Distance Dependent Infinite Latent Feature Models

Samuel J. Gershman, Peter I. Frazier, and David M. Blei

Abstract—Latent feature models are widely used to decompose data into a small number of components. Bayesian nonparametric variants of these models, which use the Indian buffet process (IBP) as a prior over latent features, allow the number of features to be determined from the data. We present a generalization of the IBP, the *distance dependent Indian buffet process* (dd-IBP), for modeling non-exchangeable data. It relies on distances defined between data points, biasing nearby data to share more features. The choice of distance measure allows for many kinds of dependencies, including temporal and spatial. Further, the original IBP is a special case of the dd-IBP. We develop the dd-IBP and theoretically characterize its feature-sharing properties. We derive a Markov chain Monte Carlo sampler for a linear Gaussian model with a dd-IBP prior and study its performance on real-world non-exchangeable data.

Index Terms—Bayesian nonparametrics, dimensionality reduction, matrix factorization, Indian buffet process, distance functions

1 Introduction

Many natural phenomena decompose into latent features. For example, visual scenes can be decomposed into objects; genetic regulatory networks can be decomposed into transcription factors; music can be decomposed into spectral components. In these examples, multiple latent features can be simultaneously active, and each can influence the observed data. Dimensionality reduction methods, such as principal component analysis, factor analysis, and probabilistic matrix factorization, provide a statistical approach to inferring latent features [3]. These methods characterize a small set of features (or dimensions) and model each data point as a weighted combination of them. Dimensionality reduction can improve predictions and identify hidden structures in observed data.

Dimensionality reduction methods typically require that the number of latent features be fixed in advance. Researchers have recently proposed a more flexible approach based on Bayesian nonparametric models, where the number of features is inferred from the data through a posterior distribution. These models are usually based on the Indian buffet process (IBP; [16], [17]), a prior over binary matrices with a finite number of rows (corresponding to data points) and an infinite number of columns (corresponding to latent features). Using the IBP as a building block, Bayesian nonparametric latent feature models have been applied to several statistical problems (e.g., [18], [19], [21], [22]). Since the number of features is effectively unbounded, these models are sometimes known as "infinite" latent feature models.

The IBP assumes that data are *exchangeable*: permuting the order of rows leaves the probability of an allocation of latent features unchanged. This assumption may be appropriate for some data sets, but for many others we expect dependencies between data points and, consequently, between their latent representations. As examples, the latent features describing human motion are autocorrelated over time; the latent features describing environmental risk factors are autocorrelated over space. In this paper, we present a generalization of the IBP—the *distance dependent IBP* (dd-IBP)—that addresses this limitation. The dd-IBP allows infinite latent feature models to capture non-exchangeable structure.

The problem of adapting nonparametric models to non-exchangeable data has been studied extensively in the mixture-modeling literature. In particular, variants of the Dirichlet process mixture model allow dependencies between data points (e.g., [1], [6], [8], [11], [15], [23]). These dependencies may be spatial, temporal or more generally covariate-dependent; the effect of such dependencies is to induce sharing of latent features between nearby data points.

Among these methods is the *distance dependent Chinese restaurant process* (dd-CRP; [5]). The dd-CRP is a non-exchangeable generalization of the Chinese restaurant process (CRP), the prior over partitions of data that emerges in Bayesian nonparametric mixture modeling [4], [13], [24]. The dd-CRP models non-exchangeability by using distances between data points—nearby data points (e.g., in time or space) are more likely to be assigned to the same mixture component. The dd-IBP extends these ideas to infinite latent feature models, where distances between data points influence feature-sharing, and nearby data points are more likely to share latent features.

We review the IBP in Section 2.1 and develop the dd-IBP in Section 2.2. Like the dd-CRP, the dd-IBP lacks *marginal invariance*, which means that removing one observation changes the distribution over the other observations. We

- *S.J. Gershman is with the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA. E-mail: sjgershm@mit.edu.*
- *P.I. Frazier is with the School of Operations Research and Information Engineering, Cornell University, Ithaca, NY. E-mail: pf98@cornell.edu.*
- *D.M. Blei is with the Department of Computer Science, Princeton University, 35 Olden Street, Princeton, NJ. E-mail: blei@cs.princeton.edu.*

Manuscript received 3 Sept. 2012; revised 17 Mar. 2014; accepted 13 Apr. 2014. Date of publication 30 Apr. 2014; date of current version 14 Jan. 2015. Recommended for acceptance by R.P. Adams, E. Fox, E. Sudderth, and Y.W. Teh. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TPAMI.2014.2321387

discuss this property further in Section 2.3. Although many Bayesian nonparametric models have this property, it is a particular modeling choice that may be appropriate for some problems but not for others.

Several other infinite latent feature models have been developed to capture dependencies between data in different ways, for example using phylogenetic trees [20] or latent Gaussian processes [28]. Of particular relevance to this work is the dependent hierarchical Beta process (dHBP; [30]), which uses a hierarchical beta process (bp) to couple data. These and other related models are discussed further in Section 3. In Section 4, we characterize the feature-sharing properties of the dd-IBP and compare it to those of the dHBP [30]. We find that the different models capture qualitatively distinct dependency structures.

Exact posterior inference in the dd-IBP is intractable. We present an approximate inference algorithm based on Markov chain Monte Carlo (MCMC; [26]) in Section 5, and we apply this algorithm in Section 6 to infer the latent features in a linear-Gaussian model. We then use the dd-IBP to model human brain imaging data in which age is predictive of dependencies between data from different individuals (Section 7). We show that the dd-IBP induces features that are useful for a supervised classification task.

2 The Distance Dependent Indian Buffet Process

We first review the definition of the IBP and its role in defining infinite latent feature models. We then introduce the dd-IBP.

2.1 The Indian Buffet Process

The IBP is a prior over binary matrices $\mathbf{Z}$ with an infinite number of columns [16], [17]. In the Indian buffet metaphor, rows of $\mathbf{Z}$ correspond to customers and columns correspond to dishes (see Fig. 1). In data analysis, the customers represent data points and the dishes represent features. Let z_{ik} denote the entry of $\mathbf{Z}$ at row i and column k. Whether customer i has decided to sample dish k (that is, whether $z_{ik} = 1$) corresponds to whether data point i possesses feature k.

The IBP is defined as a sequential process. The first customer enters the restaurant and samples the first $\lambda_1 \sim \text{Poisson}(\alpha)$ number of dishes, where the hyperparameter α is a scalar. In the binary matrix, this corresponds to the first row being a contiguous block of ones, whose length is the number of dishes sampled (λ_1), followed by an infinite block of zeros.

Subsequent customers $i = 2, \ldots, N$ enter, sampling each previously sampled dish according to its popularity,

$$p(z_{ik} = 1 \mid \mathbf{z}_{1:(i-1)}) = m_{ik}/i, \tag{1}$$

where $m_{ik} = \sum_{j<i} z_{jk}$ is the number of customers that sampled dish k prior to customer i. (We emphasize that Eq. (1) applies only to dishes k that were previously sampled, i.e., for which $m_{ik} > 0$.) Then, each customer samples $\lambda_i \sim \text{Poisson}(\alpha/i)$ new dishes. Again these are represented as a contiguous block of ones in the columns beyond the last dish sampled by a previous customer.

Though described sequentially, Griffiths and Ghahramani [16] showed that the IBP defines an allocation of dishes to customers that is *exchangeable*. This means that the order of the customers does not affect the probability of the resulting allocation of dishes to customers.

To state this notion formally, define $lof(\mathbf{Z})$ to be the *left-ordered* binary matrix obtained from $\mathbf{Z}$ by sorting its columns in decreasing order, interpreting each column as an integer represented in binary with highest bit at row 1. Then define $[\mathbf{Z}]$ to be the set of $N \times \infty$ binary matrices $\mathbf{Z}'$ with the property that $lof(\mathbf{Z}') = lof(\mathbf{Z})$. This is an equivalence class of binary matrices that imply the same allocations of features to customers, differing only in how these features are labeled. Exchangeability of the IBP's allocation of dishes to customers means that the distribution of the random equivalence class $[\mathbf{Z}]$ is the same as that of $[\mathbf{Z}^\pi]$, where $\mathbf{Z}^\pi$ is obtained by permuting the rows of $\mathbf{Z}$. In the next section, we develop a generalization of the IBP that relaxes this assumption.

2.2 The Distance Dependent Indian Buffet Process

Like the IBP, the dd-IBP is a distribution over binary latent feature matrices with a finite number of rows and an infinite number of columns. Each pair of customers has an associated distance, e.g., distance in time or space, or based on a covariate. Two customers that are close together in this distance will be more likely to share the same dishes (that is, features) than two customers that are far apart.

