An Alternative Prior Process for Nonparametric Bayesian Clustering

Hanna M. Wallach
Department of Computer Science
University of Massachusetts Amherst

Shane T. Jensen
Department of Statistics
The Wharton School, University of Pennsylvania

Lee Dicker
Department of Biostatistics
Harvard School of Public Health

Katherine A. Heller
Engineering Department
University of Cambridge

Abstract

Prior distributions play a crucial role in Bayesian approaches to clustering. Two commonly-used prior distributions are the Dirichlet and Pitman-Yor processes. In this paper, we investigate the predictive probabilities that underlie these processes, and the implicit "rich-get-richer" characteristic of the resulting partitions. We explore an alternative prior for nonparametric Bayesian clustering—the uniform process—for applications where the "rich-get-richer" property is undesirable. We also explore the cost of this process: partitions are no longer exchangeable with respect to the ordering of variables. We present new asymptotic and simulation-based results for the clustering characteristics of the uniform process and compare these with known results for the Dirichlet and Pitman-Yor processes. We compare performance on a real document clustering task, demonstrating the practical advantage of the uniform process despite its lack of exchangeability over orderings.

1 Introduction

Nonparametric Bayesian models provide a powerful and popular approach to many difficult statistical problems, including document clustering (Zhang *et al.*, 2005), topic modeling (Teh *et al.*, 2006b), and clustering motifs in DNA sequences (Jensen and Liu, 2008). The key assumption underlying nonparametric Bayesian models is the existence of a set of random variables drawn from some unknown probability distribution. This unknown probability distribution is itself drawn from some prior distribution. The Dirichlet process is one such prior for unknown probability distributions that has become ubiquitous in Bayesian nonparametric modeling, as reviewed by Muller and Quintana (2004). More recently, Pitman and Yor (1997) introduced the Pitman-Yor process, a two-parameter generalization of the Dirichlet process. These processes can also be nested within a hierarchical structure (Teh *et al.*, 2006a; Teh, 2006). A key property of any model based on Dirichlet or Pitman-Yor processes is that the posterior distribution provides a partition of the data into clusters, without requiring that the number of clusters be pre-specified in advance. However, previous work on nonparametric Bayesian clustering has paid little attention to the implicit *a priori* "rich-get-richer" property imposed by both the Dirichlet and Pitman-Yor process. As we explore in section 2, this property is a fundamental characteristic of partitions generated by these processes, and leads to partitions consisting of a small number of large clusters, with "rich-get-richer" usage. Although "rich-get-richer" cluster usage is appropriate for some clustering applications, there are others for which it is undesirable. As pointed out by Welling (2006), there exists a need for alternative priors in clustering models.

In this paper, we explore one such alternative prior—the *uniform process*—which exhibits a very different set of clustering characteristics to either the Dirichlet process or the Pitman-Yor process. The uniform process was originally introduced by Qin *et al.* (2003) (page 438) as an *ad hoc* prior for DNA motif clustering. However, it has received little attention in the subsequent statistics and machine learning literature and its clustering characteristics have remained largely unex-

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

plored. We therefore compare the uniform process to the Dirichlet and Pitman-Yor processes in terms of asymptotic characteristics (section 3) as well as characteristics for sample sizes typical of those found in real clustering applications (section 4). One fundamental difference between the uniform process and the Dirichlet and Pitman-Yor processes is the uniform process's lack of exchangeability over cluster assignments—the probability $P(\boldsymbol{c})$ of a particular set of cluster assignments $\boldsymbol{c}$ is not invariant under permutations of those assignments. Previous work on the uniform process has not acknowledged this issue with respect to either inference or probability calculations. We demonstrate that this lack of exchangeability is not a significant problem for applications where a more balanced prior assumption about cluster sizes is desired. We present a new Gibbs sampling algorithm for the uniform process that is correct for a fixed ordering of the cluster assignments, and show that while $P(\boldsymbol{c})$ is not invariant to permuted orderings, it can be highly robust.

We also consider the uniform process in the context of a real text processing application: unsupervised clustering of a set of documents into natural, thematic groupings. An extensive and diverse array of models and procedures have been developed for this task, as reviewed by Andrews and Fox (2007). These approaches include nonparametric Bayesian clustering using the Dirichlet process (Zhang *et al.*, 2005) and the hierarchical Dirichlet process (Teh *et al.*, 2006a). Such nonparametric models are popular for document clustering since the number of clusters is rarely known *a priori*, and these models allow the number of clusters to be inferred along with the assignments of documents to clusters. However, as we illustrate below, the Dirichlet process still places prior assumptions on the clustering structure: partitions will typically be dominated by a few very large clusters, with overall "rich-get-richer" cluster usage. For many applications, there is no *a priori* reason to expect that this kind of partition is preferable to other kinds of partitions, and in these cases the uniform process can be a better representation of prior beliefs than the Dirichlet process. We demonstrate that the uniform process leads to superior document clustering performance (quantified by the probability of unseen held-out documents under the model) over the Dirichlet process using a collection of carbon nanotechnology patents (section 6).

