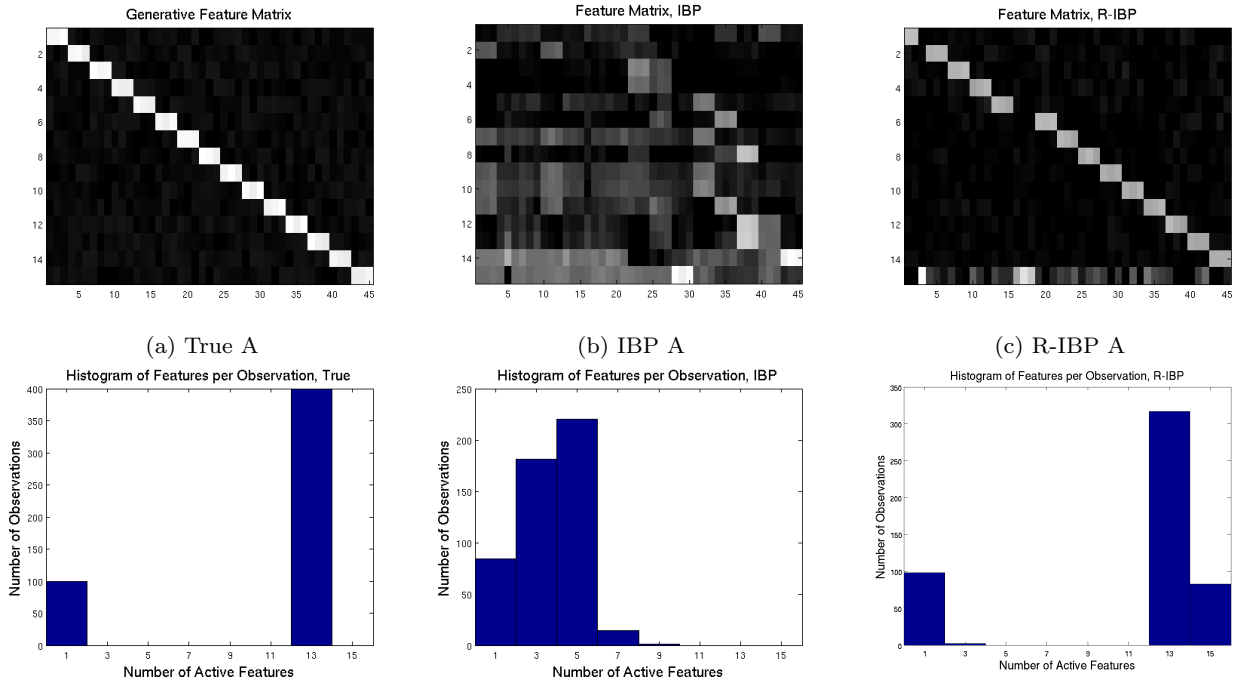
Figure 5 shows qualitative results on the toy data. The first column shows the true features and the true distribution on the number of active features in each observation. Because many of the features occur in many of the datasets, the IBP (center column) does not recover the true features, nor does it recover a distribution of active features that is close to the true distribution. In contrast, the R-IBP (right column) recovers a latent structure that is much closer to true parameters.

7.1.2 Knowledge about the Feature Distribution Assists with Predictive Performance

While interpretable features are desirable, they should not come at the expense of predictive performance. To evaluate predictive performance, we considered 500 observations from a one-inflated Poisson model where 80% of the observations have one associated latent feature and the remaining 20% have a Poisson-distributed number of associated latent features with mean λ. Such a model might be relevant when modeling patients in a typical clinical practice, where most patients might have simple complaints and a few patients may have many complications. We apply the Gibbs sampler for $\lambda = \{3, 6, 9, 12\}$; the concentration parameter for the IBP was set to the mean number of features per observation in each setting.

We explore two variants of the R-IBP: in the fully exchangeable version, we know that observations come from a mixture distribution but we do not know whether the observation is associated with the spike or the slab; all observations have the same $f_n = 0.8\delta_1 + 0.2\text{Poisson}(\lambda)$. In the partially exchangeable version, we know to which mixture component the observation belongs. If the observation belongs to the spike, we have $f_n = \delta_1$, otherwise we have $f_n = \text{Poisson}(\lambda)$ This assumption may be reasonable in many domains; for example, it may be easy to tell if a patient has a simple or complex condition without knowing explicitly what diseases a patient with complex diseases has.