The dd-IBP can be understood in terms of the following sequential construction. First, each customer selects a Poisson-distributed number of dishes (feature columns). The dishes selected by a customer in this phase of the construction are said to be "owned" by this customer. A dish is either unowned, or is owned by exactly one customer. This step is akin to the selection of new dishes in the IBP.

Then, for each owned dish, customers connect to one another. The probability that one customer connects to another decreases with the distance between them. Note that customers do not sample each dish, as in the IBP, but rather connect to other customers. Thus, each dish is associated with a graph of connections between customers.

These per-dish graphs of customers determine dish inheritance: A customer inherits a dish if its owner (from the first step) is reachable in the connectivity graph for that dish. This inheritance is computed deterministically from the connections generated in the previous step.[1] The dishes that each customer samples (i.e., the active features) are those that he inherits or owns. Thus, distance-dependent connection probabilities induce similarity of sampled dishes between nearby customers.

An example of customer assignments sampled from the dd-IBP is shown in Fig. 2. In this example, customer 1 owns dish 1; customers 2-4 all reach customer 1 for dish 1, either directly or through a chain, and thereby inherit the dish (indicated by gray shading). Consequently, feature 1 is active for customers 1-4. Dish 2 is owned by customer 2; only customer 1 reaches customer 2 for dish 2, and hence

1. If one insists upon a complete gastronomical metaphor, customer connectivity can be thought of as "I'll have what he's having."

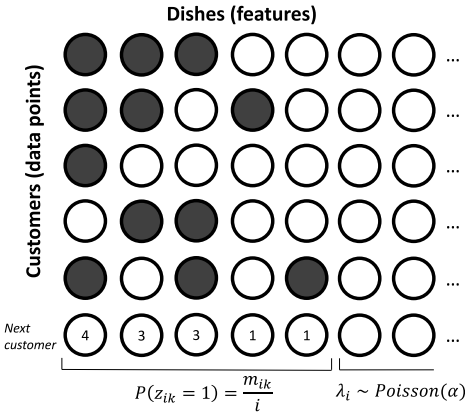

Fig. 1. Schematic of the IBP. An example of a latent feature matrix ($\mathbf{Z}$) generated by the IBP. Rows correspond to customers (data points) and columns correspond to dishes (features). Gray shading indicates that a feature is active for a given data point. The last row illustrates the assignment process for a new customer; the counts for each feature (m_{ik}) are shown inside the circles for previously sampled features.

feature 2 is active for customers 1 and 2. Dish 3 is owned by customer 2, but no other customers reach customer 2 for dish 3, and hence feature 3 is active only for that customer.

We now more formally describe the probabilistic generative process of the binary matrix $\mathbf{Z}$. First, we introduce some notation and terminology.

- Dishes (columns of $\mathbf{Z}$) are identified with the natural numbers $\mathbb{N} = \{1, 2, \ldots\}$. The number of dishes owned by customer i is λ_i, and dishes are labeled in order, so that the set of dishes owned by this customer is $\mathcal{K}_i = (\sum_{j<i} \lambda_j, \sum_{j \leq i} \lambda_i]$. The total number of owned dishes is $K = \sum_{i=1}^{N} \lambda_i$. The set of dishes owned by customers excluding i is $\mathcal{K}_{-i} = \cup_{j \neq i} \mathcal{K}_j$.
- Each dish is associated with a set of customer-to-customer assignments, specified by the $N \times K$ *connectivity matrix* $\mathbf{C}$, where $c_{ik} = j$ indicates that customer i connects to customer j for dish k. Given $\mathbf{C}$, the customers form a set of (possibly cyclic) directed graphs, one for each dish. The *ownership vector* is $\mathbf{c}^*$ and has length K, where $c_k^* \in \{1, \ldots, N\}$ indicates the customer who owns dish k, so $c_k^* = i \Longleftrightarrow k \in \mathcal{K}_i$.
- The $N \times N$ distance matrix between customers is $\mathbf{D}$, where the distance between customers i and j is d_{ij}. A customer's self-distance is 0: $d_{ii} = 0$. We call the distance matrix *sequential* when $d_{ij} = \infty$ for $j > i$. In this special case, customers can only connect to previous customers.
- The *decay function* $f : \mathbb{R} \mapsto [0, 1]$ maps distance to a quantity, which we call *proximity*, that controls the probabilities of customer links. We require that $f(0) = 1$ and $f(\infty) = 0$. We obtain the *normalized proximity matrix* $\mathbf{A}$ by applying the decay function to each customer pair and normalizing by customer. That is, $a_{ij} = f(d_{ij})/h_i$, where $h_i = \sum_{j=1}^{N} f(d_{ij})$.

Using this notation, we generate the feature indicator matrix $\mathbf{Z}$ as follows:

1. *Assign dish ownership*. For each customer i, let $\lambda_i \sim \text{Poisson}(\alpha/h_i)$, and set $\mathcal{K}_i = (\sum_{j<i} \lambda_j, \sum_{j \leq i} \lambda_i]$, thus allocating λ_i dishes to this customer. For each $k \in \mathcal{K}_i$, set the ownership $c_k^* = i$.
2. *Assign customer connections*. For each customer i and dish $k \in \{1, \ldots, K\}$, draw a customer assignment according to $P(c_{ik} = j \mid \mathbf{D}, f) = a_{ij}, j = 1, \ldots, N$. Note that customers can connect to themselves. In this case, they do not inherit a dish unless they own it (see the next step).
3. *Compute dish inheritance*. We say that customer j *inherits* dish k if there exists a path along the directed graph for dish k from customer j to the dish's owner c_k^* (i.e., if c_k^* is reachable from j), where the directed graph is defined by column k of $\mathbf{C}$. The owner of a dish automatically inherits it.[2] We encode reachability with $\mathcal{L}$. If customer j is reachable from customer i for dish k then $\mathcal{L}_{ijk} = 1$. Otherwise $\mathcal{L}_{ijk} = 0$.
4. *Compute the feature indicator matrix*. For each customer i and dish $k \leq K$ we set $z_{ik} = 1$ if i inherits k, otherwise $z_{ik} = 0$. For $k > K$ we set $z_{ik} = 0$.

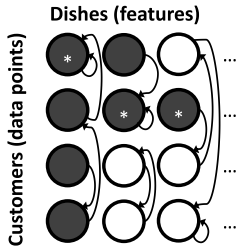

Fig. 2. Schematic of the dd-IBP. An example of a latent feature matrix generated by the dd-IBP. Rows correspond to customers (data points) and columns correspond to dishes (features). Customers connect to each other, as indicated by arrows. Customers inherit a dish if the owner of that dish (c_k^*, indicated by stars) is reachable by a sequence of connections. Gray shading indicates that a feature is active for a given data point.

The generative process of the dd-IBP defines the following joint distribution of the ownership vector and connectivity matrix,

$$P(\mathbf{C}, \mathbf{c}^* \mid \mathbf{D}, \alpha, f) = P(\mathbf{c}^* \mid \alpha) P(\mathbf{C} \mid \mathbf{c}^*, \mathbf{D}, f). \tag{2}$$

Consider the first term. The probability of the ownership vector is

$$P(\mathbf{c}^* \mid \alpha) = \prod_{i=1}^{N} P(\lambda_i \mid \alpha), \tag{3}$$

where $\mathbf{c}^*$ is a deterministic function of $\lambda_1, \ldots, \lambda_N$.

Consider the second term. The conditional distribution of the connectivity matrix $\mathbf{C}$ depends on the total number of owned dishes K and the normalized proximity matrix $\mathbf{A}$ (derived from the distances and decay function),

$$P(\mathbf{C} \mid \mathbf{c}^*, \mathbf{D}, f) = \prod_{i=1}^{N} \prod_{k=1}^{K} a_{i c_{ik}}. \tag{4}$$

The dependence on $\mathbf{c}^*$ comes from K, which is determined by the ownership vector $\mathbf{c}^*$.

2. Although customer i can link to other customers for dish k even if $k \in \mathcal{K}_i$, these connections are ignored in determining dish inheritance when $k \in \mathcal{K}_i$.

