Conditionally Dependent Dirichlet Processes for Modelling Naturally Correlated Data Sources

Dinh Phung[†], XuanLong Nguyen[‡], Hung Bui[*], Vu Nguyen[†] and S. Venkatesh[†]

[†] *Centre for Pattern Recognition and Data Analytics (PRaDA),*
Deakin University, Australia.
{dinh.phung,tvnguye,svetha.venkatesh}@deakin.edu.au

[‡] *Department of Statistics, Dept of Electrical Engineering and Computer Science*
University of Michigan. xuanlong@umich.edu

[*] *Laboratory for Natural Language Understanding,*
Nuance Communications, Sunnyvale, CA 94085, USA. bui.h.hung@gmail.com

Pattern Recognition and Data Analytics
School of Information Technology, Deakin University, Australia.
Locked Bag 20000, Geelong VIC 3220, Australia.
Tel: +61 3 5227 2150
Report Number: TR-PRaDA-02/12, Octocber, 2012.

Conditionally Dependent Dirichlet Processes for Modelling Naturally Correlated Data Sources

Dinh Phung[†], XuanLong Nguyen[‡], Hung Bui[*][*], Vu Nguyen[†] and S. Venkatesh[†]

[†]*Centre for Pattern Recognition and Data Analytics (PRaDA),*
Deakin University, Australia.
{dinh.phung,tvnguye,svetha.venkatesh}@deakin.edu.au

[‡]*Department of Statistics, Dept of Electrical Engineering and Computer Science*
University of Michigan. xuanlong@umich.edu

[*]*Laboratory for Natural Language Understanding,*
Nuance Communications, Sunnyvale, CA 94085, USA. bui.h.hung@gmail.com

Abstract

We introduce a new class of conditionally dependent Dirichlet processes (CDP) for hierarchical mixture modelling of naturally correlated data sources. This class of models provides a Bayesian nonparametric approach for modelling a range of challenging datasets which typically consists of heterogeneous observations from multiple correlated data channels. Some typical examples include annotated social media, networks in community where information about friendship and relation coexist with user's profiles, medical records where patient's information exists in several dimension (demographic information, medical history, drug uses and so on). The proposed framework can easily be tailored to model multiple data sources which are correlated by some latent underlying processes, whereas most of existing topic models, notably hierarchical Dirichlet processes (HDP), is designed for only a single data observation channel. In these existing approaches, data are grouped into documents (e.g., text documents or they are grouped according to some covariates such as time or location). Our approach is different: we view context as distributions over some index space and model both topics and contexts jointly. Distributions over topic parameters are modelled according to the usual Dirichlet processes. Stick-breaking representation gives rise to explicit realizations of topic atoms which we use as an indexing mechanism to induce conditional random mixture distributions on the context observation spaces – loosely speaking, we use a stochastic process, being DP, to conditionally 'index' other stochastic processes. The later can be designed on any suitable family of stochastic processes to suit modelling needs or data types of contexts (such as Beta or Gaussian processes). Dirichlet process is of course an obvious choice. Our model can be viewed as an integration of the hierarchical Dirichlet process (HDP) and the recent nested Dirichlet process (nDP)

[*]Part of the work was done when the author was at the Artificial Intelligence Center, SRI International, Menlo Park, USA.

with shared mixture components. In fact, it provides an interesting interpretation whereas, under a suitable parameterization, integrating out the topic components results in a nested DP, whereas integrating out the context components results in a hierarchical DP. Different approaches for posterior inference exist. This paper focus on the development of an auxiliary conditional Gibbs sampling in which both topic and context atoms are marginalized out. We demonstrate the framework on synthesis datasets for temporal topic modelling and trajectory discovery in videos surveillance. We then demonstrate an application on a current visual category classification challenge in computer vision for which we significantly outperform the current reported state-of-the-art results. Finally, it is worthwide to note that our proposed approach can be easily twisted to accommodate different forms of supervision (weakly annotated data, semi-supervision) and to perform prediction.

1 Introduction

Bayesian nonparametric methods have recently emerged in machine learning and data mining as an extremely useful modeling framework due to their model flexibility capable of fitting a wide range of data types. A widely-used application of Bayesian nonparametric is clustering data where models for inducing discrete distributions on a primary parameter space – the (hierarchical) Dirichlet process and Beta processes are two noticeable examples. In these clustering models, when multiple covariates are present, they are often treated as independent factors in a given the cluster; more generally, one has to make a choice of a parametric model of the covariates inside each cluster. Addressing more realistic problems in machine learning and data mining requires a need to advance Bayesian nonparametric modeling, both in theory and computation, to accommodate richer types of data in a principled way.

When one considers realistic multimodal data, covariates are rich, and yet tend to have a natural correlation with one another; for example: tags and their associated multimedia contents; patient's demographic information, medical history and drug usage; social user's profile and friendship network. The presence of rich and naturally correlated covariates calls for the need to model their correlation with nonparametric models, without reverting to making parametric assumptions. These needs have been recognized in various recent discussions and is reflected in MacEachern's remark: " [...] current nonparametric models are inadequate in that they do not easily accommodate covariates. Currently, two distinct distributions of random effects are (conditionally) independent realizations from a nonparametric prior distribution; either a single elaboration or several independent elaborations are conducted. The remedy for these, and many other inadequacies, lies in correlated, or dependent nonparametric processes. In particular, extension of the Dirichlet process provides a class of models that are attractive conceptually and computationally, and that capture many fundamental modelling strategies which have heretofore been inaccessible." [1]

This paper presents a full Bayesian nonparametric approach to the problem of jointly clustering

[1] MacEachern's abstract to Workshop on Bayesian Nonparametrics
(available online: http://www.stats.bris.ac.uk/~guy/Research/WORKSHOP/speakers_abstracts.html#gelfand)

data sources and modeling their correlation. To simplify presentation, throughout the paper, we consider the case of a pair of primary naturally correlated data sources, abstractly referred to as *content* and *context* (which covariates constitute content and which constitute context will be decided per actual application).

In our approach, we view context as distributions over some index space, governed by the topics discovered from the primary data source (content), and model both contents and contexts jointly. We impose a conditional structure in which contents provide the topics, upon which contexts are conditionally distributed. Distributions over topic parameters are modelled according to a Dirichlet processes (DP). Stick-breaking representation gives rise to explicit realizations of topic atoms which we use as an indexing mechanism to induce conditional random mixture distributions on the context observation spaces. Loosely speaking, we use a stochastic process, being DP, to conditionally 'index' other stochastic processes. The later can be designed on any suitable family of stochastic processes to suit modelling needs or data types of contexts (such as Beta or Gaussian processes). Dirichlet process is of course an obvious choice and will be again employed in this paper. In typical hierarchical Bayesian style, we also provide the model in grouped data setting, where contents and contexts appear in groups (for example, a collection of text documents or images embedded in time or space).

Our model can be viewed as a generalization of the hierarchical Dirichlet process (HDP) [28] and the recent nested Dirichlet process (nDP) [21]. In fact, it provides an interesting interpretation whereas, under a suitable parameterization, integrating out the topic components results in a nested DP, whereas integrating out the context components results in a hierarchical DP. Different approaches for posterior inference exist. This paper focus on the development of an auxiliary conditional Gibbs sampling in which both topic and context atoms are marginalized out. We demonstrate the framework on synthesis datasets for temporal topic modelling and trajectory discovery in videos surveillance. We then demonstrate an application on a current image classification challenge in computer vision for which we significantly outperform the current reported state-of-the-art results. Finally, it is worthwide to note that our proposed approach can be easily twisted to accommodate different forms of supervision (weakly annotated data, semi-supervision) and to perform prediction.

In brief, our key contributions in this paper includes: a) a new Bayesian nonparametric approach for modeling topics and nested contexts linked to data; b) an interesting model interpretation that provides a connection between two most popular classes of Bayesian nonparametric models, namely the hierarchical DP and the nested DP; c) an efficient auxiliary conditional Gibbs sampling approach for this models, c) a demonstration of the proposed modeling approach on various applications, both with simulated and real-world problems, d) the proposed model is also flexible and can readily be extended for more challenging data structure, making it attractive for many data modelling tasks, especially in the presence of heterogeneous, high-dimensional and richly connected data sources which we shall further highlight in the discussion section.

2 Related Background

There has been a very large body of work on hierarchical mixture modeling for text and image data, which can be placed under a very broad umbrella known as "topic models". The Latent Dirichlet Allocation (LDA) is perhaps the most well-known. There has been extensions to the LDA to incorporate contextual information, specially both time and space. Hierarchical mixing distributions that vary over time [4, 31] or over space [30]. [3] provided an excellent review of recent work on topic modeling.

A notable strand in both recent machine learning and statistics literature focused on construction of Dirichlet process-based models that enable infinite mixture distributions that vary over an indexed set, where the index might represent time or spatial information, see a recent book edited by [16]. While there are generally a number of methods for doing this via the dependent DP framework of [13], we highlight several approaches that build on the nonparametric and hierarchical modeling framework advocated by [28]. For instance, the dynamic HDP [20, 19] constructs a sequence of HDP-distributed mixing distribution that varies over time, while the nested HDP [15] constructs a collection of HDP-distributed mixing distribution that varies over general covariate space.

All above work can be viewed as "context-sensitive" topic models: data are grouped according to context (such as time or location), where each group is described by a mixture model (whether parametric or nonparametric). Our approach is different: we view context as distribution over some index space (such as time or locations), and model both topic and context jointly. To link context distributions with topic distributions we utilize the nonparametric and hierarchical modeling: A distribution over context can be viewed as random, conditionally on the topic, where the topic variables are distributed according to some Dirichlet process mixture. The resultant model can be described in terms of a collection of DP processes that are hierarchically linked, a modeling idea that were advocated by [27]. However, because we are modeling jointly the topic and the context distribution, our model is related to fundamentally different class of model known as the nested Dirichlet process [21]. In fact, it provides an interesting interpretation whereas, under a suitable parameterization, integrating out the topic components results in a nested DP, whereas integrating out the context components results in a hierarchical DP.

To provide the background for our paper, we first briefly review the Dirichlet process and its related models including the Dirichlet process mixture model (DPM) and the hierarchical Dirichlet processes (HDP). The description of the proposed model then follows.

2.1 Dirichlet Processes and Hierarchical Dirichlet Processes

A Dirichlet process DP (γ, H) is a distribution of a random probability measure G over the measurable space $(\Theta, \mathcal{B})$ where H is a *base* probability measure and $\gamma > 0$ is the *concentration* parameter. It is defined such that, for any finite measurable partition $(A_k : k = 1, \ldots, K)$ of Θ, the resultant finite-dimensional random vector $(G(A_1), \ldots, G(A_k))$ is distributed according to a Dirichlet distribution with parameters $(H(A_1), \ldots, H(A_k))$.