We randomly held out 1% of the data. Figure 6 shows the negative log-likelihoods on the held-out data averaged over 5 runs of 500 iterations each (lower is better). When the mean number of latent features in the slab distribution $\lambda = 3$, all observations have few features, and the R-IBP variants performs slightly worse than the IBP – something we attribute to slower mixing and therefore slower convergence, due to the lack of conjugacy. However, as the slab mean λ in-



(a) True A (b) IBP A (c) R-IBP A

(d) True Histogram over active features per observation (e) IBP Histogram over active features per observation (f) R-IBP Histogram over active features per observation

Fig. 5: Examples of features and counts of active features found by the variational inference for the R-IBP and the IBP on the toy data. The R-IBP recovers patterns much closer to the true features than the IBP, in which observations with just one feature tend to get assigned no features, while observations with many get a few generic features corresponding to most dimensions being active. In contrast, the R-IBP recovers a histogram of features per observation that is much closer to the true distribution.

creases, the R-IBP variants consistently out-perform the IBP. As expected, the partially-exchangeable variant, in which each observation contains a covariate describing whether it is a member of the spike or the slab, does the best.

7.2 Comparison on Multiple Real Datasets

We compare the two inference approaches for the R-IBP from sections 6.1 and 6.2 to three IBP baselines. The hybrid variational IBP applies the same hybrid variational approach to inference in the IBP as was developed for the R-IBP in section 6.2. We also compare to Gibbs sampling in the IBP [Griffiths and Ghahramani, 2011] and the standard variational inference approach for the IBP [Doshi et al., 2009]. Both the Gibbs sampler and the variational methods were run for 300 iterations. For the hybrid variational methods, the weights were resampled every 25 iterations of the coordinate ascent. All methods were run 5 times. A random 1% of the data was held-out for evaluation.

We compare these methods on several datasets, using a linear Gaussian likelihood (see Section 6.2) in all cases:

- The chord dataset consists of a collection of all combinations of single, double, and pairs of the 12 single notes for the octave containing middle C (1442 total observations). Notes and chords were synthesized into wav files using MIDIUtil and FluidSynth; the power spectrum of these wav files was evaluated at every 5Hz between 250 and 750Hz, resulting in a dataset with 100 dimensions. For the R-IBP, the prior on number of features underlying each observation was set to a uniform distribution over $\{1, 2, 3\}$. (Note that this prior was still mis-specified in that there were many more observations with 2 and 3 notes than 1 note).
- The newsgroup dataset is the subset of the 20 newsgroups dataset from http://www.cs.nyu.edu/~roweis/data.html, consisting of the counts for the top 100 words for 5000 documents. We used a partially-exchangeable model with the number of features proportional to length. We arbitrarily set $k_n = \frac{L_n}{150}$, where L_n was

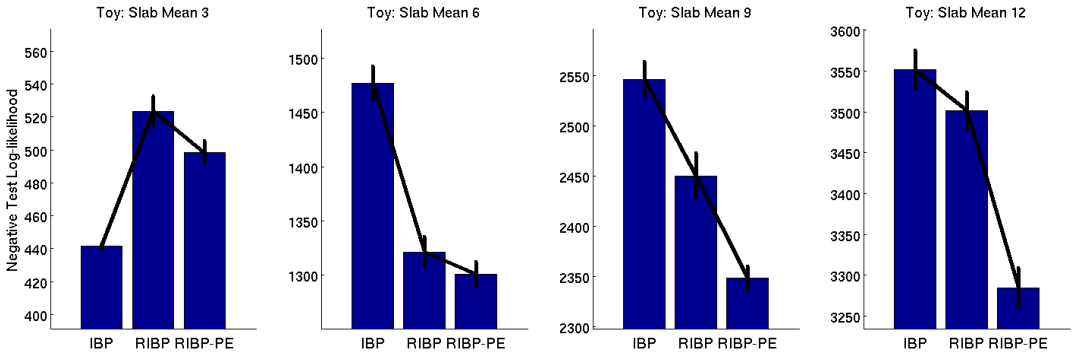


Fig. 6: Negative log-likelihoods (lower is better) for data from a one-inflated Poisson model with the mean of the Poisson $\lambda = \{3, 6, 9, 12\}$. R-IBP is the fully exchangeable R-IBP model, R-IBP-PE is the partially-exchangeable R-IBP model where each observation has a covariate describing the distribution it comes from.

the length of the document, and used a uniform distribution over $\{k_n - 1, k_n, k_n + 1\}$.