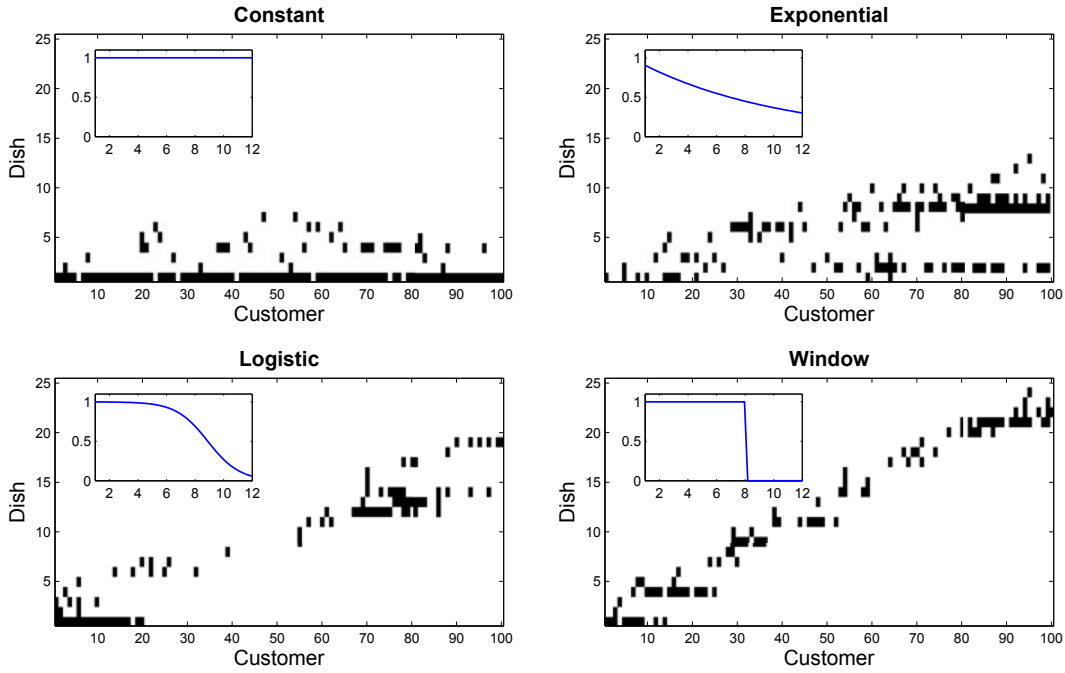

Fig. 3. Decay functions. Each panel presents a different latent feature matrix, sampled from the dd-IBP with sequential distances. Decay functions are shown in the insets.

Random feature models (and the traditional IBP) operate with a random binary matrix $\mathbf{Z}$, and the feature allocation that it induces. In the dd-IBP, $\mathbf{Z}$ is a (deterministic) many-to-one function of the random variables $\mathbf{C}$ and $\mathbf{c}^*$, which we denote by ϕ. We compute the probability of a binary matrix by marginalizing out the appropriate configurations of these variables

$$P(\mathbf{Z} \mid \mathbf{D}, \alpha, f) = \sum_{(\mathbf{c}^*, \mathbf{C}):\phi(\mathbf{c}^*, \mathbf{C})=\mathbf{Z}} P(\mathbf{c}^*, \mathbf{C} \mid \mathbf{D}, \alpha, f). \quad (5)$$

The dd-IBP reduces to the standard IBP in the special case when $f(d) = 1$ for all $d < \infty$ and the distance matrix is sequential. (Recall: $\mathbf{D}$ is sequential if $d_{ij} = \infty$ for $j > i$.) To see this, consider the probability that the kth dish is sampled by the ith customer (that is, $z_{ik} = 1$), conditioned the dish having been sampled by a previous customer (which occurs iff $c_k^* < i$ when $\mathbf{D}$ is sequential). This probability is the proportion of previous customers that already reach c_k^* because the probability of connecting to each customer is proportional to one. This probability is m_{ki}/i, which is the same as in the IBP. This is akin to the relationship between the dd-CRP and the traditional CRP under the same condition [5].

Many different decay functions are possible within this framework. Fig. 3 shows samples of $\mathbf{Z}$ using four decay functions and a sequential distance defined by absolute temporal distance ($d_{ij} = i - j$ for $i \geq j$ and $d_{ij} = \infty$ for $j > i$).

- The *constant*, $f(d) = 1$ if $d < \infty$ and $f(\infty) = 0$. This is the standard IBP.
- The *exponential*, $f(d) = \exp(-\beta d)$.
- The *logistic*, $f(d) = 1/(1 + \exp(\beta d - \nu))$.
- The *window*, $f(d) = \mathbf{1}[d < \nu]$.

Each decay function encourages the sharing of features across nearby rows in a different way.

When combined with an observation model, which specifies how the latent features give rise to observed data, the dd-IBP functions as a prior over latent feature representations of a data set. In Section 6, we consider a specific example of how the dd-IBP can be used to analyze data.

2.3 Marginal Invariance and Exchangeability

We now examine some of the theoretical differences between the dd-IBP and the IBP. Unlike the traditional IBP, the dd-IBP is not (in general) *marginally invariant*, the property that removing a customer leaves the distribution over latent features for the remaining customers unchanged. (The dd-IBP builds on the dd-CRP, which is not marginally invariant either.) In some circumstances, marginal invariance is desirable for computational reasons. For example, the conditional distributions over missing data for models lacking marginal invariance require computing ratios of normalization constants. In contrast, marginally invariant models, due to their factorized structure, require less computation for conditional distributions over missing data. In other circumstances, such as exploratory analysis of fully observed data sets, this computational concern is less important.

Beyond computational considerations, in some situations marginal invariance is an inappropriate assumption. For example, imagine you are studying the spread of disease. You believe that there are latent features which play a causal role in disease transmission, such as certain genes in the case of genetically transmitted diseases. In this case, observations correspond to individuals, and distance is naturally defined over a family tree. Removing an individual from a family tree can dramatically alter the distribution

over genes for other individuals in the same family tree. This is an example where marginal invariance is an incorrect assumption. In contrast, the distance-dependent effect of genes on individuals can be captured by an appropriate distance function using the dd-IBP.

Also unlike the traditional IBP, the dd-IBP's feature allocations are not exchangeable in general (except in some special cases). To state this formally, let $\mathbf{Z}$ be drawn from the dd-IBP with distance matrix $\mathbf{D}$, mass parameter α and decay function f. Then, let π be a permutation of the integers $\{1, \ldots, N\}$, and let $\mathbf{Z}^\pi$ be the matrix created by permuting the rows of $\mathbf{Z}$ according to π. Then, except in certain special cases (such as when $\mathbf{D}$ recovers the traditional IBP),

$$P([\mathbf{Z}] = [\mathbf{Z}'] \,|\, \mathbf{D}, \alpha, f) \neq P([\mathbf{Z}^\pi] = [\mathbf{Z}'] \,|\, \mathbf{D}, \alpha, f),$$

where $\mathbf{Z}'$ is a (non-random) $N \times \infty$ binary matrix. Permuting the data changes its distribution, and so the dd-IBP's feature allocations are not exchangeable in general. This property stems from the distance function, which induces sequential dependencies between data points.

Although the dd-IBP does not induce an exchangeable distribution over equivalence classes $[\mathbf{Z}]$, it does have a related symmetry. Let $\mathbf{D}^\pi$ be the $N \times N$ matrix $\mathbf{D}$ with both its rows and its columns permuted according to π. (We retain the same values for α and f). Then, in general,

$$P([\mathbf{Z}^\pi] = [\mathbf{Z}'] \mid \mathbf{D}, \alpha, f) = P([\mathbf{Z}] = [\mathbf{Z}'] \mid \mathbf{D}^\pi, \alpha, f).$$

Thus, if we permute both the data and the distance matrix, probabilities remain unchanged. Permuting both the data and the distance matrix is like first relabeling the data, and then explicitly altering the probability distribution to account for this relabeling. If the dd-IBP's feature allocations were exchangeable, one would not need to alter the probability distribution to account for relabeling.

3 Related Work

In this section we describe related work on infinite latent feature models that capture external dependence between the data. We focus on the most closely related model, which is the *dependent hierarchical beta process* [30]. As a prelude to describing the dHBP, we review the connection between the IBP and the beta process.

3.1 The Beta Process

Recall that the IBP induces an exchangeable distribution over feature allocations. Thibaux and Jordan [27] showed that the de Finetti mixing distribution underlying the IBP is the *beta process*, parameterized by a positive *concentration parameter* c and a *base measure* B_0 on Ω. A draw $B \sim \mathrm{BP}(c, B_0)$ is defined by a countably infinite collection of weighted atoms,

$$B = \sum_{k=1}^{\infty} p_k \delta_{\omega_k}, \tag{6}$$

where δ_ω is a probability distribution that places a single atom at $\omega \in \Omega$, and the $p_k \in [0, 1]$ are independent random variables whose distribution is described as follows. If B_0 is continuous, then the atoms and their weights are drawn from a nonhomogeneous Poisson process defined on the space $\Omega \times [0, 1]$ with rate measure

$$\nu(d\omega, dp) = cp^{-1}(1-p)^{c-1} dp B_0(d\omega). \tag{7}$$

If B_0 is discrete and of the form $B_0 = \sum_{k=1}^{\infty} q_k \delta_{\omega_k}$, $q_k \in [0, 1]$, then B has atoms at the same locations as B_0, with $p_k \sim \mathrm{Beta}(cq_k, c(1-q_k))$. Following Thibaux and Jordan [27], we define the *mass parameter* as $\gamma = B_0(\Omega)$. Note that B_0 is not necessarily a probability measure, and hence γ can take on non-negative values different from 1.

