Hierarchical Dirichlet Processes
AMS 241 Project, Fall 2010

Vadim von Brzeski, vvonbrze@ucsc.edu

December 7, 2010

1 Introduction

This work is part review and part experimental investigation of the work done by Teh, Jordan, Beal, and Blei on *Hierarchical Dirichlet Processes*, first published as a technical report in 2004 [7]. In this report, we review the theory and motivation behind hierarchical Dirichlet processes (HDP), and we study HDP inference using simulated data and a Gibbs sampler we developed for this purpose. We assume the reader is already familiar with the theory behind (non-hierarchical) Dirichlet processes and Dirichlet process mixture models, and focus specifically on the hierarchical extension of those models.

Hierarchical Dirichlet process models deal with the problem of modeling data that is divided into *groups* which share some common traits, e.g. data from counties in a given state. They are a flexible, non-parametric extension to the standard Bayesian parametric hierarchical models. Parametric (Bayesian) hierarchical models assume that the data distribution in each group $j = 1, ..., J$ has the *same form*, e.g. $Normal(\mu, \sigma^2)$, albeit it with different underlying parameters for each group, e.g. group dependent means $N(\mu_j, \sigma^2)$. The sharing of information in such models can be achieved by giving the group level distributions a common variance parameter, which is itself a random variable with a prior distribtion. The flexibility in HDP models comes from the fact that data distribution in each group is driven by a group-specific non-parametric Dirichlet process prior, thus allowing each group's distribution to take on a completely *different form*. The sharing of information across different groups comes about by making the group level distributions dependent on a common global measure, which is also driven by a Dirichlet process. We formalize this idea below.

The application that motivates our study of hierarchical Dirichlet processes is in the field of information retrieval (IR) : modeling text documents as mixtures of *topics*, i.e. *topic modeling*. In this scenario, each document (in a fixed collection of documents, i.e. a corpus) is treated as a "bag-of-words" : all the words in a given document are independent and exchangeable (this is obviously a very strong assumption, but it is nevertheless a standard one in the IR domain). We assume that the words in a document are generated from a number of latent mixture components or topics, and each topic is typically taken to be a (multinomial) distribution over a set of words from a finite and given vocabulary [4]. The goal is to discover the latent topics in a given set of documents and subsequently use them to generate a compact representation of each document in the corpus; this representation can be used to discover and/or cluster related documents. Mapping this to our HDP model, each document is a group of observations (words), and we model it as a mixture of topics (distributions); furthermore, we are allowed to share (re-use) topics among different groups (documents).

The report is organized as follows. Section 2.1 formally defines the HDP and our data model; section 3 discusses various representations of the HDP and examples of the prior distributions it induces; section 4 describes a Gibbs sampling procedure for inference in HDP models; section 5 describes our experimental setup and results; section 6 concludes the report.

2 Definitions

2.1 Hierarchical Dirichlet Processes

Equations (1) and (2) formally define the hierarchical Dirichlet process. Given a concentration parameter γ and a base measure H, the top level (global) measure G_0 is a draw from a Dirichlet process DP(γ, H). Given G_0, each

measure G_j in group $j = 1, ..., J$, is a draw from a DP(α_0, G_0), a Dirichlet process with concentration parameter α_0 and base measure G_0.

$$G_0 \mid \gamma, H \sim DP(\gamma, H) \tag{1}$$
$$G_j \mid \alpha_0, G_0 \sim DP(\alpha_0, G_0) \tag{2}$$

We can now observe how the sharing of information among groups comes about. Regardless of whether H is continuous or discrete, G_0 is discrete with probability one since it is a draw from a DP. This means that G_0 only has support at a (infinitely) countable set of locations, $\{\theta_k\}_{k=1}^{\infty}$. Since each G_j is a draw from a DP with base measure G_0, it is also discrete and it must have support at *exactly the same set of locations*. Thus, the individual groups $j = 1, ..., J$, have no choice but to share θ_k atoms.

We can contrast the above HDP formulation with a simpler hierarchical model where each G_j is a *conditionally independent* draw from a global Dirichlet process DP(α_0, $G_0(\tau)$), where $G_0(\tau)$ is a parametric distribution with random parameter τ. The sharing of information among groups is not possible in this model, for example, in the case where $G_0(\tau)$ is continuous. Since each G_j is by definition a conditionally independent (*not* i.i.d.) draw from DP(α_0, $G_0(\tau)$), the support each G_j has (the set of atoms $\{\theta_k\}_{k=1}^{\infty}$) will necessarily be different from group to group given the continuous nature of $G_0(\tau)$.

The HDP model is also a special case of the *analysis of densities* (AnDe) framework by Tomlinson and Escobar [8]. The AnDe model treats the global base measure G_0 as a draw from a *mixture of DPs* as opposed to a draw from a single DP. This produces a G_0 which is continuous in general, and therefore again does not permit the sharing of atoms among groups.

2.2 Data Model

Given the above definition of an HDP, we can now develop an *HDP mixture model* of data. In this setting, the observations (data) are organized into groups, and we assume that the observations are exchangeable within a group. We let $j = 1, ..., J$, index the J groups and we let $\boldsymbol{x}_j = (x_{ji})$, $i = 1, ..., n_j$, denote the n_j observations in group j. We also assume the $\boldsymbol{x}_j$, $j = 1, ..., J$, are exchageable at the group level. Each x_{ji} is drawn from a mixture model, whose composition (i.e. which exact components are mixed and in what proportion) is drawn once per group.

Equations (3) and (4) define our sampling distributions. The parameter $\boldsymbol{\phi}_j$ specifies the composition of the mixture model in group j, and thus each ϕ_{ji} specifies which particular mixture component is used to draw observation x_{ji}; the variables ϕ_{ji} can be thought of as "factors". Let $F(\phi_{ji})$ denote the distribution of x_{ji} given factor ϕ_{ji}. The prior for the factors $\boldsymbol{\phi}_j$ is G_j (equation (2)).

$$\phi_{ji} \mid G_j \sim G_j \tag{3}$$
$$x_{ji} \mid \boldsymbol{\phi}_j \sim F(\phi_{ji}) \tag{4}$$

To make the above model more concrete given our topic modeling scenario, suppose that we just have two documents ($J = 2$) with five words each ($n_j = 5$) sampled from a ten word vocabulary. Also, suppose that we have $K = 3$ underlying topics (mixture components), where each topic k specifies the parameters θ_k of a multinomial distribution F over the words in the vocabulary. For example, if $\boldsymbol{\phi}_1 = (2, 3, 3, 2, 1, 1)$ and $\boldsymbol{\phi}_2 = (2, 2, 3, 3, 2, 1)$, x_{11} will be a draw from a multinomial distribution with parameters θ_2, $x_{11} \sim F(\theta_2)$, x_{12} will be a draw from $F(\theta_3)$, etc. Note that in this example each document (group) has a different distribution (mixture composition) of the same three underlying topics. This is precisely the effect we're trying to model using the HDP mixture model. Loosely speaking, topics correspond to the locations of support (atoms) in a single realization of G_0, and the document specific measures G_j define the mixtures of topics in each document.

3 HDP Representations

We consider two representations of a hierarchical Dirichlet process (HDP) which will be useful for our analysis. The first is an extension of the stick-breaking construction developed by Sethuraman [6], and the second is an extension of the Polya urn sampling scheme knows as the Chinese Restaurant Franchise.

