

Functional Dirichlet Process

Lijing Qin, Xiaoyan Zhu

State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Dept. of Computer Science and Technology, Tsinghua University, Beijing, China
qinlj09@mails.tsinghua.edu.cn, zxy-dcs@tsinghua.edu.cn

ABSTRACT

Dirichlet process mixture (DPM) model is one of the most important Bayesian nonparametric models owing to its efficiency of inference and flexibility for various applications. A fundamental assumption made by DPM model is that all data items are generated from a single, shared DP. This assumption, however, is restrictive in many practical settings where samples are generated from a collection of dependent DPs, each associated with a point in some covariate space. For example, documents in the proceedings of a conference are organized by year, or photos may be tagged and recorded with GPS locations. We present a general method for constructing dependent Dirichlet processes (DP) on arbitrary covariate space. The approach is based on restricting and projecting a DP defined on a space of continuous functions with different domains, which results in a collection of dependent random measures, each associated with a point in covariate space and is marginally DP distributed. The constructed collection of dependent DPs can be used as a nonparametric prior of infinite dynamic mixture models, which allow each mixture component to appear/disappear and vary in a subspace of covariate space. Furthermore, we discuss choices of base distributions of functions in a variety of settings as a flexible method to control dependencies. In addition, we develop an efficient Gibbs sampler for model inference where all underlying random measures are integrated out. Finally, experiment results on temporal modeling and spatial modeling datasets demonstrate the effectiveness of the method in modeling dynamic mixture models on different types of covariates.

Categories and Subject Descriptors

G.3 [Probability and Statistics]: Nonparametric Statistics; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*clustering*; I.2.6 [Artificial Intelligence]: Learning

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

Copyright is held by the owner/author(s). Publication rights licensed to ACM.

ACM 978-1-4503-2263-8/13/10 ...\$15.00.

<http://dx.doi.org/10.1145/2505515.2505537>.

Keywords

Hierarchical Topic Modeling, Bayesian Nonparametric models

1. INTRODUCTION

Bayesian nonparametric models are a family of techniques that help bypass the difficult model selection problems, i.e., they allow the models to grow in size to accommodate the complexity of the data. In machine learning and a wide range of data mining applications, such as text modeling[15, 18, 17], computer vision[25], recommender system [9, 31], Bayesian nonparametric models are playing an increasingly important role.

Dirichlet process mixture (DPM) model is one of the most important Bayesian nonparametric models owing to its efficiency of inference and flexibility for various applications [27, 12, 23, 30]. DPM works by assigning Dirichlet process (DP) as prior distribution on an infinite number of mixture components, which to some extent is analog to the way Dirichlet distribution works in parametric topic model: Latent Dirichlet Allocation(LDA)[6]. Another similar point between LDA and DPM is exchangeability assumption. For DPM model, exchangeability indicates all data items are generated from a single, shared DP, and the infinite components of DPM are exchangeable, i.e., under any permutation of their ordering, the corresponding probability remains the same. This implicit simplicity is consistent with mining latent semantic of each sample, however, is restrictive in many practical settings where samples are not independent, but correlated in some covariate space.

Consider a generative model for news documents, where a piece of news is assumed to be generated from one or a few topics, and each topic is a probability distribution of words. Over time, for a particular news topic, it may emerge at a specific time point and disappear later. Meanwhile its word distribution may continuously vary over time. Over different locations, a news topic may appear only in a certain local region, while the associated word distribution is slightly different within this region. Lots of efforts have been made to model such dynamics of topics over covariate space[26, 29, 32, 19]. Most, if not all, proposals focus on the evolution of topics of sequential organized corpus of documents. An intuitive method is to construct a Markov chain of DPs, which, however, can only deal with discrete time covariates due to constraints of Markov models.

We argue that temporal correlation and spatial correlation are not unrelated, instead they are consistent, i.e., without loss of generality, time can be formulated as one-dimensional

Euclidean space, while location can be formulated as two-dimensional Euclidean space. It is natural to consider constructing a collection of dependent Dirichlet processes on a given covariate space, such that each topic has a lifespan, which is a subset of the covariate space, while the topic parameters vary smoothly within its lifespan. Note that the “lifespan” includes but not limits to timeline.

Dirichlet process (DP) is a stochastic process, meaning that it is a distribution over random distributions [11]. Specifically, it is a distribution over probability measures, which are functions that can be interpreted as distributions over some probability space. We denote a Dirichlet process as $\text{DP}(\alpha G_0)$, where parameter G_0 is called base measure and α is the concentration parameter. The base measure G_0 , which is basically regarded as the mean of the DP, can be an arbitrary distribution. A draw from DP will return a random distribution over values drawn from G_0 . In this paper, motivated by the correspondence between dynamics of topics and functions: *the domain of functions corresponds to the life-span of topics; the change of topic distribution is characterized by function values*, we present a general method to construct dependent DPs on arbitrary covariate spaces utilizing functional Dirichlet process (FDP) which is a DP over function space. That is to say, base measure G_0 is defined on a functional space, and a draw from this DP is a countably infinite collection of continuous functions. These functions thus naturally describe the correlation between covariate space and atom space (the output space). Given a realization of FDP, at any covariate index, restriction and projection operations are applied on FDP to find an induced probability measure supported by valid functions and their corresponding values.

The main contributions of this paper can be listed as follows:

- We propose a general method for constructing dependent Dirichlet processes on arbitrary covariate space based on restricting and projecting functional Dirichlet processes, where the atoms can appear/disappear and vary in subspaces of covariate space. We discuss the properties of constructed model, and prove that the output random measures are guaranteed to be marginally DP distributed.
- We develop a principled method to specify the base distribution of functions, such that the model can be applied in more general setting.
- We present an efficient Gibbs sampler inference algorithm by marginalizing out the underlying random measures.
- We conduct experiments on dynamic Gaussian mixture model, image segmentation, and dynamic topic model. The results demonstrate the effectiveness of our approach.

2. BACKGROUNDS

In this section, we give some mathematical backgrounds on Dirichlet process mixture model.

Dirichlet process is a random process, whose domain is also a random process over some probability space Θ [11]. Typically, a DP can be denoted as $\text{DP}(\alpha G_0)$, where G_0 is a distribution over Θ , α is a positive real number. Given a random distribution G over probability space Θ , for any finite

measurable partition of Θ : A_1, A_2, \dots, A_n , if random vector $(G(A_1), \dots, G(A_n))$ is Dirichlet distributed. Formally,

$$(G(A_1), \dots, G(A_r)) \sim \text{Dir}(\alpha G_0(A_1), \dots, \alpha G_0(A_r)).^1$$

We call the random distribution G is Dirichlet process distributed, and it is usually denoted as $G \sim \text{DP}(\alpha G_0)$.