Conditional on a draw from the beta process, the feature representation X_i of data point i is generated by drawing from the *Bernoulli process* (BeP) with base measure B: $X_i \mid B \sim \mathrm{BeP}(B)$. If B is discrete, then $X_i = \sum_{k=1}^{\infty} z_{ik} \delta_{\omega_k}$, where $z_{ik} \sim \mathrm{Bernoulli}(p_k)$. In other words, feature k is activated with probability p_k independently for all data points. Sampling $\mathbf{Z}$ from the compound beta-Bernoulli process is equivalent (in the sense of providing the same distribution over equivalence classes $[\mathbf{Z}]$) to sampling $\mathbf{Z}$ directly from the IBP when $c = 1$ and $\gamma = \alpha$ [27].

3.2 Dependent Hierarchical Beta Processes

The dHBP [30] builds external dependence between data points using the above BP construction. The dependencies are induced by mixing independent BP random measures, weighted by their proximities $\mathbf{A}$.

The dHBP is based on the following generative process,

$$\begin{aligned} X_i \mid B^*_{g_i} &\sim \mathrm{BeP}(B^*_{g_i}), \quad g_i \sim \mathrm{Multinomial}(\mathbf{a}_i), \\ B^*_j \mid B &\sim \mathrm{BP}(c_1, B), \quad B \sim \mathrm{BP}(c_0, B_0). \end{aligned} \tag{8}$$

This is equivalent to drawing X_i from a Bernoulli process whose base measure is a linear combination of BP random measures,

$$X_i \,|\, B_i \sim \mathrm{BeP}(B_i), \quad B_i = \sum_{j=1}^{N} a_{ij} B^*_j. \tag{9}$$

Dependencies between data points are captured in the dHBP by the proximity matrix $\mathbf{A}$, as in the dd-IBP.[3] This allows proximal data points (e.g., in time or space) to share more latent features than distant ones.

In Section 4, we compare the feature-sharing properties of the dHBP and dd-IBP. Using an asymptotic analysis, we show that the dd-IBP offers more flexibility in modeling the proportion of features shared between data points, but less flexibility in modeling uncertainty about these proportions.

3.3 Other Non-Exchangeable Variants

Several other non-exchangeable priors for infinite latent feature models have been proposed (see [14] for a comprehensive review). Williamson et al. [28] used a hierarchical Gaussian process to couple the latent features of data in a covariate-dependent manner. They named this model the *dependent Indian buffet process* (dIBP). Their framework is flexible: It can couple columns of $\mathbf{Z}$ in addition to rows,

3. Zhou et al. [30] formalize dependencies in an equivalent manner using a normalized kernel function defined over pairs of covariates associated with the data points.

while the dd-IBP cannot. However, this flexibility comes at a computational cost during inference: Their algorithm requires sampling an extra layer of variables.

Miller et al. [20] proposed a "phylogenetic IBP" that encodes tree-structured dependencies between data. Doshi-Velez and Ghahramani [10] proposed a "correlated IBP" that couples data points and features through a set of latent clusters. Both of these models relax exchangeability, but they do not allow dependencies to be specified directly in terms of distances between data. Furthermore, inference for these models requires more intensive computation than does the standard IBP. The MCMC algorithm presented by Miller et al. [20] for the phylogenetic IBP involves both dynamic programming and auxiliary variable sampling. Similarly, the MCMC algorithm for the correlated IBP involves sampling latent clusters in addition to latent features. Our model also incurs extra computational cost relative to the traditional IBP due to the computation of reachability (quadratic in the number of observations); however, it permits a richer specification of the dependency structure between observations than either the phylogenetic or the correlated IBP.

Recently, Ren et al. [25] presented a novel way of introducing dependency into latent feature models based on the beta process. Instead of defining distances between customers, each dish is associated with a latent covariate vector, and distances are defined between each customer's (observed) covariates and the dish-specific covariates. Customers then choose dishes with probability proportional to the customer-dish proximity. This construction comes with a significant computational advantage for data sets where the time complexity is tied predominantly to the number of observations. The downside of this construction is that the MCMC algorithm used for inference must sample a separate covariate vector for each dish, which may scale poorly if the covariate dimensionality is high.

4 Characterizing Feature-Sharing

In this section, we compare the feature-sharing properties of the dHBP and dd-IBP. Two data points share a feature if that feature is active for both (i.e., $z_{ik} = z_{jk} = 1$ for $i \neq j$ and a given feature k). This analysis is useful for understanding the types of dependencies induced by the different models, and can help guide the choice of model and hyperparameter settings for particular data analysis problems. We consider an asymptotic regime in which the mass parameter is large (α for the dd-IBP and γ for the dHBP), which simplifies feature-sharing properties. Proofs of all propositions in this section are contained in the Supplementary Materials, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2014.2321387.

4.1 Feature-Sharing in the dd-IBP

We first characterize the limiting distributional properties of feature-sharing in the dd-IBP as $\alpha \to \infty$. We drop the feature index k in the reachability indicator $\mathcal{L}_{ijk}$, writing it $\mathcal{L}_{ij}$. We do this because columns of the connectivity matrix $\mathbf{C}$ are identically distributed under the dd-IBP and, consequently, the distribution of the random vector $(\mathcal{L}_{ijk} : i, j = 1, \ldots, n)$ is invariant across k.

Let $R_i = \sum_{k=1}^{\infty} z_{ik}$ denote the number of features held by data point i, and let $R_{ij} = \sum_{k=1}^{\infty} z_{ik} z_{jk}$ denote the number of features shared by data points i and j, where $i \neq j$.

Proposition 1. *Under the dd-IBP,*

$$R_i \sim \text{Poisson}\left(\alpha \sum_{n=1}^{N} h_n^{-1} P(\mathcal{L}_{in} = 1)\right), \tag{10}$$

$$R_{ij} \sim \text{Poisson}\left(\alpha \sum_{n=1}^{N} h_n^{-1} P(\mathcal{L}_{in} = 1, \mathcal{L}_{jn} = 1)\right). \tag{11}$$

The probabilities $P(\mathcal{L}_{in} = 1)$ and $P(\mathcal{L}_{in} = 1, \mathcal{L}_{jn} = 1)$ depend strongly on the distribution of the connectivity matrix $\mathbf{C}$, but do not depend on the ownership vector $\mathbf{c}^*$, since $\mathcal{L}$ is independent of dish ownership.

We then use this result to derive the limiting properties of R_i and R_{ij} from properties of the Poisson distribution. In this and following results, $\xrightarrow{P}$ indicates convergence in probability.

Proposition 2. *Let $i \neq j$. R_i and R_{ij} converge in probability under the dd-IBP to the following constants as $\alpha \to \infty$:*

$$\frac{R_i}{\alpha} \xrightarrow{\text{P}} \sum_{n=1}^{N} h_n^{-1} P(\mathcal{L}_{in} = 1), \tag{12}$$

$$\frac{R_{ij}}{\alpha} \xrightarrow{\text{P}} \sum_{n=1}^{N} h_n^{-1} P(\mathcal{L}_{in} = 1, \mathcal{L}_{jn} = 1), \tag{13}$$

$$\frac{R_{ij}}{R_i} \xrightarrow{\text{P}} \frac{\sum_{n=1}^{N} h_n^{-1} P(\mathcal{L}_{in} = 1, \mathcal{L}_{jn} = 1)}{\sum_{n=1}^{N} h_n^{-1} P(\mathcal{L}_{in} = 1)}. \tag{14}$$

This proposition shows that the limiting fraction of shared features R_{ij}/R_i in the dd-IBP is a constant that may be different for each pair of data points i and j. In contrast, we show below that the same limiting fraction under the dHBP is random, and takes one of two values. These two values are fixed, and do not depend upon the data points i and j.

4.2 Feature-Sharing in the dHBP

Here we characterize the limiting distributional properties of feature sharing in the dHBP as B_0 becomes infinitely concentrated (i.e., $\gamma \to \infty$, analogous to $\alpha \to \infty$). In this limit, feature-sharing is primarily attributable to dependency induced by the proximity matrix $\mathbf{A}$.

