Clustering From Categorical Data Sequences

Harry CRANE

The three-parameter cluster model is a combinatorial stochastic process that generates categorical response sequences by randomly perturbing a fixed clustering parameter. This clear relationship between the observed data and the underlying clustering is particularly attractive in cluster analysis, in which supervised learning is a common goal and missing data is a familiar issue. The model is well equipped for this task, as it can handle missing data, perform out-of-sample inference, and accommodate both independent and dependent data sequences. Moreover, its clustering parameter lies in the unrestricted space of partitions, so that the number of clusters need not be specified beforehand. We establish these and other theoretical properties and also demonstrate the model on datasets from epidemiology, genetics, political science, and legal studies.

KEY WORDS: Classification; Random partition; Supreme court; Voting data

1. INTRODUCTION

Cluster analysis is useful for detecting structural relationships among statistical units, whether taxonomic classification in phylogenetics, separation into parties in political science, or division of judges into ideological groups in legal studies. The utility of cluster analysis underlies many methods developed in the literature, including distance-based algorithms (Lloyd 1982), parametric Bayesian models (Binder 1978; McCullagh and Yang 2008), nonparametric Bayesian methods (Hjort 1990; Barry and Hartigan 1992; Lau and Green 2007), Gaussian mixture models (Banfield and Raftery 1993), point processes (Yang, Miescke, and McCullagh 2012), and product partition models (Hartigan 1990; Crowley 1997). While many applications entail categorical observations, for example, phylogenetics (Felsenstein 2004), political science (Thurstone and Degan 1951; Sirovich 2003), linguistics (Efron and Thisted 1976, 1987), and population genetics (Ewens 1972; Kingman 1978), the most viable clustering methods specialize to data lying in a continuous space. Furthermore, some approaches offer no interpretation in terms of a data-generating process, which limits the meaning of the inferred clustering and causes uneasiness when attempting out-of-sample inference based on training data.

We address these concerns with a combinatorial stochastic process model for sequences of categorical measurements in a finite set, as in Tables 1 and 2. Such data arrays occur most obviously in genetics, but they appear more widely in legal studies, political science, and standardized testing (Thurstone and Degan 1951; Lord 1980; Sirovich 2003). Under our model, the columns in these tables are repeated draws from a probability distribution with fixed clustering and variance parameters. Units within the same cluster are more likely to exhibit the same response, establishing a clear link between the true clustering and the observed data.

Harry Crane is Professor, Department of Statistics and Biostatistics, Rutgers University, 110 Frelinghuysen Road, Room 501, Hill Center, Piscataway, NJ 08854 (E-mail: *hcrane@stat.rutgers.edu*). Author is partially supported by NSF grant DMS-1308899 and NSA grant H98230-13-1-0299. The author is indebted to several people. Hon. Jud Crandal's many suggestions improved the exposition. Marcin Hitczenko provided references to Bork (1997) and Toobin (2008) and enriched the Supreme Court example through lively discussions. Steinback Polinski provided early insights into the applications in Sections 7.1 and 7.2. Adrian Di Antonio lent some biological expertise for Section 7.3.

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

We unveil theoretical and empirical properties of our model in due course. In Section 2, we describe our setting in further detail and give motivating examples; in Section 3, we discuss some preliminaries for partition modeling; in Section 4, we introduce the *three-parameter cluster model*; in Section 5, we highlight key properties of our model; in Section 6, we show simulation results for maximum likelihood estimation and supervised learning from our model; in Section 7, we apply the model to datasets from legal studies, political science, and phylogenetics; in Section 8, we make some concluding remarks; in Appendix A, we describe stochastic search algorithms for approximate inference when the sample size is large; and in Appendix B, we prove our main theorem.

2. EXPOSITION AND MOTIVATING EXAMPLES

We propose a parametric family of probability measures for data arrays with repeated measurements from a finite set of $k = 1, 2, \ldots$ labels. For example, in Table 1, $k = 2$ corresponds to *concurrence* and *dissent* in Supreme Court rulings; in Table 2, $k = 4$ corresponds to the four DNA nucleotides, *adenine* (A), *cytosine* (C), *guanine* (G), and *thymine* (T); in standardized testing, $k = 2$ might correspond to *correct* and *incorrect* for a collection of test items, or perhaps $k = 5$ corresponds to the number of multiple choice answers.

Our model incorporates a clustering of the sample as a fixed parameter, which relates to the data in a natural way: units in the same cluster are more likely to produce the same response. This relationship between data and clustering endows the inferred parameters with a straightforward interpretation for applications. Importantly, we need not specify the number of clusters beforehand, unlike some classical clustering methods, for example, k-means (Lloyd 1982) and Gaussian mixture models (Banfield and Raftery 1993). We discuss these properties further in Sections 4 and 5. We now describe some motivating examples and datasets, which we analyze in Sections 6 and 7.

