PRIOR DISTRIBUTIONS FOR PARTITIONS IN BAYESIAN NONPARAMETRICS

LEE DICKER[1] and SHANE T. JENSEN[2]

Abstract

Prior distributions for unknown data distributions play an important role in nonparametric Bayesian statistics. A commonly-used prior distribution for an unknown data distribution is the Dirichlet process, which induces a random partition on the observations from the unknown data distribution. We investigate the prediction rule that underlies the Dirichlet process prior and the implicit "rich-get-richer" characteristics of random partitions generated by this process. To provide more flexibility for the modeling of random partitions, we present two alternative prior distributions for random partitions: the Pitman-Yor process and a uniform process. We present several asymptotic results for partitions under each process as well as a simulation-based evaluation of partition properties in small samples. We also discuss the exchangeability of partitions under each prediction rule. We give special focus to the uniform process which does not share the same "rich-get-richer" property as the Dirichlet process, which would be advantageous in applications where that implicit property is not reasonable.
Keywords: Dirichlet process, Non-Parametric Bayes, Random Partitions

1 Introduction

Fully parametric Bayesian models are a powerful and popular approach to many difficult statistical problems. However, in many applied situations, practitioners are uncomfortable with the assumption of a parametric distribution for each level of the model, and a non-parametric or semi-parametric approach is instead preferred. For a recent review of Bayesian non-parametric modeling, see Muller and Quintana (2004). A common characteristic of Bayesian non-parametric or semi-parametric models is that we have a set of observations $\mathbf{y}$ from an unknown probability distribution P. In multi-level models, these observations can also represent latent variables or unknown parameters. A Bayesian model also requires the use of a

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115 ldicker@hsph.harvard.edu

[2]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA 19104 stjensen@wharton.upenn.edu

prior distribution for the unknown probability distribution P. Dirichlet processes are a class of prior distributions that have become ubiquitous in Bayesian nonparametric modeling. Let μ be a finite measure on a sample space $\mathbf{S}$. A probability measure P follows a Dirichlet process $\mathrm{DP}(\mu)$ with characteristic measure μ if $(P(A_1), ..., P(A_n))$ follows a Dirichlet distribution with parameter $(\mu(A_1), ..., \mu(A_n))$ for every finite partition $A_1, ..., A_n$ of S with $A_i \in \mathbf{S}$. Let $\theta = \mu(\mathbf{S}) > 0$ and define $\nu(\cdot) = \mu(\cdot)/\theta$.

Ferguson (1973) proved the existence of Dirichlet processes, first using the Kolmogorov extension theorem and then giving a constructive proof. Ferguson (1973) also demonstrated the following key theorem about Dirichlet processes:

Theorem 1.1. *Suppose that $X_1, \ldots, X_n$ is a sample from P and that $P \sim \mathrm{DP}(\mu)$, then*

$$P|X_1, \ldots, X_n \sim \mathrm{DP}\left(\mu + \sum_{j=1}^{n} \delta_{X_j}\right) \tag{1}$$

where δ_x is a measure that gives mass 1 to the point x.

A crucial consequence of this theorem is that the posterior distribution of P is inherently discretized by the point masses at each unique X_i. In many applications where density estimation is the desired result, this discreteness is problematic and so convolution with smooth kernels is typically employed. Escobar and West (1995) discuss the use of Dirichlet mixtures of normals for density estimation. However, in many other applications, this discreteness can be utilized to cluster data or latent variables, effectively using the Dirichlet process as a prior distribution for partitions of random variables. In this article, we will focus entirely on the use of this discreteness property and its consequences in problems concerned with the partitioning of random variables.

The Dirichlet process has become a popular tool for clustering within hierarchical Bayesian models. Some recent examples in the literature include Liu (1996), Green and Richardson (2001) and Medvedovic and Sivaganesan (2002), and Jensen and Liu (2007). However, in most applications, very little attention is given to the implicit assumptions about the structure of partitions generated by a Dirichlet process prior distribution. As we explore in Section 2 below, a fundamental characteristic of partitions generated by the Dirichlet process is a "rich-get-richer" property that leads to a priori partitions consisting of a small number of large clusters. This may be an undesirable property in many applications, and for these situations, a practitioner is confronted by the following questions:

1. Are there alternatives to the Dirichlet process without this "rich-get-richer" property?

2. What are the corresponding properties of these alternative prior processes?

In this work, we address these questions by exploring two alternative prior processes, the Pitman-Yor process and a uniform process, which show characteristics differing substantively from the Dirichlet process. We focus primarily on the uniform process as a particularly intriguing alternative for the modeling of random partitions, since it generates a dramatically different set of clustering characteristics compared to the Dirichlet process. We compare the uniform process to the Dirichlet and Pitman-Yor processes both in terms of asymptotic characteristics (Section 3) as well as characteristics in reasonable sized samples (Section 4). We briefly discuss some exchangeability issues in Section 5 and conclude with a brief discussion.

2 Prediction Rules for Partition Priors

We are interested in the partitioning of n variables $X_1, \ldots, X_n$. It is often convenient to describe this partitioning process by a "prediction rule": the sequence of generative conditional probabilities implied by a particular prior distribution. In this framework, we observe random variables $X_1, \ldots, X_n$ one at a time, and our partition is constructed sequentially. From Theorem 1.1, if we use a Dirichlet process prior for P, then the conditional distribution of a new observation X_{n+1} is a mixture of point masses at the previous observations $X_1, \ldots, X_n$ and the underlying measure μ. If we define X_i and X_j to be in the same cluster if $X_i = X_j$, then we see that this prediction rule formulation sequentially constructs a partition, since X_{n+1} joins an existing cluster if $X_{n+1} = X_i$ for some $i \leq n$, or alternatively, X_{n+1} is drawn from the measure $\nu(\cdot) = \mu(\cdot)/\theta$, which creates a new cluster consisting only of X_{n+1}. The parameter θ plays the role of a prior weight for the formation of a new cluster.