By assigning DP as a prior of a hierarchical model, we get the Dirichlet process mixture model. DPM can be used for nonparametric clustering [22], i.e., the model need not know beforehand the number of clusters, but automatically grow in size to adapt the dataset. DPM can be formalized as a generative process:

$$\begin{aligned} G|\alpha, G_0 &\sim \text{DP}(\alpha G_0), \\ \theta_i|G &\sim G, \\ x_i|\theta_i &\sim F(\theta_i), \end{aligned} \quad (1)$$

where $\{x_1, \dots, x_n\}$ denote a set of observations, and $F(\cdot)$ is a given likelihood function parameterized by $\{\theta_1, \dots, \theta_n\}$. Each parameter θ_i is drawn independently and identically from the Dirichlet process G , and θ_i is hence exchangeable. This implicit assumption, however, is inappropriate in many collections of interest as we have discussed in Section 1. In this paper, we are to design a dependent Dirichlet process to display G in Eq. 1.

3. FUNCTIONAL DIRICHLET PROCESS

In this section, we present our construction of dependent Dirichlet processes using functional Dirichlet process (FDP), a DP defined on the space of continuous and bounded functions.

Model construction. Let (Ω, Σ) be probability space of component parameters, termed *atom space* and Θ as the *covariate space*. Our proposal is to construct a probability measure D^* over a countably infinite collection of continuous functions, regarded as functional atoms, where the domain of each function is a subspace of covariate space Θ . Given D^* and any covariate index $\phi \in \Theta$, *restrict and renormalize* D^* to include functional atoms whose domain contains ϕ , and *project* each of these functional atoms f (in function space) to atom (in atom space) located at $f(\phi)$. This gives us an induced probability measure D_ϕ over Ω .

Putting these steps together and posing a Dirichlet process prior on D^* , we have:

$$\begin{aligned} D^* &= \sum_i p_i \delta_{f_i} \sim \text{DP}(\alpha G^*), \\ D_\phi^* &= \frac{1}{Z_\phi} \sum_{i: \phi \in \text{domain}(f_i)} p_i \delta_{f_i}, \\ D_\phi &= \frac{1}{Z_\phi} \sum_{i: \phi \in \text{domain}(f_i)} p_i \delta_{f_i(\phi)}, \end{aligned} \quad (2)$$

where G^* is a probability distribution over \mathcal{F} , which is the space of continuous functions whose domains are subsets of covariate space, $\{f_i\}$ is a collection of countably infinite functions independent sampled from G^* , and the normalization constant $Z_\phi = \sum_{i: \phi \in \text{domain}(f_i)} p_i$ ensures that resulting measure D_ϕ is a properly normalized probability measure, for any element ϕ in the covariate space. We defer the discussion of construction G^* to Section 4. Figure 1 depicts

¹Basic concepts about a modicum of measure theory and Dirichlet distributions are necessary.

intuitions of the construction on several exemplary covariate spaces.

Marginal distribution. D_ϕ^* and D_ϕ are automatically marginal DP distributed by the general property that DP is closed under restriction and projection operations which is established through the following lemma.

Lemma 3.1. *Let $D = \sum p_i \delta_{\theta_i} \sim \text{DP}(\mu)$ be a Dirichlet process on (Ω, Σ) . Then:*

1. *The random probability measure obtained by restricting and renormalize D on X is DP distributed, where $X \subseteq \Omega$ is measurable subset on Ω ,*

$$\begin{aligned} D|_X &\triangleq \frac{1}{D(X)} \sum_{\theta_i \in X} p_i \delta_{\theta_i} \sim \text{DP}(\mu|_X), \\ D(X) &= \sum_{\theta_i \in X} p_i. \end{aligned} \quad (3)$$

2. *The random probability measure obtained by projecting D through measurable mapping function $g : \Omega \rightarrow \Omega'$ is also DP distributed,*

$$\begin{aligned} g(D) &\triangleq \sum_i p_i \delta_{g(\theta_i)} \sim \text{DP}(g\mu), \\ \forall B \in \Sigma, \quad g\mu(B) &\triangleq \mu(g^{-1}(B)). \end{aligned} \quad (4)$$

We refer interested readers to [16] for a detailed treatment on the advanced properties of DP. Applying the lemma, we found the marginal distribution of D_ϕ is,

$$D_\phi \sim \text{DP}(\mu_\phi), \quad (5)$$

where μ_ϕ is a measure defined on (Ω, Σ) such that, $\forall A \in \Sigma$,

$$\mu_\phi(A) = \mu^*(\{g : \phi \in \text{domain}(g), g(\phi) \in A\}). \quad (6)$$

Infinite dynamic mixture model. The resulting collection of dependent DPs $\{D_\phi\}_{\phi \in \Theta}$ is able to act as a nonparametric prior for a set of infinite mixture models. Formally, given a set of covariate indices ϕ_1, \dots, ϕ_n (may have repetition), the observed data $(x_1, \phi_1), \dots, (x_n, \phi_n)$ are generated by the following generative processes.

$$\forall i \in [n], \quad \theta_i | D_{\phi_i} \sim D_{\phi_i}, \quad x_i | \theta_i \sim F_{\theta_i}(\cdot), \quad (7)$$

where $F_\theta(\cdot)$ is a probability distribution parametrized by θ . Observation x_i at covariate θ_i is drawn from a mixture component with parameter θ_i whose value is determined by the underlying function f_i which is marginalized out in this representation.

4. BASE DISTRIBUTION OF FUNCTIONS

We have described a general solution to constructing a collection of dependent Dirichlet processes on arbitrary covariate space using functional Dirichlet process. In order to use the construction in machine learning applications, we need to consider a general question on choosing the base distribution of function G^* . This distribution essentially determines the dependency between constructed DPs and is application-dependent. In this section, we develop a principled method to construct such a distribution G^* . Our construction of G^* is applicable on arbitrary covariate space and does not preclude efficient inference. Furthermore, we give examples

on several types of covariate space with different types of application data.

For simplicity, we assume that a random function $g : \text{domain}(g) \rightarrow \Omega$ is generated by restricting a stochastic process on the *entire* covariate space to an independent random subset of covariate space. In particular, 1) The function domain is determined by drawing $\text{domain}(g)$ from a distribution over subsets. 2) The function values are determined by drawing $g_0 : \Theta \rightarrow \Omega$ as a realization from a Ω -valued stochastic process on entire covariate space: $\{g_0(\phi)\}_{\phi \in \Theta}$, called *base process*. The random function can be then obtained by restricting $g = g_0|_{\text{domain}(g)}$. We note, in this procedure, the assumption we made is that function domain is independent with function values.

