Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

Correlated Random Measures

Rajesh Ranganath & David M. Blei

To cite this article: Rajesh Ranganath & David M. Blei (2018) Correlated Random Measures, Journal of the American Statistical Association, 113:521, 417-430, DOI: 10.1080/01621459.2016.1260468

To link to this article: https://doi.org/10.1080/01621459.2016.1260468

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 521, 417–430, Theory and Methods
https://doi.org/10.1080/01621459.2016.1260468

Taylor & Francis
Taylor & Francis Group

Correlated Random Measures

Rajesh Ranganath[a] and David M. Blei[b]

[a]Department of Computer Science, Princeton University, Princeton, NJ; [b]Departments of Computer Science and Statistics, Columbia University, New York, NY

ABSTRACT

We develop correlated random measures, random measures where the atom weights can exhibit a flexible pattern of dependence, and use them to develop powerful hierarchical Bayesian nonparametric models. Hierarchical Bayesian nonparametric models are usually built from completely random measures, a Poisson-process-based construction in which the atom weights are independent. Completely random measures imply strong independence assumptions in the corresponding hierarchical model, and these assumptions are often misplaced in real-world settings. Correlated random measures address this limitation. They model correlation within the measure by using a Gaussian process in concert with the Poisson process. With correlated random measures, for example, we can develop a latent feature model for which we can infer both the properties of the latent features and their dependency pattern. We develop several other examples as well. We study a correlated random measure model of pairwise count data. We derive an efficient variational inference algorithm and show improved predictive performance on large datasets of documents, web clicks, and electronic health records. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2015
Revised April 2016

KEYWORDS
Bayesian nonparametrics; Gaussian processes; Health records; Hierarchical models; Poisson processes; Random measures

1. Introduction

Hierarchical Bayesian nonparametric models (Teh and Jordan 2010) have emerged as a powerful approach to analyzing complex data (Williamson et al. 2010; Fox et al. 2011; Zhou et al. 2012). These models assume there are a set of patterns, or components, that underlie the observed data; each data point exhibits each component with different nonnegative weight; the number of components is unknown and new data can exhibit still unseen components. Given observed data, the posterior distribution reveals the components (including how many there are), reveals how each data point exhibits them, and allows for this representation to grow as more data are seen. These kinds of assumptions describe many of the most common hierarchical Bayesian nonparametric models, such as the hierarchical Dirichlet process (Teh et al. 2006), the Gamma-poisson process (Titsias 2008), the beta-Bernoulli process (Thibaux and Jordan 2007), and others.

For example, in Section 6 we analyze patient data from a large hospital; each patient is described by the set of diagnostic codes on her chart. Potentially, the full dataset reflects patterns in diagnostic codes, each pattern a set of diagnoses that often occurs together. Further, some patients will exhibit multiple patterns—they simultaneously suffer from different clusters of symptoms. With these data, a Bayesian nonparametric model can uncover and characterize the underlying pattern and describe each patient in terms of which patterns she exhibits. Recent innovations in approximate posterior inference let us analyze such data at large scale, uncovering useful characterizations of disease and injury for both exploration and prediction. In our study on medical data, we discover components that summarize conditions such as congestive heart failure, diabetes, and depression (Table 1).

But there is a limitation to the current state of the art in Bayesian nonparametric models. To continue with the example, each patient is represented as an infinite vector of nonnegative weights, one per component. (There are a countably infinite number of components.) Most hierarchical Bayesian nonparametric models assume that these weights are uncorrelated—that is, the presence of one component is unrelated to the presence (or absence) of the others. But this assumption is usually unfounded. For example, in the medical data we find that type 2 diabetes is related to congestive heart failure (Table 4).

In this article, we solve this problem. We develop correlated random measures, a general-purpose construction for infusing covariance into the distribution of weights of both random measures and hierarchical Bayesian nonparametric models. Our approach can capture that a large positive weight for one component might covary with a large positive weight in another, a type of pattern that is out of reach for most hierarchical Bayesian nonparametric models. We demonstrate that bringing such correlations into the model both improves prediction and reveals richer exploratory structures. Correlated random measures can be used as a model for a collection of observed weighted point processes and can be adapted to a wide variety of proven Bayesian nonparametric settings, such as language modeling (Teh 2006), time series analysis (Fox et al. 2011), dictionary learning (Zhou et al. 2009), and nested models (Paisley et al. 2015).

CONTACT Rajesh Ranganath rajeshr@cs.princeton.edu Department of Computer Science, Princeton University, Princeton, NJ 08544.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

Table 1. The top 20 components on the Mayo Clinic data. We find that each factor forms a medically meaningful grouping of diagnosis codes. For example, there are allergy, pregnancy, and alcohol dependence components.

1: Mammogram, Routine medical exam, Lumbago, Cervical Cancer Screening, Hypothyroidism
2: Hypertension, Hyperlipidemia, Coronary atherosclerosis, Prostate Cancer Screening, Vaccine for influenza
3: Acute pharyngitis, Cough, Myopia, Vaccine for influenza, Joint pain-shoulder
4: Child Exam, Vaccine for flu, Otitis media, Upper respiratory infection, pharyngitis
5: Long-term anticoagulants, Atrial fibrillation, Hypertension, Congestive Heart Failure, Chronic airway obstruction
6: Normal pregnancy, Normal first pregnancy, Cervical cancer screening, Delivery, Conditions antepartum
7: Diabetes-2, Hypertension, Hyperlipidemia, Uncontrolled Diabetes-2, Diabetes-2 with ophthalmic manifestations
8: Depression, Dysthymia, Anxiety state, Generalized anxiety disorder, Major depressive affective disorder
9: Joint pain lower leg, Arthritis lower leg, Local arthritis lower leg, Post-procedural status, Follow-up surgery
10: Allergic rhinitis, Desensitization to allergens, Asthma, Chronic allergic conjunctivitis, Chronic sinusitis
11: Heart valve replacement, Prostate cancer, Lung and bronchus cancer, Secondary bone cancer, Other lung disease
12: Morbid obesity, Obesity, Obstructive sleep apnea, Sleep apnea, Intestinal bypass status
13: Acne, Convulsions, Abnormal involuntary movements, Cerebral palsy, Long-term use meds
14: Abnormality of gait, Personality change, Persistent mental disorders, Lack of coordination, Debility
15: Attention disorder w hyperactivity, Attention disorder w/o hyperactivity, Adjustment disorder, Opposition defiant disorder, Conduct disturbance
16: Diseases of nail, Corns and callosities, Dermatophytosis of nail, Ingrowing nail, Other states following surgery of eye and adnexa
17: Alcohol dependence, Tobacco use disorder, Alcohol abuse, Other alcohol dependence-in remission, Other alcohol dependence-continuous
18: Schizophrenia-Paranoid, Long-term use meds, Schizophrenia, Schizophrenia-paranoid-chronic, Drug monitor
19: Female breast cancer, Personal history of breast cancer, Lymph cancer, Carcinoma in situ of breast, lymphedema
20: Child health exam, Vaccination for disease, Vaccinations against pneumonia, Need for prophylactic vaccination against viral hepatitis, Procedure

How do we achieve this? Most Bayesian nonparametric models are built on completely random measures (Kingman 1967) and the independence of the weights is an artifact of this construction. To create correlated random measures, we infuse a Gaussian process (Rasmussen and Williams 2005) into the construction with a latent kernel between components. This lets us relax the strict independence assumptions. The details involve showing how to use the Gaussian process in concert with the Poisson process, and without sacrificing the technicalities needed to define a proper random measure. As a result of the general construction, we can build correlated variants of many hierarchical Bayesian nonparametric models.

We will describe four correlated random measures. The first is a correlated nonparametric version of Poisson factorization (Canny 2004; Gopalan et al. 2014). This is a model of count data, organized in a matrix, and it will be the model on which we focus our study. We show how to derive an efficient variational inference algorithm to approximate the posterior and use it to analyze both medical data and text data. We also describe a correlated analog of the beta process (Hjort 1990) and two correlated binary latent feature models, each expanding on the hierarchical beta-Bernoulli process (Griffiths and Ghahramani 2006; Thibaux and Jordan 2007). We note that the discrete infinite logistic normal model in Paisley, Wang, and Blei (2012) is a normalized correlated random measure, a correlated adaptation of the hierarchical Dirichlet process (Teh et al. 2006).

Related work. Correlated random measures (CorrRMs) can capture general covariance between the measure of two sets, while also being atomic and extendible to hierarchical models that share atoms. In the Bayesian nonparametric literature, researchers have proposed several other random measures with covariance. We survey this work.

Cox and Isham (1980) introduced the Cox process, a Poisson random measure whose mean measure is also stochastic. Cox processes can capture covariance if the stochastic mean measure exhibits covariance. Unlike CorrRMs, however, Cox processes do not allow for noninteger atom weights. Furthermore, most common examples of Cox processes, such as the log-Gaussian Cox process (Møller, Syversveen, and Waagepetersen 1998), do not allow for atom sharing. We note that CorrRMs share the doubly stochastic construction of the Cox process; in Appendix A.1 we show that CorrRMs can be alternatively viewed as a stochastic transformation of a Poisson process.

Determinantal point processes (Borodin 2009) are point processes where the number of points is related to the determinant of a kernel function. Determinantal point processes exhibit "anti-clumping" (i.e., negative correlation) because atoms that are close together will not appear together. In contrast, hierarchical correlated random measures do not rely on the atom values to determine their correlation, and can capture both negative and positive correlation among their weights.

Doshi-Velez and Ghahramani (2009) presented a specific correlated feature model by positing a higher level grouping of features. In their model, observations exhibit correlation through these groups. However, groups only contain features and thus these features can only express positive correlation. In the latent feature models based on CorrRMs, the latent locations of two features can induce a negative correlation between their co-occurrence.

