Supplementary Materials for *Distance Dependent Infinite Latent Feature Models*

Samuel J. Gershman[1], Peter I. Frazier[2] and David M. Blei[3]
[1] Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology
[2] School of Operations Research and Information Engineering,
Cornell University
[3] Department of Computer Science,
Princeton University

In Section 1, we present proofs of the propositions and lemmas that appeared in the main paper. In Section 2, we present a Markov chain Monte Carlo algorithm for approximate inference. Finally, in Section 3, we present analysis of data in which a non-exchangeable model might be expected to help, but does not.

1 Proofs

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

Proposition 1

Under the dd-IBP,

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

$$R_{ij} \sim \text{Poisson}\left(\alpha \sum_{n=1}^{N} h_n^{-1} P(\mathcal{L}_{in} = 1, \mathcal{L}_{jn} = 1)\right). \tag{2}$$

Proof. Recall that the number of features owned by each customer is $\lambda_n \sim \text{Poisson}(\mu_n)$, where $\mu_n = \alpha h_n^{-1}$. The total number of features across all data points is $K = \sum_{n=1}^{N} \lambda_n \sim \text{Poisson}(\mu)$, where $\mu = \sum_{n=1}^{N} \mu_n = \alpha \sum_{n=1}^{N} h_n^{-1}$.

Let π be a uniform (conditioned on K) random permutation of $\{1,\ldots,K\}$. Then, since $z_{ik}=0$ for all $k>K$ under the dd-IBP,

$$R_i=\sum_{k=1}^{K} z_{ik}=\sum_{k=1}^{K} z_{i,\pi(k)}, \qquad R_{ij}=\sum_{k=1}^{K} z_{ik}z_{jk}=\sum_{k=1}^{K} z_{i,\pi(k)}z_{j,\pi(k)}.$$

Conditioned on K, for $k\leq K$, the event $\{z_{i,\pi(k)}=1\}$ can be rewritten

$$\{z_{i,\pi(k)}=1\}=\bigcup_{n=1}^{N}\left(\{c^*_{\pi(k)}=n\}\cap\{\mathcal{L}_{i,n,\pi(k)}=1\}\right), \tag{3}$$

since feature $\pi(k)$ is on for customer i if and only if customer i links to the owner of that feature.

Conditioned on K, the collection of random variables $\{c^*_{\pi(k)}:k\leq K\}$ are iid (independent and identically distributed), with $P(c^*_{\pi(k)}=n|K)=\mu_n/\mu=h_n^{-1}/\sum_{j=1}^{N}h_j^{-1}$ for $k\leq K$. Conditioned on K, the random variables $\mathcal{L}_{i,n,\pi(k)}$ are also iid across k. Thus, by (3), and still conditioned on K, the collection of Bernoulli random variables $\{z_{i,\pi(k)}:k\leq K\}$ are iid with

$$\begin{aligned}P\left(z_{i,\pi(k)}=1\middle|\,K\right)&=\sum_{n=1}^{N}P\left(c^*_{\pi(k)}=n|K\right)P\left(\mathcal{L}_{i,n,\pi(k)}=1|K\right)\\&=\sum_{n=1}^{N}(\mu_n/\mu)P\left(\mathcal{L}_{in}=1\right)=q_i\end{aligned}$$

for $k\leq K$, where we have dropped the subscript $\pi(k)$ in $\mathcal{L}_{i,n,\pi(k)}$ in the last line because $P\left(\mathcal{L}_{i,n,\pi(k)}=1\right)$ does not depend on k, and we have dropped the conditioning on K because links are drawn independenly of K. In the last line, we have defined a quantity q_i, noting that this quantity does not depend on K.

Thus, R_i is a sum of a Poisson(μ) number of conditionally independent Bernoulli(q_i) random variables, and so is itself Poisson distributed with mean

$$\mu q_i=\sum_{n=1}^{N}\mu_n P\left(\mathcal{L}_{in}=1\right)=\alpha\sum_{n=1}^{N}h_n^{-1}P\left(\mathcal{L}_{in}=1\right).$$

This shows the statement (1).