4.1 Prior distribution of function domains

In our model, function domains characterize the life-spans of mixture components on covariate space. Determining the properties of life-span depends on domain knowledge of applications. For example, when covariate is time, a proper choice of life-span is time-interval (segment on R), in which the two endpoints represent “born” and “die” of certain component. On the other hand, in image segmentation applications, the life-span of a component would be a contiguous region on R^2 . To model the life-span of mixture components, we need an effective way to control function domains via prior distribution.

By Kolmogorov extension theorem, the distribution of function domains can be constructed indirectly via its finite-dimensional marginal distributions $\varphi \triangleq \Pr(\{\phi_{1:n}\} \subseteq \text{domain}(f))$. Through different characterizations of φ , we can construct arbitrary distributions of function domains. In particular, the *diameter dependent* φ can be used for covariate space with metric defined $d : \Omega \times \Omega \rightarrow R$,

$$\varphi(\{\phi_{1:n}\}) \triangleq \exp\left(-\gamma \max_{i,j \in [n]} \{d(\phi_i, \phi_j)\}\right), \quad \gamma > 0. \quad (8)$$

We note the above characterization recovers the discrete case discussed by Lin et al [20]: $\Theta = \{1, 2, \dots\}$ in which the covariate space is regarded as time slots. Their construction corresponds to a special case of FDP where each domain has the form $\{a_i, a_i + 1, \dots, b_i\}$, in which $|b_i - a_i| \sim \text{Geometric}(q)$. It is not hard to see this is consistent with (8) with $d(a, b) = |a - b|$ and $\gamma = \ln(1 - q)$. This can be easily extended to continuous covariate space R with little modification. In this scenario, each domain has form $[a_i, b_i]$ and $|b_i - a_i| \sim \text{Exponential}(q)$, $\gamma = q \ln q$. With general Euclidean covariate space R^d , (8) also ensures the diameter of domain have statistical pressure to be kept small.

4.2 Base processes

When using FDP as a nonparametric prior for infinite dynamic mixture models, the role of base processes closely resembles that of base distributions played in DPM model: both of them served as the prior of component parameters, while the base processes have an extra flexibility to allow the parameters changing over covariate space. Similar with the choice of base distribution in DPM, the choice of base processes is dependent on the application and the type of data. In particular, we are interested with the base processes that are conjugate with data likelihood. In the rest of this part, we investigate three different choices of conju-

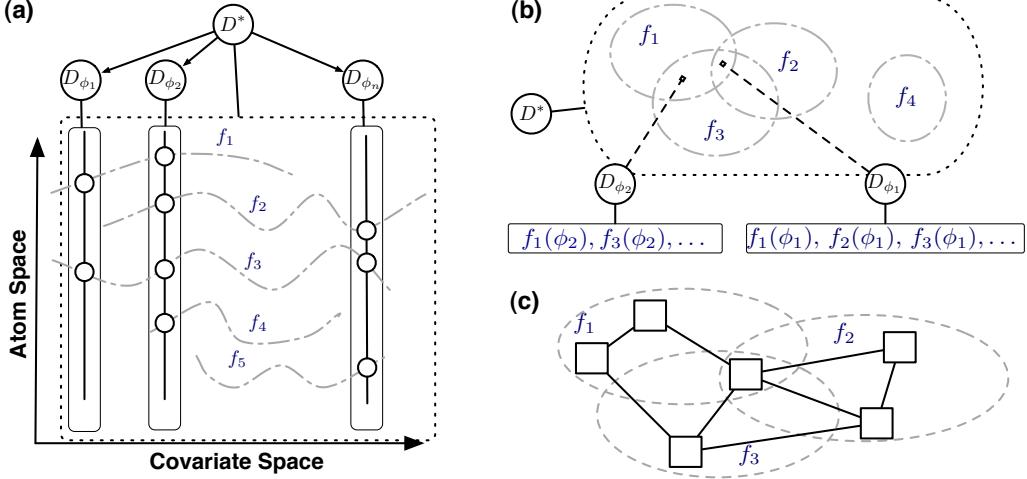


Figure 1: Illustrations of constructing dependent DPs using functional Dirichlet process on different covariate spaces. a) Both covariate space and atom space are R . D^* is a DP defined on function space. Each D_{ϕ_i} is constructed by finding all functions f in D^* whose domain contains ϕ_i and their corresponding values $f(\phi_i)$. b) Covariate space is R^2 and each function domain is a subset on R^2 . c) Covariate space is the set of all vertices of an undirected network and each function domain is a set of vertices.

gate base process that cover the need of a range of common applications.

Constant functions. Constant functions are among the simplest choices of base process which precludes the flexibility of allowing variation of parameters. In particular, $\forall \phi, g_0(\phi) = c, c \sim H_0$, where H_0 is a probability distribution over Ω . In practice, H_0 can be chosen as the conjugate prior respected to data likelihood which can enjoy a collapsed Gibbs sampler.

Gaussian process. When the data distribution $F_\theta(x)$ is such that the parameter θ can be represented as a real-valued vector $\theta \in R^M$, it is natural to use a Gaussian distribution to describe the marginal distribution of θ and a Gaussian process to encode the dependency between different covariates. That is, given a mean function $m(\phi)$ and a covariance function $K(\phi, \phi')$, the base process is a given by $g_0 \sim GP(m, K)$.

Dependent Dirichlet distribution. When each datum is an element from W discrete ones, similar with many finite topic models, a categorical model may be suitable. In this case, $g_0(\phi)$ takes value in the surface of the W -simplex. Motivated by the fact that a Dirichlet distribution is a special case of DP with a finite space, we construct a collection of dependent Dirichlet distributions $\{g(\phi)\}_{\phi \in \Theta}$ using a special case of FDP where $\Omega = [W]$. In particular, let $\bar{D}^* \sim DP(\bar{\mu}^*)$ where $\bar{\mu}^*$ is a finite measure support over the space of constant functions, but with different domains. By construction, we obtain induced dependent probability measures $\{\bar{D}_\phi\}_{\phi \in \Theta}$. $(\bar{D}_\phi(\delta_1), \dots, \bar{D}_\phi(\delta_W)) \sim \text{Dir}(\mu_\phi(\delta_1), \dots, \mu_\phi(\delta_W))$ are Dirichlet distributed. This completes the construction of base process as dependent Dirichlet distributions $g_0(\phi) = \bar{D}_\phi$. As we would show in Section 6, this construction does not preclude efficient inference as the predictive distribution $\theta_{\phi_{n+1}} | \theta_{\phi_1}, \dots, \theta_{\phi_n}$ is tractable to

evaluate. We contrast this construction with previous work on dynamic topic models that are based on transforming from underlying Gaussian process [5, 2].

