The IBP Compound Dirichlet Process and its Application to Focused Topic Modeling

Sinead Williamson SAW56@CAM.AC.UK
Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, UK

Chong Wang CHONGW@CS.PRINCETON.EDU
Department of Computer Science, Princeton University, 35 Olden Street, Princeton, NJ 08540, USA

Katherine A. Heller HELLER@GATSBY.UCL.AC.UK
Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, UK

David M. Blei BLEI@CS.PRINCETON.EDU
Department of Computer Science, Princeton University, 35 Olden Street, Princeton, NJ 08540, USA

Abstract

The hierarchical Dirichlet process (HDP) is a Bayesian nonparametric mixed membership model—each data point is modeled with a collection of components of different proportions. Though powerful, the HDP makes an assumption that the probability of a component being exhibited by a data point is positively correlated with its proportion within that data point. This might be an undesirable assumption. For example, in topic modeling, a topic (component) might be rare throughout the corpus but dominant within those documents (data points) where it occurs. We develop the *IBP compound Dirichlet process* (ICD), a Bayesian nonparametric prior that decouples across-data prevalence and within-data proportion in a mixed membership model. The ICD combines properties from the HDP and the Indian buffet process (IBP), a Bayesian nonparametric prior on binary matrices. The ICD assigns a subset of the shared mixture components to each data point. This subset, the data point's "focus", is determined *independently* from the *amount* that each of its components contribute. We develop an ICD mixture model for text, the *focused topic model* (FTM), and show superior performance over the HDP-based topic model.

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010.

1. Introduction

Finite mixture models are widely used for clustering data (McLachlan & Peel, 2000). When fit to data, the components of a mixture model reflect similarity patterns, and each data point is probabilistically assigned to one of the components. When data are groups, i.e., each data point is a collection of observations, then mixed membership models are appropriate. Mixed membership models are a hierarchical variant of finite mixtures for grouped data where each data point exhibits multiple components. The components are shared across all data, and each data point exhibits them with different proportions. Mixed membership models are an effective tool for capturing complex data heterogeneity (Erosheva et al., 2004).

Mixed membership models, like finite mixtures, require an a priori choice of the number of components. To address this issue, Teh et al. (2006) developed the hierarchical Dirichlet process (HDP), a Bayesian nonparametric mixed membership model. The HDP allows for a potentially infinite number of components a priori so that, when conditioned on data, its posterior places a distribution over how many are exhibited. The HDP provides more flexible mixed membership modeling, avoiding costly model comparisons in order to determine an appropriate number of components.

However, the HDP makes a hidden assumption: A component's overall prevalence *across* data is positively correlated with the component's average proportion *within* data. The reason for this is that the HDP centers the random component proportions for each data point around the same global proportions.

This assumption may not be sensible. Consider modeling

a text corpus with an HDP. This is called "topic modeling" because the posterior components (called "topics") tend to reflect the semantic themes of the documents (Blei et al., 2003). The HDP assumption is that a frequent topic will, on average, occur frequently within each document. However, there is no reason to correlate the number of articles on a topic, such as baseball, with how much that topic contributes to any particular article. Baseball may be a rare topic, but articles about baseball often devote themselves exclusively to it. In this paper, we build a Bayesian nonparametric mixed membership model that allays this assumption. Our model decorrelates prevalence and proportion, allowing rarely seen components to occur with high proportion and frequently seen components to occur with low proportion.

We develop the IBP compound Dirichlet process (ICD), a Bayesian nonparametric mixed membership model that decorrelates across-data prevalence and within-data proportion. The ICD uses a random binary matrix drawn from the Indian buffet process (IBP, Griffiths & Ghahramani, 2005) to select *which* components are used in each data point (across-data prevalence), and an infinite series of gamma random variables to model *how much* they are used (within-data proportions). We use the ICD in the *focused topic model* (FTM), a generative model of document collections.

The central challenge in using the ICD is posterior inference. Sampling the IBP-distributed binary matrix directly leads to slow convergence, but integrating it out exactly is intractable due to the infinite combinatorial space of latent variables. We present an approximation to this integral, based on a technique used in Wang & Blei (2009), and use this approximation to develop an efficient collapsed Gibbs sampler.

We compare the FTM to the HDP topic model on three text corpora. We see that the FTM reduces the correlation between across-data prevalence and within-data proportion, which allows for a more compact representation of the data than the HDP provides. As a consequence, the FTM obtains a better fit to language and achieves substantially better perplexity on held out data.

2. IBP Compound Dirichlet Process

In this section we present the *IBP compound Dirichlet process* (ICD), a Bayesian nonparametric prior which addresses the limitations of the hierarchical Dirichlet process (HDP). We develop the *focused topic model* (FTM), an application of the ICD to document analysis. We assume the reader is familiar with the Dirichlet process (DP); for a review, see Ghosal (2010).

2.1. Hierarchical Dirichlet Processes

The hierarchical Dirichlet process (HDP, Teh et al., 2006) is a prior appropriate for Bayesian nonparametric mixed membership modeling. In an HDP, each data point is associated with a draw from a Dirichlet process (DP), which determines how much each member of a shared set of mixture components contributes to that data point. The base measure of this data-level DP is itself drawn from a DP, which ensures that there is a single discrete set of components shared across the data. More precisely, the generative process for the per-data distribution G_m is:

$$G_0 \sim \mathrm{DP}(\zeta, H),$$
$$G_m \sim \mathrm{DP}(\tau, G_0) \quad \text{for each } m.$$

Each distribution G_m is a sample from a DP with concentration parameter τ and base probability measure G_0. This base measure G_0 is itself a draw from a DP, with concentration parameter ζ and base measure H. The base measure G_0 is thus discrete and, consequently, the per-data distributions G_m are also discrete with common support determined by the locations of the atoms of G_0.

