Biometrika Advance Access published December 2, 2014

Biometrika (2014), *pp.* 1–8 doi: 10.1093/biomet/asu052
Printed in Great Britain

Generalized Ewens–Pitman model for Bayesian clustering

BY HARRY CRANE

Department of Statistics & Biostatistics, Rutgers University, 110 Frelinghuysen Road,
Piscataway, New Jersey 08854, U.S.A.

hcrane@stat.rutgers.edu

SUMMARY

We propose a Bayesian method for clustering from discrete data structures that commonly arise in genetics and other applications. This method is equivariant with respect to relabelling units; unsampled units do not interfere with sampled data; and missing data do not hinder inference. Cluster inference using the posterior mode performs well on simulated and real datasets, and the posterior predictive distribution enables supervised learning based on a partial clustering of the sample.

Some key words: Clustering; Discrete parameter; Ewens–Pitman distribution; Partition data; Random partition.

1. INTRODUCTION

We consider clustering from data that result from observing a sequence of categorical measurements for each sampled unit. A prime example is DNA sequence data, in which each individual is associated to a sequence of DNA nucleotides. Similar data structures arise in species sampling (Bissiri et al., 2013; Good & Toulmin, 1956), ecology (Fisher et al., 1943), phylogenetics (Felsenstein, 2004), political science (Spirling & Quinn, 2010), and legal studies (Sirovich, 2003; Thurstone & Degan, 1951), where the suite of tools includes distance-based methods (Sokal & Sneath, 1963), nonparametric Bayesian methods (Spirling & Quinn, 2010), singular value decomposition (Sirovich, 2003), and factor analysis (Thurstone & Degan, 1951). Our approach incorporates the desired clustering as a parameter in a hierarchical Bayesian model, which affords the data sequence an easy interpretation as repeated noisy observations of the true clustering.

Specifically, we propose a hierarchical Bayesian method that is equivariant with respect to relabelling and whose posterior distribution exhibits lack of interference, meaning the cluster membership of unsampled individuals does not affect the inferred clustering for the sample. Both properties are essential to the logical structure of statistical models (McCullagh, 2002), but many methods fail to satisfy one or more of these, e.g., product partition models (Barry & Hartigan, 1992; Crowley, 1997; Hartigan, 1990; Park & Dunson, 2010; Quintana & Iglesias, 2003) do not generally exhibit lack of interference, because they are not consistent under subsampling.

Clustering problems abound in scientific applications, including epidemiology (Yang et al., 2012), linguistics (Efron & Thisted, 1976, 1987), and machine learning (Blei et al., 2003). Many established methods involve Bayesian techniques, e.g., finite mixture models (Banfield & Raftery, 1993; Richardson & Green, 1997), Bayesian product partition models (Crowley, 1997; Hartigan, 1990; Park & Dunson, 2010; Quintana & Iglesias, 2003), model-based clustering (Raftery & Fraley, 2002), and other approaches (Binder, 1978; Booth et al., 2008; McCullagh & Yang, 2008). Outside the realm of statistical methods, distance-based algorithms (Lloyd, 1982) dominate. We stress that all of these methods are tailored to data lying in a continuous space, and so are not amenable to our setting.

2. Generalized Ewens–Pitman cluster model

2·1. *Partition data*

In our applications, the response for each unit u_i is a sequence $y_i = y_i^1 y_i^2 \cdots$ of values in a finite set, which we take to be $[k] = \{1, \ldots, k\}$ for some $k = 1, 2, \ldots$. For the population $\mathcal{N} = \{1, 2, \ldots\}$, the data comprise an infinite array

$$Y = \begin{array}{cc} & \begin{array}{cccc} 1 & 2 & 3 & \cdots \end{array} \\ \begin{array}{c} u_1 \\ u_2 \\ \vdots \\ u_n \\ \vdots \end{array} & \begin{pmatrix} y_1^1 & y_1^2 & y_1^3 & \cdots \\ y_2^1 & y_2^2 & y_2^3 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ y_n^1 & y_n^2 & y_n^3 & \cdots \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix} \end{array},$$

of which we observe the first n rows and M columns, with $n = 1, 2, \ldots$ denoting the sample size and $M = 1, 2, \ldots$ denoting the sequence length. Many applications suit this framework: in genetics, units are individuals and y_i is a DNA sequence; in phylogenetics, units are species and y_i is a mitochondrial DNA sequence; in legal studies, units are judges and y_i is a sequence of court rulings; in political science, units are legislators and y_i is a voting record.