Ammann and Thall (1978) studied infinitely divisible random measures that are not completely random. These measures are referred to as having "aftereffects" in that every atom in the random measure has more effect on the measure than just its size. Correlated random measures are random measures with aftereffects, but they are not necessarily infinitely divisible.

As we have mentioned, the discrete infinite logistic normal (DILN; Paisley, Wang, and Blei 2012), a Bayesian nonparametric topic model, is a normalized instance of a correlated random measure. DILN first generates top level shared atoms from a Dirichlet process, along with latent locations for each. It then draws each document with a gamma process from those atoms and a Gaussian process evaluated at their locations. Finally, it convolves these processes and normalizes to form a probability measure. We discuss DILN in detail in Section 4.3.

Finally, there has been a lot of research in Bayesian nonparametrics about dependent random measures, originating from the work by MacEachern (1999), broadly surveyed by Foti and Williamson (2015), and used in applications such as for dynamic ordinal data (DeYoreo and Kottas 2017), neuron spikes (Gasthaus et al. 2009), and images (Sudderth and Jordan 2009). Dependent random measures select atoms for each observation through a priori covariates, such as a timestamp associated with the observation. Atoms are correlated, but only

through these observed covariates. The main ideas behind correlated random measures and dependent random measures are different. Correlations in CorrRMs are not based on side information, but rather are recovered through a random function associated with each observation. One dependent random measure that is close in construction to correlated random measures is the dependent Poisson process thinning measure by Foti et al. (2013). This measure can be reinterpreted as a type of correlated random measure; we discuss this connection with technical details in Section 3. Another construction, compound random measures (Griffin and Leisen 2017), builds dependent random measures by using score functions to generate a set of measures conditional on a shared Poisson process. Compound random measures and the dependent Poisson process thinning measure share with our approach the idea of separating out the atom generation from the independence breaking portion.

2. Background: Completely Random Measures

In this section, we review random measures (Kingman 1967; Cinlar 2011). We describe the Poisson random measure, completely random measures, and normalized random measures. This sets the stage for our construction of the correlated random measure in Section 3.

A random measure M is a stochastic process that is indexed by a sigma algebra. Let $(E, \mathcal{E})$ be a measurable space, for example, E is the real line and $\mathcal{E}$ are the Borel sets. A random measure is a collection of random variables $M(A) \in [0, \infty]$, one for each set $A \in \mathcal{E}$. The expectation of a random measure is called the mean measure, which we denote $\nu(A) \triangleq \mathrm{E}[M(A)]$.

One subclass of random measures is the class of completely random measures (Kingman 1967). A completely random measure is a random measure $M(\cdot)$ such that for any disjoint finite collection of sets $A_1, A_2, \ldots, A_n$, the corresponding realizations of the measure on those sets $M(A_1), M(A_2), \ldots, M(A_n)$ are independent random variables. Completely random measures encompass many of the constructions in Bayesian nonparametric statistics. Some examples include the Poisson process, the beta process (Hjort 1990), the Bernoulli process (Thibaux and Jordan 2007), and the gamma process (Ferguson 1973).

We begin by describing the simplest example of a completely random measure, the Poisson random measure. The Poisson random measure is constructed from a Poisson process. It is characterized solely by its mean measure $\nu(\cdot) : \mathcal{E} \to [0, \infty]$, which is an arbitrary measure on $(E, \mathcal{E})$. The complete characterization of a Poisson random measure $M(\cdot)$ is that the marginal distribution of $M(A)$ is a Poisson with rate $\nu(A)$.

We represent a Poisson random measure with a set of atoms a_i in E and a sum of delta measures on those atoms (Cinlar 2011),

$$M(A) = \sum_{i=1}^{\infty} \delta_{a_i}(A).$$

The delta measure $\delta_{a_i}(A)$ equals one when $a_i \in A$ and zero otherwise. Note there can be a countably infinite set of atoms, but only if $\nu(E) = \infty$. (This fact follows from the marginal distribution $M(E) \sim \text{Poisson}(\nu(E))$ and that a Poisson random variable with rate equal to ∞ is ∞ almost surely.) The distribution of the atoms comes from the mean measure $\nu(\cdot)$. For each finite measurable set A, the atoms in A are distributed according to $\nu(\cdot)/\nu(A)$.

We now expand the simple Poisson random measures to construct more general completely random measures. Consider a Poisson process on the cross product of E and the positive reals, $E \times \mathbb{R}_+$. It is represented by a set $\{(a_i, w_i)\}_{i=1}^{\infty}$; each pair contains an atom a_i and corresponding weight $w_i \in \mathbb{R}_+$. The completely random measure is

$$M(A) = \sum_{i=1}^{\infty} w_i \delta_{a_i}(A). \tag{1}$$

This Poisson process is characterized by its mean measure, called the Levy measure, which is defined on the corresponding cross product of sigma algebras,

$$\nu(\cdot, \cdot) : \mathcal{E} \times \mathcal{B}(\mathbb{R}_+) \to [0, \infty].$$

We note that completely random measures also have fixed components, where the atoms are fixed in advance and the weights are random. But we will not consider fixed components here.

We call the process homogenous when the Levy measure factorizes, $\nu(A, R) = H(A)\hat{\nu}(R)$; we call H the base measure. For example, in a nonparametric mixture of Gaussians the base measure is a distribution on the mixture locations (Escobar and West 1995); in a nonparametric model of text, the base distribution is a Dirichlet over distributions of words (Teh et al. 2006).

We confirm that $M(\cdot)$ in Equation (1) is a measure. First, $M(\emptyset) = 0$. Second, $M(A) \geq 0$ for any A. Finally, $M(\cdot)$ satisfies countable additivity. Define A to be the union of disjoint sets $\{A_1, A_2, \ldots\}$. Then $M(A) = \sum_k M(A_k)$. This follows from a simple argument,

$$\sum_{k=1}^{\infty} M(A_k) = \sum_{k=1}^{\infty} \sum_{i=1}^{\infty} w_i \delta_{a_i}(A_k) = \sum_{i=1}^{\infty} w_i \sum_{k=1}^{\infty} \delta_{a_i}(A_k)$$
$$= \sum_{i=1}^{\infty} w_i \delta_{a_i}(A) = M(A).$$

We used Tonelli's theorem to interchange the summations.

One example of a completely random measure is the gamma process (Ferguson 1973). It has Levy measure

$$\nu(da, dw) \triangleq H(da) e^{-cw} / w dw.$$

This is called the gamma process because if $M \sim \text{Gamma—Process}(H, c)$ the random measure $M(A)$ on any set $A \in \mathcal{E}$ is gamma distributed $M(A) \sim \text{Gamma}(H(A), c)$, where $H(A)$ is the shape and c is the rate (Cinlar 2011). The gamma process has an infinite number of atoms—its Levy measure integrates to infinity—but the weights of the atoms are summable when the base measure is finite ($H(E) < \infty$) because $\mathrm{E}[M(E)] = \frac{H(E)}{c}$. Finally, when $M(E) < \infty$, we can normalize a completely random measure to obtain a random probability measure. For example, we construct the Dirichlet process (Ferguson 1973) by normalizing the gamma process.

3. Correlated Random Measures

The main limitation of a completely random measure is articulated in its definition—the random variables $M(A_i)$ are

independent. (Because they are normalized, random probability measures exhibit some negative correlation between the $M(A_i)$, but cannot capture other types of relationships between the probabilities.) This limitation comes to the fore particularly when we see repeated draws of a random measure, such as in hierarchical Bayesian nonparametric models (Teh and Jordan 2010). In these settings, we may want to capture and infer a correlation structure among $M(A_i)$ but cannot do so with the existing methods (e.g., the hierarchical Dirichlet process). To this end, we construct correlated random measures. Correlated random measures build on completely random measures to capture rich correlation structure between the measure at disjoint sets, and this structure can be estimated from data.

We built completely random measures from a Poisson process by extending the space from simple atoms (in the Poisson process) to the space of atoms and weights (in a completely random measure). We build correlated random measures from completely random measures by extending the space again. As for a completely random measure, there is a set of atoms and uncorrelated weights. We now further supply each tuple with a "location," a vector in $\mathbb{R}^d$, and extend the mean measure of the Poisson process appropriately. A correlated random measure is built from a Poisson process on the extended space of tuples $\{(a_i, w_i, \ell_i)\}_{i=1}^{\infty}$.

In the completely random measure of Equation (1), the uncorrelated weights w_i give the measure at each atom. In a correlated random measure, there is an additional layer of variables x_i, called the *transformed weights*. These transformed weights depend on both the uncorrelated weights w_i and a random function on the locations $F(\ell_i)$. In the random measure, they are used in place of the uncorrelated weights,

$$M(A) = \sum_{i=1}^{\infty} x_i \delta_{a_i}(A). \tag{2}$$

It is through the random function $F(\cdot)$, which is drawn from a Gaussian process (Rasmussen and Williams 2005), that the weights exhibit correlation.

We first review the Gaussian process (GP) and then describe how to construct the transformed weights. A Gaussian process is a random function $F(\ell_i)$ from $\mathbb{R}^d \to \mathbb{R}$. It is specified by a positive-definite kernel function (This means that for a finite collection of inputs, the kernel produces a positive definite matrix.) $K(\ell_i, \ell_j)$ and mean function $\mu(\ell_i)$. The defining characteristic of a GP is that each joint distribution of a collection of values is distributed as a multivariate normal,

$$(F(\ell_1), \ldots, F(\ell_n)) \sim \mathcal{N}(m, \Sigma),$$

where $m_i = \mu(\ell_i)$ and $\Sigma_{ij} = K(\ell_i, \ell_j)$.

