Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics
Benidorm (Alicante, Spain), June 1st–6th, 2006

Bayesian nonparametric latent feature models

Zoubin Ghahramani
University of Cambridge, UK
zoubin@eng.cam.ac.uk
Thomas L. Griffiths
University of California at Berkeley, USA
tom_griffiths@berkeley.edu
Peter Sollich
King's College London, UK
peter.sollich@kcl.ac.uk

Summary

We describe a flexible nonparametric approach to latent variable modelling in which the number of latent variables is unbounded. This approach is based on a probability distribution over equivalence classes of binary matrices with a finite number of rows, corresponding to the data points, and an unbounded number of columns, corresponding to the latent variables. Each data point can be associated with a subset of the possible latent variables, which we refer to as the latent *features* of that data point. The binary variables in the matrix indicate which latent feature is possessed by which data point, and there is a potentially infinite array of features. We derive the distribution over unbounded binary matrices by taking the limit of a distribution over $N \times K$ binary matrices as $K \to \infty$. We define a simple generative processes for this distribution which we call the Indian buffet process (IBP; Griffiths and Ghahramani, 2005, 2006) by analogy to the Chinese restaurant process (Aldous, 1985; Pitman, 2002). The IBP has a single hyperparameter which controls both the number of feature per object and the total number of features. We describe a two-parameter generalization of the IBP which has additional flexibility, independently controlling the number of features per object and the total number of features in the matrix. The use of this distribution as a prior in an infinite latent feature model is illustrated, and Markov chain Monte Carlo algorithms for inference are described.

Keywords and Phrases: Non-parametric methods; MCMC; Indian buffet process; latent variable models.

Zoubin Ghahramani is in the Department of Engineering, University of Cambridge, and the Machine Learning Department, Carnegie Mellon University; Thomas L. Griffiths is in the Psychology Department, University of California at Berkeley; Peter Sollich is in the Department of Mathematics, King's College London.

1. INTRODUCTION

Latent or hidden variables are an important component of many statistical models. The role of these latent variables may be to represent properties of the objects or data points being modelled that have not been directly observed, or to represent hidden causes that explain the observed data.

Most models with latent variables assume a finite number of latent variables per object. At the extreme, mixture models can be represented via *a single* discrete latent variable, and hidden Markov models (HMMs) via a single latent variable evolving over time. Factor analysis and independent components analysis (ICA) generally use more than one latent variable per object but this number is usually assumed to be small. The close relationship between latent variable models such as factor analysis, state-space models, finite mixture models, HMMs, and ICA is reviewed in (Roweis and Ghahramani, 1999).

Our goal is to describe a class of latent variable models in which each object is associated with a (potentially unbounded) vector of latent features. Latent feature representations can be found in several widely-used statistical models. In Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003) each object is associated with a probability distribution over latent features. LDA has proven very successful for modelling the content of documents, where each feature indicates one of the topics that appears in the document. While using a probability distribution over features may be sensible to model the distribution of topics in a document, it introduces a conservation constraint—the more an object expresses one feature, the less it can express others—which may not be appropriate in other contexts. Other latent feature representations include binary vectors with entries indicating the presence or absence of each feature (e.g., Ueda & Saito, 2003), continuous vectors representing objects as points in a latent space (e.g., Jolliffe, 1986), and factorial models, in which each feature takes on one of a discrete set of values (e.g., Zemel & Hinton, 1994; Ghahramani, 1995).

While it may be computationally convenient to define models with a small finite number of latent variables or latent features per object, it may be statistically inappropriate to constrain the number of latent variables a priori. The problem of finding the number of latent variables in a statistical model has often been treated as a model selection problem, choosing the model with the dimensionality that results in the best performance. However, this treatment of the problem assumes that there is a single, finite-dimensional representation that correctly characterizes the properties of the observed objects. This assumption may be unreasonable. For example, when modelling symptoms in medical patients, the latent variables may include not only presence or absence of known diseases but also any number of environmental and genetic factors and potentially unknown diseases which relate to the pattern of symptoms the patient exhibited.

The assumption that the observed objects manifest a sparse subset of an unbounded number of latent classes is often used in nonparametric Bayesian statistics. In particular, this assumption is made in Dirichlet process mixture models, which are used for nonparametric density estimation (Antoniak, 1974; Escobar & West, 1995; Ferguson, 1983; Neal, 2000). Under one interpretation of a Dirichlet process mixture model, each object is assigned to a latent class, and each class is associated with a distribution over observable properties. The prior distribution over assignments of objects to classes is specified in such a way that the number of classes used by the model is bounded only by the number of objects, making Dirichlet process mixture models "infinite" mixture models (Rasmussen, 2000). Recent work

has extended these methods to models in which each object is represented by a distribution over features (Blei, Griffiths, Jordan, & Tenenbaum, 2004; Teh, Jordan, Beal, & Blei, 2004). However, there are no equivalent methods for dealing with other feature-based representations, be they binary vectors, factorial structures, or vectors of continuous feature values.

In this paper, we take the idea of defining priors over infinite combinatorial structures from nonparametric Bayesian statistics, and use it to develop methods for unsupervised learning in which each object is represented by a sparse subset of an unbounded number of features. These features can be binary, take on multiple discrete values, or have continuous weights. In all of these representations, the difficult problem is deciding which features an object should possess. The set of features possessed by a set of objects can be expressed in the form of a binary matrix, where each row is an object, each column is a feature, and an entry of 1 indicates that a particular objects possesses a particular feature. We thus focus on the problem of defining a distribution on infinite sparse binary matrices. Our derivation of this distribution is analogous to the limiting argument in (Neal 2000; Green and Richardson, 2001) used to derive the Dirichlet process mixture model (Antoniak, 1974; Ferguson, 1983), and the resulting process we obtain is analogous to the *Chinese restaurant process* (CRP; Aldous, 1985; Pitman, 2002). This distribution over infinite binary matrices can be used to specify probabilistic models that represent objects with infinitely many binary features, and can be combined with priors on feature values to produce factorial and continuous representations.

The plan of the paper is as follows. Section 2 discusses the role of a prior on infinite binary matrices in defining infinite latent feature models. Section 3 describes such a prior, corresponding to a stochastic process we call the *Indian buffet process* (IBP). Section 4 describes a two-parameter extension of this model which allows additional flexibility in the structure of the infinite binary matrices. Section 6 illustrates several applications of the IBP prior. Section 7 presents some conclusions.

2. LATENT FEATURE MODELS

Assume we have N objects, represented by an $N \times D$ matrix $\mathbf{X}$, where the ith row of this matrix, $\mathbf{x}_i$, consists of measurements of D observable properties of the ith object. In a latent feature model, each object is represented by a vector of latent feature values $\mathbf{f}_i$, and the properties $\mathbf{x}_i$ are generated from a distribution determined by those latent feature values. Latent feature values can be continuous, as in principal component analysis (PCA; Jolliffe, 1986), or discrete, as in cooperative vector quantization (CVQ; Zemel & Hinton, 1994; Ghahramani, 1995). In the remainder of this Section, we will assume that feature values are continuous. Using the matrix $\mathbf{F} = \left[\mathbf{f}_1^T \; \mathbf{f}_2^T \; \cdots \; \mathbf{f}_N^T\right]^T$ to indicate the latent feature values for all N objects, the model is specified by a prior over features, $p(\mathbf{F})$, and a distribution over observed property matrices conditioned on those features, $p(\mathbf{X}|\mathbf{F})$. As with latent class models, these distributions can be dealt with separately: $p(\mathbf{F})$ specifies the number of features, their probability, and the distribution over values associated with each feature, while $p(\mathbf{X}|\mathbf{F})$ determines how these features relate to the properties of objects. Our focus will be on $p(\mathbf{F})$, showing how such a prior can be defined without placing an upper bound on the number of features.

We can break the matrix $\mathbf{F}$ into two components: a binary matrix $\mathbf{Z}$ indicating which features are possessed by each object, with $z_{ik} = 1$ if object i has feature

latent variable	finite model ($K < \infty$)	infinite model ($K = \infty$)
$f_i \in \{1...K\}$	finite mixture model	DPM
$\mathbf{f}_i \in [0,1]^K, \sum_k f_{ik} = 1$	LDA	HDP
$\mathbf{f}_i \in \{0,1\}^K$	factorial models, CVQ	IBP
$\mathbf{f}_i \in \Re^K$	FA, PCA, ICA	derivable from IBP

Table 1: Some different latent variable models and the set of values their latent variables can take. DPM: Dirichlet process mixture; FA: factor analysis; HDP: Hierarchical Dirichlet process; IBP: Indian buffet process (described in this paper). Other acronyms are defined in the main text. "Derivable from IBP" refers to different choices for the distribution of $\mathbf{V}$.

k and 0 otherwise, and a second matrix $\mathbf{V}$ indicating the value of each feature for each object. $\mathbf{F}$ can be expressed as the elementwise (Hadamard) product of $\mathbf{Z}$ and $\mathbf{V}$, $\mathbf{F} = \mathbf{Z} \otimes \mathbf{V}$, as illustrated in Figure 1. In many latent feature models, such as PCA and CVQ, objects have non-zero values on every feature, and every entry of $\mathbf{Z}$ is 1. In *sparse* latent feature models (e.g., sparse PCA; d'Aspremont, Ghaoui, Jordan, & Lanckriet, 2005; Jolliffe & Uddin, 2003; Zou, Hastie, & Tibshirani, 2006) only a subset of features take on non-zero values for each object, and $\mathbf{Z}$ picks out these subsets. Table 1 shows the set of possible values that latent variables can take in different latent variable models.