Each column of Y provides a basis for comparing units. For example, in DNA sequence data the columns represent sites on the chromosome, so that $y_i^j = y_{i'}^j$ indicates genetic similarity between units i and i'. In this way, each column determines a partition of $\mathcal{N}$ into k labelled classes; but, in most applications, the labels $[k]$ are superfluous and inference depends on Y only through its induced sequence of partitions.

A partition $\pi = B_1/\cdots/B_r$ of $A \subseteq \mathcal{N}$ is an unlabelled collection of nonempty, disjoint subsets $B_1, \ldots, B_r$ for which $\bigcup_{j=1}^r B_j = A$. For $A \subseteq \mathcal{N}$, $\mathcal{P}_A$ denotes the set of partitions of A. The data array Y induces a partition sequence $\Pi = (\Pi^1, \Pi^2, \ldots)$ through

$$i \text{ and } i' \text{ are in the same block of } \Pi^j \quad \text{if and only if} \quad y_i^j = y_{i'}^j \quad (j = 1, 2, \ldots). \tag{1}$$

For a finite sample $[n] \subset \mathcal{N}$, we observe the restriction of Y to its first n rows and M columns and (1) yields a length-M sequence of partitions of $[n]$, denoted $\Pi_{[n]}^{[M]} = (\Pi_{[n]}^1, \ldots, \Pi_{[n]}^M)$.

2·2. *Partition models*

Though we observe only the finite restriction $\Pi_{[n]}^{[M]}$ of the infinite sequence $\Pi = (\Pi^1, \Pi^2, \ldots)$, we model Π as a conditionally independent, identically distributed sequence of random partitions of $\mathcal{N}$, given a clustering parameter B. In a hierarchical set-up, B is a Ewens–Pitman(α^*, θ^*) partition, whose finite-dimensional distributions on $\mathcal{P}_{[n]}$, for each $n = 1, 2, \ldots$, are

$$P_n^{\alpha^*,\theta^*}(\pi) = \mathrm{pr}(B_{[n]} = \pi) = (\theta^*/\alpha^*)^{\uparrow\#\pi} \left\{ \prod_{b\in\pi} -(-\alpha^*)^{\uparrow\#b} \right\} \Big/ \theta^{*\uparrow n} \qquad (\pi \in \mathcal{P}_{[n]}), \tag{2}$$

where $\alpha^{*\uparrow j} = \alpha^*(\alpha^* + 1)\cdots(\alpha^* + j - 1)$, $\#\pi$ is the number of blocks of π, and (α^*, θ^*) satisfies either (I) $\alpha^* < 0$ and $\theta^* = -k^*\alpha^*$, for some $k^* = 1, 2, \ldots$, or (II) $0 \leqslant \alpha^* \leqslant 1$ and $\theta^* > -\alpha^*$.

Given B, we assume that $\Pi^1, \Pi^2, \ldots$ are conditionally independent and obey the finite-dimensional distributions of the (α, k, B)-cluster model (Crane, 2014),

$$P_n^{\alpha,k,B}(\pi) = \mathrm{pr}(\Pi_{[n]}^i = \pi) = k^{\downarrow\#\pi} \left(\prod_{b\in B} \prod_{b'\in\pi} \alpha^{\uparrow\#(b\cap b')} \right) \Big/ \prod_{b\in B} (k\alpha)^{\uparrow\#b} \qquad (\pi \in \mathcal{P}_{[n]}), \tag{3}$$

where $k^{\downarrow j} = k(k-1)\cdots(k-j+1)$ and k is the size of the response space from § 2·1. By convention, we put $\alpha^{\uparrow 0} = 1$.

Remark 1. Notionally, the finite-dimensional distributions (3) depend on the entire clustering B of the infinite population $\mathcal{N}$; however, only the restriction of B to a partition of $[n]$ affects the right-hand side of (3). This attribute relates to lack of interference, which we discuss in § 2·4.

Definition 1. *The* (α^*, θ^*)*-generalized Ewens–Pitman* (α, k, B)*-cluster model is defined by*

$$B \sim \textit{Ewens–Pitman}(\alpha^*, \theta^*),$$
$$\Pi^1, \Pi^2, \ldots \mid B \sim (\alpha, k, B)\text{-}\textit{cluster model}, \tag{4}$$

where (α^*, θ^*) *satisfies* (I) *or* (II)*,* $\alpha > 0$*, and* $k = 1, 2, \ldots$ *is the size of the response space. The sequence in* (4) *is conditionally independent and identically distributed given* B.

