Department of Computer Science
University of Toronto
http://learning.cs.toronto.edu

6 King's College Rd, Toronto
M5S 3G4, Canada
fax: +1 416 978 1455

March 25, 2005

UTML TR 2005–003

Time-Varying Topic Models using Dependent Dirichlet Processes

Nathan Srebro **Sam Roweis**
Department of Computer Science, University of Toronto

Abstract

We lay the ground for extending Dirichlet Processes based clustering and factor models to explicitly include variability as a function of time (or other known covariates) by integrating a Dependent Dirichlet Processes into existing hierarchical topic models.

Time-Varying Topic Models using Dependent Dirichlet Processes

Nathan Srebro **Sam Roweis**
Dept. of Computer Science, University of Toronto, Canada
{nati,roweis}@cs.toronto.edu

Abstract

We lay the ground for extending Dirichlet Processes based clustering and factor models to explicitly include variability as a function of time (or other known covariates) by integrating a Dependent Dirichlet Processes into existing hierarchical topic models.

1 INTRODUCTION

A standard approach for modeling a corpus of documents is to identify a pool of "topics" (word distributions) such that the distribution of vocabulary words used in each document follows one of the topics (in a clustering model [AMT04]) or a mixture of the topics (in a factor model such as models derived from the Aspect model [Hof01]).

We would like to model systematic changes in topic usage over time, assuming we have a corpus in which each document is associated with a time-stamp. To do so, we start with models where the topic pool is modeled as a Dirichlet Process over all word distributions. We replace this Dirichlet Process with a time-varying Dependent Dirichlet Process. That is, we have a different pool of topics at every time, such that the topic pools at all times are marginally identically distributed, yet are correlated with topic pools at nearby times.

Two basic mechanisms exist for modeling topic drift: (1) adding topics, removing topics, and changing the proportions of topics in the pool over time; (2) continuously drifting the word distributions of the topics in the pool.

In this paper we focus on the first mechanism, which leads us to dependent Dirichlet processes with varying "weights" rather then varying "locations" as we describe below. However, the ideas we develop are applicable to both, as well as to modeling topic changes within a single document and to modeling variability with respect to more complex (e.g. higher dimensional) covariates.

This paper reports our initial investigations into the problem of introducing time dependence into topic models. We present the setup for our problem, discuss in detail various modeling considerations, and analyze the appropriateness of various models. In Section 2, we review the infinite topic clustering and factor models which we use as a basis for our further constructions. In Section 3 we introduce the idea of time dependence (or dependence on some other covariate) and discuss how it can be integrated into these topic models and what properties we would like such an extension to have. We also introduce the Dependent Dirichlet Process [Mac00] as a key modeling element. In Section 4 we present several different Dependent Dirichlet Process models, both reviewing models recently suggested in the statistics literature and introducing our own novel model. Finally, we have implemented two of the models for introducing time variability into clustering models, and we thre results of initial exploratory experiments in Section 5.

2 CLUSTERING AND FACTOR MODELS

We begin by describing non-time-varying topic models, which we use as a starting point for introducing time variability. We consider generative models for a corpus of n "documents", each document being a collection of "words". We view documents as "bags of words", disregarding their order inside a document. That is, the distribution of words in a document is exchangeable, and a document is fully characterized by its (usually very sparse) word-count vector $Y_i \in \mathbb{N}^d$, where d is the vocabulary size and $Y_{i,a}$ is the number of times word "a" appears in document i.

2.1 CLUSTERING MODELS

We first describe models in which each document is associated with a *single* topic. Here, a "topic" is a distribution over words, determining the word counts

for all documents associated with the topic. The topic of each document is chosen according to some corpus-wide distribution H over topics. That is, H is a distribution over word distributions and represents the "pool" of available topics by defining a prior probability for any given topic.

One approach to clustering is to constrain the capacity of the model by limiting the number of available topics (clusters) to some predetermined limit K. In such models, the distribution H is a discrete distribution placing nonzero mass on only k word distributions and can be written as:

$$H = \sum_{r=1}^{K} S_r \delta_{V_r} \tag{1}$$

where V_r are word distributions, S_r are non-negative weights, and δ_Θ represents a point-mass at Θ. It can be generated by, e.g. choosing the k topics V_r independently according to some prior distribution over word distributions, and then choosing the weights S_r, e.g. from a Dirichlet distribution.

Finite topics models as described above are limited in their modeling flexibility and require some way of selecting the number of topics K. Instead, in what follows, we focus on infinite models in which the capacity is controlled via a prior distribution over H that does not explicitly limit the number of possible topics, but still encourages "concentrated" topic pools (i.e. distributions H with more of the mass on fewer topics generally have a higher prior then those where the mass is spread over more topics). Specifically, we let H be a Dirichlet Process [Ant74] over word distributions.