We can also write this prediction rule in terms of the current clusters in the partition. Let $\tilde{X}_1, \ldots, \tilde{X}_K$ be the K distinct values (ie. K clusters) observed in the set of variables $X_1, \ldots, X_n$, and define $= N_1, \ldots, N_K$ such that $N_k = \sum_{i=1}^{n} \mathrm{I}(X_i = \tilde{X}_k)$ ie. N_k is the size of cluster k. Then we have

Corollary 2.1. *Suppose that $X_1, ..., X_{n+1} \sim P$ and $P \sim \mathrm{DP}(\mu)$. Then*

$$\begin{aligned} \Pr(X_{n+1} = \tilde{X}_k | X_1, ..., X_n) &= \frac{N_k}{n+\theta}, \quad 1 \leq k \leq K, \\ \Pr(X_{n+1} \notin \{X_1, ..., X_n\} | X_1, ..., X_n) &= \frac{\theta}{n+\theta} \end{aligned} \tag{2}$$

This formulation is evident in the popular Chinese restaurant construction of the Dirichlet process. Imagine a Chinese restaurant with an infinite number of infinitely large tables. The restaurant is initially empty and the first customer enters the restaurant and sits at a table by himself. Customers continue to enter the restaurant, one at a time, and let the probability that the $(n+1)$-th customer to enter the restaurant sits next to the i-th customer ($i \leq n$) at their table be $1/(n+\theta)$. The probability that the $(n+1)$-th customer sits at a previously unoccupied table is $\theta/(n+\theta)$. With this construction, we see that the probability of the $(n+1)$-th customer joining a particular table k is proportional to the number of customers N_k already sitting at table k, which leads us to (2). The most obvious characteristic of this prediction rule is the "rich get richer" property: the probability of joining a cluster is proportional to the size of that cluster, which means that new observations have a strong preference towards already large clusters. This preferential attachment property has been observed in a wide variety of natural settings, such as the study of scale-free networks (Barabasi and Albert, 1999). However, this preferential attachment property may be disadvantageous in other applications, a fact that is not commonly acknowledged by practitioners using the Dirichlet process.

This predictive rule framework allows us to consider a generalization of the Dirichlet process of the following form:

$$\begin{aligned} \Pr(X_{n+1} = \tilde{X}_k | X_1, ..., X_n) &\propto f(N_k), \quad 1 \leq k \leq K, \\ \Pr(X_{n+1} \notin \{X_1, ..., X_n\} | X_1, ..., X_n) &\propto g(\theta) \end{aligned} \tag{3}$$

That $\Pr(X_{n+1} = \tilde{X}_k | X_1, ..., X_n)$ can only depend on the sample through N_k has been referred to as "Johnson's sufficientness postulate" by Zabell (1992). In the remainder of this section, we focus on two particular sets of choices for functions f and g that give us alternatives to the Dirichlet process.

2.1 Uniform prediction rule

The prediction rule (2) suggests that the use of a Dirichlet process prior will lead to partitions that are dominated by a few large clusters, since larger clusters will tend to attract new observations in the sequential creation of a partition. We might prefer a prior distribution over partitions that more uniformly spreads observations between clusters by letting $f_{\mathrm{UN}}(N_k) = 1$ and $g_{\mathrm{UN}}(\theta) = \theta$. The use of these functions gives us the following uniform prediction rule:

$$\begin{aligned} \Pr(X_{n+1} = \tilde{X}_k | \mathbf{X}, \mathrm{UN}) &= \frac{1}{K+\theta}, \quad 1 \leq k \leq K, \\ \Pr(X_{n+1} \notin \{X_1, ..., X_n\} | \mathbf{X}, \mathrm{UN}) &= \frac{\theta}{K+\theta} \end{aligned} \tag{4}$$

Here, the probability that the $(n+1)-th$ observation joins one of the K existing clusters is a discrete uniform over these clusters, and is not related to the current size of each cluster. This prior over partitions was used in Qin *et al.* (2003) and Jensen and Liu (2007), but the theoretical properties of this prior distribution have been relatively unexplored. This exploration is a major focus of this paper.

2.2 Pitman-Yor prediction rule

We also consider a final prediction rule that serves as a compromise between the prediction rules (2) and (4) in terms of the preference towards sequential addition of new observations into already large clusters. Consider the functions $f_{\mathrm{PY}}(N_k) = N_k - \alpha$ and $g_{\mathrm{PY}}(\theta) = \theta + K \cdot \alpha$, in which case we have the Pitman-Yor prediction rule:

$$\begin{aligned} \Pr(X_{n+1} = \tilde{X}_k | \mathbf{X}, \mathrm{PY}) &= \frac{N_k - \alpha}{n+\theta}, \quad 1 \leq k \leq K, \\ \Pr(X_{n+1} \notin \{X_1, ..., X_n\} | \mathbf{X}, \mathrm{PY}) &= \frac{\theta + K \cdot \alpha}{n+\theta} \end{aligned} \tag{5}$$

We see a "rich-get-richer" property similar to the Dirichlet process, but with an additional parameter α ($0 \leq \alpha < 1$) which serves to reduce the probability of adding to an existing cluster. This compromise prediction rule arose from a process studied by Pitman and Yor (1997). Although this process has not received much attention in the statistics literature, it has been used in clustering applications in natural language processing (eg. Teh (2006)).

In the remainder of this paper, we will compare these two alternative processes to the popular Dirichlet process in terms of several characteristics of interest of their resulting partitions. Specifically, we will focus on the number of clusters K_n as well as the distribution of cluster sizes that arise from the partitioning of n observations. When analyzing the distribution of cluster sizes for n observations, we do not focus on the raw sizes of each cluster $\mathbf{N}_n = (N_1, \ldots, N_{K_n})$ directly, but rather concentrate on summary variables $\mathbf{C}_n = (C_{1,n}, C_{2,n}, \ldots, C_{n,n})$ where $C_{j,n}$ is the number of clusters of size j in the partition of n observations. In our experience, the number of clusters K_n and the histogram of cluster sizes $\mathbf{C}_n$ are the primary statistics of interest when summarizing partitions.