2·3. *Interplay of parameters*

There are five parameters in the generalized Ewens–Pitman cluster model: (α^*, θ^*) determines the prior distribution of B and, given B, (α, k, B) models the data sequence. The parameters (α^*, θ^*) and α are fixed nuisance parameters, k is the size of the response space, and B is the clustering parameter. In applications, k is fixed and known ahead of time, (α^*, θ^*) is specified by the user, and (α, B) can be estimated jointly by maximizing the posterior probability of B over a range of α values.

In Definition 1, B is a Ewens–Pitman(α^*, θ^*) partition, whose parameter space splits into regimes (I) and (II) above. Under (I), B is restricted to partitions with k^* or fewer blocks; under (II), B is unrestricted. The choice of regime depends on the application. For example, in a case-control analysis, units cluster into $k^* = 2$ classes, cases and controls; in many other applications, the number of clusters in B is best left unspecified and regime (II) is appropriate. Aside from choice of regime (I) or (II), sensitivity analysis in § 3 shows that inference from the model is robust to choice of (α^*, θ^*). Note that k^* in (I) is unrelated to k, the size of the response space.

Taking $\alpha^* \to 0$ and holding θ^* fixed in (2) gives the Ewens distribution,

$$P_n^{0,\theta^*}(\pi) = \text{pr}(\Pi_n = \pi) = \theta^{*^{\#\pi}} \left\{ \prod_{b \in \pi} (\#b - 1)! \right\} / \theta^{*^{\uparrow n}} \qquad (\pi \in \mathcal{P}_{[n]}),$$

which plays an important role in the theory of neutral allele sampling (Ewens, 1972), prime factorizations (Donnelly & Grimmett, 1993), and species sampling (Fisher et al., 1943). Note that $\theta^* = 1$ is a critical value at which the model switches from favouring partitions with fewer blocks, $\theta^* < 1$, to favouring partitions with more blocks, $\theta^* > 1$. Without strong prior information, we take $(\alpha^*, \theta^*) = (-1/k^*, 1)$ in regime (I) and $(\alpha^*, \theta^*) = (0, 1)$ in regime (II) as objective choices of parameters for B.

In (3), B is the true clustering from which observations deviate according to mutation rate $\alpha > 0$. Together, (α, k, B) models the data sequence $\Pi_{[n]}^{[M]}$ of random partitions whose distribution is peaked around B. If i and i' are in different clusters of B, then they behave independently in the (α, k, B)-cluster distribution, that is, pr(i and i' in the same block) $= 1/k$. On the other hand, if i and i' are in the same cluster of B, then pr(i and i' in the same block) $= (1 + \alpha)/(1 + k\alpha)$. When α is large, pr(i and i' in the same block) $\approx 1/k$, the probability that two independent random draws from $[k]$ coincide, and when α is small, pr(i and i' in the same block) ≈ 1. Informally, α acts as a variance parameter that controls deviations from B.

2·4. *Equivariance and lack of interference*

Since units are labelled arbitrarily, any permutation $\sigma : [n] \to [n]$ also determines a valid labelling of the sample and acts on $\pi \in \mathcal{P}_{[n]}$ by relabelling, $\pi \mapsto \pi^\sigma$, where i and i' are in the same block of π^σ if and only if $\sigma^{-1}(i)$ and $\sigma^{-1}(i')$ are in the same block of π. For $m = 1, \ldots, n$, the restriction of $\pi = B_1/\cdots/B_r$ to $\mathcal{P}_{[m]}$ is the partition $\pi_{[m]}$ obtained by subsampling $[m] \subseteq [n]$. For example, with $n = 7$, the relabelling of $\pi = \{1, 4, 6\}/\{2, 7\}/\{3, 5\}$ by $\sigma = (145)(23)(67)$ is $\pi^\sigma = \{1, 2\}/\{3, 6\}/\{4, 5, 7\}$ and the restriction to $\{1, 2, 3, 4\}$ is $\pi_{[4]} = \{1, 4\}/\{2\}/\{3\}$.