Just as a Dirichlet distributed random vector $(S_1, \ldots, S_K)$ can be seen as a random distribution over a finite number (K) of elements (i.e. a random point in a finite-dimensional probability simplex), a Dirichlet Process (DP) distributed measure H can be seen as random distribution over a non-finite domain. In our case, H will be a random distribution over the domain of word distributions.

A Dirichlet Process (denoted $H \sim \mathrm{DP}(\alpha, \mu)$) is parameterized by a base probability measure μ and a concentration parameter $\alpha > 0$, such that for any event A in the domain, $H(A) \sim \mathrm{Beta}(\alpha\mu(A), \alpha\mu(\overline{A}))$ (H is a random distribution, and so $H(A)$, the probability H assigns to A, is a random variable). An important property of the Dirichlet Process is that even though the domain of H might be continuous (as in our case), H is (almost surely) a discrete distribution, concentrated on a countable number of point masses. That is, H assigns positive probability only to a discrete selection of topics.

In fact, the Dirichlet processes can also be derived as the limit, as $K \to \infty$, of (1) where $(S_1, \ldots, S_K)$ are Dirichlet distributed with uniform shape parameters $(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K})$ and $V_k \sim \mu$. Another very useful discrete characterization of the DP is the stick breaking construction [Set94]:

$$\begin{gathered} H = \sum_{r=1}^{\infty} S_r \delta_{V_r} \\ V_r \sim \mu \quad ; \quad B_r \sim \mathrm{Beta}(1, \alpha) \quad \text{i.i.d} \\ S_r = B_r \prod_{s=1}^{r-1} (1 - B_r) \end{gathered} \tag{2}$$

The topics V_r are chosen independently according to the measure μ, as before, and the weights are constructed from i.i.d. Beta distributed random variables, which are also independent of the chosen topics. Written in this form, the weights S_r are not identically distributed, but rather have decaying expectations with $\mathbf{E}[S_r] = \frac{1}{1+\alpha}(\frac{\alpha}{1+\alpha})^{r-1}$. This form highlights the fact that most mass is likely concentrated on only a few topics (the first few V_r), and that the parameter α controls the expected number of topics with significant mass. We will denote the distribution of the random infinite vector S, as in (2), by $S \sim \mathrm{Stick}(\alpha)$.

We are now ready to describe the basic "clustering" topic-model. As a base measure for the Dirichlet processes H describing the topic pool, we must supply a prior distribution over word distributions V_r. To ensure conjugacy, we will use a Dirichlet distribution over words. The model is thus parametrized by two concentration parameters: the concentration of topics α_S and the concentration of the prior over word distributions in each topic α_V. The generative model for the words w in the documents can thus be written as:

$$\begin{gathered} H \sim \mathrm{DP}(\alpha_S, \mathrm{Dir}(\frac{\alpha_V}{d}, \ldots, \frac{\alpha_V}{d})) \\ X_i | H \sim H \ \text{ independently for each document i} \\ w_{i,j} | X, H \sim X_i \quad \begin{array}{l}\text{independently for each} \\ \text{word j in document i}\end{array} \end{gathered} \tag{3}$$

Note that H is a (random) distribution over word distributions, and so $X_i \sim H$ is a (random) word distribution. The word *counts* Y_i are multinomially distributed given X_i, and we can use (3) as a generative model for word counts Y_i directly:

$$Y_i | X, H \sim \mathrm{Multinomial}(N_i, X_i) \ \text{ ind. for each i} \tag{4}$$

where N_i is the length of document i, which we assume to be known.

2.2 FACTOR MODELS

In a factor model, each document is composed from a mixture of topics with varying proportions. Document

i is characterized by a distribution U_i of topics, generated from the overall topic "pool" H. Each word in a document is then generated by a different topic, chosen according to the document's private topic proportions U_i.

A factor model allowing an unbounded number of topics can be constructed using a Hierarchical Dirichlet Process [TJBB04]. The document pool H is a Dirichlet Process as before. For each document, U_i is itself a Dirichlet Process, with base measure H and concentration parameter α_U. That is, all topics share the same set of countably infinite topics V_k from the discrete topic pool, but the proportions in which topics appear vary between documents (although topics with smaller weights S_r also tend to have lower proportions in each documents U_i). The model can be written as:

$$\begin{aligned} H &\sim \mathrm{DP}(\alpha_S, \mathrm{Dir}(\tfrac{\alpha_V}{d}, \ldots, \tfrac{\alpha_V}{d})) \\ U_i|H &\sim \mathrm{DP}(\alpha_U, H) \text{ ind. for each doc i} \\ c_{i,j}|U, H &\sim U_i \quad \text{ind. for each} \\ w_{i,j}|c, U, H &\sim c_{i,j} \text{ word j in doc i} \end{aligned} \tag{5}$$

where c_{ij} is the (latent) topic of word j in document i.