2 Predictive Probabilities for Clustering Priors

Clustering involves partitioning random variables $\boldsymbol{X} = (X_1, \ldots, X_N)$ into clusters. This procedure is often performed using a mixture model, which assumes that each variable was generated by one of K mixture components characterized by parameters $\Phi = \{\phi_k\}_{k=1}^K$:

$$P(X_n \mid \Phi) = \sum_{k=1}^{K} P(c_n = k)\, P(X_n \mid \phi_k, c_n = k), \qquad (1)$$

where c_n is an indicator variable such that $c_n = k$ if and only if data point X_n was generated by component k with parameters ϕ_k. Clustering can then be characterized as identifying the set of parameters responsible for generating each observation. The observations associated with parameters ϕ_k are those X_n for which $c_n = k$. Together, these observations form cluster k. Bayesian mixture models assume that the parameters Φ come from some prior distribution $P(\Phi)$. Nonparametric Bayesian mixture models further assume that the probability that $c_n = k$ is well-defined in the limit as $K \to \infty$. This allows for more flexible mixture modeling, while avoiding costly model comparisons in order to determine the "right" number of clusters or components K. From a generative perspective, in nonparametric Bayesian mixture modeling, each observation is assumed to have been generated by first selecting a set of component parameters ϕ_k from the prior and then generating the observation itself from the corresponding component. Clusters are therefore constructed sequentially. The component parameters responsible for generating a new observation are selected using the *predictive probabilities*—the conditional distribution over component parameters implied by a particular choice of priors over Φ and c_n. We next describe three priors—the Dirichlet, Pitman-Yor, and uniform processes—using their predictive probabilities. For notational convenience we define ψ_n to be the component parameters for the mixture component responsible for observation X_n, such that $\psi_n = \phi_k$ when $c_n = k$.

2.1 Dirichlet Process

The Dirichlet process prior has two parameters: a *concentration parameter* θ, which controls the formation of new clusters, and a *base distribution* G_0. Under a Dirichlet process prior, the conditional probability of the mixture component parameters ψ_{N+1} associated with a new observation X_{N+1} given the component parameters $\psi_1, \ldots, \psi_N$ associated with previous observations $X_1, \ldots, X_N$ is a mixture of point masses at the locations of $\psi_1, \ldots, \psi_N$ and the base distribution G_0. Variables X_n and X_m are said to to belong to the same cluster if and only if $\psi_n = \psi_m$.[1] This predictive probability formulation therefore sequentially constructs a partition, since observation X_{N+1} belongs to an existing cluster if $\psi_{N+1} = \psi_n$ for some $n \leq N$ or a new cluster consisting only of X_{N+1} if ψ_{N+1} is drawn directly from G_0. If $\phi_1, \ldots, \phi_K$ are the K distinct values

[1] Assuming a continuous G_0.

in $\psi_1, \ldots, \psi_N$ and $N_1, \ldots, N_K$ are the corresponding cluster sizes (*i.e.*, $N_k = \sum_{n=1}^{N} \mathrm{I}(\psi_n = \phi_k)$, then

$$P(\psi_{N+1} \,|\, \psi_1, \ldots, \psi_N, \theta, G_0) = \begin{cases} \frac{N_k}{N+\theta} & \psi_{N+1} = \phi_k \in \{\phi_1, \ldots, \phi_K\} \\ \frac{\theta}{N+\theta} & \psi_{N+1} \sim G_0. \end{cases} \quad (2)$$

New observation X_{N+1} joins existing cluster k with probability proportional to N_k (the number of previous observations in that cluster) and joins a new cluster, consisting of X_{N+1} only, with probability proportional to θ. This predictive probability is evident in the *Chinese restaurant process* metaphor (Aldous, 1985).

The most obvious characteristic of the Dirichlet process predictive probability (given by (2)) is the "rich-get-richer" property: the probability of joining an existing cluster is proportional to the size of that cluster. New observations are therefore more likely to join already-large clusters. The "rich-get-richer" characteristic is also evident in the *stick-breaking* construction of the Dirichlet process (Sethuraman, 1994; Ishwaran and James, 2001), where each unique point mass is assigned a random weight. These weights are generated as a product of Beta random variables, which can be visualized as breaks of a unit-length stick. Earlier breaks of the stick will tend to lead to larger weights, which again gives rise to the "rich-get-richer" property.