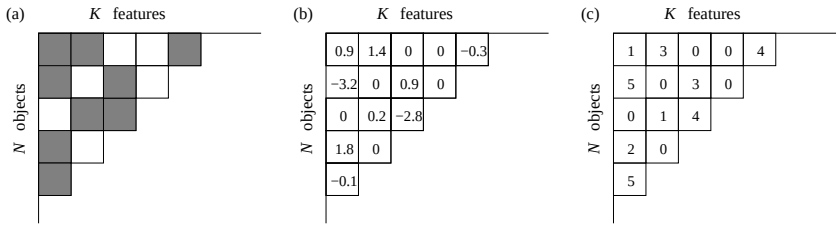

Figure 1: *Feature matrices. A binary matrix* $\mathbf{Z}$*, as shown in (a), can be used as the basis for sparse infinite latent feature models, indicating which features take non-zero values. Elementwise multiplication of* $\mathbf{Z}$ *by a matrix* $\mathbf{V}$ *of continuous values gives a representation like that shown in (b). If* $\mathbf{V}$ *contains discrete values, we obtain a representation like that shown in (c).*

A prior on $\mathbf{F}$ can be defined by specifying priors for $\mathbf{Z}$ and $\mathbf{V}$ separately, with $p(\mathbf{F}) = P(\mathbf{Z})p(\mathbf{V})$. We will focus on defining a prior on $\mathbf{Z}$, since the effective dimensionality of a latent feature model is determined by $\mathbf{Z}$. Assuming that $\mathbf{Z}$ is sparse, we can define a prior for infinite latent feature models by defining a distribution over infinite binary matrices. We have two desiderata for such a distribution: objects should be exchangeable, and inference should be tractable. The literature on nonparametric Bayesian models suggests a method by which these desiderata can be satisfied: start with a model that assumes a finite number of features, and consider the limit as the number of features approaches infinity (Neal, 2000; Green and Richardson, 2001).

3. A DISTRIBUTION ON INFINITE BINARY MATRICES

In this Section, we derive a distribution on infinite binary matrices by starting with a simple model that assumes K features, and then taking the limit as $K \to \infty$. The resulting distribution corresponds to a simple generative process, which we term the Indian buffet process.

3.1. *A finite feature model*

We have N objects and K features, and the possession of feature k by object i is indicated by a binary variable z_{ik}. Each object can possess multiple features. The z_{ik} thus form a binary $N \times K$ feature matrix, $\mathbf{Z}$. We will assume that each object possesses feature k with probability π_k, and that the features are generated independently. The probabilities π_k can each take on any value in $[0, 1]$. Under this model, the probability of a matrix $\mathbf{Z}$ given $\pi = \{\pi_1, \pi_2, \ldots, \pi_K\}$, is

$$P(\mathbf{Z}|\pi) = \prod_{k=1}^{K} \prod_{i=1}^{N} P(z_{ik}|\pi_k) = \prod_{k=1}^{K} \pi_k^{m_k} (1 - \pi_k)^{N - m_k}, \tag{1}$$

where $m_k = \sum_{i=1}^{N} z_{ik}$ is the number of objects possessing feature k.

We can define a prior on π by assuming that each π_k follows a beta distribution. The beta distribution has parameters r and s, and is conjugate to the binomial. The probability of any π_k under the Beta(r, s) distribution is given by

$$p(\pi_k) = \frac{\pi_k^{r-1}(1 - \pi_k)^{s-1}}{B(r, s)}, \tag{2}$$

where $B(r, s)$ is the beta function,

$$B(r, s) \quad = \quad \int_0^1 \pi_k^{r-1}(1 - \pi_k)^{s-1} \, d\pi_k = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r + s)}. \tag{3}$$

We take $r = \frac{\alpha}{K}$ and $s = 1$, so Equation 3 becomes

$$B(\tfrac{\alpha}{K}, 1) = \frac{\Gamma(\frac{\alpha}{K})}{\Gamma(1 + \frac{\alpha}{K})} = \frac{K}{\alpha}, \tag{4}$$

exploiting the recursive definition of the gamma function. The effect of varying s is explored in Section 4.

The probability model we have defined is

$$\begin{aligned} \pi_k \,|\, \alpha &\sim \text{Beta}(\tfrac{\alpha}{K}, 1) \\ z_{ik} \,|\, \pi_k &\sim \text{Bernoulli}(\pi_k) \end{aligned}$$

Each z_{ik} is independent of all other assignments, conditioned on π_k, and the π_k are generated independently. Having defined a prior on π, we can simplify this model by integrating over all values for π rather than representing them explicitly. The

marginal probability of a binary matrix $\mathbf{Z}$ is

$$P(\mathbf{Z}) = \prod_{k=1}^{K} \int \left(\prod_{i=1}^{N} P(z_{ik}|\pi_k) \right) p(\pi_k)\, d\pi_k \tag{5}$$

$$= \prod_{k=1}^{K} \frac{B(m_k + \frac{\alpha}{K}, N - m_k + 1)}{B(\frac{\alpha}{K}, 1)} \tag{6}$$

$$= \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \frac{\alpha}{K})}. \tag{7}$$

This result follows from conjugacy between the binomial and beta distributions. This distribution is exchangeable, depending only on the counts m_k.

This model has the important property that the expectation of the number of non-zero entries in the matrix $\mathbf{Z}$, $E\left[\mathbf{1}^T\mathbf{Z}\mathbf{1}\right] = E\left[\sum_{ik} z_{ik}\right]$, has an upper bound for any K. Since each column of $\mathbf{Z}$ is independent, the expectation is K times the expectation of the sum of a single column, $E\left[\mathbf{1}^T\mathbf{z}_k\right]$. This expectation is easily computed,

$$E\left[\mathbf{1}^T\mathbf{z}_k\right] = \sum_{i=1}^{N} E(z_{ik}) = \sum_{i=1}^{N} \int_0^1 \pi_k p(\pi_k)\, d\pi_k = N\frac{\frac{\alpha}{K}}{1 + \frac{\alpha}{K}}, \tag{8}$$

where the result follows from the fact that the expectation of a Beta(r, s) random variable is $r/(r + s)$. Consequently, $E\left[\mathbf{1}^T\mathbf{Z}\mathbf{1}\right] = KE\left[\mathbf{1}^T\mathbf{z}_k\right] = N\alpha/(1 + (\alpha/K))$. For any K, the expectation of the number of entries in $\mathbf{Z}$ is bounded above by $N\alpha$.

3.2. *Equivalence classes*

In order to find the limit of the distribution specified by Equation 7 as $K \to \infty$, we need to define equivalence classes of binary matrices. Our equivalence classes will be defined with respect to a function on binary matrices, $lof(\cdot)$. This function maps binary matrices to *left-ordered* binary matrices. $lof(\mathbf{Z})$ is obtained by ordering the columns of the binary matrix $\mathbf{Z}$ from left to right by the magnitude of the binary number expressed by that column, taking the first row as the most significant bit. The left-ordering of a binary matrix is shown in Figure 2. In the first row of the left-ordered matrix, the columns for which $z_{1k} = 1$ are grouped at the left. In the second row, the columns for which $z_{2k} = 1$ are grouped at the left of the sets for which $z_{1k} = 1$. This grouping structure persists throughout the matrix.

The *history* of feature k at object i is defined to be $(z_{1k}, \ldots, z_{(i-1)k})$. Where no object is specified, we will use *history* to refer to the full history of feature k, $(z_{1k}, \ldots, z_{Nk})$. We will individuate the histories of features using the decimal equivalent of the binary numbers corresponding to the column entries. For example, at object 3, features can have one of four histories: 0, corresponding to a feature with no previous assignments, 1, being a feature for which $z_{2k} = 1$ but $z_{1k} = 0$, 2, being a feature for which $z_{1k} = 1$ but $z_{2k} = 0$, and 3, being a feature possessed by both previous objects. K_h will denote the number of features possessing the history h, with K_0 being the number of features for which $m_k = 0$ and $K_+ = \sum_{h=1}^{2^N-1} K_h$ being the number of features for which $m_k > 0$, so $K = K_0 + K_+$. This method of denoting histories also facilitates the process of placing a binary matrix in left-ordered form, as it is used in the definition of $lof(\cdot)$.

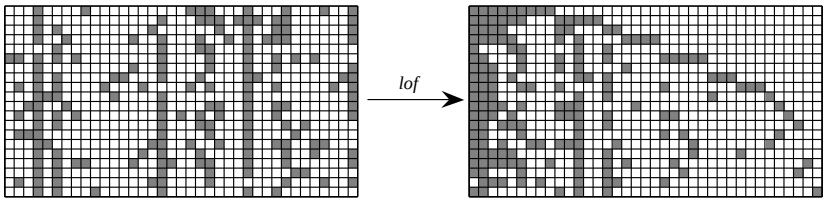

Figure 2: *Binary matrices and the left-ordered form. The binary matrix on the left is transformed into the left-ordered binary matrix on the right by the function $lof(\cdot)$. This left-ordered matrix was generated from the exchangeable Indian buffet process with $\alpha = 10$. Empty columns are omitted from both matrices.*

$lof(\cdot)$ is a many-to-one function: many binary matrices reduce to the same left-ordered form, and there is a unique left-ordered form for every binary matrix. We can thus use $lof(\cdot)$ to define a set of equivalence classes. Any two binary matrices $\mathbf{Y}$ and $\mathbf{Z}$ are lof-equivalent if $lof(\mathbf{Y}) = lof(\mathbf{Z})$, that is, if $\mathbf{Y}$ and $\mathbf{Z}$ map to the same left-ordered form. The lof-equivalence class of a binary matrix $\mathbf{Z}$, denoted $[\mathbf{Z}]$, is the set of binary matrices that are lof-equivalent to $\mathbf{Z}$. lof-equivalence classes are preserved through permutation of either the rows or the columns of a matrix, provided the same permutations are applied to the other members of the equivalence class. Performing inference at the level of lof-equivalence classes is appropriate in models where feature order is not identifiable, with $p(\mathbf{X}|\mathbf{F})$ being unaffected by the order of the columns of $\mathbf{F}$. Any model in which the probability of $\mathbf{X}$ is specified in terms of a linear function of $\mathbf{F}$, such as PCA or CVQ, has this property.