Dirichlet process and its existence was established by Ferguson [7] who has also showed that draws from a DP are discrete with probability one. Sethuraman [24] provides an alternative constructive definition which makes the discreteness property of a Dirichlet process explicitly via a stick-breaking construction:

$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \tag{1}$$

where $\phi_k \overset{\text{iid}}{\sim} H, k = 1, \ldots, \infty$ and $\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty}$ are the weights constructed through a 'stick-breaking' process $\beta_k = v_k \prod_{s<k} (1 - v_s)$ with $v_k \overset{\text{iid}}{\sim} \text{Beta}(1, \gamma), k = 1, \ldots, \infty$. It can be shown that $\sum_{k=1}^{\infty} \beta_k = 1$ with probability one, and as a convention [18], we hereafter write $\boldsymbol{\beta} \sim \text{GEM}(\gamma)$.

Yet another useful interpretation for the Dirichlet process is given by the Polya urn scheme [2] which shows that draws from the Dirichlet process are not only discrete, but also exhibit a clustering property. More concretely, let $\theta_1, \theta_2, \ldots, \theta_{n+1}$ be iid draws from G, Blackwell and MacQueen [2] showed that G can be integrated out to give the following marginal conditional distribution form:

$$\theta_{n+1} \mid \theta_1, \ldots, \theta_n, \gamma, H \sim \sum_{i=1}^{n} \frac{1}{n+\gamma} \delta_{\theta_i} + \frac{\gamma}{n+\gamma} H \tag{2}$$

If we further group identical values in the set $\{\theta_1, \ldots, \theta_n\}$ together and let K be the number of such distinct values, each represented by ϕ_k with n_k be its count, then Eq (2) is equivalent to:

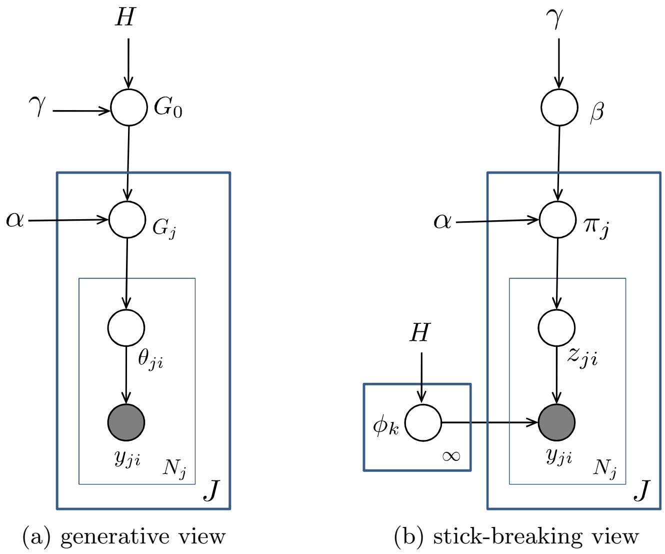
$$\theta_{n+1} \mid \theta_1, \ldots, \theta_n, \gamma, H \sim \sum_{k=1}^{K} \frac{n_k}{n+\gamma} \delta_{\phi_k} + \frac{\gamma}{n+\gamma} H$$

This expression is clearly showing the clustering property induced by G: a future draw θ is likely to return to an existing atom ϕ_k and it does so with a probability proportional to the popularity n_k of the respective atom; however it may also pick on a new value with a probability proportional to the concentration parameter γ – a view which is also known as the Chinese restaurant process.

However, due to its discreteness, the Dirichlet process is often not applied directly to model data (e.g., it is unable to model continuous data) instead it can be effectively used as a nonparametric prior on the mixture components θ, which in turn serves as the parameters within another likelihood function F to generate data - a model which is known as Dirichlet process mixture model (DPM) [1, 6]. To be precise, under a DPM formalism an observation x_n is generated from a two-step process: $x_n \sim F(x_n \mid \theta_n)$ where $\theta_n \sim G$. Using the stick-breaking representation in Eq 1, it is not hard to see that DPM yeilds an *infinite* mixture model representation:

$$p(x \mid \gamma, H) = \sum_{k=1}^{\infty} \beta_k f(x \mid \phi_k) \tag{3}$$

where f denotes the density function for F. Dirichlet process mixture models have been embraced with a great success and enthusiasm recently [8, 14]. The crucial advantage is its ability to naturally address the problem of model selection - a major obstacle encountered in several parametric mixture modeling, such as the Gaussian mixture models whose number of mixtures cannot be specified apriori in a principal way.



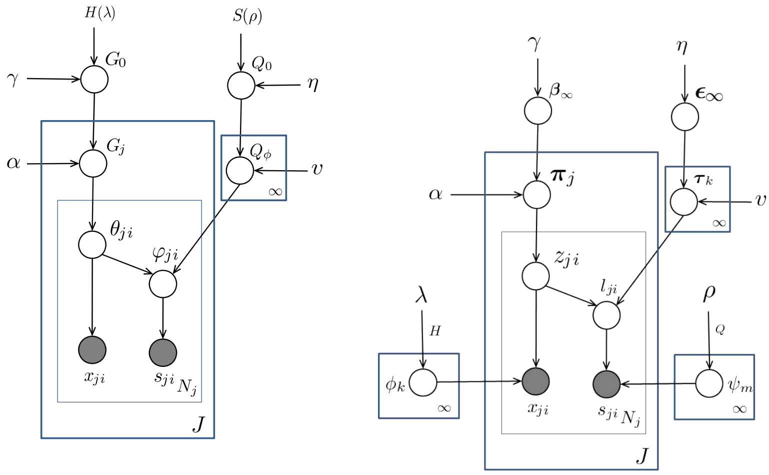
(a) generative view (b) stick-breaking view

Figure 1: Graphical model representation for HDP [28].

The Dirichlet process can also be utilized as nonparametric prior for modelling of grouped data. Under this setting, each group is modelled as a Dirichlet process mixture model and these models are 'linked' together to reflect the dependency among them. The goal is to exploit the mutual statistical strength across groups, and at the same time provide the clustering flexibility at the group level - a formalism which is generally known as dependent Dirichlet process [13]. One particular attractive formalism is the hierarchical Dirichlet processes [28, 27] which posits the dependency among the group-level DPM by another Dirichlet process (Figure 1). Specifically, let J be the number of groups and $\{x_{j1}, \ldots, x_{jN_j}\}$ be N_j observations associated with the group j which are assumed to be exchangeable within the group. Under HDP framework, each group j is endowed with a random group-specific mixture distribution G_j which is statistically connected with other mixture distributions via another Dirichlet process sharing the same base probability measure G_0:

$$G_j \mid \alpha, G_0 \overset{\text{iid}}{\sim} \text{DP}\left(\alpha, G_0\right), j = 1, \ldots, J \tag{4}$$

This generative process further suggests that G_j (s) are exchangeable at the group level and conditionally independent given the base measure G_0 , which is also a random probability measure distributed



(a) generative view

(b) stick-breaking view

Figure 2: Graphical representation for the proposed Conditionally Dependent Dirichlet Processes model.

according to another Dirichlet process

$$G_0 \mid \gamma, H \sim \mathrm{DP}(\gamma, H) \tag{5}$$

It is clear from the definition of the hierarchical Dirichlet process that G_j's ,G_0 and H share the same support Θ.

3 Conditionally Dependent Dirichlet Processes

3.1 Model Description

We start the model description with a basic setting as in the HDP whose modeling goal is to provide random mixture distributions over groups of data x_{ji}'s as described before (cf. Section 2.1). Recall that HDP models each group j with a random mixture distribution G_j, which are linked together via a shared DP-distributed random measure G_0 with the support provided by global atoms ϕ_k $(k = 1, 2, \ldots)$ drawn iid from the base measure H. These global atoms ϕ_k's are also commonly known as topics. In this paper, we refer to observations x_{ji}'s as *content* observations and ϕ_k's as *topic* atoms.

Now consider the case where each content observation x_{ji} is further augmented with a corresponding *context* observation s_{ji}, providing extra information about x_{ji}. Different from existing work [19, 20, 15, 30], we consider context as a distribution over some index space and we wish to model both x_{ji} and s_{ji} jointly. To link context distributions with topic distributions we utilize the nonparametric and hierarchical modeling: *a distribution over context can be viewed as random, conditionally on the topic.* Specifically, for each topic atom ϕ_k there corresponds a conditional distribution Q_{ϕ_k} to generative its associated contexts. These random context mixture distributions are hierarchically linked in a manner similar to [28].

Since we impose an HDP over the content observations, each x_{ji} is generated from a local factor $\theta_{ji} \in \Theta$ denoting the parameter for its corresponding mixture component. Since each cluster has several observations, it is convenient to separate the set of local factors $\{\theta_{ji} : \forall j, i\}$ into groups of distinct values in which members of the same group has the same value. Assume there are such K distinct values represented by $\phi_1, \ldots, \phi_K$, which are our global topic atoms.

To specify the context, let $(\Omega, \mathcal{L})$ be another measurable space to provide support for context observation s_i(s). Furthermore, let L be a fixed probability base measure on $(\Omega, \mathcal{L})$ and $\eta > 0$ be a context concentration parameter. Corresponding to each topic atom $\phi_k \in \Theta$, we impose a conditional random context mixture distribution Q_{ϕ_k} to explain the context observations s_{ji}'s associated with the set of content observations associated with ϕ_k. These mixture distributions Q_{ϕ_k}'s are further connected together via another Dirichlet process sharing a common base measure Q_0:

$$Q_{\phi_k} \mid v, Q_0 \overset{\text{iid}}{\sim} \text{DP}\left(v, Q_0\right)$$

To achieve an effect of hierarchical sharing as in the HDP, Q_0 is also a random probability measure distributed according another Dirichlet process with the base measure L and concentration parameter η:

$$Q_0 \mid \eta, S \sim \text{DP}\left(\eta, S\right) \tag{6}$$

Gathering the specification so far gives us the Topic-dependent Context Models whose graphical model representation is given in Figure 2a:

$$\begin{aligned}
G_0 \mid \gamma, H &\sim \text{DP}\left(\gamma, H\right) & Q_0 \mid \eta, S &\sim \text{DP}\left(\eta, S\right) && (7)\\
G_j \mid \alpha, G_0 &\overset{\text{iid}}{\sim} \text{DP}\left(\alpha, G_0\right) & Q_{\phi_k} \mid v, Q_0 &\overset{\text{iid}}{\sim} \text{DP}\left(v, Q_0\right) && (8)\\
\theta_{ji} \mid G_j &\overset{\text{iid}}{\sim} G_j & \varphi_{ji} \mid \theta_{ji}, \{Q_{\phi_k}\} &\overset{\text{iid}}{\sim} Q_{ji} \text{ where } Q_{ji} = Q_{\theta_{ji}} && (9)\\
x_{ji} \mid \theta_{ji} &\sim F\left(\cdot \mid \theta_{ji}\right) & s_{ji} \mid \varphi_{ji} &\sim Y\left(\cdot \mid \varphi_{ji}\right) && (10)
\end{aligned}$$

Note that, in Eq (9) the local topic atom θ_{ji} (left) has been used as the index in random context mixture distribution $Q_{\theta_{ji}}$, abbreviated by Q_{ji}, to generate the local context atom φ_{ji} which is then used to generate the context observation s_{ji}.