2.2 Pitman-Yor Process

The Pitman-Yor process (Pitman and Yor, 1997) has three parameters: a concentration parameter θ, a base distribution G_0, and a *discount parameter* $0 \leq \alpha < 1$. Together, θ and α control the formation of new clusters. The Pitman-Yor predictive probability is

$$P(\psi_{N+1} \,|\, \psi_1, \ldots, \psi_N, \theta, \alpha, G_0) = \begin{cases} \frac{N_k - \alpha}{N+\theta} & \psi_{N+1} = \phi_k \in \{\phi_1, \ldots, \phi_K\} \\ \frac{\theta + K\alpha}{N+\theta} & \psi_{N+1} \sim G_0. \end{cases} \quad (3)$$

The Pitman-Yor process also exhibits the "rich-get-richer" property. However, the discount parameter α serves to reduce the probability of adding a new observation to an existing cluster. This prior is particularly well-suited to natural language processing applications (Teh, 2006; Wallach *et al.*, 2008) because it yields power-law behavior (cluster usage) when $0 < \alpha < 1$.

2.3 Uniform Process

Predictive probabilities (2) and (3) result in partitions that are dominated by a few large clusters, since new observations are more likely to be assigned to larger clusters. For many tasks, however, a prior over partitions that induces more uniformly-sized clusters is desirable. The uniform process (Qin *et al.*, 2003; Jensen and Liu, 2008) is one such prior. The predictive probability for the uniform process is given by

$$P(\psi_{N+1} \,|\, \psi_1, \ldots, \psi_N, \theta, G_0) = \begin{cases} \frac{1}{K+\theta} & \psi_{N+1} = \phi_k \in \{\phi_1, \ldots, \phi_K\} \\ \frac{\theta}{K+\theta} & \psi_{N+1} \sim G_0. \end{cases} \quad (4)$$

The probability that new observation X_{N+1} joins one of the existing K clusters is uniform over these clusters, and is unrelated to the cluster sizes. Although the uniform process has been used previously for clustering DNA motifs (Qin *et al.*, 2003; Jensen and Liu, 2008), its usage has otherwise been extremely limited in the statistics and machine learning literature and its theoretical properties have thus-far not been explored.

Constructing prior processes using predictive probabilities can imply that the underlying prior results in nonexchangeability. If $\boldsymbol{c}$ denotes a partition or set of cluster assignments for observations $\boldsymbol{X}$, then the partition is exchangeable if the calculation of the full prior density of the partition $P(\boldsymbol{c})$ via the predictive probabilities is unaffected by the ordering of the cluster assignments. As discussed by Pitman (1996) and Pitman (2002), most sequential processes will fail to produce a partition that is exchangeable. The Dirichlet process and Pitman-Yor process predictive probabilities ((2) and (3)) both lead to exchangeable partitions. In fact, their densities are special cases of "exchangeable partition probability functions" given by Ishwaran and James (2003). Green and Richardson (2001) and Welling (2006) discuss the relaxation of exchangeability in order to consider alternative prior processes. The uniform process does not ensure exchangeability: the prior probability $P(\boldsymbol{c})$ of a particular set of cluster assignments $\boldsymbol{c}$ is not invariant under permutation of those cluster assignments. However, in section 5, we demonstrate that the nonexchangeability implied by the uniform process is not a significant problem for real data by showing that $P(\boldsymbol{c})$ is robust to permutations of the observations and hence cluster assignments.

3 Asymptotic Behavior

In this section, we compare the three priors implied by predictive probabilities (2), (3) and (4) in terms of the asymptotic behavior of two partition characteristics: the number of clusters K_N and the distribution of cluster sizes $\boldsymbol{H}_N = (H_{1,N}, H_{2,N}, \ldots, H_{N,N})$ where $H_{M,N}$ is the number of clusters of size M in a partition of N observations. We begin by reviewing previous results for the Dirichlet and Pitman-Yor processes, and then present new results for the uniform process.

3.1 Dirichlet Process

As the number of observations $N \to \infty$, the expected number of unique clusters K_N in a partition is

$$\mathbb{E}\left(K_N \,|\, \mathrm{DP}\right) = \sum_{n=1}^{N} \frac{\theta}{n-1+\theta} \simeq \theta \log N. \qquad (5)$$

The expected number of clusters of size M is

$$\lim_{N\to\infty} \mathbb{E}\left(H_{M,N} \,|\, \mathrm{DP}\right) = \frac{\theta}{M}. \qquad (6)$$

This well-known result (Arratia *et al.*, 2003) implies that as $N \to \infty$, the expected number of clusters of size M is inversely proportional to M regardless of the value of θ. In other words, in expectation, there will be a small number of large clusters and *vice versa*.