We need to evaluate the cardinality of $[\mathbf{Z}]$, being the number of matrices that map to the same left-ordered form. The columns of a binary matrix are not guaranteed to be unique: since an object can possess multiple features, it is possible for two features to be possessed by exactly the same set of objects. The number of matrices in $[\mathbf{Z}]$ is reduced if $\mathbf{Z}$ contains identical columns, since some re-orderings of the columns of $\mathbf{Z}$ result in exactly the same matrix. Taking this into account, the cardinality of $[\mathbf{Z}]$ is

$$\left(\begin{array}{c} K \\ K_0 \dots K_{2^N-1} \end{array} \right) = \frac{K!}{\prod_{h=0}^{2^N-1} K_h!}, \tag{9}$$

where K_h is the count of the number of columns with full history h.

The binary matrix $\mathbf{Z}$ can be thought of as a generalization of class matrices used in defining mixture models; since each object can only belong to one class, class matrices have the constraint $\sum_k z_{ik} = 1$, whereas the binary matrices in latent feature models do not have this constraint (Griffiths and Ghahramani, 2005).

3.3. *Taking the infinite limit*

Under the distribution defined by Equation 7, the probability of a particular lof-equivalence class of binary matrices, $[\mathbf{Z}]$, is

$$P([\mathbf{Z}]) = \sum_{\mathbf{Z} \in [\mathbf{Z}]} P(\mathbf{Z}) = \frac{K!}{\prod_{h=0}^{2^N-1} K_h!} \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \frac{\alpha}{K})}. \tag{10}$$

In order to take the limit of this expression as $K \to \infty$, we will divide the columns of $\mathbf{Z}$ into two subsets, corresponding to the features for which $m_k = 0$ and the features for which $m_k > 0$. Re-ordering the columns such that $m_k > 0$ if $k \leq K_+$, and $m_k = 0$ otherwise, we can break the product in Equation 10 into two parts, corresponding to these two subsets. The product thus becomes

$$\prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \frac{\alpha}{K})}$$

$$= \left(\frac{\frac{\alpha}{K}\Gamma(\frac{\alpha}{K})\Gamma(N+1)}{\Gamma(N+1+\frac{\alpha}{K})}\right)^{K-K_+} \prod_{k=1}^{K_+} \frac{\frac{\alpha}{K}\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \frac{\alpha}{K})} \tag{11}$$

$$= \left(\frac{\frac{\alpha}{K}\Gamma(\frac{\alpha}{K})\Gamma(N+1)}{\Gamma(N+1+\frac{\alpha}{K})}\right)^{K} \prod_{k=1}^{K_+} \frac{\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(\frac{\alpha}{K})\Gamma(N+1)} \tag{12}$$

$$= \left(\frac{N!}{\prod_{j=1}^{N}(j+\frac{\alpha}{K})}\right)^{K} \left(\frac{\alpha}{K}\right)^{K_+} \prod_{k=1}^{K_+} \frac{(N-m_k)!\prod_{j=1}^{m_k-1}(j+\frac{\alpha}{K})}{N!}. \tag{13}$$

Substituting Equation 13 into Equation 10 and rearranging terms, we can compute our limit

$$\lim_{K\to\infty} \frac{\alpha^{K_+}}{\prod_{h=1}^{2^N-1} K_h!} \cdot \frac{K!}{K_0!\, K^{K_+}} \cdot \left(\frac{N!}{\prod_{j=1}^{N}(j+\frac{\alpha}{K})}\right)^{K} \cdot \prod_{k=1}^{K_+} \frac{(N-m_k)!\prod_{j=1}^{m_k-1}(j+\frac{\alpha}{K})}{N!}$$

$$= \frac{\alpha^{K_+}}{\prod_{h=1}^{2^N-1} K_h!} \cdot \quad 1 \quad \cdot \quad \exp\{-\alpha H_N\} \quad \cdot \prod_{k=1}^{K_+} \frac{(N-m_k)!(m_k-1)!}{N!}, \tag{14}$$

where H_N is the Nth harmonic number, $H_N = \sum_{j=1}^{N} \frac{1}{j}$. The details of the steps taken in computing this limit are given in the appendix of (Griffiths and Ghahramani, 2005). Again, this distribution is exchangeable: neither the number of identical columns nor the column sums are affected by the ordering on objects.

3.4. *The Indian buffet process*

The probability distribution defined in Equation 14 can be derived from a simple stochastic process. This stochastic process provides an easy way to remember salient properties of the probability distribution and can be used to derive sampling schemes for models based on this distribution. This process assumes an ordering on the objects, generating the matrix sequentially using this ordering. Inspired by the Chinese restaurant process (CRP; Aldous, 1985; Pitman, 2002), we will use a culinary metaphor in defining our stochastic process, appropriately adjusted for geography. Many Indian restaurants in London offer lunchtime buffets with an apparently infinite number of dishes. We can define a distribution over infinite binary matrices by specifying a procedure by which customers (objects) choose dishes (features).

In our Indian buffet process (IBP), N customers enter a restaurant one after another. Each customer encounters a buffet consisting of infinitely many dishes

arranged in a line. The first customer starts at the left of the buffet and takes a serving from each dish, stopping after a Poisson(α) number of dishes as his plate becomes overburdened. The ith customer moves along the buffet, sampling dishes in proportion to their popularity, serving himself with probability m_k/i, where m_k is the number of previous customers who have sampled a dish. Having reached the end of all previous sampled dishes, the ith customer then tries a Poisson(α/i) number of new dishes.

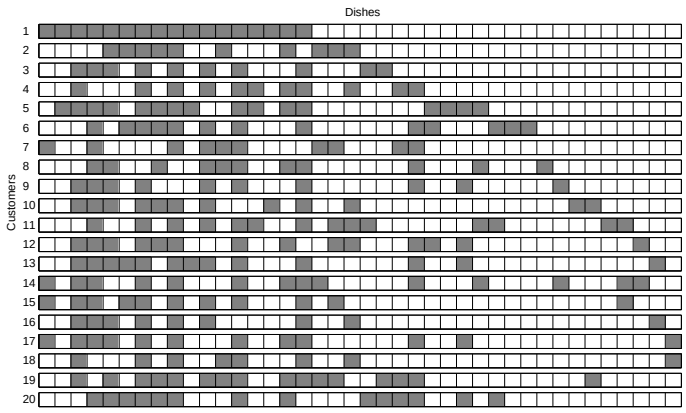

Figure 3: *A binary matrix generated by the Indian buffet process with $\alpha = 10$.*

We can indicate which customers chose which dishes using a binary matrix $\mathbf{Z}$ with N rows and infinitely many columns, where $z_{ik} = 1$ if the ith customer sampled the kth dish. Figure 3 shows a matrix generated using the IBP with $\alpha = 10$. The first customer tried 17 dishes. The second customer tried 7 of those dishes, and then tried 3 new dishes. The third customer tried 3 dishes tried by both previous customers, 5 dishes tried by only the first customer, and 2 new dishes. Vertically concatenating the choices of the customers produces the binary matrix shown in the figure.

Using $K_1^{(i)}$ to indicate the number of new dishes sampled by the ith customer, the probability of any particular matrix being produced by this process is

$$P(\mathbf{Z}) = \frac{\alpha^{K_+}}{\prod_{i=1}^{N} K_1^{(i)}!} \exp\{-\alpha H_N\} \prod_{k=1}^{K_+} \frac{(N-m_k)!(m_k-1)!}{N!}. \tag{15}$$

As can be seen from Figure 3, the matrices produced by this process are generally not in left-ordered form. However, these matrices are also not ordered arbitrarily because the Poisson draws always result in choices of new dishes that are to the right of the previously sampled dishes. Customers are not exchangeable under this distribution, as the number of dishes counted as $K_1^{(i)}$ depends upon the order in which the customers make their choices. However, if we only pay attention to the lof-equivalence classes of the matrices generated by this process, we obtain the exchangeable distribution $P([\mathbf{Z}])$ given by Equation 14: $(\prod_{i=1}^{N} K_1^{(i)}!)/(\prod_{h=1}^{2^N-1} K_h!)$ matrices generated via this process map to the same left-ordered form, and $P([\mathbf{Z}])$ is

obtained by multiplying $P(\mathbf{Z})$ from Equation 15 by this quantity. It is also possible to define a similar sequential process that directly produces a distribution on left-ordered binary matrices in which customers are exchangeable, but this requires more effort on the part of the customers. We call this the *exchangeable IBP* (Griffiths and Ghahramani, 2005).

3.5. *Some properties of this distribution*

These different views of the distribution specified by Equation 14 make it straightforward to derive some of its properties. First, the effective dimension of the model, K_+, follows a Poisson(αH_N) distribution. This is easily shown using the generative process described in previous Section, since under this process K_+ is the sum of Poisson(α), Poisson($\frac{\alpha}{2}$), Poisson($\frac{\alpha}{3}$), etc. The sum of a set of Poisson distributions is a Poisson distribution with parameter equal to the sum of the parameters of its components. Using the definition of the Nth harmonic number, this is αH_N.

A second property of this distribution is that the number of features possessed by each object follows a Poisson(α) distribution. This also follows from the definition of the IBP. The first customer chooses a Poisson(α) number of dishes. By exchangeability, all other customers must also choose a Poisson(α) number of dishes, since we can always specify an ordering on customers which begins with a particular customer.

