Random Function Priors for Correlation Modeling

Aonan Zhang [1] **John Paisley** [1]

Abstract

The likelihood model of high dimensional data X_n can often be expressed as $p(X_n|Z_n, \theta)$, where $\theta := (\theta_k)_{k\in[K]}$ is a collection of hidden features shared across objects, indexed by n, and Z_n is a non-negative factor loading vector with K entries where Z_{nk} indicates the strength of θ_k used to express X_n. In this paper, we introduce random function priors for Z_n for modeling correlations among its K dimensions Z_{n1} through Z_{nK}, which we call *population random measure embedding* (PRME). Our model can be viewed as a generalized paintbox model (Broderick et al., 2013) using random functions, and can be learned efficiently with neural networks via amortized variational inference. We derive our Bayesian nonparametric method by applying a representation theorem on separately exchangeable discrete random measures.

1. Introduction

Let $X = [X_1, \ldots, X_N]$ be a group of exchangeable high dimensional observations, where $X_n \in \mathbb{R}^d$. In this paper, we assume X is generated by the model

$$p(X) = \int p(X, Z, \theta)\, dZ d\theta = \int p(Z)p(\theta) \prod_{n\in[N]} p(X_n|Z_n, \theta)\, dZ d\theta, \quad (1)$$

where $p(X_n|Z_n, \theta)$ is a likelihood model conditioned on latent features $\theta := (\theta_k)_{k\in[K]}$ that are shared across the population. $Z_n := [Z_{n1}, \ldots, Z_{nK}]$ is a non-negative vector for the nth observation, where Z_{nk} determines the extent to which θ_k is used to express X_n. For example, in topic models (Blei et al., 2003), Z_n is a discrete distribution over topics, where Z_{nk} represents the proportion of words in document n sampled from topic k. In sparse factor models (Griffiths & Ghahramani, 2011), Z_n is a binary vector such that latent feature θ_k contributes to the likelihood if and only if $Z_{nk} = 1$. We generically refer to Z_n as a "non-negative feature loading vector." For exchangeable X, it is often assumed the Z_n are exchangeable as well. If we take Z as a feature loading matrix with Z_n as its rows, then Z is *row exchangeable*. By de Finetti's theorem, we can represent

$$p(Z) = \int \prod_{n\in\mathbb{N}} p(Z_n|\zeta)p(\zeta)d\zeta, \quad (2)$$

for some random object ζ. (We let $N = \infty$ in order to apply de Finetti's theorem.) The goal of this paper is to model complex correlations among entries of Z_n. Following a common practice, we put an independent prior on θ, $p(\theta) = \prod_k p(\theta_k)$ and focus on modeling $p(Z)$.

A straightforward way to model correlation structure is to let $p(Z_n|\zeta)$ be a *parametric* exponential family model. By defining the mean/natural parameters for the model, one can handle correlations to various degrees. For example, Z_n may follow a log-normal distribution (Lafferty & Blei, 2006), where correlations are modeled through a covariance matrix. However, exponential family models (Wainwright et al., 2008) can be rigid, since the number of free parameters is fixed for a certain K. To get a more flexible model, it is tempting to consider higher-order moments $\mathbb{E}[Z_n^{\otimes M}]$ for a large M up to K, where $u^{\otimes M}$ denotes an M-th order outer product of a vector u. but in this case the number of free parameters increases exponentially, leading to intractable inference.

In this paper, we use an alternative *Bayesian nonparametric* method to model Z_n as an outcome of random functions, which can handle complex correlations even when K and M go to infinity. Moreover, those random functions can be learned efficiently through inference/decoder networks via amortized variational inference (Kingma & Welling, 2013). In principle, arbitrarily complex neural networks can be applied to model correlations in our setting.

To give intuition why random function priors are powerful, we first show in Figure 1 an existing feature paintbox model for binary Z_n that illustrates how to model arbitrarily

[1]Department of Electrical Engineering & Data Science Institute, Columbia University, New York, USA. Correspondence to: Aonan Zhang <az2385@columbia.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

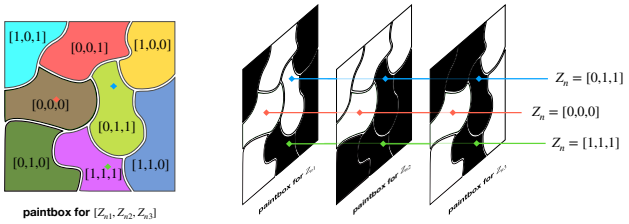


Figure 1. Example of two equivalent representations: (left) Paintbox model for $[Z_{n1}, Z_{n2}, Z_{n3}]$ by partitioning a unit square into eight regions, one for each distinct value. (right) Factorize the paintbox into three feature paintboxes, one each for a latent feature. Three examples of u_n that determine Z_n are demonstrated as dots in the partition paintbox, and as lines across feature paintboxes.

complex correlations using binary random functions (Broderick et al., 2013). For simplicity, let $K = 3$. First, select a compact set S in Euclidean space, on which we can define a uniform distribution. For example let $S = [0, 1]^2$. Then randomly partition S into eight regions. Each partition represents a possible value for $Z_n = [Z_{n1}, Z_{n2}, Z_{n3}]$, as shown in Figure 1. Given the partition, we uniformly sample a point $u_n \sim U(S)$ and assign Z_n be the value defined by the region in which u_n falls. Thus, we translate the problem of modeling distributions on Z_n to modeling *the random partition of* S. Following the classic analogy, we call this a partition paintbox model (Kingman, 1978; Pitman, 2006). One can further factorize the partition paintbox into "feature paintboxes" (Broderick et al., 2013). According to Figure 1, each feature paintbox for the k-th feature is randomly partitioned into two regions denoted as S_k (black) and S_k^c (white). Let $Z_{nk} = \mathbf{1}(u_n \in S_k)$. One can check that the feature paintbox model is the equivalent to the partition paintbox model for arbitrary finite K. (Note that here Z_n is a random indicator function.)

The feature paintbox model is redundant but flexible. The arbitrary order moment $\mathbb{E}[\prod_{k\in\mathcal{J}} Z_{nk}] = \mathbb{E}[\text{vol}(\cap_{k\subset\mathcal{J}} S_k)]$ for any $\mathcal{J} \subset [K]$ can be modeled once we have enough freedom for S_k. We summarize the generative process for the feature paintbox model in Algorithm 1 for arbitrary K, including $K = \infty$.