3.2 Pitman-Yor Process

Pitman (2002) showed that as $N \to \infty$, the expected number of unique clusters K_N in a partition is

$$\mathbb{E}\left(K_N \,|\, \mathrm{PY}\right) \approx \frac{\Gamma(1+\theta)}{\alpha\Gamma(\alpha+\theta)} N^{\alpha}. \qquad (7)$$

Pitman's result can also be used to derive the expected number of clusters of size M in a partition:

$$\mathbb{E}\left(H_{M,N} \,|\, \mathrm{PY}\right) \approx \frac{\Gamma(1+\theta)\prod_{m=1}^{M-1}(m-\alpha)}{\Gamma(\alpha+\theta)\,M!} N^{\alpha}. \qquad (8)$$

3.3 Uniform Process

Previous literature on the uniform process does not contain any asymptotic results. We therefore present the following novel result for the expected number of unique clusters K_N in a partition as $N \to \infty$:

$$\mathbb{E}\left(K_N \,|\, \mathrm{UP}\right) \approx \sqrt{2\theta} \cdot N^{\frac{1}{2}}. \qquad (9)$$

A complete proof is given in the supplementary materials. In section 4, we also present simulation-based results that suggest the following conjecture for the expected number of clusters of size M in a partition:

$$\mathbb{E}\left(H_{M,N} \,|\, \mathrm{UP}\right) \approx \theta. \qquad (10)$$

This result corresponds well to the intuition underlying the uniform process: observations are *a priori* equally likely to join any existing cluster, regardless of size.

3.4 Summary of Asymptotic Results

The distribution of cluster sizes for the uniform process is dramatically different to that of either the Pitman-Yor or Dirichlet process, as evidenced by the results above, as well as the simulation-based results in section 4. The uniform process exhibits a uniform distribution of cluster sizes. Although the Pitman-Yor process can be made to behave similarly to the uniform process in terms of the expected number of clusters (by varying α, as described below), it cannot be configured to exhibit a uniform distribution over cluster sizes, which is a unique aspect of the uniform process.

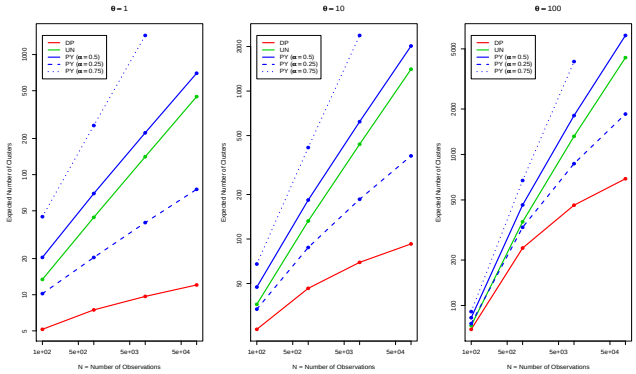

Figure 1: Expected number of clusters $\hat{K}_N$ versus sample size N for different θ. Axes are on a log scale.

Under the Dirichlet process, the expected number of clusters grows logarithmically with the number of observations N. In contrast, under the uniform process, the expected number of clusters grows with the square root of the number of observations N. The Pitman-Yor process implies that the expected number of clusters grows at a rate of N^{α}. In other words, the Pitman-Yor process can lead to a slower or faster growth rate than the uniform process, depending on the value of the discount parameter α. For $\alpha = 0.5$, the expected number of clusters grows at the same rate for both the Pitman-Yor process and the uniform process.

4 Simulation Comparisons: Finite N

The asymptotic results presented in the previous section are not necessarily applicable to real data where the finite number of observations N constrains the distribution of cluster sizes, $\sum_M M \cdot H_{M,N} = N$. In this section, we appraise the finite sample consequences for the Dirichlet, Pitman-Yor, and uniform processes via a simulation study. For each of the three processes, we simulated 1000 independent partitions for various values of sample size N and concentration parameter θ, and calculated the number of clusters K_N and distribution of cluster sizes $\boldsymbol{H}_N$ for each of the partitions.

4.1 Number of Clusters K_N

In figure 1, we examine the relationship between the number of observations N and the average number of

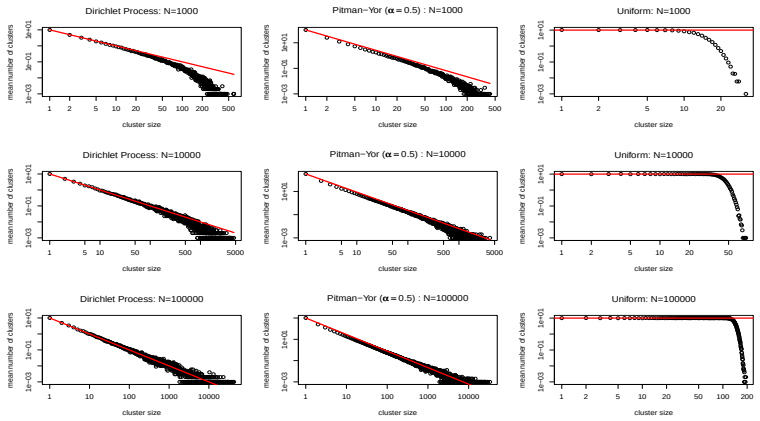

Figure 2: Cluster sizes $H_{M,N}$ as a function of M for different values of N for the Dirichlet, Pitman-Yor, and uniform processes. Data are plotted on a log-log scale and the red lines indicate the asymptotic relationships. Each point is the average number of clusters (across 1000 simulated partitions) of a particular cluster size.

clusters $\hat{K}_N$ (averaged over the 1000 simulated partitions). For $\alpha = 0.5$, the Pitman-Yor process exhibits the same rate of growth of $\hat{K}_N$ as the uniform process, confirming the equality suggested by (7) and (9) when $\alpha = 0.5$. As postulated in section 3.2, the Pitman-Yor process can exhibit either slower (*e.g.*, $\alpha = 0.25$) or faster (*e.g.*, $\alpha = 0.75$) rates of growth of $\hat{K}_N$ than the uniform process. The rate of growth of $\hat{K}_N$ for the Dirichlet process is the slowest, as suggested by (5).