Finally, it is possible to show that $\mathbf{Z}$ remains sparse as $K \to \infty$. The simplest way to do this is to exploit the previous result: if the number of features possessed by each object follows a Poisson(α) distribution, then the expected number of entries in $\mathbf{Z}$ is $N\alpha$. This is consistent with the quantity obtained by taking the limit of this expectation in the finite model, which is given in Equation 8: $\lim_{K\to\infty} E\left[\mathbf{1}^T\mathbf{Z}\mathbf{1}\right] = \lim_{K\to\infty} \frac{N\alpha}{1+\frac{\alpha}{K}} = N\alpha$. More generally, we can use the property of sums of Poisson random variables described above to show that $\mathbf{1}^T\mathbf{Z}\mathbf{1}$ will follow a Poisson($N\alpha$) distribution. Consequently, the probability of values higher than the mean decreases exponentially.

3.6. *Inference by Gibbs sampling*

We have defined a distribution over infinite binary matrices that satisfies one of our desiderata – objects (the rows of the matrix) are exchangeable under this distribution. It remains to be shown that inference in infinite latent feature models is tractable, as was the case for infinite mixture models. We will derive a Gibbs sampler for latent feature models in which the exchangeable IBP is used as a prior. The critical quantity needed to define the sampling algorithm is the full conditional distribution

$$P(z_{ik} = 1|\mathbf{Z}_{-(ik)}, \mathbf{X}) \propto p(\mathbf{X}|\mathbf{Z})P(z_{ik} = 1|\mathbf{Z}_{-(ik)}), \tag{16}$$

where $\mathbf{Z}_{-(ik)}$ denotes the entries of $\mathbf{Z}$ other than z_{ik}, and we are leaving aside the issue of the feature values $\mathbf{V}$ for the moment. The likelihood term, $p(\mathbf{X}|\mathbf{Z})$, relates the latent features to the observed data, and will depend on the model chosen for the observed data. The prior on $\mathbf{Z}$ contributes to this probability by specifying $P(z_{ik} = 1|\mathbf{Z}_{-(ik)})$.

In the finite model, where $P(\mathbf{Z})$ is given by Equation 7, it is straightforward to compute the full conditional distribution for any z_{ik}. Integrating over π_k gives

$$P(z_{ik} = 1|\mathbf{z}_{-i,k}) = \int_0^1 P(z_{ik}|\pi_k)p(\pi_k|\mathbf{z}_{-i,k})\, d\pi_k = \frac{m_{-i,k} + \frac{\alpha}{K}}{N + \frac{\alpha}{K}}, \tag{17}$$

where $\mathbf{z}_{-i,k}$ is the set of assignments of other objects, not including i, for feature k, and $m_{-i,k}$ is the number of objects possessing feature k, not including i. We need only condition on $\mathbf{z}_{-i,k}$ rather than $\mathbf{Z}_{-(ik)}$ because the columns of the matrix are generated independently under this prior.

In the infinite case, we can derive the conditional distribution from the exchangeable IBP. Choosing an ordering on objects such that the ith object corresponds to the last customer to visit the buffet, we obtain

$$P(z_{ik} = 1|\mathbf{z}_{-i,k}) = \frac{m_{-i,k}}{N}, \tag{18}$$

for any k such that $m_{-i,k} > 0$. The same result can be obtained by taking the limit of Equation 17 as $K \to \infty$. Similarly the number of new features associated with object i should be drawn from a Poisson(α/N) distribution.

4. A TWO-PARAMETER EXTENSION

As we saw in the previous section, the distribution on the number of features per object and on the total number of features are directly coupled, through α. This seems an undesirable constraint. We now present a two-parameter generalization of the IBP which lets us tune independently the average number of features each object possesses and the overall number of features used in a set of N objects. To understand the need for such a generalization, it is useful to examine some samples drawn from the IBP. Figure 4 shows three draws from the IBP with $\alpha = 3$, $\alpha = 10$, and $\alpha = 30$ respectively. We can see that α controls both the number of latent features per object, and the amount of overlap between these latent features (i.e. the probability that two objects will possess the same feature). It would be desirable to remove this restriction, for example so that it is possible to have many latent features but little variability across objects in the feature vectors.

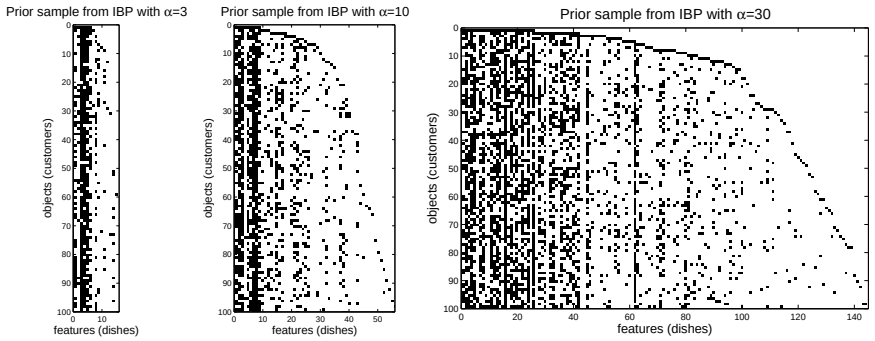

Figure 4: *Draws from the Indian buffet process prior with $\alpha = 3$ (left), $\alpha = 10$ (middle), and $\alpha = 30$ (right).*

Keeping the average number of features per object at α as before, we will define a model in which the overall number of features can range from α (extreme stickiness/herding, where all features are shared between all objects) to $N\alpha$ (extreme repulsion/individuality), where no features are shared at all. Clearly neither of these

extreme cases is very useful, but in general it will be helpful to have a prior where the overall number of features used can be specified.

The required generalization is simple: one takes $r = (\alpha\beta)/K$ and $s = \beta$ in Equation 2. Setting $\beta = 1$ then recovers the one-parameter IBP, but the calculations go through in basically the same way also for other β.

Equation (7), the joint distribution of feature vectors for finite K, becomes

$$P(\mathbf{Z}) = \prod_{k=1}^{K} \frac{B(m_k + \frac{\alpha\beta}{K}, N - m_k + \beta)}{B(\frac{\alpha\beta}{K}, \beta)} \tag{19}$$

$$= \prod_{k=1}^{K} \frac{\Gamma(m_k + \frac{\alpha\beta}{K})\Gamma(N - m_k + \beta)}{\Gamma(N + \frac{\alpha\beta}{K} + \beta)} \frac{\Gamma(\frac{\alpha\beta}{K} + \beta)}{\Gamma(\frac{\alpha\beta}{K})\Gamma(\beta)} \tag{20}$$

The corresponding probability distribution over equivalence classes in the limit $K \to \infty$ is (compare Equation 14):

$$P([\mathbf{Z}]) = \frac{(\alpha\beta)^{K_+}}{\prod_{h\geq 1} K_h!} e^{-\overline{K}_+} \prod_{k=1}^{K_+} B(m_k, N - m_k + \beta) \tag{21}$$

with the constant $\overline{K}_+$ defined below.

As the one-parameter model, this two-parameter model also has a sequential generative process. Again, we will use the Indian buffet analogy. Like before, the first customer starts at the left of the buffet and samples Poisson(α) dishes. The ith customer serves himself from any dish previously sampled by $m_k > 0$ customers with probability $m_k/(\beta + i - 1)$, and in addition from Poisson($\alpha\beta/(\beta + i - 1)$) new dishes. The customer-dish matrix is a sample from this two-parameter IBP. Two other generative processes for this model are described in the Appendix.

The parameter β is introduced above in such a way as to preserve the average number of features per object, α; this result follows from exchangeability and the fact that the first customer samples Poisson(α) dishes. Thus, also the average number of nonzero entries in $\mathbf{Z}$ remains $N\alpha$.

More interesting is the expected value of the overall number of features, i.e. the number K_+ of k with $m_k > 0$. One gets directly from the buffet interpretation, or via any of the other routes, that the expected overall number of features is $\overline{K}_+ = \alpha \sum_{i=1}^{N} \frac{\beta}{\beta+i-1}$, and that the distribution of K_+ is Poisson with this mean. We can see from this that the total number of features used increases as β increases, so we can interpret β as the feature *repulsion*, or $1/\beta$ as the feature *stickiness*. In the limit $\beta \to \infty$ (for fixed N), $\overline{K}_+ \to N\alpha$ as expected from this interpretation. Conversely, for $\beta \to 0$, only the term with $i = 1$ contributes in the sum and so $\overline{K}_+ \to \alpha$, again as expected: in this limit features are infinitely sticky and all customers sample the same dishes as the first one.

For finite β, one sees that the asymptotic behavior of $\overline{K}_+$ for large N is $\overline{K}_+ \sim \alpha\beta \ln N$, because in the relevant terms in the sum one can then approximate $\beta/(\beta + i - 1) \approx \beta/i$. If $\beta \gg 1$, on the other hand, the logarithmic regime is preceded by linear growth at small $N < \beta$, during which $\overline{K}_+ \approx N\alpha$.

We can confirm these intuitions by looking at a few sample matrices drawn from the two-parameter IBP prior. Figure 5 shows three matrices all drawn with $\alpha = 10$, but with $\beta = 0.2$, $\beta = 1$, and $\beta = 5$ respectively. Although all three matrices have

roughly the same number of 1s, the number of features used varies considerably. We can see that at small values of β, features are very sticky, and the feature vector variance is low across objects. Conversely, at high values of β there is a high degree of feature repulsion, with the probability of two objects possessing the same feature being low.

