Coupled Dirichlet Processes: Beyond HDP

Dahua Lin
MIT CSAIL
dhlin@mit.edu

John Fisher
MIT CSAIL
fisher@csail.mit.edu

Abstract

Dirichlet process mixture models (DPMMs) have become an important tool to describe complex data in the past decade. However, learning multiple DPMMs over related sets of observations remains an open question. A popular approach to this problem is the Hierarchical Dirichlet Process (HDP) [10], which is limited in that models are required to be organized as a tree. In this paper, we present a generic framework to construct dependent DP mixtures, drawing on a new formulation that we proposed recently [5]. This framework breaks the limitations of HDP, allowing each mixture model to inherit atoms from multiple sources with covariate-dependent probabilities. We show, through experiments on real data, that the proposed framework allows one to devise models that capture the dependency between different data sets more accurately – this capability is important in a context with multiple data sources, such as autonomous planning and distributed sensing.

1 Introduction

A mixture model is a probabilistic formulation where the probability density function is represented as a convex combination of component densities as

$$f_{mix}(x) = \sum_{\theta \in \Theta} \pi_\theta f(x|\theta). \tag{1}$$

Here, Θ is a set of component parameters. A classical formulation that has been widely used in practice is the *finite mixture model*, which requires the number of components $|\Theta|$ to be specified a priori. Whereas this setting results in simple methods such as expectation-maximization (EM) to learn model parameters, estimating $|\Theta|$ is often a nontrivial problem.

Dirichlet process mixture models (DPMMs) from Bayesian nonparametrics provide an elegant solution to this problem, which considers the component parameters as generated from a Dirichlet process (DP). Each realization of a DP, as shown by Sethuraman [8], is almost surely a distribution over a countably infinite set, thus naturally providing a pool of infinitely many components.

Real world applications often involve multiple sets of observations that are related to each other. Modeling such data generally requires multiple mixture models (one for each set), with statistical connections between them. Various methods have been proposed to address this problem. Hierarchical Dirichlet process (HDP) [10] is one of the most popular. An HDP is a tree of DPs, with each parent DP serving as the base measure for its children. Consequently, atoms are shared among DPs with the same parent. HDPs have been extended in a variety of ways. Kim and Smyth [4]incorporated group-specific random perturbations, allowing component parameters to vary across different groups. Ren *et al.* [7] proposed dynamic HDPs, where the DP at each time step combines the DP at a previous time step and a new one. Fox *et al.* [1] introduced a sticky extension of the HDP-HMM, making it more robust to estimate nonparametric hidden Markov chains.

In this paper, we first revisit the formulation of HDP. Detailed analysis reveals a key limitation that need to be addressed – the tree structure restricts the flexibility of model design. With an aim to break this limitation, we consider a generic framework that allows more flexible structure, which is based on a new construction of dependent DPs that we recently proposed [5]. In this framework, each mixture can inherit atoms from multiple sources, and more importantly, one can freely control the contribution of each source to a mixture.

It is worth noting that the capability of relating multiple mixture models in a flexible way is important in a number of real world applications, including planning and distributed sensing. For example, in the process of planning, agents would acquire information from different sources at different time steps. Data are often aggregated in the form of mixture models. If dependencies between these models can be reliably captured, they can be utilized to improve the planner. Moreover, the proposed framework supports a variety of model dynamics, which can also be leveraged by a planner to adapt to a changing environment.

2 Analysis of HDP

Problems with grouped observations arise in many applications, *e.g.* categorized documents and sensor networks. Specifically, suppose there are M different groups: $X_1, \ldots, X_M$. We describe these observations via a set of mixture models that respectively use Dirichlet processes $D_1, \ldots, D_M$ as non-parametric priors to the mixture components. These groups are often related. The desire to leverage such relations to share statistical strength among different groups motivates the development of dependent DPs.