Using the stick breaking representation (2) of H, we can also write the distributions U_i as (re-weighted) sums of a common set of point-masses V_r:

$$U_i = \sum_{r=1}^{\infty} U_{i,r}\delta_{V_r}. \tag{6}$$

We can think of U as an $n \times \infty$ infinite matrix and V as a $\infty \times d$ infinite matrix, with $V_{r,a}$ the probability of word "a" under topic r and $U_{i,r}$ the weight of topic r in document i. A row X_i of the (finite size) matrix product $X = UV$ specifies the generative word distribution of document i (given H and U), and we can again write a generative model for Y directly:

$$Y_i|U, H \sim \mathrm{Multinomial}(N_i, X_i) \tag{7}$$

The Hierarchical Dirichlet Process factor model can be thought of as two-stage generative process: a distribution over infinite dimensional matrix factorizations $X = UV$, and a distribution over observations Y given the matrix X, as specified by (7). Furthermore, the distribution over matrix factorizations $X = UV$ factorizes into independent distributions over U and V, $P(UV) = P(U)P(V)$, specified by[1]:

$$(V_{r,1}, \ldots, V_{r,d}) \sim \mathrm{Dir}(\tfrac{\alpha_V}{d}, \ldots, \tfrac{\alpha_V}{d}) \text{ ind. for each } r$$

and:

$$\begin{aligned} S &\sim \mathrm{Stick}(\alpha_S) \\ (U_{i,1}, U_{i,2}, \ldots)|S &\sim \mathrm{DP}(\alpha_U, S) \text{ ind. for each } i \end{aligned}$$

When $\alpha_U \to 0$, the entire mass of each U_i will be concentrated on a single topic, and the factor model (5) approaches a clustering model (3). In terms of the matrix factorization, each row of U is pushed to a sparsity extreme and is only allowed one positive entry.

3 COVARIATE TOPIC MODELS

In the topic models described above, the pool of available topics (the topic distribution H) is the same for all documents. Suppose each document in our corpus is associated with some covariate t_i, which for the purpose of this discussion we will refer to as "time". We would like our model to reflect the fact that the pool of available topics may exhibit correlation over time. To do so, we refer to $H(t)$: the distribution of available topics at time t. We would like to replace the dependence $X_i \sim H$ with

$$X_i|H \sim H(t_i) \tag{8}$$

in the clustering model (4) and replace the dependence $U_i \sim \mathrm{DP}(\alpha_U, H)$ with

$$U_i|H \sim \mathrm{DP}(\alpha_U, H(t_i)) \tag{9}$$

in the factor model (5). For each time t, $H(t)$ would still be a random distribution over word distributions (topics). For different times $t_1 \neq t_2$, the (random) topic distributions $H(t_1)$ and $H(t_2)$ would be dependent on each other, but not necessarily identical.

It should be stressed that $H(t_i)$ and U_i (in the factor model) are different and serve different purposes. U_i is the topic composition of a specific document, and although centered around $H(t_i)$, might deviate from it significantly. $H(t_i)$ is the topic distribution present in the corpus at the time relevant for document i, but is much less specific then U_i. As an extreme case, several documents with identical time stamps t might each be on very different topics, and $H(t)$ will represent all those topics, and probably other topics as well. The relationship between $H(t_i)$ and U_i is akin to the relationship between the hidden state and the observation in Hidden Markov Models (HMMs). In particular, the posterior over $H(t_i)$ takes into account documents before and after time t_i (note though, that $H(t)$ is not necessarily Markov—see below).

3.1 DEPENDENT DIRICHLET PROCESSES

Processes $H(t)$, such that for any t, $H(t)$ is marginally a Dirichlet Processes, are known as "Dependent

[1]We slightly overload the notation U: In the discussion above, U is a random distribution over word distributions, and is dependent on V, which are the locations of its mass. Here we refer to U as only the weights of the word distributions, and not the actual word distributions, which are still specified by V. These weights are independent of V.

Dirichlet Processes" (DDPs) [Mac00]. Although there is not necessarily a strong requirement that $H(t)$ be marginally a Dirichlet Process (it might be distributed according to some other distribution over distributions of topics), this choice does seem attractive to us. Our understanding of Dirichlet Process make understanding properties of Dependent Dirichlet Processes easier; furthermore, when used in mixture models such as those described in Section 2, Dirichlet Processes can also lead to computational advantages due to conjugacy.

Writing the Dependent Dirichlet Process $H(t)$ using the stick breaking construction (2), $H(t) = \sum_{r=0}^{\infty} S_r(t)\delta_{V_r(t)}$, two types of variations can be identified: the weights (S_r) can be varied with t and the point masses (word distributions) (V_r) can be varied. From a modeling perspective, varying the weights (S_r) captures topic appearances, disappearances and rise and fall in popularity. Varying (V_r) captures "topic drift", where a topic might persist over time, but its word composition changes in a continuous fashion.