In a correlated random measure, we draw a random function from a GP, evaluate it at the locations of the tuples ℓ_i, and use these values to define the transformed weights. We specify the *transformation distribution* of $x_i \in \mathbb{R}^+$ denoted $T(x_i \mid w_i, F(\ell_i))$. It depends on both the uncorrelated weights w_i and the GP evaluated at ℓ_i. For example, one transformation distribution we will consider below is the gamma,

$$x_i \sim \text{Gamma}(w_i, \exp\{-F(\ell_i)\}). \tag{3}$$

But we will consider other transformation distributions as well. What is important is that the x_i are positive random variables, one for each atom, that are correlated through their dependence on the GP F.

We have now fully defined the distribution of the transformed weights x_i that are used in the correlated random measure of Equation (2). We emphasize that in a completely random measure the weights are independent. The arguments that $M(\cdot)$ is a measure, however, only relied on its form, and not on the independence of the weights (see Equation (2)).

In summary, we build a correlated random measure by specifying the following: the mean measure of a Poisson process on atoms, weights, and locations $\nu(da, dw, d\ell)$; a kernel function $K(\ell_i, \ell_j)$ between latent locations and a mean function $m(\ell_i)$; and the conditional transformation distribution $T(\cdot \mid w_i, F(\ell_i))$ over positive values. With these elements, we draw a correlated random measure as follows:

$$\{(a_i, w_i, \ell_i)\}_{i=1}^{\infty} \sim \text{Poisson-Process}(\nu) \tag{4}$$

$$F \sim \text{Gaussian-Process}(m, K) \tag{5}$$

$$x_i \sim T(\cdot \mid w_i, F(\ell_i)). \tag{6}$$

The random measure $M(\cdot)$ is in Equation (2).

Before turning to some concrete examples, we set up some useful notation for correlated random measures. We denote the infinite set of tuples from the Poisson process (Equation (4)) with

$$\mathcal{C} \triangleq \{(a_i, w_i, \ell_i)\}_{i=1}^{\infty}.$$

Given these tuples, the process for generating a correlated random measure first draws from a Gaussian process (Equation (5)), then transforms the weights (Equation (6)), and finally constructs the measure from an infinite sum (Equation (2)). We shorthand this process with $M \sim \text{CorrRM}(\mathcal{C}, K, \mu, T)$. (As in the construction of completely random measures, correlated random measure can also include fixed components, where the tuples are fixed, but the x_i are random.)

We note that correlated random measures generalize completely random measures. Specifically, we can construct a completely random measure from a correlated random measure by setting the mean measure $\nu(dx, dw, d\ell)$ to match the corresponding completely random measure mean measure $\nu(dx, dw)$ (i.e., the location distribution does not matter) and asserting that $x_i = w_i$ with probability one.

With the full power of the correlated random measure, we can construct correlated versions of common random measures such as the gamma process, the beta process, and normalized measures such as the Dirichlet process. We give two examples below.

Example: Correlated Gamma process. We discussed the gamma process as an example of a completely random measure. We now extend the gamma process to a correlated gamma process. First, we must extend the mean measure to produce atoms, weights, and locations. We specify an additional distribution of locations $L(\ell)$—we typically use a multivariate Gaussian—and expand the mean measure of the gamma process to a product,

$$\nu(da, dw, d\ell) = L(d\ell)H(da)e^{-cw}/w dw.$$

Second, for the transformation distribution, we choose the gamma in Equation (3). Finally, we define the GP parameters, the kernel $K(\ell_i, \ell_j) = \ell_i^\top \ell_j$ and a zero mean $\mu(\ell_i) = 0$. With these components in place, we draw from Equation (4), Equation (5), and Equation (6). This is one example of a correlated random measure.

Example: Correlated Dirichlet process. We can normalize the measure to construct a *correlated random probability measure* from a correlated random measure. If $M(\cdot)$ is a correlated random measure, then

$$G(A) = M(A)/M(E)$$

is a correlated random probability measure. As we discussed in Section 2, the Dirichlet process is a normalized gamma process (Ferguson 1973). When we normalize the correlated gamma process, we obtain a correlated Dirichlet process.

This construction requires that $M(E) < \infty$. Proposition 2 in Appendix A.2 describes conditions for well-defined normalization (i.e., $M(E) < \infty$) in correlated random measures. Roughly, these conditions require finiteness of the expected value of the transformed weights against the product of the transformation distribution and the Levy measure. This condition plays a central role in constructing useful Bayesian nonparametric models.

The correlation structure. Finally, we calculate the correlation structure of a correlated random measure. Consider a measure, $M \sim \text{CorrRM}(\mathcal{C}, K, m, T)$. To understand the nature of the correlation, we compute $\text{cov}(M(A), M(B))$ for two sets A and B.

We express this covariance in terms of the covariance between atom weights,

$$\text{cov}(M(A), M(B)) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \text{cov}(X_i, X_j)\delta_{a_i}(A)\delta_{a_j}(B).$$

In other words, the covariance between the measure of two sets is the sum of the covariances between the transformed weights of the atoms in the two sets. In a completely random measure, the atom covariance is zero except when $a_i = a_j$. Thus, its covariance depends on the overlap between sets, and nothing else. In a correlated random measure, however, there may be nonzero covariance between the transformed weights.

For now we are holding the underlying Poisson process $\mathcal{C}$ fixed, that is, the atoms, untransformed weights, and locations. The covariance between transformed weights is

$$\text{cov}(X_i, X_j \mid \mathcal{C}) = \text{E}[X_i X_j] - \text{E}[X_i]\text{E}[X_j]. \tag{7}$$

These expectations are driven by two sources of randomness. First there is a Gaussian process F, a random function evaluated at the fixed locations. Second there is the transformation distribution T. This is a distribution of an atom's transformed weight, conditional on its untransformed weight and the value of the Gaussian process at its location (see Equation (6)).

Using iterated expectation, we write the conditional covariance in Equation (7) in terms of the conditional mean of the transformed weights, $\mu_i \triangleq \text{E}[X_i \mid F(\ell_i), w_i]$. This is a function of the Gaussian process F. We rewrite the conditional covariance,

$$\text{cov}(X_i, X_j \mid \mathcal{C}) = \text{E}[\mu_i \mu_j] - \text{E}[\mu_i]\text{E}[\mu_j], \tag{8}$$

where the expectations are taken with respect to the Gaussian process. For the first term, the distribution is governed by the distribution of the pair $(F(\ell_i), F(\ell_j))$, which is a bivariate normal. For the second term, the marginals are governed by $F(\ell_i)$ and $F(\ell_j)$, which are univariate normal distributions.

4. Hierarchical Correlated Random Measures

A correlated random measure takes us from a set of tuples to a random measure by way of a Gaussian process and a transformation distribution. When used in a downstream model of data, we can infer the latent correlation structure from repeated realizations of measures from the same set of tuples. It is thus natural to build *hierarchical correlated random measures*. Hierarchical correlated random measures are the central use of this new construction.

In a hierarchical correlated random measure, we first produce a set of tuples $\{(a_i, w_i, \ell_i)\}_{i=1}^{\infty}$ from a Poisson process and then reuse that set in multiple realizations of a correlated random measure. In each realization, we fix the tuples (weights, atoms, and locations) but draw from the Gaussian process anew; thus we redraw the transformed weights for each realization.

As for the simple correlated random measure, we first specify the mean measure of the Poisson process $\nu(\cdot)$, the kernel and mean for the Gaussian process $K(\cdot, \cdot)$, and the conditional transformation distribution $T(\cdot \mid w_i, G(\ell_i))$. We then draw n hierarchical correlated random measures as follows:

$$\begin{aligned} \mathcal{C} &\sim \text{Poisson-Process}(\nu) \\ M_j(A) &\sim \text{CorrRM}(\mathcal{C}, m, K, T). \end{aligned}$$

This is a hierarchical Bayesian nonparametric model (Teh and Jordan 2010). There are multiple random measures M_j. Each shares the same set of atoms, locations, and weights, but each is distinguished by its own set of transformed weights. (In our empirical study of Section 6, we will also endow each with its own mean function to the Gaussian process, $m_j(\cdot)$. Here, we omit this detail to keep the notation clean.) The correlation structure of these transformed weights is shared across measures. We note that this construction generalizes the discrete infinite logistic normal (Paisley, Wang, and Blei 2012), which is an instance of a normalized correlated random measure.

We use this construction in a model of groups of observations y_j, for which we must construct a likelihood conditional on the correlated RM. To construct a likelihood, many hierarchical Bayesian nonparametric models in the research literature use the integral with respect to the random measure. (This is akin to an unnormalized "expectation.") Define $Ma \triangleq \int aM(da)$, and note that in a discrete random measure this integral is an infinite sum,

$$Ma = \sum_i x_i a_i. \tag{9}$$

The jth observations are drawn from a distribution parameterized from this sum, $y_j \sim p(\cdot \mid M_j a)$. For example, we will study models where Ma is a collection of rates for independent Poisson distributions.

We present several examples of hierarchical correlated random measures. First, we develop *correlated nonparametric Poisson factorization* (CNPF) for factorizing matrices of discrete data. This is the example we focus on for posterior inference (Section 5) and our empirical study (Section 6). We then illustrate the breadth of correlated random measures with two other examples, both of which are latent feature models that build correlations into the class of models introduced by Griffiths and Ghahramani (2006). Finally, we discuss the discrete infinite logistic normal (DILN) by Paisley, Wang, and Blei (2012). We show that DILN is a type of normalized correlated random measure.