3.1 Stick Breaking Representation

The stick-breaking representation for the HDP starts out by observing that since $G_0 \sim DP(\gamma, H)$, it has the typical stick-breaking representation given by:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \qquad \theta_k \sim H \tag{5}$$

$$\beta_k^{'} \sim Beta(1, \gamma) \qquad \beta_k = \beta_k^{'} \prod_{l=1}^{k-1} (1 - \beta_l^{'}) \tag{6}$$

Furthermore, G_j is also distributed as a Dirichlet process, namely $G_j \sim DP(\alpha_0, G_0)$. Thus it also has it's own stick-breaking representation. However, as mentioned in section 2.1, G_0 is discrete with probability one, i.e. G_0 has support at a countable set of locations $\boldsymbol{\theta} = (\theta)_{i=1}^{\infty}$. Therefore, each G_j must also have support at this same set of locations, and we can write:

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k} \tag{7}$$

Going back to the original definition of a Dirichlet process as a probability measure on the space of probability measures, and letting $(A_1, ..., A_r)$ be a measurable partition of the sample space Θ, for each j we can write:

$$(G_j(A_1), ..., G_j(A_r)) \sim Dirichlet\left(\alpha_0 G_0(A_1), ..., \alpha_0 G_0(A_r)\right)$$

However, since both G_0 and G_j are discrete, each $G_j(A_l)$ and $G_0(A_l)$, $l = 1, ..., r$, is just the sum of the weights π_{jk} and β_k (respectively) that correspond to locations falling into the partition A_l, specifically:

$$\left(\sum_{k \in K_1} \pi_{jk}, ..., \sum_{k \in K_r} \pi_{jk}\right) \sim Dirichlet\left(\alpha_0 \sum_{k \in K_1} \beta_k, ..., \alpha_0 \sum_{k \in K_r} \beta_k\right), \qquad where\ K_l = \{k : \theta_k \in A_l\}$$

Given the above and the additive properties of the Dirichlet distribution, we can obtain the specific relationship between the two sets of weights β_k and π_{jk}:

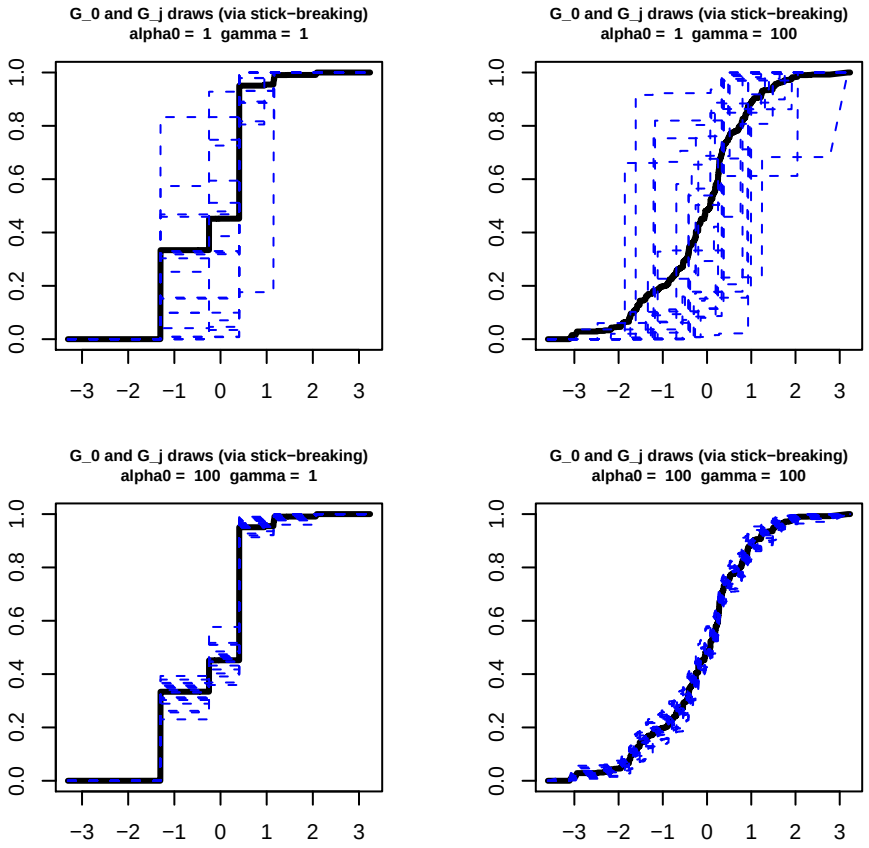
$$\pi_{jk}^{'} \sim Beta\left(\alpha_0 \beta_k, \alpha_0 \left(1 - \sum_{l=1}^{k} \beta_l\right)\right) \qquad \pi_{jk} = \pi_{jk}^{'} \prod_{l=1}^{k-1} (1 - \pi_{jl}^{'}) \tag{8}$$

Therefore, given a realization of G_0 with weights $\boldsymbol{\beta}$ and locations $\boldsymbol{\theta}$, we can quickly produce a number of realizations of G_j using those same locations $\boldsymbol{\theta}$ but with weights $\boldsymbol{\pi}_j$ that are a function of $\boldsymbol{\beta}$. Equation (8) shows that the weights $\boldsymbol{\pi}_j$ are dependent on the weights $\boldsymbol{\beta}$. For example, if γ is small, then only the first few $\boldsymbol{\beta}$ weights will be significant, and thus regardless of the value of α_0, only the first few weights of $\boldsymbol{\pi}_j$ will be significant; however, if γ is large, then many $\boldsymbol{\beta}$ weights will have an appreciable value, and how many $\boldsymbol{\pi}_j$ weights are significant will depend on the value of α_0.

Figure 1 illustrates a number of G_j draws (CDFs) given a G_0 draw for four combinations of concentration parameters α_0 and γ, with the global base measure $H = Normal(0, 1)$. We can review Figure 1 starting from the bottom-right, and going counter-clockwise:

- $\gamma = 100$, $\alpha_0 = 100$: Both concentration parameters are relatively large, and we can see the familiar shape of the N(0,1) CDF in G_0 (solid black line). Given the high value of α_0, we see that the G_j realizations (dashed blue lines) are tightly concentrated around G_0 and also approach the N(0,1) CDF.
- $\gamma = 100$, $\alpha_0 = 1$: Here we allow the G_j realizations to drift away from G_0 by making α_0 relatively small.
- $\gamma = 1$, $\alpha_0 = 1$: In this case, both parameters are relatively small. The G_0 realization is now "very discrete", i.e. it has support at only a handful of locations. α_0 is small, so the G_j realizations vary quite a bit, *but note that their support is exactly the support of* G_0.

Figure 1: Sample draws from an HDP prior with H = N(0,1). Solid black line : 1 realization of G_0. Dashed blue lines : 20 realizations of G_j.