5. COMPARISON WITH PREVIOUS WORK

A few of DDP constructions are based on varying stick breaking weights over covariates of stick breaking representation of DP. Dunson and Park [10] proposed kernel stick breaking process (KSBP), where the covariate dependency is involved via setting stick breaking weights as weighted sums of independent beta variables and kernel function on covariate space. Ren et al [25] proposed logistic stick breaking process (LSBP) in which the stick breaking weights are determined by a hierarchy of spatial logistic regressions with sparseness promoting priors placed on regression coefficients. Both works are not marginally DP distributed, and their inference algorithms are based on truncated approximation. Alternatively, the π DDP, proposed by Griffin and Steel [14], allows the change of distribution over atoms but still preserves marginally DP distributed. In their work, the dependency is implemented by inducing a different ordering of stick breaking ratios at each time slot. Different from our model, these works assumed atom locations are fixed across different covariates. Moreover, these constructions missed interpretations of life-spans of atom, which are realized through function domains in our model.

Another line of works is based on Chinese restaurant process (CRP) representation of DP. The recurrent Chinese restaurant processes (RCRP) [1] and generalized Polya Urn scheme [8] both generalized the CRP to include time dependency and still maintain marginal DP property. However, these constructions are only defined for the real line as covariate space. Blei and Fraizer developed distance dependent Chinese restaurant processes (DCRP) [4] generalized

RCRP to arbitrary covariate spaces. Fundamentally different from Bayesian nonparametric models, DCRP is not constructed based on random measures and does not exhibit marginal invariance.

Lin et al [20] is one of the closest works to our approach. They proposed a Markov chain of DPs (DDP with discrete covariate space) constructed via superposition, subsampling and transition operations of DP. FDP recovers their model, by choosing the shape distribution (geometric distributed life-span) and covariate space (discrete) defined in Section 4.1. Moreover, our developed sampling algorithm is able to exploit global information of data in every observed covariate, in contrast with sequential sampler derived in [20]. Another closely related work is the spatially normalized Gamma process (SNTP) developed by Rao and Teh [24] which enables to construct dependent DPs on arbitrary covariate space. Their work does not have the model flexibility for allowing atom transitions. In addition, their construction methods are different from our model (they use normalized restricted Gamma process, we use restricted Dirichlet process). Despite these differences, their resulting random measures are equivalent to a special case of FDP which only allows constant functions in G^* and the dependency is induced by overlapping function domains. Finally, we note they did not develop distribution of function domains for covariate space other than R .

6. MODEL INFERENCE

We develop an efficient sampling algorithm for FDP where all underlying random measures are integrated out. Assuming we have observed n data, $x_{1:n}$ with covariates $\phi_{1:n}$ respectively, we wish to infer the partition of data with an unknown number of partitions and parameters, where each partition corresponds to a function in FDP formulation. In particular, we are interested with the *functional assignments* $z_{1:n}$ of data, which resembles the *table assignments* in traditional Chinese restaurant process representations.

To formulate the problem, consider the following alternative generative process of $x_{1:n}$,

$$\begin{aligned} \forall i \in [n], \quad f_i | D_{\phi_i}^* &\sim D_{\phi_i}^*, \\ \theta_i | f_i &= f_i(\phi_i), \quad x_i \sim F_{\theta_i}(\cdot). \end{aligned} \quad (9)$$

Denote $f_{1:K}^*$ to be the unique functions among $f_{1:n}$ in this representation (9). Each unique function f_l^* appears c_l times. Then the assignments of functions are denoted as $z_i \in [K]$, s.t. $f_{z_i}^* = f_i$.

6.1 Predictive distribution

Predictive distribution plays a central role in implementing a Gibbs sampler for model inference. The aim of this part is to derive the predictive distribution $z, \theta | \{z_{1:(n-1)}, \theta_{1:(n-1)}\}$ for datum x with covariate index ϕ . The derivation relies on the fact that the posterior of D^* given $\{z_{1:(n-1)}, \theta_{1:(n-1)}\}$ is a *mixture of Dirichlet process* (MDP) [3]. The predictive distribution follows by simplifying and marginalizing out DPs in the posterior. We leave the derivation of the Gibbs sampler in the appendix. The predictive distributions

$z, \theta | \{z_{1:(n-1)}, \theta_{1:(n-1)}\}$ is given by

$$z | \{z_{1:(n-1)}, \theta_{1:(n-1)}\} = \mathbb{E}_{t_{1:K}^{(\phi)}, \theta_{1:K}^{(\phi)}} \left[\frac{\mu_\phi(\Omega) \delta_{K+1} + \sum_{i \in [K]} t_i^{(\phi)} c_i \delta_i}{\mu_\phi(\Omega) + \sum_{i \in [K]} t_i^{(\phi)} c_i} \right],$$

for $i \in [K]$,

$$\begin{aligned} t_i^{(\phi)} &\stackrel{i.i.d.}{\sim} \text{Bernoulli}(q_i^{(\phi)}), \\ \theta_i^{(\phi)} &\stackrel{i.i.d.}{\sim} T_i^{(\phi)}(\cdot), \\ \theta | z, \{z_{1:(n-1)}, \theta_{1:(n-1)}\} &= T_z^{(\phi)}(\cdot) \end{aligned}$$

where δ_{K+1} indicates the event of assigning to a new function, $q_i^{(\phi)}$ and $T_i^{(\phi)}$ are induced from G^* and are given by,

$$\begin{aligned} q_i^{(\phi)} &= \Pr(t_i^{(\phi)} = 1 | \{z_{1:(n-1)}, \theta_{1:(n-1)}\} \subseteq \text{domain}(f)) \\ &= \frac{\varphi(\{\phi_{1:(n-1)}, \phi\})}{\varphi(\{\phi_{1:(n-1)}\})}, \end{aligned} \quad (10)$$

$$T_i^{(\phi)}(d\theta) = \Pr(f(\phi) = \theta | f(\phi_1) = \theta_1, \dots, f(\phi_{n-1}) = \theta_{n-1}).$$

Sampling z using the predictive distributions is made tractable by treating $t_{1:K}^{(\phi)}, \theta_{1:K}^{(\phi)}$ as auxiliary variables to be explicitly sampled rather than to be marginalized out. We also note that the auxiliary variables have clear interpretations: $t_i^{(\phi)} \sim \text{Bernoulli}(q_i)$ is a random event that the domain of a function f includes ϕ given that its domain include $\phi_{1:(n-1)}$. Meanwhile $\theta_i^{(\phi)} \sim T_i^{(\phi)}$ is a random variable equals to $f(\phi)$ conditioned on $f(\phi_i) = \theta_i$.