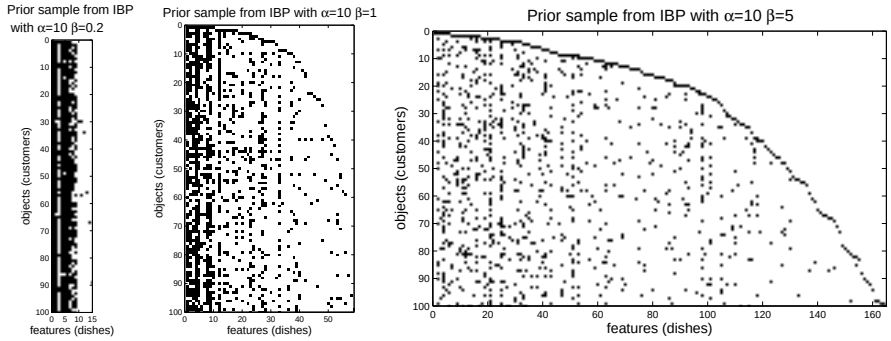

Figure 5: *Draws from the two-parameter Indian buffet process prior with* $\alpha = 10$ *and* $\beta = 0.2$ *(left),* $\beta = 1$ *(middle), and* $\beta = 5$ *(right).*

5. AN ILLUSTRATION

The Indian buffet process can be used as the basis of non-parametric Bayesian models in diverse ways. Different models can be obtained by combining the IBP prior over latent features with different generative distributions for the observed data, $p(\mathbf{X}|\mathbf{Z})$. We illustrate this using a simple model in which real valued data $\mathbf{X}$ is assumed to be linearly generated from the latent features, with Gaussian noise. This linear-Gaussian model can be thought of as a version of factor analysis with binary, instead of Gaussian, latent factors, or as a factorial model (Zemel and Hinton, 1994; Ghahramani 1995) with infinitely many factors.

5.1. *A linear Gaussian model*

We motivate the linear-Gaussian IBP model with a toy problem of modelling simple images (Griffiths and Ghahramani, 2005; 2006). In the model, greyscale images are generated by linearly superimposing different visual elements (objects) and adding Gaussian noise. Each image is composed of a vector of real-valued pixel intensities. The model assumes that there are some unknown number of visual elements and that each image is generated by choosing, for each visual element, whether the image possesses this element or not. The binary latent variable z_{ik} indicates whether image i possesses visual element k. The goal of the modelling task is to discover both the identities and the number of visual elements from a set of observed images.

We will start by describing a finite version of the simple linear-Gaussian model with binary latent features used here, and then consider the infinite limit. In the finite model, image i is represented by a D-dimensional vector of pixel intensities, $\mathbf{x}_i$ which is assumed to be generated from a Gaussian distribution with mean $\mathbf{z}_i\mathbf{A}$ and covariance matrix $\mathbf{\Sigma}_X = \sigma_X^2\mathbf{I}$, where $\mathbf{z}_i$ is a K-dimensional binary vector, and $\mathbf{A}$ is a $K \times D$ matrix of weights. In matrix notation, $E[\mathbf{X}] = \mathbf{ZA}$. If $\mathbf{Z}$ is a feature

matrix, this is a form of binary factor analysis. The distribution of $\mathbf{X}$ given $\mathbf{Z}$, $\mathbf{A}$, and σ_X is matrix Gaussian:

$$p(\mathbf{X}|\mathbf{Z},\mathbf{A},\sigma_X) = \frac{1}{(2\pi\sigma_X^2)^{ND/2}} \exp\{-\frac{1}{2\sigma_X^2}\mathrm{tr}((\mathbf{X}-\mathbf{Z}\mathbf{A})^T(\mathbf{X}-\mathbf{Z}\mathbf{A}))\} \tag{22}$$

where $\mathrm{tr}(\cdot)$ is the trace of a matrix. We need to define a prior on $\mathbf{A}$, which we also take to be matrix Gaussian:

$$p(\mathbf{A}|\sigma_A) = \frac{1}{(2\pi\sigma_A^2)^{KD/2}} \exp\{-\frac{1}{2\sigma_A^2}\mathrm{tr}(\mathbf{A}^T\mathbf{A})\}, \tag{23}$$

where σ_A is a parameter setting the diffuseness of the prior. This prior is conjugate to the likelihood which makes it possible to integrate out the model parameters $\mathbf{A}$.

Using the approach outlined in Section 3.6, it is possible to derive a Gibbs sampler for this finite model in which the parameters $\mathbf{A}$ remain marginalized out. To extend this to the infinite model with $K \to \infty$, we need to check that $p(\mathbf{X}|\mathbf{Z},\sigma_X,\sigma_A)$ remains well-defined if $\mathbf{Z}$ has an unbounded number of columns. This is indeed the case (Griffiths and Ghahramani, 2005) and a Gibbs sampler can be defined for this model.

We applied the Gibbs sampler for the infinite binary linear-Gaussian model to a simulated dataset, $\mathbf{X}$, consisting of 100 6×6 images. Each image, $\mathbf{x}_i$, was represented as a 36-dimensional vector of pixel intensity values[1]. The images were generated from a representation with four latent features, corresponding to the image elements shown in Figure 6 (a). These image elements correspond to the rows of the matrix $\mathbf{A}$ in the model, specifying the pixel intensity values associated with each binary feature. The non-zero elements of $\mathbf{A}$ were set to 1.0, and are indicated with white pixels in the figure. A feature vector, $\mathbf{z}_i$, for each image was sampled from a distribution under which each feature was present with probability 0.5. Each image was then generated from a Gaussian distribution with mean $\mathbf{z}_i\mathbf{A}$ and covariance $\sigma_X\mathbf{I}$, where $\sigma_X = 0.5$. Some of these images are shown in Figure 6 (b), together with the feature vectors, $\mathbf{z}_i$, that were used to generate them.

The Gibbs sampler was initialized with $K_+ = 1$, choosing the feature assignments for the first column by setting $z_{i1} = 1$ with probability 0.5. σ_A, σ_X, and α were initially set to 1.0, and then sampled by adding Metropolis steps to the MCMC algorithm (see Gilks et al., 1996). Figure 6 shows trace plots for the first 1000 iterations of MCMC for the log joint probability of the data and the latent features, $\log p(\mathbf{X},\mathbf{Z})$, the number of features used by at least one object, K_+, and the model parameters σ_A, σ_X, and α. The algorithm reached relatively stable values for all of these quantities after approximately 100 iterations, and our remaining analyses will use only samples taken from that point forward.

The latent feature representation discovered by the model was extremely consistent with that used to generate the data (Griffiths and Ghahramani, 2005). The posterior mean of the feature weights, $\mathbf{A}$, given $\mathbf{X}$ and $\mathbf{Z}$ is

$$E[\mathbf{A}|\mathbf{X},\mathbf{Z}] = (\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_X^2}{\sigma_A^2}\mathbf{I})^{-1}\mathbf{Z}^T\mathbf{X}. \tag{24}$$

[1] This simple toy example was inspired by the "shapes problem" in (Ghahramani, 1995); a larger scale example with real images is presented in (Griffiths and Ghahramani, 2006)

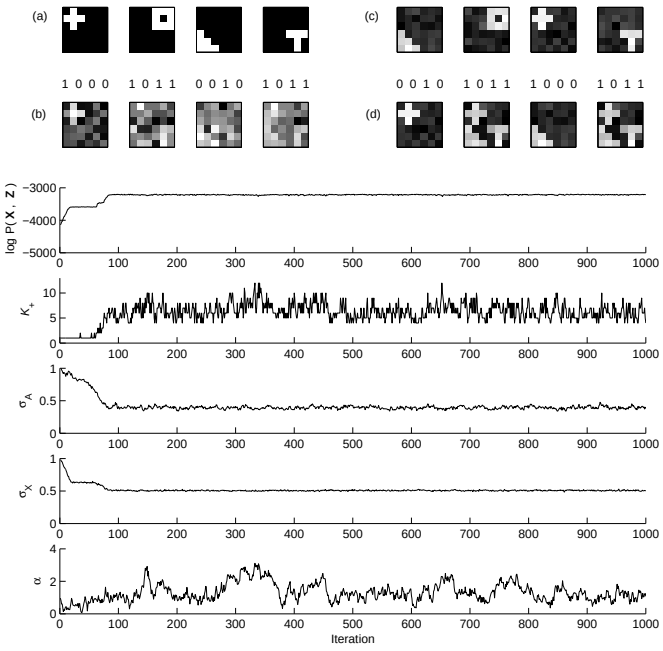

Figure 6: *Stimuli and results for the demonstration of the infinite binary linear-Gaussian model. (a) Image elements corresponding to the four latent features used to generate the data. (b) Sample images from the dataset. (c) Image elements corresponding to the four features possessed by the most objects in the* 1000*th iteration of MCMC. (d) Reconstructions of the images in (b) using the output of the algorithm. The lower portion of the figure shows trace plots for the MCMC simulation, which are described in more detail in the text.*

Figure 6 (c) shows the posterior mean of $\mathbf{a}_k$ for the four most frequent features in the 1000th sample produced by the algorithm, ordered to match the features shown in Figure 6 (a). These features pick out the image elements used in generating the data. Figure 6 (d) shows the feature vectors $\mathbf{z}_i$ from this sample for the four images in Figure 6 (b), together with the posterior means of the reconstructions of these images for this sample, $E[\mathbf{z}_i\mathbf{A}|\mathbf{X},\mathbf{Z}]$. Similar reconstructions are obtained by averaging over all values of $\mathbf{Z}$ produced by the Markov chain. The reconstructions provided by the model clearly pick out the relevant features, despite the high level of noise in the original images.

6. APPLICATIONS

We now outline five applications of the IBP, each of which uses the same prior over infinite binary matrices, $P(\mathbf{Z})$, but different choices for the likelihood relating latent features to observed data, $p(\mathbf{X}|\mathbf{Z})$. These applications will hopefully provide an indication for the potential uses of this distribution.

6.1. *A model for choice behavior*

Choice behavior refers to our ability to decide between several options. Models of choice behavior are of interest to psychology, marketing, decision theory, and computer science. Our choices are often governed by features of the different options. For example, when choosing which car to buy, one may be influenced by fuel efficiency, cost, size, make, etc. Görür et al. (2006) present a non-parametric Bayesian model based on the IBP which, given the choice data, infers latent features of the options and the corresponding weights of these features. The IBP is the prior over these latent features, which are assumed to be binary (either present or absent). Their paper also shows how MCMC inference can be extended from the conjugate IBP models to non-conjugate models.