DEFINITION 2. *For each $n = 1, 2, \ldots$ and $\theta \in \Theta$, let $P_n(\cdot\,; \theta)$ be a probability distribution on $\mathcal{P}_{[n]}$, where Θ is a generic parameter space. Assume further that every permutation $\sigma : [n] \to [n]$ determines an operation $\theta \mapsto \theta^\sigma \in \Theta$ and to every $n = 1, 2, \ldots$ there is a restriction mapping $\theta \mapsto \theta_n \in \Theta$. The model $P_\Theta = \{P_n(\cdot\,; \theta)\}_{n=1,2,\ldots;\, \theta\in\Theta}$ is label equivariant if, for all $n = 1, 2, \ldots$, $\pi \in \mathcal{P}_{[n]}$, and $\theta \in \Theta$,*

$$P_n(\pi^\sigma; \theta^\sigma) = P_n(\pi; \theta) \quad \text{for all permutations } \sigma : [n] \to [n],$$

and exhibits lack of interference if, for all $n = 1, 2, \ldots$, $\pi \in \mathcal{P}_{[n]}$, and $\theta \in \Theta$,

$$P_n(\pi; \theta) = P_n(\pi; \theta_n).$$

If $\theta \mapsto \theta^\sigma = \theta$ is the identity for every permutation σ, then P_Θ is called exchangeable; and if $\theta \mapsto \theta_n = \theta$ for every $n = 1, 2, \ldots$, then P_Θ is called consistent under subsampling.

For the (α, k, B)-cluster model, we write $\theta = (\alpha, k, B)$, whose image by σ is $\theta^\sigma = (\alpha, k, B^\sigma)$ and whose restriction is $\theta_n = (\alpha, k, B_{[n]})$. In words, label equivariance ensures that cluster inference is unaffected by arbitrary labelling of the sample and lack of interference implies that the observed sequence $\Pi_{[n]}^{[M]}$ is sufficient for inference about $B_{[n]}$.

We prove the following two theorems in the Supplementary Material.

THEOREM 1. *The posterior distribution of the (α^*, θ^*)-generalized Ewens–Pitman (α, k, B)-cluster model is label equivariant and exhibits lack of interference.*

THEOREM 2. *For $k, k^* = 1, 2, \ldots$ and $\alpha > 0$, the Ewens–Pitman$(-\alpha, k^*\alpha)$ prior and the $(\alpha/k, k, B)$-cluster distribution are conjugate; that is, the posterior distribution of B is the $(\alpha/k, k^*, \Pi^1)$-cluster distribution. Moreover, the marginal distribution of Π^1 is the Ewens–Pitman$(-k^*\alpha/k, k^*\alpha)$ distribution.*

Remark 2. Conjugacy in Theorem 2 is limited to the case of a single observation and the specific relationship between parameter values.

2·5. *Choice of prior distribution*

For finite parameter spaces, the uniform distribution is often a natural prior, unless the parameter space has additional structure (Berger et al., 2012). Since $[n]$ is sampled from an infinite population $\mathcal{N}$, $B_{[n]}$ is the restriction of the overall clustering B. Thus, the prior for $B_{[n]}$ should be the distribution induced on $\mathcal{P}_{[n]}$ by subsampling $[n] \subset \mathcal{N}$; that is, the family of finite-dimensional priors should be consistent under subsampling. Also, since labels are assigned to units arbitrarily, the prior distribution should be exchangeable. The family of uniform distributions on finite partitions is exchangeable but not consistent under subsampling. On the other hand, the Ewens–Pitman distribution is exchangeable, consistent under subsampling, and appears in many applications of random partitions.

2·6. *Posterior prediction and supervised learning*

In some applications, an initial sample $[n]$ is observed along with its correct clustering $B_{[n]}$. Given $B_{[n]}$ and data for an augmented sample with $m = 1, 2, \ldots$ new individuals, we wish to infer the clustering $B_{[n+m]}$ of the whole sample $[n+m]$, which should agree with $B_{[n]}$ upon restriction. For example, with $n = 4$ and $m = 1$, we observe a data sequence for the sample $\{1, 2, 3, 4, 5\}$ and a partial clustering $B_{[4]} = \{1, 2\}/\{3, 4\}$ of the subsample $\{1, 2, 3, 4\}$. Based on data for $\{1, 2, 3, 4, 5\}$, the supervised learning task determines to which cluster element 5 belongs, that is, whether $B_{[5]}$ is $\{1, 2, 5\}/\{3, 4\}$, $\{1, 2\}/\{3, 4, 5\}$, or $\{1, 2\}/\{3, 4\}/\{5\}$.