4.1. Correlated Nonparametric Poisson Factorization

Bayesian nonparametric Poisson matrix factorization (Gopalan et al. 2014) combines gamma processes (Ferguson 1973) with Poisson likelihoods to factorize discrete data organized in a matrix. The number of factors is unknown and is inferred as a consequence of the Bayesian nonparametric nature of the model.

For concreteness we will use the language of *patients* getting *diagnoses* (e.g., patients going to the hospital and getting marked for medical conditions). In these data, each cell of the matrix y_{uj} is the number of times patient u was marked for diagnosis j. The goal is to factorize users into their latent "health statuses" and factorize items into their latent "condition groups." These inferences then let us form predictions about which unseen codes a patient might have. Though we focus our attention here on patients getting diagnoses, we emphasize that discrete matrices are widely found in modern data analysis problems. In our empirical study, we will also examine matrices of documents (rows) organized into word counts (columns) from a fixed vocabulary and user (rows) clicks over a fixed collection of items (columns).

We will use a hierarchical correlated random measure to model these data, where each group is a patient and the group-specific data are her vector of per-diagnosis counts. An atom a_i is a vector of positive weights for each diagnosis, drawn from independent gamma distributions, $H(a) = \text{Gamma}(\alpha, \beta)$. When the posterior of these atoms is estimated from diagnosis counts, they will represent semantic groups of conditions such as "diabetes," "heart disease," or 'cancer." Table 1 displays some of the atoms inferred from a set of patients from the Mayo Clinic.

In using a correlated random measure, the idea is that patients' expression for these conditions are represented by the per-group weights x_i. Intuitively, these exhibit correlation. A patient who has "heart" conditions is more likely to also have "vascular" conditions than "cancer." (To be clear, these groupings are the latent components of the model. There are an unbounded number of them, they are discovered in posterior inference, and their labels are not known.) Using these correlations, and based on her history, a correlated model should better predict which diagnoses a patient will have.

We now set up the model. We set the mean measure for the shared atoms to be

$$\nu(da, dw, d\ell) \triangleq \text{Gamma}(da, \alpha, \beta) \times \text{Normal}(d\ell, 0, I_d\sigma_l^2) e^{-cw}/w dw.$$

We define the GP mean function to be a per-patient constant $m_u(\ell_i) = \mu_u$, where $\mu_u \sim \mathcal{N}(0, \sigma_m^2)$. These per-patient GP means account for data where some patients tend to be sicker than others. We define the GP kernel function to be $K(\ell_i, \ell_j) = \ell_i^\top \ell_j$.

Finally, we consider two different transformation distributions. The first transformation distributions is as in Equation (3),

$$x_i \sim \text{Gamma}(w_i, \exp\{-F(\ell_i)\}).$$

The second is

$$x_i \sim \text{Gamma}\left(w_i, \frac{1}{\log(1 + \exp\{F(\ell_i)\}}\right),$$

where $\log(1 + \exp(\cdot))$ is known as the softplus function. With these definitions, we can compute the conditional covariance for x_i and x_j using Equation (8). Table 4 displays some of positive correlations between atoms found on patient diagnosis counts. These correlations are captured by the locations, which are shared across patients, associated with each atom. Atoms with positive covariance in this model will have inferred locations that have a large inner product.

With these components in place, correlated nonparametric Poisson factorization is

$$\begin{aligned} \mathcal{C} &\sim \text{Poisson-Process}(\nu) \\ \mu_u &\sim \mathcal{N}(0, \sigma_m^2) \\ M_u &\sim \text{CorrRM}(\mathcal{C}, \mu_u, K, T) \\ y_u &\sim p(\cdot \mid M_u a). \end{aligned}$$

The distribution of y_u is a collection of Poisson variables, one for each diagnosis j, where

$$y_{uj} \sim \text{Poisson}\left(\sum_{i=1}^{\infty} x_{ui} a_{ij}\right). \tag{10}$$

Recall that the atoms a_i are each a vector of gamma variables, one per diagnosis, and so a_{ij} is the value of atom i for diagnosis j. For this model to be well defined each rate in Equation (10) must be finite. Using proposition 2 in Appendix A.2, it is finite almost surely if $\sigma_l < 1$.

The sum that defines the rate of y_{ui} is an infinite sum of patient weights and condition weights. Thus, this model amounts to a factorization distribution for y_{ui}. Given observed data, the posterior distribution of the atoms a_i and transformed patient weights x_i gives a mechanism to form predictions. Note that the atoms a_i are shared across patients, but through x_i each patient exhibits them to different degree. We discuss how to approximate this posterior distribution in Section 5.

4.2. Correlated Latent Feature Models

Mixture models, such at the Dirichlet process mixture (Antoniak 1974), are the most commonly used Bayesian nonparametric model. In mixture models, each observation exhibits only a single class. Many data, such as images of multiple objects, are better characterized as belonging to multiple classes. Latent feature models posit that each observation is associated with some number of latent classes, taken from a set

of features shared by all observations. For each observation, its likelihood depends on parameters attached to its active features (e.g., a sum of those parameters). Examples of latent feature models include factorial mixtures (Ghahramani 1995) and factorial hidden Markov models (Ghahramani and Jordan 1996). (Latent feature models are closely connected to spike and slab priors (Ishwaran and Rao 2005).)

Bayesian nonparametric latent feature models allow the number of features to be unbounded. As an example, consider analyzing a large dataset of images. Latent features could correspond to image patches, such as recurring objects that appear in the images. In advance, we might not know how many objects will appear in the dataset. BNP latent feature models attempt to solve this problem. BNP latent feature models have been used in many domains such as image denoising (Zhou et al. 2011) and link prediction in graphs (Miller, Griffiths, and Jordan 2009).

The most popular BNP latent feature model is the hierarchical beta-Bernoulli process (Thibaux and Jordan 2007). This process was originally developed as the Indian Buffet process, which marginalized out the beta process (Griffiths and Ghahramani 2006). Before developing the correlated version, we review the beta-Bernoulli process.

The beta process is a completely random measure with atom weights in the unit interval (0,1). Its Levy measure is

$$\nu(da, dw) = H(da)\alpha w^{-1}(1-w)^{\alpha-1}.$$

We use the beta process in concert with the Bernoulli process, which is a completely random measure parameterized by a random measure with weights in the unit interval, that is, a collection of atoms and corresponding weights. A draw from a Bernoulli process selects each atom with probability equal to its weight. This forms a random measure on the underlying space, where each weight is one or zero (i.e., where only a subset of the atoms are activated). Returning to latent feature models, the beta-Bernoulli process is

$$\begin{aligned} B &\sim \text{Beta-Process}(H, \alpha) \\ B_n &\sim \text{Bernoulli-Process}(B) \\ y_n &\sim p(\cdot \mid B_n) \end{aligned}$$

The beta process generates the feature atoms; the Bernoulli processes chooses which features are active in each observation.

This model is built on completely random measures. Thus, the appearances of features in each observation are independent of one other. Correlated random measures relax this assumption. Consider a latent feature model of household images with image patch features. The completely random assumption here implies that the appearance of a spoon is independent of the appearance of a fork. Our construction can account for such dependencies between the latent features. Below we will give two examples of correlated nonparametric latent feature models, one based on the beta process and the other based on the gamma process.

One method to develop a correlated beta-Bernoulli process is to define transformed weights at the Bernoulli process level. We define the transformation distribution to be

$$x_{ni} \sim \text{Bernoulli}(\sigma(\sigma^{-1}(w_i) + F(\ell_i))),$$

where σ is the sigmoid function $\sigma(x) = 1/(1+\exp\{-x\})$. Thus, the beta-Bernoulli correlated latent feature model is

$$\begin{aligned} \mathcal{C} &\sim \text{Poisson-Process}(H(da)L(d\ell)\alpha w^{-1}(1-w)^{\alpha-1}) \\ M_n &\sim \text{CorrRM}(\mathcal{C}, \mu, K, T). \end{aligned}$$

(We defer defining μ and K, as they will be application specific.)

We do not need to use the beta process to define a correlated latent feature model; what is important is that the per-observation weights are either one or zero. For example, if the top level process is a gamma process, which produces positive weights, then we can define the transformation distribution to be

$$x_{ni} \sim \text{Bernoulli}\left(\frac{w_i \exp(F(\ell_i))}{1 + w_i \exp(f(\ell_i))}\right).$$

The resulting gamma-Bernoulli correlated latent feature model is

$$\begin{aligned} \mathcal{C} &\sim \text{Poisson-Process}(H(da)L(d\ell)e^{-cw}/w dw) \\ M_n &\sim \text{CorrRM}(\mathcal{C}, \mu, K, T). \end{aligned}$$

The beta-Bernoulli process uses only a finite number of features to generate a finite number of observations. In Appendix A.3, we give some conditions under which the correlated latent feature models do the same.

4.3. Discrete Infinite Logistic Normal

The correlated random measure construction that we developed generalizes the discrete infinite logistic normal (DILN; Paisley, Wang, and Blei 2012). DILN is an example of a normalized hierarchical correlated random measure; its atom weights come from a normalized gamma random measure, that is, a Dirichlet process.

DILN was developed as a Bayesian nonparametric mixed-membership model, or topic model, of documents. In DILN, each document mixes a set of latent topics (distributions over terms), where the per-document topic proportions can exhibit arbitrary correlation structure. This is in contrast to a hierarchical Dirichlet process topic model (Teh et al. 2006), where the topic proportions are nearly independent.