3 Asymptotic Behaviour of Prior Prediction Rules

In this section, we compare the prior distributions implied by (2), (4), and (5) in terms of the asymptotic behaviour of our partition characteristics, the number of

clusters K_n and the histogram of cluster sizes $\mathbf{C}_n$. We use the conditional notation $(\cdot\,|\mathrm{DP})$, $(\cdot\,|\mathrm{PY})$, and $(\cdot\,|\mathrm{UN})$ to indicate that the random variables $X_1, X_2, ...$ follow the Dirichlet process, Pitman-Yor, and uniform prediction rule, respectively, though this notation is sometimes suppressed if the prediction rule under consideration is clear from the context. We first review the asympotic expectation for K_n and $C_{j,n}$, which have been studied extensively for the Dirichlet process prediction rule. We then review some previous results for the Pitman-Yor process prediction rule which are less well known in the statistical community. Finally, we present possibly novel asymptotic results for the Uniform prediction rule, which has not been studied extensively in the previous literature.

3.1 $\mathrm{E}(K_n)$ and $\mathrm{E}(C_{j,n})$ for the Dirichlet process prediction rule

Assume that $X_1, \ldots, X_n$ are generated using the Dirichlet Process prediction rule. Observe that $K_n = \sum_{j=1}^{n} \mathrm{I}(X_j \notin \{X_1, ..., X_{j-1}\})$,

$$E(K_n|\mathrm{DP}) = \sum_{j=1}^{n} \Pr(X_j \notin \{X_1, ..., X_{j-1}\}) = \sum_{j=1}^{n} \frac{\theta}{j-1+\theta}.$$

and as $n \to \infty$,

$$E(K_n|\mathrm{DP}) = \sum_{j=1}^{n} \frac{\theta}{j-1+\theta} \approx \theta \log n. \tag{6}$$

Arratia *et al.* (2003) demonstrated the following result for the asymptotic expectation of the cluster sizes $\mathbf{C}$ under the Dirichlet process prediction rule. For arbitrary $\theta > 0$,

$$\lim_{n\to\infty} E(C_{j,n}|\mathrm{DP}) = \frac{\theta}{j} \tag{7}$$

This simple result implies that, regardless of the value of θ, the expected number of clusters of a given size j is inversely proportional to that size j. In fact, for the Dirichlet process, we have a more general characterization of the asympotic behaviour of cluster sizes $\mathbf{C}$. It is also shown in Arratia *et al.* (2003) that, as $n \to \infty$, $(C_{1,n}, \ldots, C_{n,n})|\mathrm{DP}$ converges in distribution to $(Z_1, \ldots, Z_n)$ where each Z_j is independent and

$$Z_j \sim \mathrm{Poisson}(\theta/j) \tag{8}$$

It is also worth noting that since $K_n = \sum_{j=1}^{n} C_{j,n}$, one can use (7) to also obtain the asymptotic result for $E(K_n|\mathrm{DP})$ given in (6).

3.2 $\mathrm{E}(K_n)$ and $\mathrm{E}(C_{j,n})$ for the Pitman-Yor prediction rule

Suppose $0 < \alpha < 1$ and $X_1, \ldots, X_n$ are generated using the Pitman-Yor prediction rule. Again, using the observation that $K_n = \sum_{j=1}^{n} \mathrm{I}(X_j \notin \{X_1, ..., X_{j-1}\})$, we have

$$\begin{aligned} E(K_n) &= E\left[E(K_n|K_{n-1})\right] = E\left[K_{n-1} + P(X_n \notin \{X_1, ..., X_{n-1}\}|K_{n-1})\right] \\ &= E(K_{n-1}) + E\left(\frac{\theta + \alpha K_{n-1}}{\theta + n - 1}\right) \\ &= \frac{\theta}{\theta + n - 1} + \frac{\alpha + \theta + n - 1}{\theta + n - 1} \cdot E(K_{n-1}). \end{aligned}$$

With the initial condition $E(K_1) = 1$, this recursive relationship is solved by

$$E(K_n|\mathrm{PY}) = \frac{\Gamma(1+\theta)}{\alpha\Gamma(\alpha+\theta)} \frac{\Gamma(\alpha+\theta+n)}{\Gamma(\theta+n)} - \frac{\theta}{\alpha} \approx \frac{\Gamma(1+\theta)}{\alpha\Gamma(\alpha+\theta)} n^{\alpha} \tag{9}$$

and so, as $n \to \infty$,

$$E(K_n|\mathrm{PY}) \approx K(\alpha, \theta) \cdot n^{\alpha} \tag{10}$$

where $K(\alpha, \theta) = \Gamma(1+\theta)/\alpha\Gamma(\alpha+\theta)$. This result is also given in Pitman (2002), along with the following result for the distribution of cluster sizes $\mathbf{C}_n$,

$$\frac{C_{j,n}}{K_n} \xrightarrow{a.s.} p_{\alpha,j} = \frac{\alpha \prod_{i=1}^{j-1}(i-\alpha)}{j!} \qquad \text{for every } j = 1, 2, \ldots \tag{11}$$

Combining (11) together with (10) suggests that, as $n \to \infty$,

$$E(C_{j,n}|\mathrm{PY}) \approx \frac{\Gamma(1+\theta)\prod_{i=1}^{j-1}(i-\alpha)}{\Gamma(\alpha+\theta)j!} \cdot n^{\alpha} \qquad \text{for every } j = 1, 2, \ldots \tag{12}$$

3.3 $\mathrm{E}(K_n)$ and $\mathrm{E}(C_{j,n})$ for the Uniform prediction rule

The previous literature is far more sparse with regards to the Uniform prediction rule. We provide the following potentially novel result for the expected number of clusters under the Uniform prediction rule,

Theorem 3.1.

$$\frac{E(K_n)}{\sqrt{n}} \to \sqrt{2\theta}.$$

which is proven in Appendix A. Thus, we see that as $n \to \infty$,

$$E(K_n|\mathrm{UN}) \approx K(\theta) \cdot n^{\frac{1}{2}} \tag{13}$$

where $K(\theta) = \sqrt{2\theta}$. We also suggest the following conjecture for the distribution of cluster sizes $\mathbf{C}_n$, based on our simulation results from the following Section 4.