Each atom represents a component and is described by a location, a weight in G_m, and a weight in G_0. The location is identical in both G_0 and G_m; it gives the parameters associated with the component, e.g., a Gaussian mean or a distribution over terms. The weight in G_m gives the proportion for that component in the mth data point.

The weight of a component in G_m is drawn from a distribution centered around the corresponding weight in G_0. Thus, the weight for any given component is drawn from the same distribution across all the data, and that distribution controls both how prevalent the component is and its proportion within each data point. For example, if a component has low weight in G_0 then it will also have low weight in most G_m. That component is unlikely to contribute to data points and, when it does, that contribution will be very small.

As mentioned in the introduction, this is not necessarily a desirable modeling assumption. Rather than control these two properties via a single variable, as is the case in the HDP, we wish to model them separately. We develop a model where an infrequently occurring component can still have high proportion when it does occur, and vice versa.

2.2. Indian Buffet Process

Our model uses the Indian buffet process (IBP, Griffiths & Ghahramani, 2005) to control component occurrence separately from component proportion. The IBP defines a distribution over binary matrices with an infinite number of columns, only a finite number of which contain non-zero entries. It can be derived by taking the limit as $K \to \infty$

of a finite $M \times K$ binary matrix $\mathbf{B}$, with elements b_{mk} distributed according to,

$$\begin{aligned} \pi_k &\sim \text{Beta}(\alpha/K, 1), \\ b_{mk} &\sim \text{Bernoulli}(\pi_k) \quad \text{for each } m, \end{aligned}$$

where the mth row of $\mathbf{B}$ is $\mathbf{b}_m$, the kth cell of $\mathbf{b}_m$ is b_{mk}, and π_k is the probability of observing a non-zero value in column k. As K tends to infinity, we can obtain a strictly decreasing ordering of the latent probabilities π_k by starting with a "stick" of unit length and recursively breaking it at a point $\text{Beta}(\alpha, 1)$ along its length, discarding the excess (Teh et al., 2007), for $k = 1, 2, \ldots$:

$$\begin{aligned} \mu_k &\sim \text{Beta}(\alpha, 1), \\ \pi_k &= \textstyle\prod_{j=1}^{k} \mu_j, \\ b_{mk} &\sim \text{Bernoulli}(\pi_k) \quad \text{for each } m. \end{aligned} \tag{1}$$

In our model, the rows of the IBP matrix represent data points, the columns represent components, and the cells indicate which components contribute to which data points.

2.3. IBP compound Dirichlet process

We now develop a prior over a set of discrete probability distributions that decorrelates which components occur and in what proportion. Rather than assigning positive probability mass to all components for every data point, as in the HDP, our model assigns positive probability to only a subset of components, selected independently of their masses.

The IBP provides a method for selecting subsets from a countably infinite set of components. Thus, one way of achieving our goal is to introduce the IBP directly into the HDP, using the mth row of the IBP to determine a subset of the infinite set of atoms present in the top level DP sample G_0. This defines an (unnormalized) measure that can be used as the base measure for a data-specific DP.

This can be seen as an infinite spike and slab model. Spike and slab models describe a mixture model between a continuous distribution (the "slab")[1] and the measure degenerate at zero. A "spike" distribution determines which variables are drawn from the slab, and which are zero. In the model above, the spikes are provided by the IBP, and the slab is provided by the top level DP.

However, better choices for the top-level "slab" distribution are available. Draws from the DP "slab" are constrained to sum to one, which restricts the distribution over component proportions and introduces dependencies between atom masses which can lead to difficulties in developing inference schemes. While we wish to ensure that the base measure of the lower level DP is still normalizable (i.e. is drawn from a convergent process), we do not need to rely on the slab to enforce this constraint. Since the IBP selects a finite number of components for each data point, it is sufficient merely to ensure that the sum of any finite subset of top-level atoms is finite. Thus, rather than drawing the atoms of the slab from a DP, we sample their masses as independent gamma random variables. This eliminates the restrictions on component proportions imposed by a DP and, since the resulting component proportions are independent, makes inference much easier.

The model assumes the following generative process,

1. for $k = 1, 2, \ldots,$
   (a) Sample the stick length π_k according to Eq. 1.
   (b) Sample the relative mass $\phi_k \sim \text{Gamma}(\gamma, 1)$.
   (c) Sample the atom location $\beta_k \sim H$.
2. for $m = 1, \ldots, M,$
   (a) Sample a binary vector $\mathbf{b}_m$ according to Eq. 1.
   (b) Sample the lower level DP, $G_m \sim \text{DP}(\sum_k b_{mk}\phi_k, \frac{\sum_k b_{mk}\phi_k\delta_{\beta_k}}{\sum_k b_{mk}\phi_k})$.