The proof of the statement (2) is similar. Conditioned on K,

$$\{z_{i,\pi(k)}z_{j,\pi(k)}=1\}=\bigcup_{n=1}^{N}\left(\{c^*_{\pi(k)}=n\}\cap\{\mathcal{L}_{i,n,\pi(k)}=1,\mathcal{L}_{j,n,\pi(k)}=1\}\right)$$

and so the collection of Bernoulli random variables $\{z_{i,\pi(k)}z_{j,\pi(k)}:k\leq K\}$ are iid with

$$\begin{aligned}P\left(z_{i,\pi(k)}z_{j,\pi(k)}=1\middle|\,K\right)&=\sum_{n=1}^{N}P\left(c^*_{\pi(k)}=n|K\right)P\left(\mathcal{L}_{i,n,\pi(k)}=1,\mathcal{L}_{j,n,\pi(k)}=1|K\right)\\&=\sum_{n=1}^{N}(\mu_n/\mu)P\left(\mathcal{L}_{in}=1,\mathcal{L}_{jn}=1\right)=q_{ij}\end{aligned}$$

for $k \leq K$. Thus, R_{ij} is a sum of a Poisson(μ) number of conditionally independent Bernoulli(q_{ij}) random variables, and so is itself Poisson distributed with mean

$$\mu q_{ij} = \sum_{n=1}^{N} \mu_n P\left(\mathcal{L}_{in} = 1, \mathcal{L}_{jn} = 1\right) = \alpha \sum_{n=1}^{N} h_n^{-1} P\left(\mathcal{L}_{in} = 1, \mathcal{L}_{jn} = 1\right).$$

□

Lemma used in proofs of Propositions 2 and 4

The proofs of Propositions 2 and 4 rely on the following lemma.

Lemma 1. *Let* $X_\alpha \sim \text{Poisson}(\alpha\lambda_X)$ *and* $Y_\alpha \sim \text{Poisson}(\alpha\lambda_Y)$ *with* $\lambda_Y > 0$. *Then, as* $\alpha \to \infty$,

$$X_\alpha/\alpha \xrightarrow{P} \lambda_X,$$
$$Y_\alpha/\alpha \xrightarrow{P} \lambda_Y,$$
$$X_\alpha/Y_\alpha \xrightarrow{P} \lambda_X/\lambda_Y.$$

Proof. Pick any $\epsilon > 0$. $\mathbb{E}[\frac{1}{\alpha}X_\alpha] = \lambda_X$ and so Chebyshev's inequality implies

$$P\left[\left|\frac{1}{\alpha}X_\alpha - \lambda_X\right| > \epsilon\right] \leq \frac{\text{Var}[\frac{1}{\alpha}X_\alpha]}{\epsilon^2} = \frac{1}{\alpha\epsilon^2},$$

which converges to 0 as $\alpha \to \infty$. Thus, X_α/α converges in probability to λ_X. By a similar argument, Y_α/α converges in probability to λ_Y.

Then, by the continuous mapping theorem, the supposition $\lambda_Y > 0$, and the continuity of the mapping $(x, y) \mapsto x/y$ over $\mathbb{R} \times (0, \infty)$, we have that $X_\alpha/Y_\alpha = (X_\alpha/\alpha)/(Y_\alpha/\alpha)$ converges in probability to λ_X/λ_Y. □

Proposition 2

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

$$\frac{R_i}{\alpha} \xrightarrow{P} \sum_{n=1}^{N} h_n^{-1} P(\mathcal{L}_{in} = 1), \tag{4}$$

$$\frac{R_{ij}}{\alpha} \xrightarrow{P} \sum_{n=1}^{N} h_n^{-1} P(\mathcal{L}_{in} = 1, \mathcal{L}_{jn} = 1), \tag{5}$$

$$\frac{R_{ij}}{R_i} \xrightarrow{P} \frac{\sum_{n=1}^{N} h_n^{-1} P(\mathcal{L}_{in} = 1, \mathcal{L}_{jn} = 1)}{\sum_{n=1}^{N} h_n^{-1} P(\mathcal{L}_{in} = 1)}. \tag{6}$$