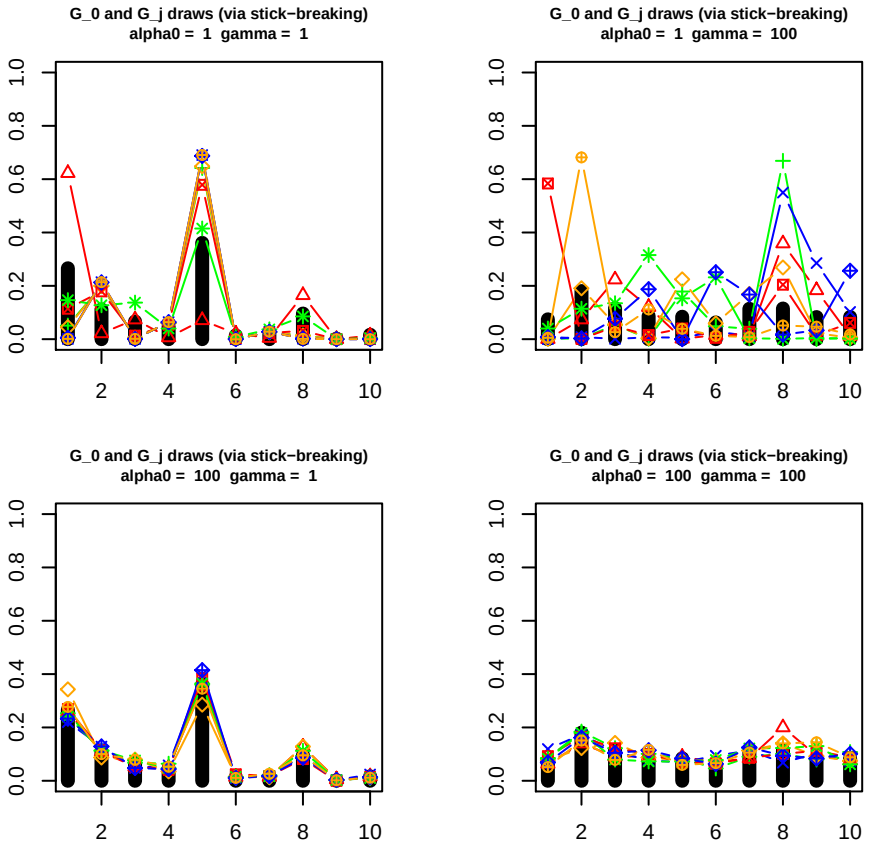
- $\gamma = 1$, $\alpha_0 = 100$: Here we have the same "very discrete" G_0 realization, but now with a large α_0. The G_j realizations are tightly concentrated around G_0.

Figure 2 also illustrates a number of G_j draws given a G_0 draw, but with a 10-dimensional global base measure $H = Dirichlet\,(0.1, 0.1, ..., 0.1)$, which is the H used in our experiments below (see section 5). The analysis is the same as in the case of $H = N(0, 1)$ above, so we will not repeat it, except to mention a few interesting points:

- The Dirichlet(0.1,0.1,...,0.1) distribution is a symmetric distribution, but a sparse one which places most of it's mass at a few of verteces of the 9-dimensional simplex, i.e. in a given draw, most of the indeces will have probabilities near 0, and only a few indeces will have appreciable values. Moreover, when $\gamma = 1$ only a few of the θ_k locations in the stick-breaking representation will have significant weights β_k (note: each θ_k is now a 10-dimensional vector of probabilities). This produces "non-uniform" probabilities in the final stick-breaking sum - see the heavy black lines in the left panels of figure 2. However, when $\gamma = 100$ many weights and independent locations contribute to the stick-breaking representation of G_0, and this has the effect of making the probabilities in the stick-breaking sum more "uniform" - see the heavy black lines in the right panels of figure 2.

- Again, note how large values of α_0 force G_j draws to be tightly clustered around G_0, whether its for a small or large value of γ.

Figure 2: Sample draws from HDP prior with H = Dirichlet(0.1,0.1,...,0.1), dim=10. Solid black lines : 1 realization of G_0. Colored lines and point : 10 realizations of G_j.



3.2 Polya Urn Sampling and the Chinese Restaurant Franchise

Draws from a hierarchical Dirichlet process can be obtained by an extension of the well known Polya urn scheme [3], in which the infinite dimensional DP has been integrated out. The Polya urn scheme for a single DP is related to a distribution on partitions known as the *Chinese Restaurant Process* (CRP) [1]. In the CRP metaphor, a new customer (observation) arriving at the (one) restaurant is seated at an existing occupied table t with probability proportional to n_t, the number of customers seated at that table. With probability proportional to α_0 (concentration parameter of the DP), he is seated at a previously unoccupied, newly allocated table. Each table has a distinct dish on it, and all customers at the same table share the single dish at that table. One can think of the distinct tables as corresponding to the distinct $\theta_k \sim DP(\cdot)$ draws in a Polya urn sampling scheme.

The single restaurant CRP metaphor can be extended to a multiple restaurant setting known as the *Chinese Resturant Franchise* (CRF). In this scenario, there are two Polya urn sampling schemes at work simultaneously : one for the tables and one for the dishes served at the tables. A customer arriving at restaurant j will be seated at a table based on the same Polya-urn sampling scheme outlined above for a single restaurant CRP. However, whereas in the CRP a new distinct table always meant a *new distinct* dish, here another Polya urn draw is made to select the dish for a new table. In the CRF, there exists a global menu of dishes shared among all restaurants, and a new table is assigned one of the existing dishes k with probability proportional to m_k, the number of tables currently serving dish k over all restaurants $j = 1, ..., J$. With probability proportional to γ, a new, previously unseen dish is created and assigned to the new table.

The formal CRF sampling equations for the HDP as defined in section 2.1 are shown below. The random

variables (factors) ϕ_{ji} correspond to customers (observations) and specify at which of the T_j tables in restaurant j a new customer x_{ji} will sit; from section 2.2, we recall that the ϕ_{ji} variables are distributed according to G_j. To simplify our analysis, we introduce the T_j random variables ψ_{jt} that correspond to tables in restaurant j. The ψ_{jt} are i.i.d. distributed according to G_0, and each ψ_{jt} specifies the mixture component for table jt. Finally, we have K random variables θ_k that correspond to dishes and specify the parameters of mixture component k; θ_k are i.i.d. distributed according H. Note that one or more ϕ_{ji} "map" to one ψ_{jt}, and that one or more ψ_{jt} "map" to one θ_k.

Given the above definitions, we can arrive at equation (9) by integrating out G_j (ala [3]) from equations (3) and (2), thus obtaining a Polya urn representation of the DP used to assign the i^{th} customer in restaurant j, ϕ_{ji}, to some table ψ_{jt} :

$$\phi_{ji}|\phi_{j1},...,\phi_{ji-1},\alpha_0,G_0 \sim \sum_{t=1}^{T_j} \frac{n_{jt}}{i-1+\alpha_0}\delta_{\psi_{jt}} + \frac{\alpha_0}{i-1+\alpha_0}G_0 \tag{9}$$

where n_{jt} is the number of (customers) ϕ_{ji}'s associated with (table) ψ_{jt}, and T_j is the total number of tables in restaurant j.

We obtain a Polya urn representation of the DP used to assign the t^{th} table to some mixture component θ_k by integrating out G_0. Since the ψ_{jt} draws from G_0 arise only from the top level $DP(\gamma, H)$, we can immediately write down the Polya urn sampler for the ψ_{jt} variables :

$$\psi_{jt}|\psi_{11},\psi_{12},...,\psi_{21},...,\psi_{jt-1},\gamma,H \sim \sum_{k=1}^{K} \frac{m_k}{\sum_k m_k+\gamma}\delta_{\theta_k} + \frac{\gamma}{\sum_k m_k+\gamma}H \tag{10}$$

where $m_k = \sum_j m_{jk}$, and m_{jk} is the number of (tables) ψ_{jt}'s associated with (mixture component) θ_k, and K is the total number of distinct θ_k's.

Sampling using (9) and (10) is straightforward. To obtain samples of $x_{ji} \sim F(\phi_{ji})$, first sample a value of ϕ_{ji} according to the proportions set out in (9). If a new table is required (RHS of (9)), then sample a mixture component for the new table according to the proportions set out in (10). If a new mixture component is required (RHS of (10)), draw the values for the new mixture component from H.