Proposition 3. *If B_0 is continuous, then under the dHBP,*

$$R_i \mid \mathbf{g}_{1:N} \sim \text{Poisson}(\gamma), \tag{15}$$

$$R_{ij} \mid \mathbf{g}_{1:N} \sim \begin{cases} \text{Poisson}\left(\gamma \frac{c_0 + c_1 + 1}{(c_0+1)(c_1+1)}\right) & \text{if } g_i = g_j, \\ \text{Poisson}\left(\gamma \frac{1}{c_0+1}\right) & \text{if } g_i \neq g_j. \end{cases} \tag{16}$$

We derive the limiting properties of R_i and R_{ij} from properties of the Poisson distribution.

Proposition 4. *Let $i \neq j$ and assume B_0 is continuous. Conditional on $\mathbf{g}_{1:N}$, R_i and R_{ij} converge in probability under the*

dHBP to the following constants as $\gamma \to \infty$:

$$\frac{R_i}{\gamma} \xrightarrow{\mathrm{P}} 1, \tag{17}$$

$$\frac{R_{ij}}{\gamma} \xrightarrow{\mathrm{P}} \begin{cases} \frac{c_0+c_1+1}{(c_0+1)(c_1+1)} & \text{if } g_i = g_j, \\ \frac{1}{c_0+1} & \text{if } g_i \neq g_j, \end{cases} \tag{18}$$

$$\frac{R_{ij}}{R_i} \xrightarrow{\mathrm{P}} \begin{cases} \frac{c_0+c_1+1}{(c_0+1)(c_1+1)} & \text{if } g_i = g_j, \\ \frac{1}{c_0+1} & \text{if } g_i \neq g_j. \end{cases} \tag{19}$$

Thus, the expected fraction of object i's features shared with object j, R_{ij}/R_i, is a factor of $\frac{c_0+c_1+1}{c_1+1}$ bigger when $g_i = g_j$. As $c_0 \to \infty$, this fraction goes to ∞. As $c_0 \to 0$, it goes to 1. We can obtain the unconditional fraction by marginalizing over g_i and g_j:

Corollary 1. *Let $i \neq j$ and assume B_0 is continuous. R_{ij}/R_i converges in probability under the dHBP as $\gamma \to \infty$ to a random variable M_{ij} defined by*

$$M_{ij} = \begin{cases} \frac{c_0+c_1+1}{(c_0+1)(c_1+1)} & \text{with probability } P(g_i = g_j), \\ \frac{1}{c_0+1} & \text{with probability } P(g_i \neq g_j), \end{cases} \tag{20}$$

where $P(g_i = g_j) = \sum_{n=1}^{N} a_{in} a_{jn}$.

This corollary shows that as γ grows large, the fraction of shared features becomes one of two values (determined by c_0 and c_1), with a mixing probability determined by the dependency structure. Thus, the dHBP affords substantial flexibility in specifying the mixing probability (via $\mathbf{A}$), but is constrained to two possible values of the limiting fraction.

4.3 Feature-Sharing in the IBP

For comparison, we briefly describe the feature-sharing properties under the traditional IBP.

Under the traditional IBP, by exchangeability of equivalence classes of $\mathbf{Z}$ and the fact that R_i and R_{ij} are the same for all $\mathbf{Z}$ in the same equivalence class, R_i and R_{ij} together have the same joint distribution as R_1 and R_{12}. The first customer draws a $\mathrm{Poisson}(\alpha)$ number of dishes. The second customer then chooses whether to sample each of these dishes independently and with probability $1/2$. Thus, the number of dishes sampled by both the first and second customers is $R_{12} \sim \mathrm{Poisson}(\alpha/2)$.

This shows (using argument similar to those used to show Propositions 2 and 4) that, under the traditional IBP, as $\alpha \to \infty$ with $i \neq j$,

$$\frac{R_i}{\alpha} \xrightarrow{\mathrm{P}} 1, \qquad \frac{R_{ij}}{\alpha} \xrightarrow{\mathrm{P}} \frac{1}{2}, \qquad \frac{R_{ij}}{R_i} \xrightarrow{\mathrm{P}} \frac{1}{2}. \tag{21}$$

4.4 Discussion

Using an asymptotic analysis, the preceding theoretical results show that the dd-IBP and dHBP provide different forms of flexibility in specifying the way in which features are shared between data points. This asymptotic analysis takes the limit as the mass parameters α and γ become large. This limit is taken for theoretical tractability, and removes much of the uncertainty that is otherwise present in these models. While such limiting dd-IBP and dHBP models are not intended for practical use, their simplicity provides insight into behavior in non-asymptotic regimes.

Under the dd-IBP, Proposition 2 shows that the modeler is allowed a great deal of flexibility in specifying the proportions of features shared by data points. Given a matrix specifying the proportion of features that are believed to be shared by pairs of data points, one can (if this matrix is sufficiently well-behaved) design a distance matrix that causes the dd-IBP to concentrate on the desired proportions. While the dd-IBP cannot model an arbitrary modeler-specified matrix of proportions, the set of matrices that can be modeled is large.

In contrast, under the dHBP, Corollary 1 shows that the modeler has less flexibility in specifying the proportions of features shared. Under the dHBP, the modeler chooses two values, $(c_0 + c_1 + 1)/(c_0 + 1)(c_1 + 1)$ and $1/(c_0 + 1)$, and the proportion of features shared by any pair of data points in the asymptotic regime must be one of these two values.

Section 4.3 shows that the traditional IBP has the least flexibility. In the asymptotic regime, the proportion of features shared by each pair of data points is a constant.

While the dd-IBP has more flexibility in specifying values of the feature-sharing-proportions than the dHBP, it has less flexibility (at least in this asymptotic regime) in modeling uncertainty about these feature-sharing proportions. Under the dd-IBP, the proportion of features shared by a pair of data points in the asymptotic regime is a deterministic quantity. Under the dHBP, the proportion of features shared is a random quantity, even in the asymptotic regime. One could extend the dd-IBP to allow uncertainty about the feature-sharing-proportions by specifying a hyperprior over distance matrices, but we do not consider this extension further.

Fig. 4 illustrates the difference in asymptotic feature-sharing behavior between the dHBP and dd-IBP. Here we use the exponential decay function with $\beta = 1$ and $d_{ij} = (i - j)^2$. Subfigures in the upper row are draws from the dHBP, and subfigures in the bottom row are draws from the dd-IBP. Within a single subfigure, the shade in the cell (i, j) is the fraction R_{ij}/R_i. (The diagonals $R_{ii}/R_i = 1$ have been set to 0 to bring out other aspects of the matrix.) Each of the four columns represents a pair of independent draws. To approximate the asymptotic regime considered by the theory, the mass parameters for the two models are set to large values of $\gamma = \alpha = 1{,}000$. The figure shows that, in draws from the dHBP, off-diagonal cells have one of two shades, corresponding to the two possible limiting values for R_{ij}/R_i. In the different columns, corresponding to different independent draws, the patterns are different, showing that R_{ij}/R_i remains random under the dHBP, even in the asymptotic regime. In contrast, in draws from the dd-IBP, off-diagonal cells take a variety of different values, but remain unchanged across independent draws.

Fig. 5 illustrates non-asymptotic feature-sharing behavior in a simple setting with only two data points. The figure shows the feature-sharing behavior of the dHBP (top) and dd-IBP (bottom) at two values for the mass parameter: $\alpha = \gamma = 15$ (top row) and $\alpha = \gamma = 30$ (bottom row). Each subfigure shows the probability mass function $P(R_{ij})$ as a

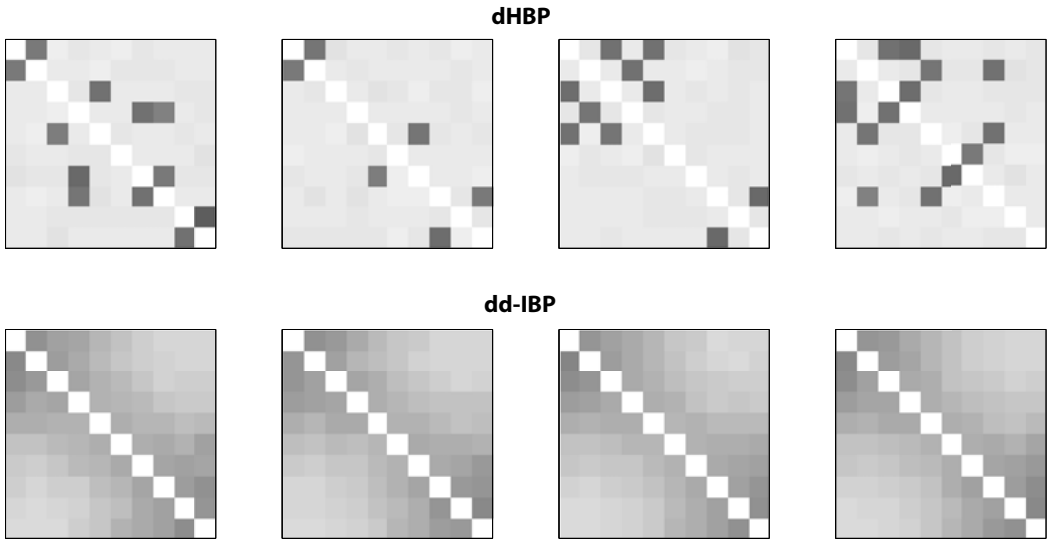

Fig. 4. Feature-sharing in the dHBP and dd-IBP, limiting case. Along the horizontal axis, we show four independent draws from the dHBP (*Top*) and dd-IBP (*Bottom*). Within each subfigure, the shade of a cell (i, j) shows the fraction R_{ij}/R_i, where R_i is the number of features held by data point i, and R_{ij} is the number held by both i and j. Diagonals $R_{ii}/R_i = 1$ have been set to 0 for clarity. Here, $\alpha = \gamma = 1,000$. Limiting results from Section 4 explain the behavior for such large α and γ: for the dHBP the feature-sharing proportion R_{ij}/R_i is random and equal to one of two constants; for the dd-IBP the proportion is non-random and takes a range of values. The dd-IBP models feature-sharing proportions that differ across data points, but does not model uncertainty about these proportions when mass parameters are large.