3.2 Stick-breaking representation

We present a stick-breaking representation for the proposed model, starting with the definitions of DP-distributed probability measures G_0, Q_0 and Q_ϕ's given in Eq (7) and (8). Since $G_0 \sim \text{DP}\left(\gamma, H\right)$ and $Q_0 \sim \text{DP}\left(\eta, S\right)$ are DP-distributed, they admit the following stick-breaking representations [24]:

$$
\begin{aligned}
\phi_k &\overset{\text{iid}}{\sim} H, \quad k = 1, \ldots, \infty \\
\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty} &\sim \text{GEM}(\gamma) \\
G_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}
\end{aligned}
$$

Assuming that the base measure H is non-atomic, G_0 is distributed according to a Dirichlet process $G_0 \sim \text{DP}(\gamma, H)$, and since G_0 has the support at at global content atoms $\boldsymbol{\phi} = (\phi_k)_{k=1}^{\infty}$, G_j admits the following stick-breaking representation [28]:

$$
G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k} \quad \boldsymbol{\pi}_j = (\pi_{jk})_{k=1}^{\infty} \sim \text{DP}(\alpha, \boldsymbol{\beta}), \quad \phi_k \mid H \overset{\text{iid}}{\sim} H. \tag{11}
$$

In addition, for each atom ϕ_k a corresponding conditional distribution $Q_{\phi_k} \sim \text{DP}(v, Q_0)$ is drawn according the definition given in Eq 8. Since $Q_0 \sim \text{DP}(\eta, S)$, it admits a stick-breaking representation:

$$
\begin{aligned}
\lambda_m &\overset{\text{iid}}{\sim} S, \qquad m = 1, 2, \ldots \\
\boldsymbol{\epsilon} = (\epsilon_m)_{m=1}^{\infty} &\sim \text{GEM}(\eta) \\
Q_0 &= \sum_{m=1}^{\infty} \epsilon_m \delta_{\lambda_m}
\end{aligned}
$$

For each ϕ_k, the random mixture measure $Q_{\phi_k} \sim \text{DP}(v, Q_0)$, hence using the property of HDP [28], it admits the following form:

$$
\begin{aligned}
\boldsymbol{\tau}_k &\sim \text{DP}(v, \boldsymbol{\epsilon}) \\
Q_{\phi_k} &= \sum_{m=1}^{\infty} \tau_{km} \delta_{\lambda_m}
\end{aligned} \tag{12}
$$

Finally, the content observation and x_{ji} and context observation s_{ji} are generated respectively as follows:

$$
\begin{aligned}
\theta_{ji} \mid G_j &\sim G_j \\
x_{ji} \mid \theta_{ji} &\sim F(\cdot \mid \theta_{ji}) \\
\varphi_{ji} \mid \theta_{ji}, \{Q_{\phi_k}\} &\sim Q_{\theta_{ji}} \\
s_{ji} &\sim Y(\cdot \mid \varphi_{ji})
\end{aligned}
$$

4 Inference for Conditionally Dependent Dirichlet Processes

We detail model inference in this section. We use an auxiliary conditional approach, assuming conjugacy between F and H in the content distribution and between Y and S in the context distribution. Base on the stick-breaking representation shown in Figure 2b, we develop a MCMC posterior inference which will be a collapsed Gibbs by integrating out the global topic atoms ϕ_k (s) and context atoms λ_m (s). The state space to be sampled consists of $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_J, \boldsymbol{\beta}, \boldsymbol{l}_1, \ldots, \boldsymbol{l}_J, \boldsymbol{\epsilon}\}$. Hyperparameters $\{\gamma, \alpha, \eta, v\}$ are further endowed with Gamma distributions and will be resampled at each Gibbs round. Given a sample realization $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_J, \boldsymbol{\beta}, \boldsymbol{l}_1, \ldots, \boldsymbol{l}_J, \boldsymbol{\epsilon}\}$, following notations are used:

- $w_{k,m}$: count for context m observed within content k
- n_{jk}: count for the content k observed within document j
- $\boldsymbol{s}_k$: collection of context observations for topic k, i.e, $\boldsymbol{s}_k := \{s_{ji} : z_{ji} = k, \forall j, i\}$
- $\boldsymbol{s}_k^{-ji}(m) := \left\{s_{j'i'} : s_{j'i'} = m, z_{j'i'} = k, j' \neq j, i' \neq i\right\}$: collection of context observations for topic k and context m, excluding at position i in document j.

4.1 Sampling content variables

Sampling the topic indicator $\boldsymbol{z}$

Different from HDP, sampling z_{ji} needs to take into account the influence of the corresponding contexts.

$$p\left(z_{ji} = k \mid \boldsymbol{z}_{-ji}, \boldsymbol{l}, \boldsymbol{x}, \boldsymbol{s}\right) \propto \underbrace{p\left(z_{ji} = k \mid \boldsymbol{z}_{-ji}, \alpha, \boldsymbol{\beta}\right)}_{\text{CRP}} \underbrace{p\left(x_{ji} \mid z_{ji} = k, \boldsymbol{z}_{-ji}, \boldsymbol{x}_{-ji} H\right)}_{\text{content predictive likelihood}} \underbrace{p\left(l_{ji} \mid z_{ji} = k, \boldsymbol{l}_{-ji}, \boldsymbol{\epsilon}\right)}_{\text{context preditive likelihood}} \tag{13}$$

The first term can easily be recognized as a form of Chinese restaurant process. The second term is the predictive likelihood from the content observations under the content mixture component k. Specifically, let $f(\cdot \mid \phi)$ and $h(\cdot)$ be respectively the density function for $F(\phi)$ and H, the conjugacy between F and H allows us to integrate out the mixture component parameter ϕ_k , leaving us the conditional density of x_{ji} under the mixture component k given all the content data items except x_{ji}:

$$p\left(x_{ji} \mid z_{ji} = k, \boldsymbol{z}_{-ji}, \boldsymbol{x}_{-ji}\right) = \frac{\int_{\phi_k} f\left(x_{ji} \mid \phi_k\right) \prod\limits_{j' \neq j, i' \neq i, z_{j'i'} = k} f\left(x_{j'i'} \mid \phi_k\right) h\left(\phi_k\right) d\phi_k}{\int_{\phi_k} \prod\limits_{j' \neq j, i' \neq i, z_{j'i'} = k} f\left(x_{j'i'} \mid \phi_k\right) h\left(\phi_k\right) d\phi_k} := f_k^{-x_{ji}}\left(x_{ji}\right)$$

Finally, the last term is the contribution from the context observation. Since $l_{ji} \mid z_{ji} = k \sim \operatorname{Mult}\left(\boldsymbol{\tau}_k\right)$ where $\boldsymbol{\tau}_k \sim \operatorname{Dir}\left(v\epsilon_1, \ldots, v\epsilon_M, \epsilon_{\text{new}}\right)$, the Multinomial-Dirichlet conjugacy property allows us to compute the last term in Eq (13) as:

$$p\left(l_{ji}=m \mid z_{ji}=k, \boldsymbol{l}_{-ji}, \boldsymbol{\epsilon}\right)=\begin{cases}\frac{v\epsilon_m+w_{k,m}}{w_{k,\bullet}+v} & \text{if } k \text{ previousely used}\\ \frac{v\epsilon_m}{v}=\epsilon_m & \text{if } k=k_{\text{new}}\end{cases}$$

In summary, the conditional distribution to sample z_{ji} is given as:

$$p\left(z_{ji}=k \mid \boldsymbol{z}_{-ji}, \boldsymbol{l}, \boldsymbol{x}, \boldsymbol{s}\right) \propto \begin{cases}(n_k^{-ji}+\alpha\beta_k) f_k^{-x_{ji}}\left(x_{ji}\right) \frac{v\epsilon_m+w_{k,m}}{w_{k,\bullet}+v} & \text{if } k \text{ previousely used}\\ \alpha\beta_{\text{new}} f_{k_{\text{new}}}^{-x_{ji}}\left(x_{ji}\right)\epsilon_m & \text{if } k=k_{\text{new}}\end{cases}$$

where again we recall that the value of l_{ji} is denoted by m for readability.

Sampling stick weights $\boldsymbol{\beta}$

Sampling $\boldsymbol{\beta}$ is similar to the HDP in [28] and proceed as follows.

$$p\left(\boldsymbol{\beta} \mid \boldsymbol{z}, \gamma, \alpha\right) \propto p\left(\boldsymbol{\beta}, \boldsymbol{z} \mid \gamma, \alpha\right)=p\left(\boldsymbol{z} \mid \boldsymbol{\beta}, \alpha, \gamma\right) p\left(\boldsymbol{\beta} \mid \gamma\right)$$

Integrating out $\boldsymbol{\pi}_j$ using the conjugacy property of Multinomial-Dirichlet and recall that $\boldsymbol{\pi}_j \sim \text{Dir}\left(\alpha\beta_1,\ldots,\alpha\beta_K\right)$ and $\sum_{k=1}^{K}\beta_k=1$ the first term becomes

$$\begin{aligned}p\left(\boldsymbol{z} \mid \beta_{1:K}\right)&=\prod_{j=1}^{J}\left[p\left(\boldsymbol{z}_j \mid \beta_{1:K}\right)\right]=\prod_{j=1}^{J}\int_{\boldsymbol{\pi}_j} p\left(\boldsymbol{z}_j \mid \boldsymbol{\pi}_j\right) p\left(\boldsymbol{\pi}_j \mid \alpha\beta_1,\ldots,\alpha\beta_K\right) d\boldsymbol{\pi}_j\\ &=\prod_{j=1}^{J}\frac{\Gamma\left(\sum_k \alpha\beta_k\right)}{\Gamma\left(\sum_k \alpha\beta_k+N_j\right)}\prod_{k=1}^{K}\frac{\Gamma\left(\alpha\beta_k+n_{jk}\right)}{\Gamma\left(\alpha\beta_k\right)}=\prod_{j=1}^{J}\frac{\Gamma\left(\alpha\right)}{\Gamma\left(\alpha+N_j\right)}\prod_{k=1}^{K}\frac{\Gamma\left(\alpha\beta_k+n_{jk}\right)}{\Gamma\left(\alpha\beta_k\right)}\end{aligned}$$

For the second term, note that $\boldsymbol{\beta}=\left(\beta_1,\ldots,\beta_K,\beta_{\text{new}}\right) \sim \text{Dir}\left(\frac{\gamma}{L},\ldots,\frac{\gamma}{L},\frac{L-K}{L}\gamma\right)$, let $\gamma_r=\frac{\gamma}{L}$ and $\gamma_{\text{new}}=\frac{L-K}{L}\gamma$ then

$$p\left(\boldsymbol{\beta} \mid \gamma\right)=\frac{\Gamma\left(\overbrace{K\gamma_r+\gamma_{\text{new}}}^{\gamma}\right)}{\left[\Gamma\left(\gamma_r\right)\right]^K\Gamma\left(\gamma_{\text{new}}\right)}\left(\prod_{k=1}^{K}\beta_k^{\gamma_r-1}\right)\beta_{\text{new}}^{\gamma_{\text{new}}-1}$$