4 Inference

Given the Chinese Restaurant Franchise (CRF) sampling scheme outlined in section 3.2, we can implement an MCMC (Gibbs) routine for posterior sampling (inference) in an HDP model given a set of observations. First, we restate the definitions of all variables of interest. We have observations x_{ji} arising from a distribution $F(\phi_{ji})$, and let $F(\cdot)$ have density $f(\cdot)$. Each factor ϕ_{ji} is associated with the table t_{ji}, namely $\phi_{ji} = \psi_{jt_{ji}}$. Also, each ψ_{jt} specifies (is an instance of) the mixture component θ_k at table jt, namely $\psi_{jt} = \theta_{k_{jt}}$. The prior for θ_k is H with density $h(\cdot)$. The quantities n_{jt}, T_j, m_k, and K are defined in section 3.2 above. Finally, define the set of all observation-to-table assignments $\boldsymbol{t} = (t_{ji} : \text{all } j,i)$, the set of all table-to-component assignments $\boldsymbol{k} = (k_{jt} : \text{all } j,t)$, and the set of all distinct mixture component values $\boldsymbol{\theta} = (\theta_1,...,\theta_K)$. Note that while $\boldsymbol{\theta}$ contains actual values, $\boldsymbol{t}$ and $\boldsymbol{k}$ are simply sets of *index variables*. A "-" superscript attached to a set of variables refers to that particular set with the superscripted variable (index) removed.

The state of the sampler at any one point consists of the variables $(\boldsymbol{t}, \boldsymbol{k}, \boldsymbol{\theta})$ and the latest values of the concentration parameters γ and α_0. One iteration of the sampler consists of : (a) for all (j,i), sample t_{ji}, (b) for all (j,t), sample k_{jt}, (c) for all k, sample θ_k, and finally (d) update γ and α_0.

Sampling t. To sample a value for t_{ji} we need an expression for the conditional posterior for t_{ji} given the remainder of the variables. Therefore we need : (a) the conditional prior for t_{ji} and (b) the likelihood of generating x_{ji}. The conditional prior is just equation (9). The likelihood of x_{ji} given an existing $t_{ji} = t$ is just $f(x_{ji}|\theta_{k_{jt}})$; the likelihood given a new table $t_{ji} = t^{new}$ is $f(x_{ji}|\theta_{k_{jt^{new}}})$, where θ for the new table, $\theta_{k_{jt^{new}}}$, is drawn according to equation (10). Combining the conditional prior and likelihood we obtain the form of the conditional posterior as in equation (11) below.

$$p(t_{ji} = t|\boldsymbol{t}^{-ji}, \boldsymbol{k}, \boldsymbol{\theta}, \boldsymbol{x}) \propto \begin{cases} \alpha_0 f(x_{ji}|\theta_{k_{jt^{new}}}) & if\ t = t^{new} \\ n_{jt}^{-i} f(x_{ji}|\theta_{k_{jt}}) & if\ t\ is\ previously\ used \end{cases} \tag{11}$$

In our implementation of the sampler, we maintain a data structure of all table and component assignments represented by $(\boldsymbol{t}, \boldsymbol{k})$, as well as the component values $\boldsymbol{\theta}$. If after updating a particular table $n_{jt} = 0$, i.e. table t is now empty, we remove this table from our data structure. If as a result of removing table t, the mixture component associated with this table is no longer associated with any table, i.e. $m_{k_{jt}} = 0$, we also delete component k_{jt} from the data structure.

Sampling k. To sample a value for k_{jt}, we first draw a value for $\theta_k^{new} \sim H$. To arrive at an expression for the conditional posterior for k_{jt} given the rest of the variables, we start with the conditional prior given by equation (10). The data likelihood at table t given component k is given by $\prod_{S_t} f(x_{ji}|\theta_k)$, where S_t is the set of observations at table t, $S_t : \{i : t_{ji} = t\}$. Combining the prior and likelihood gives us the conditional posterior in equation (12) below.

$$p(k_{jt} = k|\boldsymbol{t}, \boldsymbol{k}^{-jt}, \boldsymbol{\theta}, \boldsymbol{x}) \propto \begin{cases} \gamma \prod_{x_{ji} \in S_t} f(x_{ji}|\theta_{k^{new}}) & if\ k = k^{new} \\ m_k^{-t} \prod_{x_{ji} \in S_t} f(x_{ji})|\theta_k) & if\ k\ is\ previously\ used \end{cases} \tag{12}$$

Sampling θ. The conditional posterior for each mixture component k only depends on the observations associated with component k. The prior density is given by $h(\theta_k)$, and the data likelihood by $\prod_{ji \in S_k} f(x_{ji}|\theta_k)$, where S_k is the set of observations (in all groups) associated with component k, $S_k : \{ji : k_{jt_{ji}} = k\}$.

$$p(\theta_k|\boldsymbol{t}, \boldsymbol{k}, \boldsymbol{\theta}_{-k}, \boldsymbol{x}) \propto h(\theta_k) \prod_{x_{ji} \in S_k} f(x_{ji}|\theta_k) \tag{13}$$

Sampling γ, α_0. Teh et al. also derive expressions for posterior sampling of the concentration parameters γ and α_0. We have implemented those routines in our sampler, but we do not review them in detail here. Posterior sampling for γ is identical to the auxiliary variable approach of Escobar and West [5]; posterior sampling for α_0 is a slight modification of the same basic approach.

MCMC inference based on the Chinese Restaurant Franchise is relatively simple to implement and understand. Furthermore, since (a) it updates the mixture component for a given table and thus for multiple observations simultaneously, and (b) it re-mixes the values of each component at each iteration, it may lead to better mixing and convergence. One can also speed up the above algorithm by integrating out θ_k in equations (11) and (12), and skip sampling θ altogether. Also the CRF approach is not the only possibility for MCMC sampling in HDP models. Another approach, based on the representation of the HDP as the infinite limit of finite mixture models, is detailed in [7]; however, the authors experimentally show that neither of the two sampling schemes consistently outperforms the other.

5 Experiments and Results

5.1 Simulated Data

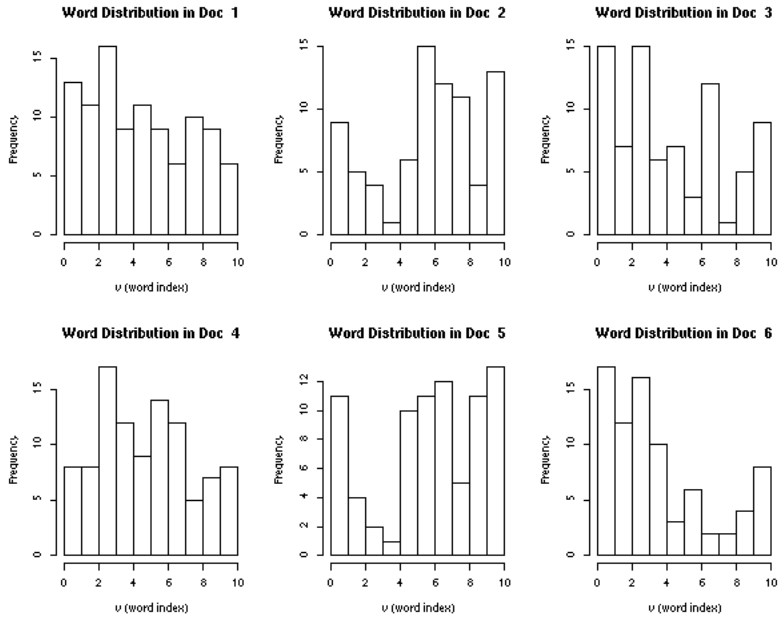
We experiment with inference in the topic modeling setting using a small set of simulated data in order to better understand the behavior of the model, and to measure how well it could recover the true data generating distributions (topics). We fixed our vocabulary size, V, at ten (10) unique words, and generated a data set consisting of $J = 6$ documents with $n_j = 80$ or 100 words arising from 2 - 3 mixture components per document. The procedure was as follows:

- We fixed the number of topics at three ($K = 3$) : θ_1, θ_2, θ_3. Each θ_k, $k = \{1, 2, 3\}$, is a parameter vector of a 10-dimensional multinomial distribution over the words (word indeces) in the vocabulary. Each θ_k was given a known fixed value for this experiment. Figure 4 shows the true values of each multinomial mixture component θ_k.
- We fixed the global topic proportions at $\pi = [0.4, 0.3, 0.3]$.
- For each document $j = 1, 2, ..., J$:
  - For each word $i = 1, 2, ..., n_j$:

* We sampled a topic (index) $\phi_{ji} \sim Multinomial(\pi_j^{'})$, where $\pi_j^{'} = \pi$ with one of it's values possibly set to 0, i.e. $\pi_{jk}^{'} \leftarrow 0$, to simulate documents composed of just two mixtures : documents 2 & 5 and 3 & 6 are only composed of mixture components (2,3) and (1,3), respectively.
* We sampled a word (index) $x_{ji} \sim Multinomial(\theta_{\phi_{ji}})$

The data (word) distribution in each simulated document is shown as a histogram in Figure 3.