6.2 Gibbs sampler

We have so far proposed the FDP model and analyze its corresponding posterior and predictive distributions. Hereby, we employ a Gibbs sampler to perform an approximate inference.

The Gibbs sampler maintains all assignments $z_{1:n}$, $z_i \in [K]$, where K is the number of represented functions and will change over iterations. It also maintains $c_{1:K}$ as the numbers of occurrences of different functions: $c_j = |\{z_i = j\}|$. During each iteration, the sampler iterates over all assignment variables $z_{1:n}$ and resample each of them at a time. In addition, for some choices of G^* and F , the Gibbs sampler can operate in a collapsed mode where parameters are integrated out. We discuss both Gibbs sampler for general case and that for collapsed model. In what follows, we describe the procedure of resampling z_n , the assignment variable of datum x_n with covariate ϕ_n , without loss of generality.

1. Sample auxiliary variables. For each $k \in [K]$, resample $t_k^{(\phi_n)} \sim \text{Bernoulli}(q_k^{(\phi_n)})$ to determine whether the domain of f_k^* contains ϕ_n , where $q_k^{(\phi_n)}$ is defined in (16).

(general) 2. Sample assignments. First for each $k \in [K]$, draw parameter $\theta_k^{(\phi_n)} \sim T_k^{(\phi_n)}(\cdot)$. Then probability of assigning to existing functions is $P(z_n = k) \propto t_k^{(\phi_n)} c_k F_{\theta_k^{(\phi_n)}}(x_n)$, while the probability of assigning to a new function is $P(z_n = K+1) \propto \mu_{\phi_n}(\Omega) F(x_n)$, where $F(x_n) \triangleq \mathbb{E}_{\theta \sim H_{\phi_n}(\cdot)}(f_\theta(x_n))$ and $H_{\phi_n}(\cdot)$ is the marginal distribution of base process at ϕ_n .

(collapsed) 2. Sample assignment. In certain cases, one can further integrate out parameters $\theta_{1:K}^{(\phi_n)}$ which leads to a collapsed sampling method for assignment variables. Denote the conditional density of x_n be generated by function f_k^* given all its data items as,

$$F(x_n | \{x_j\}_{z_j=k}) = \frac{\int_{\theta_k^{(\phi_n)}} F_{\theta_k^{(\phi_n)}}(x_n) \prod_{z_j=k} F_{\theta_k^{(\phi_n)}}(x_j) p(\theta_k^{(\phi_n)}) d\theta_k^{(\phi_n)}}{\int_{\theta_k^{(\phi_n)}} \prod_{z_j=k} F_{\theta_k^{(\phi_n)}}(x_j) p(\theta_k^{(\phi_n)}) d\theta_k^{(\phi_n)}}.$$

The probability of joining a existing function is $P(z_n = k) \propto t_k^{(\phi_n)} c_k F(x_i | \{x_j\}_{z_j=k})$, while the probability of assigning to a new function remains unchanged. Each of the three cases discussed in Section 4.2 can use this collapsed sampler.

7. EXPERIMENTS

In this section, we evaluate our methods on both synthetic and real-world datasets.

7.1 Synthetic Dataset

Temporally evolving Gaussian mixtures. In the first experiment, we simulate a collection of evolving Gaussian mixtures. Initially, 2 Gaussian components are generated. New components are introduced following a Poisson process (on average 0.4 new components per time phase). The lifespan of each component is geometric distributed with mean 5. In addition, for each component, its mean evolves follows a Brownian motion with variance 1. We simulated for 30 phase with 200 data points being drawn independently from each component at each phase. We applied our model on generated data points with covariates as the time phases that data points be generated. The base process is chosen as a zero-mean Gaussian Process. The distribution of function domain is constructed as (8). We run the Gibbs sampler for 5000 iterations and collected the resulting assignment variables as partition results.

We evaluated the results by measuring the dissimilarity between inferred partition with ground truth using the variation of information criteria[21]. Formally, the variation of information is defined as follows:

$$d_{VI}(\mathcal{C}_1, \mathcal{C}_2) = H(\mathcal{C}_1) + H(\mathcal{C}_2) - 2I(\mathcal{C}_1, \mathcal{C}_2), \quad (11)$$

where H and I denote the entropies of and the mutual information between the two clusters, respectively. d_{VI} measures the distance between two clusterings in terms of the information difference between the two clusters. Hence, the lower the value of d_{VI} , the better the performance is.

We compared our model (F-DPM) with covariate-independent DP mixture model (DPM) and a state-of-the-art technique: Markov chain of DPs (Markov-DPM) described in [20]. We note that, in this scenario, Markov-DPM is equivalent to F-DPM, with discrete covariate and suitable base distribution of functions, in the sense that the constructed random measures are identically distributed. The difference is that Markov-DPM uses a sequential sampling algorithm for inference. Based on a different analysis, F-DPM can use a batch inference algorithm as described in Section 6.

We can observe from Figure 2 that F-DPM consistently outperforms Markov-DP since the batch inference algorithm for F-DPM can exploit global information while sequential filtering algorithm for Markov-DP is not able to use further observations. The performance gain of F-DPM over

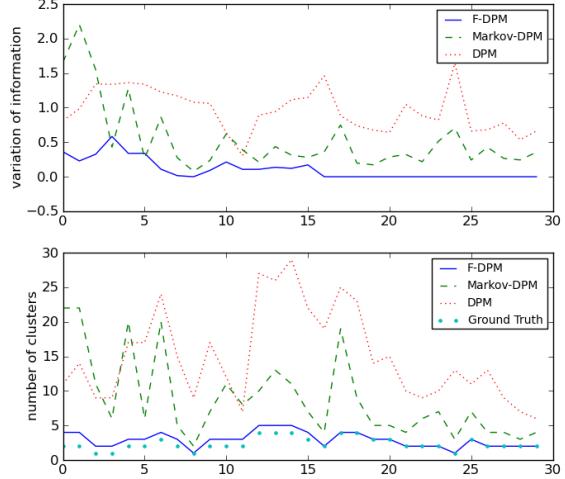


Figure 2: Experiment results: Compare F-DPM with Markov-DPM and DPM for evolving Gaussian mixture model. The top graph shows the median of distance between the resulting clusters and the ground truth at each phase. The bottom graph shows the actual numbers of clusters.

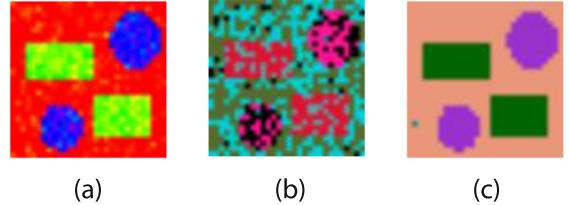


Figure 3: Experiment results: Image segmentation results by F-DPM and DPM.