2.1 Dirichlet Process Mixture Models

We first provide a brief review of Dirichlet processes and DP mixture models. Let Ω be a measurable space. The *Dirichlet process* $\mathrm{DP}(\alpha B)$ over Ω is a probability measure of measures, characterized by two parameters: the *concentration parameter* α and the *base measure* B. Sethuraman [8] showed that each sample $D \sim \mathrm{DP}(\alpha B)$ is almost surely discrete and can be expressed via a stick-breaking process:

$$v_k \sim \mathrm{Beta}(1, \alpha), \quad \pi_k = v_k \prod_{l=1}^{k-1} (1 - v_l), \quad D = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}. \tag{2}$$

This discrete nature makes a DP suited as a prior for mixture components. Specifically, a DP mixture model is formulated as below.

$$D \sim \mathrm{DP}(\alpha, B), \quad \theta_i | D \sim D, \; x_i | \theta_i \sim \mathcal{G}(\theta_i), \; \text{for } i = 1, \ldots, n. \tag{3}$$

Here, $\mathcal{G}(\theta_i)$ denotes a component model with parameter θ_i. In this formulation, each sample x_i is associated with a component parameter θ_i, which, in turn, is an atom generated from D, the prior DP. An atom can be repeatedly generated from D, resulting in clusters of observations, each associated with the same atom.

2.2 Hierarchical Dirichlet Processes

A *Hierarchical Dirichlet Process (HDP)* generalizes the formulation above to describe multiple groups of observations with shared atoms. Particularly, it introduces a latent DP, which we denote by D_0, as the base measure for the DPs associated with observed groups. The formulation is given as follows:

$$D_0 \sim \mathrm{DP}(\gamma, B); \qquad D_t | D_0 \sim \mathrm{DP}(\alpha_0, D_0), \; \text{for } t = 1, \ldots, M. \tag{4}$$

A HDP is characterized by three hyper-parameters: (1) B, which provides the prior distribution for the component parameters, (2) γ, which governs the variability of D_0, and (3) α_0, which governs how closely the child DPs are coupled. A key aspect of this formulation is that D_0 is discrete, which can be understood as a universal pool of atoms to be shared by the child DPs, and all child DPs acquire atoms from this pool. This

Recent years has witnessed wide adoption of HDP in practice [2, 9, 11], which, in our view, is partly ascribed to the fact that it provides an elegant solution for atom sharing. Yet, as a modeling tool, it is limited in several aspects:

First, groups have to be organized into a tree, which is not necessary the optimal structure. Consider a case with three groups: A, B, and C, where B is closely related to both A and C, while the relation between A and C is weaker. In this case, one may consider two HDP configurations: (1) $\{A, B, C\}$: all groups share the same parent. This setting does not reflect the difference that the relation between A and C is weaker than that between A and B. (2) $\{\{A, B\}, C\}$: A and B share a parent, while C has a different parent. Here, the strong relation between B and C is not captured. Neither setting is able to capture the structure properly.

Second, all atoms of a mixture model are inherited from a unique source – the model's parent. Whereas this provides an effective way to share atoms, a machinery is still needed to differentiate between atoms to be shared across all groups and those only used by specific groups.

3 A Generic Framework based on Poisson Processes

Recently, we proposed a new approach to constructing dependent DPs [5]. Drawing on the intrinsic connection between Dirichlet and Poisson processes, this approach provides three primitive operations for DP construction and transformation. This method was used in [5] to derive a Markov chain of DPs for describing phenomena that evolve over time. In this paper, we extend this method to a generic framework that can be applied in a much broader context.

3.1 Poisson-based DDP

We first briefly review the theoretical connection between Dirichlet, Gamma, and Poisson processes, as well as the three primitive operatons derived thereon.

Consider $D \sim \mathrm{DP}(\alpha B)$. Underlying this DP is a *Poisson process* over Ω^* with mean measure $\alpha B \times \gamma$, where γ is a special measure over $\mathbb{R}^+$ defined by $\gamma(dw) = w^{-1} e^{-w} dw$. Let Π^* be a realization of this Poisson process. Then Π^* is a countably infinite set of pairs as $\Pi^* = \{(\theta_k, w_k)\}_{k=1}^{\infty}$, where $\theta_k \in \Omega$ is an atom, and w_k is the associated weight. From Π^*, we can derive a measure over Ω, given by $G = \sum_{k=1}^{\infty} w_k \delta_{\phi_k}$. It can be shown that G is a sample path of a *Gamma process*, as $G \sim \mathrm{Gamma}(\alpha B)$. Then, one can obtain a sample path of a DP, as given in Eq.(2), by normalizing the weights (*i.e.* setting $\pi_k = w_k / \sum_l w_l$).