Conjecture 3.2.

$$E(C_{j,n}|\mathrm{PY}) \approx \theta. \qquad \textit{as } n \to \infty \qquad \textit{for every } j = 1, 2, \ldots$$

This conjecture fits nicely with the underlying intuition of the uniform prediction rule, that new observations are equally likely to join any of the previously existing clusters, regardless of size.

To summarize our asymptotic results, the expected number of clusters K_n grows logarithmically with the sample size under a Dirichlet process prediction rule, whereas the uniform prediction rule leads the expected number of clusters K_n to grow with the square root of the sample size. Interestingly, the Pitman-Yor prediction rule implies that the expected number of clusters grows at a rate of n^α, which means that depending on the value of the additional parameter α, the Pitman-Yor prediction rule can lead to a slower or faster growth rate for K_n than the uniform prediction rule. For $\alpha = 0.5$, the expected number of clusters grows at the same rate for the Pitman-Yor and uniform prediction rules, though the distribution of cluster sizes $\mathbf{C}_n$ for these two rules is clearly quite different, as seen in the results above as well as our simulation results in Section 4 below. In fact, the distribution of cluster sizes $\mathbf{C}_n$ for the Pitman-Yor prediction rule show similar characteristics to $\mathbf{C}_n$ from the Dirichlet process prediction rules, despite the closer similarity between the Pitman-Yor and uniform prediction rules with respect to the number of clusters K_n. The general conclusion is that the Pitman-Yor process can be configured to look similar to a uniform process in terms of the expected number of clusters, but not in terms of the uniform distribution of cluster sizes, which is a unique aspect of the uniform process.

4 Simulation Comparisons in Finite Samples

The asymptotic results presented in the previous section are not necessarily applicable to finite samples, where the distribution of cluster sizes are more sensitive to the finite sample constraint

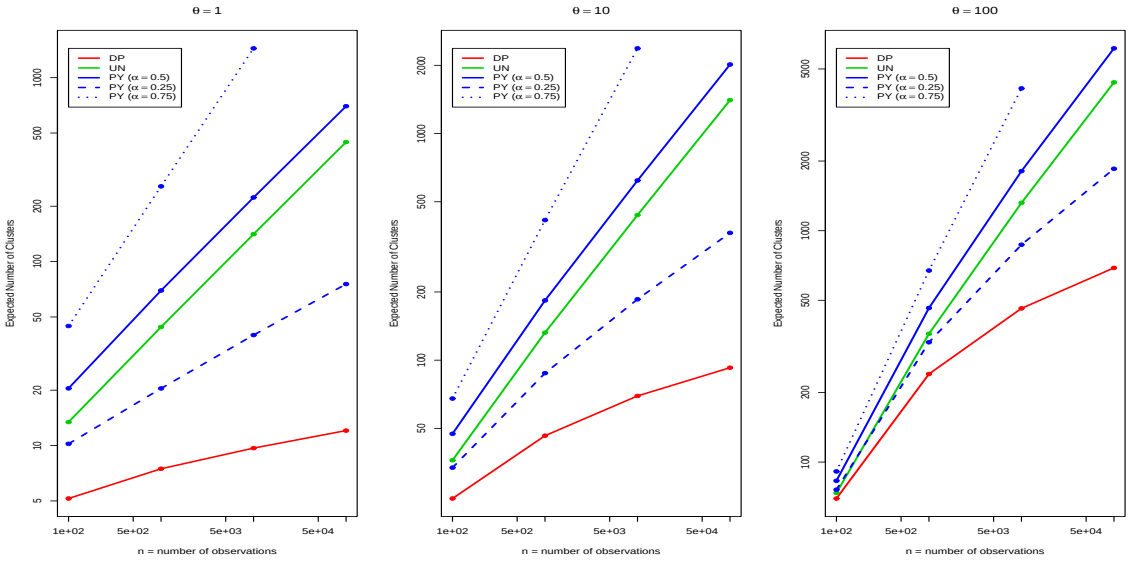
$$\sum_{j=1}^{n} j \cdot C_{j,n} = n. \tag{14}$$

We can appraise the consequences of these three prediction rules in finite samples by simulation. For various values of n and θ and for the each of the Dirichlet process, Pitman-Yor (with $\alpha = 0.5$) and uniform prediction rules, we simulated 1000 independent partitions. Each of these partitions gives us a simulated number of clusters K_n and distribution of cluster sizes $\mathbf{C}_n = (C_{1,n}, C_{2,n}, \ldots, C_{n,n})$ under our three prediction rules.

4.1 Comparison of K_n between prediction rules

In Figure 1, we examine the relationship between the number of observations n and the average number of clusters $\hat{K}_n$ (averaged over the 1000 simulated partitions). We see that the rate of growth of $\hat{K}_n$ is the same for the uniform and the Pitman-Yor ($\alpha = 0.5$) prediction rules, which agrees with the equality suggested by (10) and (13) when $\alpha = 0.5$. Also, as postulated in Section 3.2, we see that depending on the value of α, the Pitman-Yor prediction rule can show either slower ($\alpha = 0.25$) or faster ($\alpha = 0.75$) rates of growth of $\hat{K}_n$ compared to the uniform prediction rule. The rate of growth of $\hat{K}_n$ for the Dirichlet process prediction rule is slower than any of the other processes, as suggested by the logarithmic rate in (6).

Figure 1: Expected number of clusters $\hat{K}_n$ as a function of sample size n for different values of θ. Both axes are plotted on the log scale.