We propose a model that can be treated as a generalization of the feature paintbox model from binary to non-negative Z according to a function $Z_{nk} = f_n(\vartheta_k)$. There are two key differences between our model and the feature paintbox model. First, we use data-specific *random functions* f_n, instead of points u_n, to represent each observation. Second, we use points ϑ_k from a *Poisson process*, instead of S_k, to index each latent feature. A nice property of our model compared to the paintbox model is that we can use deep learning to model f_n through inference and decoder networks (Kingma & Welling, 2013), allowing for efficient amortized variational inference.

Algorithm 1 Feature paintboxes model

for $k \in [K]$ **do**
  Generate a random subset $S_k \subset S$.
end for
Guarantee that $\sum_{k\in[K]} \text{vol}(S_k) < \infty$ almost surely.
for $n = 1, 2, \ldots$ **do**
  Independently generate $u_n \sim U(S)$.
  Let $Z_n = [Z_{n1}, \ldots, Z_{nK}]$. Set $Z_{nk} = \mathbf{1}(u_n \in S_k)$.
end for

In what follows, Section 2 sets up the problem of modling Z from a random matrix point of view. In Section 3, we embed Z as a random measure and derive the functional form of $Z_{nk} = f_n(\vartheta_k)$ through a representation theorem. In Section 4, we present a concrete example for Bayesian nonparametric topic modeling together with its amortized variational inference algorithm, and show empirical results in Section 5. Finally, we discuss related work in Section 6 and conclude in Section 7.

2. Z as a random matrix?

We will rely on representation theorems to derive the functional form of our models. This usually works out by finding an infinite dimensional random object paired with an exchanegability assumption on that random object. The choice of random objects is the key step, and we will see below that it can be hard to derive an interesting model when choosing a bad random object.

Consider modeling Z as a random matrix. Equation (2) above is one example that derives a mixture representation by assuming row exchangeability of Z. However, Equation (2) is uninformative in that, first, it does not tell us what random object ζ is, and second, it does not determine the connection between Z_n and ζ through $p(Z_n|\zeta)$. Our discussion in Section 1 will show that this provides too much freedom to choose ζ and $p(Z_n|\zeta)$.

We further restrict Z by assuming it is *column exchangeable* as well. This requires allowing both N and K to equal infinity. We call Z *separately exchangeable* if it is both row and column exchangeable. Once $K = \infty$, we need to guarantee series convergence for rows. That is, $\sum_{k\in\mathbb{N}} Z_{nk} < \infty$ with probability 1, for any $n \in \mathbb{N}$. Row sum convergence is always considered necessary. (For example, in a topic model we want to normalize Z_n.) However, the following proposition says that when Z is separately exchangeable, we will get an empty model even for a binary Z.

Proposition 1. *An infinite binary matrix Z (i) is separately exchangeable, and (ii) has finite row sums almost surely, if and only if $Z = \mathbf{0}$ almost surely.*

Proof (Sketch). One can prove that Z is a graphon model if

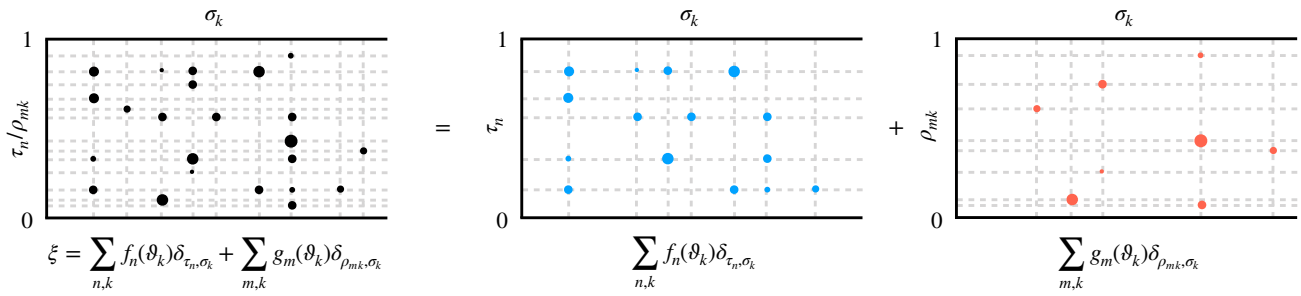


Figure 2. Decoupling a separately exchangeable discrete random measure ξ into two parts.

it is separately exchangeable (Hoover, 1979; Aldous, 1985; Orbanz & Roy, 2015). A graphon model satisfies finite row sums if and only if $Z = \mathbf{0}$. □

When choosing a bad random object, one can either get a vacuous or an empty model through representation theorems. In the next section, we fix this problem by introducing a nice random object ξ generated by embedding Z as a random measure. Then we apply representation theorems on ξ.

3. Z as a random measure

3.1. Population random measure embedding

In this section, we embed the random matrix Z as a *discrete random measure* $\xi = \sum_{n,k} Z_{nk}\delta_{\tau_n,\sigma_k}$ on an infinite strip $[0,1] \times \mathbb{R}_+$, where $(\tau_n)_{n\in\mathbb{N}} \subset [0,1]$ distinguishes objects and $(\sigma_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ distinguishes latent features. Both $(\tau_n)_{n\in\mathbb{N}}$ and $(\sigma_k)_{k\in\mathbb{N}}$ are random as well, and are not necessarily ordered. Note that ξ preserves the matrix structure as demonstrated in Figure 2; the intersection points of horizontal/vertical dashed lines indexed by $(\tau_n)_{n\in\mathbb{N}}$ and $(\sigma_k)_{k\in\mathbb{N}}$ form an "equivalent class" of matrix Z up to a re-ordering of rows and columns. The infinite strip is an abstract space introduced solely for applying representation theorems.

Next, we assume ξ is separately exchangeable. That is, $\xi(\mathcal{T}_1(A)\times\mathcal{T}_2(B)) =_d \xi(A\times B)$ for any measure-preserving transformations $\mathcal{T}_1$, $\mathcal{T}_2$ on $[0,1]$ and $\mathbb{R}_+$ separately for arbitrary Borel sets A, B. Even though the notion of separate exchangeability is different for ξ than for random matrix Z, they are conceptually similar, since interchanging row/column indices will not affect the joint distribution. It turns out that we can represent ξ precisely as follows:

Proposition 2. *A discrete random measure ξ on $[0,1] \times \mathbb{R}_+$ is separately exchangeable if and only if*

$$\xi = \sum_{n,k} f_n(\vartheta_k)\delta_{\tau_n,\sigma_k} + \sum_{m,k} g_m(\vartheta_k)\delta_{\rho_{mk},\sigma_k}, \qquad (3)$$

almost surely for some random measurable functions $f_n, g_m \geq 0$ on $\mathbb{R}_+^2$, a unit rate Poisson process $\{(\vartheta_k,\sigma_k)\}$ on $\mathbb{R}_+^2$, and independent $U(0,1)$ arrays (τ_n) and (ρ_{mk}).

