HHS Public Access

Author manuscript

*J Am Stat Assoc*. Author manuscript; available in PMC 2018 April 12.

Published in final edited form as:

*J Am Stat Assoc*. 2017 ; 112(518): 721–732. doi:10.1080/01621459.2016.1165103.

# Random Partition Distribution Indexed by Pairwise Information

**David B. Dahl**[a], **Ryan Day**[b], and **Jerry W. Tsai**[c]

[a]Department of Statistics, Brigham Young University, Provo, UT

[b]Livermore Computing, Lawrence Livermore National Laboratory, Livermore, CA

[c]Department of Chemistry, University of the Pacific, Stockton, CA

## Abstract

We propose a random partition distribution indexed by pairwise similarity information such that partitions compatible with the similarities are given more probability. The use of pairwise similarities, in the form of distances, is common in some clustering algorithms (e.g., hierarchical clustering), but we show how to use this type of information to define a prior partition distribution for flexible Bayesian modeling. A defining feature of the distribution is that it allocates probability among partitions within a *given* number of subsets, but it does *not* shift probability among sets of partitions with *different* numbers of subsets. Our distribution places more probability on partitions that group similar items yet keeps the total probability of partitions with a given number of subsets constant. The distribution of the number of subsets (and its moments) is available in closed-form and is not a function of the similarities. Our formulation has an explicit probability mass function (with a tractable normalizing constant) so the full suite of MCMC methods may be used for posterior inference. We compare our distribution with several existing partition distributions, showing that our formulation has attractive properties. We provide three demonstrations to highlight the features and relative performance of our distribution.

### Keywords

Bayesian nonparametrics; Chinese restaurant process; Cluster analysis; Nonexchangeable prior; Product partition model

## 1. Introduction

We propose a random partition distribution indexed by pairwise information for flexible Bayesian modeling. By way of introduction, consider Gibbs-type priors (De Blasi et al. 2015) which lead to a broad class of Bayesian nonparametric models for data $y_1, y_2 \ldots$:

$$y_i|\theta_i \sim p\left(y_i|\theta_i\right), \theta_i|F \sim F, F \sim Q, \quad (1)$$

**CONTACT** David B. Dahl, dahl@stat.byu.edu, Department of Statistics, Brigham Young University, Provo, UT 84602.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.

where $p(y|\theta)$ is a sampling distribution indexed by $\theta$, $F$ is a discrete random probability measure, and $Q$ is an infinite-dimensional prior distribution termed the de Finetti measure. The model can be enriched by indexing the sampling model by other parameters or by placing priors on hyperparameters defining the prior distribution $Q$. The sequence $\theta_1, \theta_2, \ldots$ in (1) is exchangeable and the discrete nature of $F$ implies that $\theta_1, \theta_2, \ldots$ will have ties with positive probability. Therefore, for any finite $n$, we can reparameterize $\theta_1, \ldots, \theta_n$ in terms of the unique values $\phi = (\phi_1, \ldots, \phi_{q_n})$ and a partition $\pi_n = \{S_1, \ldots, S_{q_n}\}$, a set whose subsets $S_1, \ldots, S_{q_n}$ are nonempty, mutually exclusive, and exhaustive such that $\cup_{S \in \pi_n} S = \{1, \ldots, n\}$. Two integers $i$ and $i'$ belong to $S_j$ if and only if $\theta_i = \theta_{i'} = \phi_j$. The parameters $\phi_1, \ldots, \phi_{q_n}$ are independent and identically distributed $G0$, the centering distribution of $Q$. $F$ in (1) implies a prior on $\pi_n$ having support $\mathscr{P}_n$ (the set of all possible partitions of $n$ items). A distribution over $\mathscr{P}_n$ is discrete, but the size of the space — which grows according to the Bell (1934) number — makes exhaustive calculations impossible except for very small $n$.

The choice of $Q$ leads to different exchangeable random partition models. For example, when $Q$ is the Dirichlet process (Ferguson 1973), the partition distribution $p(\pi_n)$ is the Ewens distribution (Ewens 1972; Pitman 1995, 1996) and the model in (1) is a Dirichlet process mixture model (Antoniak 1974). Or, when $Q$ is the Poisson-Dirichlet process (Pitman and Yor 1997), the partition distribution $p(\pi_n)$ is the Ewens-Pitman distribution (Pitman and Yor 1997) and the model in (1) is a Poisson-Dirichlet process mixture model.

In some situations, the random probability measure $F$ and the de Finetti measure $Q$ are not of interest and the model in (1) maybe marginalized as:

$$y_i|\theta_i \sim p\left(y_i|\theta_i\right), \theta_i = \sum_{j=1}^{q_n} \phi_j I\left\{i \in S_j\right\}, \phi_j \sim G_0, \ \pi_n \sim p\left(\pi_n\right). \tag{2}$$

Popular models for the partition distribution $p(\pi_n)$ include product partition models, species sampling models, and model-based clustering. These are reviewed by Quintana (2006) and Lau and Green (2007).

Exchangeable random partition models, which follow from the formulation in (1), have many attractive properties. For example, in exchangeable random partition models, the sequence of partition distributions with increasing sample size is marginally invariant: The partition distribution of $n$ items is identical to the marginal distribution of the first $n$ items after integrating out the last observation in the partition distribution of $n + 1$ items. Insisting on an exchangeable random partition distribution, however, imposes limits of the formulation of partition distributions (Lee et al. 2013).

The presence of item-specific information makes the exchangeability assumption on $\theta_1, \ldots, \theta_n$ unreasonable. Indeed the aim of this article is to explicitly explore a random probability model for partitions that uses pairwise information to a priori influence the partitioning. Since our partition distribution is nonexchangeable, there is no notion of an underlying de Finetti measure $Q$ giving rise to our partition distribution and our model lacks marginal invariance. We will show, however, how to make the data analysis invariant to the order in

which the data are observed. The use of pairwise distances is common in many ad hoc clustering algorithms (e.g., hierarchical clustering), but we show how to use this type of information to define a prior partition distribution for flexible Bayesian modeling.

Recent work has developed other nonexchangeable random partition models. A common thread is the use of covariates to influence a priori the probability for random partitions. Park and Dunson (2010) and Shahbaba and Neal (2009) include clustering covariates as part of an augmented response vector to obtain a prior partition model for inference on the response data. Park and Dunson (2010) build on product partition models and focus on continuous covariates treated as random variables, whereas Shahbaba and Neal (2009) use the Dirichlet process as the random partition and model a categorical response with logistic regression. Müller, Quintana, and Rosner (2011) proposed the PPMx model, a product partition model with covariates. In their simulation study of several of these approaches, they found no dominant method and suggested choosing among them based on the inferential goals. More recently, Airoldi et al. (2014) provided a general family of nonexchangeable species sampling sequences dependent on the realizations of a set of latent variables.

Our proposed partition distribution—which we call the Ewens-Pitman attraction (EPA) distribution—is indexed by pairwise similarities among the items, as well as a mass parameter $\alpha$ and a discount parameter $\delta$ which control the distribution of the number of subsets and the distribution of subset sizes. Our distribution allocates items based on their attraction to existing subsets, where the attraction to a given subset is a function of the pairwise similarities between the current item and the items in the subset. A defining feature of our distribution is that it allocates probability among partitions within a *given* number of subsets, but it does *not* shift probability among sets of partitions with *different* numbers of subsets. The distribution of the number of subsets (and its moments) induced by our distribution is available in closed-form and is invariant to the similarity information.

We compare our EPA distribution with several existing distributions. We draw connections with the Ewens and Ewens-Pitman distributions which result from the Dirichlet process (Ferguson 1973) and Pitman-Yor process (Pitman and Yor 1997), respectively. Of particular interest are the distributions in the proceedings article of Dahl (2008) and the distance dependent Chinese restaurant process (ddCRP) of Blei and Frazier (2011). Whereas our distribution *directly* defines a distribution over partitions through sequential allocation of items to subsets in a partition, both Dahl (2008) and Blei and Frazier (2011) *implicitly define* the probability of a partition by summing up the probabilities of all associated directed graphs whose nodes each have exactly one edge or loop. We will see that, although these other distributions use the same similarity information, our distributions behavior is substantially different. We will also contrast our approach with the PPMx model of Müller, Quintana, and Rosner (2011). Unlike the ddCRP and PPMx distributions, our partition distribution has both an explicit formula for the distribution of the number of subsets and a probability mass function with a tractable normalizing constant. As such, standard MCMC algorithms may be easily applied for posterior inference on the partition $\pi_n$ and any hyperparameters that influence partitioning. A demonstration, an application, and a simulation study all help to show the properties of our proposal and investigate its performance relative to leading alternatives.

## 2. Ewens-Pitman Attraction Distribution

### 2.1. Allocating Items According to a Permutation

Our EPA distribution can be described as sequentially allocating items to subsets to form a partition. The order in which items are allocated is not necessarily their order in the dataset; the permutation $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ of $\{1, \ldots, n\}$ gives the sequence in which the $n$ items are allocated, where the $t$th item allocated is $\sigma_t$. The sequential allocation of items yields a sequence of partitions and we let $\pi(\sigma_1, \ldots, \sigma_{t-1})$ denote the partition of $\{\sigma_1, \ldots, \sigma_{t-1}\}$ at time $t-1$. Let $q_{t-1}$ denote the number of subsets in $\pi(\sigma_1, \ldots, \sigma_{t-1})$. For $t = 1$, we take $\{\sigma_1, \ldots, \sigma_{1-1}\}$ to mean the empty set and item $\sigma_1$ is allocated to a new subset. At time $t > 1$, item $\sigma_t$ is allocated to one of the $q_{t-1}$ subsets in $\pi(\sigma_1, \ldots, \sigma_{t-1})$ or is allocated to a new subset. If $S$ denotes the subset to which item $\sigma_t$ will be allocated, then the partition at time $t$ is obtained from the partition at time $t-1$ as follows: $\pi(\sigma_1, \ldots, \sigma_t) = (\pi(\sigma_1, \ldots, \sigma_{t-1}) \setminus \{S\}) \cup S \cup \{\sigma_t\}\}$. Note that $\pi(\sigma_1, \ldots, \sigma_n)$ is equivalent to the partition $\pi_n$.