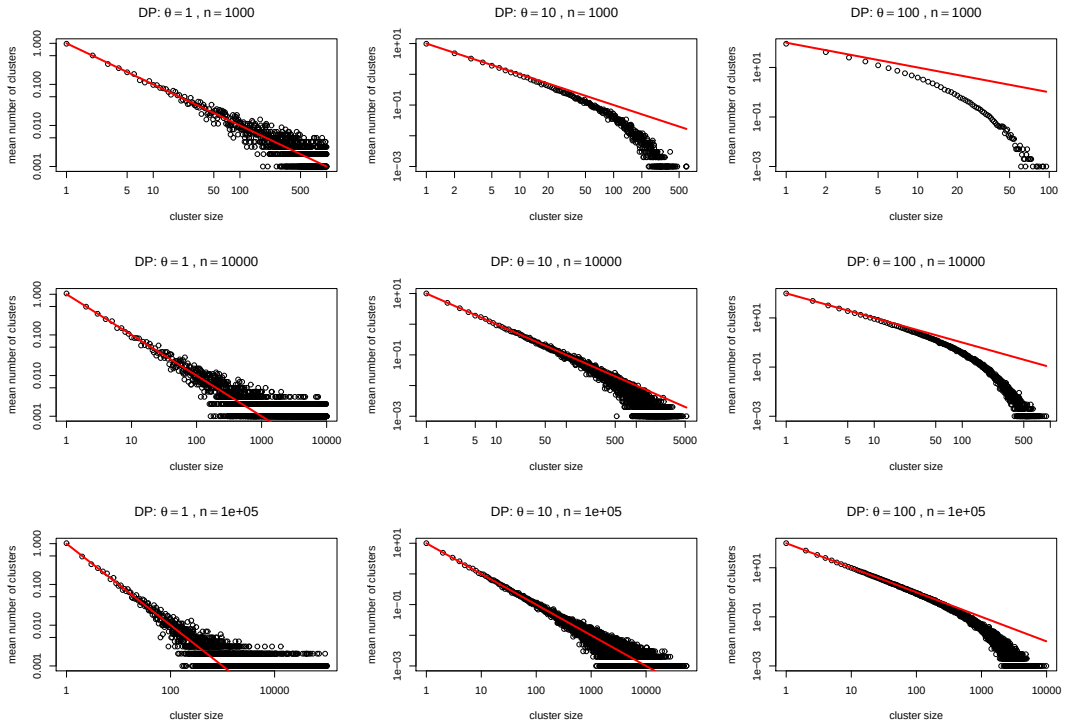


4.2 Distribution of Cluster Sizes under DP prediction rule

Figure 2 is a plot of $\hat{C}_{j,n}$, the number of clusters of size j, as a function of j. $\hat{C}_{j,n}$ was calculated as the average over the 1000 simulated independent partitions of $C_{j,n}$ under the Dirichlet process prediction rule. Data is plotted on a log-log scale and the line in each plot from the upper left corner to the lower right corner is the relationship $f(j) = \theta/j$. By equation (7), as $n \to \infty$, the relationship between the

points $(j, \hat{C}_{j,n})$ should approach this line, which we do observe when comparing the top row ($n = 1000$) to the middle row ($n = 10000$) to the bottom ($n = 100000$) row in Figure 2. However, it is interesting to observe the substantial divergence from this relationship due to the finite sample size constraint, especially in the plots with increasing values of θ ($\theta = 10$ and $\theta = 100$).

Figure 2: Dirichlet process prediction rule: $\hat{C}_{j,n}$ as a function of j for different values of θ and n. Both axes are plotted on the log scale. The red line indicates the line $f(j) = \theta/j$.

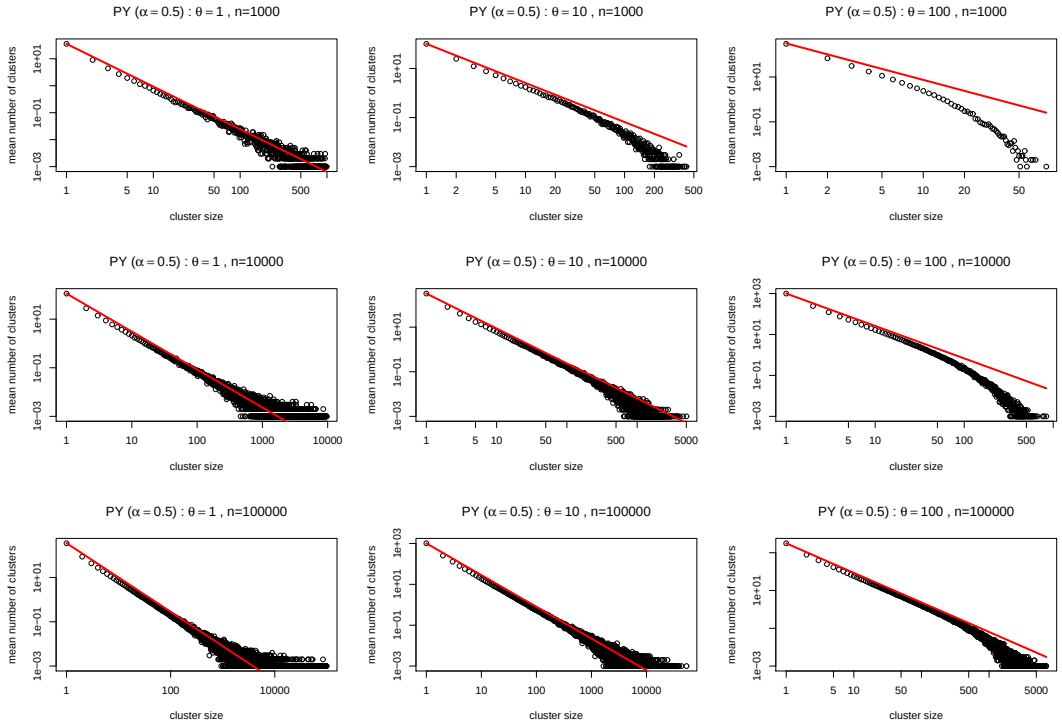


4.3 Distribution of Cluster Sizes under PY prediction rule

Figure 3 is a plot of $\hat{C}_{j,n}$ as a function of j, where $\hat{C}_{j,n}$ is now calculated as the average over the 1000 simulated independent partitions of $C_{j,n}$ under the Pitman-Yor prediction rule with $\alpha = 0.5$. Points are again plotted on a log-log scale and the line in each plot from the upper left corner to the lower right corner is the aymptotic relationship of $g(j)$ given in (12). We again observe the same divergence from the asymptotic relationship due to the finite sample size constraint, which is

more substantial in our simulations with $n = 1000$ compared to $n = 10000$ or $n = 100000$.

