This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2318721, IEEE Transactions on Pattern Analysis and Machine Intelligence

Combinatorial Clustering and the Beta Negative Binomial Process

Tamara Broderick, Lester Mackey, John Paisley, Michael I. Jordan

Abstract

We develop a Bayesian nonparametric approach to a general family of latent class problems in which individuals can belong simultaneously to multiple classes and where each class can be exhibited multiple times by an individual. We introduce a combinatorial stochastic process known as the *negative binomial process* (NBP) as an infinite-dimensional prior appropriate for such problems. We show that the NBP is conjugate to the beta process, and we characterize the posterior distribution under the beta-negative binomial process (BNBP) and hierarchical models based on the BNBP (the HBNBP). We study the asymptotic properties of the BNBP and develop a three-parameter extension of the BNBP that exhibits power-law behavior. We derive MCMC algorithms for posterior inference under the HBNBP, and we present experiments using these algorithms in the domains of image segmentation, object recognition, and document analysis.

Index Terms

beta process, admixture, mixed membership, Bayesian, nonparametric, integer latent feature model

1 Introduction

In traditional clustering problems the goal is to induce a set of latent classes and to assign each data point to one and only one class. This problem has been approached within a model-based framework via the use of finite mixture models, where the mixture components characterize the distributions associated with the classes, and the mixing proportions capture the mutual exclusivity of the classes (Fraley and Raftery, 2002; McLachlan and Basford, 1988). In many domains in which the notion of latent classes is natural, however, it is unrealistic to assign each individual to a single class. For example, in genetics, while it may be reasonable to assume the existence of underlying ancestral populations that define distributions on observed alleles, each individual in an existing population is likely to be a blend of the patterns associated with the ancestral populations. Such a genetic blend is known as an *admixture* (Pritchard et al., 2000). A significant literature on model-based approaches to admixture has arisen in recent years (Blei et al., 2003; Erosheva and Fienberg,

- *T. Broderick and M. I. Jordan are with the Department of Statistics and the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94705. L. Mackey is with the Department of Statistics, Stanford University, Stanford, CA 94305. J. Paisley is with the Department of Electrical Engineering at Columbia University, New York, NY 10027.*

2005; Pritchard et al., 2000), with applications to a wide variety of domains in genetics and beyond, including document modeling and image analysis.[1]

Model-based approaches to admixture are generally built on the foundation of mixture modeling. The basic idea is to treat each individual as a collection of data, with an exchangeability assumption imposed for the data within an individual but not between individuals. For example, in the genetics domain the intra-individual data might be a set of genetic markers, with marker probabilities varying across ancestral populations. In the document domain the intra-individual data might be the set of words in a given document, with each document (the individual) obtained as a blend across a set of underlying "topics" that encode probabilities for the words. In the image domain, the intra-individual data might be visual characteristics like edges, hue, and location extracted from image patches. Each image is then a blend of object classes (e.g., grass, sky, or car), each defining a distinct distribution over visual characteristics. In general, this blending is achieved by making use of the probabilistic structure of a finite mixture but using a different sampling pattern. In particular, mixing proportions are treated as random effects that are drawn once per individual, and the data associated with that individual are obtained by repeated draws from a mixture model having that fixed set of mixing proportions. The overall model is a hierarchical model, in which mixture components are shared among individuals and mixing proportions are treated as random effects.

Although the literature has focused on using finite mixture models in this context, there has also been a growing literature on Bayesian nonparametric approaches to admixture models, notably the *hierarchical Dirichlet process* (HDP) (Teh et al., 2006), where the number of shared mixture components is infinite. Our focus in the current paper is also on nonparametric methods, given the open-ended nature of the inferential objects with which real-world admixture modeling is generally concerned.

Although viewing an admixture as a set of repeated draws from a mixture model is natural in many situations, it is also natural to take a different perspective, akin to latent trait modeling, in which the individual (e.g., a document or a genotype) is characterized by the set of "traits" or "features" that it possesses, and where there is no assumption of mutual exclusivity. Here the focus is on the individual and not on the "data" associated with an individual. Indeed, under the exchangeability assumption alluded to above it is natural to reduce the repeated draws from a mixture model to the counts of the numbers of times that each mixture component is selected, and we may wish to model these counts directly. We may further wish to consider hierarchical models in which there is a linkage among the counts for different individuals.

This idea has been made explicit in a recent line of work based on the *beta process*. Originally developed for survival analysis, where an integrated form of the beta process was used as a model for random hazard functions (Hjort, 1990), more recently it has been observed that the beta process also provides a natural framework for latent feature modeling (Thibaux and Jordan, 2007). In particular, as we discuss in detail in

[1] While we refer to such models generically as "admixture models," we note that they are also often referred to as *topic models* or *mixed membership models*.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2318721, IEEE Transactions on Pattern Analysis and Machine Intelligence

Section 2, a draw from the beta process yields an infinite collection of coin-tossing probabilities. Tossing these coins—a draw from a *Bernoulli process*—one obtains a set of binary features that can be viewed as a description of an admixed individual. A key advantage of this approach is the conjugacy between the beta and Bernoulli processes: this property allows for tractable inference, despite the countable infinitude of coin-tossing probabilities. A limitation of this approach, however, is its restriction to binary features; indeed, one of the virtues of the mixture-model-based approach is that a given mixture component can be selected more than once, with the total number of selections being random.

We develop a model for admixture that meets all of the desiderata outlined thus far. Unlike the Bernoulli process likelihood, our featural model allows each feature to be exhibited any non-negative integer number of times by an individual. Unlike admixture models based on the HDP, our model cohesively includes a random total number of features (e.g., words or traits) per individual (e.g., a document or genotype).

As inspiration, we note that in the setting of classical random variables, beta-Bernoulli conjugacy is not the only form of conjugacy involving the beta distribution—the negative binomial distribution is also conjugate to the beta. Anticipating the value of conjugacy in the setting of nonparametric models, we define and develop a stochastic process analogue of the negative binomial distribution, which we refer to as the *negative binomial process* (NBP),[2] and provide a rigorous proof of its conjugacy to the beta process. We use this process as part of a new model—the *hierarchical beta negative binomial process* (HBNBP)—based on the NBP and the hierarchical beta process (Thibaux and Jordan, 2007). Our theoretical and experimental development focus on the usefulness of the HBNBP in the admixture setting, where flexible modeling of feature totals can lead to improved inferential accuracy (see Figure 3a and the surrounding discussion). However, the utility of the HBNBP is not limited to the admixture setting and should extend readily to the modeling of latent factors and the identification of more general latent features. Moreover, the negative binomial component of our model offers addtional flexibility in the form of a new parameter unavailable in either the Bernoulli or multinomial likelihoods traditionally explored in Bayesian nonparametrics.

The remainder of the paper is organized as follows. In Section 2 we present the framework of completely random measures that provides the formal underpinnings for our work. We discuss the Bernoulli process, introduce the NBP, and demonstrate the conjugacy of both to the beta process in Section 3. Section 4 focuses on the problem of modeling admixture and on general hierarchical modeling based on the negative binomial process. Section 5 and Section 6 are devoted to a study of the asymptotic behavior of the NBP with a beta process prior, which we call the beta-negative binomial process (BNBP). We describe algorithms for posterior inference in Section 7. Finally, we present experimental results. First, we use the BNBP to define a generative model for summaries of terrorist incidents with the goal of identifying the perpetrator of a given terrorist attack in Section 8. Second, we demonstrate the utility of a finite approximation to the BNBP in the domain

[2] Zhou et al. (2012) have independently investigated negative binomial processes in the context of integer matrix factorization. We discuss their concurrent contributions in more detail in Section 4.

of automatic image segmentation in Section 9. Section 10 presents our conclusions.

2 Completely Random Measures

In this section we review the notion of a completely random measure (CRM), a general construction that yields random measures that are closely tied to classical constructions involving sets of independent random variables. We present CRM-based constructions of several of the stochastic processes used in Bayesian nonparametrics, including the beta process, gamma process, and Dirichlet process. In the following section we build on the foundations presented here to consider additional stochastic processes.

Consider a probability space $(\Psi, \mathcal{F}, \mathbb{P})$. A *random measure* is a random element μ such that $\mu(A)$ is a nonnegative random variable for any A in the sigma algebra $\mathcal{F}$. A *completely random measure* (CRM) μ is a random measure such that, for any disjoint, measurable sets $A, A' \in \mathcal{F}$, we have that $\mu(A)$ and $\mu(A')$ are independent random variables (Kingman, 1967). Completely random measures can be shown to be composed of at most three components:

1) A *deterministic measure*. For deterministic μ_{det}, it is trivially the case that $\mu_{det}(A)$ and $\mu_{det}(A')$ are independent for disjoint A, A'.
2) A *set of fixed atoms*. Let $(u_1, \ldots, u_L) \in \Psi^L$ be a collection of deterministic locations, and let $(\eta_1, \ldots, \eta_L) \in \mathbb{R}_+^L$ be a collection of independent random weights for the atoms. The collection may be countably infinite, in which case we say $L = \infty$. Then let $\mu_{fix} = \sum_{l=1}^{L} \eta_l \delta_{u_l}$. The independence of the η_l implies the complete randomness of the measure.
3) An *ordinary component*. Let ν_{PP} be a Poisson process intensity on the space $\Psi \times \mathbb{R}_+$. Let $\{(v_1, \xi_1), (v_2, \xi_2), \ldots\}$ be a draw from the Poisson process with intensity ν_{PP}. Then the ordinary component is the measure $\mu_{ord} = \sum_{j=1}^{\infty} \xi_j \delta_{v_j}$. Here, the complete randomness follows from properties of the Poisson process.

One observation from this componentwise breakdown of CRMs is that we can obtain a countably infinite collection of random variables, the ξ_j, from the Poisson process component if ν_{PP} has infinite total mass (but is still sigma-finite). Consider again the criterion that a CRM μ yield independent random variables when applied to disjoint sets. In light of the observation about the collection $\{\xi_j\}$, this criterion may now be seen as an extension of an independence assumption in the case of a finite set of random variables. We cover specific examples next.

2.1 Beta process

The *beta process* (Hjort, 1990; Kim, 1999; Thibaux and Jordan, 2007) is an example of a CRM. It has the following parameters: a *mass parameter* $\gamma > 0$, a *concentration parameter* $\theta > 0$, a purely atomic measure $H_{fix} = \sum_l \rho_l \delta_{u_l}$ with $\gamma \rho_l \in (0, 1)$ for all l a.s., and a purely continuous probability measure H_{ord} on Ψ. Note that we have explicitly separated out the mass parameter γ so that, e.g., H_{ord} is a probability measure; in Thibaux and Jordan (2007), these two parameters are expressed as a single measure with total mass equal to γ. Typically,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2318721, IEEE Transactions on Pattern Analysis and Machine Intelligence

though, the normalized measure H_{ord} is used separately from the mass parameter γ (as we will see below), so the notational separation is convenient. Often the final two measure parameters are abbreviated as their sum: $H = H_{fix} + H_{ord}$.

Given these parameters, the beta process has the following description as a CRM:

1) The deterministic measure is uniformly zero.
2) The fixed atoms have locations $(u_1, \ldots, u_L) \in \Psi^L$, where L is potentially infinite though typically finite. Atom weight η_l has distribution

$$\eta_l \overset{\text{ind}}{\sim} \text{Beta}\left(\theta\gamma\rho_l, \theta(1-\gamma\rho_l)\right), \tag{1}$$

where the ρ_l parameters are the weights in the purely atomic measure H_{fix}.
3) The ordinary component has Poisson process intensity $H_{ord} \times \nu$, where ν is the measure

$$\nu(db) = \gamma\theta b^{-1}(1-b)^{\theta-1}\, db, \tag{2}$$

which is sigma-finite with finite mean. It follows that the number of atoms in this component will be countably infinite with finite sum.

As in the original specification of Hjort (1990) and Kim (1999), Eq. (2) can be generalized by allowing θ to depend on the Ψ coordinate. The homogeneous intensity in Eq. (2) seems to be used predominantly in practice (Thibaux and Jordan, 2007; Fox et al., 2009) though, and we focus on it here for ease of exposition. Nonetheless, we note that our results below extend easily to the non-homogeneous case.

The CRM is the sum of its components. Therefore, we may write a draw from the beta process as

$$B = \sum_{k=1}^{\infty} b_k \delta_{\psi_k} \triangleq \sum_{l=1}^{L} \eta_l \delta_{u_l} + \sum_{j=1}^{\infty} \xi_j \delta_{v_j}, \tag{3}$$

with atom locations equal to the union of the fixed atom and ordinary component atom locations $\{\psi_k\}_k = \{u_l\}_{l=1}^{L} \cup \{v_j\}_{j=1}^{\infty}$. Notably, B is a.s. discrete. We denote a draw from the beta process as $B \sim \text{BP}(\theta, \gamma, H)$. The provenance of the name "beta process" is now clear; each atom weight in the fixed atomic component is beta-distributed, and the Poisson process intensity generating the ordinary component is that of an improper beta distribution.