The permutation $\boldsymbol{\sigma}$ can be fixed (e.g., in the order the observations are recorded). Note, however, that our partition distribution does indeed depend on the permutation $\boldsymbol{\sigma}$ and it can be awkward that a data analysis depends on the order the data are processed. We recommend using the uniform distribution on the permutation, that is, $p(\boldsymbol{\sigma}) = 1/n!$ for all $\boldsymbol{\sigma}$. This has the effect of making analyses using the EPA distribution symmetric with respect to permutation of the sample indices, that is, the data analysis then does not depend on the order of the data.

### 2.2. Pairwise Similarity Function and Other Parameters

Our proposed EPA distribution uses available pairwise information to influence the partitioning of items. In its most general form, this pairwise information is represented by a similarity function $\lambda$. such that $\lambda(i, j) > 0$ for any $i, j \in \{1, \ldots, n\}$ and $\lambda(i, j) = \lambda(j, i)$. We note that the similarity function can involve unknown parameters and we later discuss how to make inference on these parameters. A large class of similarity functions can be defined as $\lambda(i, j) = f(d_{ij})$, where $f$ is a nonincreasing function of pairwise distances $d_{ij}$ between items $i$ and $j$. The metric defining the pairwise distances and the functional form of $f$ are modeling choices. For example, the reciprocal similarity is $f(d) = d^{-\tau}$ for $d > 0$. If $d_{ij} = 0$ for some $i \neq j$, one could add a small constant to the distances or consider another similarity function such as the exponential similarity $f(d) = \exp(-\tau d)$. We call the exponent $\tau \geq 0$ the temperature, as it has the effect of dampening or accentuating the distances. In addition to $\boldsymbol{\sigma}$ and $\lambda$, the EPA distribution is also indexed by a discount parameter $\delta \in [0, 1)$ and a mass parameter $\alpha > -\delta$, which govern distribution of the number of subsets and distribution of subset sizes.

### 2.3. Probability Mass Function

The probability mass function (p.m.f.) for a partition $\pi_n$ having the EPA distribution is the product of increasing conditional probabilities:

$$p(\pi_n | \alpha, \delta, \lambda, \sigma) = \prod_{t=1}^{n} p_t(\alpha, \delta, \lambda, \pi(\sigma_1, \ldots, \sigma_{t-1})), \quad (3)$$

where $p_t(\alpha, \delta, \lambda, \pi(\sigma_1, \ldots, \sigma_{t-1}))$ is one for $t = 1$ and is otherwise defined as

$$\begin{aligned} & p_t(\alpha, \delta, \lambda, \pi(\sigma_1, \ldots, \sigma_{t-1})) \\ & \quad = \mathrm{pr}(\sigma_t \in S | \alpha, \delta, \lambda, \pi(\sigma_1, \ldots, \sigma_{t-1})) \\ & \quad = \begin{cases} \frac{t-1-\delta q_{t-1}}{\alpha+t-1} \cdot \frac{\sum_{\sigma_s \in S} \lambda(\sigma_t, \sigma_s)}{\sum_{s=1}^{t-1} \lambda(\sigma_t, \sigma_s)} & \text{for } S \in \pi(\sigma_1, \ldots, \sigma_{t-1}) \\ \frac{\alpha+\delta q_{t-1}}{\alpha+t-1} & \text{for } S \text{ being a new subset.} \end{cases} \end{aligned} \quad (4)$$

At each step, $\sum_{s=1}^{t-1} \lambda(\sigma_t, \sigma_s)$ is the total attraction of item $\sigma_t$ to the previously allocated items. The ratio of the sums of the similarity function $\lambda$ in (4) gives the proportion of the total attraction of item $\sigma_t$ to those items allocated to subset $S$. As such, item $\sigma_t$ is likely to be allocated to a subset having items to which it is attracted. We note that our distribution is invariant to scale changes in the similarity $\lambda$, which aligns with the idea that similarity is a relative rather than an absolute concept.

### 2.4. Marginal Invariance

A sequence of random partition distributions in the sample size $n$ is marginally invariant (also known as consistent or coherent) if the probability distribution for partitions of {1, …, $n$} is the same as the distribution obtained by marginalizing out $n + 1$ from the probability distribution for partitions of {1, …, $n + 1$}. For a nontrivial similarity function λ, the proposed EPA distribution is not marginally invariant. We argue, however, that insisting on marginal invariance is too limiting in the context of pairwise similarity information.

Consider the following simple example with $n = 3$ items. Let $p_0$ be the partition distribution for $\pi_3$ obtained from (3), let $p_1$ be the distribution of the partition $\pi_2$ obtained by marginalizing $p_0$ over item 3, and let $p_2$ be the distribution of the partition $\pi_2$ in (3) assuming $n = 2$. Without loss of generality, assume $\alpha = 1$, $\lambda(1, 2) = 1$, $\lambda(1, 3) = a$, and $\lambda(2, 3) = b$. Using reciprocal similarity (i.e., distances $1/a$ and $1/b$) and the uniform distribution on the permutation $\sigma$, algebra shows that marginal invariance requires the similarities $a$ and $b$ are reciprocals of each other. This constraint is displayed graphically in Figure 1. Whereas one would like to be able to consider any placement of $x_3$, marginal invariance requires that $x_3$ lie on the Cassini oval. The conclusion is that requiring marginal invariance severely constrains the similarity information in ways that are not likely to be seen in practice. Of course, saying that two items are similar is relative to the other items being considered and, hence, the distribution should be allowed to change as more items are added. Marginal invariance should not be expected, or imposed, in the presence of pairwise similarity information. As such, a data analysis based on $n$ observations using our EPA distribution should be viewed as an analysis of just those observations.

### 2.5. Distributions on the Parameters

The EPA distribution is indexed by the mass parameter $\alpha$, the discount parameter δ, the similarity function $\lambda$, and the permutation $\sigma$. These parameters can be treated as known fixed quantities, or they may be treated as unknown random quantities having distributions.

The values at which they are fixed or their distributions are modeling choices. Here we give some suggestions. We recommend a gamma distribution for the mass parameter $\alpha$. Since the discount parameter $\delta \in [0, 1)$, one may consider a mixture of a point mass at zero and a beta distribution. A distribution may be placed on the parameters defining the similarity function $\lambda$. For example, if $\lambda(i, j) = d_{ij}^{-\tau}$, then a distribution for the temperature $\tau \geq 0$ could be a gamma distribution. As stated previously, we recommend a uniform distribution on the permutation.

### 2.6. Sampling Independent and Identically Distributed Partitions

Section 5 discusses posterior simulation for our EPA distribution. Here we describe prior simulation, specifically, how to sample independent and identically distributed (iid) partitions from the EPA distribution. To obtain a single random partition $\pi_n$, first sample values for any of the parameters $\alpha$, $\delta$, $\lambda$, and $\sigma$ that are not fixed. In the case of the uniform distribution on the permutations, a random permutation $\sigma$ is obtained by sorting 1, …, $n$ according to uniformly-distributed random numbers on the unit interval or through standard functions in software. Finally, sample the partition $\pi_n$ itself from (3) by sequentially applying the increasing conditional probabilities in (4). This process can be repeated many times to obtain multiple iid partitions and the process can easily be parallelized over multiple computational units. In a similar manner, iid samples can also be obtained from the ddCRP and PPMx priors.

## 3. Influence of the Parameters

### 3.1. Mass and Discount Govern the Distribution of Number of Subsets

The proposed EPA distribution is a probability distribution for a random partition $\pi_n = \{S_1, \ldots S_{q_n}\}$ and, therefore, produces a probability distribution on the number of subsets $q_n$. The distribution of $q_n$ has a recursive expression that we now give. Note that the mass parameter $\alpha$, together with the discount parameter $\delta$ and the number of subsets at time $t-1$ (i.e., $q_{t-1}$), governs the probability of opening a new subset for the $t$th allocated item. Taken over the subsets in $\pi(\sigma_1, \ldots, \sigma_{t-1})$, the similarity proportions in (4) sum to one, and consequently the probability that $\sigma_t$ is allocated to an existing subset is $(t - 1 - \delta_{q_{t-1}})/(\alpha + t - 1)$ and the probability that it is allocated to an empty subset is $(\alpha + \delta_{q_{t-1}})/(\alpha + t - 1)$. Applying this for every $\sigma_1, \ldots, \sigma_n$, we have the p.m.f. for the number of subsets $q_n$ being:

$$\text{pr}(q_n = k | \alpha, \delta) = \frac{\alpha f(n-1,\ k-1,\ 1,\ 1)}{\prod_{t=1}^{n}(\alpha t - 1)},\ \text{for } k = 1, \ldots, n, \tag{5}$$

where $f(a, 0, c, d) = \prod_{t=0}^{a-1}(c + t - \delta d)$, $f(a, a, c, d) = \prod_{t=0}^{a-1}(\alpha + \delta(d + t))$, and otherwise:

$$f(a, b, c, d) = (\alpha + \delta d) f(a - 1, b - 1,\ c+1,\ d+1) + (c - \delta d) f(a - 1,\ b,\ c+1,\ d).$$

Note that the distribution of $q_n$ does not depend on the similarity function $\lambda$ nor on the permutation $\sigma$. Thus, our EPA distribution uses pairwise similarity information to allocate

probability among partitions within a *given* number of subsets, but it does *not* shift probability among sets of partitions with *different* numbers of components. In modeling a random partition $\pi_n$, this fact provides a clear separation of the roles of: (i) the mass parameter $\alpha$ and discount parameter $\delta$ and (ii) the pairwise similarity function $\lambda$ and permutation $\sigma$.