Much of the literature on DDPs is concerned with models where the weights S_r are fixed and only the points (V_r) vary; often each point $V_r(t)$ is modeled as a process independent of the other points (e.g. [MQR04, IMRM04]). For example, in a DDP over R^d (i.e. where each $H(t)$ is a distribution over R^d, and so each V_r is a point in R^d), each $V_r(t)$ can be modeled as an independent multivariate Gaussian processes [MQR04]. The DDP can thus be viewed as a Dirichlet process over Gaussian processes. The variations are modeled exclusively by the (typically well established) model for each $V_r(t)$.

Taking such an approach in our case would require us to model $V_r(t)$ and since $V_r(t) \sim \mathrm{Dir}(\frac{\alpha_V}{d}, \ldots, \frac{\alpha_V}{d})$ is a random distribution (over a finite but large domain—the word vocabulary), this would leave us with the same problem of describing a time-varying distribution (i.e. a process over the probability simplex). We therefore choose to begin our study of time-varying topic models with models in which the weights (S_r) are varied while the topic word distributions (V_r) remained fixed. Another advantage of this line-of-attack is that a direct representation of the weights $S_r(t)$ at each relevant time point is much smaller and easier to handle then a representation of the word distributions $V_r(t)$ for each topic and each relevant time, had the word distributions also varied with time. We expect our understanding of how the weights (S_r) (a distribution over an infinite domain) can be varied, to assist us later when attempting to create models with varying topic word-distributions (V_r).

3.2 DESIRED PROPERTIES

Below, we outline several properties we would like our depedent process $H(t)$ to possess and discuss how they might be achieved.

3.2.1 Stationarity

We would like the process $H(t)$ to be stationary: $H(t)$ should be be marginally identically distributed, and furthermore, the process $H'(t) \doteq H(t + \triangle)$ should be distributed identically to $H(t)$. That is, we do not want to impose an a-priori bias on how the topic pool should evolve over time.

This requirement rules-out, for example, a chain topology hierarchy of Dirichlet processes, where (in discrete time), $H(1) \to H(2) \to \cdots$ is a Markov chain with $H(t+1)|H(t) \sim \mathrm{DP}(\alpha, H(t))$. In such a chain, $H(t)$ is likely to be concentrated on less support as t grows. This corresponds to an effective bias of having fewer topics as time goes on.

It might be argued that contraction or expansion of topic breadth are reasonable a-priori assumptions. However, when these assumptions are indeed appropriate, we would like to include them explicitly, and not as an artifact of a time variability mechanism.

3.2.2 Correlation Decay

We would like the correlation between $H(t_1)$ and $H(t_2)$ to decay monotonically as $|t_1 - t_2|$ grows, and for $H(t_1)$ and $H(t_2)$ to approach independence as $|t_1 - t_2| \to \infty$. Since $H(t)$ is a random distribution, and not a random scalar variable, the notion of "correlation" requires further clarification. We can require that for any event A (any subset of topics), the correlation between $H(t_1)(A)$ (the probability of event A under $H(t_1)$ and $H(t_2)(A)$ monotonically decreases from one to zero as $|t_1 - t_2|$ goes from zero to infinity.

The Hierarchical Dirichlet Process U_i as in model (5) is a dependent collection of random distributions that does not poses this property: all U_is are (a-priori) dependent in the same way. Although this type of relationship is suitable for modeling "a bag of documents" with no a-prior known structure, it is not suitable for covariates with a known order or metric, such as time.

3.2.3 Other properties

We will consider models which are both Markov in time, and not necessarily Markov. By "Markov" we mean that $\{H(t')\}_{t'>t}$ is independent of $\{H(t')\}_{t'<t}$ conditioned on $H(t)$. If $H(t)$ is Markov, then, when sorted by time, X_i in the clustering model (4) and U_i in the factor model, form a hidden Markov chain,

where $H(t_i)$ is the "hidden" state. Of course, since we only observe the counts Y_i, the "outputs" X_i or U_i of the HMM are also "hidden" to us.

We are generally interested in models with a continuous covariate, and that can be naturally extended also to higher dimensional covariates. All models considered below are also time reversible: $H'(t) \doteq H(-t)$ is distributed identically to $H(t)$; this may or may not be desirable from a modeling perspective.

4 SPECIFIC MODELS

4.1 TRANSFORMED GAUSSIAN PROCESS CONSTRUCTION

We describe here a process, approaching a Dependent Dirichlet Process at the limit, which uses underlying Gaussian processes.