– The NPR dataset consisted of the 365 daily features and the 365 daily summaries from www.npr.org from April 2013 to April 2014.[7] The stories were processed through NLTK clean and we kept the 1964 most common words. We let the number of topics in a feature story be uniformly distributed over $\{1, 2, 3\}$ and the number of topics in a summary be uniformly distributed over $\{4, 5, 6\}$, since a feature covers a single news story and a summary covers multiple news stories.

Likelihoods and training times for the toy problem and other problems are shown in tables 1, 2 and 3. Here we see that the auxiliary information provided by the R-IBP also translates into better likelihoods—the R-IBP Gibbs sampler performs the best among all datasets—and not just qualitatively better parameter recovery. In particular, even though our distributions were either mis-specified or somewhat arbitrary guesses, just providing some information about the likely support of the distribution on the number of features, such as the number of possible notes in the chords or the fact that some articles are likely to have more topics than others, improves predictive performance even if the values for those numbers are not carefully specified (e.g. uniform). We see the least benefit for cases where there isn't a clear source of auxiliary information, such at the newsgroups dataset. As expected, the variational inference also runs significantly faster than the MCMC-based approaches; however in some of the experiments the variational approach yielded a lower quality estimate (seen in the NPR dataset).

In addition to the above experiments, we also tried modeling the newsgroup data using an R-IBP with a negative binomial restricting function. This was motivated by the fact that word occurrence in natural languages often follows a heavy-tailed distribution. Further, in [Williamson et al., 2013], using a negative binomial-restricted R-IBP was found to outperform an IBP in directly modeling word occurrence in this dataset. However, when using a negative binomial distribution over the latent features (rather than directly on the word counts), we obtained similar performance to the IBP, and worse than the R-IBP with length-specific uniform distributions. This may be because the IBP does a sufficiently good job in this setting: A negative binomial distribution is not significantly overdispersed, and can be approximated using the Poisson-binomial conditional distribution obtained in the predictive distribution of the IBP. It may also be that heavy-tailed distributions are not actually that useful in a latent feature context. While such distributions are popular in directly modeling word counts or degree distributions [Teh, 2006, Caron, 2012], there is little work in the literature to support the benefit of a heavy tailed *latent* distribution in these cases.

[7] Source: http://www.npr.org/api/queryGenerator.php

8 Discussion and Future Work

The Restricted Indian Buffet Process is a useful tool for latent feature modeling with a non-Poissonian number of latent features per data point. In this article, we have expanded on the original exposition [Williamson et al., 2013] by providing new representations that connect the R-IBP to tilted CRMs and the scaled beta-prime process. We also provide several alternatives for exact and approximate simulation from the R-IBP, as well as new inference algorithms, including a computationally efficient variational/MCMC hybrid algorithm.

Table 1: Comparison of training set likelihoods for the R-IBP and the IBP.

	Chord	Newsgroups	NPR
Hybrid-Var. R-IBP	-1.37e+05 (-1.37e+05, -1.37e+05)	-2.33e+06 (-2.33e+06, -2.33e+06)	-1.65e+07 (-1.66e+07, -1.65e+07)
Hybrid-Var. IBP	-3.02e+05 (-3.02e+05, -3.02e+05)	-2.38e+06 (-2.38e+06, -2.38e+06)	-6.73e+06 (-6.78e+06, -6.68e+06)
Variational IBP	-3.02e+05 (-3.02e+05, -3.02e+05)	-2.39e+06 (-2.39e+06, -2.39e+06)	-6.48e+06 (-6.49e+06, -6.48e+06)
Gibbs R-IBP	-1.39e+05 (-1.39e+05, -1.39e+05)	-2.34e+06 (-2.34e+06, -2.34e+06)	-5.06e+06 (-5.07e+06, -5.06e+06)
Gibbs IBP	-1.48e+05 (-1.48e+05, -1.48e+05)	-2.34e+06 (-2.34e+06, -2.34e+06)	-5.21e+06 (-5.23e+06, -5.19e+06)

Table 2: Comparison of test set likelihoods for the R-IBP and the IBP.