From the above description, the beta process provides a prior on a potentially infinite vector of weights, each in $(0, 1)$ and each associated with a corresponding parameter $\psi \in \Psi$. The potential countable infinity comes from the Poisson process component. The weights in $(0, 1)$ may be interpreted as probabilities, though not as a distribution across the indices as we note that they need not sum to one. We will see in Section 4 that the beta process is appropriate for feature modeling (Thibaux and Jordan, 2007; Griffiths and Ghahramani, 2006). In this context, each atom, indexed by k, of B corresponds to a feature. The atom weights $\{b_k\}$, which are each in $[0, 1]$ a.s., can be viewed as representing the frequency with which each feature occurs in the dataset. The atom locations $\{\psi_k\}$ represent parameters associated with the features that can be used in forming a likelihood.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2318721, IEEE Transactions on Pattern Analysis and Machine Intelligence

In Section 5, we will show that an extension to the beta process called the *three-parameter beta process* has certain desirable properties beyond the classic beta process, in particular its ability to generate power-law behavior (Teh and Görür, 2009; Broderick et al., 2012), which roughly says that the number of features grows as a power of the number of data points. In the three-parameter case, we introduce a *discount parameter* $\alpha \in (0,1)$ with $\theta > -\alpha$ and $\gamma > 0$ such that:

1) There is again no deterministic component.
2) The fixed atoms have locations $(u_1, \ldots, u_L) \in \Psi^L$, with L potentially infinite but typically finite. Atom weight η_l has distribution $\eta_l \overset{\text{ind}}{\sim} \text{Beta}\left(\theta\gamma\rho_l - \alpha, \theta(1-\gamma\rho_l) + \alpha\right)$, where the ρ_l parameters are the weights in the purely atomic measure H_{fix} and we now have the constraints $\theta\gamma\rho_l - \alpha, \theta(1-\gamma\rho_l) + \alpha \geq 0$.
3) The ordinary component has Poisson process intensity $H_{ord} \times \nu$, where ν is the measure:

$$\nu(db) = \gamma \frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} b^{-1-\alpha}(1-b)^{\theta+\alpha-1}\, db.$$

Again, we focus on the homogeneous intensity ν as in the beta process case though it is straightforward to allow θ to depend on coordinates in Ψ.

In this case, we again have the full process draw B as in Eq. (3), and we say $B \sim 3\text{BP}(\alpha, \theta, \gamma, H)$.

2.2 Full beta process

The specification that the atom parameters in the beta process be of the form $\theta\gamma\rho_l$ and $\theta(1-\gamma\rho_l)$ can be unnecessarily constraining; $\theta\gamma\rho_l - \alpha$ and $\theta(1-\gamma\rho_l)+\alpha$ are even more unwieldy in the power-law case. Indeed, the classical beta distribution has two free parameters. Yet, in the beta process as described above, θ and γ are determined as part of the Poisson process intensity, so there is essentially one free parameter for each of the beta-distributed weights associated with the atoms (Eq. (1)). A related problematic issue is that the beta process forces the two parameters in the beta distribution associated with each atom to sum to θ, which is constant across all of the atoms.

One way to remove these restrictions is to allow $\theta = \theta(\psi)$, a function of the position $\psi \in \Psi$ as mentioned above. However, we demonstrate in Appendix A that there are reasons to prefer a fixed concentration parameter θ for the ordinary component; there is a fundamental relation between this parameter and similar parameters in other common CRMs (e.g., the Dirichlet process, which we describe in Section 2.4). Moreover, the concern here is entirely centered on the behavior of the fixed atoms of the process, and letting θ depend on ψ retains the unusual—from a classical parametric perspective—form of the beta distribution in Eq. (1). As an alternative, we provide a generalization of the beta process that more closely aligns with the classical perspective in which we allow two general beta parameters for each atom. As we will see, this generalization is natural, and indeed necessary, in considering conjugacy.

We thus define the *full beta process* (FBP) as having the following parameterization: a *mass parameter* $\gamma > 0$, a *concentration parameter* $\theta > 0$, a number of fixed atoms $L \in \{0, 1, 2, \ldots\} \cup \{\infty\}$ with locations $(u_1, \ldots, u_L) \in \Psi^L$, two sets of strictly positive atom weight parameters $\{\rho_l\}_{l=1}^L$ and $\{\sigma_l\}_{l=1}^L$, and a purely continuous measure H_{ord} on Ψ. In this case, the atom weight parameters satisfy the simple condition $\rho_l, \sigma_l > 0$ for all $l \in \{1, \ldots, L\}$. This specification is the same as the beta process specification introduced above with the sole exception of a more general parameterization for the fixed atoms. We obtain the following CRM:

1) There is no deterministic measure.
2) There are L fixed atoms with locations $(u_1, \ldots, u_L) \in \Psi^L$ and corresponding weights $\eta_l \overset{\text{ind}}{\sim} \text{Beta}\,(\rho_l, \sigma_l)$.
3) The ordinary component has Poisson process intensity $H_{ord} \times \nu$, where ν is the measure $\nu(db) = \gamma\theta b^{-1}(1-b)^{\theta-1}\,db$.

As discussed above, we favor the homogeneous intensity ν in exposition but note the straightforward extension to allow θ to depend on Ψ location.

We denote this CRM by $B \sim \text{FBP}(\theta, \gamma, \mathbf{u}, \boldsymbol{\rho}, \boldsymbol{\sigma}, H_{ord})$.

2.3 Gamma process

While the beta process provides a countably infinite vector of frequencies in $(0, 1]$ with associated parameters ψ_k, it is sometimes useful to have a countably infinite vector of positive, real-valued quantities that can be used as rates rather than frequencies for features. We can obtain such a prior with the *gamma process* (Ferguson, 1973), a CRM with the following parameters: a *concentration parameter* $\theta > 0$, a *scale parameter* $c > 0$, a purely atomic measure $H_{fix} = \sum_l \rho_l \delta_{u_l}$ with $\forall l, \rho_l > 0$, and a purely continuous measure H_{ord} with support on Ψ. Its description as a CRM is as follows (Thibaux, 2008):

1) There is no deterministic measure.
2) The fixed atoms have locations $(u_1, \ldots, u_L) \in \Psi^L$, where L is potentially infinite but typically finite. Atom weight η_l has distribution $\eta_l \overset{\text{ind}}{\sim} \text{Gamma}(\theta\rho_l, c)$, where we use the shape-inverse-scale parameterization of the gamma distribution and where the ρ_l parameters are the weights in the purely atomic measure H_{fix}.
3) The ordinary component has Poisson process intensity $H_{ord} \times \nu$, where ν is the measure:

$$\nu(d\tilde{g}) = \theta\tilde{g}^{-1}\exp\left(-c\tilde{g}\right)\,d\tilde{g}. \tag{4}$$

As in the case of the beta process, the gamma process can be expressed as the sum of its components: $\tilde{G} = \sum_k \tilde{g}_k \delta_{\psi_k} \triangleq \sum_{l=1}^L \eta_l \delta_{u_l} + \sum_j \xi_j \delta_{v_j}$. We denote this CRM as $\tilde{G} \sim \Gamma\text{P}(\theta, c, H)$, for $H = H_{fix} + H_{ord}$.

2.4 Dirichlet process

While the beta process has been used as a prior in featural models, the Dirichlet process is the classic Bayesian nonparametric prior for clustering models (Ferguson, 1973; MacEachern and Müller, 1998; McCloskey, 1965; Neal, 2000; West, 1992). The Dirichlet process itself is not a CRM; its atom weights, which represent cluster frequencies, must sum to one and are therefore correlated. But it can be obtained by normalizing the gamma process (Ferguson, 1973).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2318721, IEEE Transactions on Pattern Analysis and Machine Intelligence

In particular, using facts about the Poisson process (Kingman, 1993), one can check that, when there are finitely many fixed atoms, we have $\tilde{G}(\Psi) < \infty$ a.s.; that is, the total mass of the gamma process is almost surely finite despite having infinitely many atoms from the ordinary component. Therefore, normalizing the process by dividing its weights by its total mass is well-defined. We thus can define a *Dirichlet process* as

$$G = \sum_k g_k \delta_{\psi_k} \triangleq \tilde{G}/\tilde{G}(\Psi),$$

where $\tilde{G} \sim \Gamma\mathrm{P}(\theta, 1, H)$, and where there are two parameters: a *concentration parameter* θ and a *base measure* H with finitely many fixed atoms. Note that while we have chosen the scale parameter $c = 1$ in this construction, the choice is in fact arbitrary for $c > 0$ and does not affect the G distribution (Eq. (4.15) and p. 83 of Pitman (2006)).

From this construction, we see immediately that the Dirichlet process is almost surely atomic, a property inherited from the gamma process. Moreover, not only are the weights of the Dirichlet process all contained in $(0, 1)$ but they further sum to one. Thus, the Dirichlet process may be seen as providing a probability distribution on a countable set. In particular, this countable set is often viewed as a countable number of clusters, with cluster parameters ψ_k.

3 Conjugacy and combinatorial clustering

In Section 2, we introduced CRMs and showed how a number of classical Bayesian nonparametric priors can be derived from CRMs. These priors provide infinite-dimensional vectors of real values, which can be interpreted as feature frequencies, feature rates, or cluster frequencies. To flesh out such interpretations we need to couple these real-valued processes with discrete-valued processes that capture combinatorial structure. In particular, viewing the weights of the beta process as feature frequencies, it is natural to consider binomial and negative binomial models that transform these frequencies into binary values or nonnegative integer counts. In this section we describe stochastic processes that achieve such transformations, again relying on the CRM framework.

The use of a Bernoulli likelihood whose frequency parameter is obtained from the weights of the beta process has been explored in the context of survival models by Hjort (1990) and Kim (1999) and in the context of feature modeling by Thibaux and Jordan (2007). After reviewing the latter construction, we discuss a similar construction based on the negative binomial process. Moreover, recalling that Thibaux and Jordan (2007), building on work of Hjort (1990) and Kim (1999), have shown that the Bernoulli likelihood is conjugate to the beta process, we demonstrate an analogous conjugacy result for the negative binomial process.

3.1 Bernoulli process

One way to make use of the beta process is to couple it to a *Bernoulli process* (Thibaux and Jordan, 2007). The Bernoulli process, denoted $\mathrm{BeP}(\tilde{H})$, has a single parameter, a *base measure* $\tilde{H}$; $\tilde{H}$ is any discrete measure with atom weights in $(0, 1]$. Although our focus will be on models in which $\tilde{H}$ is a draw from a beta process, as a matter of the general definition of the Bernoulli process the base measure $\tilde{H}$ need not be a CRM or even random—just as the Poisson distribution is defined relative to a parameter that may or may not be random in general but which is sometimes given a gamma distribution prior. Since $\tilde{H}$ is discrete by assumption, we may write

$$\tilde{H} = \sum_{k=1}^{\infty} b_k \delta_{\psi_k} \tag{5}$$

with $b_k \in (0, 1]$. We say that the random measure I is drawn from a Bernoulli process, $I \sim \mathrm{BeP}(\tilde{H})$, if $I = \sum_{k=1}^{\infty} i_k \delta_{\psi_k}$ with $i_k \overset{\text{ind}}{\sim} \mathrm{Bern}(b_k)$ for $k = 1, 2, \ldots$. That is, to form the Bernoulli process, we simply make a Bernoulli random variable draw for every one of the (potentially countable) atoms of the base measure. This definition of the Bernoulli process was proposed by Thibaux and Jordan (2007); it differs from a precursor introduced by Hjort (1990) in the context of survival analysis.

One interpretation for this construction is that the atoms of the base measure $\tilde{H}$ represent potential features of an individual, with feature frequencies equal to the atom weights and feature characteristics defined by the atom locations. The Bernoulli process draw can be viewed as characterizing the individual by the set of features that have weights equal to one. Suppose $\tilde{H}$ is derived from a Poisson process as the ordinary component of a completely random measure and has finite mass; then the number of features exhibited by the Bernoulli process, i.e. the total mass of the Bernoulli process draw, is a.s. finite. Thus the Bernoulli process can be viewed as providing a Bayesian nonparametric model of sparse binary feature vectors.