Proof. This follows from the general representation theorem for separately exchangeable random measures on $[0,1] \times \mathbb{R}_+$ (Kallenberg, 2006) by removing the non-atomic parts. Details are given in the appendix. □

We briefly look at the two parts of this representation:

1. $\sum_{n,k} f_n(\vartheta_k)\delta_{\tau_n,\sigma_k}$: This is the part we are interested in. Correlations are learned through coupling of random functions f_n with a Poisson process.

2. $\sum_{m,k} g_m(\vartheta_k)\delta_{\rho_{mk},\sigma_k}$: This part is less important since the double index in ρ_{mk} means each row (object) slice $\xi(\{\rho_{mk}\},\cdot)$ contains at most one atom. We drop this part in our model.

Thus, we can represent $\xi = \sum_{n,k} f_n(\vartheta_k)\delta_{\tau_n,\sigma_k}$ as a coupling of a 2d Poisson process (ϑ_k,σ_k) and random functions f_n. As mentioned in Section 1, we derive $Z_{nk} = f_n(\vartheta_k)$. Since we model the entire population through Z by a random measure embedding, we call our model *population random measure embedding* (PRME).

3.2. Construction via completely random measures

Once we have a representation for Z_{nk}, we still need to guarantee series convergence $\sum_k Z_{nk} = \sum_k f_n(\vartheta_k) < \infty$. This is not obvious, since ϑ_k spans uniformly on $\mathbb{R}_+$. One remedy is to introduce a transformation $\widetilde{\vartheta}_k = T(\vartheta_k)$ that maps almost every $\widetilde{\vartheta}_k$ close to zero, leaving only finite number of $\widetilde{\vartheta}_k$ above any positive threshold. The method to introduce such a transformation T is via completely random measures (CRM) (Kingman, 1967). In the appendix, we show the construction of T via CRMs. In addition, we show that the well-known Indian buffet process (Ghahramani & Griffiths, 2006; Griffiths & Ghahramani, 2011), its extensions (Teh & Gorur, 2009), hierarchical Dirichlet processes (HDP) (Teh et al., 2005) and the discrete infinite logistic normal distribution (DILN) (Paisley et al., 2012b) are instances of population random measure embeddings. However, these models have restrictions in their model capacity. For example, (Paisley et al., 2012b) relies on a linear kernel to model

correlations and there is no obvious extension to complex kernels. As we will show, a PRME can be more flexible by using nonlinear object-specific functions f_n such as deep neural networks.

4. An illustration on topic modeling

4.1. The model

In a topic model, we use Z_n to represents an un-normalized discrete distribution over topics, where Z_{nk} is the strength of topic k for document n. We use a PRME to model Z_{nk}, with the following construction,

$$\begin{aligned}
&Z_{nk} \sim \text{Gamma}(\beta p_k, \exp(f(h_n, \ell_k))),\\
&p_k = V_k \prod_{k'=1}^{k-1}(1-V_{k'}), \quad V_k \sim \text{Beta}(1,\alpha),\\
&h_n \sim \mathcal{N}(0, aI), \quad \ell_k \sim \mathcal{N}(0, bI),\\
&f(h_n, \ell_k) \sim \mathcal{N}(\mu_f(h_n,\ell_k), \sigma_f^2(h_n,\ell_k)).
\end{aligned} \tag{4}$$

We now explain how Equation (4) relates to the original PRME equation $Z_{nk} = f_n(\vartheta_k)$, via four steps.

1. $\boxed{f_n(\vartheta_k) \;\to\; f(h_n, \vartheta_k)}$
We use a parametric function $f(h_n, \cdot)$ to represent $f_n(\cdot)$, where f is a random function, and h_n is an observation-specific random vector. This decomposition is necessary, since we model f as a normal distribution parameterized by *decoder networks* μ_f, σ_f^2, and h_n as the output of an *inference network*.

2. $\boxed{f(h_n, \vartheta_k) \;\to\; f(h_n, \widetilde{\vartheta}_k)}$
We transform $\widetilde{\vartheta}_k = T(\vartheta_k)$ by transforming the original Poisson process (θ_k, σ_k) to a hierarchical Gamma process (Teh et al., 2005; Wang et al., 2011). Then we use a stick-breaking construction over $\widetilde{\vartheta}_k$ (Sethuraman, 1994), where $\widetilde{\vartheta}_k \sim \text{Gamma}(\beta p_k, 1)$. β is a hyperparameter and p_k is generated by the second line of Equation (4).

3. $\boxed{f(h_n, \widetilde{\vartheta}_k) \;\to\; f(h_n, \widetilde{\vartheta}_k, \ell_k)}$
We augment $\widetilde{\vartheta}_k$ to $(\widetilde{\vartheta}_k, \ell_k)$ to introduce extra randomness via ℓ_k. This operation is equivalent to augmenting the original 2d Poisson process (θ_k, σ_k) to a higher dimensional Poisson process $(\theta_k, \sigma_k, \ell_k)$.

4. $\boxed{f(h_n, \widetilde{\vartheta}_k, \ell_k) \;\to\; \widetilde{\vartheta}_k \cdot \exp(f(h_n, \ell_k))}$
We represent $f(h_n, \widetilde{\vartheta}_k, \ell_k)$ as $\widetilde{\vartheta}_k \cdot \exp(f(h_n, \ell_k))$ and assign priors for h_n and ℓ_k (line 3 in Equation (4)). We get Equation (4) by absorbing $\exp(f(h_n, \ell_k))$ into the Gamma scale parameter.

In our construction, series convergence $\sum_{k=1}^{\infty} Z_{nk} < \infty$ can be achieved by bounding μ_f and σ_f^2 through a truncation layer in the decoder network. Given Z_n, we sample words in a document, X_{nm} for $m \in [M_n]$, by first sampling its topic assignment $C_{nm} \sim \text{Disc}(\frac{Z_n}{\sum_k Z_{nk}})$, and then sampling the word from that topic, $X_{nm} \sim \text{Disc}(\theta_{C_{nm}})$, with topic prior $\theta_k \sim \text{Dir}(\gamma_0)$. We recall that in topic models, θ_k (topic k) is a discrete distribution over the vocabulary.