	Chord	Newsgroups	NPR
Hybrid-Var. R-IBP	-3.03e+03 (-3.06e+03, -3.00e+03)	-2.38e+04 (-2.38e+04, -2.37e+04)	-1.64e+05 (-1.66e+05, -1.62e+05)
Hybrid-Var. IBP	-3.28e+03 (-3.30e+03, -3.26e+03)	-2.43e+04 (-2.43e+04, -2.42e+04)	-7.01e+04 (-7.08e+04, -6.94e+04)
Variational IBP	-3.28e+03 (-3.30e+03, -3.26e+03)	-2.43e+04 (-2.43e+04, -2.42e+04)	-7.29e+04 (-7.41e+04, -7.17e+04)
Gibbs R-IBP	-3.20e+03 (-3.23e+03, -3.17e+03)	-2.37e+04 (-2.37e+04, -2.37e+04)	-5.59e+04 (-5.63e+04, -5.56e+04)
Gibbs IBP	-4.30e+03 (-4.34e+03, -4.25e+03)	-2.37e+04 (-2.38e+04, -2.37e+04)	-5.81e+04 (-5.87e+04, -5.75e+04)

Table 3: Comparison of running times (in seconds) for the R-IBP and the IBP.

	Chord	Newsgroups	NPR
Hybrid-Var. R-IBP	1.48e+03 (1.42e+03, 1.53e+03)	1.56e+05 (1.55e+05, 1.58e+05)	1.13e+04 (1.11e+04, 1.14e+04)
Hybrid-Var. IBP	1.02+03 (9.82e+02, 1.05e+03)	3.21e+04 (3.19e+04, 3.23e+04)	1.58e+04 (1.56e+04, 1.60e+04)
Variational IBP	1.11e+03 (1.08e+03, 1.15e+03)	3.69e+04 (3.67e+04, 3.72e+04)	1.68e+04 (1.65e+04, 1.71e+04)
Gibbs R-IBP	3.12e+03 (3.01e+03, 3.23e+03)	2.04e+04 (1.99e+04, 2.08e+04)	5.61e+03 (5.50e+03, 5.73e+03)
Gibbs IBP	1.96e+03 (1.89e+03, 2.03e+03)	1.33e+04 (1.31e+04, 1.35e+04)	5.13e+03 (4.98e+03, 5.28e+03)

While the IBP often has reasonable performance on datasets with arbitrary distributions over the number of features—rather than a Poisson distribution—we find that additional knowledge about the number of features can be very helpful if it is available. In particular, a common challenge when performing inference with the IBP is that it often learns combinations of features as a single feature, especially when there are correlations between features. While these feature combinations may reasonably represent the data, a latent variable model that learns such grouped features will do poorly if asked to make predictions on observations where that correlation is not present. With the R-IBP, it is possible to specify the expected number of features in an observation, allowing us to discover features with both better interpretability and generalization.

In general, we see the most pronounced differences in situations where we had strong prior knowledge about the number of features in a dataset—such as the chord and toy examples. Differences were less pronounced in datasets such as newsgroups, where we made somewhat arbitrary decisions about the potential number of features based on document lengths; in general the IBP is a sufficiently flexible prior to capture posteriors with non-Poisson distributions on the number of latent features. An interesting direction for further research would be try to leverage less strong prior information—such as the information in the NPR dataset where some stories are features and some stories are collections of multiple news summaries.

Empirically, we found that the R-IBP latent feature model was most successful when used with an under-, rather than over-dispersed distributions over the num-

ber of features. This is also the niche that cannot be addressed with alternative approaches such as that of [Caron, 2012]. While the focus of this paper is on *latent* feature models, an example of the R-IBP successfully capturing heavy-tailed behavior when directly modeling work counts is given in [Williamson et al., 2013].

While we have focused on the Indian Buffet Process, the concepts described in this paper are applicable to other nonparametric models such as the beta-negative Binomial process or gamma-Poisson process. As we discussed in Section 4, the variety of possible restrictions is much broader when considering non-binary counting measures. It will be interesting to explore where restricted models can be effectively used in this context; in principle different restrictions can allow domain experts to encode a rich number of kinds of prior knowledge.

Finally, there is much to be explored on approaches for incorporating the kinds of observation-specific restrictions described in this work. The R-IBP has a natural interpretation as an IBP with arbitrary distributions on the number of features in each observation. However, as we discussed in Section 3.4, there is an extra degree of freedom when we specify the R-IBP with a beta process or a beta-prime process. Intuitively, this invariance arises because conditioned on the number of latent features in an observation, the scale of the weights no longer matters. Any restriction that can be viewed as conditioning will result in this property. In theory, working with a normalized beta-prime process would remove this invariance; in practice, working with a normalized beta-prime process is intractable.