These relations suggest a way to construct dependent DPs, namely to transform the underlying Poisson processes. In [5], we developed three such operations: *superposition*, *sub-sampling*, and *transition*. While these operations were derived via operations on the underlying Poisson processes, they can be directly applied to DPs without making the Poisson processes explicit.

(1) Superposition. Let $D_k \sim \mathrm{DP}(\alpha_k B_k)$ for $k = 1, \dots, K$ be independent DPs and $(c_1, \dots, c_K) \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_K)$. Then the stochastic convex combination of the DPs as below is also a DP:

$$D_1 \oplus \cdots \oplus D_K \triangleq \sum_{k=1}^{k} c_k D_k \;\sim\; \mathrm{DP}(\alpha_s, \tilde{\alpha}_1 B_1 + \cdots + \tilde{\alpha}_K B_K). \tag{5}$$

Here, $\alpha_s = \sum_{k=1}^{K} \alpha_k$ and $\tilde{\alpha}_k = \alpha_k / \alpha_s$ for each k.

(2) Sub-sampling. Let $D = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \sim \mathrm{DP}(\alpha, B)$. We obtain a new DP by sampling a Bernoulli random variable r_k, $P(r_k = 1) = q$, retaining those atoms with $r_k = 1$, and renormalizing the coefficients, as

$$q \circ D \triangleq \sum_{k: r_k = 1}^{\infty} \pi'_k \delta_{\phi_k} \sim \mathrm{DP}(\alpha q, B). \tag{6}$$

Here, $\pi'_k = \pi_k / \sum_k r_k \pi_k$ is the renormalized coefficient for the k-th atom.

(3) Transition. Consider D as in Eq.(2). Perturbing each atom ϕ_k independently following a probabilistic transition kernel T would yield a new DP, given by $T(D) \triangleq \sum_{k=1}^{\infty} b_k \delta_{T(\phi_k)}$.

3.2 A Generic Formulation

These operations together lead to a generic construction of new DP that depends on existing ones, as stated by the theorem below.

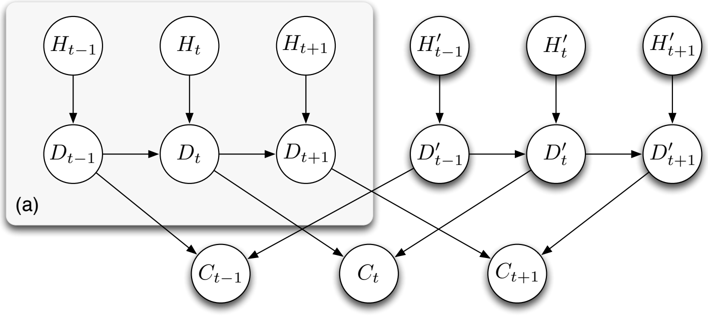

Figure 1: This figure shows the graphical structure of the chain models. Part (a) shows the structure for Eq.(8), while the whole diagram shows the extended structure as given by Eq.(10).

Theorem 1. *Let B be a probability measure over Ω, $H_1, \ldots, H_M$ be independent DPs with $H_s \sim \mathrm{DP}(\alpha_s B)$, and $q_1, \ldots, q_M \in [0, 1]$, then*

$$D = (q_1 \circ H_1) \oplus \cdots \oplus (q_M \circ H_M) \;\sim\; \mathrm{DP}(\beta B), \quad \textit{with } \beta = \sum_{s=1}^{M} \alpha_s q_s. \tag{7}$$

Here, the coefficients $q_1, \ldots, q_M$ are called *inheritance probabilities*, which governs the contribution of each source to D. Like HDP, this formulation allows atoms to be shared across DPs that inherit from common sources. More importantly, it provides greater flexibility in two aspects: (1) each dependent DP can acquire atoms from multiple sources with different weights; and (2) one can control the range of sharing by setting up multiple sources connected to different sets of groups.