4.2. Amortized variational inference

Assume we have N documents and the posterior is truncated to K topics. The joint likelihood is

$$p(\ell, V, \theta, h, Z, C, X) = \prod_{k=1}^{K} p(\ell_k)p(V_k)p(\theta_k) \prod_{n=1}^{N} \Big[p(h_n) \prod_{k=1}^{K} p(Z_{nk}|V, h_n, \ell_k) \prod_{m=1}^{M_n} p(C_{nm}|Z_n) p(X_{nm}|C_{nm}, \theta) \Big]. \tag{5}$$

We use variational inference to approximate the model posterior by optimizing the variational objective function

$$\max_q \; \mathcal{L} = \max_q \; \mathbb{E}_q \Big[\ln \frac{p(\ell, V, \theta, h, Z, C, X)}{q(\ell, V, \theta, h, Z, C)} \Big], \tag{6}$$

where we restrict q to the factorized family

$$q(\ell, V, \theta, h, Z, C) = \prod_{k=1}^{K} q(\ell_k)q(V_k)q(\theta_k) \prod_{n=1}^{N} \Big[q(h_n|X_n) \prod_{k=1}^{K} q(Z_{nk}) \prod_{m=1}^{M_n} q(C_{nm}) \Big]. \tag{7}$$

Further, for global variables we let

$$q(\ell_k) = \delta_{\widehat{\ell}_k}, \quad q(V_k) = \delta_{\widehat{V}_k}, \quad q(\theta_k) = \text{Dir}(\gamma_k). \tag{8}$$

For local variables, we introduce an inference network g and let $q(h_n|X_n) = \delta_{g(X_n)}$. For the remaining variables

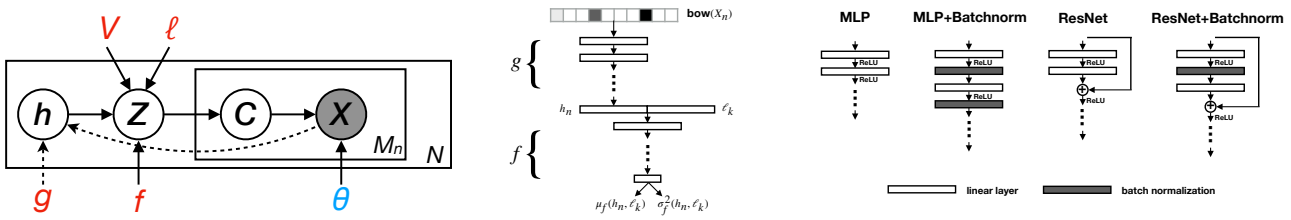
$$q(Z_{nk}) = \text{Gam}(a_{nk}, b_{nk}), \; q(C_{nm}) = \text{Disc}(\phi_{nm}). \tag{9}$$

We use coordinate ascent to update q. Each of these updates is guaranteed to improve the objective when the gradient descent step size is small enough (Nesterov, 2013). More details are given in the appendix.

For $q(Z_{nk})$, we maximize a lower bound for $\mathcal{L}$ similar to Paisley et al. (2012b), giving updates

$$\begin{aligned}
a_{nk} &= \beta \widehat{p}_k + \sum_{m=1}^{M_n} \phi_{nm}(k),\\
b_{nk} &= 1/\Big(\mathbb{E}\Big[\exp(-f(h_n, \ell_k)) \Big] + \frac{M_n}{\varepsilon_n} \Big),
\end{aligned} \tag{10}$$

where $\varepsilon_n = \sum_{k=1}^{K} \mathbb{E}[Z_{nk}]$.



(a) (b)

Figure 3. (a) Graphical representation of our proposed model. Solid arrows represent the generative process and dashed arrows show the VAE part of the posterior. We organize local parameters that belong to a document/word into boxes and remove all sub-indices. We use stochastic natural gradient ascent for θ and use stochastic gradient ascent for $[\ell, V, g, f]$ (b) Left: The architecture we used in our experiments. Right: Various layer designs.

For $q(C_{nm})$ and $q(\theta)$, we have respective updates

$$\phi_{nm}(k) \propto \exp\Big(\mathbb{E}[\ln \theta_{k,X_{nm}}] + \mathbb{E}[\ln Z_{nk}]\Big), \tag{11}$$

$$\gamma_{kd} = \gamma_0 + \sum_{n=1}^{N}\sum_{m=1}^{M_n} \phi_{nm}(k)\cdot \mathbf{1}(X_{nm}=d). \tag{12}$$

For $[\ell, V, g, f]$, we do gradient ascent on $\mathcal{L}$. Batch variational inference can be done via coordinate ascent by iteratively updating the above variables. Dependencies among variables are shown in Figure 3(a).

For stochastic inference, in each global iteration we sample a subset $N_t \subset [N]$ and compute the noisy variational objective

$$\mathcal{L}_t = \mathbb{E}\Big[\ln p(\ell, V, \theta)\Big] + \frac{N}{|N_t|}\sum_{n\in N_t}\mathbb{E}\Big[\ln p(h_n, Z_n, C_n, X_n)\Big]$$
$$+ \mathbb{H}\Big[q(\theta)\Big] + \frac{N}{|N_t|}\sum_{n\in N_t}\mathbb{H}\Big[q(Z_n, C_n)\Big]. \tag{13}$$

Optimizing local variables Z, C can be done via closed-form updates exactly as in the batch case. For the other parameters we use stochastic gradient methods. Let $\rho^{(t)} \propto (t_0 + t)^{-\kappa}$ be the step size with some constant t_0 and $\kappa \in (0.5, 1]$. We apply the stochastic natural gradient method (Hoffman et al., 2013) for θ

$$\widetilde{\gamma}_{kd}^{(t)} = \gamma_0 + \sum_{n\in N_t}\sum_{m=1}^{M_n}\phi_{nm}(k)\cdot\mathbf{1}(X_{nm}=d),$$
$$\gamma_{kd}^{(t)} = (1-\rho^{(t)})\gamma_{kd}^{(t-1)} + \rho^{(t)}\widetilde{\gamma}_{kd}^{(t)}. \tag{14}$$

and stochastic gradient method for the rest,

$$[\ell, V, g, f]^{(t)} = [\ell, V, g, f]^{(t-1)} + \rho^{(t)}\nabla_{[\ell,V,g,f]}\mathcal{L}_t. \tag{15}$$

Since in each iteration we only do one gradient step, the cost is low. Note that through the variational autoencoder (VAE) (Kingma & Welling, 2013) we transfer local updates for h_n to global update for g, which will significantly speed-up inference. We summarize the stochastic inference algorithm in Algorithm 2.