Markov-DPM is especially significant for data with small timestamps, as a result of the sequential nature of Markov-DPM inference algorithm. In addition, both F-DPM and Markov-DPM outperforms DPM since they better captured the dynamic nature of data. Moreover, we can see that F-DPM almost exactly recovered the number of clusters. This is remarkable because the state of the art Bayesian nonparametric method DPM, which is often applied for its capability of estimating the number of clusters, produced much more inaccurate result on this quantity.

Gaussian mixture with other covariate spaces. We further tested our model with different covariate spaces: R^2 and undirected network. On R^2 covariate space, we demonstrated a simple image segmentation application. We synthesize the original gray-scaled image with four continuous sub-regions on a background with different intensities. The background has an intensity 5 and the two pairs of sub-region have intensities of 10 and 25. Each pixel is added

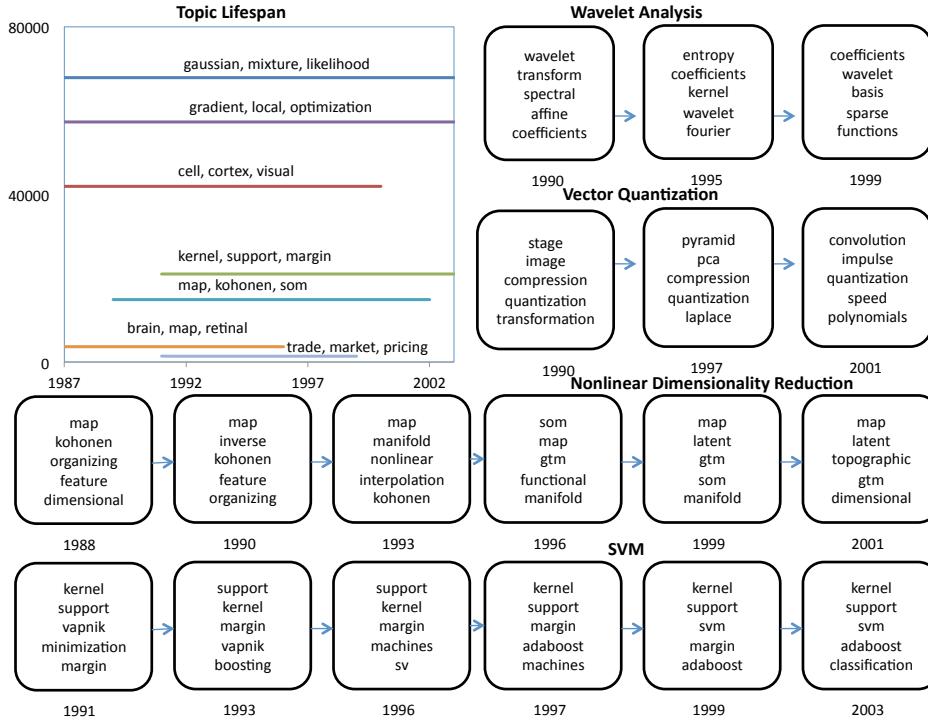


Figure 4: Experiment results: Recovered topic timeline by F-DPM model on NIPS dataset.

a independent white Gaussian noise with standard deviation of 1.5. We applied our model to infer partition of all pixels with covariate space R^2 . We take the base process as constant functions with H_0 and distributions of function domain as defined in (8). We compared the segmentation results with covariate-independent DP mixture model.

The results is shown in Figure 3. We found the segmentation results generated from our model are clearer than that of DPM, since FDP accommodates spatial dependency such that closer pixels are more likely to be clustered together.

7.2 Real-world Dataset

Evolving topic model. We analyze the proceedings of the NIPS conference from 1987 to 2001, which consist of 2484 documents. As in hierarchical Dirichlet process (HDP), we consider a topic as a distribution over words and each document as a distribution over topics. In order to model the evolution of topics, we place a FDP prior over topics. In this setting, each topic has a life-span over time and its word distribution evolves as well. In particular, we set the length of life-span to have an exponential distribution with parameter 0.6. In addition, the base process is set to be dependent Dirichlet distributions in order to model the continuous transition. The recovered timeline of topics is demonstrated in Figure 4. From topic lifespan, we can see the model indeed captures the death and birth of various topics. The top keywords showed in the same figure also demonstrate some topics (chains) evolve over time. In Table 2 (NIPS dataset), we have also compared the performance of DPM and F-DPM in terms of held-out log likelihood, which is a standard metric for measuring modeling performance (cf.

[7, 27]). Held-out log likelihood is defined as log-likelihood of the held-out data, $L = \log p(w_{\text{heldout}} | m_{\text{trained}})$. In particular, we randomly separate 90% words to train the topic model and use the remaining data for testing. We can see that the F-DPM obtained a higher held-out log likelihood than DPM, since the additional temporal information leads to a more preferable model.

Location based topic model. In the last experiment, we use Flickr dataset to evaluate our approach over two-dimensional location space. Specifically, we use F-DPM to mine latent topics from image tags with GPS information. We crawl images with GPS locations from Flickr website². Three representative datasets, including Landscape, Activity, and National Park, whose time spans range from 2009/01/01 to 2010/01/01 and GPS locates in USA territory are kept. We crawl images using keyword *landscape* around USA for Landscape dataset, *surfing* and *hiking* around USA for Activities dataset, and *nationalpark* around USA for National Park dataset. Tags occurring less than 15 times are removed. Totally, we obtain 1505 + 11868 + 2109 images and 2313 + 2381 + 2374 tags corresponding to three datasets.

We compare our approach with DPM, which assumes all tags are sampled from a set of shared topics, i.e., without considering location information. Table 1 shows the number of topics discovered by DPM and F-DPM from three datasets. The result shows that F-DPM recovers more topics than DPM, which is reasonable since F-DPM requires a slight constriction on the topic generation and separates more topics. Furthermore, three examples shown in Figure 5 implicit topics found by F-DPM are less ambiguous,

²<http://www.flickr.com/services/api>



(a) **Topic 1** Landscape(DPM): land- (b) **Topic 2** Nationalpark (DPM): (c) **Topic 3** Activity (DPM): hiking, scape, paisaje, mexico, tanawin, land- nationalpark, yosemite, nature, yel- water, waking, park, green schaft lowstone, osprey



(d) **Topic 1** Landscape (F-DPM): (e) **Topic 2** Nationalpark (F-DPM): (f) **Topic 3** Activity (F-DPM): surf- landscape, paisaje, mexico, tanawin, landscape, yellowstone, wyoming, ing, water, waking, nature, coast mekcnka sunset, sky

Figure 5: Experiment results: Photos in the topics discovered by DPM and F-DPM.