In what follows, we will demonstrate, through two examples, how such a flexible formulation can be utilized to address modeling problems arising in real world applications.

3.3 Chain Models

In [5], we considered a chain model which combines the three operations above to derive a Markov chain of DPs, as

$$D_t = (q \circ D_{t-1}) \oplus H_t. \tag{8}$$

Here, at each time step t, D_t inherits atoms from D_{t-1} with probability q, and obtains new atoms from an independent DP: $H_t \sim \mathrm{DP}(\alpha_h B)$. The graphical representation is shown in part (a) of Figure 1. Let α_t be the concentration parameter for D_t, then we have $\alpha_t = q\alpha_{t-1} + \alpha_h$. In modeling a long sequence, it is often desirable to have a constant concentration over time, this can be easily achieved by setting $\alpha_h = \alpha_0(1 - q)$. Here, we can use q to control the balance between innovation and inheritance.

Note that it is possible to construct a Markov chain of DPs using HDP, as follows.

$$D_t | D_{t-1} \;\sim\; \mathrm{DP}(\gamma_t D_{t-1}). \tag{9}$$

However, this formulation suffers from an important drawback that makes it unsuitable for practical use: D_t only contains a subset of atoms from D_{t-1} and there is no way to introduce new atoms dynamically. This also implies that all atoms that have ever appeared throughout the entire sequence must be contained in D_0.

In the formulation given by Eq.(8), the expected lift span of every atom is the same, which is $1/(1 - q)$. To allow atoms to have different life spans, one can extend this formulation to incorporate multiple chains $D_t^{(1)}, \ldots, D_t^{(m)}$, each with a different inheritance probability, and express the mixture models associated with the observed data, which we denote by C_t, as a combination of these chains, as

$$D_t^{(k)} = (q^{(k)} \circ D_{t-1}^{(k)}) \oplus H_t^{(k)}, \;\text{ for } k = 1, \ldots, m, \quad C_t = D_t^{(1)} \oplus \cdots \oplus D_t^{(m)}. \tag{10}$$

Figure 1 illustrates the model structure of this extension.

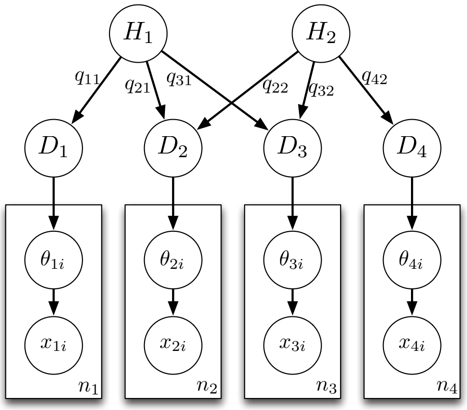

Figure 2: This shows the graphical model of the coupled DP formulation on a case with four groups and two latent DPs. Each mixture model D_t inherits atoms from H_s with a probability q_{ts}

3.4 Coupled Mixture Models with Latent DPs

Groups of data are not necessarily organized as a chain in many applications. *Can we further extend the boundary of DP mixture models for such applications?* Let's consider a set of covariates $\mathcal{T}$. Our primary goal here is to develop a joint formulation over a group of DPs $\{D_t : t \in \mathcal{T}\}$, each for a covariate $t \in \mathcal{T}$, where components are shared by different groups and the weights of each component vary across groups.

Here, we propose a formulation as illustrated in Figure 2, which introduces a set of *latent DPs* $\{H_s : s \in \mathcal{S}\}$ as sources of atoms. Each covariate $t \in \mathcal{T}$ is associated with a DP D_t that inherits atoms from the latent sources. In particular, the atoms in H_s are inherited by D_t with probability q_{ts}. This formulation can be expressed by the formula below

$$D_t = \bigoplus_{s \in \mathcal{S}} (q_{ts} \circ H_s), \quad \text{for } t \in \mathcal{T}, \qquad \text{with } H_s \sim \mathrm{DP}(\alpha_s B). \tag{11}$$

We derive the following theorem that characterizes the covariance between the dependent DPs formulated above.