Algorithm 2 Stochastic inference algorithm

for $t = 1, 2, \ldots$ **do**
  Sample a subset $N_t \subset [N]$
  Update local variables
  while not converge **do**
    Closed-form update $q(Z_n)$ for $n \in N_t$. Eq. (10)
    Closed-form update $q(C_n)$ for $n \in N_t$. Eq. (11)
  end while
  Update global variables
  Noisy natural gradient step for $q(\theta)$. Eq. (14)
  Noisy gradient step for ℓ, V, g, f. Eq. (15)
end for

4.3. Network architectures

The flexibility of our model comes from the inference and decoder networks g and f. As we show in the experiments, these allow us to learn complex non-linear "paintboxes" in order to capture complex topic correlations. Since optimizing over deep neural networks is still a challenging problem in theory, we design our networks with architectures that work well in practice. Rather than directly applying multilayer perceptrons (Rumelhart et al., 1985), we instead use more complex layer designs such as batch normalization (Ioffe & Szegedy, 2015) and deep residual networks (ResNet) (He et al., 2016) to speed-up training. For inference network g, we use the bag-of-words representation of X_n as the input feature. For decoder network f, we concatenate $h_n = g(X_n)$ and ℓ_k as inputs. Detailed architecture design is shown in Figure 3(b).

5. Experiments

5.1. Batch experiments

We show empirical results on three text datasets: a 5K subset of New York Times, 20Newsgroups, and NeurIPS. Their basic statistics are shown in Table 2. For each test document X_n, we do a $90\%/10\%$ split into training words $X_{n,TR}$ and testing words $X_{n,TS}$. The perplexity is calculated based on

Table 1. Perplexity result for text data sets with different dictionary sparsity levels controlled by γ_0.

Model	New York Times				20Newsgroups				NeurIPS			
	$\gamma_0=0.2$	$\gamma_0=0.4$	$\gamma_0=0.6$	$\gamma_0=0.8$	$\gamma_0=0.2$	$\gamma_0=0.4$	$\gamma_0=0.6$	$\gamma_0=0.8$	$\gamma_0=0.2$	$\gamma_0=0.4$	$\gamma_0=0.6$	$\gamma_0=0.8$
HDP	2436.51	2464.74	2482.61	2501.82	5317.68	5845.90	6294.68	6665.68	1973.39	1962.90	1981.83	2009.58
DILN	2231.16	2295.12	2418.16	2509.24	5164.93	5732.12	6143.64	6389.99	1853.89	1902.88	1944.90	**1947.94**
PRME	**2203.00**	**2247.25**	**2299.60**	**2338.38**	**5102.08**	**5531.04**	**5878.39**	**5975.12**	**1753.61**	**1850.37**	**1917.21**	1953.85

Table 2. Dataset description.

Corpus	# train	# test	# vocab	# tokens
New York Times	5,000	500	8,000	1.4M
20Newsgroups	11,269	7,505	53,975	2.2M
NeurIPS	2,183	300	14,086	3.3M

Table 3. Network layer configurations for New York Times dataset.

Depth	Inference Network	Decoder Network
2 layers	$[8000 \times d_h]$	$[(d_h + d_\ell) \times 80]$ $[80 \times 2]$
4 layers	$[8000 \times 1000]$ $[1000 \times d_h]$	$[(d_h + d_\ell) \times 80]$ $[80 \times 80]$ $[80 \times 2]$
6 layers	$[8000 \times 1000]$ $[1000 \times 1000]$ $[1000 \times d_h]$	$[(d_h + d_\ell) \times 80]$ $[80 \times 80]$ $[80 \times 80]$ $[80 \times 2]$
8 layers	$[8000 \times 1000]$ $[1000 \times 1000]$ $[1000 \times 1000]$ $[1000 \times d_h]$	$[(d_h + d_\ell) \times 80]$ $[80 \times 80]$ $[80 \times 80]$ $[80 \times 80]$ $[80 \times 2]$

the prediction of $X_{n,TS}$ given the model and $X_{n,TR}$,

$$\text{perplexity} = \exp\Big(-\frac{\sum_{m \in X_{n,TS}} \ln p(X_{nm}|X_{n,TR})}{|X_{n,TS}|}\Big). \quad (16)$$

Lower perplexity means better predictive performance.

In Table 1, we compare three Bayesian nonparametric models: hierarchical Dirichlet process (HDP) (Teh et al., 2005), discrete infinite logistic normal (DILN) (Paisley et al., 2012b), and our population random measure embedding (PRME) using 4-layer MLP with batch normalization.[1] We tune γ_0 and fix the truncation level $K = 100$ and set the $a = 1, b = 1, \alpha = 1, \beta = 5$ for fair comparisons. All gradient updates are done via Adam (Kingma & Ba, 2014) with learning rate 10^{-4}. As Table 1 shows, PRME consistently perform better than HDP and DILN. Where DILN was designed to outperform HDP by learning topic correlation structure, PRME improves upon DILN by learning a more complex kernel structure.

Since PRME encodes complex correlation patterns with a neural network, we further consider the influence of network architecture on perplexity for the New York Time dataset.

[1]The number of layers includes inference network and decoder network. We ignore the last layer of the decoder network.

Table 4. Perplexity result for various network depths.

Depth	MLP	MLP+BN	ResNet	ResNet+BN
2 layers	2325.84	2327.81	N/A	N/A
4 layers	2228.62	2203.00	2214.02	2195.72
6 layers	2219.06	**2184.44**	2202.79	2194.74
8 layers	**2196.35**	2195.68	**2199.07**	**2184.56**

Table 5. Perplexity result for various size of h_n/ℓ_k.