Figure 3: Simulated data X. Actual word distributions in documents.



5.2 Posterior Estimates

We implemented a Gibbs sampler (in R) for posterior inference based on the Chinese Restaurant Franchise approach as described in section 4, including sampling for the concentration parameters γ and α_0. We ran for 1000 burn-in iterations and collected data for 4000 subsequent iterations. The priors for the concentration parameters were $\gamma \sim Gamma(2, 4)$ and $\alpha_0 \sim Gamma(1, 1)$, favoring smaller values of γ and α_0. Our global base distribution was a 10-dimensional symmetric Dirichlet distribution, namely H = Dirichlet(1/V,1/V,...,1/V), $V = 10$. The results are shown in figures 5, 6, 8, 9, and 10. The R code is available from www.numberjack.net/download/ucsc/classes/ams241/project/R.

Figure 5 shows the point estimates $\theta_k^{(est)}$ of the three components θ_k, $k = \{1, 2, 3\}$. Since we are using simulated data and we actually know what the truth is, i.e. for each x_{ji} we know the true $\phi_{ji} = k_{x_{ji}}$, we can "cheat" and compute each $\theta_k^{(est)}$ as follows:

- For each observation x_{ji}, compute the mean of the θ values associated with this observation over the B Gibbs sampling iterations, namely : $\theta_{ji}^{(\bar{B})} = \frac{1}{B} \sum_{b=1}^{B} \theta_{ji}^{(b)}$
- Compute $\theta_k^{(est)} = \frac{1}{|S_k|} \sum_{x_{ji} \in S_k}^{|S_k|} \theta_{ji}^{(\bar{B})}$, where set S_k is the set of observations x_{ji} whose true component (per the simulated data) is component k.

Comparing figures 4 and 5 (true values versus estimates, respectively), we see that inference was able to recover the relative proportions in the parameters of each component quite well : *within* each component, the estimated parameter values seem to be scaled (smoothed) versions of the true values, with "peaks" and "valleys" in the correct places. Furthermore, at each word index point v, we were also able to recover the relationship *between* different

Figure 4: True word distributions per topic θ_k, $k = \{1, 2, 3\}$ used to generate simulated data X.

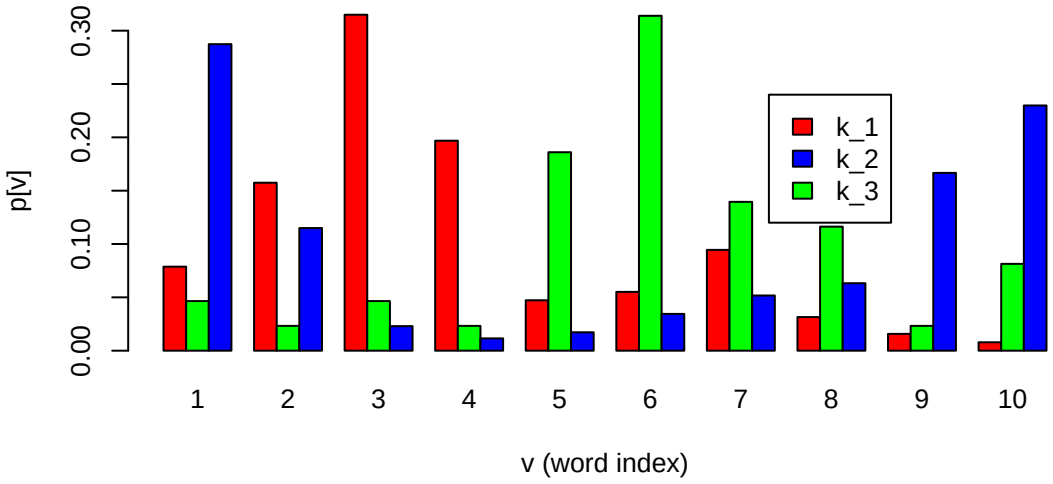
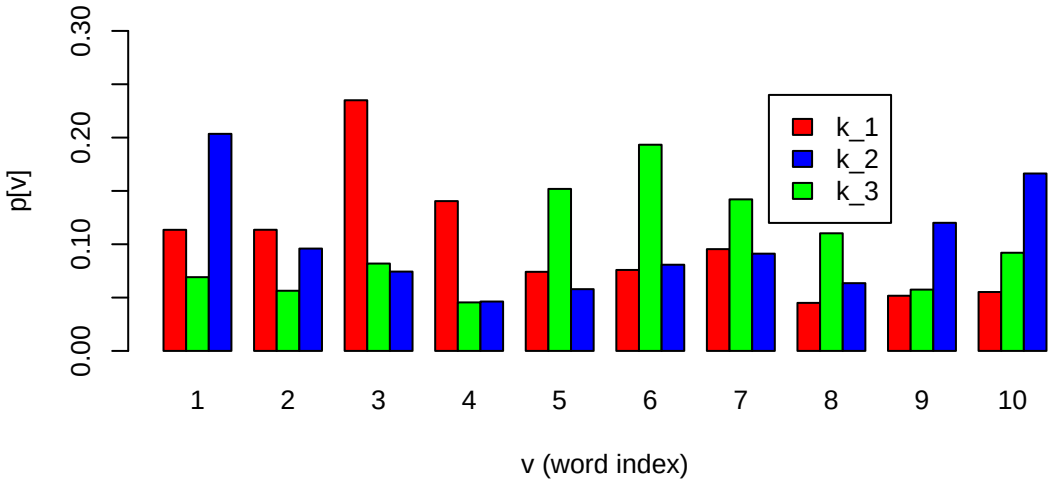


Figure 5: Estimated word distributions per topic θ_k, $k = \{1, 2, 3\}$, for $H = Dir(0.1, 0.1, ..., 0.1)$.

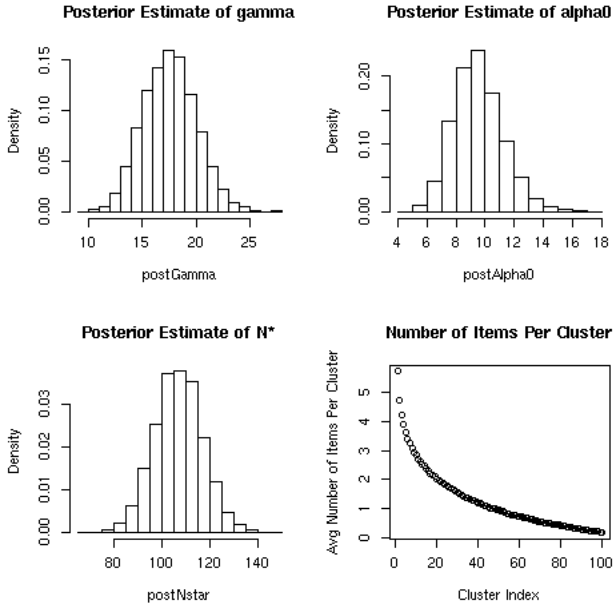


components : for each index point v, the component having the highest estimated parameter value at v is the same one which has the highest true parameter value at v. Of course had this been a real (non-simulated) data set, we would not have been able to do this kind of component comparison.