Posterior lack of interference permits clustering of newly observed units using the posterior predictive distributions. For $\Pi_{[n+m]}^1, \ldots, \Pi_{[n+m]}^M$ from the generalized Ewens–Pitman cluster model (4), the joint clustering $B_{[n+m]}$, given $B_{[n]}$, has conditional probability

$$\begin{aligned}&\mathrm{pr}(B_{[n+m]} = \pi^* \mid B_{[n]} = \pi,\ \Pi_{[n+m]}^{[M]} = \pi') \\ &\quad = \mathrm{pr}(\Pi_{[n+m]}^{[M]} = \pi' \mid B_{[n+m]} = \pi^*)\, \mathrm{pr}(B_{[n+m]} = \pi^*) / \mathrm{pr}(B_{[n]} = \pi,\ \Pi_{[n+m]}^{[M]} = \pi'). \qquad (5)\end{aligned}$$

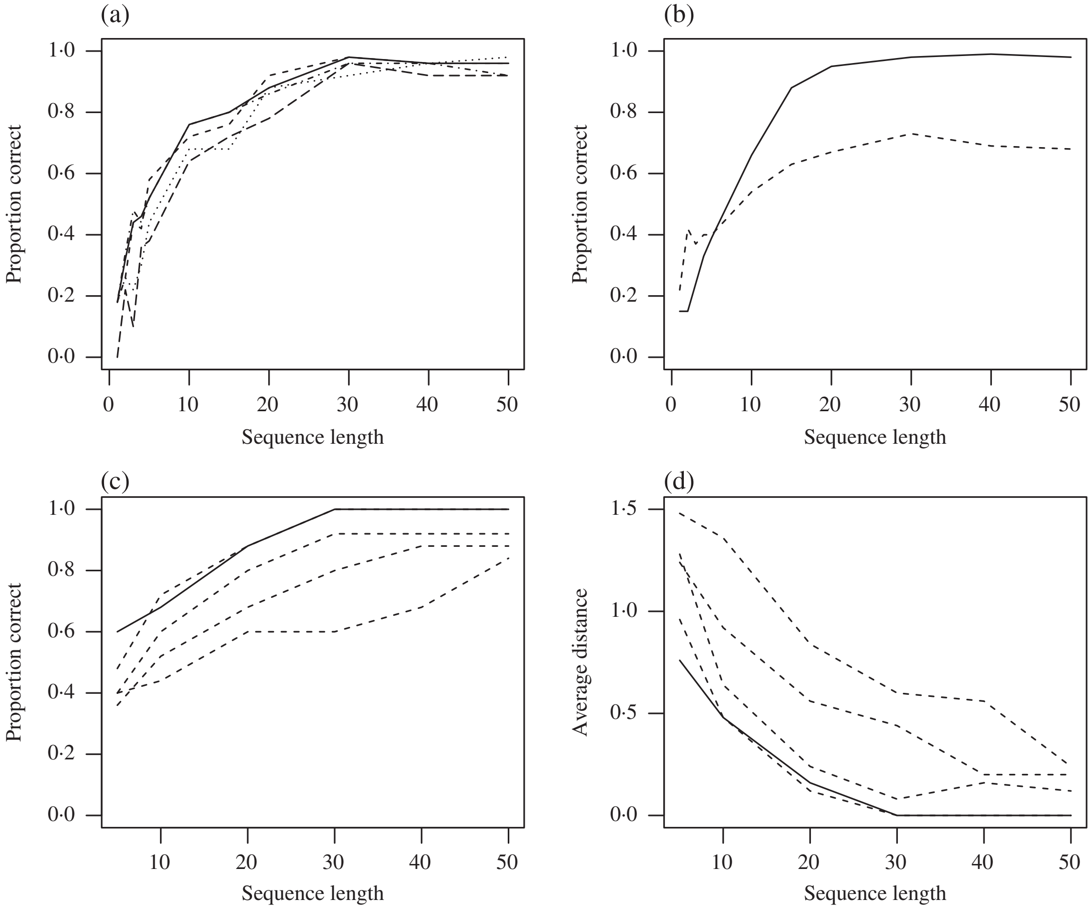

Fig. 1. Simulation results for generalized Ewens–Pitman model. (a): Proportion of trials in which the posterior mode coincided with the true clustering under misspecifications $\alpha^* = 0$, $\theta^* = 1/10, 1/2, 1, 2, 5$ for data simulated from (0, 1)-generalized Ewens–Pitman $(1, 4, B)$-cluster model with $n = 5$ and $k = 4$. (b): Comparison of the worst-case performance of our method (solid) to the best-case performance of the interchange algorithm (dashed). (c): Proportion of trials the posterior mode coincided with the true clustering under specifications of $\alpha = 1/5, 1/2, 1, 2, 5$, where data is simulated from the (0, 1)-generalized Ewens–Pitman $(1, 4, B)$-cluster model. (d): Average interchange distance between modal clustering and true clustering under different specifications of α. In both (c) and (d), the solid line corresponds to $\alpha = 1$ and the others correspond to $\alpha = 1/5, 1/2, 2, 5$.

3. Simulation study

Figure 1(a) shows the proportion of times in 100 replications that the posterior mode coincided with the true clustering under different specifications of nuisance parameters (α^*, θ^*). The panels in Fig. 1 illustrate that our model is robust to misspecification of the prior distribution and the posterior mode finds the correct clustering with high probability even for relatively short data sequences. Though not shown, we also observe robustness when the true prior distribution of B lies outside the model, e.g., if $B_{[n]}$ is drawn uniformly from $\mathcal{P}_{[n]}$.

For the clustering task, α is a nuisance parameter that can be specified or not, depending on the objective. Figures 1(c) and 1(d) show that estimation of B based on a fixed value of α is robust to misspecification. If α is left unspecified, then we infer (α, B) jointly by maximizing the posterior over both α and B. Such estimation requires more computational resources, but not prohibitively so. For instance, joint inference of (α, B) when $n = 8$, $(\alpha^*, \theta^*) = (-1, 2)$, $\alpha = 1$, and $k = 4$ yields the estimates in Table 1. This demonstrates empirically that $(\hat{\alpha}, \hat{B})$ is jointly consistent as sequence length increases. While estimation of B with α fixed is reasonably robust, Table 1 suggests leaving α unspecified, unless it is known.