Now suppose that the base measure parameter is a draw from a beta process with parameters $\theta > 0$, $\gamma > 0$, and base measure H. That is, $B \sim \mathrm{BP}(\theta, \gamma, H)$ and $I \sim \mathrm{BeP}(B)$. We refer to the overall process as the *beta-Bernoulli process* (BBeP). Suppose that the beta process B has a finite number of fixed atoms. Then we note that the finite mass of the ordinary component of B implies that I has support on a finite set. That is, even though B has a countable infinity of atoms, I has only a finite number of atoms. This observation is important since, in any practical model, we will want an individual to exhibit only finitely many features.

Hjort (1990) and Kim (1999) originally established that the posterior distribution of B under a constrained form of the BBeP was also a beta process with known parameters. Thibaux and Jordan (2007) went on to extend this analysis to the full BBeP. We cite the result by Thibaux and Jordan (2007) here, using the completely random measure notation established above.

Theorem 1 (The beta process prior is conjugate to the Bernoulli process likelihood). *Let H be a measure with atomic component $H_{fix} = \sum_{l=1}^{L} \rho_l \delta_{u_l}$ and continuous component H_{ord}. Let θ and γ be strictly positive scalars. Consider N conditionally-independent draws from the Bernoulli process: $I_n = \sum_{l=1}^{L} i_{fix,n,l} \delta_{u_l} + \sum_{j=1}^{J} i_{ord,n,j} \delta_{v_j} \overset{\text{iid}}{\sim}$ $\mathrm{BeP}(B)$, for $n = 1, \ldots, N$ with $B \sim \mathrm{BP}(\theta, \gamma, H)$. That is, the Bernoulli process draws have J atoms that are not located at the atoms of H_{fix}. Then, $B | I_1, \ldots, I_N \sim \mathrm{BP}(\theta_{post}, \gamma_{post}, H_{post})$ with $\theta_{post} = \theta + N$, $\gamma_{post} = \gamma \frac{\theta}{\theta+N}$, and $H_{post,ord} = H_{ord}$. Further, $H_{post,fix} = \sum_{l=1}^{L} \rho_{post,l} \delta_{u_l} + \sum_{j=1}^{J} \xi_{post,j} \delta_{v_j}$, where $\rho_{post,l} = \rho_l + (\theta_{post} \gamma_{post})^{-1} \sum_{n=1}^{N} i_{fix,n,l}$ and*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2318721, IEEE Transactions on Pattern Analysis and Machine Intelligence

$\xi_{post,j} = (\theta_{post}\gamma_{post})^{-1}\sum_{n=1}^{N} i_{ord,n,j}$.

Note that the posterior beta-distributed fixed atoms are well-defined since $\xi_{post,j} > 0$ follows from $\sum_{n=1}^{N} i_{ord,n,j} > 0$, which holds by construction. As shown by Thibaux and Jordan (2007), if the underlying beta process is integrated out in the BBeP, we recover the *Indian buffet process* of Griffiths and Ghahramani (2006).

Since the FBP and BP only differ in the fixed atoms, where conjugacy reduces to the finite-dimensional case, Theorem 1 immediately implies the following.

Corollary 2 (The FBP prior is conjugate to the Bernoulli process likelihood)**.**
Assume the conditions of Theorem 1, and consider N conditionally-independent Bernoulli process draws: $I_n = \sum_{l=1}^{L} i_{fix,n,l}\delta_{u_l} + \sum_{j=1}^{J} i_{ord,n,j}\delta_{v_j} \overset{\text{iid}}{\sim} \mathrm{BeP}(B)$, for $n = 1, \ldots, N$ with $B \sim \mathrm{FBP}(\theta, \gamma, \mathbf{u}, \boldsymbol{\rho}, \boldsymbol{\sigma}, H_{ord})$ and $\{\rho_l\}_{l=1}^{L}$ and $\{\sigma_l\}_{l=1}^{L}$ strictly positive scalars. Then, $B|I_1, \ldots, I_N \sim \mathrm{FBP}(\theta_{post}, \gamma_{post}, \mathbf{u}_{post}, \boldsymbol{\rho}_{post}, \boldsymbol{\sigma}_{post}, H_{post,ord})$, for $\theta_{post} = \theta + N$, $\gamma_{post} = \gamma\frac{\theta}{\theta+N}$, $H_{post,ord} = H_{ord}$, and $L + J$ fixed atoms, $\{u_{post,l'}\} = \{u_l\}_{l=1}^{L} \cup \{v_j\}_{j=1}^{J}$. The $\boldsymbol{\rho}_{post}$ and $\boldsymbol{\sigma}_{post}$ parameters satisfy $\rho_{post,l} = \rho_l + \sum_{n=1}^{N} i_{fix,n,l}$ and $\sigma_{post,l} = \sigma_l + N - \sum_{n=1}^{N} i_{fix,n,l}$ for $l \in \{1, \ldots, L\}$ and $\rho_{post,L+j} = \sum_{n=1}^{N} i_{ord,n,j}$ and $\sigma_{post,L+j} = \theta + N - \sum_{n=1}^{N} i_{ord,n,j}$ for $j \in \{1, \ldots, J\}$.

The usefulness of the FBP becomes apparent in the posterior parameterization; the distributions associated with the fixed atoms more closely mirror the classical parametric conjugacy between the Bernoulli distribution and the beta distribution. This is an issue of convenience in the case of the BBeP, but it is more significant in the case of the negative binomial process, as we show in the following section, where conjugacy is preserved only in the FBP case (and not for the traditional, more constrained BP).

3.2 Negative binomial process

The Bernoulli distribution is not the only distribution that yields conjugacy when coupled to the beta distribution in the classical parametric setting; conjugacy holds for the negative binomial distribution as well. As we show in this section, this result can be extended to stochastic processes via the CRM framework.

We define the *negative binomial process* as a CRM with two parameters: a shape parameter $r > 0$ and a discrete base measure $\tilde{H} = \sum_k b_k\delta_{\psi_k}$ whose weights b_k take values in $(0, 1]$. As in the case of the Bernoulli process, $\tilde{H}$ need not be random at this point. Since $\tilde{H}$ is discrete, we again have a representation for $\tilde{H}$ as in Eq. (5), and we say that the random measure I is drawn from a negative binomial process, $I \sim \mathrm{NBP}(r, \tilde{H})$, if $I = \sum_{k=1}^{\infty} i_k\delta_{\psi_k}$ with $i_k \overset{\text{ind}}{\sim} \mathrm{NB}(r, b_k)$ for $k = 1, 2, \ldots$. That is, the negative binomial process is formed by simply making a single draw from a negative binomial distribution at each of the (potentially countably infinite) atoms of $\tilde{H}$. This construction generalizes the geometric process studied by Thibaux (2008).

As a Bernoulli process draw can be interpreted as assigning a set of features to a data point, so can we interpret a draw from the negative binomial process as assigning a set of feature counts to a data point. In particular, as for the Bernoulli process, we assume that each data point has its own draw from the negative binomial process. Every atom with strictly positive mass in this draw corresponds to a feature that is exhibited by this data point. Moreover, the size of the atom, which is a positive integer by construction, dictates how many times the feature is exhibited by the data point. For example, if the data point is a document, and each feature represents a particular word, then the negative binomial process draw would tell us how many occurrences of each word there are in the document.

If the base measure for a negative binomial process is a beta process, we say that the combined process is a *beta-negative binomial process* (BNBP). If the base measure is a three-parameter beta process, we say that the combined process is a *three-parameter beta-negative binomial process* (3BNBP). When either the BP or 3BP has a finite number of fixed atoms, the ordinary component of the BP or 3BP still has an infinite number of atoms, but the number of atoms in the negative binomial process is a.s. finite. We prove this fact and more in Section 5.

We now suppose that the base measure for the negative binomial process is a draw B from an FBP with parameters $\theta > 0, \gamma > 0$, $\{u_l\}_{l=1}^L$, $\{\rho_l\}_{l=1}^L$, $\{\sigma_l\}_{l=1}^L$, and H_{ord}. The overall specification is $B \sim \mathrm{FBP}(\theta, \gamma, \mathbf{u}, \boldsymbol{\rho}, \boldsymbol{\sigma}, H_{ord})$ and $I \sim \mathrm{NBP}(r, B)$. The following theorem characterizes the posterior distribution for this model. The proof is given in Appendix E.

Theorem 3 (The FBP prior is conjugate to the negative binomial process likelihood.)**.** *Let θ and γ be strictly positive scalars. Let $(u_1, \ldots, u_L) \in \Psi^L$. Let the members of $\{\rho_l\}_{l=1}^L$ and $\{\sigma_l\}_{l=1}^L$ be strictly positive scalars. Let H_{ord} be a continuous measure on Ψ. Consider the following model for N draws from a negative binomial process: $I_n = \sum_{l=1}^L i_{fix,n,l}\delta_{u_l} + \sum_{j=1}^J i_{ord,n,j}\delta_{v_j} \overset{\text{iid}}{\sim} \mathrm{NBP}(B)$, for $n = 1, \ldots, N$ with $B \sim \mathrm{FBP}(\theta, \gamma, \mathbf{u}, \boldsymbol{\rho}, \boldsymbol{\sigma}, H_{ord})$. That is, the negative binomial process draws have J atoms that are not located at the atoms of H_{fix}. Then, $B|I_1, \ldots, I_N \sim \mathrm{FBP}(\theta_{post}, \gamma_{post}, \mathbf{u}_{post}, \boldsymbol{\rho}_{post}, \boldsymbol{\sigma}_{post}, H_{post,ord})$ for $\theta_{post} = \theta + Nr$, $\gamma_{post} = \gamma\frac{\theta}{\theta+Nr}$, $H_{post,ord} = H_{ord}$, and $L + J$ fixed atoms, $\{u_{post,l}\} = \{u_l\}_{l=1}^L \cup \{v_j\}_{j=1}^J$. The $\boldsymbol{\rho}_{post}$ and $\boldsymbol{\sigma}_{post}$ parameters satisfy $\rho_{post,l} = \rho_l + \sum_{n=1}^N i_{fix,n,l}$ and $\sigma_{post,l} = \sigma_l + Nr$ for $l \in \{1, \ldots, L\}$ and $\rho_{post,L+j} = \sum_{n=1}^N i_{ord,n,j}$ and $\sigma_{post,L+j} = \theta + Nr$ for $j \in \{1, \ldots, J\}$.*

For the posterior measure to be a BP, we must have $\rho_{post,k} + \sigma_{post,k} = \theta_{post}$ for all k, but this equality can fail to hold even when the prior is a BP. For instance, whenever there are new fixed atom locations in the posterior relative to the prior, this equality will fail. So the BP is not conjugate to the negative binomial process likelihood.

4 Mixtures and admixtures

We now assemble the pieces that we have introduced and consider Bayesian nonparametric models of admixture. Recall that the basic idea of an admixture is that an individual (e.g., an organism, a document, or an image) can belong simultaneously to multiple classes. This can be represented by associating a binary-valued vector with each individual; the vector has value one in components corresponding to classes to which the individual belongs and zero in components corresponding to classes to which the individual does not belong. More generally, we wish to remove the restriction to binary values and consider a general notion of admixture in which an individual is represented by a nonnegative, integer-valued vector. We refer to such vectors as *feature vectors*, and view the components of such vectors as counts representing the number of times the corresponding feature is exhibited by a given individual. For example, a document may exhibit a given word zero or more times.

As we discussed in Section 1, the standard approach to modeling an admixture is to assume that there is an exchangeable set of data associated with each individual and to assume that these data are drawn from a finite mixture model with individual-specific mixing proportions. There is another way to view this process, however, that opens the door to a variety of extensions. Note that to draw a set of data from a mixture, we can first choose the number of data points to be associated with each mixture component (a vector of counts) and then draw the data point values independently from each selected mixture component. That is, we randomly draw nonnegative integers i_k for each mixture component (or *cluster*) k. Then, for each k and each $n = 1, \dots, i_k$, we draw a data point $x_{k,n} \sim F(\psi_k)$, where ψ_k is the parameter associated with mixture component k. The overall collection of data for this individual is $\{x_{k,n}\}_{k,n}$, with $N = \sum_k i_k$ total points. One way to generate data according to this decomposition is to make use of the NBP. We draw $I = \sum_k i_k \delta_{\psi_k} \sim \mathrm{NBP}(r, B)$, where B is drawn from a beta process, $B \sim \mathrm{BP}(\theta, \gamma, H)$. The overall model is a BNBP mixture model for the counts, coupled to a conditionally independent set of draws for the individual's data points $\{x_{k,n}\}_{k,n}$.