Figure 3: Pitman-Yor ($\alpha = 0.5$) prediction rule: $\hat{C}_{j,n}$ as a function of j for different values of θ and n. Both axes are plotted on the log scale. The red line indicates the relationship $g(j)$ suggested in (12).

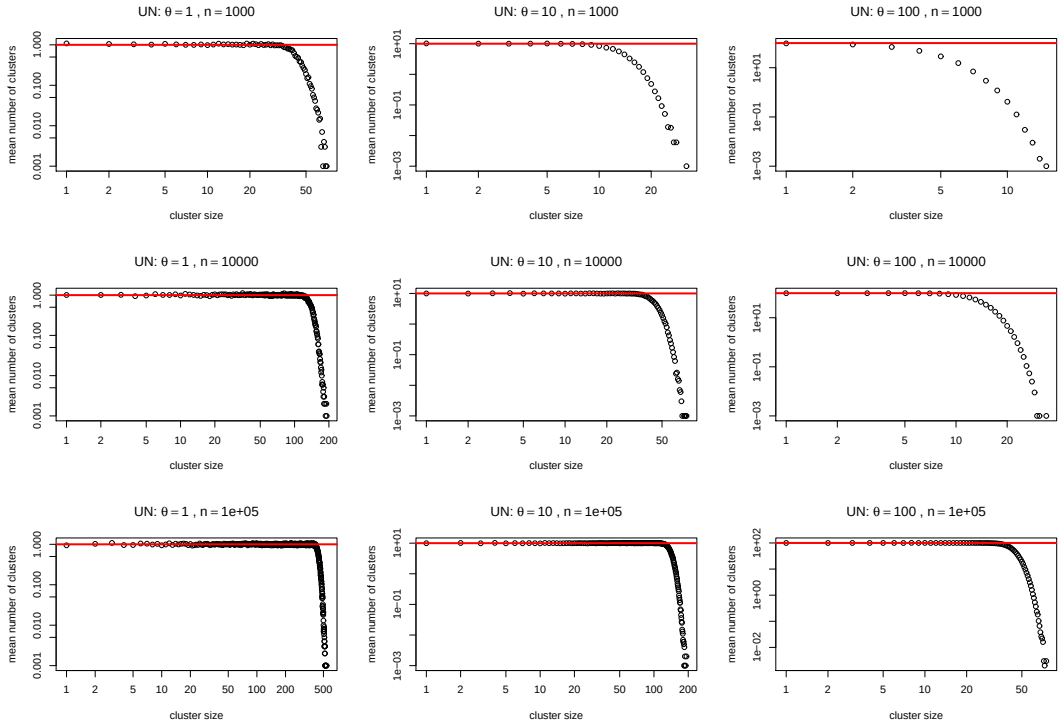


4.4 Distribution of Cluster Sizes under UN prediction rule

Figure 4 is a plot of $\hat{C}_{j,n}$ as a function of j, where $\hat{C}_{j,n}$ is now calculated as the average over the 1000 simulated independent partitions of $C_{j,n}$ under the uniform prediction rule. Unlike Figures 2-3, the points in Figure 4 are plotted on the log-log scale, and the line along the top of each plot is the relationship $f(j) = \theta$. Based on Figure 4, we conjecture that $\lim_{n\to\infty} E(C_{j,n}|\mathrm{UF}) = \theta$, though clearly we again see a deviation due to the finite sample size constraint. We again observe that the deviation from the conjectured relationship is most substantial in the upper-right corner, when n is smaller ($n = 1000$) and θ is larger ($\theta = 100$).

We draw attention to several conclusions from our simulation results. The first is

Figure 4: Uniform prediction rule: $\hat{C}_{j,n}$ as a function of j for different values of θ and n. Both axes are plotted on the log scale. The red line indicates the line $h(j) = \theta$.



the interesting similarities and differences between the Uniform and Pitman-Yor prediction rules. When $\alpha = 0.5$, both the Uniform and Pitman-Yor processes show the same rate of growth for the expected number of clusters $\mathrm{E}(K_n)$. However, the two processes are quite different in terms of their resulting distributions of cluster sizes $C_{j,n}$, as can be seen in Figures 3 and 4. In addition, from examining Figures 2-4, we note that there is substantial deviation from the asymptotic relationship for larger cluster sizes, especially in smaller samples (e.g. $n = 1000$). This deviation is due to the finite sample constraint (14), which is not acknowledged by the asymptotic relationships presented in Section 3.