6.2. *A model for protein interaction screens*

Proteomics aims to understand the functional interactions of proteins, and is a field of growing importance to modern biology and medicine. One of the key concepts in proteomics is a *protein complex*, a group of several interacting proteins. Protein complexes can be experimentally determined by doing high-throughput protein-protein interaction screens. Typically the results of such experiments are subjected to mixture-model based clustering methods. However, a protein can belong to multiple complexes at the same time, making the mixture model assumption invalid. Chu et al. (2006) propose a Bayesian approach based on the IBP for identifying protein complexes and their constituents from interaction screens. The latent binary feature z_{ik} indicates whether protein i belongs to complex k. The likelihood function captures the probability that two proteins will be observed to bind in the interaction screen, as a function of how many complexes they both belong to, $\sum_{k=1}^{\infty} z_{ik} z_{jk}$. The approach automatically infers the number of significant complexes from the data and the results are validated using affinity purification/mass spectrometry experimental data from yeast RNA-processing complexes.

6.3. *A model for the structure of causal graphs*

Wood et al. (2006) use the infinite latent feature model to learn the structure of directed acyclic probabilistic graphical models. The focus of this paper is on learning the graphical models in which an unknown number of hidden variables (e.g. diseases) are causes for some set of observed variables (e.g. symptoms). Rather than defining a prior over the number of hidden causes, Wood et al. use a non-parametric Bayesian approach based on the IBP to model the structure of graphs with countably infinitely many hidden causes. The binary variable z_{ik} indicates whether hidden variable k has a direct causal influence on observed variable i; in other words whether k is a parent of i in the graph. The performance of MCMC inference is evaluated both on simulated data and on a real medical dataset describing stroke localizations.

6.4. *A model for dyadic data*

Many interesting data sets are *dyadic*: there are two sets of objects or entities and observations are made on pairs with one element from each set. For example, the two sets might consist of movies and viewers, and the observations are ratings given by viewers to movies. The two sets might be genes and biological tissues and the observations may be expression levels for particular genes in different tissues. Models of dyadic data make it possible to predict, for example, the ratings a viewer might give to a movie based on ratings from other viewers, a task known as *collaborative*

filtering. A traditional approach to modelling dyadic data is *bi-clustering*: simultaneously cluster both the rows (e.g. viewers) and the columns (e.g. movies) of the observation matrix using coupled mixture models. However, as we have discussed, mixture models provide a very limited latent variable representation of data. Meeds et al. (2007) present a more expressive model of dyadic data based on the infinite latent feature model. In this model, both movies and viewers are represented by binary latent vectors with an unbounded number of elements, corresponding to the features they might possess (e.g. "likes horror movies"). The two corresponding infinite binary matrices interact via a real-valued weight matrix which links features of movies to features of viewers. Novel MCMC proposals are defined for this model which combine Gibbs, Metropolis, and split-merge steps.

6.5. *Extracting features from similarity judgments*

One of the goals of cognitive psychology is to determine the kinds of representations that underlie people's judgments. In particular, a method called "additive clustering" has been used to infer people's beliefs about the features of objects from their judgments of the similarity between them (Shepard and Arabie, 1979). Given a square matrix of judgments of the similarity between N objects, where s_{ij} is the similarity between objects i and j, the additive clustering model seeks to recover a $N \times K$ binary feature matrix $\mathbf{F}$ and a vector of K weights associated with those features such that $s_{ij} \approx \sum_{k=1}^{K} w_k f_{ik} f_{jk}$. A standard problem for this approach is determining the value of K, for which a variety of heuristic methods have been used. Navarro and Griffiths (2007) present a nonparametric Bayesian solution to this problem, using the IBP to define a prior on $\mathbf{F}$ and assuming that s_{ij} has a Gaussian distribution with mean $\sum_{k=1}^{K_+} w_k f_{ik} f_{jk}$. Using this method provides a posterior distribution over the effective dimension of $\mathbf{F}$, K_+, and gives both a weight and a posterior probability for the presence of each feature. Samples from the posterior distribution over feature matrices reveal some surprisingly rich representations expressed in classic similarity datasets.

7. CONCLUSIONS

We have derived a distribution on infinite binary matrices that can be used as a prior for models in which objects are represented in terms of a set of latent features. While we derived this prior as the infinite limit of a simple distribution on finite binary matrices, we also showed that the same distribution can be specified in terms of a simple stochastic process—the Indian buffet process. This distribution satisfies our two desiderata for a prior for infinite latent feature models: objects are exchangeable, and inference via MCMC remains tractable. We described a two-parameter extension of the Indian buffet process which has the added flexibility of decoupling the number of features per object from the total number of features. This prior on infinite binary matrices has been useful in a diverse set of applications, ranging from causal discovery, to choice modelling, and proteomics.

APPENDIX

The generative process for the one-parameter IBP described in Section 3.4, and the process described in Section 4 for the two-parameter model, do not result in matrices which are in left-ordered form. However, as in the one-parameter IBP (Griffiths and Ghahramani, 2005) an exchangeable version can also be defined for the two-

parameter model which produces left-ordered matrices. In the *exchangeable* two-parameter Indian buffet process, the first customer samples a Poisson(α) number of dishes, moving from left to right. The ith customer moves along the buffet, and makes a single decision for each set of dishes with the same history. If there are K_h dishes with history h, under which m_h previous customers have sampled each of those dishes, then the customer samples a Binomial($m_h/(\beta + i - 1), K_h$) number of those dishes, starting at the left. Having reached the end of all previously sampled dishes, the ith customer then tries a Poisson($\alpha\beta/(\beta + i - 1)$) number of new dishes. Attending to the history of the dishes and always sampling from the left guarantees that the resulting matrix is in left-ordered form, and the resulting distribution over matrices is exchangeable across customers.

As in the one-parameter IBP, the generative process for the two-parameter model also defines a probability distribution directly over the feature histories (c.f. Section 4.5 of Griffiths and Ghahramani, 2005). Recall that the history of feature k is the vector $(z_{1k}, \ldots z_{Nk})$, and that for each of the 2^N possible histories h, K_h is the number of features possessing that history. In this two-parameter model, the distribution of K_h (for $h > 0$) is Poisson with mean $\alpha\beta B(m_h, N - m_h + \beta) = \alpha\beta\Gamma(m_h)\Gamma(N - m_h + \beta)/\Gamma(N + \beta)$.

REFERENCES

Aldous, D. (1985) Exchangeability and related topics. *École d'été de probabilités de Saint-Flour, XIII*, Lecture Notes in Mathematics **1117**, (A. Dold and B. Eckmann, eds.) Berlin: Springer, 1–198.

Antoniak, C. (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**:1152–1174.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian theory.* New York: Wiley.

Blei, D. M., Griffiths, T.L., Jordan, M.I, Tenenbaum, J. (2004) Hierarchical topic models and the nested Chinese restaurant process. *Advances in Neural Information Processing Systems* **16**.

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

Chu, W., Ghahramani, Z., Krause, R., and Wild, D.L. (2006) Identifying Protein Complexes in High-Throughput Protein Interaction Screens using an Infinite Latent Feature Model. *BIOCOMPUTING 2006: Proceedings of the Pacific Symposium* (Altman et al., eds.) **11**:231–242.

d'Aspremont, A., El Ghaoui, L. E., Jordan, M. I., Lanckriet, G. R. G. (2005). A Direct Formulation for Sparse PCA using Semidefinite Programming. *Advances in Neural Information Processing Systems* **17**, Cambridge, MA: MIT Press.

Escobar, M. D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90**:577-588.

Ferguson, T. S. (1983) Bayesian density estimation by mixtures of normal distributions. *Recent advances in statistics* (M. Rizvi, J. Rustagi, and D. Siegmund, eds.) New York: Academic Press, 287–302

Ghahramani, Z. (1995) Factorial learning and the EM algorithm. *Advances in Neural Information Processing Systems* **7**. (G. Tesauro, D. S. Touretzky and T. K. Leen, eds.) San Francisco, CA: Morgan Kaufmann, 617–624.

Gilks, W., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov chain Monte Carlo in practice.* Suffolk: Chapman and Hall.

Görür, D., Jäkel, F. and Rasmussen, C. (2006) A Choice Model with Infinitely Many Latent Features. *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006).* (W. W. Cohen and A. Moore, eds.) 361–368.

Green P.J., and Richardson S. (2001) Modelling heterogeneity with and without the Dirichlet process. *Scand J Stat.* **28**: 355–375.

Griffiths, T.L. and Ghahramani, Z. (2005) *Infinite Latent Feature Models and the Indian Buffet Process.* Gatsby Unit Technical Report GCNU-TR-2005-001.

Griffiths, T.L., and Ghahramani, Z. (2006) Infinite Latent Feature Models and the Indian Buffet Process. *Advances in Neural Information Processing Systems* **18** (Y. Weiss, B. Schölkopf, J. Platt, eds.) Cambridge, MA: MIT Press, 475–482

Jolliffe, I. T. (1986) *Principal component analysis.* New York: Springer.

Jolliffe, I. T. and Uddin, M. (2003) A modified principal component technique based on the lasso. *J. Comp. Graphical Statist.* **12**: 531-547.

Meeds, E., Ghahramani, Z., Neal, R. and Roweis, S.T. (2007) Modeling Dyadic Data with Binary Latent Factors. *Advances in Neural Information Processing Systems* **19**. (B. Schölkopf, J. Platt and T. Hoffman, eds.) Cambridge, MA: MIT Press.