Theorem 2. *Let $t_1 \neq t_2$ and U be a measurable subset of Ω, then*

$$\mathrm{Cov}(D_{t_1}(U), D_{t_2}(U)) = \frac{1}{\beta_{t_1}\beta_{t_2}} \sum_{s=1}^{M_L} \frac{(\alpha_s q_{t_1 s} q_{t_2 s})^2}{\alpha_s q_{t_1 s} q_{t_2 s} + 1} B(U)(1 - B(U)). \tag{12}$$

With this formulation, the variability of each DP as well as the correlation between different DPs can be flexibly controlled via α_s and q_{ts}. In general, two models are strongly coupled, if there exists a subset of latent DPs, from which both inherit atoms with high probabilities, while their coupling is much weaker if the associated inheritance probabilities are set differently. Another important factor is $|\mathcal{S}|$, the number of latent DPs. A large number of latent DPs enables fine-grained control at the cost of increased complexity.

3.5 Comparison with SNΓP

We note that SNΓP proposed by Rao and Teh [6] also allows each mixture model to inherit atoms from multiple sources. Specifically, it defines a gamma process G over an extended space. For each group t, D_t is derived through normalized restriction of G into a measurable subset. The DPs derived on overlapped subsets are dependent. This construction can be reduced to a formulation in the form $D_t = \sum_{j \in R_t} c_{tj} H_j$, where R_t is the subset of latent DPs used for D_t. Comparing this with Eq.(7), we can see that it is essentially a special case of the construction presented above without sub-sampling (*i.e.* all q_{ts}-values equal 1). Consequently, the combination coefficients have to satisfy $(c_{tj})_{j \in R_t} \sim \mathrm{Dir}((\alpha_j)_{j \in R_t})$, implying that the relative weights of two latent sources are restricted

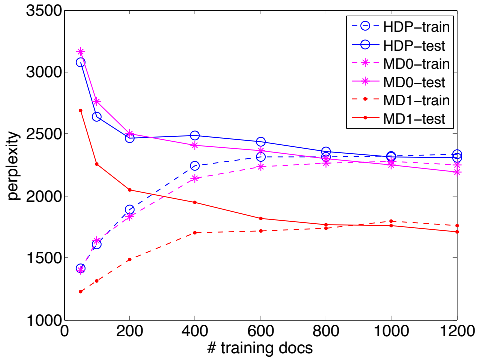

Figure 3: The results obtained on the NIPS data using different numbers of training documents. The average perplexity values (over 20 runs) of each model are evaluated on both training and testing set.

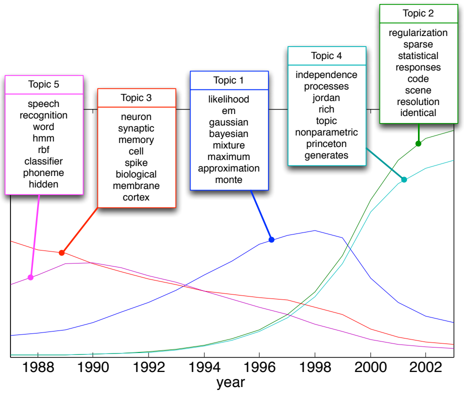

Figure 4: This shows the trends of the year-specific weights of the top five topics, and the eight leading words of them.

to be the same in all groups that inherit from both. In contrast, the approach here allows the weights of latent DPs to vary across groups.

4 Application to Document Analysis

To demonstrate the practical utility of the generic formulation presented above, we performed empirical tests on real data. Specifically, we use the NIPS (1-17) database [3], which consists of all 2484 papers published at NIPS from 1987 to 2003. These papers are partitioned into 17 groups, each for a year. Each paper is characterized by a bag of words.