Recall that a Dirichlet processes can be obtained at the limit, as $K \to \infty$ of (1), where $(S_1, \ldots, S_K) \sim \mathrm{Dir}(\frac{\alpha_S}{K}, \ldots, \frac{\alpha_S}{K})$. Furthermore, such a Dirichlet distribution can be described as a normalization $S_r = G_r / \sum_{r'} G_{r'}$ of i.i.d. Gamma random variables $G_1, \ldots, G_K$ with shape parameter $\frac{\alpha_S}{K}$. In order to introduce time-variability, it is tempting to use Gamma increment processes $G_r(t)$ (i.e. a process in which $G_r(t + \triangle) - G_r(t) \sim \mathrm{Gamma}(\alpha\triangle, 1)$ independent of $G_r(t)$) and let $S_r(t) = G_r(t) / \sum_{r'} G_{r'}(t)$. Although this would indeed lead to a Dependent Dirichlet Process at the limit, such a process is not stationary. The shape of each Gamma random variable $G_r(t)$ increases with t, and so although each resulting $H(t) = \sum_r S_r(t)\delta_{V_r}$ would indeed be marginally Dirichlet, it would be marginally Dirichlet with a concentration parameter which increases with time.

To obtain a stationary distribution, we can instead use independent stationary autoregressive Gaussian processes $Z_r(t)$, and deterministically transforming them to obtain Gamma distributed random variables. This leads to a model summarized by:

For each r, $Z_r(t)$ is an independent G.P. with:

$$\begin{aligned} \mathrm{Cov}\left[Z_r t_1, Z_r t_2\right] &= e^{-\lambda|t_1 - t_2|} \\ G_r(t) &= \mathrm{GammaCDF}^{-1}_{\frac{\alpha_S}{K}}\left(\Phi(Z_r(t))\right) \\ S_r(t) &= \frac{G_r(t)}{\sum_{r'=1}^{K} G_r(t)} \\ V_r &\sim \mathrm{Dir}(\tfrac{\alpha_V}{d}, \ldots, \tfrac{\alpha_V}{d}) \text{ i.i.d., ind. of } Z \\ H(t) &= \sum_{r=1}^{\infty} S_r(t)\delta_{V_r} \end{aligned} \tag{10}$$

where $\mathrm{GammaCDF}^{-1}_{\alpha}$ is the inverse CDF of the Gamma distribution with shape α and scale 1, and Φ is the normal Gaussian CDF.

At any time t, $(S_1, \ldots, S_K) \sim \mathrm{Dir}(\frac{\alpha_S}{K}, \ldots, \frac{\alpha_S}{K})$, and so as $K \to \infty$, $H(t)$ approaches a Dirichlet processes, and H is a Dependent Dirichlet Process. $H(t)$ inherits its desirable properties from the underlying Gaussian processes. In particular, it is stationary, Markov, and its decorrelation is controlled by λ.

It is more common to transform Gaussian random variables to points on the simplex by other transformations, such a logit or probit transformations [Bor02]. Another approach which is somewhat similar to the one suggested here, but simpler computationally (as it does not involve the hard-to-compute inverse of the Gamma CDF), is to exponentiate the Gaussian, thus creating log-normal distributed variables, then then normalizing these. However, such transformations do not yield distributions S in the simplex which are Dirichlet distributed. A practical problem with this deficiency is that the behavior of the resulting distribution as $K \to \infty$ is not understood. It is not clear (to the best of our knowledge) how to change the parameters of the transformation so as to get a behavior which approaches a sensible limit, e.g. where the distribution of the number of topics with significant mass is controlled, or if this is at all possible. On the other hand, the Gamma transformation described above ensures a sensible, well understood, limit as $K \to \infty$, and when K is reasonably larger then the number of topics with significant mass, the distribution of the top topics is mostly independent of K.

The choice of the limit representation (1) is somewhat problematic computationally, compared to the stick-breaking representation, since K should be chosen to be significantly higher then the number of topics with significant mass. However, using a stick breaking representations and varying the beta random variables B_r would not achieve the desired effect, as the first few topics would tend to be strong for all times, and the process will not completely decorate and introduce new topics. A possible solution, discussed in the next section, is to vary the order in which the variables B_r are combined, rather then their value.

4.2 ORDER BASED DEPENDENT DIRICHLET PROCESSES

We describe here the Order-Based Dependent Dirichlet Process, recently suggested by Griffin and Steel [GS04], as applied to topic distributions. (Below we summarize their idea; see the references for further details. Griffin and Steel also describe a Markov Chain Monte Carlo procedure for the model, which we have implemented in the context of the clustering topic model (4) and discuss in Section 5.)