function of the proximity a_{ij}, where $a_{ij} = 1/d_{ij}$ for the dd-IBP. Because there are only two data points, with $a_{ii} = 1$ and $a_{ij} = a_{ji}$, specifying a_{ij} is sufficient for specifying the full proximity matrix $\mathbf{A}$. For the dHBP, we set $c_0 = 10$ and $c_1 = 1$. Also facilitating comparison, $\mathbb{E}[R_i]$ is the same between both models (when $\alpha = \gamma$).

Fig. 5 shows that as the proximity a_{ij} increases to 1, the number of shared features R_{ij} tends to increase under both models. More precisely, $P(R_{ij})$ concentrates on larger values of R_{ij} as a_{ij} increases. However, the way in which the probability mass functions change with a_{ij} differs between the two models. In the dd-IBP, the most likely value of R_{ij} increases smoothly, while under the dHBP it remains roughly constant and then jumps. As one varies a_{ij} across its full range, the set of most likely values for R_{ij} under the dd-IBP spans its full range from 0 to 20, while under the dHBP the most likely value for R_{ij} takes only a few values. Instead, varying a_{ij} under the dHBP allows a variety of bimodal distributions centered near the values from the asymptotic analysis.

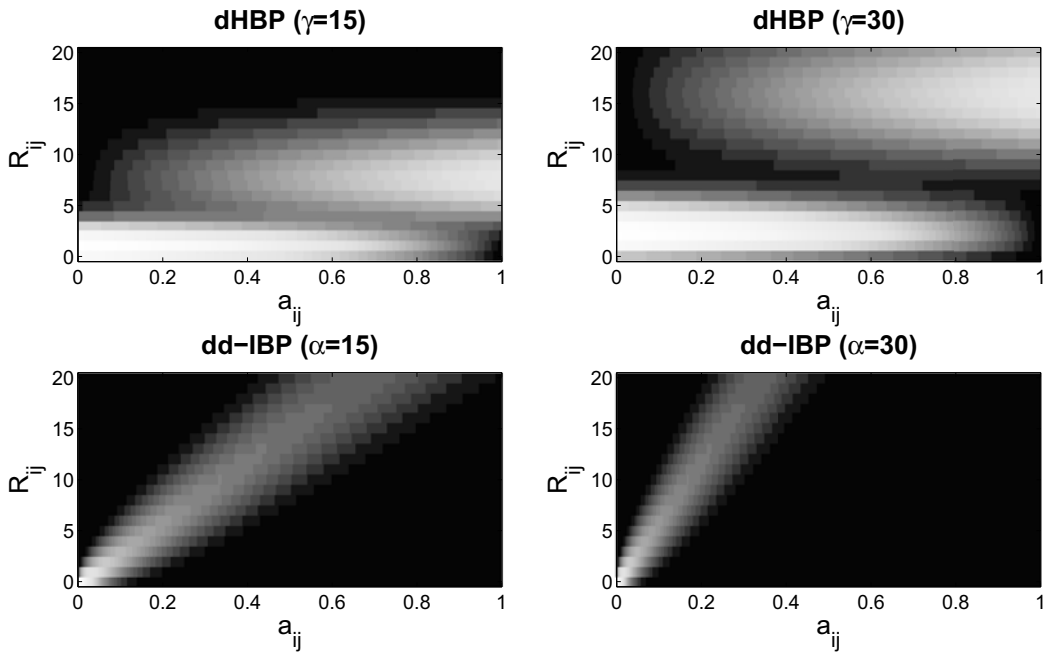

Fig. 5. Feature-sharing in the dHBP and dd-IBP. Heatmaps of the probability mass function over the number of shared features R_{ij} (*y*-axis) as a function of proximity a_{ij} (x-axis) in a data set consisting of two data points. Black indicates a probability mass of 0, with lighter shades indicating larger values. For the dHBP, we set $c_0 = 10$ and $c_1 = 1$. Note that $\mathbb{E}[R_i]$ is the same for both the dHBP and dd-IBP in these examples (when $\alpha = \gamma$).

This difference in non-asymptotic behaviors mirrors the difference between the two models in the asymptotic regime, where the dd-IBP allows feature-sharing-proportions to be specified almost arbitrarily but allows little flexibility in modeling uncertainty about them, and the dHBP limits the number of possible values for the feature-sharing proportions, but allows uncertainty over these values. This difference may have consequences for data analysis. For example, latent features underlying preferences for movies or music may be shared in highly predictable but diverse ways across a set of individuals, in which case the dd-IBP would be a more suitable model. In contrast, latent features underlying disease may be shared in a smaller number of ways but be highly unpredictable (e.g., two individuals either share or don't share a particular stochastically occurring gene mutation), in which case the dHBP would be more suitable.

5 Inference Using Markov Chain Monte Carlo Sampling

Given a data set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ and a latent feature model $P(\mathbf{X} \mid \mathbf{Z}, \theta)$ with parameter θ, the goal of inference is to compute the joint posterior over the customer assignment matrix $\mathbf{C}$, the dd-IBP hyperparameter α, and likelihood parameter θ, as given by Bayes' rule:

$$P(\mathbf{C}, \mathbf{c}^*, \theta, \alpha \mid \mathbf{X}, \mathbf{D}, f) \propto P(\mathbf{X} \mid \mathbf{C}, \mathbf{c}^*, \theta) P(\theta) P(\mathbf{C} \mid \mathbf{D}, f) P(\mathbf{c}^* \mid \alpha) P(\alpha), \tag{22}$$

where the first term is the likelihood (recall that $\mathbf{Z}$ is a deterministic function of $\mathbf{C}$ and $\mathbf{c}^*$), the second term is the prior over parameters, the third term is the dd-IBP prior over the connectivity matrix $\mathbf{C}$, the fourth term is the prior over the ownership vector $\mathbf{c}^*$, and the last term is the prior over α.

Exact inference in this model is computationally intractable. We therefore use MCMC sampling [26] to approximate the posterior with L samples. Details of this algorithm can be found in the Supplementary Materials, available online. The algorithm can be adapted to different data sets by choosing an appropriate likelihood function. In the next section, we present a simple linear-Gaussian model.[4]

6 A Linear-Gaussian Model

As an example of how the dd-IBP can be used in data analysis, we incorporate it into a linear-Gaussian latent feature model (Fig. 6). This model was originally studied for the IBP by Griffiths and Ghahramani [16], [17]. The observed data $\mathbf{X} \in \mathbb{R}^{N \times M}$ consist of N objects, each of which is a M-dimensional vector of real-valued object properties. We model $\mathbf{X}$ as a linear combination of binary latent features corrupted by Gaussian noise:

$$\mathbf{X} = \mathbf{Z}\mathbf{W} + \epsilon, \tag{23}$$

where $\mathbf{W}$ is a $K \times M$ matrix of real-valued weights, and ϵ is a $N \times M$ matrix of independent, zero-mean Gaussian noise terms with standard deviation σ_x. We place a zero-mean Gaussian prior on $\mathbf{W}$ with covariance $\sigma_w^2 \mathbf{I}$. Intuitively, the weights capture how the latent features interact to produce the observed data. For example, if each latent feature corresponds to a person in an image, then the weight w_{km} captures the contribution of person k to pixel m. Algorithmic details for performing inference in this model are provided in the Supplementary Materials, available online.