Put them together, we get:

$$p\left(\boldsymbol{\beta}, \boldsymbol{z} \mid \gamma, \alpha\right)=\underbrace{\frac{\Gamma\left(\gamma\right)}{\left[\Gamma\left(\gamma_r\right)\right]^K\Gamma\left(\gamma_{\text{new}}\right)}\beta_{\text{new}}^{\gamma_{\text{new}}-1}\prod_{j=1}^{J}\frac{\Gamma\left(\alpha\right)}{\Gamma\left(\alpha+N_j\right)}\prod_{k=1}^{K}\beta_k^{\gamma_r-1}}_{\text{(term)}}\frac{\Gamma\left(\alpha\beta_k+n_{jk}\right)}{\Gamma\left(\alpha\beta_k\right)}$$

Using the results from [28], let $\boldsymbol{m} = (m_{jk} : \text{ for all } j \text{ and } k)$ and $\text{Stirl}\,(n, k)$ is the Stirling number of the second kind, we have:

$$\frac{\Gamma\left(\alpha\beta_k + n_{jk}\right)}{\Gamma\left(\alpha\beta_k\right)} = \sum_{m_{jk}=0}^{n_{jk}} \text{Stirl}\left(n_{ij}, m_{jk}\right)\left(\alpha\beta_k\right)^{m_{jk}}$$

$$p\left(\boldsymbol{\beta}, \boldsymbol{z} \mid \gamma, \alpha\right) = (\text{term}) \times \sum_{m_{jk}=0}^{n_{jk}} \text{Stirl}\left(n_{ij}, m_{jk}\right)\left(\alpha\beta_k\right)^{m_{jk}}$$

Dropping the summation over m_{jk}, it is easy to see that

$$\sum_{\boldsymbol{m}} (\text{term}) \times \text{Stirl}\left(n_{ij}, m_{jk}\right)\left(\alpha\beta_k\right)^{m_{jk}} = p\left(\boldsymbol{\beta}, \boldsymbol{z} \mid \gamma, \alpha\right)$$

This defines a joint distribution over $\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{m}$:

$$\begin{aligned} p\left(\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{m}\right) &= (\text{term}) \times \text{Stirl}\left(n_{ij}, m_{jk}\right)\left(\alpha\beta_k\right)^{m_{jk}} \\ &= \frac{\Gamma\left(\gamma\right)}{\left[\Gamma\left(\gamma_r\right)\right]^K \Gamma\left(\gamma_{\text{new}}\right)} \beta_{\text{new}}^{\gamma_{\text{new}}-1} \prod_{j=1}^{J} \frac{\Gamma\left(\alpha\right)}{\Gamma\left(\alpha + N_j\right)} \prod_{k=1}^{K} \beta_k^{\gamma_r - 1} \text{Stirl}\left(n_{ij}, m_{jk}\right)\left(\alpha\beta_k\right)^{m_{jk}} \end{aligned}$$

We sample $\boldsymbol{\beta}$ jointly with the auxilary variable $\boldsymbol{m}$:

$$\begin{aligned} p\left(m_{jk} = m \mid \boldsymbol{z}, \boldsymbol{m}_{-jk}, \boldsymbol{\beta}\right) &\propto \text{Stirl}\left(n_{ij}, m_{jk}\right)\left(\alpha\beta_k\right)^{m} \\ p\left(\boldsymbol{\beta} \mid \boldsymbol{m}, \boldsymbol{z}, \alpha, \gamma\right) &\propto \beta_{\text{new}}^{\gamma_{\text{new}}-1} \prod_{k=1}^{K} \beta_k^{\sum_j m_{jk} + \gamma_r - 1} \overset{\infty}{\cong} \beta_{\text{new}}^{\gamma-1} \prod_{k=1}^{K} \beta_k^{\sum_j m_{jk} - 1} (\text{as } L \to \infty) \end{aligned}$$

where we note in the last equation that $\gamma_{\text{new}} = \left(\frac{L-K}{L}\right)\gamma \to \gamma$ and $\gamma_r = \frac{K}{L} \to 0$ and $L \to \infty$.

4.2 Sampling context variables

Sampling context indicator $\boldsymbol{l}$

The key idea behind sampling l_{ji} is group those l_{ji} (s) indexed by the same content assignment z_{ji} together and then HDP inference as in [26] can be utilized. Given $\boldsymbol{z}$, let $\boldsymbol{s}_k$ be the set of all context observations indexed by the same k, i.e., $\boldsymbol{s}_k := \{s_{ji} : z_{ji} = k, \forall j, i\}$, in addition let $\boldsymbol{s}_k^{-ji}$ be the same set as $\boldsymbol{s}_k$ but *excluding* s_{ji}, i.e., $\boldsymbol{s}_k^{-ji} := \left\{s_{j'i'} : z_{j'i'} = k, j' \neq j, i' \neq i\right\}$, we can then write:

$$p\left(l_{ji} = m \mid \boldsymbol{l}_{-ji}, \boldsymbol{z}, \boldsymbol{s}, v, \boldsymbol{\epsilon}\right) \propto \underbrace{p\left(l_{ji} = m \mid z_{ji} = k, \boldsymbol{l}_{-ji}, \boldsymbol{\epsilon}\right)}_{\text{conditional CRP}} \underbrace{p\left(s_{ji} \mid l_{ji} = m, z_{ji} = k, \boldsymbol{s}_{-ji}\right)}_{\text{predictive likelihood}} \tag{14}$$

Since conditional on $z_{ji} = k$ the collection of context observations $\boldsymbol{s}_k = \{s_{ji} : z_{ji} = k, \forall j, i\}$ sharing the same k are modeled by a DPM distributed according the random mixture distribution $Q_k \sim \text{DP}\left(v, Q_0\right)$

whose stick-breaking is given the Eq (12), the first term in Eq (14) can be computed using the Chinese restaurant process, or equivalently its Polya-urn characterization, to give:

$$p\left(l_{ji}=m \mid z_{ji}=k, \boldsymbol{l}_{-ji}, \boldsymbol{\epsilon}\right) \propto \begin{cases} \left(w_{k,m}+v\epsilon_m\right) & \text{if } m \text{ previously used} \\ v\epsilon_{\text{new}} & \text{if } m=m_{\text{new}} \end{cases} \tag{15}$$

The second term in Eq (14) is recognized to be a form of predictive likelihood in a standard Bayesian setting whose likelihood function is Y, conjugate prior S and a set of observation $\boldsymbol{s}_k^{-ji}(m) := \left\{s_{j'i'} : s_{j'i'}=m, z_{j'i'}=k, j'\neq j, i'\neq i\right\}$. If the context likelihood distribution function $Y(\lambda)$ and its conjugate prior S have $a(\cdot \mid \lambda)$ and $b(\cdot)$ as their density functions respectively, then this term can be expressed as

$$p\left(s_{ji} \mid l_{ji}=m, z_{ji}=k, \boldsymbol{s}_{-ji}\right) = \frac{\int_{\lambda_m} a\left(s_{ji} \mid \lambda_m\right)\left[\prod_{s\in \boldsymbol{s}_k^{-ji}(m)} a\left(s \mid \lambda_m\right)\right] b\left(\lambda_m\right) d\lambda_m}{\int_{\lambda_m}\left[\prod_{s\in \boldsymbol{s}_k^{-ji}(m)} a\left(s \mid \lambda_m\right)\right] b\left(\lambda_m\right) d\lambda_m} := y_{k,m}^{-s_{ji}}\left(s_{ji}\right) \tag{16}$$

Substitute Eqs (16) and (15) into Eq (14) give us the final form to sample l_{ji}:

$$p\left(l_{ji}=m \mid \boldsymbol{l}_{-ji}, \boldsymbol{z}, \boldsymbol{s}, v, \boldsymbol{\epsilon}\right) \propto \begin{cases} \left(w_{k,m}+v\epsilon_m\right) y_{k,m}^{-s_{ji}}\left(s_{ji}\right) & \text{if } m \text{ previously used} \\ v\epsilon_{\text{new}} y_{k,m_{\text{new}}}^{-s_{ji}}\left(s_{ji}\right) & \text{if } m=m_{\text{new}} \end{cases}$$

Sampling stick weights $\boldsymbol{\epsilon}$

Different from HDP, sampling $\boldsymbol{\epsilon}$ requires more works as it is dependent on both $\boldsymbol{z}$ and $\boldsymbol{l}$. Let start with this factorization:

$$\begin{aligned} p\left(\boldsymbol{\epsilon} \mid \boldsymbol{l}, \boldsymbol{z}, v, \eta\right) &\propto p\left(\boldsymbol{\epsilon}, \boldsymbol{l} \mid \boldsymbol{z}, v, \eta\right) \\ &= p\left(\boldsymbol{l} \mid \boldsymbol{\epsilon}, \boldsymbol{z}, v, \eta\right) p\left(\boldsymbol{\epsilon} \mid v, \eta\right) \end{aligned} \tag{17}$$

Isolating those context variables l_{ji}^k generated by the same topic $z_{ji}=k$ into one group $\boldsymbol{l}_j^k := \left\{l_{ji} : 1\leq i\leq N_j, z_{ji}=k\right\}$, the first term of Eq (17) can be expanded as

$$\begin{aligned} p\left(\boldsymbol{l} \mid \boldsymbol{\epsilon}, \boldsymbol{z}, v, \eta\right) &= \prod_{j=1}^{J}\prod_{k=1}^{K}\int_{\boldsymbol{\tau}_k} p\left(\boldsymbol{l}_j^k \mid \boldsymbol{\tau}_k\right) p\left(\boldsymbol{\tau}_k \mid \boldsymbol{\epsilon}\right) d\boldsymbol{\tau}_k \\ &= \prod_{j=1}^{J}\prod_{k=1}^{K}\frac{\Gamma(v)}{\Gamma\left(v+u_{jk}^{\cdot}\right)}\prod_{t=1}^{M}\frac{\Gamma\left(v\epsilon_t+u_{jk}^t\right)}{\Gamma\left(v\epsilon_t\right)} \end{aligned}$$

where $u_{jk}^t := \left|\left\{l_{ji} \mid l_{ji}=t, z_{ij}=k, i=1,\ldots,N_j\right\}\right|$ is the count of seeing the content-context pair (k,t) in document k and $u_{jk}^{\cdot} := \sum_{t=1}^{M} h_{jk}^{\cdot}$, which is also the number of elements in $\boldsymbol{l}_j^k$.