Recall the stick breaking description (2) of Dirich-

let Processes. In an Order-Based DDP, each topic is associated with a word distribution V_r and beta-distributed random variable B_r. Both V_r and B_r are fixed over time. What changes over time is the *order* in which the B_r variables are combined:

$$H(t) = \sum_{q=1}^{\infty} S_q(t)\delta_{V_{\pi_q(t)}}$$
$$S_q(t) = B_{\pi_q(t)} \prod_{q'=1}^{q-1} \left(1 - B_{\pi_{q'}(t)}\right) \quad (11)$$
$$V_r \sim \mathrm{Dir}(\tfrac{\alpha_V}{d}, \ldots, \tfrac{\alpha_V}{d}) o \text{ i.i.d.}$$
$$B_r \sim \mathrm{Beta}(1, \alpha_S) \text{ i.i.d.}$$

where $\pi(t)$ is a time-dependent infinite permutation.

First note that for any time t, (11) describes a stick-breaking construction, and hence marginally $H(t)$ is a Dirichlet processes with $H(t) \sim \mathrm{DP}(\alpha_S, \mathrm{Dir}(\frac{\alpha_V}{d}, \ldots, \frac{\alpha_V}{d}))$. All times share the same topics V_r, hence the DPs are clearly dependent. However, a topic r that appears "late" in the permutation $\pi(t)$ (i.e. for which q such that $\pi_q(t) = r$, is high) would have many $(1 - B_{\pi_{q'}(t)})$ terms multiplied into its weight $S_{\pi_q(t)}$. Topics therefore change their weight according to their place in the permutation $\pi(t)$.

To determine the permutation $\pi(t)$ for every time t, each topic r is associated with a time τ_r. The permutation π_t lists the topics in increasing order of absolute distance of τ_r from t: $\pi_1(t)$ is the topic closest in time to t, and so on (formally, $|t - \tau_\pi(q)| < |t - \tau_\pi(q')|$ for $q < q'$).

Rouhgly speaking, the most relevant topics at time t are the topics close in time to t. As we move through time, a topic becomes "stronger" as we approach it, then gradually weakens as we get further away from it. Although the actual magnitudes of B_r might mean that the closest topic in time is not the strongest, it would be extremely difficult for a document far away in time to have any meaningful contribution.

4.2.1 Distribution of Topics Through Time

To complete the description of the Order-Based DDP, we must specify how topics are distributed through time. The times at which topics appear follow a Poisson process with intensity λ, which is a parameter of the model. That is, the time between any two "consecutive" topics is exponentially distributed with mean $1/\lambda$. An infinite number of topics appear throughout the infinite time line, though only topics reasonably close to the observed document times would have any significant effect (truncation of the time line for practical computation is analogous to truncation of the stick-breaking representation).

The intensity parameter λ controls how quickly topics change. A higher value of λ yields more densely packed topics, and so the permutation $\pi(t)$ changes more quickly in time, and $H(t)$ deccorolates faster. The parameter λ does *not* effect the effective number of topics at any time: regardless of λ we have $H(t) \sim \mathrm{DP}(\alpha_S, \mathrm{Dir}(\frac{\alpha_V}{d}, \ldots, \frac{\alpha_V}{d}))$ and only α_S controls the marginal concentration of $H(t)$. In fact, α_S and λ together control the decorrelation of $H(t)$: a higher values of α_S makes more leading terms of the stick-breaking relevant, and so implies more topics need to change place in the permutation before the resulting distributions decorrelate. Griffin and Steel [GS04] provide an explicit expression of the correlation $\rho(H(t_1), H(t_2))$ in terms of λ and α_S.

4.2.2 Topic Persistence & Assymmetry

One problem of the Order-Based DDP as suggested by Griffin and Steel is that topics *must* change when time passes. A topic cannot persist while other topics change positions. A possible extension to the model allowing persistence is to associate with each topic r also a *volatility* ν_r, which are i.i.d. and independent of the Vs, Bs and τs. The permutation $\pi(t)$ is then determined according to the *weighted* distances $\nu_r|t - \tau_r|$. A topic with a volatility close to zero would persist in the top positions of the permutation, while more volatile topics change. It is also possible to make the procedure assymmetric (non-reversible) in time; Griffin and Steel describe some examples of this.

The Order-Based DDP fulfills our requirements of being stationary and having decorrelating with time. It is also time-reversible (but is not Markov) and it can be readily extended to covariates of higher dimensions.

4.3 DISCRETE-TIME DDP VIA LATENT MULTINOMIALS

Pitt *et al* [PCW02] suggest an approach which can be used to construct a discrete time stationary Markov chain which is marginally a Dirichlet Processes (i.e. a Dependent Dirichlet Processes over an ordered discrete covariate). In order to describe a stationary Markov chain $H(1) \to H(2) \to \cdots$, it is enough to describe the transition probability $H(t+1)|H(t)$. We specify a transition generatively as follws: first sample L samples according to $H(t)$, and then sample $H(t+1)$ according to the posterior distribution of the parameters of the sampling processes.