In sampling the lower level DP, masses are assigned to the atoms δ_{β_k} independent of their locations. Since the number of locations selected by the binary vector $\mathbf{b}_m$ is finite almost surely, these masses can be sampled from a Dirichlet distribution defined over the selected ϕ_k:

$$\begin{aligned} \boldsymbol{\theta}_m &\sim \text{Dirichlet}(\mathbf{b} \cdot \boldsymbol{\phi}) \\ G_m &= \textstyle\sum_k \theta_{mk}\delta_{\beta_k}, \end{aligned}$$

where $\mathbf{b} \cdot \boldsymbol{\phi}$ is the Hadamard product of $\mathbf{b}$ and $\boldsymbol{\phi}$. If we marginalize out the sparse binary matrix $\mathbf{B}$ and the gamma random variables ϕ_k, the atom masses are distributed according to a mixture of Dirichlet distributions governed by the IBP:

$$p(\boldsymbol{\theta}_m|\gamma, \alpha) = \int \mathrm{d}\boldsymbol{\phi} \textstyle\sum_{\mathbf{B}} p(\boldsymbol{\theta}_m|\mathbf{B}, \boldsymbol{\phi}) p(\mathbf{B}|\alpha) p(\boldsymbol{\phi}|\gamma), \tag{2}$$

where,

$$\begin{aligned} \mathbf{B} &\sim \text{IBP}(\alpha), \\ \phi_k &\sim \text{Gamma}(\gamma, 1), \\ \boldsymbol{\theta}_m &\sim \text{Dirichlet}(\mathbf{b}_m \cdot \boldsymbol{\phi}). \end{aligned}$$

We call this model the IBP compound Dirichlet process (ICD), since the IBP provides the mixing measure for a mixture of Dirichlet distributions. Like the HDP, this model is a form of dependent Dirichlet process (MacEachern, 2000).

The ICD achieves our goal of decoupling how often the components occur and in what proportion. The IBP draw

[1] In its original form, the slab was a uniform distribution. However, the concept and terminology have also been employed in models where the slab is not the uniform distribution – see for example (Ishwaran & Rao, 2005) – and it is in this more general sense that we use the term.

$\mathbf{B}$ selects a subset of atoms for each distribution, and the gamma random variables $\boldsymbol{\phi}$ determine the relative masses associated with these atoms.

2.4. Focused Topic Models

Suppose H parametrizes distributions over words. Then, the ICD defines a generative topic model, where it is used to generate a set of sparse distributions over an infinite number of components, called "topics." Each topic is drawn from a Dirichlet distribution over words. In order to specify a fully generative model, we sample the number of words for each document from a negative binomial distribution, $n_{\cdot}^{(m)} \sim \mathrm{NB}(\sum_k b_{mk}\phi_k, 1/2)$.[2]

The generative model for M documents is

1. for $k = 1, 2, \ldots,$
   (a) Sample the stick length π_k according to Eq. 1.
   (b) Sample the relative mass $\phi_k \sim \mathrm{Gamma}(\gamma, 1)$.
   (c) Draw the topic distribution over words, $\boldsymbol{\beta}_k \sim \mathrm{Dirichlet}(\eta)$.
2. for $m = 1, \ldots, M,$
   (a) Sample a binary vector $\mathbf{b}_m$ according to Eq. 1.
   (b) Draw the total number of words, $n_{\cdot}^{(m)} \sim \mathrm{NB}(\sum_k b_{mk}\phi_k, 1/2)$.
   (c) Sample the distribution over topics, $\boldsymbol{\theta}_m \sim \mathrm{Dirichlet}(\mathbf{b}_m \cdot \boldsymbol{\phi})$.
   (d) For each word $w_{mi}, i = 1, \ldots, n_{\cdot}^{(m)}$,
      i. Draw the topic index $z_{mi} \sim \mathrm{Discrete}(\boldsymbol{\theta}_m)$.
      ii. Draw the word $w_{mi} \sim \mathrm{Discrete}(\boldsymbol{\beta}_{z_{mi}})$.

We call this the *focused topic model* (FTM) because the infinite binary matrix $\mathbf{B}$ serves to focus the distribution over topics onto a finite subset (see Figure 1). The number of topics within a single document is almost surely finite, though the total number of topics is unbounded. The topic distribution for the mth document, $\boldsymbol{\theta}_m$, is drawn from a Dirichlet distribution over the topics selected by $\mathbf{b}_m$. The Dirichlet distribution models uncertainty about topic proportions while maintaining the restriction to a sparse set of topics.

The ICD models the distribution over the global topic proportion parameters $\boldsymbol{\phi}$ separately from the distribution over the binary matrix $\mathbf{B}$. This captures the idea that a topic may appear infrequently in a corpus, but make up a high proportion of those documents in which it occurs. Conversely, a topic may appear frequently in a corpus, but only with low proportion.

[2]Notation $n_k^{(m)}$ is the number of words assigned to the kth topic of the mth document, and we use a dot notation to represent summation - i.e. $n_{\cdot}^{(m)} = \sum_k n_k^{(m)}$.

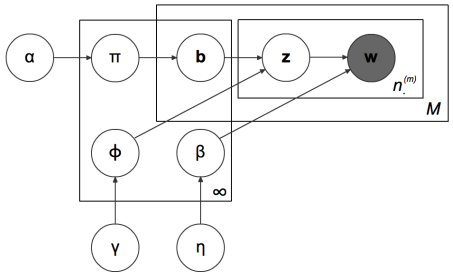

Figure 1. Graphical model for the focused topic model

3. Related Models

Titsias (2007) introduced the infinite gamma-Poisson process, a distribution over unbounded matrices of non-negative integers, and used it as the basis for a topic model of images. In this model, the distribution over features for the mth image is given by a Dirichlet distribution over the non-negative elements of the mth row of the infinite gamma-Poisson process matrix, with parameters proportional to the values at these elements. While this results in a sparse matrix of distributions, the number of zero entries in any column of the matrix is correlated with the values of the non-zero entries. Columns which have entries with large values will not typically be sparse. Therefore, this model will not decouple across-data prevalence and within-data proportions of topics. In the ICD the number of zero entries is controlled by a separate process, the IBP, from the values of the non-zero entries, which are controlled by the gamma random variables.