Table 1. *Joint estimation of* (α, B) *for different sequence lengths*

Sequence length	Mean estimate of α	Standard error of α estimates	Percentage of trials B correctly estimated	Mean distance between estimate of B and true value
5	0·86	0·19	15	1·85
15	0·97	0·20	58	0·88
30	1·03	0·19	75	0·35
50	1·03	0·17	90	0·23

For clustering from discrete data, there are fewer statistically sound model-based methods than for continuous data. For instance, the product partition model is widely used, but it is not generally equipped to handle missing data or perform out-of-sample inference. The ability to handle missing data allied with computational efficiency makes distance-based methods more competitive for the data we consider. Figure 1(b) demonstrates the performance of distance-based clustering using the interchange metric, which specializes the widely used nearest neighbour interchange distance from phylogenetics (Fitch & Margoliash, 1967; Waterman & Smith, 1978). The interchange distance between partitions π, π', denoted $d_I(\pi, \pi')$, is the minimum number of elements whose block membership must be changed to transform π into π', likewise π' into π. For example, $d_I(\{1, 2\}/\{3, 4, 5, 6\}, \{1, 3, 4, 5\}/\{2, 6\}) = 1$ since we need interchange only elements 1 and 6 to move between π and π'. Likewise, $d_I(\{1, 2\}/\{3, 4, 5, 6\}, \{1, 2\}/\{3, 4, 5\}/\{6\}) = 1$ because we obtain $\{1, 2\}/\{3, 4, 5, 6\}$ by moving element 6 to the block $\{3, 4, 5\}$ in $\{1, 2\}/\{3, 4, 5\}/\{6\}$.

For clustering from the interchange distance, we assign to every $B \in \mathcal{P}_{[n]}$ a score $s(B) = \sum_{j=1}^{M} d_I(B, \pi_{[n]}^{j})$, where $\pi_{[n]} = (\pi_{[n]}^{1}, \ldots, \pi_{[n]}^{M})$ are the data. We then choose the partition, or partitions, with the lowest score. Figure 1 shows that the distance-based approach performs reasonably well in relation to our approach. When our model is incorrect for simulated data, the two methods perform comparably, but our method still holds an advantage in predictive power; see §§ 2·6 and 4·2. The interchange algorithm's apparent lack of consistency as sequence length increases reflects an averaging effect, whereby variation among different nearby clusterings is averaged to an ultimate clustering that is close, but not exactly equal, to the true value. McCullagh & Yang (2008) observe a similar phenomenon for clustering from the Gauss–Ewens process.