5.3 The Role of H

In figure 6 we plot the posterior distributions of concentration parameters γ, and α_0. We also show the distribution of the number of unique components N^* and the average number of observations per component. We see that the expected number of components is quite high, over 100, and on average (over the MCMC iterations), the "most populous" component only has approximately 5-6 observations associated with it. These results are in line with the large mean posterior value of γ.

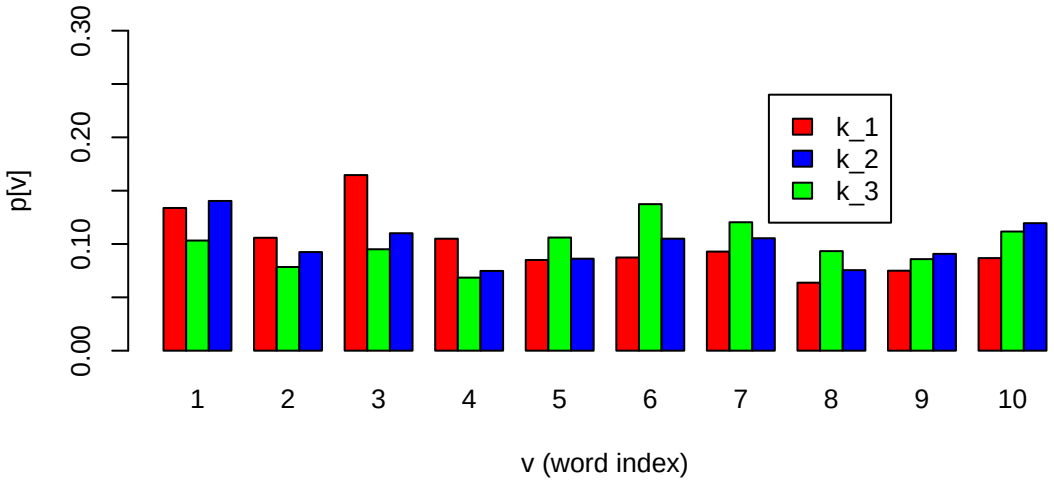
Figure 6: Posterior distributions of concentration parameters γ, and α_0, as well as the distribution of the number of unique components N^* and the average number of observations per component.



This may be explained by looking at the form of our global base distribution $H = Dir(1/V, ..., 1/V)$, where $V = 10$. As mentioned in section 3, this type of symmetric Dirichlet distribution (of dimenstion V) is "sparse", i.e. it places most of it's mass at a few of the verteces of the $V-1$ dimensional simplex, and draws from it exhibit little uniformity in their values. Therefore, in order to adequately model (support) a given document's actual mixture of topics - the more "uniform" true θ_j's pictured as black bars in figure 8 - a large number of such "non-uniform" components are required. However knowing this, suppose we try to decrease the number of components by making the draws from H more "uniform", specifically let $H = Dir(1, 1, ..., 1)$. The problem we run into is since the Dirichlet distribution is conjugate to our Multinomial sampling distribution, the parameters of H effectively act as *prior data sample sizes*, and larger parameters (e.g. 1 vs. $1/V$) imply more strength to the prior H than to the data, and the posterior differences between components are smoothed out - compare figure 7 with our previous estimates in figure 5. The only choice now is to get more data, which is not typically very feasible or even possible.

Figure 8 shows the point and [10%,90%] interval estimates of the *posterior mean* of $p(\theta_j|data)$, namely $\theta_j^{(est)} = E[\theta_j|data]$, as well as the true values of θ_j computed as $\theta_j = \sum_{k=1}^{K} \pi_{jk}^{'} \theta_k$ (see section 5.1). The $\theta_j^{(est)}$ values are computed in similar fashion to the $\theta_k^{(est)}$ values described above, i.e. $\theta_j^{(est)} = \frac{1}{n_j} \sum_{i=1}^{n_j} \theta_{ji}^{(\bar{B})}$, where $\theta_{ji}^{(\bar{B})}$ is the mean of the θ draws associated with observation x_{ji} over the B Gibbs sampling iterations. In addition to the point estimate of the mean, we also compute and plot the posterior [10%,90%] intervals of the mean as dashed lines in Figure 8. Figure 8 shows that the posterior interval estimates do a good - but not great - job of recovering the true θ_j values : in each document, there is at least one true $\theta_j[v]$ value that lies outside the posterior [10%,90%] interval for the given word index v (e.g. see $v = 1$ in document 2); nevertheless, the majority of estimates do cover the true values.

Figure 7: Estimated word distributions per topic θ_k, $k = \{1, 2, 3\}$, for $H = Dir(1, 1, ..., 1)$.



5.4 Posterior Predictive Estimates

Given our B posterior samples from $p(\boldsymbol{t}, \boldsymbol{k}, \boldsymbol{\theta}|data)$ associated with each document j, we can obtain an estimate of the mean (expected value) of the posterior predictive distribution $p(\theta_{j0}|data)$ - the predictive distribution according to which a new observation x_{j0} will be drawn as $x_{j0} \sim Multinomial(\theta_{j0})$.

We estimate $E[\theta_{j0}|data]$ as follows. First, we note that at each iteration b of the sampler, the state of the sampler reflects a unique configuration $(\boldsymbol{t}^{(b)}, \boldsymbol{k}^{(b)}, \boldsymbol{\theta}^{(b)}, \gamma^{(b)}, \alpha_0^{(b)})$ of a Chinese Restaurant Franchise with data-to-table assignments given by $\boldsymbol{t}^{(b)}$, table-to-component assignments given by $\boldsymbol{k}^{(b)}$, and distinct component values given by $\boldsymbol{\theta}^{(b)}$. Therefore at each iteration b, we can sample a predictive $\theta_{j0}^{(b)}$ value as outlined in section 3.2 using the appropriate values of $n_{jt}^{(b)}, m_k^{(b)}, \alpha_0^{(b)}, \gamma^{(b)}, \boldsymbol{\theta}^{(b)}$. Specifically, for each iteration b per document j, draw a new value for $\theta_{j0}^{(b)}$ corresponding to a new observation x_{j0} as follows. First, draw a new table assignment $\phi_{j0}^{(b)}$ according to:

$$\phi_{j0}^{(b)} \sim \sum_{t=1}^{T_j^{(b)}} \frac{n_{jt}^{(b)}}{n_j + \alpha_0^{(b)}} \delta_{\psi_{jt}^{(b)}} + \frac{\alpha_0^{(b)}}{n_j + \alpha_0^{(b)}} G_0^{(b)}$$