Proof. The proof follows Proposition 1, Lemma 1, and the fact that

$$\sum_{n=1}^{N} h_n^{-1} P(\mathcal{L}_{in} = 1) \geq h_i^{-1} P(\mathcal{L}_{ii} = 1) = h_i^{-1} > 0.$$

□

Proposition 3

If B_0 is continuous, then under the dHBP,

$$R_i | \mathbf{g}_{1:N} \sim \text{Poisson}\left(\gamma\right), \tag{7}$$

$$R_{ij} | \mathbf{g}_{1:N} \sim \begin{cases} \text{Poisson}\left(\gamma \frac{c_0+c_1+1}{(c_0+1)(c_1+1)}\right) & \text{if } g_i = g_j, \\ \text{Poisson}\left(\gamma \frac{1}{c_0+1}\right) & \text{if } g_i \neq g_j. \end{cases} \tag{8}$$

Proof. We write the random measures B and B_j^* in the generative model defining the dHBP in Section 3.2 as the following mixtures over point masses.

$$B = \sum_{k=1}^{\infty} p_k \delta_{\omega_k}, \qquad p_k \sim \text{Beta}(0, c_0), \quad \omega_k \sim B_0. \tag{9}$$

$$B_j^* = \sum_{k=1}^{\infty} p_{jk}^* \delta_{\omega_k}, \qquad p_{jk}^* \sim \text{Beta}(c_1 p_k, c_1(1-p_k)). \tag{10}$$

Recall that $X_i \sim \text{BeP}(B_{g_i}^*)$ where $g_i \sim \text{Multinomial}(\mathbf{a}_i)$.

Let z_{ik} be the random variable that is 1 if the Bernoulli process draw X_i has atom ω_k, and 0 if not. We have $z_{ik} \sim \text{Bernoulli}(p_{g_i k}^*)$. Because B_0 is continuous, $P(\omega_k = \omega_{k'}) = 0$ for $k \neq k'$ and the random variables R_i and R_{ij} satisfy

$$R_i = \sum_{k=1}^{\infty} z_{ik} \quad \text{and} \quad R_{ij} = \sum_{k=1}^{\infty} z_{ik} z_{jk}. \tag{11}$$

We first show that R_i is Poisson distributed with mean γ.

Let $q_i(\epsilon)$ denote the probability that X_i has atom ω_k conditioned on $p_k > \epsilon$ (this value does not depend on k). That is, $q_i(\epsilon) = P(z_{ik} = 1 | p_k > \epsilon)$.

For a given ϵ, the density of p_k conditioned on $p_k > \epsilon$ is:

$$P(p_k \in dp | p_k > \epsilon) = \frac{c_0 p^{-1}(1-p)^{c_0-1}}{\int_\epsilon^1 c_0 u^{-1}(1-u)^{c_0-1} du} dp, \quad p \in (\epsilon, 1). \tag{12}$$

We can use this density to calculate the success probability $q_i(\epsilon)$:

$$q_i(\epsilon) = \mathbb{E}[z_{ik}|p_k > \epsilon] = \mathbb{E}[p^*_{g_i k}|p_k > \epsilon] = \mathbb{E}[p_k|p_k > \epsilon] = \frac{\int_\epsilon^1 p c_0 p^{-1}(1-p)^{c_0-1}dp}{\int_\epsilon^1 c_0 p^{-1}(1-p)^{c_0-1}dp}, \tag{13}$$

where we have used the tower property of conditional expectation in the second and third equalities.