Hidden Size	MLP	MLP+BN	ResNet	ResNet+BN
$d_h = d_\ell = 2$	2287.40	2258.97	2265.53	2256.84
$d_h = d_\ell = 5$	2245.43	2243.26	2231.54	2225.64
$d_h = d_\ell = 10$	**2220.82**	2217.65	2227.04	2199.73
$d_h = d_\ell = 20$	2228.62	**2203.00**	**2214.02**	**2195.72**

We compare four layer designs: multilayer perceptron (MLP), MLP with batch normalization (MLP+BN), ResNet, and ResNet with batch normalization (ResNet+BN); see Figure 3(b) for details. In Table 4 and Table 5, we separately tune the depth of each network and the hidden size of h/ℓ while holding other parameters fixed. The details of layer sizes can be found in Table 3. We observe that the perplexity result tend to be better when we scale up the network depth/width. Batch normalization and ResNet both improve performance.

5.2. Online experiments

For the larger one million New York Times dataset, we show "topic paintboxes" learned with stochastic PRME in Figure 4.[2] In Figure 4, each paintbox corresponds to one topic whose top words are displayed inside the box. The color of a pixel (x, y) in the k-th paintbox ranges from blue (small value) to red (large value) and represents mean topic strength $\mathbb{E}[Z_{(x,y),k}] = \beta\widehat{p}_k\mathbb{E}[\exp(f(h_{(x,y)}, \ell_k))]$ as a function of $h_{(x,y)}$ for topic k. To define $h_{(x,y)}$ for 2d visualization, we collect the empirical embeddings $H = [h_1, \ldots, h_N]^\top = [g(X_1), \ldots, g(X_N)]^\top$ on a subset of data, subtract their mean m_h, and use the SVD to select the two most informative directions $\widetilde{h}_1, \widetilde{h}_2$ with singular values s_1, s_2. Then we plot each paintbox as the function value $\mathbb{E}[Z_{(x,y),k}] = \beta\widehat{p}_k\mathbb{E}[\exp(f(m_h + xs_1\widetilde{h}_1 + ys_2\widetilde{h}_2, \ell_k))]$ by tuning $(x, y) \in [-0.2, 0.2]^2$.

The correlation between topics can be read out from the paintboxes. Those paintboxes that have overlapping salient regions tend to be more correlated. For example, topic 13 [music, concert, orchestra], topic 20 [film, movie, films],

[2]We set $t_0 = 100, \kappa = 0.75$ and use a 6-layer MLP.

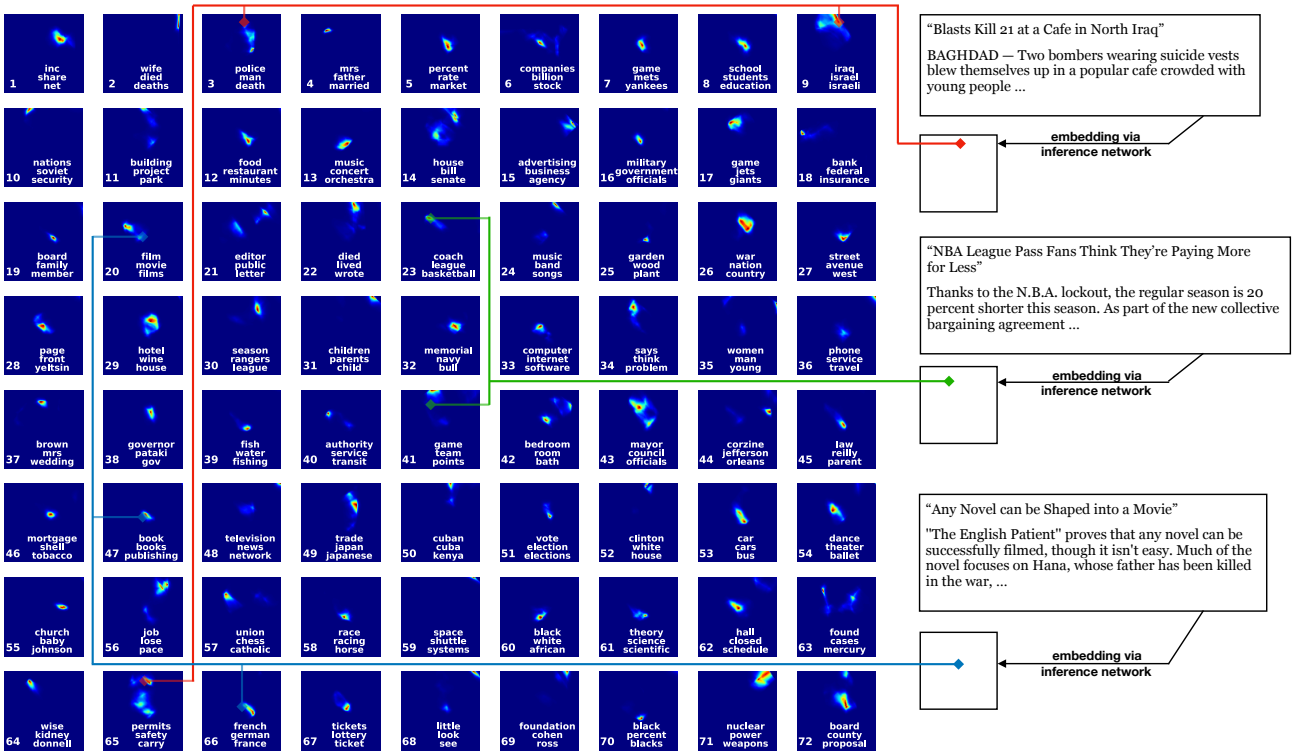


Figure 4. A paintbox demonstration of salient topics learned from the one million New York Times dataset. In each paintbox on the LHS, pixel (x, y) represents the topic strength $Z_{(x,y),k}$ as a function of $h_{(x,y)}$ for a particular topic k. We also show embeddings of three articles in the same space, as well as their projection onto selected paintboxes. Each article is connected to its most-used topics.

and topic 47 [book, books, publishing] share a salient region, which gives a third-order positive correlations over those topics. In principle, the paintbox can explain arbitrary order correlations as the neural network complexity increases. We observe that each paintbox in Figure 4 consists of multiple contiguous salient regions. This is due to the smoothness of neural networks, since $g(X_{n_1}) \approx g(X_{n_2})$ when X_{n_1} and X_{n_2} share similar words. Also, the various "modes" in each paintbox demonstrate the greater flexibility of neural networks in explaining different contexts of a topic.

In Figure 4, we also display three documents with their embeddings h_n projected onto the 2d paintbox space. Each embedding hits salient regions of several paintboxes. Thus, each document can be interpreted as a mixture of these corresponding topics. We again note that we only display the paintbox in 2d via post-processing, but the actual paintbox is in 20 dimension; a higher-dimensional paintbox can be more complex than what is shown.