Navarro, D. J. & Griffiths, T. L. (2007). A nonparametric Bayesian method for inferring features from similarity judgments. *Advances in Neural Information Processing Systems* **19**. (B. Schölkopf, J. Platt and T. Hoffman, eds.) Cambridge, MA: MIT Press.

Neal, R. M. (2000) Markov chain sampling methods for Dirichlet process mixture models. *J. Comp. Graphical Statist.* **9**: 249-265.

Pitman, J. (2002) *Combinatorial stochastic processes.* Notes for Saint Flour Summer School. Technical Report 621, Dept. Statistics, U.C. Berkeley.

Rasmussen, C. (2000) The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems* **12**. (S. A. Solla, T. K. Leen and K.-R. Muller, eds.) Cambridge, MA: MIT Press, 554–560.

Roweis, S.T. and Ghahramani, Z. (1999) A unifying review of linear Gaussian models. *Neural Computation*, **11**(2): 305-345.

Shepard, R.N. and Arabie, P. (1979). Additive clustering representations of similarities as combinations of discrete overlapping properties. *Psychological Review*, **86**(2), 87–123.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006) Hierarchical Dirichlet Processes. *J. Amer. Statist. Assoc.* **101**(476):1566–1581.

Ueda, N. and Saito, K. (2003) Parametric mixture models for multi-labeled text. In *Advances in Neural Information Processing Systems* **15.** (S. Becker, S. Thrun and K. Obermayer, eds.) Cambridge: MIT Press.

Wood, F., Griffiths, T. L., & Ghahramani, Z. (2006). A non-parametric Bayesian method for inferring hidden causes. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI 2006).* AUAI Press, 536–543.

Zemel, R. S. and Hinton, G. E. (1994) Developing population codes by minimizing description length. *Advances in Neural Information Processing Systems* **6.** (J. D. Cowan, G. Tesauro and J. Alspector, eds.) San Francisco, CA: Morgan Kaufmann, 3–10.

Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal component analysis. *J. Comp. Graphical Statist.* **15**(2):265–286.

Discussion of "Bayesian Nonparametric Latent Feature Models" by Zoubin Ghahramani

David B. Dunson

Biostatistics Branch, MD A3-03, National Institute of Environmental Health Sciences

Research Triangle Park, NC 27709, USA

1 Brief Comments

Ghahramani and colleagues have proposed an interesting class of infinite latent feature (ILF) models. The basic premise of ILF models is that there are infinitely many latent predictors represented in the population, with any particular subject having a finite selection. This is presented as an important advance over models that allow a finite number of latent variables. ILF models are most useful when all but a few of the features are very rare, so that one obtains a *sparse* representation. Otherwise, one cannot realistically hope to learn about the latent feature structure from the available data. The utility of sparse latent factor models has been compellingly illustrated in large p, small n problems by West (2003) and Carvalho et al. (2006). Given that performance is best when the number of latent features represented in the sample is much less than the sample size, it is not clear whether there are practical advantages to the ILF formulation over finite latent variable models that allow uncertainty in the dimension. For example, Lopes and West (2004) and Dunson (2006) allow the number of latent factors to be unknown using Bayesian methods.

That said, it is conceptually appealing to allow additional features to be represented in the data set as additional subjects are added, and it is also appealing to allow partial clustering of subjects. In particular, under an ILF model, subjects can have some features in common, leading to a degree of similarity based on the number of shared features and the

values of these features.

Following the notation of Ghahramani et al., the $K \times 1$ latent feature vector for subject i is denoted $\mathbf{f}_i = (f_{i1}, \ldots, f_{iK})'$, with $f_{ik} = z_{ik}v_{ik}$, where $z_{ik} = 1$ if subject i has feature k and $z_{ik} = 0$ otherwise, and v_{ik} is the value of the feature. There are then two important aspects of the specification for an infinite latent feature model: (1) the prior on the $N \times K$ binary matrix $\mathbf{Z} = \{z_{ik}\}$, with $K \to \infty$; and (2) the prior on the $N \times K$ matrix $\mathbf{V} = \{v_{ik}\}$.

The focus of Ghahramani et al. is on the prior for $\mathbf{Z}$, proposing an Indian Buffet Process (IBP) specification. The IBP follows in a straightforward but elegant manner from the following assumptions: (i) the elements of $\mathbf{Z}$ are independent and Bernoulli distributed given π_k, the probability of occurrence of the kth feature; and (ii) $\pi_k \sim \text{beta}(\alpha/K, 1)$. Because the features are treated as exchangeable in this specification, it is necessary to introduce a left ordering function, so that it is possible to base inference on a finite approximation focusing only on the more common features.

In this discussion, I briefly consider the more general problem of nonparametric modeling of both $\mathbf{Z}$ and $\mathbf{V}$, proposing an exponentiated gamma Dirichlet process (EGDP) prior. The exponentiated gamma (EG) is used as an alternative to the IBP, with some advantages, while the Dirichlet process (DP) (Ferguson, 1973; 1974) is used for nonparametric modeling of the feature scores among subjects possessing a feature.

2 Exponentiated Gamma Dirichlet Process

To provide motivation, I focus on an epidemiologic application in which an ILF model is clearly warranted. In the Agricultural Health Study (Kamel et al., 2005), interest focused on studying factors contributing to neurological symptom (headaches, dizziness, etc) occurrence in farm workers. Individual i is asked through a questionnaire to record the frequency of symptom occurrence for p different symptom types, resulting in the response vector,

$\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})'$. It is natural to suppose that the symptom frequencies, $\mathbf{y}_i$, provide measurements of latent features, $\mathbf{f}_i = (f_{i1}, \ldots, f_{iK})'$. Here, $f_{ik} = z_{ik}v_{ik}$, with $z_{ik} = 1$ if individual i has latent risk factor k and 0 otherwise, while v_{ik} denotes the severity of risk factor k for individual i. For example, feature k may represent the occurrence of an undiagnosed mild stroke, while v_{ik} represents how severe the stroke is, with more severe stroke resulting in more frequent neurological problems.

Such data would not be well characterized with a typical latent class model, which requires individuals to be grouped into a single set of classes. However, the approach of Ghahramani et al. is also not ideal in this case, as there are two important drawbacks. First, the assumption of feature exchangeability makes inferences on the latent features awkward. Thus, across posterior samples collected using an MCMC algorithm, the feature index changes meaning. This label ambiguity also also occurs in DPM models. A solution in the setting of ILF models is to choose a prior that explicitly orders the features by their frequency of occurrence, with feature one being the most common. Second, one can potentially characterize the data using fewer features by modeling the feature scores $\{v_{ik}\}$ nonparametrically. This also provides a more realistic characterization of the data. By assuming a parametric model, one artificially inflates the number of features needed to fit the data, making the latent features less likely to characterize a true unobserved risk factor.

An exponentiated gamma Dirichlet process (EGDP) prior can address both of these issues. I first define the exponentiated gamma (EG) component of the prior, which provides a probability model for the random matrix, $\mathbf{Z}$. Without loss of generality, the features are ordered, so that the first trait tends to be more common in the population, and the features decrease stochastically in population frequency with increasing index h. This is accomplished by letting

$$\pi_h = 1 - \exp(-\gamma_h), \quad \gamma_h \overset{ind}{\sim} \mathcal{G}(1, \beta_h), \quad \text{for } h = 1, \ldots, \infty, \tag{1}$$

where $\boldsymbol{\gamma} = \{\gamma_h, h = 1, \ldots, \infty\}$ is a stochastically decreasing infinite sequence of independent gamma random variables, with the stochastic decreasing constraint ensured by letting $\beta_1 < \beta_2 < \ldots < \beta_\infty$. Marginalizing over the prior for $\boldsymbol{\gamma}$, we obtain

$$\begin{aligned} \Pr(Z_{ih} = 1 \,|\, \boldsymbol{\beta}) &= 1 - \int_0^\infty \exp(-\gamma_h)\, \beta_h\, \exp(-\gamma_h \beta_h)\, d\gamma_h \\ &= \frac{1}{1+\beta_h}, \end{aligned} \tag{2}$$

which is decreasing in h for increasing $\boldsymbol{\beta} = \{\beta_h, h = 1, \ldots, \infty\}$.

Note that, unlike for the IBP, the exponentiated gamma (EG) process defined in (1) does not result in a Poisson distribution for $S_i = \sum_{h=1}^{\infty} Z_{ih}$, the number of traits per subject. Instead S_i is defined as the convolution of independent but not identically distributed Bernoulli random variables. A convenient special case corresponds to

$$\beta_h = \exp\{\psi_1 + \psi_2(h-1)\}, \quad h = 1, 2, \ldots, \infty, \tag{3}$$

which results in a logistic regression model for the frequency of trait occurrence upon marginalizing out $\boldsymbol{\gamma}$. In this case, two hyperparameters, ψ_1 and ψ_2, characterize the EG process, with ψ_1 controlling the frequency of trait one and ψ_2 controlling how rapidly traits decrease in frequency with the index h. The restriction $\psi_2 > 0$ ensures that $\beta_1 < \beta_2 < \ldots < \beta_\infty$. Assuming (1) and (3), it is straightforward to show that the distribution of S_i can be accurately approximated by the distribution of $S_{iT} = \sum_{i=1}^{T} Z_{ih}$ for sufficiently large T. In most applications, a sparse representation with few dominant features (expressed by choosing $\psi \geq 1$) may be preferred. In such cases, an accurate truncation approximation can be produced by replacing $\mathbf{F} = \mathbf{Z} \bigotimes \mathbf{V}$ with $\mathbf{F}_T = \mathbf{Z}_T \bigotimes \mathbf{V}_T$, with $\bigotimes$ denoting the element-wise product, and $\mathbf{A}_T$ denoting the submatrix of $\mathbf{A}$ consisting of the first T columns. Here, T is a finite integer, e.g., $T = 20$ or $T = 50$.