The mean number of subsets is the sum of the success probabilities of dependent Bernoulli random variables obtained by iterated expectations, yielding

$$E\left(q_n|\alpha,\delta\right)=\sum_{t=1}^{n}w_t,\ \text{where}\ w_1=1\ \text{and}$$

$$w_t=\frac{\alpha+\delta\sum_{s=1}^{t-1}w_s}{\alpha+t-1}\ \text{for t}>1. \quad (6)$$

Figure 2 shows, for various values of the mass parameter $\alpha$ and discount parameter $\delta$, how the mean number of subsets increases as the number of items $n$ increases. Note that the rate of growth can vary substantially with $\alpha$ and $\delta$. Other moments (such as the variance) can be calculated from their definitions using the p.m.f. in (5).

Whereas the expectation in (6) scales for large $n$, evaluating the p.m.f. in (5) becomes prohibitive for large $n$ and moderate $k$. Alternatively, Monte Carlo estimates of the distribution of the number of subsets and its moments can be obtained by simulation. Samples of $q_n$ can be drawn by counting the number of subsets in randomly obtained partitions using the algorithm in Section 2.6. Even faster, a random draw for $q_n$ is obtained by counting the number of successes in $n$ dependent Bernoulli trials having success probability $(\alpha + \delta r)/(\alpha + t - 1)$, where $t = 1, \ldots, n$ and $r$ initially equals 0 and increments with each success.

In the special case that the discount $\delta$ is equal to zero, (5) simplifies to

$$\text{pr}\left(q_n=k|\alpha,\delta\right)=\frac{\alpha^k\left|s\left(n,k\right)\right|}{\prod_{t=1}^{n}\left(\alpha+t-1\right)}\ \text{for}\ k=1,\ldots,n, \quad (7)$$

where $|s(n, k)|$ is the Stirling number of the first kind. Recall that $|s(n, k)| = (n - 1)|s(n - 1, k)| + |s(n - 1, k - 1)|$ with initial conditions $|s(0, 0)| = 1$ and $|s(n, 0)| = |s(0, k)| = 0$. Since the $n$ Bernoulli random variables are now independent with success probability $\alpha/(\alpha + t - 1)$, the expectation formula simplifies and the variance is available:

$$\text{E}\left(q_n\right)=\sum_{t=1}^{n}\frac{\alpha}{\alpha+t-1}\ \text{and}\ \text{var}\left(q_n\right)=\sum_{t=1}^{n}\frac{\alpha\left(t-1\right)}{\left(\alpha+t-1\right)^2}. \quad (8)$$

We review the Ewens distribution, Chinese restaurant process, and Dirichlet process in Section 4.1 and there note that the expressions in (7) and (8) are the same for these distributions.

### 3.2. Effect of Similarity Function

We now study the influence of the similarity function λ. As shown in Section 3.1, a feature of our approach is that the distribution of the number of subsets is not influenced by λ.

**Result 1—**For any number of items $n$, mass $\alpha$, discount δ, and permutation $\boldsymbol{\sigma}$, the probability that items $i$ and $j$ are in the same subset is increasing in their similarity $\lambda(i, j)$, holding all other similarities constant.

This result is proved as follows. Let $I\{W\}$ be the indicator function of the event $W$ and let $C_{i,j}$ be the event that items $i$ and $j$ are in the same subset. Recall that $\mathscr{P}_n$ is the set of all partitions of $n$ items. The task is to show that $\Pr(C_{i,j} \mid \alpha, \delta, \lambda, \boldsymbol{\sigma}) = f(\alpha, \delta, \lambda, \boldsymbol{\sigma})$ is increasing in $\lambda(i, j)$. Without loss of generality, assume that $\boldsymbol{\sigma}$ is such that $j$ is allocated before $i$. Let $t_i$ be the time in which item $i$ is allocated, and note that $\sigma_{t_i}$ Then,

$$f(\alpha, \delta, \lambda, \sigma) = \sum_{\pi_n \in \mathscr{P}_n} \mathrm{I}\{C_{i,j}\}\, p(\pi_n | \alpha, \delta, \lambda, \sigma) = \sum_{\pi_n \in \mathscr{P}_n} \mathrm{I}\{C_{i,j}\}\, c^1_{\pi_n} p_{t_i}(\alpha, \delta, \lambda, \pi(\sigma_1, \ldots, \sigma_{t_{i-1}})),$$

where $c^1_{\pi_n}$ and later $c^2_{\pi_n}$ are positive constants with respect to $\lambda(i, j)$. Let $S^{j,\pi_n}$ denote the subset in $\boldsymbol{\pi}_n$ containing $j$. By (4),

$$\begin{aligned} & f(\alpha, \delta, \lambda, \sigma) \\ &= \sum_{\pi_n \in \mathscr{P}n} c^1_{\pi_n} \mathrm{pr}\left(i \in S^{j,\pi_n} | \alpha, \delta, \lambda, \pi(\sigma_1, \ldots, \sigma_{t_{i-1}})\right) \\ &= \sum_{\pi_n \in \mathscr{P}n} c^2_{\pi_n} \frac{\sum_{\sigma_s \in j,\pi_n} \lambda(\sigma_{t_i}, \sigma_s)}{\sum_{s=1}^{t_{i-1}} \lambda(\sigma_{t_i,\sigma_s})} \\ &= \sum_{\pi_n \in \mathscr{P}_n} c^2_{\pi_n} \frac{\lambda(i, j) + \sum_{\sigma_s \in S^{j,\pi_n}} \mathrm{I}\{\sigma_s \neq j\} \lambda(i, \sigma_s)}{\lambda(i, j) + \sum_{s=1}^{t_{i-1}} \mathrm{I}\{\sigma_s \neq j\} \lambda(i, \sigma_s)}. \end{aligned} \quad (9)$$

Since the numerator in each fraction is less than or equal to the denominator, each element of the sum is increasing in $\lambda(i, j)$. The proof is completed by noting that the sum of increasing functions is also increasing.

**Result 2—**For any number of items $n$, if a distribution is placed on the mass $\alpha$, the discount $\delta$, and the permutation $\boldsymbol{\sigma}$, then the marginal probability that items $i$ and $j$ are in the same subset is increasing in their similarity $\lambda(i, j)$, holding all other similarities constant.

We establish this result as follows. Let $p(\alpha, \delta, \boldsymbol{\sigma})$ be the joint distribution of $\alpha$, $\delta$, and $\boldsymbol{\sigma}$ and let $f(\lambda)$ denote the marginal probability of interest. The task is to show that $f(\lambda)$ is increasing in $\lambda(i, j)$. It is sufficient to show that its derivative is greater than zero. Note that:

$$\frac{d}{d\lambda(i,j)} f(\lambda) = \frac{d}{d\lambda(i,j)} \int f(\alpha, \delta, \lambda, \sigma) d\alpha\, d\delta\, d\sigma = \int \left[ \frac{d}{d\lambda(i,j)} f(\alpha, \delta, \lambda, \sigma) \right] \times p(\alpha, \delta, \sigma)\, d\alpha\, d\delta\, d\sigma > 0,$$

because the derivative of $f(\alpha, \delta, \lambda, \sigma) > 0$ (since it is increasing in $\lambda(i, j)$) and the expectation of a positive random variable is positive. Switching the order of operations is justified since $f(\alpha, \delta, \lambda, \sigma)$ is continuous in $\lambda(i, j)$ for every $\alpha$, $\delta$, and $\sigma$, and $f(\alpha, \delta, \lambda, \sigma)$ $p(\alpha, \delta, \sigma)$ is nonnegative and less than or equal to $p(\alpha, \delta, \sigma)$, which is itself integrable.

Results 1 and 2 establish monotonicity in $\lambda(i, j)$ of the probability that any two items $i$ and $j$ are in the same subset. One might naively expect that $\lambda(i, j) < \lambda(i, k)$ would imply $\Pr(C_{i,j}) < \Pr(C_{i,k})$. While this generally holds, examples can be contrived that contradict this statement. The explanation is that the probability that $i$ and $j$ belong to the same subset is not only determined by $\lambda(i, j)$, but also by other parameters and the ensemble of information in the similarity function $\lambda$, including the similarities $\lambda(i, l)$ and $\lambda(j, l)$ for $l \in \{1, \ldots, n\}$.

## 4. Comparison to Other Partition Distributions

We now examine the relationship between our proposed EPA distribution for a random partition $\pi_n$ and other random partition distributions. In particular, we compare and contrast the EPA distribution with the Ewens distribution, the Ewens-Pitman distribution, and two other distributions influenced by pairwise distances. Figure 3 summarizes the relationship between our EPA distribution and the Ewens and Ewens-Pitman distributions.

### 4.1. Comparison to the Ewens and Ewens-Pitman Distributions

First, consider the special case that the discount $\delta$ is 0 and the similarity function $\lambda(i, j)$ is constant for all $i$ and $j$. The ratio of the sums of similarities in (4) reduces to $|S|/(t-1)$ and, since $\delta = 0$, (4) itself reduces to:

$$\text{pr}\ (\sigma_t \in S | \alpha, \pi(\sigma_1, \ldots, \sigma_{t-1})) = \begin{cases} \frac{|S|}{\alpha+t-1} \text{for } S \in \pi(\sigma_1, \ldots, \sigma_{t-1}) \\ \frac{\alpha}{\alpha+t-1} \text{for } S \text{ being a new subset.} \end{cases} \quad (10)$$

This is known as “Ewens’ sampling formula,” a particular predictive probability function (Pitman 1996). Its product over $\sigma$ results in a partition distribution called the Ewens distribution, which is the partition distribution from the Dirichlet process and is also known as the partition distribution of the Chinese restaurant process (CRP), a discrete-time stochastic process on the positive integers. The metaphor to a Chinese restaurant first appeared in Aldous et al. (1985, pp. 91–92) and is credited to Jim Pitman and Lester E. Dubins.