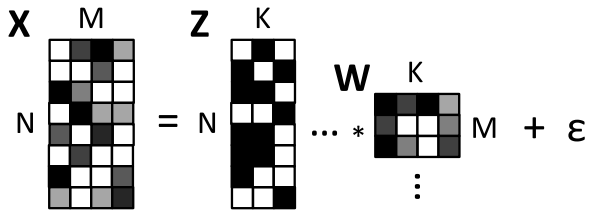

Fig. 6. Linear-Gaussian model. Matrix multiplication view of how latent features ($\mathbf{Z}$) combine with a weight matrix ($\mathbf{W}$) and white noise (ϵ) to produce observed data ($\mathbf{X}$).

7 Empirical Study

In this section we report experimental investigations of the dd-IBP and comparisons with alternative models. We show how the dd-IBP can be used as a dimensionality reduction pre-processing technique for classification tasks when the data points are non-exchangeable. In the Supplementary Materials, available online, we present an example of a situation in which a non-exchangeable model might be expected to help, but does not.

The performance of supervised learning algorithms is often enhanced by pre-processing the data to reduce its dimensionality [3]. Classical techniques for dimensionality reduction, such as principal components analysis and factor analysis, assume exchangeability, as does the infinite latent feature model based on the IBP [16]. For this reason, these techniques may not work as well for pre-processing non-exchangeable data, and this may adversely affect their performance on supervised learning tasks.

We investigated this hypothesis using a magnetic resonance imaging (MRI) data set collected from 27 patients with Alzheimer's disease and 35 healthy controls [7].[5] The observed features consist of four structural summary statistics measured in 56 brain regions of interest: (1) surface area; (2) shape index; (3) curvedness; (4) fractal dimension. The classification task is to sort individuals into Alzheimer's or control classes based on their observed features.

Age-related changes in brain structure produce natural declines in cognitive function that make diagnosis of Alzheimer's disease difficult [12]. Thus, it is important to take age into account when designing predictive models. For the dd-IBP and dHBP, age is naturally incorporated as a covariate over which we constructed a distance matrix. Specifically, we defined d_{ij} as the absolute age difference between subjects i and j, with $d_{ij} = \infty$ for $j > i$ (i.e., the distance matrix is sequential). This induces a prior belief that individuals with similar ages tend to share more latent features. In the MRI data set, ages ranged from 60 to 90 (median: 76.5).

4. Matlab software implementing this algorithm is available at the first author's homepage: `web.mit.edu/sjgershm/www`.

5. Available at: `http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_July2009_ID_NI`.

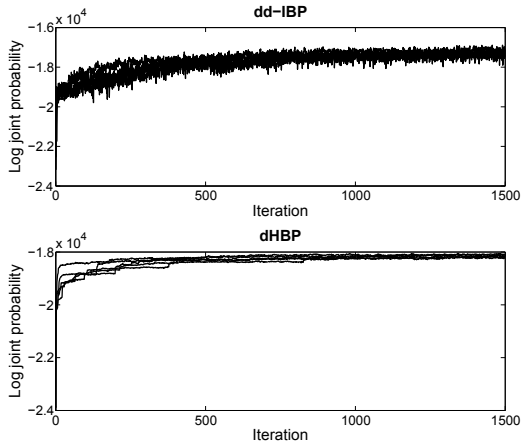

Fig. 7. Trace plots. Representative traces of the log joint probability of the Alzheimer's data and latent variables for the dd-IBP (*top*) and dHBP (*bottom*). Each iteration corresponds to a sweep over all the latent variables.

In detail, we ran 1,500 iterations of MCMC sampling on the entire data set using the linear-Gaussian observation model, and then selected the latent features of the *maximum a posteriori* sample as input to a supervised learning algorithm (L_2-regularized logistic regression, with the regularization constant set to 10^{-6}). Training was performed on half of the data, and testing on the other half.[6] The noise hyperparameters of the dd-IBP and dHBP (σ_x and σ_y) were updated using Metropolis-Hastings proposals.

We monitored the log of the joint distribution $\log P(\mathbf{X}, \alpha, \mathbf{C}, \mathbf{c}^*)$. Visual inspection of the log joint probability traces suggested that the sampler reaches a local maximum within 400-500 iterations (Fig. 7). This process was repeated for a range of decay parameter (β) values, using the exponential decay function. The same proximity matrix, $\mathbf{A}$, was used for both the dd-IBP and dHBP. We performed 10 random restarts of the sampler (initialized to draws from the prior) and recomputed the classification measure for each restart, averaging the resulting measures to reduce sampling variability. For comparison, we also made predictions using the standard IBP, the dIBP [28], and the raw observed features (i.e., no pre-processing). The dIBP was fit using the MCMC algorithm described in Williamson et al. [28], which adaptively samples the parameters controlling dependencies between observations (thus the results do not depend on β).

The inferred latent features are shown in Fig. 8, illustrating how the distribution of features shifts across values of the covariate. In this case, the data from younger Alzheimer's patients are explained by a set of latent features that are largely distinct from the latent features used to explain the data from older patients.

Classification results are shown in Fig. 9 (left), where performance is measured as the area under the receiver operating characteristic curve (AUC). Chance performance corresponds to an AUC of 0.5, perfect performance to an AUC of 1. Using features from the non-exchangeable models (dd-IBP, dHBP, dIBP) leads to better performance than the raw and IBP features. For a range of β values, the dd-IBP produces superior classification performance to the alternative models, with the exception of the dIBP (which outperformed the dd-IBP for some settings).

We also ran the dd-IBP sampler with $\beta = 0$ (in which case the dd-IBP and IBP are equivalent) and found no significant difference between it and the standard IBP sampler with respect to performance on the Alzheimer's classification.

6. A few randomly chosen individuals were removed from the test set to make it balanced.

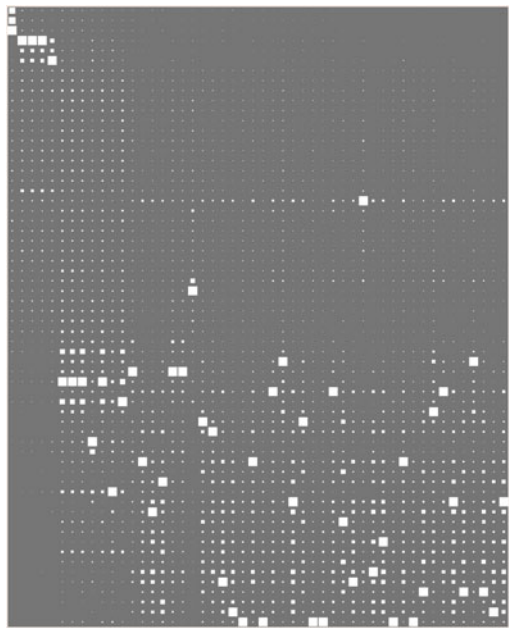

Fig. 8. Inferred latent features for the Alzheimer's data set. Hinton diagram showing the posterior expected latent feature matrix. Rows correspond to data points (ordered according to increasing age), columns correspond to latent features. The size of the square indicates the magnitude of the corresponding entry.

8 Conclusions

By relaxing the exchangeability assumption for infinite latent feature models, the dd-IBP extends their applicability

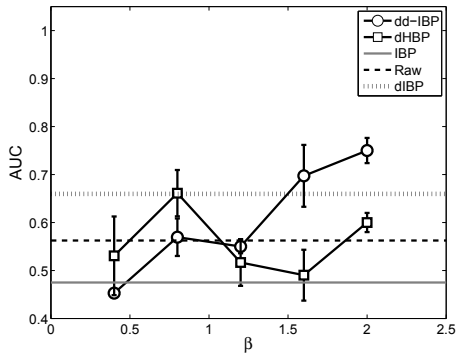

Fig. 9. Classification results for the Alzheimer's data set. Area under the curve (AUC) for binary classification (Alzheimer's versus normal control) using L_2-regularized logistic regression and features learned from a linear-Gaussian latent feature model. Each curve represents a different choice of predictor variables (latent features) for logistic regression. The x-axis corresponds to different settings of the exponential decay function parameter, β. "Raw" refers to the original data features (see text for details); the IBP, dd-IBP, dIBP and dHBP results were based on using the latent features of the *maximum a posteriori* sample following $1{,}500$ iterations of MCMC sampling. Error bars represent standard error of the mean.

to a richer class of data. We have shown empirically that this innovation fares better than the standard IBP on non-exchangeable data (e.g., timeseries).

We note that the dd-IBP is not a standard Bayesian nonparametric distribution, in the sense of arising from a de Finetti mixing distribution. For the standard IBP, the de Finetti mixing distribution is the beta process [27], but this result does not generalize to the dd-IBP due to its non-exchangeability. Nonetheless, this does not detract from our model's ability to let the data infer the number of latent features, a property that it shares with other infinite latent feature models.