We will express DILN in terms of a correlated random measure. The observations w_{uj} are categorical variables, that is, word j in document u. Set the kernel $K(\ell_i, \ell_j) = \ell_i^\top \ell_j$, and set α and β to be positive hyperparameters. Set the transformation distribution to be

$$x_i \sim \text{Gamma}(\beta w_i, \exp\{-F(\ell_i)\}).$$

With our construction, DILN is

$$\begin{aligned} \mathcal{C} &\sim \text{Dirichlet-Process}(\alpha, H(da) \times \mathcal{N}(d\ell, 0, \sigma_l^2 I_d)) \\ M_u &\sim \text{Normalized-CorrRM}(\mathcal{C}, 0, K, T) \\ z_{uj} &\sim M_u \\ w_{uj} &\sim z_{uj}. \end{aligned}$$

Note that the shared tuples come from a Dirichlet process, that is, a normalized gamma process. When modeling documents, the base distribution over atoms $H(da)$ is a Dirichlet distribution over the vocabulary.

This is a mixed-membership model—there is an additional layer of hidden variables z_{uj}, drawn from the random measure, before drawing observations w_{uj}. These hidden variables z_{ui} will be atoms, that is, distributions over the vocabulary, from the set of shared tuples and drawn with probability according to the per-document transformed weights. Each observation w_{uj} is drawn from the distribution over terms given in its atom z_{uj}.

Paisley, Wang, and Blei (2012) showed that the normalization step is well defined when $\sigma_l < 1$. Viewing DILN through the lens of correlated random measures makes clear what can be changed. For example, the top level choice of the Dirichlet process is not critical. It could be any random measure that places finite total mass, such as a gamma process or a beta process.

4.4. Connection to Dependent Random Measures

Finally, we discuss the detailed connection between correlated random measures and dependent random measures (MacEachern 1999). Dependent random measures are a collection of measures indexed by covariates. A broad class of dependent random measures can be created by thinning a Poisson process (Foti et al. 2013). Given a draw from a Poisson process $(a_i, w_i, \ell_i)_{i=1}^{\infty}$, where a are atoms, ℓ are locations in the covariate space, and w are weights, the thinned dependent random measure B for user u with covariate θ_u is

$$B_u(A) = \sum_{i=1}^{\infty} x_{ui}\delta_{a_i}(A)$$
$$x_{ui} \sim w_i \text{Bernoulli}(k(\theta_u, \ell_i)),$$

where k is a function from $T \times L \to [0, 1]$. This construction is related to CorrRMs. Consider the correlated random measure

$$x_{ui} \sim w_i \text{Bernoulli}(\sigma(F(\ell_i)))$$
$$M_u(A) \sim \text{CorrRM}\left((a_i, w_i, \ell_i)_{i=1}^{\infty}, m, K, T\right).$$

From this, we can see that $B_u(A) \overset{d}{=} M_u(A)$ when $\sigma(F_u(\ell_i)) = k(\theta_u, \ell_i)$. In other words, thinned dependent random measures are equivalent to a type of correlated random measure where the random function, F_u associated with each user is known and given by the covariate.

We note that dependent random measures map from covariates to measures. Thus they can be viewed as a type of measure-valued regression. In parallel, correlated random measures use latent covariates. In this sense, they can be viewed as measure-valued factor analysis.

5. Variational Inference for Correlated Nonparametric Poisson Factorization

Computing the posterior is the central computational problem in Bayesian nonparametric modeling. However, computing the posterior exactly is intractable. To approximate it, we use variational inference (Jordan et al. 1999; Wainwright and Jordan 2008), an alternative to Markov chain Monte Carlo. Variational inference has been used to approximate the posterior in many Bayesian nonparametric models (Kurihara, Welling, and Teh 2007; Doshi-Velez et al. 2009; Paisley and Carin 2009; Wang and Blei 2012) and has been of general interest in statistics (Braun and McAuliffe 2007; Faes, Ormerod, and Wand 2011; Ormerod and Wand 2012). Here, we develop a variational algorithm for correlated nonparametric Poisson matrix factorization (Section 4.1).

Variational inference turns approximate posterior computation into optimization. We set up a family of distributions over the latent variables $\mathcal{Q} = \{q(\cdot)\}$ and then find the member that minimizes the KL divergence to the exact posterior. Minimizing the KL divergence to the posterior is equivalent to maximizing the evidence lower bound (ELBO),

$$q^*(\cdot) = \arg\max_{q \in \mathcal{Q}} \text{E}_{q_\lambda(\xi)}[\log p(y, \xi) - \log q(\xi)], \qquad (11)$$

where ξ are the latent variables and y are the observations. In this article, we work with the mean-field family, where the approximating distribution fully factorizes. Each latent variable is independently governed by its own variational parameter.

To develop a variational method for CNPF, we give a constructive definition of the gamma process and introduce auxiliary variables for the Gaussian process. We then define the corresponding mean-field family and show how to optimize the corresponding ELBO.

Additional latent variables. We first give a constructive definition of a homogenous gamma process. We scale the stick breaking construction of Sethuraman (1994) as used by Gopalan et al. (2014); Zhou and Carin (2015). We define stick lengths v_k from a beta distribution and a scaling s from a gamma distribution. The weights of the gamma process w_k are from the following process,

$$s \sim \text{Gamma}(\alpha, c)$$
$$v_k \sim \text{Beta}(1, \alpha)$$
$$w_k = s\left(v_i \prod_{j=1}^{k-1}(1 - v_j)\right).$$

We treat the gamma shape α and rate c as latent variables (with gamma priors).

We adapt the auxiliary variable representation of zero-mean Gaussian processes with linear kernels (Paisley, Wang, and Blei 2012) to more general Gaussian processes. Suppose G_n is a Gaussian process with mean μ_n and a linear kernel. Let d be a standard Gaussian vector with same dimension as ℓ_k. We can write the process as

$$G(\ell_k) \overset{d}{=} \ell_k^\top d + \mu_n(\ell_k).$$

This lets us evaluate likelihoods without matrix inversion.

The mean-field family. With the latent variables for the gamma and Gaussian processes in hand, we now define the mean-field variational distribution. We use the following approximating family for each latent variable

$$q(x_{ku}) = \text{Gamma}\left(\alpha_{ku}^x, \beta_{ku}^x\right)$$
$$q(a_{ki}) = \text{Gamma}\left(\alpha_{ki}^a, \beta_{ki}^a\right)$$
$$q(s)q(V_k)q(\ell_k) = \delta_{\hat{s}}\delta_{\hat{V}_k}\delta_{\hat{\ell}_k}$$
$$q(d_u)q(\mu_u) = \delta_{\hat{d}_u}\delta_{\hat{\mu}_u}$$
$$q(\alpha)q(c) = \delta_{\hat{\alpha}}\delta_{\hat{c}},$$

where δ_r represents a point mass at r. As in prior work on variational inference for Bayesian nonparametrics, we use delta distributions in the top level stick components, scaling, and hyperparameters for analytic tractability (Liang et al. 2007; Paisley, Wang, and Blei 2012; Gopalan et al. 2014). (This corresponds to variational expectation-maximization, where the E step computes variational expectations and the M step takes MAP estimates of the latent variables with delta factors (Beal 2003).)

Bayesian nonparametric models contain an infinite number of latent variables. Following Blei and Jordan (2005), we truncate the variational approximation of the sticks V_k and associated tuples to T. In practice, it is straightforward to recognize if the truncation level is too small because all of the components will be populated in the fitted variational distribution. In our studies $T = 200$ was sufficient (Section 6).

The goal of variational inference is to find the variational parameters—the free parameters of q, such as $\hat{\ell}_k$—that maximize the evidence lower bound. In Appendix A.4, we describe how to optimize the ELBO (Equation (11)) with stochastic variational inference (Hoffman et al. 2013). Code will be made available on GitHub.

We have derived a variational inference algorithm for one example of a correlated random measure model. Deriving algorithms for other examples follows a similar recipe. In general, we can handle inference for covariance functions with inducing variables (Titsias 2009) and subsampling (Hensman, Fusi, and Lawrence 2013). Further, we can address models with intractable expectations– - for example, those arising from different transformation distributions or Levy measures—with recent methods for generic and nonconjugate variational inference (Salimans and Knowles 2013; Wang and Blei 2013; Ranganath, Gerrish, and Blei 2014).

6. Empirical Study

We study correlated nonparametric Poisson factorization (CNPF) and compare to its uncorrelated counterpart on a large text dataset and a large dataset of medical diagnosis codes. Quantitatively, we find that the correlated model gives better predictive performance. We also find that it reveals interesting visualizations of the posterior components and their relationships.

6.1. Study Details

Before giving the results, we describe the baseline models, the evaluation metric, and the hyperparameter settings.

Baseline models. As a baseline, we compare against the uncorrelated variant of Bayesian nonparametric Poisson factorization. As we mentioned in Section 3, uncorrelated random measures can be cast in the correlated random measure framework by setting a transformation distribution that does not depend on the Gaussian process.

Recall that x_{ik} is the weight for data point i on component k. In the simplest Bayesian nonparametric Poisson factorization model, the transformation distribution is

$$x_{ik} \sim \text{Gamma}(w_k, 1).$$

This is a two-layer hierarchical gamma process, and we abbreviate this model HGP. The first layer contains shared atoms and weights. The second layer is a gamma process for each data point (e.g., patient or document), with base measure given by the first layer's measure.

The second uncorrelated model places further hierarchy on the log of the scale parameter of the Gamma,

$$x_{ik} \sim \text{Gamma}(w_k, \exp(-m_i)).$$

Here $m_i \sim \text{Normal}(a, b)$, which captures variants in the row sums for each data point (i.e., how many total diagnoses for a patient or how many words for a document). We call this model the scaled HGP.

Appendix A.5 gives inference details for both uncorrelated models.

Evaluation metric. We compare models with held out perplexity, a standard metric from information retrieval that relates to held out predictive likelihood (Geisser 1975). We use the partial observation scenario that is now common in topic modeling (Wallach et al. 2009). The idea is to uncover components from most of the data, and then evaluate how well those components can help predict held out portions of new partially observed data.