Let $\eta_r = \frac{\eta}{R}$ and $\eta_{\text{new}} = \frac{R-M}{M}\eta$ and recall that $\boldsymbol{\epsilon} \sim \text{Dir}\left(\eta_r,\ldots,\eta_r,\eta_{\text{new}}\right)$ the second term of Eq (17) is a Dirichlet density:

$$p\left(\boldsymbol{\epsilon} \mid \eta\right) = \frac{\Gamma\left(\overbrace{M\eta_r + \eta_{\text{new}}}^{\eta}\right)}{\left[\Gamma\left(\eta_r\right)\right]^M \Gamma\left(\eta_{\text{new}}\right)} \left(\prod_{t=1}^{M} \epsilon_t^{\eta_r - 1}\right) \epsilon_{\text{new}}^{\eta_{\text{new}-1}}$$

Put them together and again use the result $\frac{\Gamma\left(v\epsilon_t + u^t_{jk}\right)}{\Gamma(v\epsilon_t)} = \sum_{h^t_{jk}=0}^{u^t_{jk}} \text{Stirl}\left(u^t_{jk}, h^t_{jk}\right)\left(v\epsilon_t\right)^{h^t_{jk}}$ we have:

$$p\left(\boldsymbol{\epsilon}, \boldsymbol{l} \mid \boldsymbol{z}, v, \eta\right) = \frac{\Gamma\left(\eta\right)}{\left[\Gamma\left(\eta_r\right)\right]^M \Gamma\left(\eta_{\text{new}}\right)} \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\Gamma\left(v\right)}{\Gamma\left(v + u^{\cdot}_{jk}\right)} \prod_{t=1}^{M} \frac{\Gamma\left(v\epsilon_t + u^t_{jk}\right)}{\Gamma\left(v\epsilon_t\right)} \epsilon_t^{\eta_r - 1} \epsilon_{\text{new}}^{\eta_{\text{new}-1}}$$

$$\frac{\Gamma\left(\eta\right)}{\left[\Gamma\left(\eta_r\right)\right]^M \Gamma\left(\eta_{\text{new}}\right)} \epsilon_{\text{new}}^{\eta_{\text{new}-1}} \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\Gamma\left(v\right)}{\Gamma\left(v + u^{\cdot}_{jk}\right)} \prod_{t=1}^{M} \epsilon_t^{\eta_r - 1} \sum_{h^t_{jk}=0}^{u^t_{jk}} \text{Stirl}\left(u^t_{jk}, h^t_{jk}\right)\left(v\epsilon_t\right)^{h^t_{jk}}$$

Now let $\boldsymbol{h} = \left(h^t_{jk} : \forall j, k, t\right)$ we arrive the following joint distribution

$$q\left(\boldsymbol{\epsilon}, \boldsymbol{l}, \boldsymbol{h}\right) \propto \epsilon_{\text{new}}^{\eta_{\text{new}-1}} \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\Gamma\left(v\right)}{\Gamma\left(v + u^{\cdot}_{jk}\right)} \prod_{t=1}^{M} \epsilon_t^{\eta_r - 1} \text{Stirl}\left(u^t_{jk}, h^t_{jk}\right)\left(v\epsilon_t\right)^{h^t_{jk}}$$

Therefore we sample $\boldsymbol{\epsilon}$ jointly with the auxiliary variable h^t_{jk} as follows

$$q\left(h^t_{jk} = h \mid \cdot\right) \propto \text{Stirl}\left(u^t_{jk}, h^t_{jk}\right)\left(v\epsilon_t\right)^{h^t_{jk}}, \quad h = 0, 1, \ldots, u^t_{jk}$$

$$q\left(\boldsymbol{\epsilon} \mid \cdot\right) \propto \epsilon_{\text{new}}^{\eta_{\text{new}-1}} \prod_{t}^{M} \epsilon_t^{\sum_j \sum_k h^t_{jk} + \eta_r - 1} \overset{\infty}{\cong} \epsilon_{\text{new}}^{\eta} \prod_{t}^{M} \epsilon_t^{\sum_j \sum_k h^t_{jk} - 1} (\text{as } R \to \infty)$$

4.3 Sampling hyperparameters

There are four hyper-parameters in our model: α, γ, v and η. Sampling α and γ is identical to HDP and therefore we refer to [28] for details.

Sampling v. The key idea here is to note that after $\boldsymbol{z}$ has been sampled, the number of active topics K plays a role of grouping contexts into K 'documents', we then utilize the results from HDP to sample v.

Let $M_k = \sum_{j,t} h^t_{jk}$ which is the number of active context atoms conditional topic atom k-th, and $M. = \sum_k M_k$, then:

$$p\left(M_1, \ldots, M_K \mid v, \cdot\right) = \prod_{k=1}^{K} \text{Stirl}\left(M., M_k\right) v^{M_k} \frac{\Gamma\left(v\right)}{\Gamma\left(v + M_k\right)}$$

Using the technique in [28], we write:

$$\frac{\Gamma(v)}{\Gamma(v+M_k)} = \int_0^1 w_k^v (1-w_k)^{M_k-1} \left(1+\frac{M_k}{v}\right) dw_k$$

Assuming $v \sim \text{Gamma}(a,b)$, define $\mathbf{w} = (w_k : \quad k=1,\ldots,K), w_k \in [0,1]$ and $\boldsymbol{o} = (o_k : \quad k=1,\ldots,K), s_k \in \{0,1\}$ we have

$$q(v,\mathbf{w},\boldsymbol{o}) \propto v^{a-1+\sum_k M_k} e^{-vb} \prod_{k=1}^{K} w_k^v (1-w_k)^{M_k-1} \left(\frac{M_k}{v}\right)^{o_k}$$

Therefore we sample v together the two auxiliary variables w_k and o_k as follows:

$$\begin{aligned} q(v \mid \cdot) &\propto v^{a-1+\sum_k (M_k - o_k)} e^{-v\left(b-\sum_k \log w_k\right)} = \text{Gamma}\left(a+\sum_k (M_k-o_k), b-\sum_k \log w_k\right) \\ q(w_k \mid v) &\propto w_k^v (1-w_k)^{M_k-1} = \text{Beta}(v+1, M_k) \\ q(o_k \mid \cdot) &\propto \left(\frac{M_k}{v}\right)^{o_k} = \text{Bernoulli}\left(\frac{M_k/v}{1+M_k/v}\right) \end{aligned}$$

Sampling η. Using similar strategy and using technique from Escobar and West [6], we write:

$$p(M \mid \eta, M.) = \text{Stirl}(M, M.)\, \eta^M \frac{\Gamma(\eta)}{\Gamma(\eta+M.)}$$

Let $\eta \sim \text{Gamma}(c,d)$ and for readability, we also replace $M.$ by U.

$$p(\eta \mid M, U) \propto p(M \mid \eta, U)\, p(\eta)$$

Recall that:

$$\frac{\Gamma(\eta)}{\Gamma(\eta+U)} = \int_0^1 w^\eta (1-w)^{U-1} \left(1+\frac{U}{\eta}\right) dw$$

Therefore,

$$\begin{aligned} p(\eta \mid w) &\propto \eta^{c-1+M} e^{-\eta d} w^\eta (1-w)^{U-1} \left(1+\frac{U}{\eta}\right) \\ &= \eta^{c-1+M} e^{-\eta(d-\log w)} (1-w)^{U-1} + \eta^{c-1+M-1} e^{-\eta(d-\log w)} (1-w)^{U-1} U \\ &\propto \eta^{c-1+M} e^{-\eta(d-\log w)} + U\eta^{c-1+M-1} e^{-\eta(d-\log w)} \\ &= \lambda \text{Gamma}(c+M, b-\log w) + (1-\lambda)\,\text{Gamma}(c+M-1, b-\log w) \end{aligned}$$

where λ satisfies this following equation to make the above expression a proper mixture density:

$$\frac{\lambda}{1-\lambda}=\frac{c+M-1}{M\left(b-\log w\right)}$$

Finally,

$$p\left(w\mid\eta\right)\propto w^{\eta}\left(1-w\right)^{U-1}=\mathrm{Beta}\left(\eta+1,U\right)$$

To re-sample η, we first sample $w\sim\mathrm{Beta}\left(\eta+1,U\right)$, compute λ as in the previous equation, and then use λ to select the correct Gamma distribution to sample η.

5 Experiments

We provide three experiments. Two with synthesis data to demonstrate the possible use of the proposed framework for topic modeling sensitive to time and trajectory discovery in surveillance data. The last experiment is an application of the model as a way to induce features for a visual category classification problem in computer vision.

5.1 Synthesis data: nonparametric Topic-Over-Time modeling

Topic-over-Time, introduced in [31], is a topic model whose topics are time-sensitive. It extends LDA [5] to deal with documents in which each word has an additional observed time information modelled by a Beta distribution customized for each topic. Our framework developed in this paper naturally be twisted to provide a nonparametric extension to this modeling providing two key advancements: a) the number of topics is unknown and to be inferred, b) the time distribution for each topic is a Dirichlet process mixture model over time axis, instead of a Beta distribution, thus it naturally fits better to many real-world problems where topic is allowed to rise and fall over time unbounded over time (whereas the Beta distribution is bounded between 0 and 1 and unimodal).

The fix is simple: we let the x_{ji} be the observed word and s_{ji} be the corresponding observed time information. We construct in this experiment a synthesis dataset to demonstrate this scenario. A similar set of simulated bar topics in [9] is used, but the topics are distributed according to different time mixture distributions of univariate Gaussian distributions. There are ten bar topics of size 5×5 as visualized in Figure 3 (top), grouped into either horizontal or vertical bars, resulting a vocabulary size of $V=25$. The time distributions are the uniform mixture of five Gaussians, being the context atoms, centered at even numbers from 2 to 10 with a fixed variance of 0.2. Corresponding to each of these centers, a set of $N=30$ documents are constructed whose topics are dictated by the groundtruth as shown in Figure 3 (middle) where H means only horizontal bar topics are used, V means only vertical bar topics and $H+V$ means a uniform mixture of both. Each document at location t is therefore

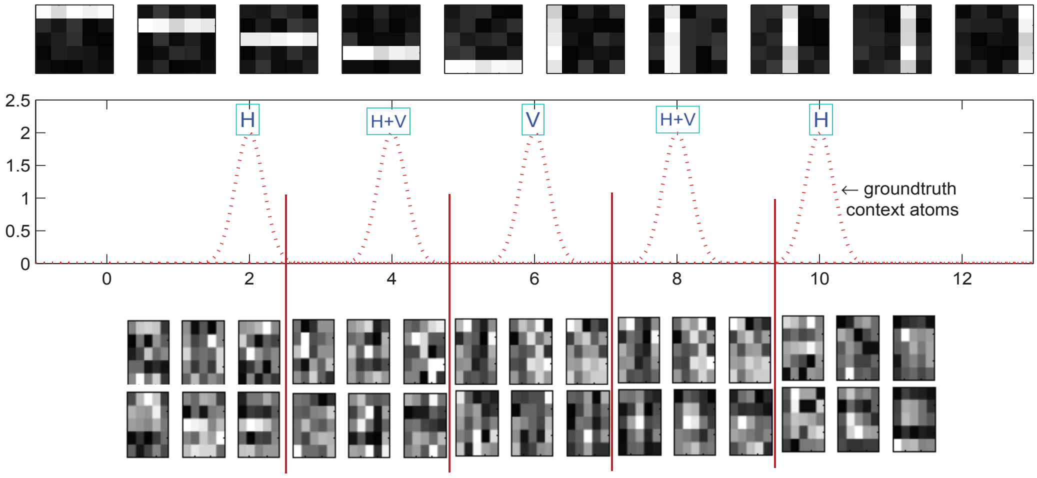


Figure 3: Groundtruth simulated data. *Top*: ten bar topics grouped into horizontal and vertical themes. *Middle*: Five Gaussians centered evenly from 2 to 10 used to generate the timestamps and words for documents: H means horizontal topics only, V vertical topics only, and $H+V$ means a uniform mixture of both. *Bottom*: examples of documents (visualized for words only) at different times.

consists of a set of $25 = 5 \times 5$ words drawn from the topics, and each word is further augmented with a timestamp drawn from the Gaussian with mean t and variance of 0.2 where $t = 2, 4, 6, 8, 10$.