For a given $\epsilon > 0$, let N_ϵ denote the number of atoms in B with $p_k > \epsilon$. This number is Poisson-distributed with mean $\lambda_\epsilon = \gamma \int_\epsilon^1 c_0 p^{-1}(1-p)^{c_0-1}dp$.

Let $R_i(\epsilon)$ be the number of such atoms that are also in X_i. Because $R_i(\epsilon)$ is the sum of N_ϵ independent Bernoulli trials that each have success probability $q_i(\epsilon)$, it follows that $R_i(\epsilon)|N_\epsilon \sim$ Binomial$(N_\epsilon, q_i(\epsilon))$ and

$$R_i(\epsilon) \sim \text{Poisson}(\lambda_\epsilon q_i(\epsilon)). \tag{14}$$

Because $R_i = \lim_{\epsilon\to 0} R_i(\epsilon)$, it follows that $R_i \sim \text{Poisson}(\lim_{\epsilon\to 0} \lambda_\epsilon q_i(\epsilon))$, where

$$\begin{aligned}\lim_{\epsilon\to 0} \lambda_\epsilon q_i(\epsilon) &= \lim_{\epsilon\to 0} \left[\gamma \int_\epsilon^1 c_0 p^{-1}(1-p)^{c_0-1}dp\right]\left[\frac{\int_\epsilon^1 p c_0 p^{-1}(1-p)^{c_0-1}dp}{\int_\epsilon^1 c_0 p^{-1}(1-p)^{c_0-1}dp}\right] \\ &= \lim_{\epsilon\to 0} \gamma \int_\epsilon^1 p c_0 p^{-1}(1-p)^{c_0-1}dp = \gamma c_0 \int_0^1 (1-p)^{c_0-1}dp = \gamma,\end{aligned}$$

where we have used that $\int_0^1 (1-p)^{c_0-1}dp = \frac{1}{c_0}$. Thus $R_i \sim \text{Poisson}(\gamma)$.

We perform a similar analysis to show the distribution of R_{ij}. Let $q_{ij}(\epsilon)$ denote the probability that X_i and X_j share atom ω_k conditional on $p_k > \epsilon$, g_i and g_j. That is,

$$q_{ij}(\epsilon) = P(z_{ik} = z_{ij} = 1|g_i, g_j, p_k > \epsilon). \tag{15}$$

Although only ϵ appears in the argument of $q_{ij}(\epsilon)$, this quantity also implicitly depends on g_i and g_j. We calculate $q_{ij}(\epsilon)$ explicitly below.

Let $R_{ij}(\epsilon)$ be the number of atoms ω_k for which $p_k > \epsilon$ and ω_k is in both X_i and X_j. We have $R_{ij}(\epsilon)|N_\epsilon, g_i, g_j \sim$ Binomial$(N_\epsilon, q_{ij}(\epsilon))$ and

$$R_{ij}(\epsilon)|g_i, g_j \sim \text{Poisson}(\lambda_\epsilon q_{ij}(\epsilon)). \tag{16}$$

Because $R_{ij} = \lim_{\epsilon\to 0} R_{ij}(\epsilon)$, it follows that $R_{ij} \sim \text{Poisson}(\lim_{\epsilon\to 0} \lambda_\epsilon q_{ij}(\epsilon))$.

To calculate $\lim_{\epsilon\to 0} \lambda_\epsilon q_{ij}(\epsilon)$, we consider two cases. In each case, we first calculate $q_{ij}(\epsilon)$ and then calculate the limit, showing that it is the same as the mean of R_{ij} claimed in the statement of the proposition.