For each dataset, we hold out 1000 examples (i.e., rows of the matrix). From the remaining examples, we run approximate posterior inference, resulting in approximate posterior components $\text{E}[a_k]$ that describe the data. With the 1000 held out examples, we then split each observation (i.e., columns) randomly into two parts, 90% in one part (y_{test}) and 10% in the other (y_{obs}). We condition on the y_{obs} (and that there is a test word) and calculate the conditional perplexity on y_{test}. A better model will assign the true observations a higher probability and thus lower perplexity. Formally, perplexity is defined as

$$\text{Perplexity} = \exp\left(\frac{-\sum_{y\in\text{held out}}\sum_{w\in y_{\text{test}}}\log p(w \mid y_{\text{obs}})}{N_{\text{held out words}}}\right).$$

Perplexity measures the average surprise of the test observations. The exponent is the average number of nats (base e bits) needed to encode the test sample.

For the models we analyze, we compute this metric as follows. For each held out data point, we hold the components fixed (i.e., $\text{E}_q[a_k]$) and use the 10% of observed columns to form a variational expectation of the per-data point weights $\text{E}_q[x_{ik}]$. In all models, we compute the held out probability of unobserved columns by using the multinomial conditioning property of Poisson. Conditional on there being a test observation, it is assigned to a particular column (e.g., a word or a diagnostic code) with probability equal to that column's normalized Poisson rates. Formally,

$$p(y_i = j) = \frac{\sum_k \text{E}_q[x_{ik}]\text{E}_q[a_{kj}]}{\sum_j\sum_k \text{E}_q[x_{ik}]\text{E}_q[a_{kj}]}.$$

We measure the probability of the y_{test} columns under this distribution. This evaluates how well the discovered components can form predictions in new and partially observed observations.

Hyperparameters. We set the hyperparameters on the base distribution to have shape 0.01 and rate 10.0. We set the truncation level T to be 200, and found that none of the studies

Table 2. A summary of the predictive results on the *New York Times*, the Mayo Clinic, and ArXiv clicks. The correlated models outperform both the uncorrelated models. Adding per observation scalings improves predictions.

Data	HGP	Scaled HGP	CNPF	Softplus-CNPF
NYT	3570	3283	2755	2768
Mayo Clinic	1251	877	779	780
ArXiv	5713	4076	2107	2120

required more than this. We set the dimensionality of the latent locations to be 25 and the prior variance to be $\frac{1}{250}$. We keep these hyperparameters fixed for all data.

In the algorithm, we use Robbins Monro learning rates, $(50 + t)^{-.9}$ for the text data and $(100 + t)^{-.9}$ for the medical codes, and click data.

6.2. Results

We evaluate our posterior fits on text, medical diagnosis data, and click data.

The New York Times. We study a large collection of text from the *New York Times*. Rows are documents; columns are vocabulary words; the cell y_{ij} is the number of times term j appeared in document i. After preprocessing, the data contains 100,000 documents over a vocabulary of 8000 words. Analyzing text data with a Poisson factorization model is a type of topic modeling (Blei 2012).

Table 2 summarizes the held-out perplexity. We find that the correlated model outperforms both of the uncorrelated models. Note that even in the uncorrelated model, adding a per-document scale parameter improves predictions.

Table 3. The top 10 pairs of negatively correlated components inferred from the *New York Times*. Each pair of components are highly unlikely to cooccur in an article.

israel, israeli, palestinian, jewish, peace league, players, sports, baseball, team
room, bedroom, bath, taxes, market news, book, magazine, editor, books
war, iraq, military, army, iraqi space, kim, koch, moon, nasa
rock, music, band, jones, album family, tax, board, paid, friend
water, plant, garden, plants, trees union, soviet, moscow, russian, gorbachev
island, water, beach, river, sea theater, broadway, play, show, production
building, housing, buildings, square, project news, book, magazine, editor, books
bush, administration, clinton, officials, house space, kim, koch, moon, nasa
room, bedroom, bath, taxes, market indian, atlantic, casino, trump, las
century, small, wine, place, white contract, los, angeles, league, chicago

The model also provides new ways to explore and summarize the data. Figure 1 is a graph of the positive correlation structure in the posterior for the top 50 components, sorted by frequency; Table 3 contains a list of the top 10 negative correlations. To explore the data, we compute correlations between components

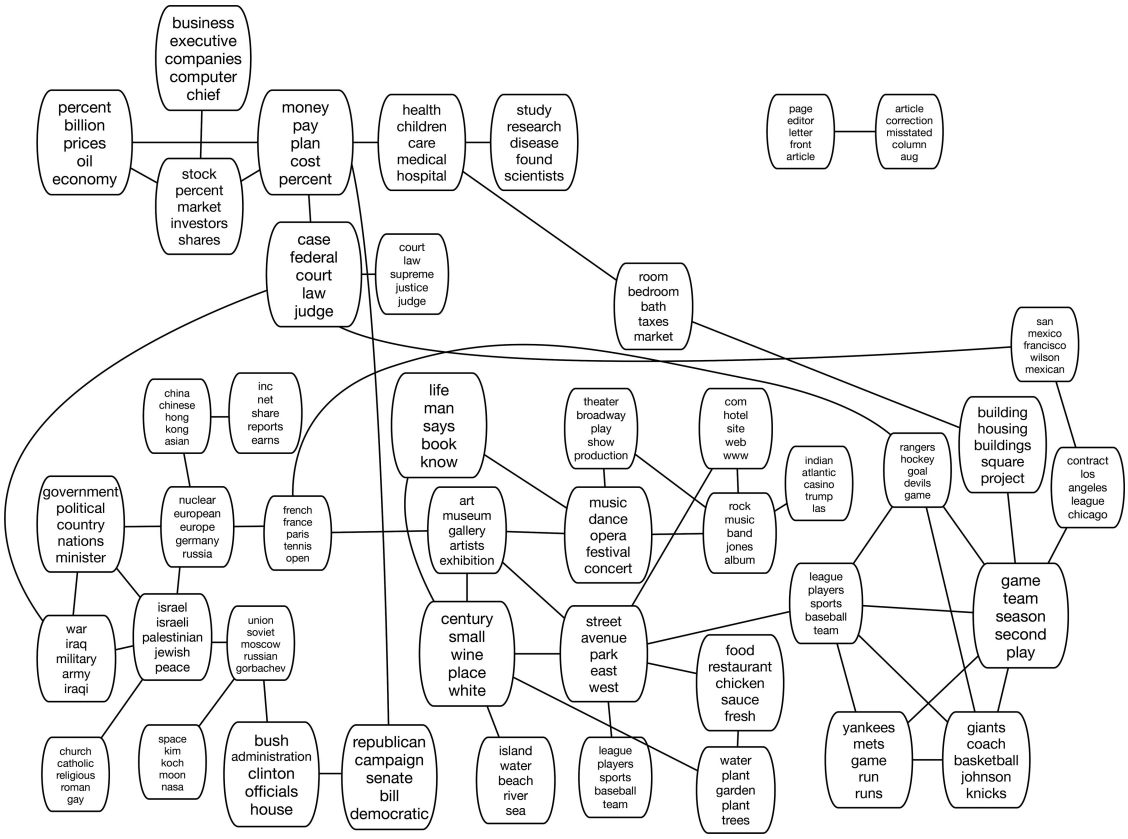

Figure 1. A graph of the latent component correlations found on the *New York Times*. This figure is best viewed on a computer with zoom. The sizes of the components are related to their frequency. The correlation structure consists of several tightly connected groups with sparse links between them.

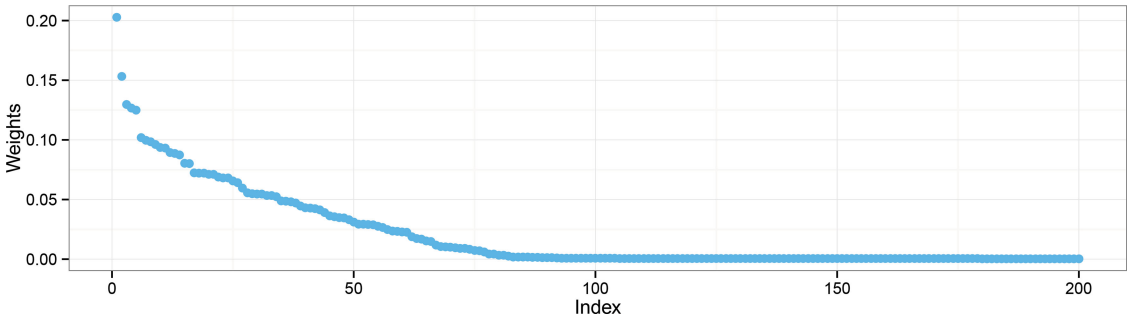

Figure 2. The weights in each tuple on the *New York Times* ordered by magnitude. Around 75 components are used.

by using their latent locations through the covariance function of the Gaussian process. For these fits, the covariance between ℓ_i and ℓ_j is $\ell_i^\top \ell_j$; the correlation between two components is thus

$$\rho_{km} = \frac{\ell_k^\top \ell_m}{\sqrt{\ell_k^\top \ell_k \, \ell_m^\top \ell_m}}.$$

We find the correlation structures contain highly connected groups, connected to each other by "glue," individual components that bridge larger groups. For example, the bottom left connected group of "international politics" is glued together with the top left group of "finance" through the "political parties" component and the "law" component.

As we said above, we set the truncation level of the approximate posterior to 200 components. Figure 2 plots the atom weights of these 200 components, ordered by size. The posterior uses about 75 components; the truncation level is appropriate.