A context-sensitive HDP is then fitted to the data whose topics are modelled as conjugate pair of Multinomial-Dirichlet, and for time information Gaussians with unknown means and variances are used together with their conjugate priors Gaussian-Gamma. The concentration parameters are resampled at each MCMC round where a Gamma prior are used: $\gamma, \eta \sim \text{Gamma}(4, 1)$, $\alpha, v \sim \text{Gamma}(3, 6)$. We collect 500 Gibbs samples after a burnin period of 100 samples.

The posterior distribution on the number of topic and context atoms (K, M) is shown in Figure 5 where the number of groundtruth atoms has been recovered correctly $K = 10$ for topics and $M = 5$ for five Gaussians (cf. Figure 3). Moreover, Figure 4 (middle) further shows that it recovers exactly 5 horizontal and 5 vertical bar topics as have been seen with the groundtruth where given a Gibbs sample $\{\alpha, \boldsymbol{\beta}, \boldsymbol{z}\}$ the recovered topics and mixture proportion are estimated as

$$\hat{\phi}_{k,v} = \frac{m_{k,v} + \lambda}{m_{k,\bullet} + \lambda V} \qquad \hat{\pi}_{jk} = \frac{n_{j,k} + \alpha\beta_k}{n_{j,\bullet} + \alpha}$$

We are also interested in inferring the conditional distribution of context given a topic; in our case each is a DP mixture of Gaussians. Given a realization of the Gibbs sample $\{v, \boldsymbol{l}, \boldsymbol{\epsilon}\}$ it is estimated as

$$p(t \mid k) = \sum_{m=1}^{M} \hat{\tau}_{k,m} \text{Normal}\left(t \left| \hat{\mu}_{k,m}, \hat{\sigma}^2_{k,m}\right.\right) \qquad \text{where} \qquad \hat{\tau}_{k,m} = \frac{[k, m] + v\epsilon_m}{[k, \bullet] + v}$$

and $\hat{\mu}_{k,m}, \hat{\sigma}^2_{k,m}$ are respectively the mean and variance from the collection of timestamps which are assigned to context m and topic k. Precisely they are computed from the following set of observation

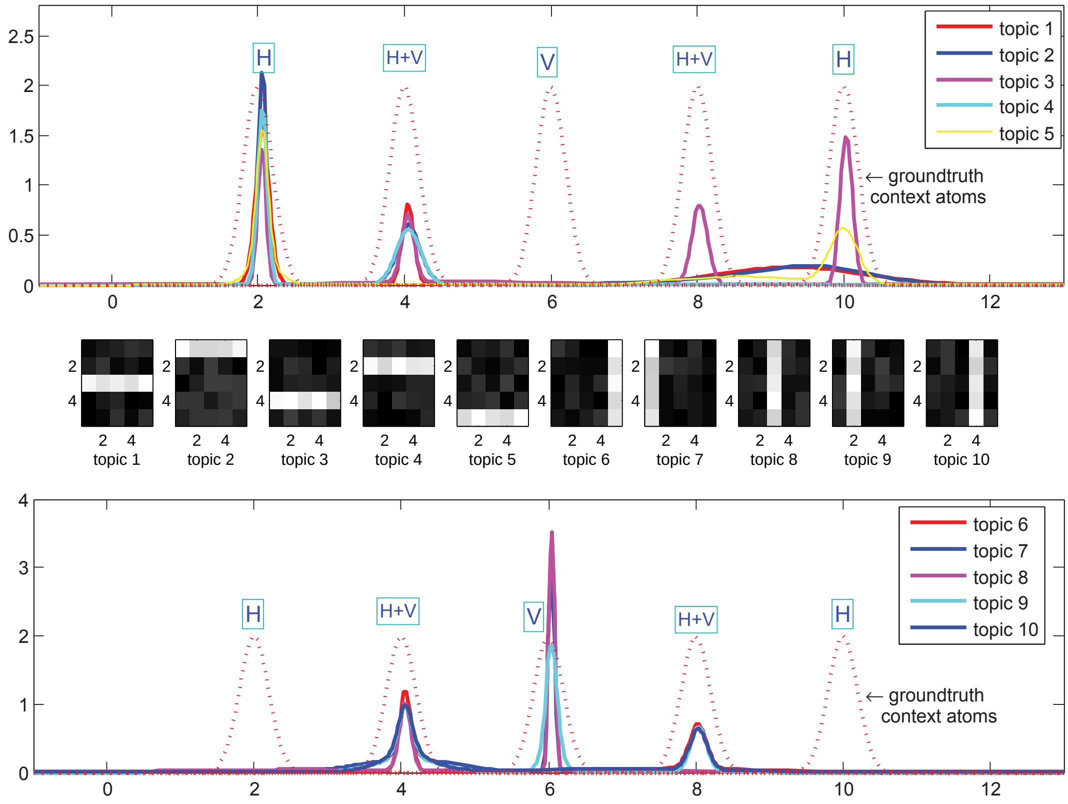


Figure 4: Results for bar topics over time experiment. *Middle*: learned topics which are able to recover exactly 10 topics grouped into horizontal and vertical bar topics. *Top*: learned conditional time distributions for learned horizontal topics which are able to recover correct periods of times they are active (note the ground where H mean horizontal topics are used). *Bottom*: learned conditional distributions for vertical topics, which are again able to recover correct times.

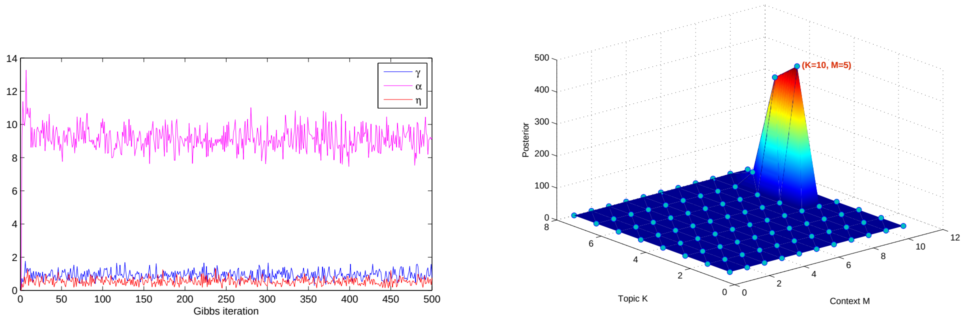


Figure 5: Concentration parameters being resampled (left) and posterior on the pair of estimated number of topics and context atoms (K, M) (right).

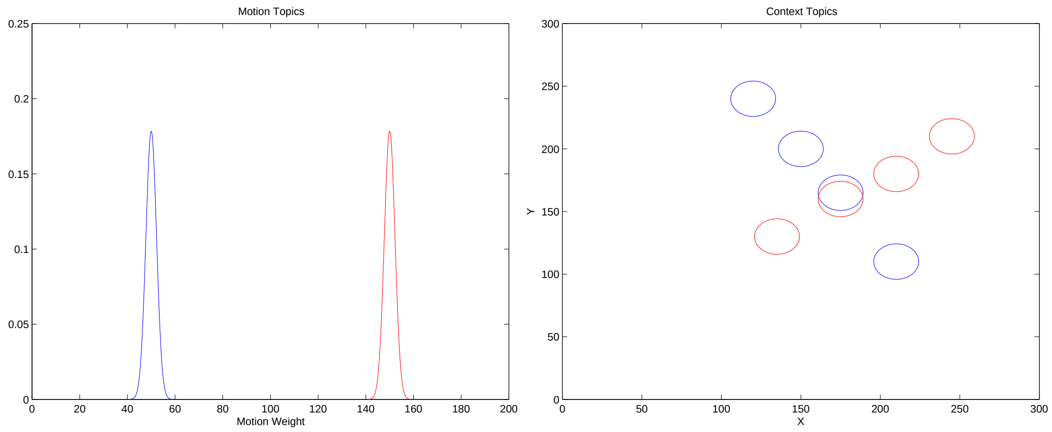


Figure 6: Left: Two motion topics: low magnitude $\mu = 50$ in red color and low magnitude $\mu = 150$ in blue color. Right: Four context atoms for low motion pattern are shown in red color and four context atoms for high motion patterns in blue noting the overlap of two context atoms in the middle which presents a challenging scenario.

$\{t_{ji} \mid l_{ji} = m, z_{ji} = k, \forall i, j\}$. Figure 4 illustrates the conditional time distributions learned for the group of horizontal bar topics (top) and vertical bar topics (bottom). We observe that, in all cases the model has correctly learned the time epochs at which the horizontal and vertical bar topics exist according to the groundtruth.

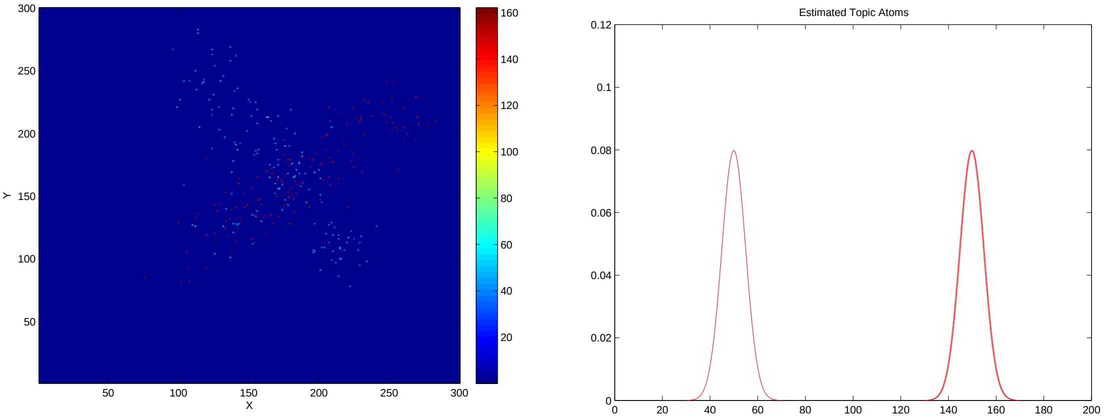
5.2 Synthesis Data: Detecting Movement Patterns in Video Surveillance

Scene understanding is a classic and challenging problem in computer vision. The work of [23] presents a state-of-the-art modeling approach in which the authors used a sequence of parametric mixture models at each time slice, which are then 'matched' through time formulated under a graph-cut formalism. We present in this section a toy example, mimicking the problem in[23] and demonstrate that our proposed model can address this problem elegantly, at least at from the modeling point of view.