- **Case 1**: $g_i = g_j$

$$\begin{aligned}q_{ij}(\epsilon) &= \mathbb{E}[z_{ik}z_{jk}|p_k > \epsilon, g_i, g_j] = \mathbb{E}[(p^*_{g_i k})^2|p_k > \epsilon, g_i, g_j] \\ &= \mathbb{E}[\mathbb{E}[(p^*_{g_i k})^2|p_k, g_i, g_j]|p_k > \epsilon, g_i, g_j] = \mathbb{E}[p_k(c_1 p_k + 1)/(c_1+1)|p_k > \epsilon, g_i, g_j] \\ &= \frac{\int_\epsilon^1 \frac{c_1 p+1}{c_1+1} p c_0 p^{-1}(1-p)^{c_0-1}dp}{\int_\epsilon^1 c_0 p^{-1}(1-p)^{c_0-1}dp} = \gamma \frac{c_0}{c_1+1}\frac{\int_\epsilon^1 (c_1 p+1)(1-p)^{c_0-1}dp}{\lambda_\epsilon}.\end{aligned} \tag{17}$$

Then the limit $\lim_{\epsilon\to 0}\lambda_\epsilon q_{ij}(\epsilon)$ can be written

$$\begin{aligned}\lim_{\epsilon\to 0}\lambda_\epsilon q_{ij}(\epsilon) &= \gamma\frac{c_0}{c_1+1}\int_0^1 (c_1p+1)(1-p)^{c_0-1}dp \\ &= \gamma\frac{c_0}{c_1+1}\left[c_1\int_0^1 p(1-p)^{c_0-1}dp + \int_0^1(1-p)^{c_0-1}dp\right] \\ &= \gamma\frac{c_0}{c_1+1}\left[\frac{c_1}{c_0(c_0+1)}+\frac{1}{c_0}\right] = \gamma\frac{c_0+c_1+1}{(c_0+1)(c_1+1)},\end{aligned}$$

where we have used that $\int_0^1(1-p)^{c_0-1}dp = \frac{1}{c_0}$ and $\int_0^1 p(1-p)^{c_0-1}dp = \frac{1}{c_0(c_0+1)}$.

- **Case 2**: $g_i \neq g_j$

$$\begin{aligned}q_{ij}(\epsilon) &= \mathbb{E}[z_{ik}z_{jk}|p_k>\epsilon, g_i, g_j] = \mathbb{E}[\mathbb{E}[p^*_{g_ik}p^*_{g_jk}|p_k, g_i, g_j]|p_k>\epsilon, g_i, g_j] \\ &= \mathbb{E}[p_k^2|p_k>\epsilon, g_i, g_j] = \gamma\frac{\int_\epsilon^1 p^2 c_0 p^{-1}(1-p)^{c_0-1}dp}{\lambda_\epsilon}.\end{aligned}$$

Then the limit $\lim_{\epsilon\to 0}\lambda_\epsilon q_{ij}(\epsilon)$ can be written

$$\lim_{\epsilon\to 0}\lambda_\epsilon q_{ij}(\epsilon) = \gamma c_0\int_0^1 p(1-p)^{c_0-1}dp = \gamma\frac{1}{c_0+1},$$

where we have used that $\int_0^1 p(1-p)^{c_0-1}dp = \frac{1}{c_0(c_0+1)}$.

□

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 $\gamma\to\infty$:

$$\frac{R_i}{\gamma} \xrightarrow{P} 1, \tag{18}$$

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

$$\frac{R_{ij}}{R_i} \xrightarrow{P} \begin{cases} \frac{c_0+c_1+1}{(c_0+1)(c_1+1)} & \text{if } g_i = g_j, \\ \frac{1}{c_0+1} & \text{if } g_i \neq g_j. \end{cases} \tag{20}$$

Proof. The proof follows from Proposition 3, Lemma 1, and by noting that both $(c_0+c_1+1)/(c_0+1)(c_1+1)$ and $1/(c_0+1)$ are strictly positive. □

2 Inference using Markov chain Monte Carlo sampling

Our algorithm combines Gibbs and Metropolis updates. For Gibbs updates, we sample a variable from its conditional distribution given the current states of all the other variables. Conjugacy allows simple Gibbs updates for θ and α. Because the dd-IBP prior is not conjugate to the likelihood, we use the Metropolis algorithm to sample $\mathbf{C}$ and $\mathbf{c}^*$. We generate proposals for $\mathbf{C}$ and $\mathbf{c}^*$, and then accept or reject them based on the likelihood ratio. We further divide these updates into two cases: updates for "owned" (active) dishes and updates of dish ownership.