An alternative approach in the same spirit is to make use of a gamma process (to obtain a set of rates) that is coupled to a Poisson likelihood process (PLP)[3] to convert the rates into counts (Titsias, 2008). In particular, given a base measure $\tilde{G} = \sum_k \tilde{g}_k \delta_{\psi_k}$, let $I \sim \mathrm{PLP}(\tilde{G})$ denote $I = \sum_k i_k \delta_{\psi_k}$, with $i_k \sim \mathrm{Pois}(\tilde{g}_k)$. We then consider a *gamma Poisson likelihood process* (ΓPLP) as follows: $\tilde{G} \sim \Gamma\mathrm{P}(\theta, c, H)$, $I = \sum_k i_k \delta_{\psi_k} \sim \mathrm{PLP}(\tilde{G})$, and $x_{k,n} \sim F(\psi_k)$, for $n = 1, \dots, i_k$ and each k.

Both the BNBP approach and the ΓPLP approach deliver a random measure, $I = \sum_k i_k \delta_{\psi_k}$, as a representation of an admixed individual.[4] While the atom locations, (ψ_k), are subsequently used to generate data points, the pattern of admixture inheres in the vector of weights (i_k). It is thus natural to view this vector as the representation of an admixed individual. Indeed, in some problems such a weight vector might itself be the observed data. In other problems, the weights may be used to generate data in some more complex way that does not simply involve conditionally i.i.d. draws.

This perspective on admixture—focusing on the vector of weights (i_k) rather than the data associated with an individual—is also natural when we consider multiple individuals. The main issue becomes that of linking these vectors among multiple individuals; this linking can readily be achieved in the Bayesian formalism via a hierarchical model. In the remainder of this section we consider examples of such hierarchies in the Bayesian nonparametric setting.

Let us first consider the standard approach to admixture in which an individual is represented by a set of

[3] We use the terminology "Poisson likelihood process" to distinguish a particular process with Poisson distributions affixed to each atom of some base distribution from the more general Poisson point process of Kingman (1993).

[4] We elaborate on the parallels and deep connections between the BNBP and ΓPLP in Appendix A.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2318721, IEEE Transactions on Pattern Analysis and Machine Intelligence

draws from a mixture model. For each individual we need to draw a set of mixing proportions, and these mixing proportions need to be coupled among the individuals. This can be achieved via a prior known as the *hierarchical Dirichlet process* (HDP) (Teh et al., 2006):

$$\begin{aligned} &G_0 \sim \mathrm{DP}(\theta, H) \\ &G_d = \sum_k g_{d,k} \delta_{\psi_k} \overset{\mathrm{ind}}{\sim} \mathrm{DP}(\theta_d, G_0), \quad d = 1, 2, \ldots, \end{aligned}$$

where the index d ranges over the individuals. Note that the global measure G_0 is a discrete random probability measure, given that it is drawn from a Dirichlet process. In drawing the individual-specific random measure G_d at the second level, we therefore resample from among the atoms of G_0 and do so according to the weights of these atoms in G_0. This shares atoms among the individuals and couples the individual-specific mixing proportions $g_{d,k}$. We complete the model specification as follows:

$$\begin{aligned} &z_{d,n} \overset{\mathrm{iid}}{\sim} (g_{d,k})_k \quad \text{for } n = 1, \ldots, N_d \\ &x_{d,n} \overset{\mathrm{ind}}{\sim} F(\psi_{z_{d,n}}), \end{aligned}$$

which draws an index $z_{d,n}$ from the discrete distribution $(g_{d,k})_k$ and then draws a data point $x_{d,n}$ from a distribution indexed by $z_{d,n}$. For instance, $(g_{d,k})$ might represent topic proportions in document d; $\psi_{z_{d,n}}$ might represent a topic, i.e. a distribution over words; and $x_{d,n}$ might represent the nth word in the dth document.

In the HDP, N_d is known for each d and is part of the model specification. We propose to instead take the featural approach as follows; we draw an individual-specific set of counts from an appropriate stochastic process and then generate the appropriate number of data points for each individual. Then the number of data points for each individual is itself a random variable and potentially coupled across individuals. In particular, one might consider the following conditional independence hierarchy involving the NBP:

$$\begin{aligned} &B_0 \sim \mathrm{BP}(\theta, \gamma, H) \\ &I_d = \sum_k i_{d,k} \delta_{\psi_k} \overset{\mathrm{ind}}{\sim} \mathrm{NBP}(r_d, B_0), \end{aligned} \tag{6}$$

where we first draw a random measure B_0 from the beta process and then draw multiple times from an NBP with base measure given by B_0.

Although this conditional independence hierarchy does couple count vectors across multiple individuals, it uses a single collection of mixing proportions, the atom weights of B_0, for all individuals. By contrast, the HDP draws individual-specific mixing proportions from an underlying set of population-wide mixing proportions—and then converts these mixing proportions into counts. We can model individual-specific, but coupled, mixing proportions within an NBP-based framework by simply extending the hierarchy by one level:

$$\begin{aligned} &B_0 \sim \mathrm{BP}(\theta, \gamma, H) \\ &B_d \overset{\mathrm{ind}}{\sim} \mathrm{BP}(\theta_d, \gamma_d, B_0/B_0(\Psi)) \end{aligned} \tag{7}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2318721, IEEE Transactions on Pattern Analysis and Machine Intelligence

$$I_d = \sum_k i_{d,k}\delta_{\psi_k} \overset{\text{ind}}{\sim} \text{NBP}(r_d, B_d).$$

Since B_0 is almost surely an atomic measure, the atoms of each B_d will coincide with those of B_0 almost surely. The weights associated with these atoms can be viewed as individual-specific feature probability vectors. We refer to this prior as the *hierarchical beta-negative binomial process* (HBNBP).

We also note that it is possible to consider additional levels of structure in which a population is decomposed into subpopulations and further decomposed into subsubpopulations and so on, bottoming out in a set of individuals. This tree structure can be captured by repeated draws from a set of beta processes at each level of the tree, conditioning on the beta process at the next highest level of the tree. Hierarchies of this form have previously been explored for beta-Bernoulli processes by Thibaux and Jordan (2007).

Comparison with Zhou et al. (2012). Zhou et al. (2012) have independently proposed a (non-hierarchical) beta-negative binomial process prior

$$B_0 = \sum_k b_k \delta_{r_k,\psi_k} \sim \text{BP}(\theta, \gamma, R \times H)$$

$$I_d = \sum_k i_{d,k}\delta_{\psi_k} \quad \text{where} \quad i_{d,k} \overset{\text{ind}}{\sim} \text{NB}(r_k, b_k),$$

where R is a continuous finite measure over $\mathbb{R}^+$ used to associate a distinct failure parameter r_k with each beta process atom. Note that each individual is restricted to use the same failure parameters and the same beta process weights under this model. In contrast, our BNBP formulation (6) offers the flexibility of differentiating individuals by assigning each its own failure parameter r_d. Our HBNBP formulation (7) further introduces heterogeneity in the individual-specific beta process weights by leveraging the hierarchical beta process. We will see that these modeling choices are particularly well-suited for admixture modeling in the coming sections.

Zhou et al. (2012) use their prior to develop a Poisson factor analysis model for integer matrix factorization, while our primary motivation is mixture and admixture modeling. Our differing models and motivating applications have led to different challenges and algorithms for posterior inference. While Zhou et al. (2012) develop an inexact inference scheme based on a finite approximation to the beta process, we develop both an exact Markov chain Monte Carlo sampler and a finite approximation sampler for posterior inference under the HBNBP (see Section 7). Finally, unlike Zhou et al. (2012), we provide an extensive theoretical analysis of our priors including a proof of the conjugacy of the full beta process and the NBP (given in Section 3) and an asymptotic analysis of the BNBP (see Section 5).

5 Asymptotics

An important component of choosing a Bayesian prior is verifying that its behavior aligns with our beliefs about the behavior of the data-generating mechanism. In models of clustering, a particular measure of interest is the *diversity*—the dependence of the number of clusters on the number of data points. In speaking of the diversity, we typically assume a finite number of fixed atoms in a process derived from a CRM, so that asymptotic behavior is dominated by the ordinary component.

It has been observed in a variety of different contexts that the number of clusters in a dataset grows as a *power law* of the size of the data; that is, the number of clusters is asymptotically proportional to the number of data points raised to some positive power (Gnedin et al., 2007). Real-world examples of such behavior are provided by Newman (2005) and Mitzenmacher (2004).

The diversity has been characterized for the Dirichlet process (DP) and a two-parameter extension to the Dirichlet process known as the *Pitman-Yor process* (PYP) (Pitman and Yor, 1997), with extra parameter $\alpha \in (0, 1)$ and concentration parameter $\theta > -\alpha$. We will see that while the number of clusters generated according to a DP grows as a logarithm of the size of the data, the number of clusters generated according to a PYP grows as a power of the size of the data. Indeed, the popularity of the Pitman-Yor process—as an alternative prior to the Dirichlet process in the clustering domain—can be attributed to this power-law growth (Goldwater et al., 2006; Teh, 2006; Wood et al., 2009). In this section, we derive analogous asymptotic results for the BNBP treated as a clustering model.

We first highlight a subtle difference between our model and the Dirichlet process. For a Dirichlet process, the number of data points N is known a priori and fixed. An advantage of our model is that it models the number of data points N as a random variable and therefore has potentially more predictive power in modeling multiple populations. We note that a similar effect can be achieved for the Dirichlet process by using the gamma process for feature modeling as described in Section 4 rather than normalizing away the mass that determines the number of observations. However, there is no such unnormalized completely random measure for the PYP (Pitman and Yor, 1997). We thus treat N as fixed for the DP and PYP, in which case the number of clusters $K(N)$ is a function of N. On the other hand, the number of data points $N(r)$ depends on r in the case of the BNBP, and the number of clusters $K(r)$ does as well. We also define $K_j(N)$ to be the number of clusters with exactly j elements in the case of the DP and PYP, and we define $K_j(r)$ to be the number of clusters with exactly j elements in the BNBP case.

For the DP and PYP, $K(N)$ and $K_j(N)$ are random even though N is fixed, so it will be useful to also define their expectations:

$$\Phi(N) \triangleq \mathbb{E}[K(N)], \quad \Phi_j(N) \triangleq \mathbb{E}[K_j(N)]. \tag{8}$$

In the BNBP and 3BNBP cases, all of $K(r)$, $K_j(r)$, and $N(r)$ are random. So we further define

$$\Phi(r) \triangleq \mathbb{E}[K(r)], \quad \Phi_j(r) \triangleq \mathbb{E}[K_j(r)], \quad \xi(r) \triangleq \mathbb{E}[N(r)]. \tag{9}$$

We summarize the results that we establish in this section in Table 1, where we also include comparisons to existing results for the DP and PYP.[5] The full statements of our results, from which the table is derived, can be found in Appendix C, and proofs are given in Appendix D.

[5] The reader interested in power laws may also note that the generalized gamma process is a completely random measure that, when normalized, provides a probability measure for clusters that has asymptotic behavior similar to the PYP; in particular, the expected number of clusters grows almost surely as a power of the size of the data (Lijoi et al., 2007).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2318721, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 1: Let N be the number of data points when this number is fixed and $\xi(r)$ be the expected number of data points when N is random. Let $\Phi(N)$, $\Phi_j(N)$, $\Phi(r)$, and $\Phi_j(r)$ be the expected number of clusters under various scenarios and defined as in Eqs. (8) and (9). The upper part of the table gives the asymptotic behavior of Φ up to a multiplicative constant, and the bottom part of the table gives the multiplicative constants. For the DP, $\theta > 0$. For the PYP, $\alpha \in (0, 1)$ and $\theta > -\alpha$. For the BNBP, $\theta > 1$. For the 3BNBP, $\alpha \in (0, 1)$ and $\theta > 1 - \alpha$.

Process	Expected number of clusters	Expected number of clusters of size j
	Function of N or $\xi(r)$	
DP	$\log(N)$	1
PYP	N^α	N^α
BNBP	$\log(\xi(r))$	1
3BNBP	$(\xi(r))^\alpha$	$(\xi(r))^\alpha$
	Constants	
DP	θ	θj^{-1}
PYP	$\frac{\Gamma(\theta+1)}{\alpha\Gamma(\theta+\alpha)}$	$\frac{\Gamma(\theta+1)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} \frac{\Gamma(j-\alpha)}{\Gamma(j+1)}$
BNBP	$\gamma\theta$	$\gamma\theta j^{-1}$
3BNBP	$\frac{\gamma^{1-\alpha}}{\alpha} \frac{\Gamma(\theta+1)}{\Gamma(\theta+\alpha)} \left(\frac{\theta+\alpha-1}{\theta}\right)^\alpha$	$\gamma^{1-\alpha} \frac{\Gamma(\theta+1)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} \frac{\Gamma(j-\alpha)}{\Gamma(j+1)} \left(\frac{\theta+\alpha-1}{\theta}\right)^\alpha$

The table shows, for example, that for the DP, $\Phi(N) \sim \theta \log(N)$ as $N \to \infty$, and, for the BNBP, $\Phi_j(r) \sim \gamma\theta j^{-1}$ as $r \to \infty$ (i.e., constant in r). The result for the expected number of clusters for the DP can be found in Korwar and Hollander (1973); results for the expected number of clusters for both the DP and PYP can be found in Pitman (2006, Eq. (3.24) on p. 69 and Eq. (3.47) on p. 73). Note that in all cases the expected counts of clusters of size j are asymptotic expansions in terms of r for fixed j and should not be interpreted as asymptotic expansions in terms of j.