4.2 Distribution of Cluster Sizes

In this section, we examine the expected distribution of cluster sizes under each process. For brevity, we focus only on concentration parameter $\theta = 10$, though the same trends are observed for other values of θ. Figure 2 is a plot of $\hat{H}_{M,N}$ (the average number of clusters of size M) as a function of M. For each process, $\hat{H}_{M,N}$ was calculated as the average over the 1000 simulated independent partitions of $H_{M,N}$ under that process. The red lines indicate the asymptotic relationships, *i.e.*, (6) for the Dirichlet process, (8) for the Pitman-Yor process, and (10) for the uniform process.

The results in figure 2 demonstrate that the simulated distribution of cluster sizes for the uniform process is quite different to the simulated distributions of clusters sizes for either the Dirichlet or Pitman-Yor processes. It is also interesting to observe the divergence from the asymptotic relationships due to the finite sample sizes, especially in the case of small N (*e.g.*, $N = 1000$).

5 Exchangeability

As mentioned in section 2, the uniform process does not lead to exchangeable partitions. Although the exchangeability of the Dirichlet and Pitman-Yor processes is desirable, these clustering models also exhibit the "rich-get-richer" property. Applied researchers are routinely forced to make assumptions when modeling real data. Even though the use of exchangeable priors can provide many practical advantages for clustering tasks, exchangeability itself is one particular modeling assumption, and there are situations in which the "rich-get-richer" property is disadvantageous. In reality, many data generating processes are not exchangeable, *e.g.*, news stories are published at different times and therefore have an associated temporal ordering. If one is willing to make an exchangeability assumption, then the Dirichlet process prior is a natural choice. However, it comes with additional assumptions about the size distribution of clusters. These assumptions will be reasonable in certain situations, but less reasonable in others. It should not be necessary to restrict applied researchers to exchangeable models, which can impose other undesired assumptions, when alternatives do exist. The uniform process sacrifices the exchangeability assumption in order to make a more balanced prior assumption about cluster sizes.

In this section, we explore the lack of exchangeability of the uniform process by first examining, for real data, the extent to which $P(\boldsymbol{c})$ is affected by permuting the observations. For any particular ordering of observations $\boldsymbol{X} = (X_1, \ldots, X_N)$, the joint probability of the corresponding cluster assignments $\boldsymbol{c}$ is

$$P(\boldsymbol{c} \,|\, \text{ordering } 1, \ldots, N) = \prod_{n=1}^{N} P(c_n \,|\, \boldsymbol{c}_{<n}) \qquad (11)$$

where "$\boldsymbol{c}_{<n}$" denotes the cluster assignments for observations $X_1, \ldots, X_{n-1}$ and $P(c_n \,|\, \boldsymbol{c}_{<n})$ is given by (4). Clearly, exhaustive evaluation of $P(\boldsymbol{c})$ for all possible orderings (permutations of observations) is not possible for realistically-sized data sets. However, we

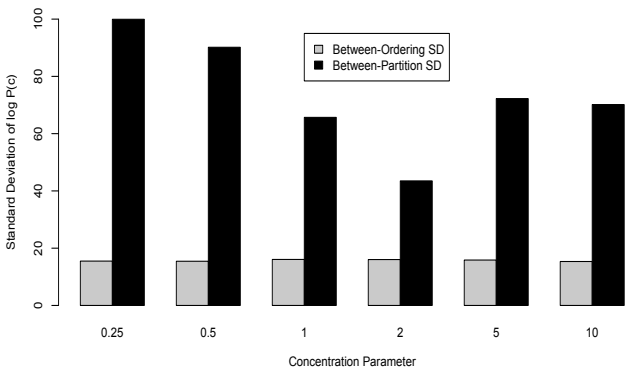

Figure 3: Comparison of the "Between-Partition SD" and the "Between-Ordering SD" (averaged over different inferred partitions) for the uniform process with six different values of concentration parameter θ.

can evaluate the robustness of $P(\boldsymbol{c})$ to different orderings as follows: for any given partition $\boldsymbol{c}$ (set of cluster assignments), we can compute the standard deviation of $\log P(\boldsymbol{c})$ over multiple different orderings of the observations. This "between-ordering SD" gives an estimate of the degree to which the ordering of observations affects $P(\boldsymbol{c})$ for a particular partition. For any given ordering of observations, we can also compute the standard deviation of $\log P(\boldsymbol{c})$ over multiple different partitions (realizations of $\boldsymbol{c}$) inferred using the Gibbs sampling algorithm described below. This "between-partition SD" gives an estimate of the variability of inferred partitions for a fixed ordering.