In what follows, we assume that $\mathbf{x}_i$ is conditionally independent of $\mathbf{z}_j$ and $\mathbf{x}_j$ for $j \neq i$ given $\mathbf{z}_i$ and θ.

Sampling θ. To sample the likelihood parameter θ, we draw from the following conditional distribution:

$$P(\theta|\mathbf{X}, \mathbf{C}, \mathbf{c}^*) \propto P(\mathbf{X}|\mathbf{C}, \mathbf{c}^*, \theta)P(\theta), \tag{21}$$

where the prior and likelihood are problem-specific. To obtain a closed-form expression for this conditional distribution, the prior and likelihood must be conjugate. For non-conjugate priors, one can use alternative updates, such as Metropolis-Hastings or slice sampling [4]. Generally, updates for θ will be decomposed into separate updates for each component of θ. In some cases (e.g., the linear-Gaussian model), θ can be marginalized analytically.

Sampling α. To sample the hyperparameter α, we draw from the following conditional distribution:

$$P(\alpha|\mathbf{c}^*, \mathbf{D}, f) \propto P(\alpha) \prod_{i=1}^{N} \text{Poisson}(\lambda_i; \alpha/h_i), \tag{22}$$

where λ_i is determined by $\mathbf{c}^*$ and the prior on α is a Gamma distribution with shape ν_α and inverse scale η_α. Using the conjugacy of the Gamma and Poisson distributions, the conditional distribution over α is given by:

$$\alpha|\mathbf{c}^*, \mathbf{D}, f \sim \text{Gamma}\left(\nu_\alpha + \sum_{i=1}^{N} \lambda_i, \eta_\alpha + \sum_{i=1}^{N} h_i^{-1}\right). \tag{23}$$

Sampling assignments for owned dishes. We update customer assignments for owned dishes (corresponding to "active" features) using Gibbs sampling. For $n = 1, \ldots, N$, $i = 1, \ldots, N$, and $k \in \mathcal{K}_n$, we draw a sample from the conditional distribution over c_{ik} given the current state of all the other variables:

$$P(c_{ik}|\mathbf{C}_{-i}, \mathbf{x}_i, \mathbf{c}^*, \theta, \mathbf{D}, f) \propto P(\mathbf{x}_i|\mathbf{C}, \mathbf{c}^*, \theta)P(c_{ik}|\mathbf{D}, f), \tag{24}$$

where $\mathbf{x}_i$ is the ith row of $\mathbf{X}$, $\mathbf{c}_i$ is the ith row of $\mathbf{C}$, and $\mathbf{C}_{-i}$ is $\mathbf{C}$ excluding row i.[1] The first factor in Eq. 24 is the likelihood,[2] and the second factor is the prior, given by $P(c_{ik} = j|\mathbf{D}, f) = a_{ij}$. In

[1] We rely on several conditional independencies in this expression; for example, $\mathbf{x}_i$ is conditionally independent of $\mathbf{X}_{-i}$ given $\mathbf{C}, \mathbf{c}^*$, and θ.

[2] In calculating the likelihood, we only include the active columns of $\mathbf{Z}$ (i.e., those for which $\sum_{n=1}^{N} z_{nk} > 0$).

considering possible assignments of c_{ik}, one of two scenarios will occur: Either data point i reaches the owner of k (in which case feature k becomes active for i as well as for all other data points that reach i), or it does not (in which case feature k becomes inactive for i as well as for all other data points that reach i). This means we only need to consider two different likelihoods when updating c_{ik}.