Remark 3. Our empirical study on simulated and real data suggests that short sequences are sufficient for reliable inference. For example, analysis of simulated data with $n = 8$ shows sufficient accuracy for sequences of length 10 or 15.

4. APPLICATION: MITOCHONDRIAL DNA DATA

4·1. *Clustering from the posterior mode*

We obtained mitochondrial DNA data for 15 species from the University of Montreal mitochondrial DNA database; see Table 1 in the Supplementary Material for a list of these 15 species. Based on substantially more information, scientific consensus classifies these species into three phyla, *Chordata*, *Platyhelminthes*, and *Arthropoda*, which we attempt to reconstruct using only mitochondrial DNA sequence data.

All sequences are about 20 000 nucleotides long, but lengths vary as a result of evolutionary factors and sequences must be aligned prior to analysis. We use ClustalOmega for sequence alignment, which finds the most parsimonious alignment by inserting missing sites into the mitochondrial DNA genome. We treat insertions as missing data.

From the generalized Ewens–Pitman cluster model with $(\alpha^*, \theta^*) = (-1/2, 1)$ and $k = 4$, the maximum posterior probability, with unnormalized log-posterior -8024, occurs at $\alpha = 0{\cdot}31$ and a clustering into two classes, one containing all 12 species from the phyla *Chordata* and *Arthropoda* and the other containing only species in the phylum *Platyhelminthes*. According to evolutionary theory, *Chordata* and *Arthropoda* are more closely related to each other than to *Platyhelminthes*, and so our findings are reasonable. The second most probable clustering, with unnormalized log-posterior -8159, is $\{$*Chordata*$\}$ and $\{$*Arthropoda*, *Platyhelminthes*$\}$, which is less biologically plausible.

The parameter choice $(\alpha^*, \theta^*) = (-1/2, 1)$ ensures that our estimate of B has at most two classes. Using the randomized search method from § 2 of the Supplementary Material, we searched for the posterior mode in the unrestricted model with $(\alpha^*, \theta^*) = (0, 1)$. In this case, the maximum posterior probability, with unnormalized log-probability -8020, occurs at $\alpha = 0{\cdot}26$ and a clustering of species into their distinct phyla *Chordata*, *Arthropoda*, and *Platyhelminthes*. The previous mode now has the second highest unnormalized log-posterior probability of -8042.

On the other hand, the minimum interchange algorithm does not detect the strong separation among these phyla. Instead, it ranks the clustering {*Chordata*, *Arthropoda*}/{*Platyhelminthes*} third.

4·2. *Supervised clustering application*

In § 2·6, we highlight the potential for supervised clustering from the posterior predictive distributions. We demonstrate this task by unclassifying certain species in the phylum *Chordata*. There are ten chordates in our mitochondrial DNA dataset, of which six are mammals. To illustrate our model's potential for supervised clustering, we unclassify the mammals wild boar and rhinoceros and assume the clustering of the eight other chordates into mammals and non-mammals is known. Using the posterior predictive distribution (5) with $(\alpha^*, \theta^*) = (0, 1)$, $\alpha = 1/4$, and $k = 4$, the maximum posterior probability, with unnormalized log-probability -816, corresponds to the correct clustering with these two species in the mammals class. This predictive probability should be compared to the second most probable clustering, which has an unnormalized log-posterior probability of -877. See Table 2 of the Supplementary Material.

Since the model is Bayesian, the posterior predictive probability assigns a quantitative value to class membership, which leaves discretion to the user when these probabilities are close. For example, when the four non-mammal species are unclassified, we obtain two competing optimal clusterings. The most probable, with unnormalized log-probability -814, clusters crocodile, fish, and iguana in one class, with snake alone as a singleton; the second most probable puts all four together in a cluster separate from the mammals and has unnormalized log-probability -815. Scientific consensus agrees with this second most probable clustering, but the closeness of these log-probabilities alerts the user that further study is appropriate.

5. Computational efficiency

The computational complexity of the unnormalized posterior probability of a single clustering scales linearly with both sample size and sequence length. Therefore, exponential growth of the space of partitions presents the main challenge in maximizing the posterior for large samples. The examples in § § 3 and 4 illustrate the model empirically on small sample sizes for which exact inference is tractable. For large samples, we use a Markov chain algorithm that converges to equilibrium in $O(\log n)$ steps and, thus, searches the partition space exponentially quickly (Crane & Lalley, 2013). In the Supplementary Material, we describe this method and apply it to senate roll call data.