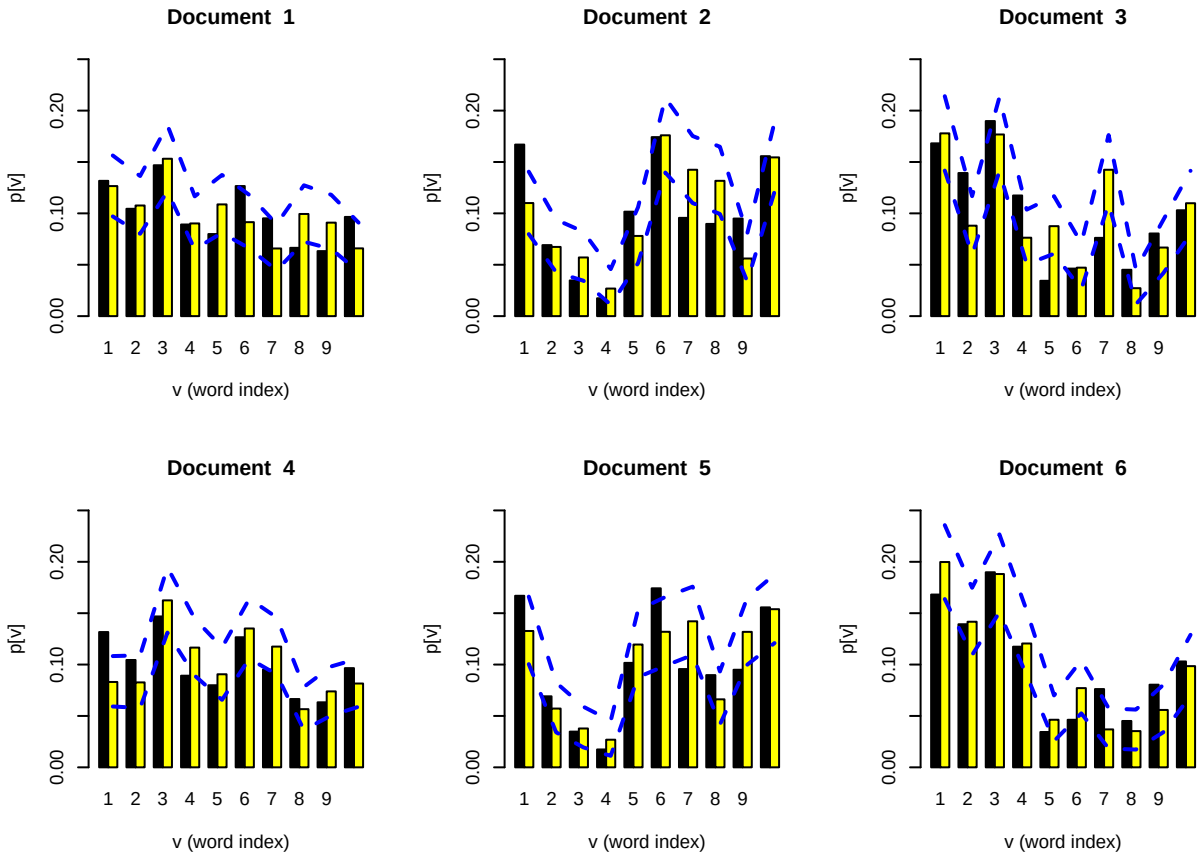
If the table assignment ends up being one of the existing tables, we're done : $\theta_{j0}^{(b)}$ will assume the value of the mixture component corresponding to this table. If the table assignment ends up being a *new* table, we draw a value for $\theta_{j0}^{(b)}$ from $G_0^{(b)}$:

$$\theta_{j0}^{(b)} \sim G_0^{(b)}(.) = \sum_{k=1}^{K^{(b)}} \frac{m_k^{(b)}}{\sum_k m_k^{(b)} + \gamma^{(b)}} \delta_{\theta_k^{(b)}} + \frac{\gamma^{(b)}}{\sum_k m_k^{(b)} + \gamma^{(b)}} H$$

Note that the value of $\theta_{j0}^{(b)}$ for the new table could come from an existing component (with probability proportional to $m_k^{(b)}$), or it could be a completely new draw from H (with probability proportional to $\gamma^{(b)}$).

Finally, we average over the B draws of $\theta_{j0}^{(b)}$ to compute $E[\theta_{j0}|data]$, and show the results in Figure 9, where we compare $E[\theta_{j0}|data]$ with the actual simulated data distribution (as a normalized histogram). Figure 9 shows that

Figure 8: True θ_j (black) vs. point and interval estimates $\theta_j^{(est)} = E[\theta_j|data]$ (yellow), $j = \{1, ..., 6\}$.



in each document j, the values of $E[\theta_{j0}|data]$ at word indeces $v = 1, .., 10$ - which are effectively parameters of a $Multinomial(\theta_{j0})$ distribution - line up well with the actual data generating distribution in that document.

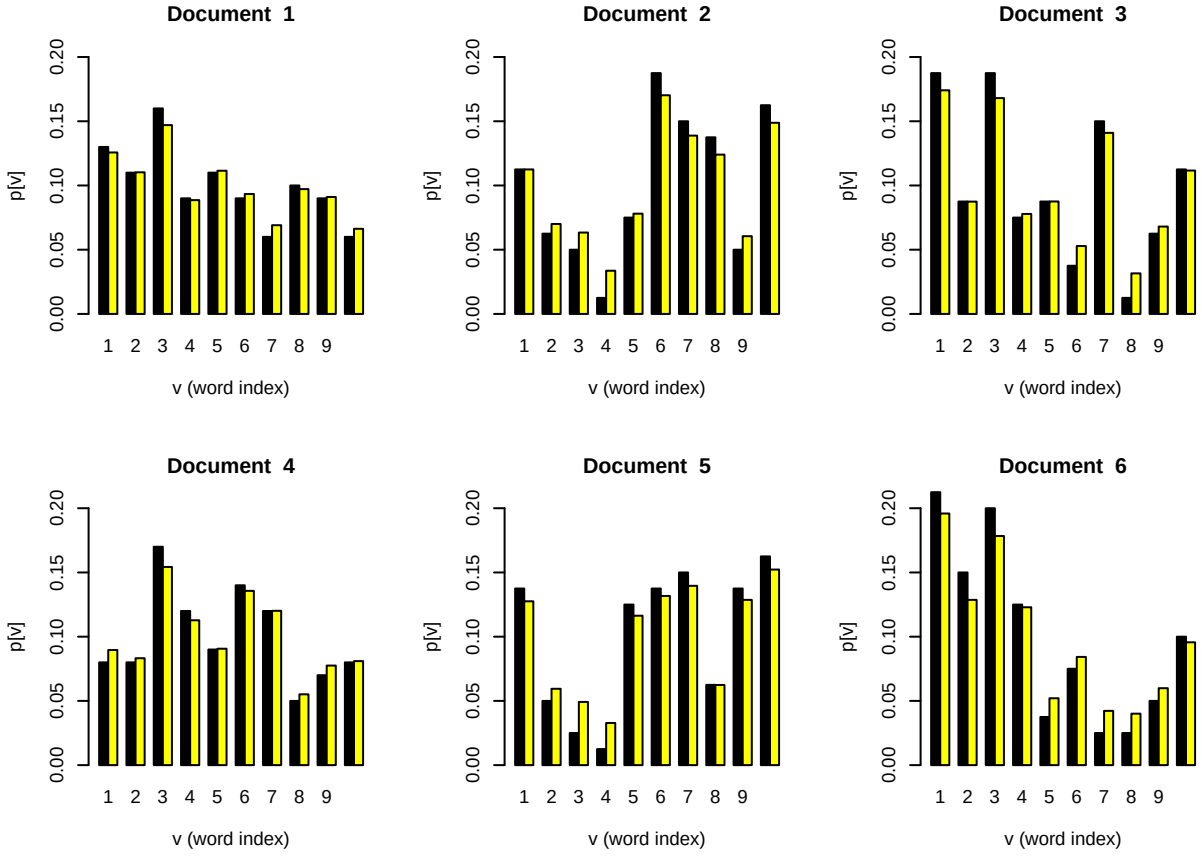
Furthermore, having already obtained B $\theta_{j0}^{(b)}$ draws above, we can now sample a new data value $x_{j0}^{(b)} \sim Multinom(\theta_{j0}^{(b)})$ at each iteration b. We use these B $x_{j0}^{(b)}$ draws to produce a density estimate of the posterior predictive distribution for a new observation in each document : $p(x_{j0}|data)$ - this is shown figure 10 (red curve) along with the density estimates of the data generating distribution in each document (black curve). Figure 10 shows that in each document, the shape of $p(x_{j0}|data)$ compares well with the actual (simulated) data density. The "peaks" and "valleys" of $p(x_{j0}|data)$ are at the expected locations and with a few exceptions, line up well with the data density. The largest discrepancy seems to be that the predictive density estimates have sharper peaks (e.g. documents 3 and 5) than their data density counterparts - this can be attributed to two factors : one, the strong peaks present in $E[\theta_{j0}|data]$ (see figure 9) from which x_{j0} are drawn, and two, the large number (1000) of x_{j0} samples taken.