We note that the distribution of the number of subsets $q_n$ in (7) and the mean and variance in (8) apply to the Ewens distribution, Chinese restaurant process, and Dirichlet process —just as they apply to our proposed EPA distribution when $\delta = 0$, for any similarity function $\lambda$ and permutation $\sigma$. In fact, Arratia, Barbour, and Tavaré (2003) provide equivalent expressions to (7) and (8) in their study of Ewens’ sampling formula. Therefore, the role of, interpretation

of, and intuition regarding the mass parameter $\sigma$ that one has for these established models carries over directly to the EPA distribution.

Second, consider the special case that, again, the similarity function $\lambda(i, j)$ is constant for all $i$ and $j$, but the discount parameter 5 is not necessarily zero. Then, (4) reduces to

$$\text{pr}\left(\sigma_t \in S \middle| \alpha, \delta, \pi\left(\sigma_1, \ldots, \sigma_{t-1}\right)\right) = \begin{cases} \frac{\left|S\right| - \delta q_{t-1} \left|S\right|(t-1)}{\alpha + t - 1} & \text{for } S \in \pi\left(\sigma_1, \ldots, \sigma_{t-1}\right) \\ \frac{\alpha + \delta q_{t-1}}{\alpha + t - 1} & \text{for } S \text{ being a new subset.} \end{cases} \tag{11}$$

Contrast that with the "two-parameter Ewens' sampling formula" of Pitman (1995):

$$\text{pr}\ \left(\delta_t \in S \middle| \alpha, \delta, \pi\left(\sigma_1, \ldots, \sigma_{t-1}\right)\right) = \begin{cases} \frac{|S| - \delta}{\alpha + t - 1} & \text{for } S\ \in\ \pi\left(\sigma_1, \ldots, \sigma_{t-1}\right) \\ \frac{\alpha + \delta q_{t-1}}{\alpha + t - 1} & \text{for } S \text{ being a new subset.} \end{cases} \tag{12}$$

Sequentially applying (12) results in what we refer to as the Ewens-Pitman distribution. This distribution is also known as the partition distribution of the two-parameter Chinese restaurant process and the partition distribution from the Poisson-Dirichlet process. Comparing (11) and (12) we see that, whereas the Ewens-Pitman distribution applies the discount 5 uniformly to small and large subsets alike, the EPA distribution under constant similarities applies the discount proportional to the relative size of the subset times the number of subsets. This difference in the application of the discount 5 leads to somewhat different large sample behavior. We use two univariate summaries of a partition $\pi_n$ to illustrate this difference: (i) entropy: $-\sum_{S \in \pi_n} \left(|S|\right)/n \log\left(|S|/n\right)$ and (ii) proportion of singleton subsets: $\sum_{S \in \pi_n} \text{I}\left\{|S| = 1\right\} / |\pi_n|$. Figure 4 illustrates the limiting behavior of these two distributions for various combinations of the mass parameter $\alpha$ and the discount parameter $\delta$.

For any mass parameter $\alpha$ and discount parameter $\delta$, the probability of assigning to the new subset is the same for both our EPA distribution and the Ewens-Pitman distribution, regardless of the similarity function $\lambda$ and the permutation $\sigma$ used for our distribution. As such, the distribution of the number of subsets $q_n$ in (5) and the mean in (6) apply to the Ewens-Pitman distribution, the two-parameter Chinese restaurant process, and the Pitman-Yor process (Teh 2006), just as they apply to our proposed EPA distribution. In summary, whereas the large sample behavior of the entropy and proportion of singletons differ, the distribution of the number of subsets is exactly the same. Therefore, the role of, interpretation of, and intuition regarding the mass parameter $\alpha$ and discount parameter $\delta$ regarding their influence on the number of subsets that one has for these established models carries over directly to the EPA distribution.

### 4.2. Comparison to Distance Dependent Chinese Restaurant Processes

Our EPA distribution resembles the distribution in the proceedings article of Dahl (2008) and the distance dependent Chinese restaurant process (ddCRP) of Blei and Frazier (2011).

The key difference between our EPA distribution and these others is how they arrive at a distribution over partitions. The EPA distribution *directly* defines a distribution over partitions through sequential allocation of items to subsets in a partition. In contrast, both Dahl (2008) and Blei and Frazier (2011) define a distribution over a directed graph in which the $n$ nodes have exactly one edge or loop, and the disjoint subgraphs form the subsets for the implied partition. There is a many-to-one mapping from these graphs to partitions and the probability of a given partition is *implicitly* defined by summing up the probabilities of graphs that map to the partition of interest. In the ddCRP distribution, the probability that item $i$ has a directed edge to item $j$ is proportional to $\lambda(i, j)$ for $j \neq i$ and is proportional to $\alpha$ for $j = i$. The similarities may be zero and non symmetric. The probabilities of a directed edge for an item is independent of all other edges. Because of the asymmetry and the many-to-one nature, the probability of a directed edge from $i$ to $j$ is not the probability that items $i$ and $j$ are in the same subset of the partition.

The size of the set of graphs is $n^n$ because each of the $n$ items can be assigned to any one of the $n$ items. Finding the probability for all possible graphs that map to a given partition $\pi_n$ quickly becomes infeasible for moderately sized $n$ and, thus, algorithms that require the evaluation of partition probabilities cannot be used. For example, although Blei and Frazier (2011) provide a Gibbs sampling algorithm for posterior inference in the ddCRP, it is not clear how to implement more general sampling strategies, for example, split-merge updates (Jain and Neal 2004,2007; Dahl 2003) which require evaluating the probability of a partition. In contrast, multivariate updating strategies can be applied to the EPA distribution because its p.m.f. for $\pi_n$ is easily calculated. Bayesian inference also requires the ability to update other parameters (e.g., those of the sampling model) and hyperparameters (e.g., mass parameter $\alpha$ and the temperature $\tau$). Standard MCMC update methods can be used for models involving the EPA distribution. In contrast, the ddCRP generally requires approximate inference for hyperparameters through Griddy Gibbs sampling (Ritter and Tanner 1992). Further, since the EPA distribution has an explicit p.m.f, the distribution of the number of subsets and its moments are available in closed form (Section 2.6), but this is not the case for the ddCRP. Computational issues aside, the ddCRP does not have a discount parameter 5 and thus does not have the same flexibility of the EPA distribution shown in Figure 4.

We illustrate stark differences between the EPA and ddCRP distributions using an example dataset in Section 6.1. There we show that, unlike the EPA distribution, the ddCRP has no clear separation between the mass parameter $\alpha$ and the similarity function $\lambda$ in determining the number of subsets. As we will see, even though the two distributions make use of the same similarity information, they arrive at fundamentally different partition distributions.

### 4.3. Comparison to PPMx Model

Müller, Quintana, and Rosner (2011) proposed the PPMx model, a product partition model in which the prior partition distribution has the form: $p(\pi_n|\mathbf{w}_1,\ldots,\mathbf{w}_n) \propto \prod_{j=1}^{qn} g\left(\mathbf{w}^j\right) c(S_j)$, where $c(\cdot)$ is a cohesion as in a standard product partition model, $g(\cdot)$ is a similarity function defined on a set of covariates, and $\mathbf{w}^j = \{\mathbf{w}_i : i \in S_j\}$. Although any cohesion may be used, the default is that of the Ewens distribution: $c(S) = \alpha\Gamma(|S|)$. If, in addition, $g(\cdot)$ is the

marginal distribution from a probability model for the covariates, then the partition distribution $p(\boldsymbol{\pi}_n|\boldsymbol{w}_1,\ldots,\boldsymbol{w}_n)$ is symmetric with respect to permutation of sample indices and is marginally invariant (as defined in Section 3.2). Müller, Quintana, and Rosner (2011) suggested default choices (depending on the type of the covariates) for the similarity function that guarantee these properties.

By way of comparison, our EPA distribution with a uniform prior of the permutation $\boldsymbol{\sigma}$ is also symmetric, but is not marginally invariant. On the other hand, the hyperparameters in the probability model on the covariates can heavily influence the partitioning process, but they are generally fixed in the PPMx model because posterior inference is complicated by an intractable normalizing constant. In contrast, posterior inference on hyperparameters in the EPA distribution is straightforward. As with the ddCRP but unlike our EPA distribution, the PPMx models does not have a clear separation between the mass parameter $\alpha$ and the covariates in determining the number of subsets. More generally, it is also not clear how to balance the relative effect of the covariates $g(\cdot)$ and the cohesion $c(\cdot)$ in the PPMx model. Finally, one can always define pairwise similarity information from item-specific covariates, but not all pairwise similarity information can be encoded as a function $g(\cdot)$ of item-specific covariates, as required by the PPMx model. As such, the EPA distribution can accommodate a wider class of information to influence partitioning.

## 5. Posterior Inference

In Bayesian analysis, interest lies in the posterior distribution of parameters given the data. The posterior distribution is not available in closed-form for the current approach, but a Markov chain Monte Carlo (MCMC) algorithm is available, as we now describe. This algorithm systematically updates parts of the parameter space at each iteration and performs many iterations to obtain samples from the posterior distribution.

First, consider the update of the partition $\boldsymbol{\pi}_n$ given the *data* $\boldsymbol{y}$ and all the other parameters. Because the model is not exchangeable, the algorithms of Neal (2000) for updating a partition $\boldsymbol{\pi}_n$ do not hold. As the p.m.f. is available, one could use a sampler that updates the allocation of many items simultaneously (e.g., a merge-split sampler (Jain and Neal 2004, 2007; Dahl 2003)). Here we use a Gibbs sampler (Gelfand and Smith 1990). To describe this sampler, suppose the current state of the partition is $\pi_n = \{S_1, \ldots, S_{q_n}\}$ and let $S_1^{-i}, \ldots, S_{qn}^{-i}$ be these subsets without item $i$. Let $\pi_n^{i\to j}$ be the partition obtained by moving $i$ from its current subset to the subset $S_j^{-i}$. Further, let $\pi_n^{i\to 0}$ denote the partition obtained by moving $i$ from its current subset to an empty subset $S_0^{-i}$. The full conditional distribution for the allocation of item $i$ is

$$p\left(i \in S_j^{-i}|\cdot\right) \propto p\left(\pi_n^{i\to j}|\alpha,\delta,\lambda,\sigma\right) p\left(y_i|\phi_j\right) \text{ for } j=0,1,\ldots,q_n, \quad (13)$$

where $\boldsymbol{\phi}_0$ is a new, independent draw from $G_0$ at each update. Note that $p\left(\pi_n^{i\to j}|\alpha,\delta,\lambda,\sigma\right)$ is calculated by evaluating (3) and (4) at the partition $\pi_n^{i\to j}$.