Medical history from the Mayo Clinic. We study medical code data from the Mayo Clinic. This dataset contains of all the International Classification of Diseases 9 (ICD-9) diagnosis codes (also called billing codes) for a collection of patients over 3 years. The diagnosis codes mark medical conditions, such as chronic ischemic heart disease, pure hypercholesterolemia, and Type 2 diabetes. The entire collection contains 142,297 patients and 11,102 codes. Patients are rows in the matrix; codes are columns; each cell marks how many times the patient was assigned to the code.

Table 2 summarizes the held-out perplexity. Again, the correlated model does best. Further, as for text modeling, it is important to allow a patient-specific scale parameter to capture their relative health. Figure 3 plots the posterior sticks, ordered by size. The approximate posterior uses about 50 components, using the first 20 more heavily.

Table 1 contains the 20 most commonly used components. The components correspond to medically meaningful groups of conditions, such as obesity (12), type 2 diabetes (7), and breast malignancy (19). The top positive correlations are in Table 4. There are several meaningful correlations, such as depression and alcohol dependency, and using anticoagulants and hypertension/lipidemia. Note that the relationship between schizophrenia and type 2 diabetes is an active area of research in medicine (Suvisaari et al. 2008; Liu et al. 2013).

ArXiv click data. Finally, we examine user click data from the ArXiv, an online repository of research articles. The ArXiv initially focused on physics articles but has now expanded to many other domains, including statistics. This dataset contains the number of times each user clicked on an article; it spans 50,000 users and 20,000 articles. Building models of such data is useful, for example, to develop recommendation systems that find interesting articles to ArXiv readers.

As for the other data, we hold out some of the clicks and try to predict them. Table 2 summarizes the results. We find similar results as on our other two datasets. The correlated models outperform the uncorrelated models on predicting unseen clicks for new users. We find that the standard CNPF model outperforms the softplus CNPF on all of our datasets.

7. Discussion

We present correlated random measures. Correlated random measures enable us to construct versions of Bayesian nonparametric models that capture correlations between their components. We construct several examples of such models, and develop the inference algorithm in detail for one of them, correlated nonparametric Poisson factorization. With this model, we find that the correlated random measure improves predictions and produces interesting interpretable results.

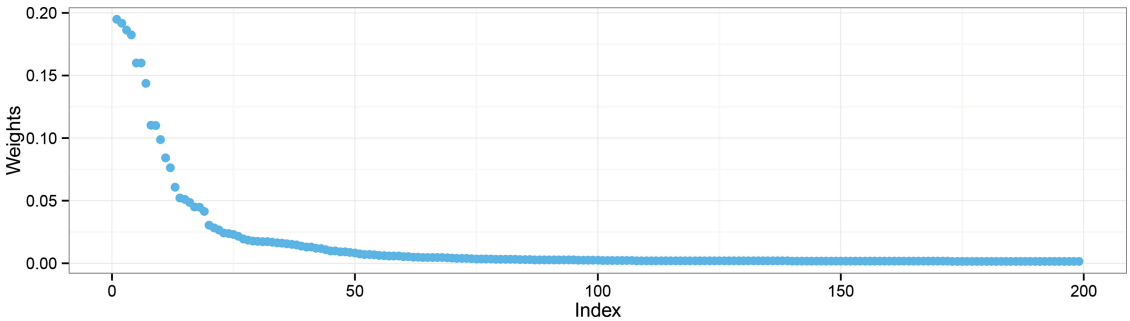

Figure 3. The weights in each tuple ordered by magnitude. Around 50 of components are used. Though similar in size to the NYT dataset, fewer components are used. The component usage has a steeper decline.

Table 4. The top 10 correlations among the heavily used components in the Mayo Clinic data. We find several medically meaningful relationships between latent components. For example, the relationships between obesity and type 2 diabetes is well established.

– Long-term anticoagulants, Atrial fibrillation, Hypertension, Congestive Heart Failure, Chronic airway obstruction – Heart valve replacement, Prostate cancer, Lung and bronchus cancer, Secondary bone cancer, Other lung disease
– Abnormality of gait, Personality change, Persistent mental disorders, Lack of coordination, Debility – Schizophrenia-Paranoid, Long-term use meds, Schizophrenia, Schizophrenia-paranoid-chronic, Drug monitor
– Attention deficit disorder with hyperactivity, Attention deficit disorder without hyperactivity, Adjustment disorder with disturbance of emotions and conduct, Opposition defiant disorder, Conduct disturbance – Schizophrenia-Paranoid, Long-term use meds, Schizophrenia, Schizophrenia-paranoid-chronic, Drug monitor
– Depression, Dysthymia, Anxiety state, Generalized anxiety disorder, Major depressive affective disorder – Alcohol dependence, Tobacco use disorder, Alcohol abuse, Other alcohol dependence-in remission, Other alcohol dependence-continuous
– Long-term anticoagulants, Atrial fibrillation, Hypertension, Congestive Heart Failure, Chronic airway obstruction – Diabetes-2, Hypertension, Hyperlipidemia, Uncontrolled Diabetes-2, Diabetes-2 with ophthalmic manifestations
– Hypertension, Hyperlipidemia, Coronary atherosclerosis, Prostate Cancer Screening, Vaccine for influenza – Long-term anticoagulants, Atrial fibrillation, Hypertension, Congestive Heart Failure, Chronic airway obstruction
– Heart valve replacement, Prostate cancer, Lung and bronchus cancer, Secondary bone cancer, Other lung disease – Female breast cancer, Personal history of breast cancer, Lymph cancer, Carcinoma in situ of breast, Lymphedema
– Depression, Dysthymia, Anxiety state, Generalized anxiety disorder, Major depressive affective disorder – Schizophrenia-Paranoid, Long-term use meds, Schizophrenia, Schizophrenia-paranoid-chronic, Drug monitor
– Diabetes-2, Hypertension, Hyperlipidemia, Uncontrolled Diabetes-2, Diabetes-2 with ophthalmic manifestations – Schizophrenia-Paranoid, Long-term use meds, Schizophrenia, Schizophrenia-paranoid-chronic, Drug monitor
– Mammogram, Routine medical exam, Lumbago, Cervical Cancer Screening, Hypothyroidism – Female breast cancer, Personal history of breast cancer, Lymph cancer, Carcinoma in situ of breast, Lymphedema

Random probability measures such as the Dirichlet process are consistent for density estimation, so why might one prefer a correlated random measure over a completely random measure? We conjecture that correlated random measures make more efficient use of the data. One promising avenue of future research is to study the rates of correlated random measures versus completely random measures.

Correlated random measures model latent correlations in the data, while dependent random measures model correlations based on observed covariates. Combining these two ideas to incorporate correlations both observed and latent yields a broader class of random measures that can model many real world phenomena. Another avenue of future research is to study this construction both methodologically and practically.

We define correlated random measures by combining Poisson and Gaussian processes. However, we note that other processes can also be used for the source of tuples (a_i, w_i, ℓ_i) and the random function. For example, the DILN model of Section 4.3 uses a Dirichlet process to form its tuples; another way to generate tuples would be through the Pitman–Yor process (Pitman and Yor 1997; Teh et al. 2006).

Similarly, though we used a Gaussian process to define a random function from latent locations to real values, there are other possibilities. For example, we can replace the GP with the student-T process (Shah, Wilson, and Ghahramani 2014). Or, if we restrict the latent locations to be positive then we can use them to subordinate, as an index to, another stochastic process, such as Brownian motion. Also, we could use discrete random functions to form feature groups. We leave these extensions for possible future research.

Supplementary Materials

The supplement contains propositions detailing finiteness and Laplace transforms of correlated random measures along with the derivation of a variational inference algorithm for a posterior inference.

Acknowledgments

The authors thank the reviewers for their helpful feedback and comments.

Funding

This work is supported by NSF IIS-1247664, ONR N00014-11-1-0651, DARPA FA8750-14-2-0009, DARPA N66001-15-C-4032, Adobe, NDSEG Fellowship, Porter Ogden Jacobus Fellowship, Seibel Foundation, and the Sloan Foundation.

References

Ammann, L. P., and Thall, P. F. (1978), "Random Measures With Aftereffects," *The Annals of Probability*, 6, 216–230. [418]

Antoniak, C. (1974), "Mixtures of Dirichlet Processes With Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2, 1152–1174. [422]

Beal, M. (2003), "Variational Algorithms for Approximate Bayesian Inference," Ph.D. dissertation, Gatsby Computational Neuroscience Unit, University College London. [425]

Blei, D. (2012), "Probabilistic Topic Models," *Communications of the ACM*, 55, 77–84. [426]

Blei, D., and Jordan, M. (2005), "Variational Inference for Dirichlet Process Mixtures," *Journal of Bayesian Analysis*, 1, 121–144. [425]

Borodin, A. (2009), "Determinantal Point Processes," *arXiv preprint arXiv:0911.1153*. [418]

Braun, M., and McAuliffe, J. (2007), "Variational Inference for Large-Scale Models of Discrete Choice," *Journal of American Statistical Association*, 105, 324–335. [424]

Canny, J. (2004), "GaP: A Factor Model for Discrete Data," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 122–129. [418]

Cinlar, E. (2011), *Probability and Stochastics*, New York: Springer. [419]

Cox, D. R., and Isham, V. (1980), *Point Processes*, Boca Raton, FL: Chapman Hall. [418]

DeYoreo, M., and Kottas, A. (2017), "Modeling for Dynamic Ordinal Regression Relationships: An Application to Estimating Maturity of Rockfish in California," *Journal of the American Statistical Association*, accepted, DOI:10.1080/01621459.2017.1328357. [418]

Doshi-Velez, F., and Ghahramani, Z. (2009), "Correlated Non-Parametric Latent Feature Models," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, AUAI Press, pp. 143–150. [418]