More formally, we will describe the Markov chain

$$H(1) \to M(1) \to H(2) \to M(2) \to H(3) \to \cdots$$

where each $M(t)$ is a (random) sample of topics (with repetitions). We will describe the Markov chain by

describing the joint distribution of $(H(t), M(t))$ and $(M(t), H(t+1))$, making sure the marginals over $H(t)$ and $M(t)$ agree. In fact, $(H(t), M(t))$ and $(H(t+1), M(t))$ will follow the same law (H, M): Marginally, $H \sim \mathrm{DP}(\alpha, \mu)$, as desired ($\mu$ is the base measure, i.e. $\mathrm{Dir}(\frac{\alpha_V}{d}, \ldots, \frac{\alpha_V}{d})$). Conditioned on H, each of the L samples in M is i.i.d. $\sim H$. This completes description of the Markov chain.

Representing M as a discrete integer valued measure over topics, describing the number of times a topic appears in the sample, we have $H|M \sim \mathrm{DP}(\alpha + L, M + \alpha\mu)$. We can now calculate:

$$\mathbf{E}[H(t+1)|H(t)] = \mathbf{E}[\mathbf{E}[H(t+1)|M(t)]|H(t)]$$
$$= \mathbf{E}\left[\left.\frac{M(t) + \alpha\mu}{L + \alpha}\right| H(t)\right] = \tfrac{L}{L+\alpha} H(t) + \tfrac{\alpha}{L+\alpha}\mu \quad (12)$$

The Markov chain is autoregressive in expectation, and the ratio $\frac{L}{L+\alpha}$ controls the decorrelation of the chain. In order for the chain to decorrelate slower, a larger number of samples must be sampled at each iteration, increasing the strength of the link between $H(t)$ and $H(t+1)$. As this description cannot readily be extended to a continuous covariate, it is interesting to study how the model should be changed when more densely sampled times are considered, i.e. if the same technique is used to construct transitions $H(t+\triangle)|H(t)$ for small increments $\triangle$. If we would like to maintain $\mathbf{E}[H(t+1)|H(t)] = \rho H(t) + (1-\rho)\mu$, the sample size for each $M(t)$ must be set to $L = \frac{\rho^{1/\triangle}}{1-\rho^{1/\triangle}}\alpha$.

4.4 WEIGHTED POLYA URN SAMPLING

Let $H \sim \mathrm{DP}(\alpha, \mu)$ and $X_1, \ldots, X_n|H \sim$ i.i.d. H. The posterior distribution of X_i conditioned on all other $X_{i'}$ (marginalized over H) is given by:

$$P(X_i = x|\{X_{i' \neq i}\}) \propto \alpha\mu(x) + \sum_{i', X_{i'} = x} 1 \quad (13)$$

This conditional distribution can be used to make Gibbs updates to X_i, marginalized over H. One might consider introducing time dependence by scaling the contribution of each i' in the update by the distance of $t_{i'}$ from t_i, e.g. considering marginals of the form:

$$P(X_i = x|\{X_{i' \neq i}\}) \propto \alpha\mu(x) + \sum_{i', X_{i'} = x} e^{-\lambda|t_i - t_{i'}|} \quad (14)$$

Unfortunately, *this set of marginals is not compatible with any joint distribution* over $X_1, \ldots, X_n$. In fact, even for three points, no set of non-uniform weights leads to a compatible joint distribution.

Nevertheless, one might attempt constructing a Markov chain over variable settings by performing Gibbs-like updates as suggested by (14), although this would lead to a Markov chain whose equilibrium distribution, if any, has an unclear interpretation. The corresponding updates are also much more computationally intensive, as all data items $X_{i'}$ need to be considered separately for each X_i update, leading to a run-time dependence of $O(n^2)$ for one update of all variables. This approach has actually been taken in the past in order to obtain a covariate-dependent gating network for a covariate-dependent infinite mixture of Gaussians [RG02].

5 EXPLORATORY EXPERIMENTS

We have implemented an MCMC posterior sampler for a clustering topic model with the Order Based DDP model we described. We compared this Dependent DP to a static (non-time-varying) DP model in which documents are considered to be unordered and identically distributed. Figure 1 shows a simple synthetic example. We generated 100 "documents" over a 20 word vocabulary; 80 of those documents have word counts drawn according to a different random distribution different for each document and 20 of them have word counts drawn from a distribution closely concentrated on a single "topic". The left panel of the figure shows the word counts as gray levels, where documents are arranged vertically (ordered in time) and words are arranged horizontally. The similar documents are nearby in time and occur near the top of the dataset (early in time). The middle panel shows (a typical posterior sample of) the clustering discovered by the fixed (static) model; it captures many but not all of the documents associated with the topic in one cluster but also clusters the remaining "noise" documents into three further clusters. The right panel shows (a typical posterior sample of) the clustering from our order based DDP which correctly identified exactly all the topic documents and put the rest into a single class.