Acknowledgement

The author's work is partially supported by the U.S. National Science Foundation and U.S. National Security Agency.

Supplementary material

Supplementary material available at *Biometrika* online discusses further aspects of the generalized Ewens–Pitman cluster model, which includes an explanation of the Ewens–Pitman distribution, a description of a randomized search algorithm, a real data example for which we use randomized search, and proofs of Theorems 1 and 2.

References

Banfield, J. & Raftery, A. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–21.
Barry, D. & Hartigan, J. A. (1992). Product partition models for change point problems. *Ann. Statist.* **20**, 260–79.

Berger, J. O., Bernardo, J. M. & Sun, D. (2012). Objective priors for discrete parameter spaces. *J. Am. Statist. Assoc.* **107**, 636–48.

Binder, D. (1978). Bayesian cluster analysis. *Biometrika* **65**, 31–8.

Bissiri, P., Ongaro, A. & Walker, S. (2013). Species sampling models: consistency for the number of species. *Biometrika* **100**, 771–7.

Blei, D., Ng, A. & Jordan, M. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022.

Booth, J. G., Casella, G. & Hobert, J. (2008). Clustering using objective functions and stochastic search. *J. R. Statist. Soc.* B **70**, 119–39.

Crane, H. (2014). Clustering from categorical data sequences. *J. Am. Statist. Assoc.*, in press. doi:10.1080/01621459.2014.983521.

Crane, H. & Lalley, S. P. (2013). Convergence rates of Markov chains on spaces of partitions. *Electron. J. Prob.* **18** (paper no. 61), 1–23.

Crowley, E. M. (1997). Product partition models for normal means. *J. Am. Statist. Assoc.* **92**, 192–8.

Donnelly, P. & Grimmett, G. (1993). On the asymptotic distribution of large prime factors. *J. Lond. Math. Soc.* **47**, 395–404.

Efron, B. & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63**, 435–47.

Efron, B. & Thisted, R. (1987). Did Shakespeare write a newly discovered poem? *Biometrika* **74**, 445–55.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.* **3**, 87–112.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates, Inc.

Fisher, R. A., Corbet, A. & Williams, C. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12**, 42–58.

Fitch, W. M. & Margoliash, E. (1967). Construction of phylogenetic trees. *Science* **155**, 279–84.

Good, I. J. & Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43**, 45–63.

Hartigan, J. A. (1990). Partition models. *Commun. Statist.* A **19**, 2745–56.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Trans. Info. Theory* **28**, 129–37.

McCullagh, P. (2002). What is a statistical model? (with discussion). *Ann. Statist.* **30**, 1225–310.

McCullagh, P. & Yang, J. (2008). How many clusters? *Bayesian Anal.* **3**, 101–20.

Park, J. H. & Dunson, D. B. (2010). Bayesian generalized product partition model. *Statist. Sinica* **20**, 1203–26.

Quintana, F. A. & Iglesias, P. L. (2003). Bayesian clustering and product partition models. *J. R. Statist. Soc.* B **65**, 557–74.

Raftery, A. & Fraley, (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Am. Statist. Assoc.* **97**, 611–31.

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. R. Statist. Soc.* B **59**, 731–92.

Sirovich, L. (2003). A pattern analysis of the second Rehnquist U.S. Supreme Court. *Proc. Nat. Acad. Sci.* **100**, 7432–73.

Sokal, R. R. & Sneath, P. H. A. (1963). *Numerical Taxonomy*. San Francisco: W. H. Freeman.

Spirling, A. & Quinn, K. (2010). Identifying intraparty voting blocs in the U.K. House of Commons. *J. Am. Statist. Assoc.* **105**, 447–57.

Thurstone, L. L. & Degan, J. W. (1951). Factorial study of the Supreme Court. *Proc. Nat. Acad. Sci.* **37**, 628–35.

Waterman, M. S. & Smith, T. F. (1978). On the similarity of dendrograms. *J. Theor. Biol.* **73**, 789–800.

Yang, J., Miescke, K. & McCullagh, P. (2012). Classification based on a permanental process with cyclic approximation. *Biometrika* **99**, 775–86.

[*Received February* 2014. *Revised August* 2014]