However, there do exist other tractable normalized random measures [James et al., 2009] such as the Dirichlet process and other and nonparametric probability measures such as the Pitman-Yor process [Pitman and Yor, 1997]. These measures could be substituted for the beta-prime process in Equation 7. The resulting model could no longer be interpreted as a restricted version of the IBP, but it is nonetheless a valid model that may have very similar properties. Having a more potentially more tractable directing measure may assist in developing robust and scalable inference techniques for restricted models.

A Impact of truncation level on approximation quality when simulating from the R-IBP

In Section 5, we described two approximate methods for sampling from the R-IBP, that made use of a finite-dimensional approximation to the beta process-distributed measure μ. As the dimensionality I of the approximation tends to infinity, these methods will give exact samples from the R-IBP; however a fixed finite I will introduce errors. In this appendix, we discuss the errors introduced in both cases, and provide an error bound for the inclusion probability sampler.

A.1 Impact of truncation level in an inclusion probability sampler

If a size-ordered stick-breaking representation is used to approximate the weights π, then we can directly bound the errors on the inclusion probabilities as functions of the truncation level I, the size of the smallest instantiated weight π_I, and the function f. To do so, we first expand the expression for the probabilities S_J^∞, starting with Equation 15:

$$\begin{aligned}
S_J^\infty &= \sum_{s \in A_J(I)} \prod_{k \in s} \pi_k \prod_{j \ni s} (1 - \pi_j) \\
&\quad + \sum_{s \ni A_J(I)} \prod_{k \in s} \pi_k \prod_{j \ni s} (1 - \pi_j) \\
&= \exp(-\pi_I \alpha) \sum_{s \in A_J(I)} \prod_{k \in s} \pi_k \prod_{j \ni s, j \leq I} (1 - \pi_j) \\
&\quad + \sum_{s \ni A_J(I)} \prod_{k \in s} \pi_k \prod_{j \ni s} (1 - \pi_j) \\
&= \exp(-\pi_I \alpha) S_J^I + \sum_{s \ni A_J(I)} \prod_{k \in s} \pi_k \prod_{j \ni s} (1 - \pi_j)
\end{aligned}$$

where $A_J(I)$ are the sets of feature allocations in which all J instantiated features are associated with one of the I largest atoms in μ. The second line follows because the probability of that none of the features associated with the remaining atoms are selected is $\exp(-\pi_I \alpha)$.

Since the probability that at least one feature outside the most significant I features appears is $1 - \exp(-\pi_I \alpha)$, the second term is bounded between 0 and $1 - \exp(-\pi_I \alpha)$. Thus we can bound the inclusion probabilities

$$\begin{aligned}
\eta_{k;J} &= \pi_k \frac{S_{J-1}^\infty(\pi_1, ..., \pi_{k-1}, \pi_{k+1}, ..., \pi_I)}{S_J^\infty(\pi_1, ..., \pi_I)} \\
&\geq \pi_k \frac{e^{-\pi_I \alpha} S_{J-1}^{I-1}(\pi_1, ..., \pi_{k-1}, \pi_{k+1}, ..., \pi_I)}{e^{-\pi_I \alpha} S_J^I(\pi_1, ..., \pi_I) + (1 - e^{-\pi_I \alpha})} \\
&\leq \pi_k \frac{e^{-\pi_I \alpha} S_{J-1}^{I-1}(\pi_1, ..., \pi_{k-1}, \pi_{k+1}, ..., \pi_I) + (1 - e^{-\pi_I \alpha})}{e^{-\pi_I \alpha} S_J^I(\pi_1, ..., \pi_I)}
\end{aligned}$$

As expected, the quality of the approximation depends not only truncation I (and associated π_I) but also on the values S_J^I. If the probability of sampling J elements from the first I is low, then the approximation will be poor because it is likely that additional features would have been required to sample J elements. These bounds can be used in situations where one can use approximate, rather than exact, probabilities.