We can compare the difference between paintboxes for PRME in Figure 4 and paintboxes for binary random measures in Figure 1. First, the paintbox for PRME is real-valued, so it is natural to use smooth functions to model it. In the binary case the paintbox is zero/one valued; in this case one can apply a threshold function over the PRME paintbox to binarize it. Second, in contrast to the binary paintbox, each PRME paintbox is unbounded. We control the area of this salient region through regularization.

Figure 5(a) demonstrates the perplexity of DILN and PRME with various decay speed κ on a held-out test set of size 3K. PRME converges after seeing one million documents, and it performs better than DILN. Also, online learning is much more efficient than batch learning with various training data size, as shown in Figure 5(b). In Figure 5(c), we compare run times for updating local parameters ($[Z, C]$ for PRME) and global parameters ($[\theta, \ell, V, g, f]$ for PRME) with batch size 500. Since the cost is very imbalanced between local and global, for demonstration purpose we compare the cost between five local iterations and one global iteration. In our experiments, local updates requires around 20 iterations to converge. Compared with DILN, PRME costs much less in local and costs more in global updates, since it uses the VAE to transfer local updates for h_n into global updates for g. The extra global cost ($\sim$0.35s) is significantly smaller than the reduced local cost ($\sim$4s), even when using a deep network architecture. Finally, Figure 5(d) demonstrates the usage proportion for all topics. PRME tends to use a subset of the 100 available topics in the truncated posterior, indicating use by the model of this nonparamteric feature.

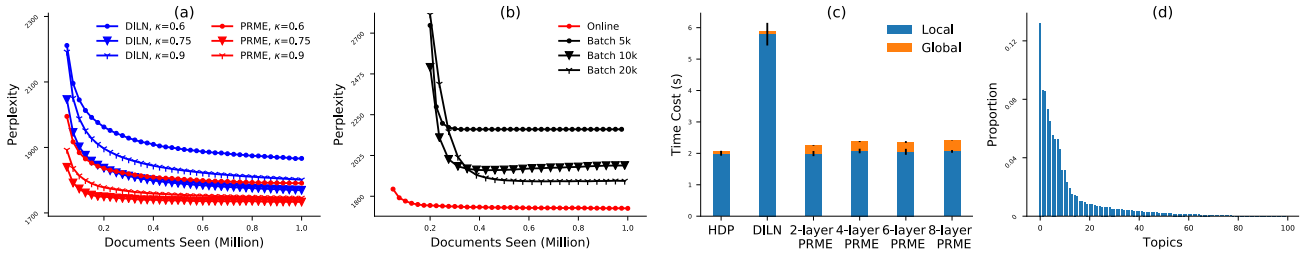


Figure 5. (a) Online performance comparisons between DILN and PRME. (b) Online versus batch. (c) Time cost comparison between updating local and global variables. (d) Ranked topic usage proportions in the posterior, indicating nonparametric functionality.

6. Discussion

6.1. Connections with other random objects

Another view is to treat $(Z_{nk})_{n\in[N],k\in[K]}$ as a bipartite graph over objects $[N]$ and atoms (features) $[K]$ with edge strength Z_{nk}. An important topic in random graph theory is to study the total strength of edges $|E| = \sum_{n\in[N],k\in[K]} \mathbb{E}[Z_{nk}]$ asymptotically as a function of N. There has been extensive work on random graphs, networks, and relational models (Roy et al., 2008; Miller et al., 2009; Caron, 2012; Lloyd et al., 2012; Veitch & Roy, 2015; Cai et al., 2016; Lee et al., 2016; Crane & Dempsey, 2017; Caron & Rousseau, 2017; Caron & Fox, 2017), but these methods mainly focus on dense graphs where $|E| \sim \mathcal{O}(N^2)$, and sparse graphs where $|E| \sim \mathcal{O}(N^{1+\alpha})$ with $0 < \alpha < 1$ or $|E| \sim \mathcal{O}(N \log N)$. Our method offers a new solution to *extremely sparse hidden graphs* where $|E| \sim \mathcal{O}(N)$, by coupling random functions and a Poisson process. Our solution cannot be trivially derived from previous representations in sparse/dense graphs. There is a developed probability theory building connections between exchangeable binary random measures and functions on combinatorial structures among atoms (Pitman, 1995; 2006; Broderick et al., 2013; 2015; Heaukulani et al., 2016; Campbell et al., 2018).

Our topic model construction is motivated by previous research on dependent random measures (Zhou et al., 2011; Paisley et al., 2012b; Chen et al., 2013; Foti et al., 2013; Zhang & Paisley, 2015; 2016). Our focus is to place *mild exchangeability assumptions* on a population random measure ξ and derive a very general random function model through representation theorems. Hence our use of neural networks to achieve this task. We mention that our method can also be adapted to non-exchangeable settings.

6.2. Deep hierarchical Bayesian models

One can scale up model capacity by stacking multiple one-layer Bayesian nonparametric models such as Dirichlet processes (Teh et al., 2005), beta processes (Thibaux & Jordan, 2007), and Gamma processes (Zhou et al., 2015; Zhou, 2018). Population random measure embedding uses a different strategy by constructing random measures as a coupling of random functions with a single Poisson process. In this way, we transfer all the model complexity into random functions f_n. Using amortized variational inference, we transfer posterior inference of discrete random measures into optimizing neural networks, which is much more efficient.

6.3. Posterior inference bottleneck

Efficient posterior inference is essential in Bayesian nonparametric methods where conjugacy often does not hold (Broderick et al., 2014; Zhang et al., 2016). In principle, one can apply a simple prior on Z and still rely on accurate posterior inference to resolve the structure. However, posterior inference for random measures is not simple because complex correlations among atoms leads to slow MCMC mixing. Instead, one can approximate the posterior using variational methods (Blei et al., 2017) and try to learn a q distribution with good approximation quality (Paisley et al., 2012a; Hoffman & Blei, 2015; Ranganath et al., 2016; Tran et al., 2017). Our method introduced a structured prior to regularize variational inference. Empirical results showed that we get an interpretable posterior.

7. Conclusion and Future Work

We presented random function priors to handle complex correlations among features via a population random measure embedding. We further derived a new Bayesian nonparametric topic model to demonstrate the effectiveness of our method for learning topic correlations through deep neural networks with amortized variational posterior inference. In future work, we will consider the more challenging task of removing the non-differentiable Poisson process and making our model fully differentiable.