We conclude that, just as for the Dirichlet process, the BNBP can achieve both logarithmic cluster number growth in the basic model and power law cluster number growth in the expanded, three-parameter model.

6 Simulation

Our theoretical results in Section 5 are supported by simulation results, summarized in Figure 1; in particular, our simulation corroborated the existence of power laws in the three-parameter beta process case examined in Section 5. The simulation was performed as follows. For values of the negative binomial parameter r evenly spaced between 1 and 1,001, we generated beta process weights according to a beta process (or three-parameter beta process) using a stick-breaking representation (Paisley et al., 2010; Broderick et al., 2012). For each of the resulting atoms, we simulated negative binomial draws to arrive at a sample from a BNBP. For each such BNBP, we can count the resulting total number of data points N and total number of clusters K. Thus, each r gives us an (r, N, K) triple.

In the simulation, we set the mass parameter $\gamma = 3$. We set the concentration parameter $\theta = 3$; in particular, we note that the analysis in Section 5 implies that we should always have $\theta > 1$. Finally, we ran the simulation for both the $\alpha = 0$ case, where we expect no power law behavior, and the $\alpha = 0.5$ case, where we do expect power law behavior. The results are shown in Figure 1. Is this figure, we scatter plot the (r, K) tuples from the generated (r, N, K) triples on the left and plot the (N, K) tuples on the right.

In the left plot, the upper black points represent the simulation with $\alpha = 0.5$, and the lower blue data points represent the $\alpha = 0$ case. The lower red line illustrates the asymptotic theoretical result corresponding to the $\alpha = 0$ case (Lemma 12 in Appendix C), and we can see that the anticipated logarithmic growth behavior agrees with our simulation. The upper red line illustrates the theoretical result for the $\alpha = 0.5$ case (Lemma 13 in Appendix C). The agreement between simulation and theory here demonstrates that, in contrast to the $\alpha = 0$ case, the $\alpha = 0.5$ case exhibits power law growth in the number of clusters K as a function of the negative binomial parameter r.

Our simulations also bear out that the expectation of the random number of data points N increases linearly with r (Lemmas 10 and 11 in Appendix C). We see, then, on the right side of Figure 1 the behavior of the number of clusters K now plotted as a function of N. As expected given the asymptotics of the expected value of N, the behavior in the right plot largely mirrors the behavior in the left plot. Just as in the left plot, the lower red line (Theorem 16 in Appendix C) shows the anticipated logarithmic growth of K and N when $\alpha = 0$. And the upper red line (Theorem 17 in Appendix C) shows the anticipated power law growth of K and N when $\alpha = 0.5$.

We can see the parallels with the DP and PYP here. Clusters generated from the Dirichlet process (i.e., Pitman-Yor process with $\alpha = 0$) exhibit logarithmic growth of the expected number of clusters K as the (deterministic) number of data points N grows. And clusters generated from the Pitman-Yor process with $\alpha \in (0, 1)$ exhibit power law behavior in the expectation of K as a function of (fixed) N. So too do we see that the BNBP, when applied to clustering problems, yields asymptotic growth similar to the DP and that the 3BNBP yields asymptotic growth similar to the PYP.

7 Posterior inference

In this section we present posterior inference algorithms for the HBNBP. We focus on the setting in which, for each individual d, there is an associated exchangeable sequence of observations $(x_{d,n})_{n=1}^{N_d}$. We seek to infer both the admixture component responsible for each observation and the parameter ψ_k associated with each component. Hereafter, we let $z_{d,n}$ denote the unknown component index associated with $x_{d,n}$, so that $x_{d,n} \sim F(\psi_{z_{d,n}})$.

Under the HBNBP admixture model introduced in Section 4, the posterior over component indices and parameters has the form

$$p(\mathbf{z}_{\cdot,\cdot}, \boldsymbol{\psi}_\cdot \mid \mathbf{x}_{\cdot,\cdot}, \Theta) \propto p(\mathbf{z}_{\cdot,\cdot}, \boldsymbol{\psi}_\cdot, \mathbf{b}_{0,\cdot}, \mathbf{b}_{\cdot,\cdot} \mid \mathbf{x}_{\cdot,\cdot}, \Theta),$$

where $\Theta \triangleq (F, H, \gamma_0, \theta_0, \boldsymbol{\gamma}_\cdot, \boldsymbol{\theta}_\cdot, \mathbf{r}_\cdot)$ is the collection of all fixed hyperparameters. As is the case with HDP admixtures (Teh et al., 2006) and earlier hierarchical beta process featural models (Thibaux and Jordan, 2007), the posterior of the HBNBP admixture cannot be obtained in analytical form due to complex couplings in the marginal $p(\mathbf{x}_{\cdot,\cdot} \mid \Theta)$. We therefore develop Gibbs sampling algorithms (Geman and Geman, 1984) to draw samples of the relevant latent variables from their joint posterior.

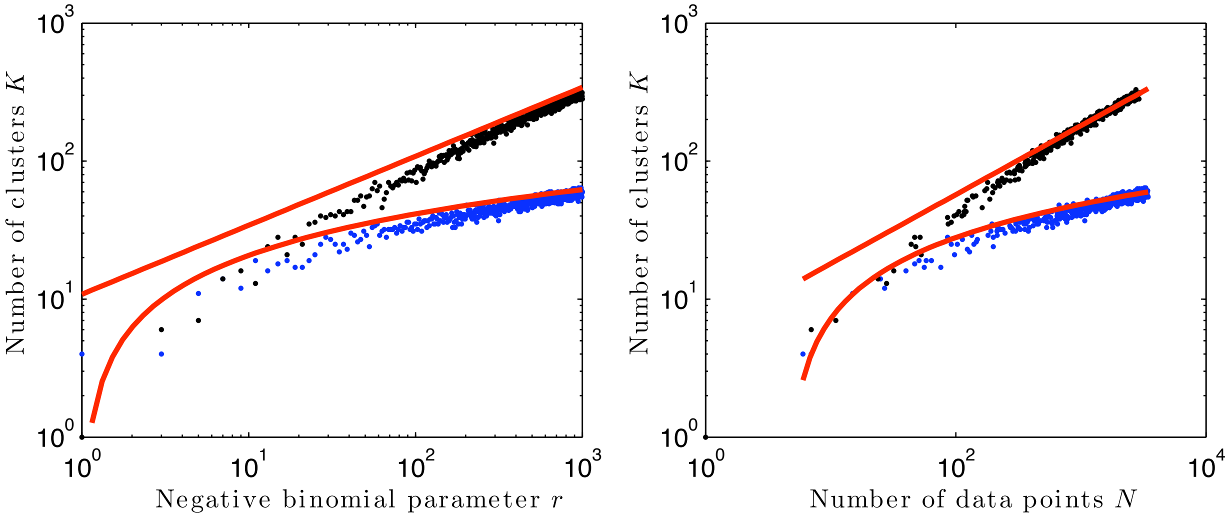


Fig. 1: For each r evenly spaced between 1 and 1,001, we simulate (random) values of the number of data points N and number of clusters K from the BNBP and 3BNBP. In both plots, we have mass parameter $\gamma = 3$ and concentration parameter $\theta = 3$. On the *left*, we see the number of clusters K as a function of the negative binomial parameter r (see Lemma 12 and Lemma 13 in Appendix C); on the *right*, we see the number of clusters K as a function of the (random) number of data points N (see Theorem 16 and Theorem 17 in Appendix C). In both plots, the upper black points show simulation results for the case $\alpha = 0.5$, and the lower blue points show $\alpha = 0$. Red lines indicate the theoretical asymptotic mean behavior we expect from Section 5.

A challenging aspect of inference in the nonparametric setting is the countable infinitude of component parameters and the countably infinite support of the component indices. We develop two sampling algorithms that cope with this issue in different ways. In Section 7.1, we use slice sampling to control the number of components that need be considered on a given round of sampling and thereby derive an exact Gibbs sampler for posterior inference under the HBNBP admixture model. In Section 7.2, we describe an efficient alternative sampler that makes use of a finite approximation to the beta process. Throughout we assume that the base measure H is continuous. We note that neither procedure requires conjugacy between the base distribution H and the data-generating distribution F.

7.1 Exact Gibbs slice sampler

Slice sampling (Damien et al., 1999; Neal, 2003) has been successfully employed in several Bayesian nonparametric contexts, including Dirichlet process mixture modeling (Walker, 2007; Papaspiliopoulos, 2008; Kalli et al., 2011) and beta process feature modeling (Teh et al., 2007). The key to its success lies in the introduction of one or more auxiliary variables that serve as adaptive truncation levels for an infinite sum representation of the stochastic process.

This adaptive truncation procedure proceeds as follows. For each observation associated with individual d, we introduce an auxiliary variable $u_{d,n}$ with conditional distribution

$$u_{d,n} \sim \mathrm{Unif}(0, \zeta_{d,z_{d,n}}),$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2318721, IEEE Transactions on Pattern Analysis and Machine Intelligence

where $(\zeta_{d,k})_{k=1}^{\infty}$ is a fixed positive sequence with $\lim_{k\to\infty} \zeta_{d,k} = 0$. To sample the component indices, we recall that a negative binomial draw $i_{d,k} \sim \text{NB}(r_d, b_{d,k})$ may be represented as a gamma-Poisson mixture:

$$\lambda_{d,k} \sim \text{Gamma}\left(r_d, \frac{1 - b_{d,k}}{b_{d,k}}\right)$$
$$i_{d,k} \sim \text{Pois}(\lambda_{d,k}).$$

We first sample $\lambda_{d,k}$ from its full conditional. By gamma-Poisson conjugacy, this has the simple form

$$\lambda_{d,k} \sim \text{Gamma}\left(r_d + i_{d,k}, 1/b_{d,k}\right).$$

We next note that, given $\boldsymbol{\lambda}_{d,\cdot}$ and the total number of observations associated with individual d, the cluster sizes $i_{d,k}$ may be constructed by sampling each $z_{d,n}$ independently from $\boldsymbol{\lambda}_{d,\cdot}/\sum_k \lambda_{d,k}$ and setting $i_{d,k} = \sum_n \mathbb{I}(z_{d,n} = k)$. Hence, conditioned on the number of data points N_d, the component parameters ψ_k, the auxiliary variables $\lambda_{d,k}$, and the slice-sampling variable $u_{d,n}$, we sample the index $z_{d,n}$ from a discrete distribution with

$$\mathbb{P}(z_{d,n} = k) \propto F(dx_{d,n} \mid \psi_k) \frac{\mathbb{I}(u_{d,n} \leq \zeta_{d,k})}{\zeta_{d,k}} \lambda_{d,k}$$

so that only the finite set of component indices $\{k : \zeta_{d,k} \geq u_{d,n}\}$ need be considered when sampling $z_{d,n}$.

Let $K_d \triangleq \max\{k : \exists n \text{ s.t. } \zeta_{d,k} \geq u_{d,n}\}$ and $K \triangleq \max_d K_d$. Then, on a given round of sampling, we need only explicitly represent $\lambda_{d,k}$ and $b_{d,k}$ for $k \leq K_d$ and ψ_k and $b_{0,k}$ for $k \leq K$. The simple Gibbs conditionals for $b_{d,k}$ and ψ_k can be found in Appendix F.1. To sample the shared beta process weights $b_{0,k}$, we leverage the size-biased construction of the beta process introduced by Thibaux and Jordan (2007):

$$B_0 = \sum_{m=0}^{\infty} \sum_{i=1}^{C_m} b_{0,m,i} \delta_{\psi_{m,i,\cdot}},$$

where

$$C_m \overset{\text{ind}}{\sim} \text{Pois}\left(\frac{\theta_0\gamma_0}{\theta_0 + m}\right), \quad b_{0,m,i} \overset{\text{ind}}{\sim} \text{Beta}(1, \theta_0 + m), \quad \text{and} \quad \psi_{m,i,\cdot} \overset{\text{iid}}{\sim} H,$$

and we develop a Gibbs slice sampler for generating samples from its posterior. The details are deferred to Appendix F.1.