5 Exchangeability of Prediction Rules

When using a sequential prediction rule framework for generating prior distributions over partitions, one also needs to address the issue of exchangeability. A par-

tition is exchangeable if the joint prior density of the partition is unaffected by the order in which the clusters were generated. As pointed out by Pitman (2002), most sequential prediction rules will fail to produce a partition that is exchangeable. Consider a partition of n observations into K clusters with sizes $\mathbf{N} = (N_1, \ldots, N_K)$ which are listed in the same order in which they were created. Pitman (2002) states that the partition generated by a particular prediction rule is exchangeable if and only if the joint density $p(\mathbf{N})$ is a symmetric function of these cluster sizes $\mathbf{N} = (N_1, \ldots, N_K)$. The Dirichlet process prediction rule (2) gives the joint density,

$$p(\mathbf{N}) = \frac{\theta^K \prod_{i=1}^{K} (N_i - 1)!}{\prod_{i=1}^{n} (\theta + i - 1)} \tag{15}$$

while the Pitman-Yor process prediction rule (5) gives the joint density,

$$p(\mathbf{N}) = \frac{\prod_{i=0}^{K-1} (\theta^K + \alpha + i \cdot \alpha) \prod_{i=1}^{K} \prod_{j=0}^{N_i - 1} (1 - \alpha + j)}{\prod_{i=1}^{n} (\theta + i - 1)} \tag{16}$$

Both of these joint densities (15) and (16) are symmetric functions of the cluster sizes $(N_1, \ldots, N_K)$, which implies that the Dirichlet process and Pitman-Yor process prediction rules both directly lead to exchangeable partitions. However, the joint density from the Uniform prediction rule (4) is

$$p(\mathbf{N}) = \frac{\theta^{K-1}(\theta + K)}{\prod_{i=1}^{K} (\theta + i)^{N_i}} \tag{17}$$

which is not a symmetric function of $(N_1, \ldots, N_K)$ in the denominator. This means that we will get different values of the prior density (17) for different, but exchangeable orderings of unequally-sized clusters. However, this lack of exchangeability for the uniform process can be addressed by defining a "signature" of the partition that is identical for exchangeable partitions (Green and Richardson, 2001). For example, if we let $p^*(\mathbf{N}) = k \cdot p(\mathbf{N}')$ where $\mathbf{N}'$ is $\mathbf{N}$ with the N_i's arranged in order from the largest cluster to the smallest, then the calculation of (17) for $\mathbf{N}'$ will be the same for all exchangeable values of $\mathbf{N}$. Comparisons of partitions generated by the uniform process should be performed on the signatures of these partitions instead of the partitions themselves.

6 Discussion

We have explored and compared the characteristics of three prior distributions for partitions of random variables. The popular Dirichlet process prior has a "rich get richer" property that leads to partitions with a small number of relatively large clusters. This property is not usually acknowledged by practitioners when using the Dirichlet process within a hierarchical model. An important consequence of the "rich get richer" property is that the number of clusters grows slowly with an increasing number of observations. In fact, the expected number of clusters grows logarithmically as the sample size increases to infinity. We have presented two other priors for random partitions as alternatives to the Dirichlet process. The Pitman-Yor process prior includes an additional parameter $0 \leq \alpha \leq 1$ that helps to dampen this "rich get richer" property. This parameter α controls the growth of the number of clusters: the expected number of clusters grows at a rate of n^{α} as the sample size goes to infinity. However, we observed that the distribution of cluster sizes under the Pitman-Yor process still shows similar characteristics to the Dirichlet process.

As a more substantial alternative to the Dirichlet process, we introduced a uniform process prior that eliminates the "rich get richer" property completely. We presented a novel result that shows that the expected number of clusters under this uniform process grows with the square root of n. Our simulation studies also demonstrated a dramatic difference in the distribution of cluster sizes between the uniform and Dirichlet process. In many applied settings, these differences in prior assumptions may not be influential on posterior inference (eg. Jensen and Liu (2007)). However, as demonstrated by Green and Richardson (2001), the "rich get richer" property of the Dirichlet process priors can persist in the posterior distribution for many datasets. When the data does not contain strong clustering relationships, the uniform prior distribution will be less likely to group variables into large clusters, which is a more conservative tendency in terms of inferring clustering relationships. We suggest that the extra cautiousness of the uniform process could be advantageous when in situations where a partition must be inferred over a large number of variables based on relatively noisy data. We hope that the results derived in this paper can provide some guidance as to the appropriate prior choice for future non-parametric Bayesian clustering applications.

Acknowledgement

The authors thank Warren Ewens and Dylan Small for many helpful discussions.

A Proof of Theorem 3.1.

We first define $T_k = \inf\{m > T_{k-1}; X_m \notin \{X_1, \ldots, X_{m-1}\}$. T_k is the "waiting time" (number of observations needed) until the k-th unique observation. Under the Uniform prediction rule, $T_k = \sum_{i=1}^{k} \tau_i$ where $\tau_i \sim \text{Geometric}(\theta/(\theta + i - 1)$ and the τ_i's are independent, so

$$\begin{aligned} \mathrm{E}(T_k) &= \sum_{i=1}^{k} \frac{\theta + i - 1}{\theta} = \frac{k^2}{2\theta} + k\left(1 - \frac{1}{2\theta}\right) \\ \mathrm{Var}(T_k) &= \sum_{i=1}^{k} \frac{(\theta + i - 1)(i-1)}{\theta^2} = \frac{k^3}{3\theta^2} + k^2\frac{1}{2\theta}\left(1 - \frac{1}{\theta}\right) + k\frac{1}{2\theta}\left(\frac{1}{3\theta} - 1\right) \end{aligned}$$

In terms of T_j,

$$K_n = \max\{j; T_j \leq n\} = \sum_{j=1}^{n} \mathrm{I}(T_j \leq n)$$

We first prove a strong law for the convergence of T_k. Let $\epsilon > 0$. From Chebychev's inequality and (18),

$$\mathrm{P}\left(|T_k - \mathrm{E}(T_k)| > \epsilon k^2\right) \leq \frac{\mathrm{Var}(T_k)}{\epsilon^2 k^4} \leq \frac{C(\theta, \epsilon)}{k} \tag{18}$$

From (18), we have that,

$$\mathrm{P}\left(|T_{k^2} - ET_{k^2}| > \epsilon k^4\right) \leq \frac{C(\theta, \epsilon)}{k^2}$$

and so by the Borel-Cantelli lemma, $\mathrm{P}\left(|T_{k^2} - \mathrm{E}(T_{k^2})| > \epsilon k^4\right) = 0$. Since $\epsilon > 0$ was chosen arbitrarily, it follows that $\frac{T_{k^2} - ET_{k^2}}{k^4} \to 0$ almost surely and hence $\frac{T_{k^2}}{k^4} \to \frac{1}{2\theta}$ almost surely. Now, let $m = \lfloor \sqrt{k} \rfloor$. Since T_k is increasing,

$$\frac{T_{m^2}}{(m+1)^4} \leq \frac{T_k}{k^2} \leq \frac{T_{(m+1)^2}}{m^4}. \tag{19}$$