For the proposed model, we place a weak prior over α_s, as $\alpha_s \sim \mathrm{Gamma}(0.1, 0.1)$. In this way, the actual value of α_s is basically determined by the data in sampling. The base distribution B is assumed to be a symmetric Dirichlet distribution with the concentration parameter set to 1. We consider two settings: (1) MD0: one "global" latent DP connecting to all groups, with inheritance probability 0.5. (2) MD1: in addition to the "global" latent DP, we add a set of "local" latent DPs, each connecting to three consecutive years with inheritance probability 0.9, and to other years with a relatively low probability 0.2. We also test HDP following the settings in [10]. Gibbs sampling is used to estimate the model parameters.

We measure the performance in terms of *perplexity*. A good model generally yields small perplexity on testing documents. We randomly divide papers into two halves, respectively for training and testing. For each test, models are estimated from a subset of specific size randomly chosen from the training set, which are then tested on both the chosen training subset and the held-out testing set. The reason why we measure performance on both sets is to study the gap between empirical and generalized performance.

From Figure 3, we observe: (1) For all models, as the training set grows, the perplexity measured on the training set increases and that on the testing set decreases. The training and testing curves converge when the training set size increases beyond 800. (2) The proposed model under MD1 setting significantly and consistently outperforms the other two across all training set sizes. When the whole training set is used, MD1 yields perplexity at 1720, while MD0 and HDP respectively yield 2200 and 2300. This improvement is largely due to the use of local latent DPs that closely couple the mixture models for consecutive years, sharing statistical strength between them.

We take a closer examination of the estimated models, by evaluating the year-specific weights of the topics. Suppose the topic ϕ_j is the k-th atom contained in the latent DP A_s, then its *year-specific weight* for year t is defined to be $\mathbf{c}_t(s) r_{sk}^t \pi_{sk}$. Here, $\mathbf{c}_t(s)$ is the probability that an atom in D_t comes from A_s, r_{sk}^t indicates whether the atom is inherited, and π_{sk} is its relative weight within A_s. Figure 4 shows the trends of such weights of the top five topics (ranked in descending order of total weights). We can see considerable changes of the weights over time, reflecting the rapid evolution taking place in the field.

5 Conclusion

We presented a generic formulation for constructing dependent Dirichlet processes, based on the primitive operations developed in [5]. As compared to HDP, where mixture models have to be organized as a tree, this formulation provides greater flexibility: each mixture model can inherit atoms from multiple groups with different probabilities. Taking advantage of this flexibility, one can devise models that capture the data dependency more accurately, as we have demonstrated on the NIPS dataset.

References

[1] Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. An HDP-HMM for Systems with State Persistence, 2008.

[2] Emily B. Fox, Erik B. Sudderth, and Alan S. Willsky. Hierarchical dirichlet processes for tracking manuevering targets. In *Proc. of International Conference on Info Fusion*, 2007.

[3] Amir Globerson, Gal Chechik, Fernando Pereira, and Naftali Tishby. Euclidean embedding of co-occurrence data. *JMLR*, 8, 2007.

[4] Seyoung Kim and Padhraic Smyth. Hierarchical dirichlet processes with random effects. In *Proc. of NIPS'06*, 2006.

[5] Dahua Lin, Eric Grimson, and John Fisher. Construction of dependent dirichlet processes based on poisson processes. In *Advances of NIPS'10*, 2010.

[6] Vinayak Rao and Yee Whye Teh. Spatial Normalized Gamma Processes. In *Proc. of NIPS'09*, 2009.

[7] Lu Ren, David B. Dunson, and Lawrence Carin. The Dynamic Hierarchical Dirichlet Process. In *Proc. of ICML'08*, New York, New York, USA, 2008. ACM Press.

[8] J. Sethuraman. A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4(2):639–650, 1994.

[9] By Kyung-Aj Sohn and Eric P. Xing. A hierarchical dirichlet process mixture model for haplotype reconstruction from multi-population data. *The Annals of Applied Stats*, 3(2):791–821, 2009.

[10] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[11] Elias Zavitsanos, Georgios Paliouras, and George A. Vouros. Non-parametric estimation of topic hierarchies from texts with hierarchical dirichlet processes. 2011.