Acknowledgements

We thank Howard Karloff and Victor Veitch for their helpful comments during the early stage of this work. This research was supported in part by funding from Capital One Labs in New York City.

References

Aldous, D. J. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pp. 1–198. Springer, 1985.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation. *Journal of machine Learning research (JMLR)*, 3(Jan):993–1022, 2003.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association (JASA)*, 112(518): 859–877, 2017.

Broderick, T., Pitman, J., Jordan, M. I., et al. Feature allocations, probability functions, and paintboxes. *Bayesian Analysis*, 8(4):801–836, 2013.

Broderick, T., Wilson, A. C., and Jordan, M. I. Posteriors, conjugacy, and exponential families for completely random measures. *arXiv preprint arXiv:1410.6843*, 2014.

Broderick, T., Mackey, L., Paisley, J., and Jordan, M. I. Combinatorial clustering and the beta negative Binomial process. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):290–306, 2015.

Cai, D., Campbell, T., and Broderick, T. Edge-exchangeable graphs and sparsity. In *Advances in Neural Information Processing Systems*, pp. 4249–4257, 2016.

Campbell, T., Cai, D., Broderick, T., et al. Exchangeable trait allocations. *Electronic Journal of Statistics*, 12(2): 2290–2322, 2018.

Caron, F. Bayesian nonparametric models for bipartite graphs. In *Advances in Neural Information Processing Systems*, pp. 2051–2059, 2012.

Caron, F. and Fox, E. B. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1295–1366, 2017.

Caron, F. and Rousseau, J. On sparsity and power-law properties of graphs based on exchangeable point processes. *arXiv preprint arXiv:1708.03120*, 2017.

Chen, C., Rao, V., Buntine, W., and Teh, Y. Dependent normalized random measures. In *International Conference on Machine Learning*, 2013.

Crane, H. and Dempsey, W. Edge exchangeable models for interaction networks. *Journal of the American Statistical Association (JASA)*, 2017.

Foti, N., Futoma, J., Rockmore, D., and Williamson, S. A unifying representation for a class of dependent random measures. In *Artificial Intelligence and Statistics*, pp. 20–28, 2013.

Ghahramani, Z. and Griffiths, T. L. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, pp. 475–482, 2006.

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

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Heaukulani, C., Roy, D. M., et al. The combinatorial structure of beta negative Binomial processes. *Bernoulli*, 22 (4):2301–2324, 2016.

Hoffman, M. D. and Blei, D. M. Structured stochastic variational inference. In *Artificial Intelligence and Statistics*, 2015.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *The Journal of Machine Learning Research (JMLR)*, 14(1):1303–1347, 2013.

Hoover, D. N. Relations on probability spaces and arrays of random variables. Technical report, Institute of Advanced Study, Princeton, 1979.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Kallenberg, O. *Probabilistic symmetries and invariance principles*. Springer Science & Business Media, 2006.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

Kingman, J. F. The representation of partition structures. *Journal of the London Mathematical Society*, 2(2):374–380, 1978.

Lafferty, J. D. and Blei, D. M. Correlated topic models. In *Advances in Neural Information Processing Systems*, pp. 147–154, 2006.

Lee, J., James, L. F., and Choi, S. Finite-dimensional BFRY priors and variational Bayesian inference for power law models. In *Advances in Neural Information Processing Systems*, pp. 3162–3170, 2016.

Lloyd, J., Orbanz, P., Ghahramani, Z., and Roy, D. M. Random function priors for exchangeable arrays with applications to graphs and relational data. In *Advances in Neural Information Processing Systems*, pp. 998–1006, 2012.

Miller, K., Jordan, M. I., and Griffiths, T. L. Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems*, pp. 1276–1284, 2009.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Orbanz, P. and Roy, D. M. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):437–461, 2015.

Paisley, J., Blei, D., and Jordan, M. Variational Bayesian inference with stochastic search. In *International Conference on Machine Learning*, pp. 1367–1374, 2012a.

Paisley, J., Wang, C., Blei, D. M., et al. The discrete infinite logistic normal distribution. *Bayesian Analysis*, 7(4): 997–1034, 2012b.

Pitman, J. Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2): 145–158, 1995.

Pitman, J. *Combinatorial Stochastic Processes: Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002*. Springer, 2006.

Ranganath, R., Tran, D., and Blei, D. Hierarchical variational models. In *International Conference on Machine Learning*, pp. 324–333, 2016.

Roy, D. M., Teh, Y. W., et al. The mondrian process. In *NIPS*, pp. 1377–1384, 2008.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

Sethuraman, J. A constructive definition of Dirichlet priors. *Statistica sinica*, pp. 639–650, 1994.

Teh, Y. W. and Gorur, D. Indian buffet processes with power-law behavior. In *Advances in Neural Information Processing Systems*, pp. 1838–1846, 2009.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems*, pp. 1385–1392, 2005.

Thibaux, R. and Jordan, M. I. Hierarchical beta processes and the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, volume 2, pp. 564–571, 2007.

Tran, D., Ranganath, R., and Blei, D. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, pp. 5523–5533, 2017.

Veitch, V. and Roy, D. M. The class of random graphs arising from exchangeable random measures. *arXiv preprint arXiv:1512.03099*, 2015.

Wainwright, M. J., Jordan, M. I., et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Wang, C., Paisley, J., and Blei, D. Online variational inference for the hierarchical Dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 752–760, 2011.

Zhang, A. and Paisley, J. Markov mixed membership models. In *International Conference on Machine Learning*, pp. 475–483, 2015.

Zhang, A. and Paisley, J. Markov latent feature models. In *International Conference on Machine Learning*, pp. 1129–1137, 2016.

Zhang, A., Gultekin, S., and Paisley, J. Stochastic variational inference for the hdp-hmm. In *Artificial Intelligence and Statistics*, pp. 800–808, 2016.

Zhou, M. Parsimonious Bayesian deep networks. *arXiv preprint arXiv:1805.08719*, 2018.

Zhou, M., Yang, H., Sapiro, G., Dunson, D., and Carin, L. Dependent hierarchical beta process for image interpolation and denoising. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 883–891, 2011.

Zhou, M., Cong, Y., and Chen, B. The Poisson Gamma belief network. In *Advances in Neural Information Processing Systems*, pp. 3043–3051, 2015.