A.2 Impact of truncation level in a rejection sampler

As $I \to \infty$, both the weak-limit approximation of Equation 12 and the stick-breaking construction of Equation 13 will give exact samples from the R-IBP. However, a finite I will introduce errors. When a stick-breaking representation for μ is used, then we know that all weights π_j, $j > I$ will be less than π_I. In particular, the iterative nature of the stick-breaking construction means that, if we exclude the first I atoms $\pi_1, \ldots, \pi_I$, and scale the remaining atoms by π_I, we are left with a (strictly ordered) sample from the beta process.

We can consider the error introduced by this construction by considering the values of z_{nj} that are excluded due to the truncation. If there are any non-zero elements z_{nj} for $j > I$, our rejection probability will not be correct. Since the weights $\pi_j, j > I$ are described by a scaled beta process, we know that the number of excluded non-zero elements will be distributed as Poisson($\alpha\pi_I$). So, with probability $1 - \text{Poisson}(0; \alpha\pi_I) = 1 - \exp(-\pi_I\alpha)$ the true sum $\sum_i^\infty z_{ni} \neq \sum_i^I z_{ni}$ and thus we may incorrectly reject or accept a proposal. Conditioned on a desired number of features J, we can further break down the probability of incorrectly rejecting a proposal with $\sum_{i=1}^I (z_i^*) < J$ of incorrectly accepting a proposal with $\sum_{i=1}^I (z_i^*) = J$ by considering the following possible scenarios:

1. $\sum_{i=1}^I z_i^* > J$: We reject the proposal. This is always correct.
2. $\sum_{i=1}^I z_i^* = J$: We accept the proposal. However, if the truncated tail has $\sum_{i=I+1}^\infty z_i^* > 0$, we should really have rejected. Our decision is correct with probability $P(\sum_{i=I+1}^\infty z_i^* = 0) = \exp(-\pi_I\alpha)$.
3. $\sum_{i=1}^I z_i^* < J$: We reject the proposal. However, if $\sum_{i=1}^I z_i^* = J - k$ but the truncated tail has $\sum_{i=I+1}^\infty z_i^* = k$, we will really should have accepted. Our decision is correct with probability $1 - P(\sum_{i=I+1}^\infty z_i^* = J - \sum_{i=1}^I z_i^*) = 1 - \text{Poisson}(J - \sum_{i=1}^I z_i^*; \pi_I\alpha)$.

Acknowledgements The authors would like to thank Ryan P. Adams for numerous helpful discussions and suggestions, and Jeff Miller for suggesting the link to tilted random measures.

References

[Aires, 1999] Aires, N. (1999). Algorithms to find exact inclusion probabilities for conditional Poisson sampling and Pareto πps sampling designs. *Methodology and Computing in Applied Probability*, 1:457–469.

[Aldous, 1983] Aldous, D. (1983). Exchangeability and related topics. In *Ecole d'Ete St Flour*, number 1117 in Springer Lecture Notes in Mathematics, pages 1–198. Springer.

[Brix, 1999] Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Adv. in Appl. Probab.*, 31(4):929–953.

[Broderick et al., 2015] Broderick, T., Mackey, L., Paisley, J., Jordan, M., et al. (2015). Combinatorial clustering and the beta negative binomial process. *Pattern Analysis and Machine Intelligence*, 37(2):290–306.

[Broderick et al., 2014] Broderick, T., Wilson, A., and Jordan, M. (2014). Posteriors, conjugacy, and exponential families for completely random measures. *arXiv:1410.6843*.

[Brostrom and Nilsson, 2000] Brostrom, G. and Nilsson, L. (2000). Acceptance-rejection sampling from the conditional distribution of independent discrete random variables, given their sum. *Statistics: A Journal of Theoretical and Applied Statistics*, 34:247–257.

[Caron, 2012] Caron, F. (2012). Bayesian nonparametric models for bipartite graphs. In *Advances in Neural Information Processing Systems*.

[Chen, 2000] Chen, S. X. (2000). General properties and estimation of conditional Bernoulli models. *Journal of Multivariate Analysis*, 74:69–87.

[Doshi et al., 2009] Doshi, F., Miller, K. T., Van Gael, J., and Teh, Y. W. (2009). Variational inference for the Indian buffet process. In *Artificial Intelligence and Statistics*.

[Doshi-Velez and Ghahramani, 2009] Doshi-Velez, F. and Ghahramani, Z. (2009). Correlated non-parametric latent feature models. In *Uncertainty in Artificial Intelligence*.

[Ferguson and Klass, 1972] Ferguson, T. S. and Klass, M. J. (1972). A representation of independent increment processes without Gaussian components. *Ann. Math. Statist.*, 43(5):1634–1643.