Because the p.m.f. of a partition $\pi_n$ is easily calculated, standard MCMC schemes are available for updating other parameters, including $\alpha$, $\delta$, $\lambda$., $\sigma$, and $\phi = (\phi_1, \ldots, \phi_{q_n})$. Herewemakea few notes. We suggest proposing a new permutation $\sigma^*$ by shuffing $k$ randomly chosen integers in the current permutation $\sigma$, leaving the other $n - k$ integers in their current positions. Being a symmetric proposal distribution, the proposed $\sigma^*$ is accepted with probability given by the minimum of 1 and the Metropolis ratio $(p(\pi_n \mid \alpha, \delta, \lambda, \sigma^*)\, p(\sigma^*))/(p(\pi_n / \alpha, \delta, \lambda, \sigma)\, p(\sigma))$, which reduces to $p(\pi_n / \alpha, \delta, \lambda, \sigma^*)/p(\pi_n / \alpha, \delta, \lambda, \sigma)$ when the prior permutation distribution $p(\sigma)$ is uniform. As $k$ controls the amount of change from the current permutation $\sigma$, the acceptance rate tends to decrease as $k$ increases. If the similarity function $\lambda$ involves hyperparameters, such as a temperature $\tau$, a Gaussian random walk is a natural sampler to use. Likewise, a Gaussian random walk can be used to update the mass parameter $\alpha$ and the discount parameter $\delta$. When $\delta = 0$, the distribution of the number of subsets is the same as in Dirichlet process mixture models and, as such, the Gibbs sampler of Escobar and West (1995) for updating the mass parameter $\alpha$ also applies to the EPA distribution.

Now consider updating $\phi = (\phi_1, \ldots, \phi_{q_n})$ given the data and the other parameters. This update is the same as in any other random partition model. For $j = 1, \ldots, q_n$, update $\phi_j$ using its full conditional distribution

$$p\left(\phi_j | y_i : i \in S_j\right) \propto p\left(\phi_j\right) \prod_{i \in S_j} p\left(y_i | \phi_j\right),$$

where $p(\phi)$ is the density of the centering distribution $G_0$. This full conditional distribution can usually be sampled directly if $G_0$ is conjugate to the sampling model $p(y / \phi)$. If not, any other valid MCMC update can be used, including a Metropolis-Hastings update.

Finally, we consider a sampling scheme for the estimation of $p(y_{n+1} \mid y_1, \ldots, y_n)$, the density of a new observation $y_{n+1}$ whose similarities $\lambda(n + 1, j)$ are available for $j = 1, \ldots, n$. Pick an initial value $y_{n+1}$. Use the posterior sampling procedure as described previously but also update the value of $y_{n+1}$ at each iteration by sampling $y_{n+1}$ using the current value of its model parameter $\theta_{n+1}$. Let $\theta_{n+1}^{(b)}$ denote the value of this model parameter for the observation $y_{n+1}$ at iteration $b$. Under squared error loss, the Bayes estimate of $p(y_{n+1} \mid y_1, \ldots, y_n)$ based on $B$ samples from the MCMC scheme is $\sum_{b=1}^{B} p\left(y_{n+1} | \theta_{n+1}^{(b)}\right) / B$.

## 6. Demonstrations

### 6.1. Arrests Dataset

In this section, we illustrate properties of the EPA distribution and compare its behavior to the ddCRP of Blei and Frazier (2011) using the “USArrests” dataset in R. We see that the two distributions use the same similarity information to arrive at fundamentally different partition distributions. As the temperature $\tau$ increases, the EPA distribution smoothly moves away from the Ewens distribution, placing more probability on partitions that group items with small distances (and that separate those with large distances), yet keeping the total

probability of partitions with a given number of subsets constant. In contrast, the ddCRP does not correspond to the Ewens when $\tau = 0$ and, as temperature goes to infinity, it collapses all probability to the partition with each item in its own singleton subset.

The "USArrests" dataset contains statistics on "arrests per 100,000 residents for assault, murder, and rape in each of the 50 United States in 1973" and "the percent of the population living in urban areas." The Euclidean distances between the four-dimensional standardized data vectors of $n = 5$ selected states are used. For both distributions, we use the exponential similarity function $\lambda(i, j) = f(d_{ij}) = \exp(-\tau d_{ij})$ and let $\alpha = 2$. In addition, for the EPA distribution, let $\delta = 0$ and $p(\sigma) = 1/n!$. We compute the probability of each of the $B(5) = 52$ possible partitions of the five states for a range of temperatures.

The evolution — as temperature $x$ increases — of the probabilities of the 52 partitions are displayed in the left panel of Figure 5. The cumulative probabilities of the partitions for the five states are displayed horizontally, and the ordering of the partitions is consistent across temperatures. For each partition, the cumulative probabilities across temperatures are joined to form the curves and the probability of a given partition is the difference between curves. The curves of several interesting partitions are identified with capital letters. Temperature $x = 0$ corresponds to the partition distribution of the Ewens distribution since $\lambda(i, j)$ is constant when $\tau = 0$. As the temperature increases, the pairwise distances become more influential and eventually the EPA distribution has appreciable probability on several partitions and virtually no probability for others. For example, whereas the partition "J" in Figure 5 has probability about 0.01 when $\tau = 0$ (corresponding to the Ewens distribution), it grows about 10 fold in probability when $\tau = 4$ because this partition matches well the pairwise distance information. Therefore, in the EPA distribution, the temperature $x$ controls the degree to which the prior distance information influences the partition distribution. The left panel of Figure 5 also shows that the aggregate probability for partitions with 1, 2, 3, 4, and 5 subsets is constant across temperature, illustrating a key feature of the EPA distribution discussed in Section 3.1: Our distribution allocates probability among partitions within a *given* number of subsets, but it does *not* shift probability among sets of partitions with *different* numbers of subsets.

The right-hand side of Figure 5 is the same plot for the ddCRP using the same value for mass parameter $\alpha$, the same distance information $d_{ij}$, and the same similarity function $\lambda(i, j) = f(d_{ij}) = exp(-\tau d_{ij})$. Capital letters label the same partitions for both the left- and right-hand sides of the figure. In contrast with the EPA distribution, the ddCRP: (i) does not correspond to the Ewens distribution with $\tau = 0$, (ii) the distribution of the number of subsets is heavily influenced by the temperature $\tau$, and (iii) partition 'K' initially dominates but partition 'A' eventually absorbs all the probability mass when $\tau \to \infty$. We thus see that, even with the same inputs, the EPA and ddCRP have fundamentally different properties and our EPA distribution adds to the set of available prior distributions that one can choose.

### 6.2. Bayesian Density Estimation for Dihedral Angles

We now demonstrate the EPA distribution as a prior partition distribution in Bayesian density estimation for protein structure prediction and find that using the EPA distribution significantly improves prediction over competing methods. A protein is a string of amino

acids that together adopt unique three-dimensional conformations (i.e., structures) to allow the protein to carry out its biochemical function. While it is relatively easy to determine the amino acid sequence of the protein, solving its structure is more challenging. A proteins structure can largely be characterized by the ($\phi$, $\psi$) torsion angles at each amino acid position. The task of protein structure prediction is simplified if, for a given protein family, the distribution of ($\phi$, $\psi$) angles at each position can be estimated. The sine model (Singh, Hnizdo, and Demchuk 2002) of the bivariate von Mises distribution is a model for ($\phi$, $\psi$) angles

$$p\left((\phi,\psi)\,|\mu,\nu,\kappa_1,\kappa_2,\lambda\right)=C\exp\left\{\kappa_1\cos\ (\phi-\mu)+\kappa_2\cos\ (\psi-\nu)+\lambda\sin(\phi-\mu)\sin\ (\psi-\nu)\right\},$$

where $C^{-1}=4\pi^2\sum_{m=0}^{\infty}\binom{2m}{m}\left(\frac{\lambda^2}{4\kappa_1\kappa_2}\right)^m I_m(\kappa_1)\,I_m(\kappa_2)$, $\phi, \psi, \mu, \nu \in (-\pi, \pi]$, $\kappa_1, \kappa_2 > 0$, and $\lambda \in (-\infty, \infty)$. Note that $I_m(x)$ is the modified Bessel function of the first kind of order $m$. Lennox et al. (2009) used the sine model as a kernel in a Dirichlet process mixture model for nonparametric density estimation of a ($\phi$, $\psi$) distribution. In the notation of (2), $y_i = (\phi_i, \psi_i)$ and $\theta_i = (\mu_i, \nu_i, \kappa_{1i}, \kappa_{2i}, \lambda_i)$. For the centering distribution $G_0$, we use the product of a bivariate uniform distribution on $(-\pi, \pi] \times (-\pi, \pi]$ (for $\mu$, $\nu$) and a bivariate Wishart distribution with shape 2 and rate matrix 0.25 $I_2$ (for $\kappa_1$, $\kappa_2$, $\lambda$), where $I_2$ is the 2 × 2 identity matrix and the mean is therefore 0.5 $I_2$.