Figure 3 shows the "between-ordering SD" and the "between-partition SD" for partitions of 1000 carbon nanotechnology patent abstracts (see next section), obtained using five Gibbs sampling chains and 5000 orderings of the data with different values of θ. The variability between orderings is considerably smaller than the variability between partitions, suggesting that uniform process clustering results are not significantly sensitive to different orderings. These results are encouraging for applications where one is willing to sacrifice exchangeability over orderings in favor of a more balanced prior assumption about cluster sizes.

6 Document Clustering Application

In this section, we compare the Dirichlet process and the uniform processes on the task of clustering real documents—specifically, the abstracts of 1200 carbon nanotechnology patents. Dirichlet processes have been used as the basis of many approaches to document clustering including those of Zhang *et al.* (2005), Zhu *et al.* (2005) and Wallach (2008). In practice, however, there is often little justification for the *a priori* "rich-get-richer" property exhibited by the Dirichlet process.

We consider a nonparametric word-based mixture model where documents are clustered into groups on the basis of word occurrences. The model assumes the following generative process: The tokens $\boldsymbol{w}_d$ that comprise each document, indexed by d, are drawn from a document-specific distribution over words $\boldsymbol{\phi}_d$, which is itself drawn from a document-specific Dirichlet distribution with base distribution $\boldsymbol{n}_d$ and concentration parameter β. The document-specific base distribution is obtained by selecting a cluster assignment from the uniform process. If an existing cluster is selected, then $\boldsymbol{n}_d$ is set to the cluster-specific distribution over words for that cluster. If a new cluster is selected, then a new cluster-specific distribution over words is drawn from G_0, and $\boldsymbol{n}_d$ is set to that distribution:

$$c_d \,|\, c_{<d} \;\sim\; \begin{cases} \frac{1}{d-1+\theta} & c_d = k \in 1, \ldots, K \\ \frac{\theta}{d-1+\theta} & c_d = k_{\text{new}} \end{cases} \tag{12}$$

$$\boldsymbol{n}_k \;\sim\; G_0 \tag{13}$$

$$\boldsymbol{\phi}_d \;\sim\; \text{Dir}\,(\boldsymbol{\phi}_d \,|\, \boldsymbol{n}_{c_d}, \beta) \tag{14}$$

$$\boldsymbol{w}_d \;\sim\; \text{Mult}\,(\boldsymbol{\phi}_d), \tag{15}$$

where c_d is the cluster assignment for the d^{th} document. Finally, G_0 is chosen to be a hierarchical Dirichlet distribution: $G_0 = \text{Dir}\,(\boldsymbol{n}_c \,|\, \beta_1 \boldsymbol{n})$, where $\boldsymbol{n} \sim \text{Dir}\,(\boldsymbol{n} \,|\, \beta_0 \boldsymbol{u})$. This model captures the fact that documents in different clusters are likely to use different vocabularies, yet allows the word distribution for each document to vary slightly from the word distribution for the cluster to which that document belongs.

The key consequence of using either a Dirichlet or uniform process prior is that the latent variables $\boldsymbol{n}_d$ are partitioned into C clusters where the value of C does not need to be pre-specified and fixed. The vector $\boldsymbol{c}$ denotes the cluster assignments for the documents: c_d is the cluster assignment for document d. Given a set of observed documents $\mathcal{W} = \{\boldsymbol{w}_d\}_{d=1}^{D}$, Gibbs sampling (Geman and Geman, 1984) can be used to infer the latent cluster assignments $\boldsymbol{c}$. Specifically, the cluster assignment c_d for document d can be resampled from

$$P(c_d \,|\, \boldsymbol{c}_{\backslash d}, \boldsymbol{w}, \theta) \propto P(c_d \,|\, \boldsymbol{c}_{\backslash d}, \theta) \cdot P(\boldsymbol{w}_d \,|\, c_d, \boldsymbol{c}_{\backslash d}, \mathcal{W}_{\backslash d}, \boldsymbol{\beta}), \tag{16}$$

where $\boldsymbol{c}_{\backslash d}$ and $\mathcal{W}_{\backslash d}$ denote the sets of clusters and documents, respectively, excluding document d. The vector $\boldsymbol{\beta} = (\beta, \beta_1, \beta_0)$ represents the concentration parameters in the model, which can be inferred from $\mathcal{W}$ using slice sampling (Neal, 2003), as described by Wallach (2008). The likelihood component of (16) is

$$P(\boldsymbol{w}_d \,|\, c_d, \boldsymbol{c}_{\backslash d}, \mathcal{W}_{\backslash d}, \boldsymbol{\beta}) = \prod_{n=1}^{N_d} \frac{N_{w_n|d}^{<d,n} + \beta \, \frac{N_{w_n|c_d}^{<d,n} + \beta_1 \, \frac{N_{w_n}^{<d,n} + \beta_0 \, \frac{1}{W}}{\sum_w N_w^{<d,n} + \beta_0}}{\sum_w N_{w|c_d}^{<d,n} + \beta_1}}{\sum_w N_{w|d}^{<d,n} + \beta}, \tag{17}$$

where the superscript "$< d, n$" denotes a quantity including data from documents $1, \ldots, d$ and positions $1, \ldots, n-1$ only for document d. $N_{w|d}$ is the number of times word type w occurs in document d, $N_{w|c_d}$ is the number of times w occurs in cluster c_d, and N_w is the number of times w occurs in the entire corpus.