Expressions (1) and (3) provide a prior for the random binary matrix, $\mathbf{Z}$, allocating features to subjects. In order to complete the EGDP specification, we let $v_{ih} = 0$ if $z_{ih} = 0$

and otherwise

$$(v_{ih} \,|\, z_{ih} = 1) \sim G_h, \quad G_h \sim DP(\alpha G_0). \tag{4}$$

Here, G_h represents a random probability measure characterizing the distribution of the hth latent feature score among those individuals with the feature. This probability measure is drawn from a Dirichlet process (DP) with base measure G_0 and precision α.

3 Nonparametric Latent Factor Models

To illustrate the EGDP, we focus on a nonparametric extension of factor analysis. For subjects $i = 1, \ldots, n$, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})'$ denote a multivariate response vector. Then, a typical factor analytic model can be expressed as:

$$\mathbf{y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}), \tag{5}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ is a mean vector, $\boldsymbol{\Lambda}$ is a $p \times K$ factor loadings matrix, $\mathbf{f}_i = (f_{i1}, \ldots, f_{iK})'$ is a $K \times 1$ vector of latent factors, and $\boldsymbol{\epsilon}_i$ is a normal residual with diagonal covariance $\boldsymbol{\Sigma}$ (see, for example, Lopes and West, 2004). In a parametric specification, one typically assumes $f_{ih} \sim N(0, 1)$, while constraining the factor loadings matrix $\boldsymbol{\Lambda}$ to ensure identifiability.

Instead we let $\mathbf{f}_i \sim F$, with $F \sim EGDP(\boldsymbol{\psi}, \alpha, G_0)$, where F denotes the unknown distribution of $\mathbf{f}_i$ and $EGDP(\boldsymbol{\psi}, \alpha, G_0)$ is shorthand notation for the exponentiated gamma Dirichlet process prior with hyperparameters $\boldsymbol{\psi} = (\psi_1, \psi_2)'$, α and G_0. Due to the constraint that the higher numbered factors correspond to rarer features that are less frequent in the population, we avoid the need to constrain $\boldsymbol{\Lambda}$. To remove sign ambiguity, we instead restrict G_0 to have strictly positive support, ensuring that $f_{ih} \geq 0$ for all i, h.

Note that this characterization has several appealing properties. First, the distributions of the factor scores are modelled nonparametrically, with subjects automatically clustered

into groups separately for each factor. One of these groups corresponds to the cluster of subjects not having the factor, while the others are formed through the discreteness property of the DP. Second, the formulation automatically allows an unknown number of factors represented among the subjects in the sample. Thus, uncertainty in the number of factors is accommodated in a very different manner from Lopes and West (2004). Third, for G_0 chosen to be truncated normal, posterior computation can proceed efficiency via a data augmentation MCMC algorithm. Using a truncation approximation (say with $T = 20$), the algorithm alternately updates: (i) $\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}$ conditionally on $\mathbf{F}$ using Gibbs sampling steps; (ii) the elements of $\mathbf{Z}$ by sampling from the Bernoulli full conditional posterior distributions; (iii) $\{\gamma_h, h = 1, \ldots, T\}$ with a data augmentation step (relying on an approach similar to Dunson and Stanford, 2005 and Holmes and Held, 2006); (iv) $\mathbf{V}$ using standard algorithms for DPMs (MacEachern and Müller, 1998). Details are excluded given space considerations.

References

Carvalho, C.M., Lucas, J., Wang, Q., Nevins, J. and West, M. (2005) High-dimensional sparse factor models and latent factor regression. *ISDS Discussion Paper*, Duke University.

Dunson, D.B. (2006) Efficient Bayesian model averaging in factor analysis. *ISDS Discussion Paper*, Duke University.

Dunson, D.B. and Stanford, J.B. (2005) Bayesian inferences on predictors of conception probabilities. *Biometrics*, 61, 126-133.

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

Ferguson, T.S. (1974) Prior distributions on spaces of probability measures. *The Annals of Statistics*, **2**, 615-629.

Holmes, C.C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* **1**, 145-168.

Lopes, H.F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14**, 41-67.

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

West, M. (2003) Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics* **7**, 723-732.

Rejoinder for "Bayesian Nonparametric Latent Feature Models"

Z. Ghahramani, P. Sollich, and T. L. Griffiths

February 16, 2007

We thank Dr. Dunson for a stimulating discussion of our paper. In his discussion, Dunson makes several comments about our paper, and then proposes an alternative approach to sparse latent feature modelling. We first address his comments, and then turn to his suggested approach.

The first comment is that although the utility of sparse latent factor models has been illustrated by West and colleagues, it is not clear whether there are practical advantages to allowing the number of latent factors to be unbounded, as in our approach, as opposed to defining a model with a finite but unknown number of latent factors.

There are two advantages, we believe, one philosophical and one practical. The philosophical advantage is what motivates the use of nonparametric Bayesian methods in the first place: If we don't really believe that the data was actually generated from a finite number of latent factors, then we should not put much or any of our prior mass on such hypotheses. It is hard to think of many real-world generative processes for data in which one can be confident that there are some small number of latent factors. On the practical side, a finite model with an unknown number of latent factors may be preferable to an infinite model if there were significant computational advantages to assuming the finite model. However, inference in finite models of unknown dimension is in fact more computationally demanding, due to the variable dimensionality of the parameter space. Our experience comparing sampling from the infinite model and using Reversible Jump MCMC to sample from an analogous finite but variable-dimension model suggests that the sampler for the infinite model is both easier to implement and faster to mix (Wood et al, 2006).

Dunson also states that for West and colleagues "performance is best when the number of latent features represented in the sample is much less than the sample size". However, West's (2003) model is substantially different from ours; it is essentially a linear Gaussian factor analysis model with a sparse prior on the factor loading matrix, while our infinite latent feature models

can be used in many different contexts and allow the factors themselves to be sparse. We do not feel that the results that West reports on a particular application and choice of model specification can be generalized to Bayesian inference in all sparse models with latent features.

A second comment is that the assumption of feature exchangeability makes inference in the latent feature space awkward. This is a similar problem to the one suffered by Dirichlet process mixture (DPM) models where feature indices can change across samples in an MCMC run. We agree that questions such as "what does latent feature k represent" are meaningless in models with exhangeable features. We would never really be interested in such questions. However, there are plenty of meaningful inferences that can be derived from such a model, such as asking how many latent features two data points share. Rather than looking at averages of $\mathbf{Z}$ across MCMC runs, which makes no sense in an model with exchangeable features, one can look at averages of the $N \times N$ matrix $\mathbf{Z}\mathbf{Z}^T$, whose elements measure the number of latent features two data points share. Dunson's proposed solution, a prior that explicitly orders features by their frequency of occurence, is interesting but probably not enough to ensure that meaningful inferences can be made about $\mathbf{Z}$. For example, if two latent features have approximately the same frequency across the data, then any reasonably well-mixing sampler will frequently permute their labels, again muddling inferences about $\mathbf{Z}$ and the parameters associated with the two latent features.

A third comment made by Dunson is that one can define a more flexible model by having a non-parametric model for the features scores v_{ik}, rather than a parametric model. We entirely agree with this last point, and we did not intend to imply that v needs to come from a parametric model. A non-parametric model for v_{ik}, for example based on the Dirichlet process, is potentially very desirable in certain contexts. One possible disadvantage of such a model is that it requires additional bookkeeping and computation in an MCMC implementation. For certain parametric models for v_{ik}, one can analytically integrate out the $\mathbf{V}$ matrix, making the MCMC sampler over other variables mix faster.

We now turn to the proposed exponentiated gamma Dirichlet process (EGDP). This is an interesting model, well worth further study and elaboration.

Our first comment on this model is that the γ_h random variables defined in equation (1) of the discussion are rather unnecessary. Pushing through the transformation of variables, we can compute the distribution on π_h implied by assuming that γ_h follows a particular distribution. In the case of the exponentiated gamma model, this gives $\pi_h \sim \text{Beta}(1, \beta_h)$. This leads us to the question of why this way around and not, e.g., $\text{Beta}(\alpha_h, 1)$? The latter would be a more natural way to generalize our $\pi_h \sim \text{Beta}(\alpha/K, 1)$ to have non-

exchangeable latent features. In this proposal, the α_h would get smaller for $h \to \infty$, with the mean frequency for feature h being $\frac{\alpha_h}{\alpha_h+1}$. Writing both models in terms of their Beta distributions over feature frequencies highlights the similarities and differences between the two proposals. The choice $\text{Beta}(\alpha_h, 1)$ provides an alternative method for producing sparseness. Of course one could also look at $\text{Beta}(\alpha_h\beta, \beta)$, to generalize our two-parameter model.

Making the features inequivalent is attractive in some respects, but on the other hand may reduce flexibility. With exponentially decreasing β's, the higher index features will be so strongly suppressed that they will be hard to "activate" even with large amounts of data.

For the factor model in equation (5) of the discussion, we disagree that making the f's all positive is necessarily a good thing—one then models data that lie in a (suitably affinely transformed) octant of the space spanned by the columns of $\mathbf{L}$, rather than in the whole space. This is not merely a method for fixing a sign indeterminacy, but makes quite a different assumption about the data than in an ordinary factor analysis model. This model with positive factors is similar to a large body of work on non-negative matrix factorization models (e.g. Paatero and Tapper, 1994; Lee and Seung 1999).

To summarize, we thank Dr. Dunson for his interesting discussion and we hope that our work, his discussion, and this rejoinder will stimulate further work on sparse latent feature models.

REFERENCES

Lee, D. D. and Seung, H. S. (1999) Learning the parts of objects by non-negative matrix factorization, *Nature*, **401**(6755):788–791.

Paatero, P. and U. Tapper (1994) Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values, *Environmetrics* **5**:111–126.