Data is simulated from two motion patterns, each is characterized by a sequence of locations modelled by a sequence of Gaussian distributions on 2D plane. However, two motion patterns are different in the intensity of movements in the scene. For example, high motion pattern might correspond traffic movements whereas low motions corresponds to people walking. To distinguish motion level, two motion intensity distributions are introduced, each serves as the *topics* whereas spatial observations are *contexts*.

Fig 6 displays true motion intensity atoms and context atoms. True motion intensity atoms are univariate Gaussian distributions at $\mu = 50$ and $\mu = 150$ with fixed variance of 5 characterizing low motion and high motion respectively. Eight context atoms are 2D Gaussian distributions, four atoms for low topic motion and the rest for high topic motion. We simulate a set of $J = 30$ images, describing a sequence of video frames. Some examples are shown Figure 7a

We fit this data to the model where motion intensity is the content x_{ji} and its (x, y) spatial location



(a) Example of simulated images of the traffic scene with two motion patterns distinguished by their intensities

(b) Two topics are learned from the data. Each represents a motion intensity level, resembling identically to the groundtruth.

Figure 7: Examples of data simulated and the model recovers exactly two topics, describing low and high motion intensity.

is the context s_{ji}. The data is relatively small and thus we learn exactly two topic atoms ϕ_k as shown in Figure 7b, resembling almost identical the groundtruth. Conditional on each topic (low and high), the distribution on the context represents the motion patterns are shown in Figure 8. Again, the model estimate these motion patterns almost identical to the groundtruth, suggesting that the proposed framework might be effective to tackling the problem of scene understanding in real-world computer vision problem.

5.3 Real-world Data Experiment: Visual Category Classification

In this section, we present an application of the proposed framework as a method for multimodal dimensionality reduction and use the reduced dimension vector, being the mixture proportion π_j as a form of feature. We use the benchmark dataset in [17] which consists of 8 visual categories collected from LabelMe dataset [22]. These include: *tall buildings*, *inside city*, *street*, *highway*, *coast*, *open country mountain* and *forest*. Each image in LabelMe is annotated. User draws one or more regions on the image and annotate each region with a label. Annotation examples are shown in Figure 9.

In our model, each image j is then treated as a document, the label of the region is used as content x_{ji} and the visual feature of that region is the context s_{ji}. In this case, we extract GIST feature [17] from each region and use in place of s_{ji}. GIST is a visual descriptor to represent perceptual dimensions and oriented spatial structures of a scene. Each GIST descriptor is a 512-dimensional vector. Figure 10 shows some examples of GIST features. We further use PCA to project GIST features into 30 dimensions, thus s_{ji} is modelled according to a 30-dim Gaussian distribution whose conjugate prior is a Gaussian-Wishart distribution for both mean and covariances.

The inference setting for sampling hyper parameter are $\gamma \sim \text{Gamma}\,(6,5)$, $\eta \sim \text{Gamma}\,(3000,100)$, $\alpha, v \sim \text{Gamma}\,(3.5,1)$. We run collapsed Gibbs MCMC for 1000 iterations and 100 burnins which

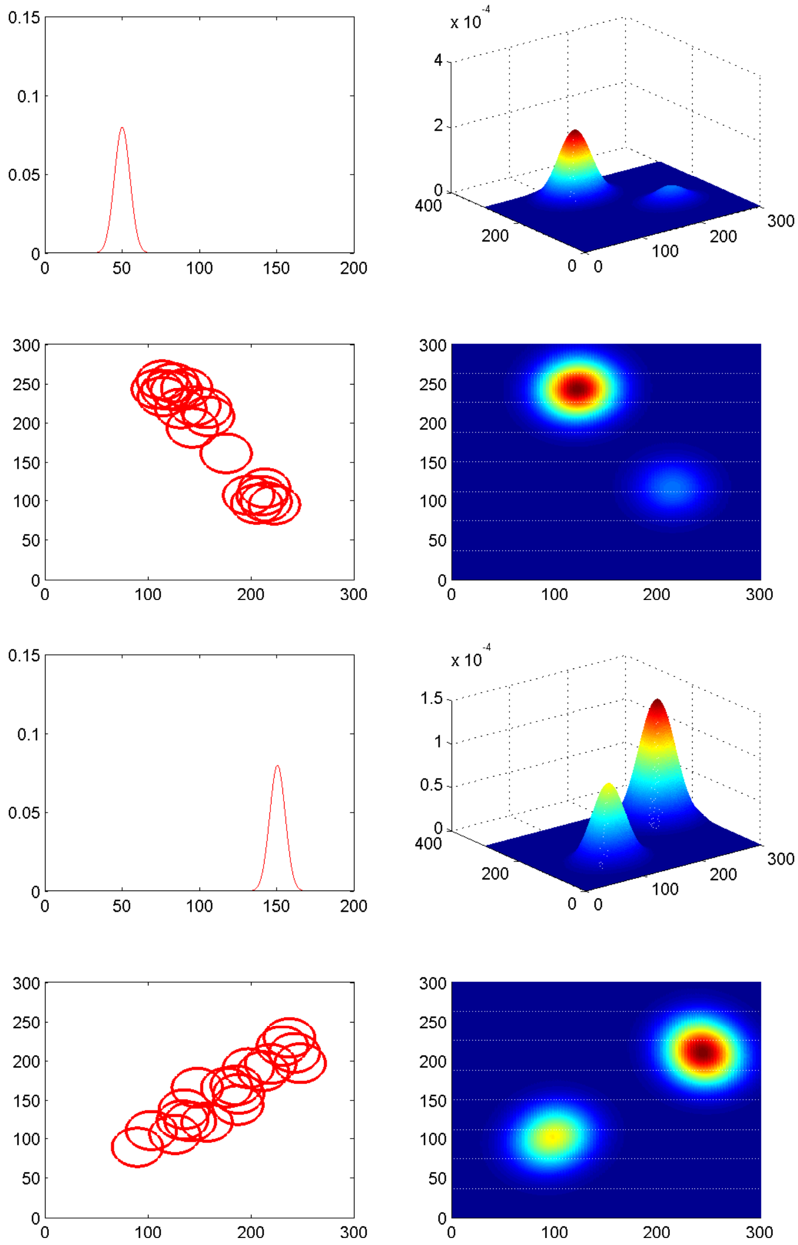


Figure 8: Learned topic atoms ϕ_k representing motion intensity and their corresponding conditional context atoms ψ_{km} and mixture distributions, each represents a motion or movement pattern.

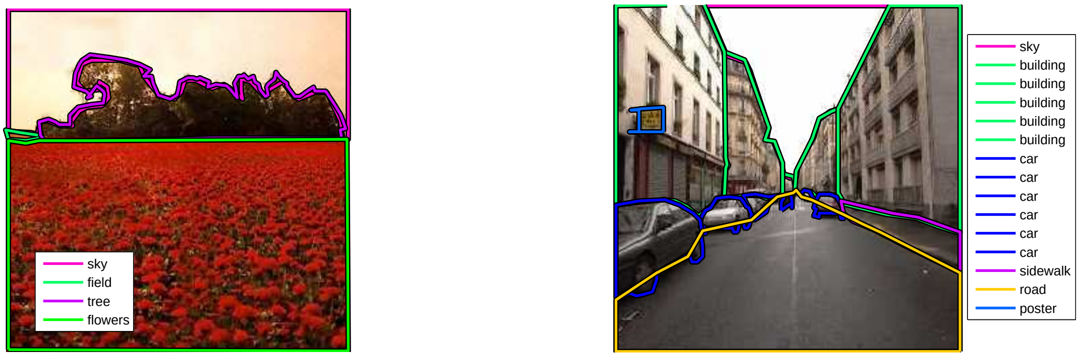


Figure 9: Examples of images used in our dataset. Each image consists of different annotated regions.

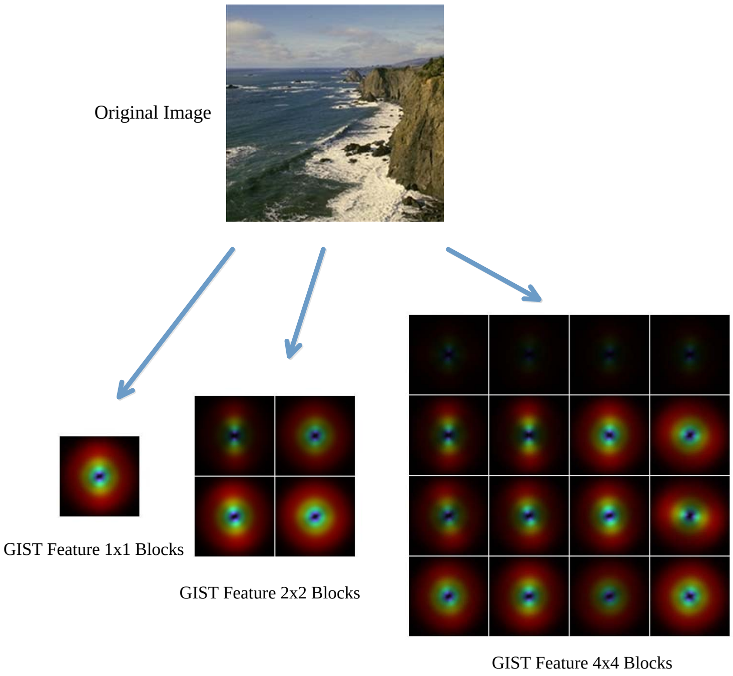


Figure 10: Visualization of GIST feature extraction [17].

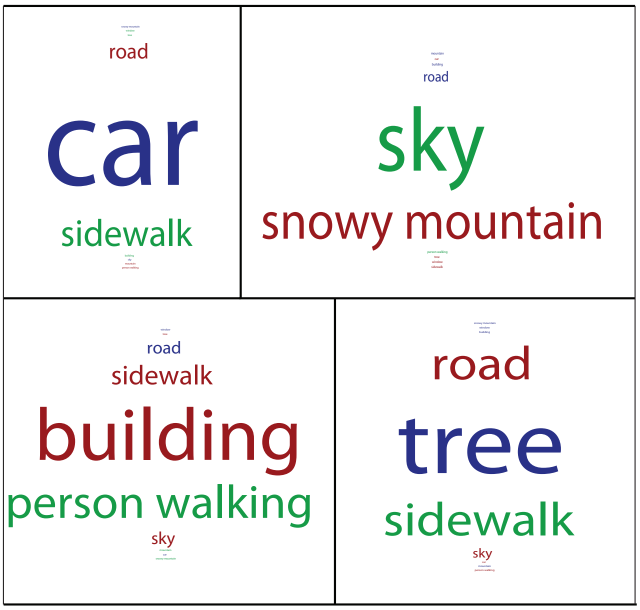


Figure 11: Top four topics learned from annotated labels.