Sampling dish ownership. We update dish ownership and customer assignments for newly owned dishes (corresponding to features going from inactive to active in the sampling step) using Metropolis sampling. Both a new ownership vector $\mathbf{c}^{*'}$ and the columns of newly allocated dishes in a new connectivity matrix $\mathbf{C}'$ are proposed by drawing from the prior, and then accepted or rejected according to a likelihood ratio. In more detail, the update proceeds as follows.

1. Propose $\lambda'_i \sim \text{Poisson}(\alpha/h_i)$ for each data point $i = 1, \ldots, N$. Let $\mathcal{K}'_i = \left(\sum_{j<i} \lambda'_j, \sum_{j \leq i} \lambda'_j\right]$, and let $\mathbf{c}^{*'}_k = i$ for all $k \in \mathcal{K}'_i$.

2. Set $\mathbf{C}' \leftarrow \mathbf{C}$. Then populate or depopulate it by performing, for each $i = 1, \ldots, N$,

   (a) If $\lambda'_i > \lambda_i$, allocate $\lambda'_i - \lambda_i$ new dishes to customer i.
   To make room for these new dishes in $\mathbf{C}'$, relabel dishes owned by later customers by moving each column $k > \sum_{j<i} \lambda'_j + \lambda_i$ to column $k + \lambda'_i - \lambda_i$ in $\mathbf{C}'$.
   Then for each new dish $k \in \left(\sum_{j<i} \lambda'_j + \lambda_i, \sum_{j \leq i} \lambda'_j\right]$ fill in the corresponding column of $\mathbf{C}'$ by sampling c'_{mk} according to $P(c'_{mk} = j) = a_{mj}$.

   (b) If $\lambda'_i < \lambda_i$, remove $\lambda_i - \lambda'_i$ randomly selected dishes from customer i.
   Do this by first choosing $\lambda_i - \lambda'_i$ dishes uniformly at random (without replacement) from $\left(\sum_{j<i} \lambda'_j, \sum_{j<i} \lambda'_j + \lambda_i\right]$. Then remove these columns from $\mathbf{C}'$, and relabel all dishes after the first removed dish by moving the corresponding columns of $\mathbf{C}'$.

3. Compute the acceptance ratio ζ. Because the prior (conditional on the current state of the Markov chain) is being used as the proposal distribution, the acceptance ratio reduces to a likelihood ratio (the prior and proposal terms cancel out):

$$\zeta = \min\left[1, \frac{P(\mathbf{X}|\mathbf{C}', \mathbf{c}^{*'}, \theta)}{P(\mathbf{X}|\mathbf{C}, \mathbf{c}^{*}, \theta)}\right]. \tag{25}$$

4. Draw $r \sim \text{Bernoulli}(\zeta)$. Set $\mathbf{C} \leftarrow \mathbf{C}'$ and $\mathbf{c}^* \leftarrow \mathbf{c}^{*'}$ if $r = 1$, otherwise leave $\mathbf{C}$ and $\mathbf{c}^*$ unchanged.

Iteratively applying these updates, the sampler will (after a burn-in period) draw samples from a distribution that approaches the posterior as the burn-in period grows large. The time complexity of this algorithm is dominated by the reachability computation, $O(KN^2)$, and the likelihood computation, which is $O(N^3)$ if coded naively (see [1] for a more efficient implementation using rank-one updates).

In simulation studies, we have found that the acceptance rate for the Metropolis-Hastings step is typically high (≈ 0.9). The main computational bottleneck is the cubic scaling of time complexity

with the number of data points. In future work, we intend to investigate alternative algorithms that scale more favorably with the number of data points. In particular, recently developed stochastic variational inference algorithms have shown great promise for fitting latent variable models to large data sets [2].