The conditional prior probability $P(c_d \,|\, \boldsymbol{c}_{\backslash d}, \theta)$ can be constructed using any of the predictive probabilities in section 2. For brevity, we focus on the (commonly-used) Dirichlet process and the uniform process. For the Dirichlet process, the conditional prior probability is given by (2). Since the uniform process lacks exchangeability over observations, we condition upon an arbitrary ordering of the documents, *e.g.*, $1, \ldots, D$. The conditional prior of c_d given $\boldsymbol{c}_{\backslash d}$ is therefore

$$P(c_d \,|\, \boldsymbol{c}_{\backslash d}, \theta, \text{ ordering } 1, \ldots, D) \propto P(c_d \,|\, c_1, \ldots, c_{d-1}, \theta) \prod_{m=d+1}^{D} P(c_m \,|\, c_1, \ldots, c_{m-1}, \theta), \quad (18)$$

where $P(c_d \,|\, c_1, \ldots, c_{d-1}, \theta)$ is given by (4). The latter terms propagate the value of c_d to the cluster assignments $c_{d+1}, \ldots, c_D$ for the documents that follow document d in the chosen ordering. With this definition of the conditional prior, the Gibbs sampling algorithm is a correct clustering procedure for $\mathcal{W}$, conditioned on the arbitrarily imposed ordering of the documents.

We compare the Dirichlet and uniform process priors by using the model (with each prior) to cluster 1200 carbon nanotechnology patent abstracts. For each prior, we use Gibbs sampling and slice sampling to infer cluster assignments $\boldsymbol{c}^{\text{train}}$ and $\boldsymbol{\beta}$ for a subset $\mathcal{W}^{\text{train}}$ of 1000 "training" abstracts. Since the results in section 5 indicate that the variability between partitions is greater than the variability between orderings, we use a single ordering of $\mathcal{W}^{\text{train}}$ and perform five runs of the Gibbs sampler. To provide insight into the role of θ, we compare results for several θ values. We evaluate predictive performance by computing the probability of a held-out set $\mathcal{W}^{\text{test}}$ of 200 abstracts given each run from the trained model. We compute $\log P(\mathcal{W}^{\text{test}} \,|\, \mathcal{D}^{\text{train}}, \theta, \boldsymbol{\beta}) = \log \sum_{\boldsymbol{c}^{\text{test}}} P(\mathcal{W}^{\text{test}}, \boldsymbol{c}^{\text{test}} \,|\, \mathcal{D}^{\text{train}}, \theta, \boldsymbol{\beta})$, where $\mathcal{D}^{\text{train}} = (\mathcal{W}^{\text{train}}, \boldsymbol{c}^{\text{train}})$ and the sum over $\boldsymbol{c}^{\text{test}}$ is approximated using a novel variant of (Wallach *et al.*, 2009)'s "left-to-right" algorithm (see supplementary materials). We average this quantity over runs of the Gibbs sampler for $\mathcal{W}^{\text{train}}$, runs of the "left-to-right" algorithm, and twenty permutations of the held-out data $\mathcal{W}^{\text{test}}$.

The left-hand plot of figure 4 compares the Dirichlet and uniform processes in terms of $\log P(\mathcal{W}^{\text{test}} \,|\, \mathcal{D}^{\text{train}}, \theta, \boldsymbol{\beta})$. Regardless of the value of concentration parameter θ, the model based on the uniform process leads to systematically higher held-out probabilities than the model based on the Dirichlet process. In other words, the uniform process provides a substantially better fit for the data in this application. The right-hand plot of figure 4 compares the Dirichlet and uniform processes in terms of the average number of clusters in a representative partition obtained using the Gibbs sampler. The uniform process leads to a greater number of clusters than the Dirichlet process for each value of θ. This is not surprising given the theoretical results for the *a priori* expected cluster sizes (section 3) and the fact that the choice of clustering prior is clearly influential on the posterior distribution in this application.