7.2 Finite approximation Gibbs sampler

An alternative to the size-biased construction of B_0 is a finite approximation to the beta process with a fixed number of components, K:

$$b_{0,k} \overset{\text{iid}}{\sim} \text{Beta}(\theta_0\gamma_0/K, \theta_0(1 - \gamma_0/K)), \quad \psi_k \overset{\text{iid}}{\sim} H, \quad k \in \{1, \ldots, K\}. \tag{10}$$

It is known that, when H is continuous, the distribution of $\sum_{k=1}^{K} b_{0,k}\delta_{\psi_k}$ converges to $\text{BP}(\theta_0, \gamma_0, H)$ as the number of components $K \to \infty$ (see the proof of Theorem 3.1 by Hjort (1990) with the choice $A_0(t) = \gamma$). Hence, we may leverage the beta process approximation (10) to develop an approximate posterior sampler for the HBNBP admixture model with an approximation level K that trades off between computational efficiency and

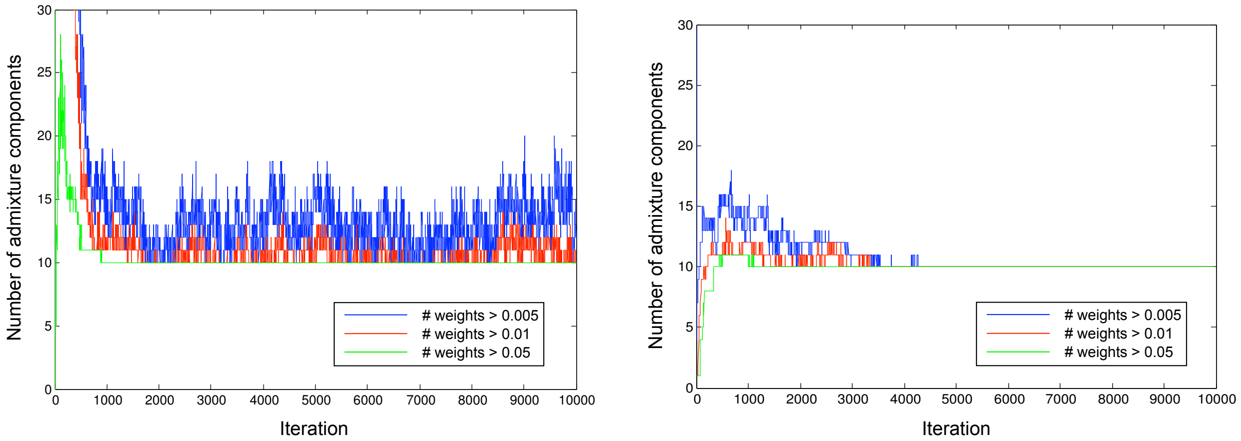


Fig. 2: Number of admixture components used by the finite approximation sampler with $K = 100$ (*left*) and the exact Gibbs slice sampler (*right*) on each iteration of HBNBP admixture model posterior inference. We use a standard "toy bars" dataset with ten underlying admixture components (cf. Griffiths and Steyvers (2004)). We declare a component to be used by a sample if the sampled beta process weight, $b_{0,k}$, exceeds a small threshold. Both the exact and the finite approximation sampler find the correct underlying structure, while the finite sampler attempts to innovate more because of the larger number of proposal components available to the data in each iteration.

fidelity to the true posterior. We defer the detailed conditionals of the resulting Gibbs sampler to Appendix F.3 and briefly compare the behavior of the finite and exact samplers on a toy dataset in Figure Figure 2. We note finally that the beta process approximation in Eq. (10) also gives rise to a new finite admixture model that may be of interest in its own right; we explore the utility of this HBNBP approximation in Section 9.

8 Document Topic Modeling

In the next two sections, we show how the HBNBP admixture model and its finite approximation can be used as practical building blocks for more complex supervised and unsupervised inferential tasks.

We first consider the unsupervised task of *document topic modeling*, in which each individual d is a document containing N_d observations (words) and each word $x_{d,n}$ belongs to a vocabulary of size V. The topic modeling framework is an instance of admixture modeling in which we assume that each word of each document is generated from a latent admixture component or *topic*, and our goal is to infer the topic underlying each word.

In our experiments, we let H_{ord}, the Ψ dimension of the ordinary component intensity measure, be a Dirichlet distribution with parameter $\eta\mathbf{1}$ for $\eta = 0.1$ and $\mathbf{1}$ a V-dimensional vector of ones and let $F(\psi_k)$ be $\mathrm{Mult}(1, \psi_k)$. We use the setting $(\gamma_0, \theta_0, \gamma_d, \theta_d) = (3, 3, 1, 10)$ for the global and document-specific mass and concentration parameters and set the document-specific negative binomial shape parameter according to the heuristic $r_d = N_d(\theta_0 - 1)/(\theta_0\gamma_0)$. We arrive at this heuristic by matching N_d to its expectation under a non-hierarchical BNBP model and solving for r_d:

$$\mathbb{E}[N_d] = r_d \mathbb{E}\left[\sum_{k=1}^{\infty} b_{d,k}/(1 - b_{d,k})\right] = \gamma_0\theta_0/(\theta_0 - 1).$$

When applying the exact Gibbs slice sampler, we let the slice sampling decay sequence follow the same pattern across all documents: $\zeta_{d,k} = 1.5^{-k}$.

TABLE 2: The number of incidents claimed by each organization in the WITS perpetrator identification experiment.

Group ID	Perpetrator	# Claimed Incidents
1	taliban	2647
2	al-aqsa	417
3	farc	76
4	izz al-din al-qassam	478
5	hizballah	89
6	al-shabaab al-islamiya	426
7	al-quds	505
8	abu ali mustafa	249
9	al-nasser salah al-din	212
10	communist party of nepal (maoist)	291

8.1 Worldwide Incidents Tracking System

We report results on the Worldwide Incidents Tracking System (WITS) dataset.[6] This dataset consists of reports on 79,754 terrorist attacks from the years 2004 through 2010. Each event contains a written summary of the incident, location information, victim statistics, and various binary fields such as "assassination," "IED," and "suicide." We transformed each incident into a text document by concatenating the summary and location fields and then adding further words to account for other, categorical fields: e.g., an incident with seven hostages would have the word "hostage" added to the document seven times. We used a vocabulary size of $V = 1{,}048$ words.

Perpetrator Identification. Our experiment assesses the ability of the HBNBP admixture model to discriminate among incidents perpetrated by different organizations. We first grouped documents according to the organization claiming responsibility for the reported incident. We considered 5,390 claimed documents in total distributed across the ten organizations listed in Table 2. We removed all organization identifiers from all documents and randomly set aside 10% of the documents in each group as test data. Next, for each group, we trained an independent, organization-specific HBNBP model on the remaining documents in that group by drawing 10,000 MCMC samples. We proceeded to classify each test document by measuring the likelihood of the document under each trained HBNBP model and assigning the label associated with the largest likelihood. The resulting confusion matrix across the ten candidate organizations is displayed in Table 3a. Results are reported for the exact Gibbs slice sampler; performance under the finite approximation sampler is nearly identical.

For comparison, we carried out the same experiment using the more standard HDP admixture model in place of the HBNBP. For posterior inference, we used the HDP block sampler code of Yee Whye Teh[7] and initialized the sampler with 100 topics and topic hyperparameter $\eta = 0.1$ (all remaining parameters were set to their default values). For each organization, we drew 250,000 MCMC samples and kept every twenty-fifth sample for evaluation. The confusion matrix obtained through HDP modeling is displayed in Table 3b. We see

[6] https://wits.nctc.gov

[7] http://www.gatsby.ucl.ac.uk/~ywteh/research/npbayes/npbayes-r1.tgz

TABLE 3: Confusion matrices for WITS perpetrator identification. See Table 2 for the organization names matching each group ID.

(a) HBNBP Confusion Matrix

		Predicted Groups									
		1	2	3	4	5	6	7	8	9	10
Actual Groups	1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2	0.00	0.38	0.00	0.02	0.00	0.00	0.29	0.29	0.02	0.00
	3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.54	0.00	0.00	0.15	0.27	0.04	0.00
	5	0.11	0.33	0.00	0.11	0.44	0.00	0.00	0.00	0.00	0.00
	6	0.02	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00
	7	0.00	0.10	0.00	0.06	0.02	0.00	0.48	0.30	0.04	0.00
	8	0.00	0.04	0.00	0.00	0.00	0.00	0.16	0.76	0.04	0.00
	9	0.00	0.10	0.00	0.05	0.10	0.00	0.29	0.43	0.05	0.00
	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(b) HDP Confusion Matrix

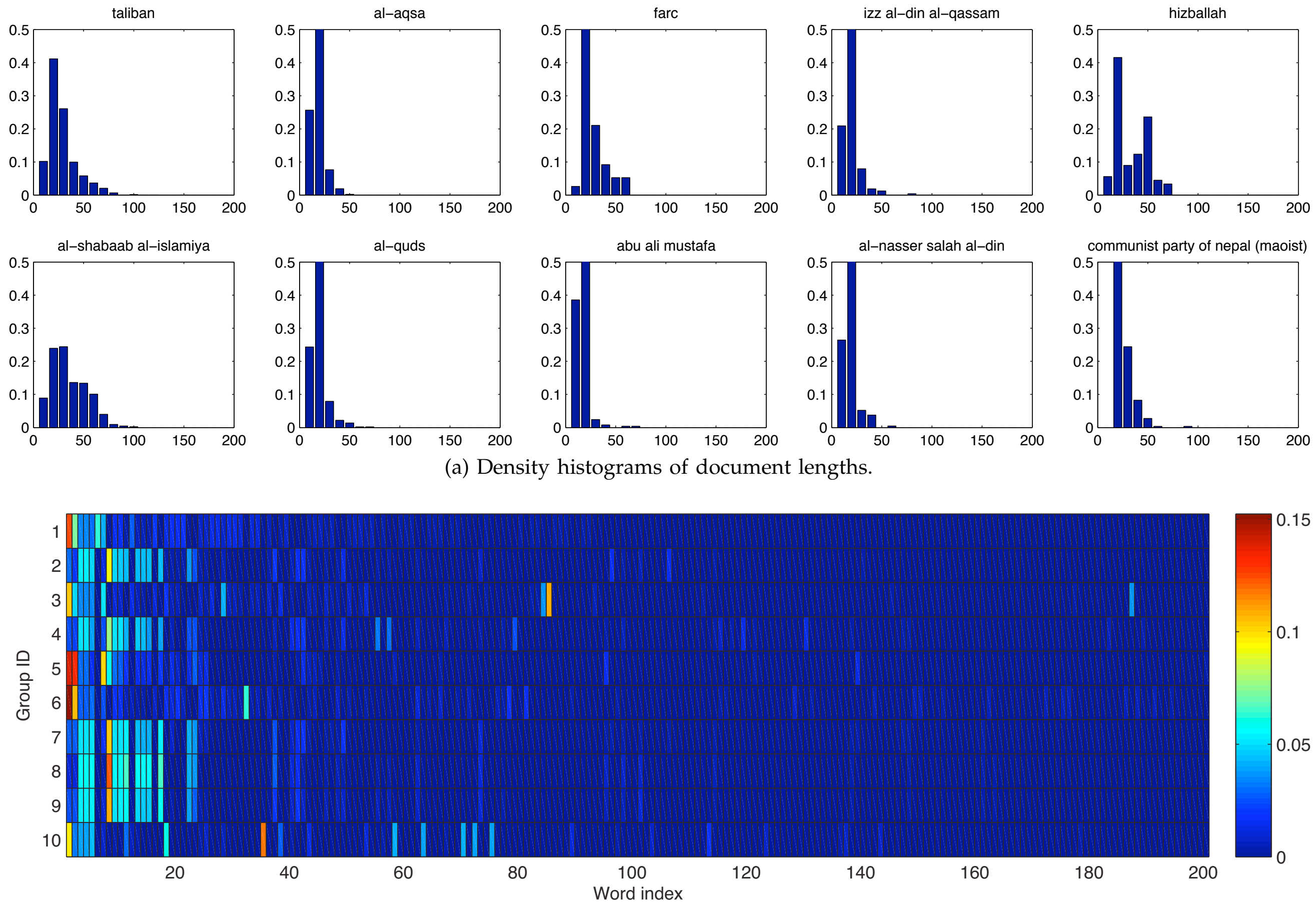
		Predicted Groups									
		1	2	3	4	5	6	7	8	9	10
Actual Groups	1	0.46	0.00	0.26	0.00	0.03	0.23	0.00	0.00	0.00	0.01
	2	0.00	0.31	0.02	0.02	0.00	0.00	0.29	0.36	0.00	0.00
	3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.52	0.04	0.00	0.06	0.31	0.06	0.00
	5	0.11	0.00	0.00	0.00	0.44	0.00	0.11	0.11	0.11	0.11
	6	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
	7	0.00	0.10	0.00	0.04	0.00	0.00	0.38	0.42	0.06	0.00
	8	0.00	0.04	0.00	0.00	0.00	0.00	0.08	0.84	0.04	0.00
	9	0.00	0.05	0.00	0.10	0.00	0.00	0.24	0.62	0.00	0.00
	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

that, overall, HBNBP modeling leads to more accurate identification of perpetrators than its HDP counterpart. Most notably, the HDP wrongly attributes more than half of all documents from group 1 (taliban) to group 3 (farc) or group 6 (al-shabaab al-islamiya). We hypothesize that the HBNBP's superior discriminative power stems from its ability to distinguish between documents both on the basis of word frequency and on the basis of document length.