The sparse topic model (SparseTM, Wang & Blei, 2009) uses a finite spike and slab model to ensure that each topic is represented by a sparse distribution over words. The spikes are generated by Bernoulli draws with a single topic-wide parameter. The topic distribution is then drawn from a symmetric Dirichlet distribution defined over these spikes. The ICD also uses a spike and slab approach, but allows an unbounded number of "spikes" (due to the IBP) and a more globally informative "slab" (due to the shared gamma random variables). We extend the SparseTM's approximation of the expectation of a finite mixture of Dirichlet distributions, to approximate the more complicated mixture of Dirichlet distributions given in Eq. 2.

Recent work by Fox et al. (2009) uses draws from an IBP to select subsets of an infinite set of states, to model multiple dynamic systems with shared states. (A state in the dynamic system is like a component in a mixed membership model.) The probability of transitioning from the ith state to the jth state in the mth dynamic system is drawn from a Dirichlet distribution with parameters $b_{mj}\gamma + \tau\delta_{i,j}$, where

γ and τ are constant. This model does not allow sharing of information about the within-data probability of a state between data points, which is modeled in the ICD via the gamma-distributed ϕ_k. An alternative inference scheme is also used, where the IBP matrix is sampled instead of being integrated out.

A number of other models have been proposed to address the rigid topic correlation structure assumed in the LDA and HDP topic models, including the correlated topic model (CTM, Blei & Lafferty, 2005) and the pachinko allocation model (PAM, Li & McCallum, 2006). Our aim is different. The FTM reduces undesirable correlations between the prevalence of a topic across the corpus and its proportion within any particular document, rather than adding new correlations between topics. Correlations among topics could be integrated into the FTM in future work.

4. Posterior Inference

We use Gibbs sampling for posterior inference over the latent variables. The algorithm cyclically samples the value for a single variable from its conditional distribution given the remaining variables. To improve mixing time, we use a collapsed Gibbs sampler, integrating out the topic-specific word distributions $\boldsymbol{\beta}$, the topic mixture distributions $\boldsymbol{\theta}$, and the sparsity pattern $\mathbf{B}$. We use an approximate method, described in the appendix, to integrate out the infinite-dimensional sparsity pattern $\mathbf{B}$. We sample only the global topic proportion variables $\boldsymbol{\phi}$, the global topic sparsity probability variables $\boldsymbol{\pi}$, and the topic assignments $\mathbf{z}$.

4.1. Sampling z

The conditional distribution of the topic assignment of the ith word in the mth document depends on the posterior distribution of the topic proportion $\boldsymbol{\theta}_m$ for that document:

$$\begin{aligned} &p(z_{mi} = k|\mathbf{z}_{\neg mi}, w_{mi}, \mathbf{w}_{\neg mi}, \Psi) \\ &\propto p(w_{mi}|z_{mi} = k, \mathbf{z}_{\neg mi}, \mathbf{w}_{\neg mi}) p(z_{mi} = k|\mathbf{z}_{\neg mi}, \Psi) \\ &\propto (n_{k,\neg i}^{(w_{mi})} + \eta) \int \mathrm{d}\boldsymbol{\theta}_m \, p(z_{mi} = k|\boldsymbol{\theta}_m) p(\boldsymbol{\theta}_m|\mathbf{z}_{\neg mi}, \Psi), \end{aligned}$$

where $\Psi = \{\boldsymbol{\pi}^\bullet, \boldsymbol{\phi}^\bullet, n_\cdot^{(m)}, \alpha, \gamma\}$, and $n_k^{(w)}$ is the number of times word w has been assigned to topic k in the vector of assignments $\mathbf{z}$. We use $\boldsymbol{\phi}^\bullet, \boldsymbol{\pi}^\bullet$ to represent those elements of $\boldsymbol{\phi}$ and $\boldsymbol{\pi}$ associated with topics which are currently represented in the corpus, and $\boldsymbol{\phi}^\circ, \boldsymbol{\pi}^\circ$ to represent the remaining elements, which are associated with unused topics and whose values are therefore unknown.

Conditioned on the sparse binary vector $\mathbf{b}_m$ and the gamma random variables $\boldsymbol{\phi}$, the topic mixture distribution, $\boldsymbol{\theta}_m$, is distributed according to a Dirichlet distribution. The sparse vector $\mathbf{b}_m$ determines the subset of topics over which the Dirichlet distribution is defined, and the gamma random variables $\boldsymbol{\phi}$ determine the values of the Dirichlet parameters at these points. If we integrate out the sparse binary vector $\mathbf{b}_m$, rather than sampling $\boldsymbol{\theta}_m$ from a single Dirichlet distribution, we must sample it from an infinite mixture of Dirichlet distributions, with the IBP determining the mixing proportions:

$$\begin{aligned} &p(\boldsymbol{\theta}_m|\mathbf{z}_{\neg mi}, \Psi) \\ &\propto \int \mathrm{d}\boldsymbol{\phi}^\circ \textstyle\sum_{\mathbf{b}_m} \mathrm{Dirichlet}(\boldsymbol{\theta}_m|(\mathbf{n}_{\neg i}^{(m)} + \boldsymbol{\phi}) \cdot \mathbf{b}_m) \\ &\qquad p(\mathbf{b}_m, \boldsymbol{\phi}^\circ, n_\cdot^{(m)}|\boldsymbol{\phi}^\bullet, \boldsymbol{\pi}^\bullet, \gamma, \alpha), \end{aligned} \tag{3}$$