7 Discussion

The Dirichlet and Pitman-Yor processes both exhibit a "rich-get-richer" property that leads to partitions with a small number of relatively large clusters and *vice versa*. This property is seldom fully acknowledged by practitioners when using either process as part of a nonparametric Bayesian clustering model. We examine the uniform process prior, which does not exhibit this "rich-get-richer" property. The uniform process prior has received relatively little attention in the statistics literature to date, and its clustering characteristics have remained largely unexplored. We provide a comprehensive comparison of the uniform process with the Dirichlet and Pitman-Yor processes, and present a new asymptotic result for the square-root growth of the expected number of clusters under the uniform process. We also conduct a simulation study for finite sample sizes that demonstrates a substantial difference in cluster size distributions between the uniform process and the Pitman-Yor and Dirichlet processes. Previous work on the uniform process has ignored its lack of exchangeability. We present new results demonstrating that although the uniform process is not invariant to permutations of cluster assignments, it is highly robust. Finally, we compare the uniform and Dirichlet processes on a real document clustering task, demonstrating superior predictive performance of the uniform process over the Dirichlet process.

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part by CIA, NSA and NSF under NSF grant #IIS-0326249, and in part by subcontract #B582467 from Lawrence Livermore National Security, LLC, prime contractor to DOE/NNSA contract #DE-AC52-07NA27344. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

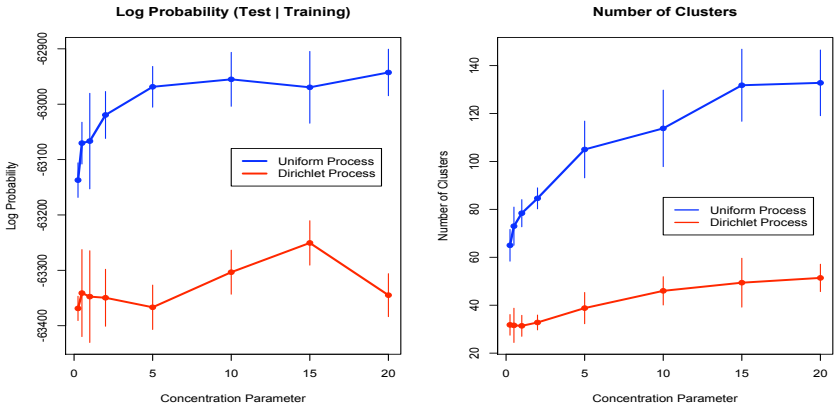

Figure 4: Left: Average log probability of held-out data given the trained model. Vertical lines indicate one SD across runs of the Gibbs sampler for $\mathcal{W}^{\text{train}}$, runs of the evaluation algorithm and (for the uniform process) twenty permutations of the held-out data. Right: The average number of clusters in a representative partition from each trained model. Vertical lines indicate one SD across runs of the Gibbs sampler for $\mathcal{W}^{\text{train}}$.

References

Aldous, D. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII-1983*, 1–198. Springer, Berlin.

Andrews, N. O. and Fox, E. A. (2007). Recent developments in document clustering. Tech. Rep. TR-07-35, Virginia Tech Department of Computer Science.

Arratia, R., Barbour, A., and Tavare, S. (2003). *Logarithmic Combinatorial Structures: a Probabilistic Approach*. Monographs in Mathematics. European Mathematical Society, Zurich, Switzerland.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence 6*, 721–741.

Green, P.J. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process *Scandinavian Journal of Statistics* **28**, 355–375.

Ishwaran, H. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

Ishwaran, H. and James, L. F. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica* **13**, 1211–1235.

Jensen, S. and Liu, J. (2008). Bayesian clustering of transcription factor binding motifs. *Journal of the American Statistical Association* **103**, 188–200.

Muller, P. and Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science* **19**, 95–110.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics* **31**, 705–767.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme In *Statistics, Probability and Game Theory*, IMS Lecture Notes - Volume 30, pp. 245–267.

Pitman, J. (2002). Combinatorial stochastic processes. Tech. Rep. 621, Department of Statistics, University of California at Berkeley.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability* **25**, 855–900.

Qin, Z. S., McCue, L. A., Thompson, W., Mayerhofer, L., Lawrence, C. E., and Liu, J. S. (2003). Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nature Biotechnology* **21**, 435–439.

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

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006a). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101**, 1566–1581.

Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *ACL 2006*.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006b). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101**, 1566–1581.

Wallach, H., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *26th International Conference on Machine Learning*.

Wallach, H., Sutton, C., and McCallum, A. (2008). Bayesian modeling of dependency trees using hierarchical Pitman-Yor priors. In *Workshop on Prior Knowledge for Text and Language*, 15–20, Finland.

Wallach, H. M. (2008). *Structured Topic Models for Language*. Ph.D. thesis, University of Cambridge.

Welling, M. (2006). Flexible priors for infinite mixture models. *Workshop on Learning with Non-parametric Bayesian Methods*.

Zhang, J., Ghahramani, Z., and Yang, Y. (2005). A probabilistic model for online document clustering with application to novelty detection. In *Advances in Neural Information Processing Systems 17*. MIT Press.

Zhu, X., Ghahramani, Z., and Lafferty, J. (2005). Time-sensitive Dirichlet process mixture models. Tech. rep., School of Computer Science, Carnegie Mellon University.