In this demonstration, our data are ($\phi$, $\psi$) angles for 94 members of the globin family at aligned positions 93,94,95,104, 105, and 106, based on the default multiple sequence alignment from MUSCLE 3.8.31 (Edgar 2004). While Lennox et al. (2009) models the ($\phi$, $\psi$) distribution of a protein at a specific position based on angular data, their use of the Ewens distribution for the prior partition distribution $p(\pi_n)$ does not take advantage of the known amino acid sequence of the protein of interest. Here we replace the Ewens distribution with several specifications of our EPA distribution, the PPMx model, and a simple data-subsetting approach, all of which use the known amino acid sequence. Thus we mimic the task of protein structure prediction by using amino acid sequences to inform a prior partition distribution, resulting in density estimates tailored to a specific protein.

For each model described below, 20 independent Markov chains were run using the MCMC sampling algorithm described in Section 5 with 27,500 scans, discarding the first 2500 as burn-in and applying 1-in-5 thinning. Half of the 20 chains were initialized with a partition having all observations in their own subsets and the other half were initialized with all observations in the same subset. For each model and position, we compute the log pseudo marginal likelihood (LPML), that is, the sum of conditional predictive ordinates (Geisser and Eddy 1979; Gelfand 1996) across the 94 proteins. This evaluation criterion employs leave-one-out cross-validation to compare the predicted densities to the actual observed angle pairs. All comparisons are relative to the model using the Ewens prior partition distribution with mass parameter $\alpha$ fixed at 1.0. Table 1 provides the difference between the mean LPML values for each model discussed below and the baseline model using the Ewens prior partition distribution. Large positive values in the table indicate better fit to the data, with differences larger than a few units generally being statistically significant.

Our baseline specification of the EPA distribution has similarity function $\lambda(i, j)$ being one plus the mean BLAST bit score between the amino acid sequences for proteins $i$ and $j$. A BLAST bit score is a pairwise measure of similarity between two proteins. It is large for a pair of proteins having similar amino acid sequences and small otherwise. Twenty-seven percent of the similarities are 1 and the remaining have a five-number summary of (9, 32, 55, 110, 306). The temperature $\tau$ has a Gamma (2, 0.5) prior (with mean 4) and we use a uniform prior on the permutation $\boldsymbol{\sigma}$. We fix the discount $\delta$ at 0.0 and, as with the Ewens distribution, the mass parameter $\alpha$ is fixed at 1.0. When updating the permutation $\boldsymbol{\sigma}$, the sampling algorithm proposes to update $k = 46$ items and the mean acceptance rate is about 30%. When updating the temperature $\tau$, a Gaussian random walk proposal is used with standard deviation 2.0 and the mean acceptance rate is 59%. Diagnostics indicate that the Markov chains mix well. The LPML results for this model are found in row 2 of Table 1. There is substantial improvement at all the positions except positions 104 and 105. At these positions, the performance is about that of the Ewens distribution because the $(\phi, \psi)$ distributions are highly concentrated in one region and, therefore, the amino acid sequence information is not helpful in prediction.

The BLAST bit scores are not compatible with the recommended similarity functions for the PPMx model because they are not individual-specific covariates, but rather a measure of similarity between two proteins. It is therefore not obvious how to incorporate them in the PPMx framework. We can, however, treat the the amino acids at positions 93, 94, 95, 99, 104, 105, and 106 as seven categorical covariates, taking one of 21 values (representing missingness or one of the 20 amino acids). We use the default Dirichlet-multinomial similarity function and follow the recommendation for setting the Dirichlet hyperparameters less than one. (Specifically, we set them at 0.5.) For the sake of comparison, we use Monte Carlo simulation from the prior to find a value for the mass parameter $\alpha$ such that the prior number of subsets is the same as that obtained by using the Ewens or EPA distribution with $\alpha = 1$. The results are found in row 5 of Table 1 and show that, for most positions, the PPMx model performs substantially worse than the Ewens distribution. We caution, however, that many of the Markov chains exhibit poor mixing. We also find poor performance with smaller values for the mass parameter $\alpha$ or when substantially increasing the burn-in period (not shown). We suspect that this PPMx prior has several local modes that dominate the likelihood. To make a direct comparison, we also consider an alternative specification of the EPA prior partition distribution where the similarity function $\lambda(i, j)$ is one plus the number of times proteins $i$ and $j$ share the same value across these seven covariates. We find that this second specification of EPA (row 3) performs substantially better than the PPMx for these same covariates (row 5) and almost as well as the original EPA specification (row 2).

We suspect that other formulations of the PPMx model may perform better. Indeed, consider the PPMx model using the default Dirichlet-multinomial similarity function based only on the amino acid at the current position. Under this formulation the PPMx model (row 6 of Table 1) performs much better than the Ewens distribution (row 1) overall. By way of comparison, consider the EPA distribution where the similarity function $\lambda(i, j)$ is 2 if proteins $i$ and $j$ have the same amino acid at that position and is 1 otherwise. This EPA formulation (row 4) also performs much better than the Ewens distribution (row 1) and

dominates the analogous PPMx formulation (row 6) at each position. While the EPA distribution dominates the PPMx distribution in this case, we suspect it may perform better in other scenarios or with non-default choices for the PPMx similarity function.

The EPA distribution allows pairwise similarity information to inform the partitioning. An ad hoc method capturing this idea uses the standard Ewens distribution but subset the data to only include those observations whose similarities to the observation of interest exceed a threshold. The subsetting threshold is analogous to the temperature $\tau$ in the EPA distribution, but there the temperature $\tau$ can be treated as random with a prior distribution whereas the threshold must be fixed to implement the subsetting approach. Further, discarding observations will likely lead to a loss of precision in estimating other parameters. We examine several thresholds for the BLAST bit scores and the results for the best thresholds are found on rows 7–9 in Table 1. For threshold $t = 15$ and $t = 25$, subsetting is usually better than not subsetting (row 1), but the PPMx model (row 6) and EPA distribution (rows 2–4) perform better.

Finally, we consider posterior inference on the hyperparameters. Let the discount 5 have a mixture prior distribution with a point-mass at 0 with probability 0.5 and a Beta(1, 3) distribution otherwise. We again run 20 independent Markov chains for each position (but do not leave out an observation). Whereas the prior probability that $\delta = 0$ is 0.5, the posterior probabilities at positions 93, 94, 95, 104, 105, and 106 are 0.47, 0.36, 0.38, 0.53, 0.51, and 0.58, respectively (all of which are statistically different from 0.5). The posterior expectations of temperature $x$ *are* 3.3, 10.4, 7.8, 4.1, 3.4, and 6.9, respectively (all of which are statistically different from the prior expectation of 4). To assess the posterior learning on the permutation $\sigma$, consider the indices of observations in $\sigma$. The uniform prior of $\sigma$ makes the prior expectation of an index be $94/2 = 47$ for all 94 observations. The five-number summary of the posterior means of the indices at position 94 is *(*11.6, 48.0, 48.9, 50.0, 57.7*)* and this pattern is consistent across independent Markov chains. At position 106, the five-number summary is (31.5, 46.9, 49.4, 51.2, 53.4). We conclude that, in some cases, there is substantial learning on these parameters whereas, in other cases, there is little difference between the prior and posterior distributions.

### 6.3. Bayesian Linear Regression with Latent Clusters

Section 6.2 demonstrates our proposed distribution in an application with 94 observations and five parameters per subset. To see how our proposal performs as the dimension and sample size grow, we now consider a simulation study with $n = 1050$ observations and 31 parameters per subset. Consider a linear regression model in which a response $y_i$ has a normal distribution with mean $\mathbf{x}_i\boldsymbol{\beta}_i$ and precision $\lambda_i$, for covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, and $p = 30$. In this simulation study, the data are generated from one of three sets of regression coefficient vectors and precisions. The inferential goal is to estimate the latent partition $\pi$ and the regression coefficient vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n$. To aid in estimation, partition covariates are available as prior information to help separate the data into subsets.

The specifics of the data generation are as follows: Set $x_{i1} = 1$ and sample all the other $x$'s from the uniform distribution of the unit interval. For $i \in S_1 = \{1, \ldots, 350\}$, set $(\boldsymbol{\beta}_i, \lambda_i)$ to be $\phi_1 = ((0, \ldots, 0), 1.0)$. For $i \in S_2 = \{351, \ldots, 700\}$, set $(\boldsymbol{\beta}_i, \lambda_i)$ to be a tuple $\phi_2$ containing: (i) a

column vector whose first 10 elements are 0.9 and the other elements are 0 and (ii) 1.0. For $i \in S_3 = \{701, \ldots, 1050\}$, set $(\beta_i, \lambda)$ to be a tuple $\phi_3$ containing: (i) a column vector $\phi_3$ whose first 6 elements are 1.0 and the other elements are 0, and (ii) 1.0. Partition covariates $\mathbf{w}_1, \ldots, \mathbf{w}_n$ *are* sampled from one of three four-dimensional multivariate normal distributions, depending on the subset to which $i$ belongs. The parameters are taken from the empirical moments of the three classes in the iris data (Fisher 1936), with the first, second, and third subsets corresponding to "setosa," "versicolor," and "virginica," respectively. Note that subsets 2 and 3 have similar coeicients and their observations have partition covariates drawn from somewhat overlapping distributions.

In the notation of (2), we wish to estimate the parameters $\pi_n = \{S_1, S_2, S_3\}$ and $\boldsymbol{\phi} = ((\phi_1, \phi_2, \phi_3)$. The prior distribution for the $\phi$*'s* is the conjugate multivariate normal-gamma distribution $\text{Ng}(\boldsymbol{\beta}, \lambda \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha, \beta)$ where $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Lambda}$ is the identity matrix, $\alpha = 1$, and $\beta = 1$. Three prior distributions for the partition $\pi_n$ are considered and the resulting performance is compared. First, we consider the Ewens distribution which ignores the partition covariates $\mathbf{w}$*'s.* Second, we consider the PPMx model (Müller, Quintana, and Rosner 2011) using their default formulation of the similarity function based on the centered and scaled versions of the $\mathbf{w}$*'s.* Finally, we consider our EPA distribution using an exponential similarity function applied to the Euclidean distance between the $\mathbf{w}$*'s.* We place a uniform prior on the permutation $\boldsymbol{\sigma}$ and the temperature $x$ has a Gamma(2, 0.5*)* prior (with mean 4). The discount 5 is fixed at 0.0. We set the mass parameter $\alpha$ to 1.0 for the Ewens and EPA distributions and to 23.0 for the PPMx model, making the prior expected number of subsets to be approximately 7.5 for all three distributions.