6 Conclusion

In this report we reviewed a number of ideas presented by Teh et al. in their work on hierarchical Dirichlet process. In addition, we implemented our own MCMC based sampler for HDP inference, and used it to study a small simulated data set. This report is by no means the complete story with respect to hierarchical Dirichlet processes, but it does explain the fundamental concepts in detail and provides a number of examples to illustrate the theory.

Hierarchical Dirichlet processes provide a suitable approach to clustering problems in grouped data, where the

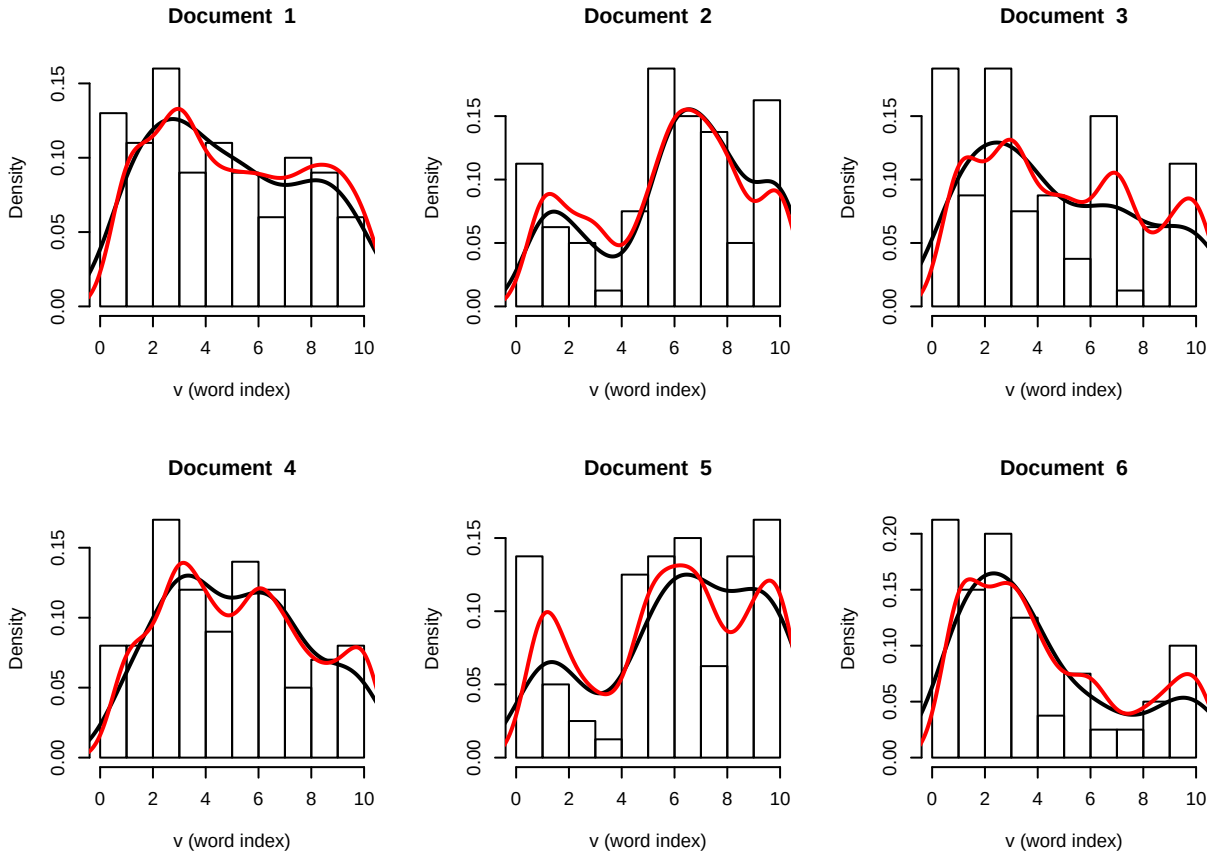
Figure 9: Normalized data histogram (black) vs. $E[\theta_{j0}|data]$ (yellow), per document $j = \{1, ..., 6\}$.



number of clusters is not known nor specified ahead of time, and where we would like to share clusters among the groups of data. The use of a Dirichlet process at the top level of the model specification frees us from apriori specifying the number of clusters (components in a mixture model) - we can infer the distribution for that quantity from the data. The use of group-specific Dirichlet processes at the next level of the model - which have as their base measure a draw from the top level DP - enables the sharing of components among the groups.

HDP models seem to be a natural fit to topic modeling in the IR domain. It is not difficult to imagine that in a corpus of documents, there is a large but countable set of topics, that each document may be a different mixture of one or more topics, and that the topics are shared among all documents in the corpus. HDP inference in our small simulated data set with distinct topics worked out well. However, trouble creeps in when one tries to actually use the inferred topics in isolation (something we did not explore in this report); specifically, the issue of how *granular* should a topic be? For example, one set of parameters in the HDP model may yield very "broad" topics that refer to "sports", "politics", etc., whereas a different set of parameters may yield very "specific" topics that refer to "baseball", "tennis", etc. The values of the parameters will be application dependent, but it is not clear how to set them (or their priors) to achieve the desired level of granularity. Finally, inference for HDP models in this scenario is straightforward given the Chinese Restaurant Franchise and the conjugacy inherent in the data model. However, Gibbs sampling as described above may be too slow given a real world document modeling scenario, where one has 100,000+ documents and a vocabulary of 20,000+ words. Nevertheless, it remains an intriguing model for such hierarchical modeling problems.

Figure 10: Data density and histogram (black) vs. density estimate of predictive distribution (red) for new word $x_{j0} \sim p(x_{j0}|data)$.



References

[1] Aldous, D. *Exchangeability and Related Topics*. Ecole d'Ete de Probilites de Saint-Flour XIII-1983, Springer, Berlin, pp.1-198, 1985.

[2] Antoniak, C. *Mixtures of Dirichler Processes with Applications to Bayesian Nonparametric Problems*. Annals of Statistics, 2(6), pp. 1152-1174, 1974.

[3] Blackwell, D. and MacQueen, J. *Ferguson Distributions via Polya Urn Schemes*. Annals of Statistics, 1, pp. 353-355, 1973.

[4] Blei, D., Jordan, M., and Ng, A. *Hierarchical Bayesian Models for Applications in Informations Retrieval*. Bayesian Statistics, vol. 7, pp 25-44, 2003.

[5] Escobar, M. and West, M. *Bayesian Density Estimation and Inference Using Mixtures*. JASA, 90, pp. 577-588, 1995.

[6] Sethuraman, J. *A Constructive Definition of Dirichlet Priors*. Statistica Sinica, 4, pp. 639-650, 1994.

[7] Teh, Y., Jordan, M., Beal, M., and Blei, D. *Hierarchical Dirichlet Processes*, Technical Report 653, Department of Statistics, University of California at Berkeley, 2004.

[8] Tomlinson, G. and Escobar, M. *Analysis of Densities*. Technical Report.

[9] Salton, G. and McGill, M. *An Introduction to Modern Information Retrieval*, New York, McGraw-Hill, 1983.