Doshi-Velez, F., Miller, K., Van Gael, J., and Teh, Y. (2009), "Variational Inference for the Indian Buffet Process," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 137–144. [424]

Escobar, M., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588. [419]

Faes, C., Ormerod, J. T., and Wand, M. P. (2011), "Variational Bayesian Inference for Parametric and Nonparametric Regression With Missing Data," *Journal of the American Statistical Association*, 106, 959–971. [424]

Ferguson, T. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230. [419,421,422]

Foti, N. J., Futoma, J. D., Rockmore, D. N., and Williamson, S. (2013), "A Unifying Representation for a Class of Dependent Random Measures," in *International Conference on Artificial Intelligence and Statistics*, pp. 20–28. [419,424]

Foti, N. J., and Williamson, S. (2015), "A Survey of Non-Exchangeable Priors for Bayesian Nonparametric Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 359–371. [418]

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011), "A Sticky hdp-hmm With Application to Speaker Diarization," *The Annals of Applied Statistics*, 5, 1020–1056. [417]

Gasthaus, J., Wood, F., Gorur, D., and Teh, Y. W. (2009), "Dependent Dirichlet Process Spike Sorting," in *Advances in Neural Information Processing Systems*, pp. 497–504. [418]

Geisser, S. (1975), "The Predictive Sample Reuse Method With Applications," *Journal of the American Statistical Association*, 70, 320–328. [425]

Ghahramani, Z. (1995), "Factorial Learning and the em Algorithm," in *Advances in Neural Information Processing Systems*, pp. 617–624. [423]

Ghahramani, Z., and Jordan, M. (1996), "Factorial Hidden Markov Models," *Advances in Neural Information Processing Systems*, 31, 472–478. [423]

Gopalan, P., Ruiz, F. J., Ranganath, R., and Blei, D. M. (2014), "Bayesian Nonparametric Poisson Factorization for Recommendation Systems," in *International Conference on Artificial Intelligence and Statistics*, pp. 275–283. [418,422,424]

Griffin, J. E., and Leisen, F. (2017), "Compound Random Measures and Their Use in Bayesian Nonparametrics," *Journal of the Royal Statistical Society*, Series B, 79, 525–545. [419]

Griffiths, T., and Ghahramani, Z. (2006), "Infinite Latent Feature Models and the Indian Buffet Process," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 475–482. [418,422,423]

Hensman, J., Fusi, N., and Lawrence, N. D. (2013), "Gaussian Processes for Big Data," in *Conference on Uncertainty in Artificial Intelligence*, p. 282. [425]

Hjort, N. (1990), "Nonparametric Bayes Estimators Based on Beta Processes in Models for Life History Data," *The Annals of Statistics*, 18, 1259–1294. [418,419]

Hoffman, M., Blei, D., Wang, C., and Paisley, J. (2013), "Stochastic Variational Inference," *Journal of Machine Learning Research*, 14, 1303–1347. [425]

Ishwaran, H., and Rao, J. S. (2005), "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies," *The Annals of Statistics*, 33, 730–773. [423]

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999), "Introduction to Variational Methods for Graphical Models," *Machine Learning*, 37, 183–233. [424]

Kingman, J. (1967), "Completely Random Measures," *Pacific Journal of Mathematics*, 21, 59–78. [418,419]

Kurihara, K., Welling, M., and Teh, Y. (2007), "Collapsed Variational Dirichlet Process Mixture Models," in *International Joint Conferences on Artificial Intelligence (IJCAI)*, pp. 2796–2801. [424]

Liang, P., Petrov, S., Klein, D., and Jordan, M. (2007), "The Infinite PCFG Using Hierarchical Dirichlet Processes," in *Empirical Methods in Natural Language Processing*, pp. 688–697. [425]

Liu, Y., Li, Z., Zhang, M., Deng, Y., Yi, Z., and Shi, T. (2013), "Exploring the Pathogenetic Association Between Schizophrenia and Type 2 Diabetes Mellitus Diseases Based on Pathway Analysis," *BMC Medical Genomics*, 6, S17. [427]

MacEachern, S. (1999), "Dependent Nonparametric Processes," in *ASA Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55. [418,424]

Miller, K., Griffiths, T., and Jordan, M. (2009), "Nonparametric Latent Feature Models for Link Prediction," in *Advances in Neural Information Processing Systems* (Vol. 22), eds. Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, pp. 1276–1284. [423]

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998), "Log Gaussian Cox Processes," *Scandinavian Journal of Statistics*, 25, 451–482. [418]

Ormerod, J. T., and Wand, M. (2012), "Gaussian Variational Approximate Inference for Generalized Linear Mixed Models," *Journal of Computational and Graphical Statistics*, 21, 2–17. [424]

Paisley, B., and Carin, L. (2009), "Nonparametric Factor Analysis With Beta Process Priors," in *International Conference on Machine Learning*, pp. 777–784. [424]

Paisley, J., Wang, C., and Blei, D. (2012), "The Discrete Infinite Logistic Normal Distribution," *Bayesian Analysis*, 7, 235–272. [418,421,423,424]

Paisley, J., Wang, C., Blei, D. M., and Jordan, M. (2015), "Nested Hierarchical Dirichlet Processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 256–270. [417]

Pitman, J., and Yor, M. (1997), "The Two-Parameter Poisson-Dirichlet Distribution Derived From a Stable Subordinator," *The Annals of Probability*, 25, 855–900. [428]

Ranganath, R., Gerrish, S., and Blei, D. (2014), "Black Box Variational Inference," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pp. 814–822. [425]

Rasmussen, C. E., and Williams, C. K. I. (2005), *Gaussian Processes for Machine Learning*, Cambridge, MA: The MIT Press. [418,420]

Salimans, T., and Knowles, D. A. (2013), "Fixed-Form Variational Posterior Approximation Through Stochastic Linear Regression," *Bayesian Analysis*, 8, 837–882. [425]

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650. [424]

Shah, A., Wilson, A. G., and Ghahramani, Z. (2014), "Student-t Processes as Alternatives to Gaussian Processes," in *Artificial Intelligence and Statistics*, pp. 877–885. [428]

Sudderth, E. B., and Jordan, M. I. (2009), "Shared Segmentation of Natural Scenes Using Dependent Pitman-Yor Processes," in *Advances in Neural Information Processing Systems*, pp. 1585–1592. [418]

Suvisaari, J., Perälä, J., Saarni, S. I., Härkänen, T., Pirkola, S., Joukamaa, M., Koskinen, S., Lönnqvist, J., and Reunanen, A. (2008), "Type 2 Diabetes Among Persons With Schizophrenia and Other Psychotic Disorders in a General Population Survey," *European Archives of Psychiatry and Clinical Neuroscience*, 258, 129–136. [427]

Teh, Y. (2006), "A Hierarchical Bayesian Language Model Based on Pitman-Yor Processes," in *Proceedings of the Association of Computational Linguistics*, pp. 985–992. [417]

Teh, Y. W., Jordan, M., Beal, M. J., and Blei, D. M. (2006), "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101, 1566–1581. [417,418,419,423,428]

Teh, Y. W., and Jordan, M. I. (2010), "Hierarchical Bayesian Nonparametric Models With Applications," in *Bayesian Nonparametrics: Principles and Practice*, eds. N. Hjort, C. Holmes, P. Müller, and S. Walker, Cambridge, UK: Cambridge University Press, pp. 1566–1581. [417,420,421]

Thibaux, R., and Jordan, M. (2007), "Hierarchical Beta Processes and the Indian Buffet Process," in *11th Conference on Artificial Intelligence and Statistics*, pp. 564–571. [417,418,419,423]

Titsias, M. K. (2008), "The Infinite Gamma-Poisson Feature Model," in *Advances in Neural Information Processing Systems*, pp. 1513–1520. [417]

——— (2009), "Variational Learning of Inducing Variables in Sparse Gaussian Processes," in *International Conference on Artificial Intelligence and Statistics*, pp. 567–574. [425]

Wainwright, M., and Jordan, M. (2008), "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends in Machine Learning*, 1, 1–305. [424]

Wallach, H., Murray, I., Salakhutdinov, R., and Mimno, D. (2009), "Evaluation Methods for Topic Models," in *International Conference on Machine Learning (ICML)*, pp. 1105–1112. [425]

Wang, C., and Blei, D. M. (2012), "Truncation-Free Stochastic Variational Inference for Bayesian Nonparametric Models," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 413–421. [424]

——— (2013), "Variational Inference for Nonconjugate Models," *Journal of Machine Learning Research*, 14, 1005–1031. [425]

Williamson, S., Wang, C., Heller, K. A., and Blei, D. M. (2010), "The ibp Compound Dirichlet Process and Its Application to Focused Topic Modeling," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 1151–1158. [417]

Zhou, M., and Carin, L. (2015), "Negative Binomial Process Count and Mixture Modeling," *Pattern Analysis and Machine Intelligence*, 307–320. [424]

Zhou, M., Chen, H., Paisley, J., Ren, L., Sapiro, G., and Carin, L. (2009), "Non-Parametric Bayesian Dictionary Learning for Sparse Image Representations," in *Advances in Neural Information Processing Systems* (Vol. 22), eds. Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, pp. 2295–2303. [417]

Zhou, M., Hannah, L., Dunson, D., and Carin, L. (2012), "Beta Negative Binomial Process and Poisson Factor Analysis," in *International Conference on Artificial Intelligence and Statistics*, pp. 1462–1471. [417]

Zhou, M., Yang, H., Sapiro, G., Dunson, D. B., and Carin, L. (2011), "Dependent Hierarchical Beta Process for Image Interpolation and Denoising," in *International Conference on Artificial Intelligence and Statistics*, pp. 883–891. [423]