where $\mathbf{n}_{\neg i}^{(m)}$ is the topic assignment statistic excluding word w_{mi}. However, we cannot integrate out the sparse binary vector $\mathbf{b}_m$ exactly due to its combinatorial nature. Fortunately, we only ever use the expected values of $\boldsymbol{\theta}_m$ given $\mathbf{z}_{\neg mi}$ and Ψ, since $\int \mathrm{d}\boldsymbol{\theta}_m p(z_{mi} = k|\boldsymbol{\theta}_m) p(\boldsymbol{\theta}_m|\mathbf{z}_{\neg mi}, \Psi) = \mathbb{E}[\theta_{mk}|\mathbf{z}_{\neg mi}, \Psi]$ (from Eq. 3). This expectation can be efficiently approximated via a procedure detailed in the appendix.

4.2. Sampling $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$

To sample $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$, we re-instantiate the binary matrix $\mathbf{B}$ as an auxiliary variable, and iteratively sample $\mathbf{B}$, $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$. We categorize the columns of $\mathbf{B}$ as "active" if $n_k^{(\cdot)} > 0$, and "inactive" otherwise.

The total number of words in the mth document assigned to the kth topic is distributed according to $\mathrm{NB}(b_{mk}\phi_k, 1/2)$. The joint probability of ϕ_k and the total number of words assigned to the kth topic is given by,

$$\begin{aligned} &p(\phi_k, n_k^{(\cdot)}|\mathbf{b}_k, \gamma) = p(\phi_k|\gamma) \textstyle\prod_{m=1}^M p(n_k^{(m)}|b_{mk}, \phi_k) \\ &= \frac{\phi_k^{\gamma-1} e^{-\phi_k}}{\Gamma(\gamma)} \textstyle\prod_{m:b_{mk}=1} \frac{\Gamma(\phi_k + n_k^{(m)})}{\Gamma(\phi_k) n_k^{(m)}! 2^{\phi_k + n_k^m}}. \end{aligned} \tag{4}$$

This is log differentiable with respect to ϕ_k and γ. Thus we use Hybrid Monte Carlo (MacKay, 2002) to sample from the posteriors of ϕ_k and γ.

To sample the π_k, we follow a similar approach to the semi-ordering stick-breaking scheme of (Teh et al., 2007). The "active" features are distributed according to:

$$p(\pi_k|\mathbf{B}) \sim \mathrm{Beta}\left(\textstyle\sum_{m=1}^M b_{mk}, 1 + M - \sum_{m=1}^M b_{mk} \right) \tag{5}$$

and the "inactive" features are strictly ordered as suggested by Eq. 1. (Note that the definition given here of "active" and "inactive" features differs slightly from that given in Teh et al. (2007), as we consider a feature where $n_k^{(\cdot)} = 0$ to be inactive, and therefore subject to strict ordering, regardless of whether $\sum_m b_{mk} > 0$.)

Since the binary matrix $\boldsymbol{B}$ is discarded after this step, and we only use the active π_k to sample the topic allocations,

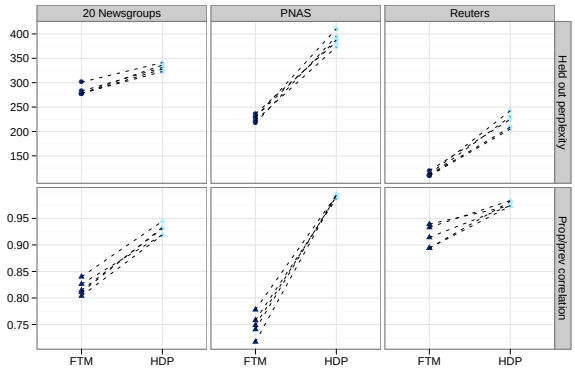

Figure 2. Experimental comparison between FTM (dark blue) and HDP (pale blue) on three datasets. Each point represents the result on one fold, and is computed with the other folds as training data. Dashed lines connect the results from the same fold. **Top.** Test set perplexities. Lower numbers indicate better performance. **Bottom.** Correlation between topic presence frequency and topic proportion. The FTM reduces the correlation between them.

we have no interest in the inactive features and only need to sample the active elements of $\mathbf{B}$ and $\boldsymbol{\pi}$. The elements of $\mathbf{B}$ are sampled according to the following probabilities:

$$p(b_{mk}|\pi_k, \phi_k, n_k^{(m)}) = \begin{cases} b_{mk} & \text{if } n_k^{(m)} > 0 \\ \frac{2^{\phi_k}(1-\pi_k)}{\pi_k + 2^{\phi_k}(1-\pi_k)} & \text{if } b_{mk} = 0 \text{ and } n_k^{(m)} = 0 \\ \frac{\pi_k}{\pi_k + 2^{\phi_k}(1-\pi_k)} & \text{if } b_{mk} = 1 \text{ and } n_k^{(m)} = 0. \end{cases} \quad (6)$$

With equations 3, 4, 5 and 6 we have specified the full Gibbs sampler for the FTM model[3]. From the states of this sampler, we can compute topics, topic proportions, and sparsity patterns.