We would expect the HBNBP to have greatest difficulty discriminating among perpetrators when both word usage frequencies and document length distributions are similar across groups. To evaluate the extent to which this occurs in our perpetrator identification experiment, for each organization, we plotted the density histogram of document lengths in Figure 3a and the heat map displaying word usage frequency across all associated documents in Figure 3b. We find that the word frequency patterns are nearly identical across groups 2, 7, 8, and 9 (al-aqsa, al-quds, abu ali mustafa, and al-nasser salah al-din, respectively) and that the document length distributions of these four groups are all well aligned. As expected, the majority of classification errors made by our HBNBP models result from misattribution among these same four groups. The same group similarity structure is evidenced in a display of the ten most probable words from the most probable HBNBP topic for each group, Table 4. There, we also find an intuitive summary of the salient regional and methodological vocabulary associated with each organization.

TABLE 4: The ten most probable words from the most probable topic in the final MCMC sample of each group in the WITS perpetrator identification experiment. The topic probability is given in parentheses. See Table 2 for the organization names matching each group ID.

HBNBP: Top topic per organization	
group 1 (0.29)	afghanistan, assailants, claimed, responsibility, armedattack, fired, police, victims, armed, upon
group 2 (0.77)	israel, assailants, armedattack, responsibility, fired, claimed, district, causing, southern, damage
group 3 (0.95)	colombia, victims, facility, wounded, armed, claimed, forces, revolutionary, responsibility, assailants
group 4 (0.87)	israel, fired, responsibility, claimed, armedattack, causing, injuries, district, southern, assailants
group 5 (0.95)	victims, wounded, facility, israel, responsibility, claimed, armedattack, fired, rockets, katyusha
group 6 (0.54)	wounded, victims, somalia, civilians, wounding, facility, killing, mortars, armedattack, several
group 7 (0.83)	israel, district, southern, responsibility, claimed, fired, armedattack, assailants, causing, injuries
group 8 (0.94)	israel, district, southern, armedattack, claimed, fired, responsibility, assailants, causing, injuries
group 9 (0.88)	israel, district, southern, fired, responsibility, claimed, armedattack, assailants, causing, injuries
group 10 (0.80)	nepal, victims, hostage, assailants, party, communist, claimed, front, maoist/united, responsibility



(a) Density histograms of document lengths.

(b) Heat map of word frequencies for the 200 most common words across all documents (best viewed in color).

Fig. 3: Document length distributions and word frequencies for each organization in the WITS perpertrator identification experiment.

9 Image Segmentation and Object Recognition

Two problems of enduring interest in the computer vision community are *image segmentation*, dividing an image into its distinct, semantically meaningful regions, and *object recognition*, labeling the regions of images according to their semantic object classes. Solutions to these problems are at the core of applications such as content-based image retrieval, video surveying, and object tracking. Here we will take an admixture modeling approach to jointly recognizing and localizing objects within images (Cao and Li, 2007; Russell et al., 2006; Sivic et al., 2005; Verbeek and Triggs, 2007). Each individual d is an image comprised of N_d image patches (observations), and each patch $\mathbf{x}_{d,n}$ is assumed to be generated by an unknown object class (a latent component of the admixture). Given a series of training images with image patches labeled, the problem of recognizing and localizing objects in a new image reduces to inferring the latent class associated with each new image patch. Since the number of object classes is typically known *a priori*, we will tackle this inferential task with the finite approximation to the HBNBP admixture model given in Section 7.2 and compare its performance with that of a more standard model of admixture, latent Dirichlet allocation (LDA) (Blei et al., 2003).

9.1 Representing an Image Patch

We will represent each image patch as a vector of visual descriptors drawn from multiple modalities. Verbeek and Triggs (2007) suggest three complementary modalities: texture, hue, and location. Here, we introduce a fourth: opponent angle. To describe hue, we use the robust hue descriptor of Van De Weijer and Schmid (2006), which grants invariance to illuminant variations, lighting geometry, and specularities. For texture description we use "dense SIFT" features (Lowe, 2004; Dalal and Triggs, 2005), histograms of oriented gradients computed not at local keypoints but rather at a single scale over each patch. To describe coarse location, we cover each image with a regular c x c grid of cells (for a total of $V^{\text{loc}} = c^2$ cells) and assign each patch the index of the covering cell. The opponent angle descriptor of Van De Weijer and Schmid (2006) captures a second characterization of image patch color. These features are invariant to specularities, illuminant variations, and diffuse lighting conditions.

To build a discrete visual vocabulary from these raw descriptors, we vector quantize the dense SIFT, hue, and opponent angle descriptors using k-means, producing V^{sift}, V^{hue}, and V^{opp} clusters respectively. Finally, we form the observation associated with a patch by concatenating the four modality components into a single vector, $\mathbf{x}_{d,n} = (x^{\text{sift}}_{d,n}, x^{\text{hue}}_{d,n}, x^{\text{loc}}_{d,n}, x^{\text{opp}}_{d,n})$. As in Verbeek and Triggs (2007), we assume that the descriptors from disparate modalities are conditionally independent given the latent object class of the patch. Hence, we define our data generating distribution and our base distribution over parameters $\boldsymbol{\psi}_k = (\psi^{\text{sift}}_k, \psi^{\text{hue}}_k, \psi^{\text{loc}}_k, \psi^{\text{opp}}_k)$ via

$$\begin{aligned} \psi^m_k &\overset{\text{ind}}{\sim} \text{Dirichlet}(\eta \mathbf{1}_{V^m}) && \text{for } m \in \{\text{sift}, \text{hue}, \text{loc}, \text{opp}\} \\ x^m_{d,n} \mid z_{d,n}, \boldsymbol{\psi}_{\cdot} &\overset{\text{ind}}{\sim} \text{Mult}(1, \psi^m_{z_{d,n}}) && \text{for } m \in \{\text{sift}, \text{hue}, \text{loc}, \text{opp}\} \end{aligned}$$

for a hyperparameter $\eta \in \mathbb{R}$ and $\mathbf{1}_{V^m}$ a V^m-dimensional vector of ones.

9.2 Experimental Setup

We use the Microsoft Research Cambridge pixel-wise labeled image database v1 in our experiments.[8] The dataset consists of 240 images, each of size 213 x 320 pixels. Each image has an associated pixel-wise ground truth labeling, with each pixel labeled as belonging to one of 13 semantic classes or to the *void* class. Pixels have a ground truth label of *void* when they do not belong to any semantic class or when they lie on the boundaries

[8] http://research.microsoft.com/vision/cambridge/recognition/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2318721, IEEE Transactions on Pattern Analysis and Machine Intelligence

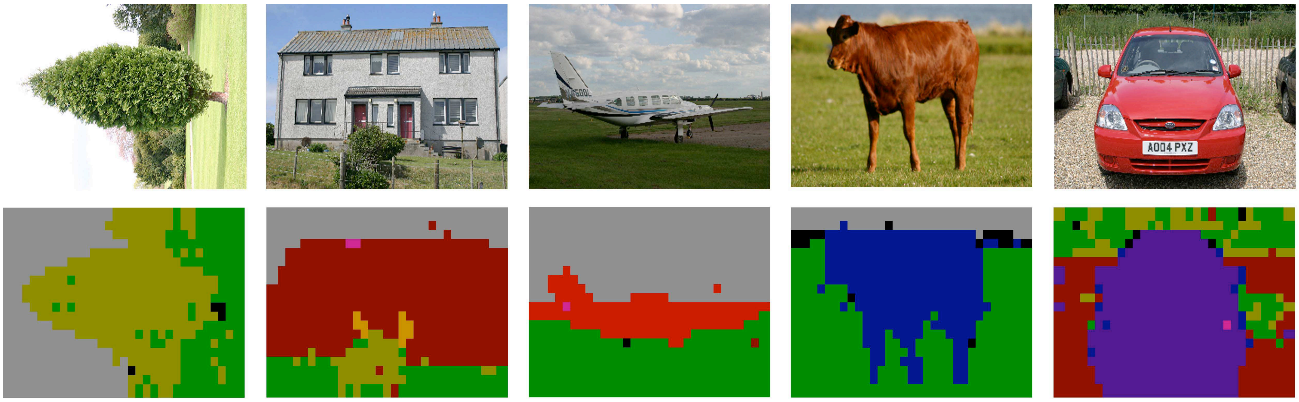


Fig. 4: MSRC-v1 test image segmentations inferred by the HBNBP admixture model (best viewed in color).

between classes in an image. The dataset provider notes that there are insufficiently many instances of *horse*, *mountain*, *sheep*, or *water* to learn these classes, so, as in Verbeek and Triggs (2007), we treat these ground truth labels as *void* as well. Thus, our general task is to learn and segment the remaining nine semantic object classes.

From each image, we extract 20 x 20 pixel patches spaced at 10 pixel intervals across the image. We choose the visual vocabulary sizes $(V^{\text{sift}}, V^{\text{hue}}, V^{\text{loc}}, V^{\text{opp}}) = (1000, 100, 100, 100)$ and fix the hyperparameter $\eta = 0.1$. As in Verbeek and Triggs (2007), we assign each patch a ground truth label $z_{d,n}$ representing the most frequent pixel label within the patch. When performing posterior inference, we divide the dataset into training and test images. We allow the inference algorithm to observe the labels of the training image patches, and we evaluate the algorithm's ability to correctly infer the label associated with each test image patch.

Since the number of object classes is known *a priori*, we employ the HBNBP finite approximation Gibbs sampler of Section 7.2 to conduct posterior inference. We again use the hyperparameters $(\gamma_0, \theta_0, \gamma_d, \theta_d) = (3, 3, 1, 10)$ for all documents d and set r_d according to the heuristic $r_d = N_d(\theta_0 - 1)/(\theta_0\gamma_0)$. We draw 10,000 samples and, for each test patch, predict the label with the highest posterior probability across the samples. We compare HBNBP performance with that of LDA using the standard variational inference algorithm of Blei et al. (2003) and maximum *a posteriori* prediction of patch labels. For each model, we set $K = 10$, allowing for the nine semantic classes plus *void*, and, following Verbeek and Triggs (2007), we ensure that the *void* class remains generic by fixing $\psi_{10}^m = (\frac{1}{V^m}, \cdots, \frac{1}{V^m})$ for each modality m.

9.3 Results

Figure 4 displays sample test image segmentations obtained using the HBNBP admixture model. Each pixel is given the predicted label of its closest patch center. Test patch classification accuracies for the HBNBP admixture model and LDA are reported in Tables 5a and 5b respectively. All results are averaged over twenty randomly generated 90% training / 10% test divisions of the dataset. The two methods perform comparably, with the HBNBP admixture model outperforming LDA in the prediction of every object class save *building*. Indeed, the mean object class accuracy is 0.79 for the HBNBP model versus 0.76 for LDA, showing that the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2318721, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 5: Confusion matrices for patch-level image segmentation and object recognition on the MSRC-v1 database. We report test image patch inference accuracy averaged over twenty randomly generated 90% training / 10% test divisions.