Data Set	# topics	
	DPM	F-DPM
Landscape	6	9
Activities	5	12
National Park	9	14

Table 1: The number of topics discovered by DPM and F-DPM.

Data Set	DPM	F-DPM
NIPS	-3.45×10^6	-3.15×10^6
Landscape	-7.74×10^4	-7.32×10^4
Activities	-5.14×10^5	-4.39×10^5
National Park	-9.21×10^4	-8.98×10^4

Table 2: Comparison of held-out log likelihood.

and thus more satisfying. In the first example for landscape dataset, Figure 5(a) and Figure 5(c) show that both DPM and F-DPM detect a clear topic about `mexico`. However, in the second and third examples, we can see that topics found by F-DPM are significantly more interpretable, namely, Figure 5 (e) shows a topic on `yellowstone`, and Figure 5 (f) shows a topic about `coast surfing`. Moreover, the quantitative evaluation results using held-out log likelihood is shown in Table. 2 for all of the three datasets. The quantitative results demonstrated that, by incorporating the

location information, the F-DPM yields a better modeling performance than DPM.

8. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we proposed a general method for constructing dependent Dirichlet processes on arbitrary covariate space based on restricting and projecting functional Dirichlet process, where the atoms can appear/disappear and vary in subspaces of covariate space. We discussed the properties of constructed model, and proved that the output random measures are guaranteed to be marginally DP distributed. We developed a principled method to specify the base distribution of functions, such that the model can be applied in more general setting. We presented an efficient Gibbs sampler inference algorithm by marginalizing out the underlying random measures. We conducted experiments on dynamic Gaussian mixture model, image segmentation, and dynamic topic model. The results demonstrated the effectiveness of our approach.

There are several interesting future directions. First, we are interested in applying the idea of constructing dependent random measures by restricting and projecting a shared underlying random measure on function space for constructing other types of dependent random measures, for example Indian Buffet process [13], in which the underlying random measure is beta process [28]. Second, we plan to explore a variational method as an alternative deterministic form of approximate inference. Finally, as a byproduct of our construction, the dependent Dirichlet distributions can be used as a component in a variety of existing dynamic topic model

applications for inducing dependency in arbitrary space and not precluding efficient sampling algorithm.

Acknowledgments

This work is supported by Canada's IDRC Research Chair in Information Technology program, Project No. 104519-006, China National Key Basic Research Program (also called 973 Program), Project No. 2013CB329403, and Tsinghua University Initiative Scientific Research Program, Project No. 2012108807. The authors would also like to express our gratitude to Shouyuan Chen for his assistance on experiments.

References

- [1] A. Ahmed and E. P. Xing. Dynamic non-parametric mixture models and the recurrent Chinese restaurant process. Citeseer, 2008.
- [2] A. Ahmed and E. P. Xing. *Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream*. 2010.
- [3] C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2(6):1152–1174, 1974.
- [4] D. M. Blei and P. Frazier. Distance dependent Chinese restaurant processes. In *Proc. of ICML'10*, pages 87–94, Haifa, Israel, June 2010.
- [5] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proc. of ICML'06*, pages 113–120, New York, NY, USA, 2006. ACM.
- [6] D. M. Blei and J. D. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35, 2007.
- [7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.
- [8] F. Caron, M. Davy, and A. Doucet. Generalized Polya urn for time-varying Dirichlet process mixtures. In *Proc. of UAI'07*, pages 33–40, Corvallis, Oregon, 2007.
- [9] N. Ding, Y. A. Qi, R. Xiang, I. Molloy, and N. Li. Nonparametric bayesian matrix factorization by power ep. In *Proc. AISTATS*, volume 9, pages 169–176, 2010.
- [10] D. B. Dunson and J. Park. Kernel stick-breaking processes. *Biometrika*, 95(2):307–323, 2008.
- [11] T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.
- [12] E. Fox, E. Sudderth, M. I. Jordan, and A. Willsky. An HDP-HMM for systems with state persistence. pages 312–319, 2008.
- [13] Z. Ghahramani and T. L. Griffiths. Infinite latent features models and the Indian buffet process. In *Proc. of NIPS'05*, 2005.
- [14] J. E. Griffin and M. F. J. Steel. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101(473):179–194, 2006.
- [15] T. Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16:106–114, 2004.
- [16] N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker. *Bayesian nonparametrics*, volume 28. Cambridge University Press, 2010.
- [17] D. Kim and A. Oh. Accounting for data dependencies within a hierarchical dirichlet process mixture model. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 873–878. ACM, 2011.
- [18] J. H. Kim, D. Kim, S. Kim, and A. Oh. Modeling topic hierarchies with the recursive chinese restaurant process. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 783–792. ACM, 2012.
- [19] P. Langley. Crafting papers on machine learning. In P. Langley, editor, *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pages 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.
- [20] D. Lin, E. Grimson, and J. Fisher. Construction of dependent Dirichlet processes based on Poisson processes. In *Proc. of NIPS'10*, pages 1396–1404, 2010.
- [21] M. Meila. Comparing clusterings - An axiomatic view. In *Proc. of ICML'05*, 2005.
- [22] R. M. Neal. Bayesian mixture modeling. *Maximum Entropy and Bayesian Methods: Proceedings of the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, pages 197–211, 1992.
- [23] Y. Qi, D. Liu, D. B. Dunson, and L. Carin. Multi-task compressive sensing with Dirichlet process priors. In *ICML'2008*, pages 768–775, 2008.
- [24] V. Rao and Y. W. Teh. Spatial normalized Gamma processes. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Proc. of NIPS'09*, pages 1554–1562. 2009.
- [25] L. Ren, L. Du, L. Carin, and D. B. Dunson. Logistic stick-breaking process. *Journal of Machine Learning Research*, 12:203–239, February 2011.
- [26] N. Srebro and S. Roweis. Time-varying topic models using dependent dirichlet processes. *UTML, TR# 2005*, 3, 2005.
- [27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
- [28] R. Thibaux and M. I. Jordan. Hierarchical Beta processes and the Indian buffet process. In *Proc. of AISTATS'07*, 2007.

- [29] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM, 2006.
- [30] S. Williamson, C. Wong, and K. A. Heller. The IBP compound Dirichlet process and its application to focused topic modeling. *Computer*, 2010.
- [31] K. Yu, S. Zhu, J. Lafferty, and Y. Gong. Fast nonparametric matrix factorization for large-scale collaborative filtering. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 211–218. ACM, 2009.
- [32] X. Zhu, Z. Ghahramani, and J. Lafferty. Time-sensitive dirichlet process mixture models. Technical report, DTIC Document, 2005.