For the linear-Gaussian model, $\theta = \mathbf{W}$. As a consequence of our Gaussian assumptions, $\mathbf{W}$ can be marginalized analytically, yielding the likelihood:

$$\begin{aligned} P(\mathbf{X}|\mathbf{Z}) &= \int_{\mathbf{W}} P(\mathbf{X}|\mathbf{Z},\mathbf{W})P(\mathbf{W})d\mathbf{W} \\ &= \frac{\exp\left\{-\frac{1}{2\sigma_x^2}\mathrm{tr}\left(\mathbf{X}^T(\mathbf{I}-\mathbf{Z}\mathbf{H}^{-1}\mathbf{Z}^T)\mathbf{X}\right)\right\}}{(2\pi)^{NM/2}\sigma_x^{(N-K)M}\sigma_w^{KM}|\mathbf{H}|^{M/2}}, \end{aligned} \tag{26}$$

where $\mathrm{tr}(\cdot)$ is the matrix trace, K is the number of active columns, and $\mathbf{H} = \mathbf{Z}^T\mathbf{Z} + \frac{\sigma_x^2}{\sigma_w^2}\mathbf{I}$. In calculating the likelihood, we only include the "active" columns of $\mathbf{Z}$ (i.e., those for which $\sum_{j=1}^{N} z_{jk} > 0$).

3 Additional experimental results: when non-exchangeability hurts

As an example of a missing data problem, we use latent feature models to reconstruct missing observations in electroencephalography (EEG) time series. The EEG data[3] are from a visual detection experiment in which human subjects were asked to count how many times a particular image appeared on the screen [3]. The data were collected as part of a larger effort to design brain-computer interfaces to assist physically disabled subjects.

Distance between data points was defined using the absolute time-difference. Data were z-scored prior to analysis. For 10 of the data points, we removed 2 of the observed features at random. We then ran the MCMC sampler for 1500 iterations, adding Gibbs updates for the missing data by sampling from the observation distribution conditional on the current values of the latent features and hyperparameters. We then used the MAP sample for reconstruction, measuring performance by the squared reconstruction error on the missing data. We repeated this procedure for 10 random restarts. Figure 1 shows the reconstruction results; in this case, the covariate-dependence appears to hurt reconstruction performance, with the IBP achieving the best performance. Thus, covariate-dependent models can suffer when there are not strong dependencies in the data.

References

[1] T.L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. *Advances in Neural Information Processing Systems*, 18, 2005.

[2] M Hoffman, D Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

[3] Available at: `http://mmspl.epfl.ch/page33712.html`

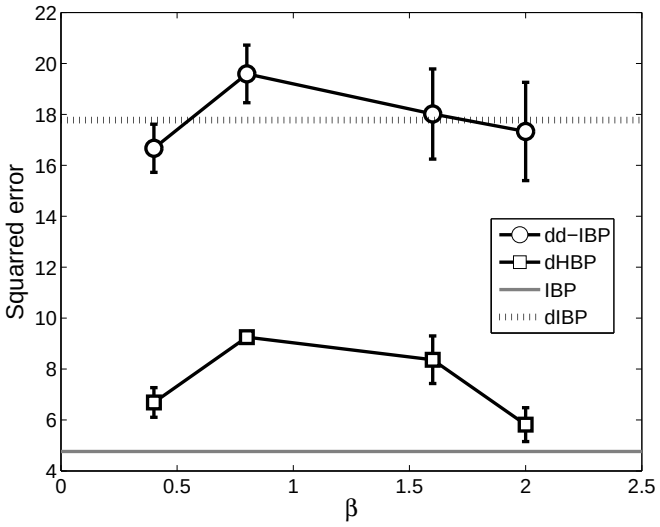

Figure 1: **Reconstruction of missing EEG data**. Reconstruction error for latent feature models as a function of the exponential decay function parameter, β. Results were based on the *maximum a posteriori* sample following 1500 iterations of MCMC sampling. Lower values indicate better performance. Error bars represent standard error of the mean.

[3] U. Hoffmann, J.M. Vesin, T. Ebrahimi, and K. Diserens. An efficient P300-based brain-computer interface for disabled subjects. *Journal of Neuroscience Methods*, 167(1):115–125, 2008.

[4] C.P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Verlag, 2004.