Methods	Accuracy
Using HDP features on annotated labels	62.385%
Using our model on annotated labels and local GIST	77.125%
Context by visual region ancestry [12]	82%
Using GIST feature extract at image level [17]	83.7%
Using Gabor-PHOG feature at image level[25]	86.6%[2]
Dense SIFT + Texton HSMK at image level [11]	89.75%
Our mixture proportion feature + GIST at image level	**91.88%**

Table 1: Comparison of classification results for 8 visual categories. This dataset is originally introduced in [17].

takes roughly 5 mins for each loop using Matlab implementation. The model returned 8 topic atoms, each is a distribution over the set of annotated labels and 20 context atoms over GIST features, each is a multivariate Gaussian distribution of dimension 30. Fig 11 shows top four topic atoms learned.

To further demonstrate the strength of the model, we consider a classification task among these 8 visual categories. Our intuition is that, by further leveraging the GIST feature together with the annotated labels, the proposed model will help to 'regularize' and learn more discriminative topics (than not using together with visual information). We use the mixture component π_j in each image as an input feature for SVM classification with RBF kernel. In each class, we split randomly 100 images for training and 100 images for testing, the accuracy comparison on this dataset is recorded at their best performances after parameter selection step in validation set.

For comparison, we also employ Hierarchical Dirichlet Processes [28] on the annotated labels, then use the mixture components as features for SVM classification in a similar manner. In this experiment, we additionally consider a setting in which the mixture proportion is used together the existing GIST features (extracted at the image level) as a combined feature vector before input to SVM. We compare our method with state-of-the-art results reported in [17], [12], and [11] on the same dataset. Table 1 presents the classification results in which the use of model has improved over the most of baseline methods, including the use of HDP. This suggests that correlation among data channels carries additional discriminative information, which has been exploited in our modeling approach.

6 Extensions and Discussions

We provide discussions on some possible and attractive extensions arrived from the proposed models.

6.1 Modeling Multiple Contexts

When multiple contexts exist for a topic, the proposed model can easily be extended to accommodate this. That is for each content x_{ji} there is a set of M context $\left\{s_{ji}^m\right\}_{m=1}^M$ correlated to it. In the nutshell, conditional on the a topic ϕ, one can imposes multiple conditionally independent context distributions

$\left\{Q_{\phi}^{m}\right\}_{m=1}^{M}$ where M is the number of contexts. The generative process for contexts be modified as follows, starting with the sticking breaking $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$

- For each $m = 1, \ldots, M$ draw $Q_0^m \sim \text{DP}\left(\eta, S^m\right)$
- For each topic ϕ_k and for $m = 1, \ldots, M$ draw $Q_{\phi_k}^{M} \sim \text{DP}\left(\eta, Q_0^m\right)$
- For each group j and each word i within this group
  - Generate content $x_{ji} \sim F\left(\cdot \mid \theta_{ji}\right)$ where $\theta_{ji} \overset{\text{iid}}{\sim} G_j$
  - For each context channel $m = 1, \ldots, M$
    - Sample m-th context $s_{ji}^{m} \sim Y^{m}\left(\cdot \mid \varphi_{ji}^{m}\right)$ where $\varphi_{ji}^{m} \overset{\text{iid}}{\sim} Q_{\theta_{ji}}^{m}$

The inference presented in Section 4 can be readily extended to accommodate for this setting. More interesting, under a Gibbs sampling approach, it is clear that we can sample the latent indicator l_{ji}^{m} each context channel m in ***parallel*** once conditioning the content topic z_{ji}, thus the computation complexity in this case should remain the same as in the single context case given enough number of core processors to execute in parallel.

6.2 Modeling Hierarchically Nesting Contexts

In the setting of multiple contexts described in the extension in Section 6.1, instead of treating context conditionally independently given the topic, one might prefer to a setting in which the contexts are hierarchically nested $x_{ji} \rightarrow s_{ji}^{(1)} \rightarrow \ldots \rightarrow s_{ji}^{(M)}$. This nested structure leads to the nested structure in the latent space $z_{ji} \rightarrow l_{ji}^{(1)} \rightarrow \ldots \rightarrow l_{ji}^{(M)}$.

Though it appears that Gibbs inference in Section 4 might be derived for this case, it is anticipated that the Gibbs sampling might be very slow due to nested structure which requires the samples from the upper level to perform sampling at the lower level.

6.3 Modelling Group-Level Context Observations

In some application, context data s_{ji} may not be available for each word x_{ji} within a group j, instead it exists at the group-level; i.e., for each document j there is group-specific context s_j. Exemplar scenarios include: timestamps attached to a document (not at word level as overly treated in [31]), GPS location recorded for a text message sent from a mobile phone, modelling patient complication progression when the patient's living location is known.

The twist is simple, by interchanging the role of context and content in our model and push the context outside the document plate, it is readily to derive a model to model group-specific context observation. The generative process can be twisted as follows where we note that H is now providing the support for context s_j and Q provides support for content x_{ji}:

- $G_0 \sim \mathrm{DP}(\gamma, H)$ and for each j: $G_j \sim \mathrm{DP}(\alpha, G_0)$
- $Q_0 \sim \mathrm{DP}(\eta, Q)$ and for each topic ϕ_k, from the stick-breaking representation $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$, generate $Q_{\phi_k} \sim \mathrm{DP}(v, Q_0)$
- For each group j:
  - generate context $s_j \sim F(\cdot \mid \theta_j)$ where $\theta_j \sim G_j$
  - for each word i within this document, generate content $x_{ji} \sim \varphi_{ji}$ where $\varphi_{ji} \overset{\text{iid}}{\sim} Q_{\theta_j}$

The inference scheme presented in Section 4 is readily to be applied to this case.

6.4 Modelling Contextual Factors

Another possible and interesting extension to the proposed framework is to consider other nonparametric stochastic processes for contexts. One attractive choice is to induce hierarchical Beta processes for the context instead of hierarchical DP. Specifically, for each topic ϕ_k we introduce a topic-specific distribution Q_{ϕ_k} which is now a Beta processes. The collection of random distributions $\{Q_\phi\}_{\phi \in \Theta}$ is now linked according to a hierarchical Beta processes introduced in the work of [29]. Conditional on the topic, sampling the hierarchical Beta processes appears to be tractable using the sampling scheme presented in [29]. Alternatively, a Restricted Hierarchical Beta process (R-HBP) in our previous work present another attractive model choice with more efficient slice sampling routine [10].

7 Conclusion

Bayesian nonparametric methods are attractive modelling choices for several problems in machine learning and data mining due to flexibility. However, addressing more realistic problems in machine learning and data mining requires a need to advance Bayesian nonparametric modeling, both in theory and computation, to accommodate richer types of data in a principled way. Most of existing work considers a single data observation type. This paper addresses more realistic multimodal data in which covariates are rich, and yet tend to have a natural correlation with one another. These setting arises in a wide array of practical applications across many domains; including, to name a few: *medical data mining* (e.g., patient profiling, modelling medical records, modeling early intervention data in children), *multimedia social media* (e.g., tags and their associated multimedia contents, network data, context-sensitive community detection, joint topic and sentiment analysis), *computer vision* (e.g., context-sensitive object recognition and patterns discovery in surveillance) and *pervasive computing* (e.g., context-aware applications on mobile devices, analysis of honest social signals). The presence of rich and naturally correlated covariates calls for the need to model their correlation with nonparametric models, without reverting to making parametric assumptions and we have proposed a flexible class of fully Bayesian nonparametric model to address these problems. We have derived an auxiliary conditional Gibbs sampling scheme and demonstrated the applicability of the models on three experiments. Finally, we have discussed various extensions that are immediately possible to carry on using the proposed framework.

References

[1] C.E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[2] D. Blackwell and J.B. MacQueen. Ferguson distributions via Pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.

[3] D. Blei, L. Carin, and D. Dunson. Probabilistic topic models. *IEEE Signal processing magazine*, pages 55–65, 2010.

[4] D.M. Blei and J.D. Lafferty. Dynamic topic models. In *Proc. Int. Conf. on Machine learning ICML'06*, pages 113–120. ACM New York, NY, USA, 2006.

[5] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] M.D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430), 1995.

[7] T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[8] A.E. Gelfand, A. Kottas, and S.N. MacEachern. Bayesian nonparametric spatial modeling with dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035, 2005.

[9] T.L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(90001):5228–5235, 2004.

[10] S. Gupta, D. Phung, and S. Venkatesh. A slice sampler for restricted hierarchical Beta process with applications to shared subspace learning. In *International Conference on Uncertainty in Artificial Intelligence (UAI)*, Catalina Island, USA, August 2012.

[11] T. Le, Y. Kang, A. Sugimoto, S. Tran, and T. Nguyen. Hierarchical spatial matching kernel for image categorization. *Image Analysis and Recognition*, pages 141–151, 2011.

[12] J.J. Lim, P. Arbeláez, C. Gu, and J. Malik. Context by region ancestry. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1978–1985. IEEE, 2009.

[13] S.N. MacEachern. Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pages 50–55, 1999.

[14] R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, pages 249–265, 2000.

[15] X.L. Nguyen. Inference of global clusters from locally distributed data. *Bayesian Analysis*, 5:817–846, 2010.

[16] Peter Müller Stephen G. Walker Nils Lid Hjort, Chris Holmes. *Bayesian nonparametrics.* Cambridge University Press, 2010.

[17] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[18] J. Pitman. Poisson–Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability and Computing*, 11(05):501–514, 2002.

[19] L. Ren, D. Dunson, S. Lindroth, and L. Carin. Dynamic nonparametric Bayesian models for analysis of music. *Journal of American Statistician Association (JASA)*, 2009.

[20] L. Ren, D.B. Dunson, and L. Carin. The dynamic hierarchical Dirichlet process. In *Proc. of Int. Conf. on Machine Learning (ICML)*, pages 824–831. ACM, 2008.

[21] A. Rodriguez, D.B. Dunson, and A.E. Gelfand. The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154, 2008.

[22] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1):157–173, 2008.

[23] Imran Saleemi, Lance Hartung, and Mubarak Shah. Scene understanding by statistical modeling of motion patterns. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 2010.

[24] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

[25] A. Sinha, S. Banerji, and C. Liu. Novel gabor-phog features for object and scene image classification. *Structural, Syntactic, and Statistical Pattern Recognition*, pages 584–592, 2012.

[26] Y.W. Teh. A hierarchical bayesian language model based on pitman-yor processes. In *Proc. of Int. Conf. on Computational Linguistics (ACL)*, pages 985–992. Association for Computational Linguistics, 2006.

[27] Y.W. Teh and M.I. Jordan. Hierarchical Bayesian nonparametric models with applications. In N. Hjort, C. Holmes, P. Müller, and S. Walker, editors, *Bayesian Nonparametrics: Principles and Practice*, page 158. Cambridge University Press, 2009.

[28] 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.

[29] R. Thibaux and M.I. Jordan. Hierarchical Beta processes and the Indian buffet process. In *Proc. of Int. Conf. on Artificial Intelligence and Statistics (AISTAT)*, volume 11, pages 564–571, 2007.

[30] X. Wang and E. Grimson. Spatial latent Dirichlet allocation. In *Advanes in Neural Information Processing Systems (NIPS)*, 2007.

[31] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proc. ACM SIGKDD Int. Conf. on Knowledge Ddiscovery and Data Mining*, pages 424–433, New York, NY, USA, 2006. ACM.