APPENDIX

A. DERIVATION OF PREDICTIVE DISTRIBUTION

In this part, we show the derivation of predictive distribution $z, \theta | \{z_{1:(n-1)}, \theta_{1:(n-1)}\}$, which is essential for developing the Gibbs sampler. Our approach is to first analyze the posterior distribution of FDP. After that we are able to integrate out all underlying random measures and compute predictive distribution, since the constructed random measure D_ϕ are marginally DP distributed. Then, we will derive the predictive distribution based on the collapsed posterior distribution of FDP.

Posterior of $D^* | \{z_{1:(n-1)}, \theta_{1:(n-1)}\}$. The existence the functional assignments $z_{1:n}$ and function values is possible if and only if that each unique function f_i^* satisfies,

$$f_i^* \in \mathcal{F}_l = \bigcap_{j \in [n-1]}^{z_j=1} \{f : \phi_j \in \text{domain}(f), \theta_j = f(\phi_j)\} \quad (12)$$

In other words, conditioning on sample $\{z_{1:(n-1)}, \theta_{1:(n-1)}\}$ is equivalent with given K different observations such that 1) the l th observation occurs c_l times 2) the l th observation is in the set \mathcal{F}_l . In short, $D^* | \{z_{1:(n-1)}, \theta_{1:(n-1)}\} \stackrel{d}{=} D^* | \{c_{1:K}, \mathcal{F}_{1:K}\}$. It is noted that this is different with the posterior analysis of traditional DP where locations of each observation are exactly known. In this model, however, the exact locations of each functional atom are subjected to *censoring* and only possible regions of each observation are known. Such family of posterior, originally analyzed by Antoniak [3], is termed as *mixture of Dirichlet processes* (MDP). The posterior can be obtained by following steps. Given functions $f_{1:K}^*$, the posterior of D^* is simply $\text{DP}(\mu^* + \sum_i c_i \delta_{f_i^*})$. Integrating these with respect to the conditional probability of $\Pr(f_i^* | f_i^* \in \mathcal{F}_i)$, we obtain the posterior of D^* given $\{f_i^* \in \mathcal{F}_i\}$ is

$$\begin{aligned} M^* &\triangleq D^* | \{z_{1:(n-1)}, \theta_{1:(n-1)}\} \\ &= \int_{f_{1:K}^*} \left[\prod_{i \in [K]} \Pr(f_i^* | f_i^* \in \mathcal{F}_i) \right] \text{DP} \left(\mu^* + \sum_{i \in [K]} c_i \delta_{f_i^*} \right) df_{1:K}^*. \end{aligned}$$

Posterior of $\{D_\phi\}_{\phi \in \Theta} | \{z_{1:(n-1)}, \theta_{1:(n-1)}\}$.

Similarly, for any $\phi \in \Theta$, given functions $f_{1:K}^*$, the posterior of D_ϕ is a DP obtained from restricting and projecting the posterior of D^* . Integrating all functionals $f_{1:K}^*$, we have the posterior of D_ϕ which is MDP distributed,

$$\begin{aligned} M_\phi &\triangleq D_\phi | \{z_{1:(n-1)}, \theta_{1:(n-1)}\} \\ &= \int_{f_{1:K}^*} \left[\prod_{i \in [K]} \Pr(f_i^* | f_i^* \in \mathcal{F}_i) \right] \text{DP} \left(\mu_\phi + \sum_{i \in [K]} c_i t_{f_i^*}^{(\phi)} \delta_{f_i^*(\phi)} \right) df_{1:K}^*, \end{aligned} \quad (13)$$

where $t_{f_i^*}^{(\phi)} = \delta_{\phi \in \text{domain}(f_i^*)}$ is an indicator variable.

Although the posterior of D_ϕ is a mixture of uncountably infinite number of DPs, it is that all functions f such that $\phi \in \text{domain}(f)$ and $f(\phi) = \theta$ for any fixed θ have exactly same effect in (13). By rearranging and grouping the components in (13), we obtain a simpler form.

$$\begin{aligned} M_\phi &= \mathbb{E}_{t_{1:K}^{(\phi)}, \theta_{1:K}^{(\phi)}} \left[\text{DP} \left(\mu_\phi + \sum_i t_i^{(\phi)} c_i \delta_{\theta_i^{(\phi)}} \right) \right], \\ t_i^{(\phi)} &\stackrel{i.i.d.}{\sim} \text{Bernoulli}(q_i^{(\phi)}), \quad \theta_i^{(\phi)} \stackrel{i.i.d.}{\sim} T_i^{(\phi)}(\cdot), \end{aligned} \quad (14)$$

where $q_i^{(\phi)}$ and $T_i^{(\phi)}$ are induced from G^* ,

$$\begin{aligned} q_i^{(\phi)} &= \Pr_{G^*} (\phi \in \text{domain}(f) | f \in \mathcal{F}_i), \\ T_i^{(\phi)}(d\theta) &= \Pr_{G^*} (f(\phi) = \theta | \phi \in \text{domain}(f), f \in \mathcal{F}_i). \end{aligned} \quad (15)$$

Applying the independent assumption made in 4, we can further simplify 15 and express them by the distribution of function domains and the base process

$$\begin{aligned} q_i^{(\phi)} &= \Pr (\phi \in \text{domain}(f) | \{\phi_{1:(n-1)}\} \subseteq \text{domain}(f)) \\ &= \frac{\varphi(\{\phi_{1:(n-1)}, \phi\})}{\varphi(\{\phi_{1:(n-1)}\})}, \end{aligned} \quad (16)$$

$$T_i^{(\phi)}(d\theta) = \Pr (f(\phi) = \theta | f(\phi_1) = \theta_1, \dots, f(\phi_{n-1}) = \theta_{n-1}).$$

Predictive distribution.

The predictive distributions $z, \theta | \{z_{1:(n-1)}, \theta_{1:(n-1)}\}$ can be readily obtained by marginalizing DPs in 14 out,

$$\begin{aligned} z | \{z_{1:(n-1)}, \theta_{1:(n-1)}\} &= \mathbb{E}_{t_{1:K}^{(\phi)}, \theta_{1:K}^{(\phi)}} \left[\frac{\mu_\phi(\Omega) \delta_{K+1} + \sum_{i \in [K]} t_i^{(\phi)} c_i \delta_i}{\mu_\phi(\Omega) + \sum_{i \in [K]} t_i^{(\phi)} c_i} \right], \\ \theta | z, \{z_{1:(n-1)}, \theta_{1:(n-1)}\} &= T_z^{(\phi)}(\cdot). \end{aligned} \quad (17)$$

where δ_{K+1} indicates the event of assigning to a new function.