5. Empirical study

We compared the performance of the FTM to the HDP with three datasets:

- *PNAS*: This is a collection of 1766 abstracts from the Proceedings of the National Academy of Sciences (PNAS) from between 1991 and 2001. The vocabulary contains 2452 words.
- *20 Newsgroups*: This is a collection of 1000 randomly selected articles from the 20 newsgroups dataset.[4] The vocabulary contains 1407 words.
- *Reuters-21578*: This is a collection of 2000 randomly selected documents from the Reuters-21578 dataset.[5] The vocabulary contains 1472 words.

[3]Matlab code is available from the authors
[4]http://people.csail.mit.edu/jrennie/20Newsgroups/
[5]http://kdd.ics.uci.edu/databases/reuters21578/

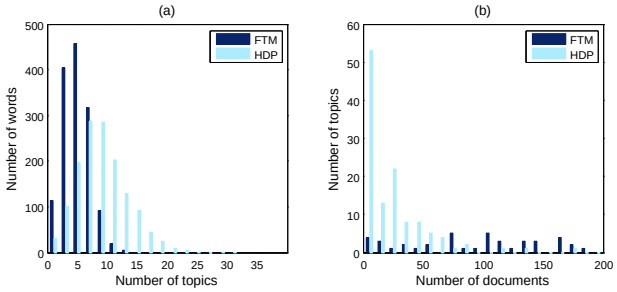

Figure 3. (**a**) Histogram of the number of topics a word appears in for the FTM (dark blue, left) and the HDP (pale blue, right) models on *20 Newsgroups* data. In the FTM, words are generally associated with fewer topics. (**b**) Histogram of the number of documents a topic appears in for the FTM (dark blue, left) and HDP (pale blue, right) models on *20 Newsgroups* data. In both models, topics appearing in more than 200 documents have been excluded to focus on the low frequency topics. The HDP has many more topics which only appear in a very few documents.

For each dataset, the vocabulary excluded stop-words and words occurring in fewer than 5 documents.

In both the FTM and HDP topic models, we fixed the topic distribution hyper-parameter η to 0.1. Following Teh et al. (2006), we used priors of $\zeta \sim \text{Gamma}(5, 0.1)$ and $\tau \sim \text{Gamma}(0.1, 0.1)$ in the HDP. In the FTM, we used prior $\gamma \sim \text{Gamma}(5, 0.1)$, and we fixed $\alpha = 5$.

First, we examined test-set perplexity, a measure of how well the models generalize to new data. Figure 2 (top) shows test set perplexities obtained on each dataset for the two models, using 5-fold cross-validation. To obtain these measurements, we ran both Gibbs samplers for 1000 iterations, discarding the first 500. In each case, the FTM achieves better (lower) perplexity on the held out data.

We developed the FTM to decorrelate the probability of a topic being active within a document and its proportion within the documents attributed to it. To consider whether this is observed in the posterior, we next compared two statistics for each topic found. The *topic presence frequency* for a given topic is the fraction of the documents within the corpus that contain at least one incidence of that topic. The *topic proportion* for a topic is the fraction of the words within the corpus attributed to that topic.

The correlation between topic presence frequencies and topic proportions is shown in Figure 2 (bottom). In each case we see lower correlation for the FTM. In fact, the dataset which exhibits the greatest improvement in held-out perplexity under the FTM, also exhibits the greatest decrease in correlation between frequency and proportion.

In our study, we observed that the FTM posterior prefers a more compact representation. It contains fewer topics overall than the HDP, but more topics per document. While

the HDP uses a larger number of topics to model the corpus than the FTM, most of these topics appear in only a handful of documents. In addition, individual words in the vocabulary are, on average, associated with more topics under the HDP, meaning that FTM topics are generally more distinct. These findings are illustrated in Figure 3. (These results are for *20 Newsgroups*. Other data exhibited similar properties.)

Finally, we note that the computational complexity of both models grows with the number of topics represented. When the number of topics is equal, a single iteration of the FTM algorithm is more costly than a single iteration of the HDP. However, the more compact representation of the FTM yields similar runtime. In the data we analyzed, the FTM analysis was as fast as the HDP analysis.

6. Discussion

We developed the *IBP compound Dirichlet process* (ICD), a Bayesian nonparametric prior over discrete, dependent distributions, which is an alternative to the hierarchical Dirichlet process (HDP). The ICD decouples the relationship between across-data prevalence and within-data proportion.

We have used the ICD to construct the *focused topic model* (FTM), a generative model for collections of documents. We demonstrated that the FTM provides a better fit to text than the HDP topic model. The representations obtained by the FTM reflect lower correlation between across-data prevalence and within-data proportion than the representations obtained by the HDP.

We have concentrated on correlation in HDPs. This correlation does not occur in finite mixed membership models with mixture components drawn from a Dirichlet($\boldsymbol{\alpha}$), where $\boldsymbol{\alpha}$ is fixed, since the draws from Dirichlet($\boldsymbol{\alpha}$) are i.i.d. However, restriction to a fixed Dirichlet distribution limits the flexibility of mixed membership models, and if $\boldsymbol{\alpha}$ is unknown, there will be a similar correlation bias as with the HDP. Although easily derived, there is currently no parametric counterpart of the ICD.

The idea of removing correlation between global and local probabilities is potentially applicable to a range of models. For example, in a state transition sequence, a certain state may be inaccessible from most other states, but occur with high probability following a small subset of states. Such a relationship will be poorly captured using the HDP-based Hidden Markov model (Teh et al., 2006), as it tends to correlate the global occurrence of a state with the probability of transitioning to it from another state. A model based on the ICD could better capture such relationships. Exploring which HDP applications benefit most from this decoupling is an avenue for further research.

Acknowledgments. We thank anonymous reviewers for valuable comments. David M. Blei is supported by ONR 175-6343 and NSF CAREER 0745520.