Since $\frac{m+1}{m} \to 1$, both sides of the inequality (19) converge to $(2\theta)^{-1}$ almost surely, and so

$$\frac{T_k}{k^2} \to \frac{1}{2\theta} \text{ almost surely.} \tag{20}$$

The strong law (20) easily implies a strong law for K_n as follows. Note that $T_{K_n} \leq n < T_{K_n+1}$ and, consequently,

$$\frac{T_{K_n}}{K_n^2} \leq \frac{n}{K_n^2} < \frac{T_{K_n+1}}{K_n^2}.$$

Since $K_n \to \infty$ almost surely and $T_k/k^2 \to 1/(2\theta)$ almost surely, it follows that the left and right hand side above both converge to $1/(2\theta)$ almost surely. Thus, $K_n^2/n \to 2\theta$ almost surely and so

$$\frac{K_n}{\sqrt{n}} \to \sqrt{2\theta} \text{ almost surely.} \tag{21}$$

From the strong law (21) and the dominated convergence theorem, we have

$$\frac{EK_n}{n} \to 0. \tag{22}$$

We combine (22) together with following result,

$$EK_n^2 = EK_n + 2\theta(n - EK_n). \tag{23}$$

to give us

$$\frac{EK_n^2}{n} \to 2\theta. \tag{24}$$

We prove result (23) in Appendix B. Finally, using (24) together with Fatou's lemma and Jensen's inequality, gives us

$$\sqrt{2\theta} \leq \liminf_{n\to\infty} \frac{EK_n}{\sqrt{n}} \leq \limsup_{n\to\infty} \frac{EK_n}{\sqrt{n}} \leq \limsup_{n\to\infty} \sqrt{\frac{EK_n^2}{n}} = \sqrt{2\theta}.$$

which proves the result

$$\frac{EK_n}{\sqrt{n}} \to \sqrt{2\theta}.$$

under the Uniform prediction rule.

B Result relating $\mathrm{E}(K_n)$ to $\mathrm{E}(K_n^2)$

Recall the definition of T_j from Appendix A and now define $M_n = K_n + 1$. Consider the "waiting time" T_{M_n} until the observation that creates our $(K_n + 1)$-th unique cluster. We relate $\mathrm{E}(K_n)$ to $\mathrm{E}(K_n^2)$ by calculating $\mathrm{E}(T_{M_n})$ in two different ways. First, observe that

$$\begin{aligned}
\mathrm{E}(T_{M_n}) &= \mathrm{E}\left(\sum_{k=1}^{\infty} \tau_k \cdot \mathrm{I}(k \leq M_n)\right) = \sum_{k=1}^{\infty} \mathrm{E}(\tau_k) \cdot \mathrm{P}(k \leq M_n) \\
&= \frac{\theta - 1}{\theta} \sum_{k=1}^{\infty} \mathrm{P}(k \leq M_n) + \frac{1}{\theta} \sum_{k=1}^{\infty} k \cdot \mathrm{P}(k \leq M_n) \\
&= \frac{\theta - 1}{\theta} \mathrm{E}(M_n) + \frac{1}{2\theta} \mathrm{E}(M_n(M_n + 1))
\end{aligned}$$

which, since $M_n = K_n + 1$, simplifies to

$$\mathrm{E}(T_{M_n}) = 1 + \mathrm{E}(K_n)\left(1 + \frac{1}{2\theta}\right) + \mathrm{E}(K_n^2)\frac{1}{2\theta} \tag{25}$$

Now, observe that $T_{M_n} = n + \sum_j \mathrm{I}(M_{n+j} = M_n)$ and so $\mathrm{E}(T_{M_n}) = n + \sum_j \mathrm{P}(M_{n+j} = M_n)$ where

$$\begin{aligned}
\mathrm{P}(M_{n+j} = M_n) &= \sum_k \mathrm{P}(T_k \le n,\ n + j < T_{k+1}) \\
&= \sum_k \mathrm{P}(T_k \le n < T_{k+1})\mathrm{P}(j < \tau_{k+1}) \\
&= \sum_k \mathrm{P}(M_n = k + 1)\mathrm{P}(j < \tau_{k+1}).
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathrm{E}(T_{M_n}) &= n + \sum_j \sum_k \mathrm{P}(M_n = k + 1)\mathrm{P}(j < \tau_{k+1}) \\
&= n + \sum_k \mathrm{P}(M_n = k + 1) \sum_j \mathrm{P}(j < \tau_{k+1}) \\
&= n + \sum_k \mathrm{P}(M_n = k + 1)\mathrm{E}(\tau_{k+1}) \\
&= n + \sum_k \mathrm{P}(K_n = k)\frac{k + \theta}{\theta}
\end{aligned}$$

which can be simplified to

$$\mathrm{E}(T_{M_n}) = n + 1 + \mathrm{E}(K_n)\frac{1}{\theta} \tag{26}$$

Combining (25) and (26) gives the result (23):

$$EK_n^2 = EK_n + 2\theta(n - EK_n).$$

References

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

Barabasi, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286**, 509–512.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

Ferguson, T. (1973). A bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.

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

Jensen, S. and Liu, J. (2007). Bayesian clustering of transcription factor binding motifs. *Journal of the American Statistical Association* **Forthcoming**.

Liu, J. (1996). Nonparametric hierarchical Bayes via sequential imputations. *Annals of Statistics* **24**, 911–930.

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture models based clustering of gene expression profiles. *Bioinformatics* **18**, 1194–1206.

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

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.

Teh, Y. W. (2006). A hierarchical bayesian language model based on pitman-yor processes. In *ACL 2006*.

Zabell, S. L. (1992). Predicting the unpredictable. *Synthese* **90**, 205–232.