Thirty independent Markov chains are run for 2000 iterations for each of the three models using our software written in Scala. The frst 500 iterations are discarded as burn-in. The performance of the models is assessed using Monte Carlo estimates of the posterior mean of the adjusted Rand index (ARI) (Hubert and Arabie 1985) with respect to the true partition. The ARI is a measure of similarity between two partitions, with 1.0 corresponding to perfect agreement. The mean ARI is 0.227, 0.505, 0.648 for the models using the Ewens, PPMx, and EPA distributions, respectively. We also compute the posterior mean of the sum of squared Euclidean distances from the true coefcient vectors and find values 8.68, 4.24, and 3.91 for the models using the Ewens, PPMx, and EPA distributions, respectively. All pairwise differences are statistically significant ($p$-value less than 0.01) based on a two-sample $t$-test. Using either evaluation criteria, the model with the EPA distribution performs the best in this simulation study and demonstrates the viability of the EPA distributions in high dimensions. The PPMx model, which also performs well, has the advantage that it runs in about 59% of the CPU time required for the EPA distribution.

## 7. Conclusion

Our proposed EPA distribution uses pairwise similarity information to define a random partition distribution. A key feature of our formulation is that probability is allocated among partitions within a *given* number of subsets, but probability is *not* shifted among sets of partitions with *different* numbers of subsets. This feature provides a clear separation of the roles of: (i) the mass parameter $\alpha$ and discount parameter 5 and (ii) the pairwise similarity

function λ and permutation $\boldsymbol{\sigma}$. Further, the distribution of the number of subsets is unchanged from the usual Ewens and Ewens-Pitman distributions, and the intuition one has regarding the $\alpha$ and δ from these familiar distributions carries over. We note that our distribution is invariant to scale changes in the similarity λ, which aligns with the idea that similarity is a relative rather than an absolute concept. Our formulation also has an explicit p.m.f. with an easily-evaluated normalizing constant, so standard MCMC samplers are available for posterior inference on the partition and hyperparameters influencing the partition distribution.

It could be argued that our proposal excessively shrinks toward the Ewens and Ewens-Pitman distributions and that the distribution of the number of subsets should be influenced by the similarity information. In a preliminary formulation, we initially considered defining $\Pr(\sigma_t \in S \mid \alpha, \delta, \lambda, \pi(\sigma_1, \ldots, \sigma_{t-1}))$ in (4) to be proportional to $\sum_{\sigma_s \in S} \lambda(\sigma_t, \sigma_s) - \delta$ for an existing subset $S$ and proportional to $\alpha + \delta q_{t-1}$ for a new subset. This makes the probability of forming a new subset depend on the similarity function and, therefore, the distribution of the number of subsets different from that of the Ewens, Ewens-Pitman, and EPA distributions. We chose to not pursue this formulation for a few reasons. First, the normalizing constant of the p.m.f. would then become intractable, making posterior inference difficult for the partition and hyperparameters. Second, we feel that the clear separation of the roles of the $\alpha$, $\delta$, and λ can be desirable and a feature that distinguishes our distribution from the PPMx and ddCRP distributions. We view our contribution as expanding the choices available for flexible Bayesian modeling. Finally, we showed in the demonstrations of Sections 6.2 and 6.3 that using the EPA distribution as a prior partition distribution can provide better statistical performance.

## Acknowledgments

The authors gratefully acknowledge Peter Müller, Fernando A. Quintana, David H. Russell, Lei Tao, Gordon B. Dahl and anonymous referees for helpful suggestions.

**Funding**

This work is supported by NIH NIGMS R01 GM104972.

## References

Airoldi EM, Costa T, Bassetti F, Leisen F, Guindani M. Generalized Species Sampling Priors With Latent Beta Reinforcements. Journal of the American Statistical Association. 2014; 109:1466–1480. [PubMed: 25870462]

Aldous, D., Ibragimov, I., Jacod, J., Aldous, D. cole d't de Probabilits de Saint-Flour XIII 1983, vol. 1117 of Lecture Notes in Mathematics. Berlin / Heidelberg: Springer; 1985. Exchangeability and Related Topics; p. 1-198.

Antoniak CE. Mixtures of Dirichlet Processes With Applications to Bayesian Nonparametric Problems. The Annals of Statistics. 1974; 2:1152–1174.

Arratia, R., Barbour, AD., Tavaré, S. Logarithmic Combinatorial Structures: A Probalistic Approach. Zurich: European Mathematical Society; 2003.

Bell ET. Exponential Numbers. American Mathematical Monthly. 1934; 41:411–419.

Blei DM, Frazier PI. Distance Dependent Chinese Restaurant Processes. Journal of Machine Learning Research. 2011; 12:2461–2488.

Dahl, DB. An Improved Merge-Split Sampler for Conjugate Dirichlet Process Mixture Models. University of Wisconsin - Madison, Department of Statistics; 2003. Technical Report 1086

Dahl, DB. JSM Proceedings, Section on Bayesian Statistical Science. Alexandria, VA: American Statistical Association; 2008. Distance-Based Probability Distribution for Set Partitions With Applications to Bayesian Nonparametrics.

De Blasi P, Favaro S, Lijoi A, Mena R, Prunster I, Ruggiero M. Are Gibbs-Type Priors the Most Natural Generalization of the Dirichlet Process? Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2015; 37:212–229.

Edgar RC. MUSCLE: Multiple Sequence Alignment With High Accuracy and High Throughput. Nucleic Acids Research. 2004; 32:1792–1797. [PubMed: 15034147]

Escobar MD, West M. Bayesian Density Estimation and Inference Using Mixtures. Journal of the American Statistical Association. 1995; 90:577–588.

Ewens W. The Sampling Theory of Selectively Neutral Alleles. Theoretical Population Biology. 1972; 3:87–112. [PubMed: 4667078]

Ferguson TS. A Bayesian Analysis of Some Nonparametric Problems. The Annals of Statistics. 1973; 1:209–230.

Fisher RA. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics. 1936; 7:179–188.

Geisser S, Eddy WF. A Predictive Approach to Model Selection. Journal of the American Statistical Association. 1979; 74:153–160.

Gelfand AE. Empirical Bayes Methods for Combining Likelihoods: Comment. Journal of the American Statistical Association. 1996; 91:551–552.

Gelfand AE, Smith AFM. Sampling-Based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association. 1990; 85:398–409.

Hubert L, Arabie P. Comparing Partitions. Journal of Classification. 1985; 2:193–218.

Jain S, Neal RM. A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model. Journal of Computational and Graphical Statistics. 2004; 13:158–182.

Jain S, Neal RM. Splitting and Merging Components of a Nonconjugate Dirichlet Process Mixture Model. Bayesian Analysis. 2007; 2:445–472.

Lau JW, Green PJ. Bayesian Model Based Clustering Procedures. Journal of Computational and Graphical Statistics. 2007; 16:526–558.

Lee J, Quintana FA, Müller P, Trippa L. Defining Predictive Probability Functions for Species Sampling Models. Statistical Science. 2013; 28:209–222. [PubMed: 24368874]

Lennox KP, Dahl DB, Vannucci M, Tsai JW. Density Estimation for Protein Conformation Angles Using a Bivariate von Mises Distribution and Bayesian Nonparametrics. Journal of the American Statistical Association. 2009; 104:586–596. [PubMed: 20221312]

Müller P, Quintana F, Rosner GL. A Product Partition Model With Regression on Covariates. Journal of Computational and Graphical Statistics. 2011; 20:260–278. [PubMed: 21566678]

Neal RM. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics. 2000; 9:249–265.

Park JH, Dunson DB. Bayesian Generalized Product Partition Model. Statistica Sinica. 2010; 20:1203–1226.

Pitman J. Exchangeable and Partially Exchangeable Random Partitions. Probability Theory and Related Fields. 1995; 102:145–158.

Pitman, J. Some Developments of the Blackwell-MacQueen Urn Scheme. In: Ferguson, TS.Shapley, LS., MacQueen, JB., editors. Statistics, Probability and Game Theory. Vol. 30. Beachwood, OH: Institute of Mathematical Statistics; 1996. p. 245-267.IMS Lecture Notes Monograph Series

Pitman J, Yor M. The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator. The Annals of Probability. 1997; 25:855–900.

Quintana FA. A Predictive View of Bayesian Clustering. Journal of Statistical Planning and Inference. 2006; 136:2407–2429.

Ritter C, Tanner MA. Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler. Journal of the American Statistical Association. 1992; 87:861–868.

Shahbaba B, Neal RM. Nonlinear Models Using Dirichlet Process Mixtures. Journal of Machine Learning Research. 2009; 10:1829–1850.

Singh H, Hnizdo V, Demchuk E. Probabilistic Model for Two Dependent Circular Variables. Biometrika. 2002; 89:719–723.

Teh, YW. A Hierarchical Bayesian Language Model Based on Pitman-Yor Processes; Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics; Sydney, Australia. Stroudsburg, PA: Association for Computational Linguistics; 2006. p. 985-992.