We can consider a number of possible future directions. First, we may exploit distance dependence to derive more efficient samplers. In particular, Doshi-Velez and Ghahramani [9] have shown that partitioning the data into subsets enables faster Gibbs sampling for the traditional IBP; the window decay function imposes a natural partition of the data into conditionally independent subsets.

Second, we can apply the dd-IBP to other likelihood functions. For example, it could be applied to relational data [19], [21] or text data [27]. As pointed out by Miller et al. [21], covariates like age or location often play an important role in link prediction. Whereas Miller et al. [21] incorporated covariates into the likelihood function, one could instead incorporate them into the prior by defining covariate-based distances between data points (e.g., the age difference between two people). A distinction of the latter approach is that it would allow one to model dependencies in terms of latent features. For instance, two people close in age or geographic location may be more likely to share latent interests, a pattern naturally captured by the dd-IBP.

Third, modeling shared dependency structure across groups is important for several applications. In brain imaging studies, for example, similar spatial and temporal dependencies are frequently observed across subjects. Modeling shared structure without sacrificing intersubject variability has been addressed with hierarchical models [2], [29]. One way to extend the dd-IBP hierarchically would be to allow the parameters of the decay function to vary across individuals while being coupled together by higher-level variables.

Acknowledgments

SJG was supported by a US National Science Foundation (NSF) graduate research fellowship and a postdoctoral fellowship from the MIT Intelligence Initiative. PIF acknowledges support from NSF Awards #142251, #1247696 and #1254298, and from AFOSR Awards FA9550-11-1-0083 and FA9550-12-1-0200. DMB acknowledges support from NSF CAREER NSF IIS-0745520, NSF BIGDATA NSF IIS-1247664, NSF NEURO NSF IIS-1009542, ONR N00014-11-1-0651, the Alfred P. Sloan foundation, and DARPA FA8750-14-2-0009. The authors thank Matt Hoffman, Chong Wang, Gungor Polatkan, Sean Gerrish and John Paisley for helpful discussions. They are also grateful to Sinead Williamson and Mingyuan Zhou for sharing their code.

References

[1] A. Ahmed and E. P. Xing, "Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream," in *Proc. 26th Conf. Uncertainty Artif. Intell.*, 2010.

[2] C. F. Beckmann, M. Jenkinson, and S. M. Smith, "General multilevel linear modeling for group analysis in FMRI," *NeuroImage*, vol. 20, no. 2, pp. 1052–1063, 2003.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[4] D. Blackwell and J. MacQueen, "Ferguson distributions via Pólya urn schemes," *The Ann. Statist.*, vol. 1, no. 2, pp. 353–355, 1973.

[5] D. M. Blei and P. I. Frazier, "Distance dependent Chinese restaurant processes," *J. Mach. Learn. Res.*, vol. 12, pp. 2461–2488, 2011.

[6] F. Caron, M. Davy, and A. Doucet, "Generalized Polya urn for time-varying Dirichlet process mixtures," in *Proc. 23rd Conf. Uncertainity Artif. Intell.*, 2007.

[7] N. Christou and I. D. Dinov, "Confidence interval based parameter estimation, a new SOCR applet and activity," *PloS One*, vol. 6, no. 5, p. e19178, 2011.

[8] Y. Chung and D. B. Dunson, "The local Dirichlet process," *Ann. Inst. Statist. Math.*, vol. 63, no. 1, pp. 59–80, 2011.

[9] F. Doshi-Velez and Z. Ghahramani, "Accelerated sampling for the Indian buffet process," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 273–280.

[10] F. Doshi-Velez and Z. Ghahramani, "Correlated non-parametric latent feature models," in *Proc. 25th Conf. Uncertainity Artif. Intell.*, 2009, pp. 143–150.

[11] J. A. Duan, M. Guindani, and A. E. Gelfand, "Generalized spatial Dirichlet process models," *Biometrika*, vol. 94, no. 4, pp. 809–825, 2007.

[12] T. Erkinjuntti, R. Laaksonen, R. Sulkava, R. Syrjäläinen, and J. Palo, "Neuropsychological differentiation between normal aging, Alzheimer's disease and vascular dementia," *Acta Neurologica Scandinavica*, vol. 74, no. 5, pp. 393–403, 1986.

[13] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures.," *J. Amer. Statist. Assoc.*, vol. 90, no. 430, pp. 577–588, 1995.

[14] N. J. Foti and S. Williamson, "A survey of non-exchangeable priors for Bayesian nonparametric models," 2012.

[15] J. E. Griffin and M. F. J. Steel, "Order-based dependent Dirichlet processes," *J. Amer. Statist. Assoc.*, vol. 101, no. 473, pp. 179–194, 2006.

[16] T. L. Griffiths and Z. Ghahramani, "Infinite latent feature models and the Indian buffet process," *in Proc. Adv. Neural Inform. Process. Syst.*, vol. 18, 2005.

[17] T. Griffiths and Z. Ghahramani, "The Indian buffet process: An introduction and review," *J. Mach. Learn. Res.*, vol. 12, pp. 1185–1224, 2011.

[18] D. Knowles and Z. Ghahramani, "Infinite sparse factor analysis and infinite independent components analysis," in *Proc. 7th Int. Conf. Ind. Compon. Anal. Signal Sep.*, 2007, pp. 381–388.

[19] E. Meeds, Z. Ghahramani, R. M. Neal, and S. T. Roweis, "Modeling dyadic data with binary latent factors," in *Proc. Adv. Neural Inform. Process. Syst.*, 2007, vol. 19, pp. 977–984.

[20] K. T. Miller, T. L. Griffiths, and M. I. Jordan, "The phylogenetic Indian buffet process: A non-exchangeable nonparametric prior for latent features," in *Proc. 24th Conf. Uncertainty Artif. Intell.*, 2008.

[21] K. T. Miller, T. L. Griffiths, and M. I. Jordan, "Nonparametric latent feature models for link prediction," in *Proc. Adv. Neural Inform. Process. Syst.*, 2009.

[22] D. J. Navarro and T. L. Griffiths, "Latent features in similarity judgments: A nonparametric Bayesian approach," *Neural Comput.*, vol. 20, no. 11, pp. 2597–2628, 2008.

[23] V. Rao and Y. W. Teh, "Spatial normalized gamma processes," in *Proc. Adv. Neural Inform. Process. Syst.*, 2009, vol. 22, pp. 1554–1562.

[24] C. E. Rasmussen, "The infinite Gaussian mixture model," in *Proc. Adv. Neural Inform. Process. Syst.*, 2000, vol. 12, pp. 554–560.

[25] L. Ren, Y. Wang, D. Dunson, and L. Carin, "The kernel beta process," in *Proc. Adv. Neural Inform. Process. Syst.*, 2011, pp. 963–971.

[26] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. New York, NY, USA: Springer Verlag, 2004.

[27] R. Thibaux and M. I. Jordan, "Hierarchical beta processes and the Indian buffet process," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2007, vol. 11, pp. 564–571.

[28] S. Williamson, P. Orbanz, and Z. Ghahramani, "Dependent Indian buffet processes," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010.

[29] M. W. Woolrich, T. E. J. Behrens, C. F. Beckmann, M. Jenkinson, and S. M. Smith, "Multilevel linear modelling for FMRI group analysis using Bayesian inference," *NeuroImage*, vol. 21, no. 4, pp. 1732–1747, 2004.

[30] M. Zhou, H. Yang, G. Sapiro, D. Dunson, and L. Carin, "Dependent hierarchical beta process for image interpolation and denoising," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2011.

Samuel J. Gershman received the PhD degree in psychology and neuroscience from Princeton University in 2013. He is currently a postdoctoral fellow in the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology. He received a graduate research fellowship from the National Science Foundation. His research focuses on computational and experimental studies of learning in humans and animals.

Peter I. Frazier received the PhD degree in operations research and financial engineering from Princeton University in 2009. He is currently an assistant professor in the School of Operations Research and Information Engineering, Cornell University. He received the AFOSR Young Investigator Award, and the US National Science Foundation (NSF) CAREER Award. He is an associate editor for *Operations Research*, *ACM Transactions on Modeling and Computer Simulation* and *IIE Transactions*, and was on program committees for ICML, NIPS and Winter Simulation Conference. He has received best paper awards from the Institute for Operations Research and the Management Sciences (INFORMS) and Winter Simulation Conference. His research interest include dynamic programming and Bayesian statistics, focusing on the optimal acquisition of information and sequential design of experiments. He works on applications in simulation, optimization, operations management, and medicine.

David M. Blei received the PhD degree in 2004 from U.C. Berkeley and was a postdoctoral fellow at Carnegie Mellon University. He is currently an associate professor of computer science at Princeton University. His research focuses on probabilistic topic models, Bayesian nonparametric methods, and approximate posterior inference. He works on a variety of applications, including text, images, music, social networks, and scientific data.