(a) HBNBP Confusion Matrix

		Predicted Class Label								
		building	grass	tree	cow	sky	aeroplane	face	car	bicycle
Actual Class Label	building	0.66	0.01	0.05	0.00	0.03	0.09	0.01	0.03	0.09
	grass	0.00	0.89	0.06	0.02	0.00	0.01	0.00	0.00	0.00
	tree	0.01	0.08	0.75	0.01	0.04	0.03	0.00	0.00	0.07
	cow	0.00	0.10	0.04	0.72	0.00	0.00	0.05	0.01	0.01
	sky	0.04	0.00	0.01	0.00	0.93	0.01	0.00	0.00	0.00
	aeroplane	0.10	0.04	0.01	0.00	0.02	0.81	0.00	0.02	0.00
	face	0.04	0.00	0.01	0.04	0.00	0.00	0.84	0.00	0.00
	car	0.20	0.00	0.01	0.00	0.01	0.01	0.00	0.73	0.02
	bicycle	0.16	0.00	0.04	0.00	0.00	0.00	0.00	0.02	0.73

(b) LDA Confusion Matrix

		Predicted Groups								
		building	grass	tree	cow	sky	aeroplane	face	car	bicycle
Actual Groups	building	0.69	0.01	0.04	0.01	0.03	0.07	0.01	0.03	0.08
	grass	0.00	0.88	0.05	0.02	0.00	0.01	0.00	0.00	0.00
	tree	0.02	0.08	0.75	0.01	0.04	0.02	0.00	0.00	0.05
	cow	0.00	0.10	0.03	0.70	0.00	0.00	0.05	0.01	0.01
	sky	0.05	0.00	0.02	0.00	0.91	0.01	0.00	0.00	0.00
	aeroplane	0.12	0.04	0.01	0.00	0.02	0.75	0.00	0.03	0.00
	face	0.04	0.00	0.01	0.05	0.00	0.00	0.80	0.00	0.00
	car	0.19	0.00	0.01	0.00	0.01	0.01	0.00	0.71	0.03
	bicycle	0.19	0.00	0.04	0.01	0.00	0.00	0.00	0.02	0.68

TABLE 6: Sensitivity of HBNBP admixture model to hyperparameter specification for joint image segmentation and object recognition on the MSRC-v1 database. Each hyperparameter is varied across the specified range while the remaining parameters are held fixed to the default values reported in Section 9.2. We report test patch inference accuracy averaged across object classes and over twenty randomly generated 90% training / 10% test divisions. For each test patch, we predict the label with the highest posterior probability across 2,000 samples.

Hyperparameter	Parameter range	Minimum accuracy	Maximum accuracy
γ_0	$[0.3, 30]$	0.786	0.787
θ_0	$[1.5, 30]$	0.786	0.786
η	$[2 \times 10^{-16}, 1]$	0.778	0.788

HBNBP provides a viable alternative to more classical approaches to admixture.

9.4 Parameter Sensitivity

To test the sensitivity of the HBNBP admixture model to misspecification of the mass, concentration, and likelihood hyperparameters, we measure the fluctuation in test set performance as each hyperparameter deviates from its default value (with the remainder held fixed). The results of this study are summarized in Table 6. We find that the HBNBP model is rather robust to changes in the hyperparameters and maintains nearly constant predictive performance, even as the parameters vary over several orders of magnitude.

10 Conclusions

Motivated by problems of admixture, in which individuals are represented multiple times in multiple latent classes, we introduced the negative binomial process, an infinite-dimensional prior for vectors of counts. We developed new nonparametric admixture models based on the NBP and its conjugate prior, the beta process, and characterized the relationship between the BNBP and preexisting models for admixture. We also analyzed the asymptotics of our new priors, derived MCMC procedures for posterior inference, and demonstrated the effectiveness of our models in the domains of image segmentation and document analysis.

There are many other problem domains in which latent vectors of counts provide a natural modeling framework and where we believe that the HBNBP can prove useful. These include the computer vision task of *multiple object recognition*, where one aims to discover which and how many objects are present in a given image (Titsias, 2008), and the problem of modeling *copy number variation* in genomic regions, where one seeks to infer the underlying events responsible for large repetitions or deletions in segments of DNA (Chen et al., 2011).

Acknowledgments

Support for this project was provided by IARPA under the "Knowledge Discovery and Dissemination" program (IARPA-BAA-09-10) and by ONR under the Multidisciplinary University Research Initiative (MURI) program (N00014-11-1-0688). Tamara Broderick was supported by a National Science Foundation Graduate Research Fellowship. Lester Mackey was supported by the National Defense Science and Engineering Graduate Fellowship.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet allocation," *The Journal of Machine Learning Research*, 3, 993–1022.

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

Cao, L., and Li, F.-F. (2007), Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes,, in *IEEE International Conference on Computer Vision*, pp. 1–8.

Chen, H., Xing, H., and Zhang, N. R. (2011), "Stochastic segmentation models for allele-specific copy number estimation with SNP-array data," *PLoS Computational Biology*, 7, e1001060.

Dalal, N., and Triggs, B. (2005), Histograms of oriented gradients for human detection,, in *Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, pp. 886–893.

Damien, P., Wakefield, J., and Walker, S. (1999), "Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables," *Journal of the Royal Statistical Society Series B*, 61, 331–344.

Erosheva, E. A., and Fienberg, S. E. (2005), Bayesian mixed membership models for soft clustering and classification,, in *Classification–The Ubiquitous Challenge*, Springer, New York, pp. 11–26.

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *Annals of Statistics*, 1(2), 209–230.

Fox, E., Sudderth, E., Jordan, M., and Willsky, A. (2009), "Sharing features among dynamical systems with beta processes," *Advances in Neural Information Processing Systems*, 22, 549–557.

Fraley, C., and Raftery, A. E. (2002), "Model-based clustering, discriminant analysis and density estimation," *Journal of the American Statistical Association*, 97, 611–631.

Geman, S., and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Pattern Analysis and Machine Intelligence*, 6, 721–741.

Gnedin, A., Hansen, B., and Pitman, J. (2007), "Notes on the occupancy problem with infinitely many boxes: General asymptotics and power laws," *Probability Surveys*, 4, 146–171.

Goldwater, S., Griffiths, T., and Johnson, M. (2006), "Interpolating between types and tokens by estimating power-law generators," in *Advances in Neural Information Processing Systems 18*, eds. Y. Weiss, B. Schölkopf, and J. Platt, Vol. 18, Cambridge, MA: MIT Press, pp. 459–466.

Griffiths, T., and Ghahramani, Z. (2006), "Infinite latent feature models and the Indian buffet process," in *Advances in Neural Information Processing Systems 18*, eds. Y. Weiss, B. Schölkopf, and J. Platt, Cambridge, MA: MIT Press, pp. 475–482.

Griffiths, T. L., and Steyvers, M. (2004), "Finding scientific topics," *Proceedings of the National Academy of Sciences*, 101(Suppl. 1), 5228–5235.

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

Kalli, M., Griffin, J. E., and Walker, S. G. (2011), "Slice sampling mixture models," *Statistics and Computing*, 21, 93–105.

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

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

Kingman, J. F. C. (1993), *Poisson Processes*, New York: Oxford University Press.

Korwar, R. M., and Hollander, M. (1973), "Contributions to the theory of Dirichlet processes," *The Annals of Probability*, pp. 705–711.

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

Lowe, D. G. (2004), "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 60(2), 91–110.

MacEachern, S. N., and Müller, P. (1998), "Estimating mixture of Dirichlet process models," *Journal of Computational and Graphical Statistics*, 7, 223–238.

McCloskey, J. W. (1965), A model for the distribution of individuals by species in an environment, PhD thesis, Michigan State University.

McLachlan, G., and Basford, K. (1988), *Mixture Models: Inference and Applications to Clustering*, New York: Dekker.

Mitzenmacher, M. (2004), "A brief history of generative models for power law and lognormal distributions," *Internet mathematics*, 1(2), 226–251.

Neal, R. M. (2000), "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, 9(2), 249–265.

Neal, R. M. (2003), "Slice sampling," *Annals of Statistics*, 31(3), 705–741.

Newman, M. (2005), "Power laws, Pareto distributions and Zipf's law," *Contemporary physics*, 46(5), 323–351.

Paisley, J., Zaas, A., Woods, C. W., Ginsburg, G. S., and Carin, L. (2010), A stick-breaking construction of the beta process,, in *International Conference on Machine Learning*, Haifa, Israel.

Papaspiliopoulos, O. (2008), A note on posterior sampling from Dirichlet mixture models,, Technical Report 8, University of Warwick, Centre for Research in Statistical Methodology.

Pitman, J. (2006), *Combinatorial Stochastic Processes*, Vol. 1875 of *Lecture Notes in Mathematics*, Berlin: Springer-Verlag.

Pitman, J., and Yor, M. (1997), "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator," *Annals of Probability*, 25, 855–900.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000), "Inference of population structure using multilocus genotype data," *Genetics*, 155(2), 945–959.

Qi, F., and Losonczi, L. (2010), "Bounds for the ratio of two gamma functions," *Journal of Inequalities and Applications*,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2318721, IEEE Transactions on Pattern Analysis and Machine Intelligence

2010, 204.

Russell, B. C., Freeman, W. T., Efros, A. A., Sivic, J., and Zisserman, A. (2006), Using multiple segmentations to discover objects and their extent in image collections,, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1605–1614.

Sivic, J., Russell, B., Efros, A., Zisserman, A., and Freeman, W. (2005), Discovering object categories in image collections,, Technical Report AIM-2005-005, MIT.

Teh, Y. W. (2006), A hierarchical Bayesian language model based on Pitman-Yor processes,, in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 985–992.

Teh, Y. W., and Görür, D. (2009), Indian buffet processes with power-law behavior,, in *Advances in Neural Information Processing Systems 22*, eds. Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, pp. 1838–1846.

Teh, Y. W., Görür, D., and Ghahramani, Z. (2007), Stick-breaking construction for the Indian buffet process,, in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, Vol. 11.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006), "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, 101(476), 1566–1581.

Thibaux, R. J. (2008), Nonparametric Bayesian models for machine learning, PhD thesis, EECS Department, University of California, Berkeley.

Thibaux, R., and Jordan, M. (2007), Hierarchical beta processes and the Indian buffet process,, in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, Vol. 11.

Titsias, M. (2008), "The infinite gamma-Poisson feature model," in *Advances in Neural Information Processing Systems 20*, eds. J. C. Platt, D. Koller, Y. Singer, and S. Roweis, Cambridge, MA: MIT Press, pp. 1513–1520.

Tricomi, F. G., and Erdélyi, A. (1951), "The asymptotic expansion of a ratio of gamma functions," *Pacific Journal of Mathematics*, 1(1), 133–142.

Van De Weijer, J., and Schmid, C. (2006), Coloring local feature extraction,, in *Proceedings of the European Conference on Computer Vision*, pp. 334–348.

Verbeek, J. J., and Triggs, B. (2007), Region classification with Markov field aspect models,, in *Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society.

Walker, S. G. (2007), "Sampling the Dirichlet mixture model with slices," *Communications in Statistics–Simulation and Computation*, 36, 45–54.

Watterson, G. A. (1974), "The sampling theory of selectively neutral alleles," *Advances in applied probability*, pp. 463–488.

West, M. (1992), Hyperparameter estimation in Dirichlet process mixture models,, Technical Report 92-A03, Institute of Statistics and Decision Sciences Discussion Paper.

Wood, F., Archambeau, C., Gasthaus, J., James, L., and Teh, Y. W. (2009), A stochastic memoizer for sequence data,, in *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, pp. 1129–1136.

Zhou, M., Hannah, L., Dunson, D., and Carin, L. (2012), Beta-negative binomial process and Poisson factor analysis,, in *International Conference on Artificial Intelligence and Statistics*.

Tamara Broderick is a Ph.D. candidate in statistics at the University of California, Berkeley. She received an AB in mathematics from Princeton University in 2007. On a Marshall Scholarship at the University of Cambridge, she received an Master of Advanced Study for completion of Part III of the Mathematical Tripos (2008) and an MPhil by research in physics (2009). Her research interests focus on developing and analyzing models for unsupervised learning using Bayesian nonparametrics.

Lester Mackey earned his BSE in Computer Science from Princeton University and his MA in Statistics and PhD in Computer Science from the University of California, Berkeley. He was a Simons Math+X postdoctoral fellow at Stanford University and is currently an assistant professor of Statistics at Stanford University. His recent work focuses on the development and analysis of algorithms for ranking, matrix factorization, compressed sensing, and admixture modeling.

John Paisley received the B.S.E. (2004), M.S. (2007) and Ph.D. (2010) in Electrical & Computer Engineering from Duke University. He was then a postdoctoral researcher in the Computer Science Department at Princeton University and in the Department of EECS at UC Berkeley. He is currently an assistant professor in the Department of Electrical Engineering at Columbia University. He works on developing Bayesian models for machine learning applications, particularly for dictionary learning and topic modeling.

Michael I. Jordan is the Pehong Chen Distinguished Professor in the Department of Electrical Engineering and Computer Science and the Department of Statistics at the University of California, Berkeley. His research in recent years has focused on Bayesian nonparametric analysis, probabilistic graphical models, spectral methods, kernel machines and applications to problems in statistical genetics, signal processing, computational biology, information retrieval and natural language processing. Prof. Jordan is a member of the National Academy of Sciences, a member of the National Academy of Engineering and a member of the American Academy of Arts and Sciences. He is a Fellow of the American Association for the Advancement of Science. He has been named a Neyman Lecturer and a Medallion Lecturer by the Institute of Mathematical Statistics. He is an Elected Member of the International Institute of Statistics. He is a Fellow of the AAAI, ACM, ASA, CSS, IMS, IEEE and SIAM.