References

Blei, D. M. and Lafferty, J. Correlated topic models. In *NIPS*, volume 18, 2005.

Blei, D. M., Jordan, M. I., and Ng, A. Y. Latent Dirichlet allocation. *JMLR*, pp. 993–1022, 2003.

Erosheva, E., Fienberg, S., and Lafferty, J. Mixed-membership models of scientific publications. *PNAS*, 101(Suppl 1):5220–5227, April 2004.

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. Sharing features among dynamical systems with beta processes. In *NIPS*, volume 22, 2009.

Ghosal, S. Dirichlet process, related priors and posterior asymptotics. In Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (eds.), *Bayesian Nonparametrics*, Cambridge Series in Statistical and Probabilistic Mathematics, chapter 2, pp. 35–79. Cambridge University Press, 2010.

Griffiths, T. L. and Ghahramani, Z. Infinite latent feature models and the Indian buffet process. In *NIPS*, volume 18, 2005.

Ishwaran, H. and Rao, J. S. Spike and slab variable selection: Frequentist and Bayesian strategies. *Annals of Statistics*, 33: 730, 2005.

Li, W. and McCallum, A. Pachinko allocation: DAG-structured mixture models of topic correlations. In *ICML*, volume 23, 2006.

MacEachern, S. N. Dependent Dirichlet processes. Technical report, Department of Statistics, Ohio State University, 2000.

MacKay, D. J. C. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, 2002.

McLachlan, G. and Peel, D. *Finite Mixture Models*. John Wiley & Sons, 2000.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical Dirichlet processes. *JASA*, 101(476):1566–1581, 2006.

Teh, Y. W., Görür, D., and Ghahramani, Z. Stick-breaking construction for the Indian buffet process. In *AISTATS*, volume 11, 2007.

Titsias, M. K. The infinite Gamma-Poisson feature model. In *NIPS*, volume 21, 2007.

Wang, C. and Blei, D. M. Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. In *NIPS*, volume 22, 2009.

A. Approximating the expectation of $\boldsymbol{\theta}_m$

First we recall that $\boldsymbol{\phi}^\bullet, \boldsymbol{\pi}^\bullet$ denote those elements of $\boldsymbol{\phi}$ and $\boldsymbol{\pi}$ associated with the topics that are currently represented in the corpus, and $\boldsymbol{\phi}^\circ, \boldsymbol{\pi}^\circ$ denote the rest. As described in Section 4.1, in order to sample the topic allocations z_i, we wish to approximate the expectation of $\boldsymbol{\theta}_m$ given the posterior distribution in Equation 3. This is an extension of the technique used to approximate the expectation of the distribution over words in Wang & Blei (2009).

The expectation of the kth element of $\boldsymbol{\theta}_m$ is:

$$\begin{aligned}
&\mathbb{E}[\theta_{mk}|\boldsymbol{z}_{\neg mi}, \Psi] \\
&\propto \int \mathrm{d}\boldsymbol{\phi}^\circ \sum_{\substack{\mathbf{b}_m^\circ: \\ b_{mk}=1}} (n_{k,\neg i}^{(m)} + \phi_k) p(\mathbf{b}_m, \boldsymbol{\phi}^\circ, n_\cdot^{(m)} | \boldsymbol{\phi}^\bullet, \boldsymbol{\pi}^\bullet, \gamma, \alpha) \\
&\propto \int \mathrm{d}\boldsymbol{\phi}^\circ \sum_{\substack{\mathbf{b}_m^\circ: \\ b_{mk}=1}} \frac{(n_{k,\neg i}^{(m)} + \phi_k) p(\mathbf{b}_m^\circ | \boldsymbol{\pi}^\bullet, \alpha) p(\boldsymbol{\phi}^\circ|\gamma)}{2^{\sum_j b_{mj}\phi_j} (n_{\cdot,\neg i}^{(m)} + \sum_j b_{mj}\phi_j)}.
\end{aligned}$$

We divide $\sum_j b_{mj}\phi_j$ into a "known" term X, corresponding to those topics with which words in the mth document are associated,and an "unknown" term, corresponding to those topics with which no words in the mth document are associated:

$$\sum_j b_{mj}\phi_j = \underbrace{\sum_{j:n_{j,\neg i}^{(m)}>0} \phi_j}_{X} + \underbrace{\sum_{j:n_{j,\neg i}^{(m)}=0} b_{mj}\phi_j}_{Y}.$$

Let $g_{X,k}(Y) \triangleq \frac{b_{mk}(n_{k,\neg i}^{(m)} + \phi_k)}{2^{X+Y}(n_{\cdot,\neg i}^{(m)} + X + Y)}$. Then we can write the expectation of θ_{mk} in terms of the expectation of $g_{X,k}(Y)$:

$$\mathbb{E}[\theta_{mk}|\Psi, \mathbf{n}_{\neg i}^m] \propto \mathbb{E}[g_{X,k}(Y)|\boldsymbol{\pi}^\bullet, \boldsymbol{\phi}^\bullet]. \tag{7}$$

Consider approximating this expectation in three scenarios:

1. $n_{k,\neg i}^{(m)} > 0$: the kth topic is currently represented in the document under consideration;
2. $n_{k,\neg i}^{(m)} = 0$ and $n_{k,\neg i}^{(\cdot)} > 0$: the kth topic is not currently represented in the document under consideration, but *is* currently represented in the corpus;
3. $n_{k,\neg i}^{(\cdot)} = 0$: the kth topic is not currently represented in the corpus.