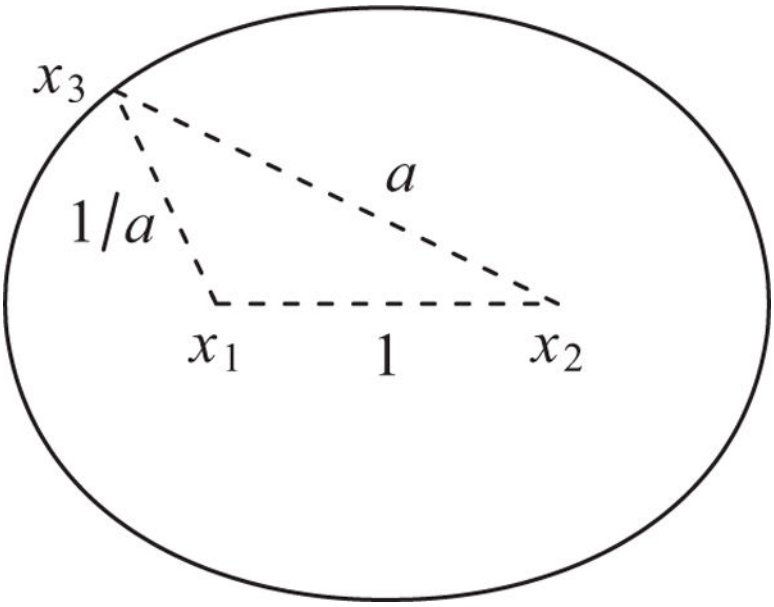


**Figure 1.**
While ideally item 3 could be placed anywhere, insisting on marginal invariance for the EPA distribution requires that item 3 be constrained to fall on this Cassini oval.

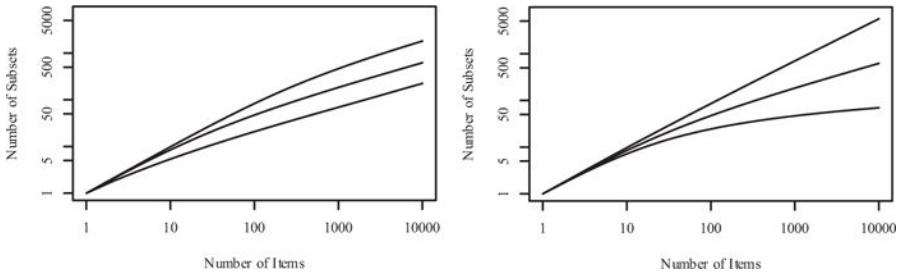


**Figure 2.**

Left: Mean number of subsets $E(q_n / a, \delta)$ as a function of the number of items $n$ displayed on the log-log scale, with discount parameter $\delta = 0.5$ and mass parameter $a = 1$ (bottom), $a = 10$ (middle), and $a = 100$ (top). Right: Same as left, except mass parameter $a = 10$ and discount parameter $\delta = 0$ (bottom), $\delta = 0.5$ (middle), and $\delta = 0.9$ (top).

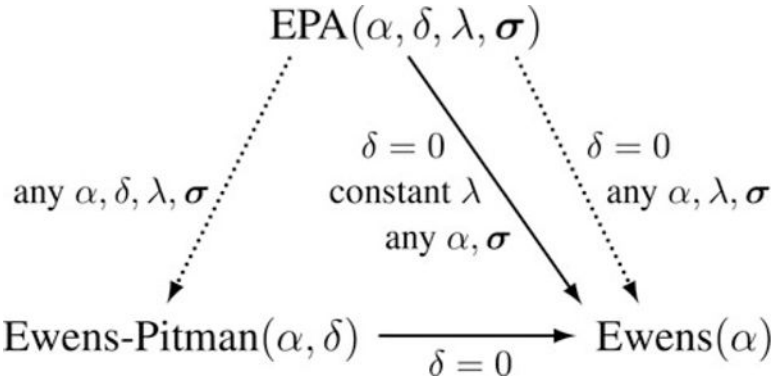


**Figure 3.**

Relationships between the EPA, Ewens-Pitman, and Ewens distributions. Solid lines indicate that, under the indicated constraints, the more general distribution reduces to a simpler distribution. Dotted lines indicate that, under the indicated constraints, the more general distribution has the same distribution on the number of subsets $q_n$.

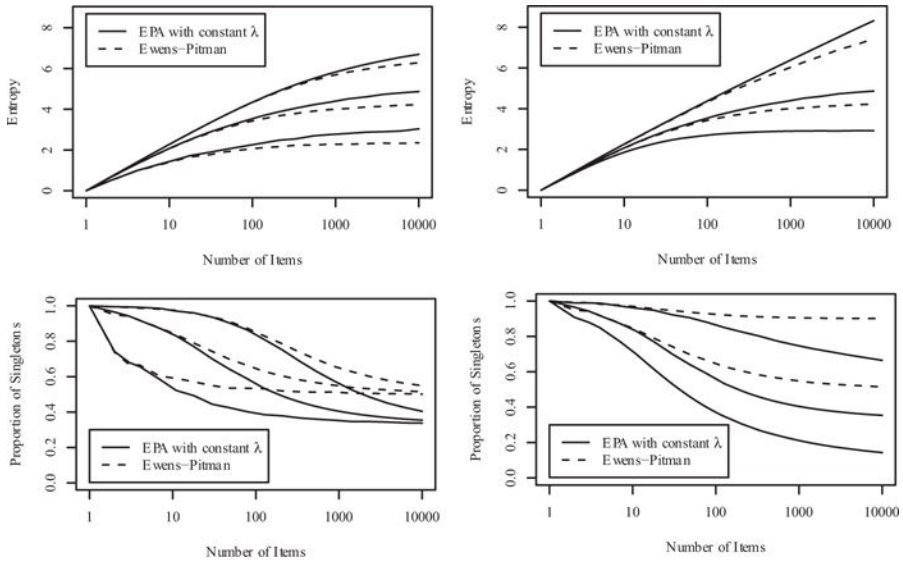


**Figure 4.**

Top left panel: Mean entropy as a function of the number of items $n$ (on the log scale), with discount parameter $\delta = 0.5$ and mass parameter $\alpha = 1$ (bottom), $\alpha = 10$ (middle), and $\alpha = 100$ (top). Top right panel: Same, with mass parameter $\alpha = 10$ and discount parameter $\delta = 0$ (bottom), $\delta = 0.5$ (middle), and $\delta = 0.9$ (top). Bottom panels show the mean proportion of subsets having only one item as a function of the number of items, using the same combinations of $\alpha$ and $\delta$ values. When $\delta = 0$, the EPA distribution with constant similarity function). and the Ewens-Pitman distribution are the same.

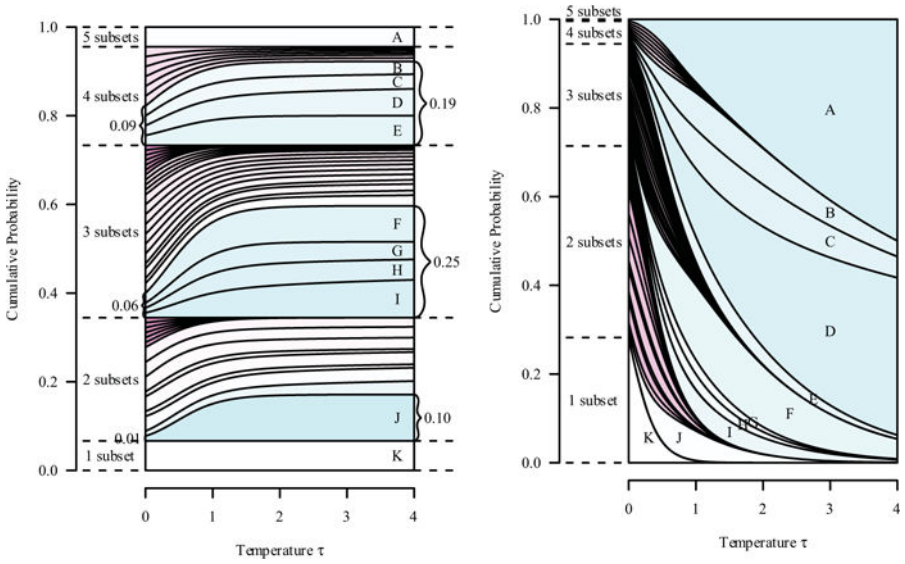


**Figure 5.**

The cumulative probabilities of the 52 partitions for the five states selected from the "USArrests"dataset for the EPA distribution (left) and the ddCRP (right). For each partition, the cumulative probabilities across temperatures are joined to form the curves and the probability of a given partition is the difference between curves. Capital letters label the same partitions for both the left- and right-hand sides.

**Table 1**

Differences in the log pseudo marginal likelihood (LPML) between several models and the model using the standard Ewens distribution.

| | Position | | | | | | |
|---|---|---|---|---|---|---|---|
| | **93** | **94** | **95** | **104** | **105** | **106** | **Total** |
| 1. Ewens | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2. EPA using BLAST similarity | 28.0 | 40.3 | 51.3 | − 0.8 | − 0.1 | 57.9 | 176.6 |
| 3. EPA using 7-covariates similarity | 27.8 | 31.8 | 45.1 | − 0.3 | − 0.6 | 57.5 | 1612 |
| 4. EPA using 1-covariate similarity | 27.9 | 17.0 | 26.7 | 0.6 | −1.3 | 22.3 | 93.3 |
| 5. PPMx using 7-covariates similarity | − 54.6 | − 30.5 | 7.1 | − 45.9 | −18.63 | − 8.6 | − 1335 |
| 6. PPMx using 1-covariate similarity | 24.3 | 10.7 | 15.6 | − 4.2 | − 5.9 | 17.3 | 57.7 |
| 7. Ewens w/ BLAST subsetting, $t = 15$ | 2.7 | 5.1 | 15.4 | − 0.4 | 1.8 | 6.2 | 30.8 |
| 8. Ewens w/ BLAST subsetting, $t = 25$ | 3.8 | 5.9 | 29.5 | − 4.7 | − 05 | 17.7 | 515 |
| 9. Ewens w/ BLAST subsetting, $t = 35$ | −19.6 | − 1.8 | 31.7 | − 26.0 | − 205 | 6.9 | − 292 |

NOTE: Large positive values indicate better fit.