We have also applied our model to a document collection in which each "document" is the title of a paper appearing the Proceedings of the National Academy of Sciences (USA). The full dataset contains 79801 paper titles from 1391 issues of the journal from 1915 to 2005. We used a vocabulary of 4752 words, which was obtained after suppressing common function words (stopwords) and words appearing in less than 10 titles. Each document has associated with it a time stamp corresponding to the publication date of the issue in which is appeared. We ran our order based DDP on a subset of 8567 cases formed by taking a constant fraction of papers from each issue (to avoid over-sampling from later years in which the volume of papers is much higher).

After generating several posterior samples of assignments of documents to classes, we can compute the marginal (summing over classes and posterior samples) probabilty of any vocabulary item occuring as a function of time. Figure 1 shows such probability curves for several terms over the period covered by the dataset. We can also take a previously unseen document and evaluate its marginal likelihood as a function of time; Figure 1 shows this for two articles. On a larger set of 10,000 previously unseen documents, the Order Based DDP gives comparable average log-likelihood to the static clustering model (-62.1 nats under the Order Based model versus -62.4 nats for the static model).

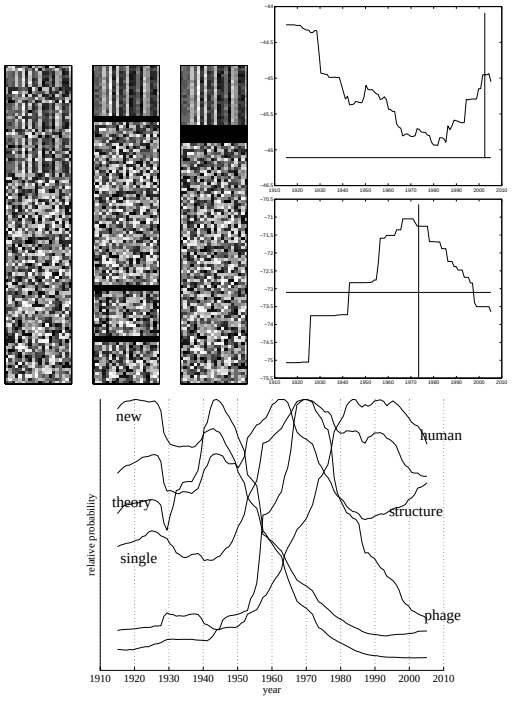

Figure 1: **Left:** A synthetic data experiment–rows are "documents", columns are "words" and intensities represent word counts. From left to right—the input data sorted by time, the clustering found by a fixed mode and the clustering found by the Order Based DDP based model. **Right:** For two held-out documents from the PNAS corpus, the log-likelihood of the documents under a non-time varying clustering model (horizontal line), the log-likelihood of the documents at different times under the Order Based DDP clustering model, and the real time of the documents. **Bottom:** Posterior word distributions as a function of time, for selected words, in the PNAS corpus.

6 SUMMARY

In this paper we present various probabilistic models for approaching the problem of introducing dependence on a covariate to clustering and factor models. We plan on continuing our explorations in this area, and implementing further models including time-dependent factor models, other DDP constructions, and models in which the topic distribution V_r vary with time. We hope that this paper will lay the ground for extending the powerful Dirichlet processes framework for topic models to model time (and other) variability effects.

References

[AMT04] Periklis Andritsos, Renee J. Miller, and Panayiotis Tsaparas. Information-theoretic tools for mining database structure from large data sets. In *SIGMOD*, 2004.

[Ant74] Charles E Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[Bor02] Vani K. Borooah. *Logit and probit: ordered and multinomial models*. Sage Publications, 2002.

[GS04] J.E. Griffin and M.F.J. Steel. Order-based dependent dirichlet processes. Research Report 430, Warwick Statistics, 2004.

[Hof01] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning Journal*, 42(1):177–196, 2001.

[IMRM04] Maria De Iorio, Peter Mueller, Gary L. Rosner, and Steven N. MacEachern. An anova model for dependent random measures. *Journal of the American Statistical Association*, 99(465):205–215, March 2004.

[Mac00] Steven N. MacEachern. Decision theoretic aspects of dependent nonparametric processes. In *The Sixth World Meeting of the International Society for Bayesian Analysis*, 2000.

[MQR04] Peter Mueller, Fernando Quintana, and Gary Rosner. A method for combining inference across related nonparametric bayesian models. *Journal of the Royal Statistical Society B*, 66:735–749, 2004.

[PCW02] Michael K. Pitt, Chris Chatfield, and Stephen G. Walker. Constructing first order stationary autoregressive models via latent processes. *Scand J Stat*, 29(4):657–657, 2002.

[RG02] Carl Edward Rasmussen and Zubin Ghahramani. Infinite mixtures of gaussian processes. In *Advances in Neural Information Processing Systems*, volume 14, 2002.

[Set94] J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[TJBB04] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. Technical Report 653, Department of Statistics, University of California, Berkeley, 2004.