In the *first case*, where $n_{k,\neg i}^{(m)} > 0$, we know that $b_{mk} = 1$, so the expectation in equation 7 is

$$\begin{aligned}
&\mathbb{E}[g_{X,k}(Y)|\boldsymbol{\pi}^\bullet, \boldsymbol{\phi}^\bullet, \mathbf{n}^{(m)}] \\
&\quad = (n_{k,\neg i}^{(m)} + \phi_k)\mathbb{E}\left[\frac{1}{2^{X+Y}(n_{\cdot,\neg i}^{(m)} + X + Y)}\right]
\end{aligned} \tag{8}$$

We divide Y into two components $Y = Y^* + Y^\dagger$, where

$$Y^* = \textstyle\sum_{j\in\mathcal{J}_m^*} b_{mj}\phi_j \quad \text{and} \quad Y^\dagger = \sum_{j\in\mathcal{J}^\dagger} b_{mj}\phi_j.$$

where $\mathcal{J}_m^*$ is the set of j such that $n_{j,\neg i}^{(m)} = 0, n_{j,\neg i}^{(\cdot)} > 0$, and $\mathcal{J}^\dagger$ is the set of j such that $n_j^{(\cdot)} = 0$. The expectation and variance of Y conditioned on $\boldsymbol{\pi}^\bullet, \boldsymbol{\phi}^\bullet, \alpha$ and γ, are

$$\begin{aligned}
\mathbb{E}[Y|\boldsymbol{\pi}^\bullet, \boldsymbol{\phi}^\bullet, \alpha, \gamma] &= \mathbb{E}[Y^*|\boldsymbol{\pi}^\bullet, \boldsymbol{\phi}^\bullet] + \mathbb{E}[Y^\dagger|\alpha, \gamma] \\
\mathbb{V}[Y|\boldsymbol{\pi}^\bullet, \boldsymbol{\phi}^\bullet, \alpha, \gamma] &= \mathbb{V}[Y^*|\boldsymbol{\pi}^\bullet, \boldsymbol{\phi}^\bullet] + \mathbb{V}[Y^\dagger|\alpha, \gamma],
\end{aligned} \tag{9}$$

where

$$\begin{aligned}
\mathbb{E}[Y^*|\boldsymbol{\pi}^\bullet, \boldsymbol{\phi}^\bullet] &= \textstyle\sum_{j\in\mathcal{J}_m^\bullet} \pi_j\phi_j \\
\mathbb{V}[Y^*|\boldsymbol{\pi}^\bullet, \boldsymbol{\phi}^\bullet] &= \textstyle\sum_{j\in\mathcal{J}_m^\bullet} \phi_j^2\pi_j(1-\pi_j) \\
\mathbb{E}[Y^\dagger|\alpha, \gamma] &= \alpha\gamma \\
\mathbb{V}[Y^\dagger|\alpha, \gamma] &= \alpha\gamma(\gamma+1) - \tfrac{\alpha^2\gamma^2}{2\alpha+1}.
\end{aligned} \tag{10}$$

In summary, Y^* is a summation over a finite number of topics that grows in expectation as $\alpha(\log(M) - 1)$, where M is the total number of documents; and $Y^\dagger$ is a summation over a countably infinite number of topics. If the number of documents, and therefore the expected total number of topics, is large, then the central limit theorem suggests that we can approximate $(X + Y)$ using a Gaussian distribution with mean and variance as described in equations 9 and 10. We can then approximate the expectation in equation 8 using a second order Taylor expansion.

In the *second case*, where $n_{k,\neg i}^{(m)} = 0$ and $n_{k,\neg i}^{(\cdot)} > 0$, the expectation in equation 7 becomes:

$$\begin{aligned}
&\mathbb{E}[g_{X,k}(Y)|\boldsymbol{\pi}^\bullet, \boldsymbol{\phi}^\bullet] \\
&\quad = \phi_k\pi_k\mathbb{E}\left[\frac{1}{2^{X+Y_{\neg k}+\phi_k}(n_{k,\neg i}^{(m)} + X + Y_{\neg k} + \phi_k)}\right],
\end{aligned}$$

where $Y_{\neg k} = Y - b_{mk}\phi_k$. We approximate the expectation as described for the first case (where $n_{k,\neg i}^{(m)} > 0$), noting the slight changes in the expectation and variance of $Y_{\neg k}$ compared with Y.

In the *third case*, where $n_{k,\neg i}^{(\cdot)} = 0$, we consider the probability of assigning the ith word to one of the infinite number of unobserved classes. Let $z = u$ indicate assignment to a previously unseen topic. Equation 7 becomes:

$$\begin{aligned}
&\mathbb{E}[g_{X,u}(Y)|\boldsymbol{\pi}^\bullet, \boldsymbol{\phi}^\bullet] \\
&\quad = \mathbb{E}\left[\frac{Y^\dagger}{2^{X+Y^*+Y^\dagger}(n_{\cdot,\neg i}^{(m)} + X + Y^* + Y^\dagger)}\right].
\end{aligned} \tag{11}$$

We approximate the expectation in equation 11 using a multivariate second order Taylor expansion, noting that Y^* and $Y^\dagger$ are independent conditioned on $\boldsymbol{\pi}^\bullet$, $\boldsymbol{\phi}^\bullet$, α and γ.

We have evaluated the quality of this approximation by comparing the approximated expectations with those estimated using Monte Carlo simulations, and found that the approximation holds well, provided γ is not large. The values of γ learned by the model in the experiments described in section 5 fall into this paradigm.