[Fox et al., 2009] Fox, E., Jordan, M., Sudderth, E., and Willsky, A. (2009). Sharing features among dynamical systems with beta processes. In *Advances in Neural Information Processing Systems*.

[Gerber and Shiu, 1993] Gerber, H. U. and Shiu, E. S. (1993). *Option pricing by Esscher transforms*. HEC Ecole des hautes études commerciales.

[Görür et al., 2006] Görür, D., Jäkel, F., and Rasmussen, C. E. (2006). A choice model with infinitely many latent features. In *International Conference of Machine Learning*.

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

[Gupta et al., 2013] Gupta, S., Phung, D., and Venkatesh, S. (2013). Factorial multi-task learning: a Bayesian nonparametric approach. In *International Conference of Machine Learning*, pages 657–665.

[Hanif and Brewer, 1983] Hanif, M. and Brewer, K. R. W. (1983). *Sampling with unequal probabilities*. Springer-Verlag.

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

[James et al., 2009] James, L., Lijoi, A., and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36(1):76–97.

[James, 2005] James, L. F. (2005). Functionals of Dirichlet processes, the Cifarelli-Regazzini identity and beta-gamma processes. *The Annals of Statistics*, 33(2):pp. 647–660.

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

[Knowles and Ghahramani, 2007] Knowles, D. and Ghahramani, Z. (2007). Infinite sparse factor analysis and infinite independent components analysis. In *Independent Component Analysis and Signal Separation*.

[Lau, 2013] Lau, J. W. (2013). A conjugate class of random probability measures based on tilting and with its posterior analysis. *Bernoulli*, 19(5B):2590–2626.

[Miller et al., 2008] Miller, K. T., Griffiths, T., and Jordan, M. I. (2008). The phylogenetic Indian buffet process: A non-exchangeable nonparametric prior for latent features. In *Uncertainty in Artificial Intelligence*.

[Miller et al., 2009] Miller, K. T., Griffiths, T. L., and Jordan, M. I. (2009). Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems*.

[Orbanz, 2009] Orbanz, P. (2009). Construction of nonparametric Bayesian models from parametric Bayes equations. In *Advances in Neural Information Processing Systems*.

[Papaspiliopoulos and Roberts, 2008] Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, 95(1):169–186.

[Pitman and Yor, 1997] Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, 25(2):855–900.

[Rosiski, 2001] Rosiski, J. (2001). Series representations of Lvy processes from the perspective of point processes. In Barndorff-Nielsen, O., Resnick, S., and Mikosch, T., editors, *Lvy Processes*, pages 401–415. Birkhuser Boston.

[Ruiz et al., 2014] Ruiz, F., Valera, I., Blanco, C., and Perez-Cruz, F. (2014). Bayesian nonparametric comorbidity analysis of psychiatric disorders. *Journal of Machine Learning Research*, 15:1215–1247.

[Saeedi and Bouchard-Côté, 2011] Saeedi, A. and Bouchard-Côté, A. (2011). Priors over recurrent continuous time processes. In *Advances in Neural Information Processing Systems*.

[Teh, 2006] Teh, Y. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *International Conference on Computational Linguistics*, pages 985–992.

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

[Teh et al., 2007] Teh, Y. W., Görür, D., and Ghahramani, Z. (2007). Stick-breaking construction for the Indian buffet process. In *Artificial Intelligence and Statistics*.

[Thibaux and Jordan, 2007] Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In *Artificial Intelligence and Statistics*.

[Titsias, 2008] Titsias, M. (2008). The infinite gamma-Poisson feature model. In *Advances in Neural Information Processing Systems*.

[Wainwright and Jordan, 2008] Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1305.

[Williamson et al., 2013] Williamson, S. A., MacEachern, S. N., and Xing, E. P. (2013). Restricting exchangeable nonparametric distributions. In *Advances in Neural Information Processing Systems*.

[Zhou et al., 2009] Zhou, M., Chen, H., Paisley, J., Ren, L., Sapiro, G., and Carin, L. (2009). Non-parametric Bayesian dictionary learning for sparse image representations. In *Advances in Neural Information Processing Systems*.

[Zhou et al., 2012] Zhou, M., Hannah, L., Dunson, D., and Carin, L. (2012). Beta-negative binomial process and poisson factor analysis. In *Artificial Intelligence and Statistics*.