2.1 Supreme Court Rulings

In the United States Supreme Court, nine justices rule on about 80 cases per year. Each case has a definitive outcome,

Journal of the American Statistical Association
June 2015, Vol. 110, No. 510, Theory and Methods
DOI: 10.1080/01621459.2014.983521

Table 1. Sample dataset for U.S. Supreme Court voting alignment during the 2012 term. Court cases are labeled 1, 2, ... and comprise the columns of the table. For each justice, C corresponds to *concurrence* with the majority opinion of the Court, D corresponds to *dissent* from the majority opinion of the Court, and - corresponds to *recusal* from the case

Justice	Case 1	2	3	4	5	6	7	8	9	···
Sotomayor	C	–	C	C	D	C	D	D	C	···
Ginsburg	C	D	C	D	D	C	D	D	C	···
Kagan	–	D	C	–	D	C	D	D	C	···
Breyer	C	D	C	C	D	C	D	D	C	···
Kennedy	C	C	C	C	C	C	C	C	C	···
Roberts	C	C	C	C	C	C	C	C	C	···
Alito	C	C	D	C	C	C	C	C	C	···
Scalia	D	C	C	C	C	D	C	C	C	···
Thomas	D	C	C	C	C	D	C	C	C	···

with each justice either *concurring with* or *dissenting from* the majority opinion. Though Supreme Court justices do not declare political or philosophical affiliations, conventional wisdom often clusters judges into two or three ideological factions, such as conservative, liberal, or swing. Justices with similar ideologies tend to agree on how cases should be decided, and so the columns in Table 1 should reflect their ideological clustering. In principle, each case stands on its own, and so we model case outcomes as repeated, independent measurements from a probability distribution on voting alignments. Missing observations in Table 1 occur as a result of recusal, whereby one or more justice refrains from ruling on a case due to conflict of interest, illness, or some other reason. We treat recusals as missing data.

2.2 Genetic Sequences and Higher Order Taxonomy

Table 2 shows mitochondrial DNA (mtDNA) sequence data for a sample of 15 species in the animal kingdom. In many species, mtDNA is inherited solely from the mother and, therefore, is more stable between generations than cellular DNA.

Table 2. Mitochondrial DNA sequences for 15 species from different taxonomic phyla. At each site, species can be partitioned into as many as four classes corresponding to each of the nucleotide bases A, C, G, and T. In some cases, we observe no nucleotide (-) as a result of an insertion during sequence alignment

Phylum	Species	Site 1	2	3	4	5	6	7	···
Chordata	*X. unicolor*	A	C	G	T	A	C	G	···
Chordata	*C. niloticus*	C	C	G	A	A	A	G	···
Chordata	*T. tinca*	A	G	G	G	G	A	G	···
Chordata	*I. iguana*	A	A	G	C	A	T	G	···
Chordata	*H. sapiens*	A	A	G	T	A	A	G	···
Chordata	*C. familiaris*	G	G	G	A	A	C	G	···
Chordata	*U. americanus*	G	T	G	A	A	T	G	···
Chordata	*B. taurus*	A	A	G	C	A	C	G	···
Chordata	*S. scrofa*	A	T	G	A	A	C	G	···
Chordata	*R. unicornus*	A	A	G	C	G	T	C	···
Platyhelminthes	*E. granulosus*	A	G	G	T	G	T	T	···
Platyhelminthes	*T. asiatica*	A	A	G	G	A	T	T	···
Arthropoda	*T. californicus*	-	T	G	G	A	T	G	···
Arthropoda	*D. pulex*	A	A	G	G	A	C	G	···
Arthropoda	*D. melanogaster*	A	A	G	A	A	T	C	···

As a result, mtDNA is more reliable for comparing different species and is widely used in phylogenetics and evolutionary research. On the basis of genetic and other biological information, taxonomists classify organisms into a hierarchy by kingdom, phylum, class, order, etc. Species that are distant evolutionary relatives are likely to have different genomes, and we expect the columns in Table 2 to reflect this classification. See Campbell, Reece, and Mitchell (1999) for an elementary biological overview.

Missing observations in the table correspond to insertions and deletions during evolution, which cause sequence lengths to differ among species. After alignment, each sequence consists of the usual DNA nucleotides (A, C, G, and T) along with insertion (-). We treat insertions as missing data.

2.3 Gene Expression dataset

Golub et al. (1999) used gene expression levels to classify two types of acute leukemia, acute lymphocytic leukemia (ALL) and acute myeloid leukemia (AML). Gene expression levels were recorded for 7129 genes from 72 tissue samples, 47 of type ALL and 25 of type AML. Based on their gene expression levels, we wish to separate tissue samples according to leukemia type.

In contrast to the data in Sections 2.1 and 2.2, gene expression levels lie in a continuous space. While nominally a model for discrete data structures, our model can treat sequences of continuous data by grouping observations into a finite number of classes, which enables comparisons to widely studied methods for continuous response data, for example, discriminant analysis, nearest neighbor, and support vector machines; see Sections 6.3 and 6.4.

2.4 Opening Remarks

2.4.1 Independence Assumption. Each of the above datasets shares the same structure as an array of repeated measurements. Within each measurement, the responses of different units are correlated with their cluster membership, but we assume independence between different measurements. Independence across columns need not hold, but the assumption simplifies methodological development without contradicting our empirical observations. In Section 8, we discuss extensions of our model to data structures with more general dependence.

2.4.2 Modeling Missing Observations. The data in Sections 2.1 and 2.2 are riddled with missing observations. In all our analyses, we assume missing observations occur completely at random, that is, independently of the data-generating process. Under this assumption, sampling consistency equips our model to handle missing data; see Section 5.1.2. Though reasonable in some cases, this assumption does not hold in general. For example, in standardized testing, missing responses are strongly correlated with items to which students do not know the answer. In this case, the missing data mechanism must be modeled, a task we do not attempt.

2.4.3 Approximate Inference. To estimate the clustering parameter by maximum likelihood, we must optimize over the discrete space of partitions, which grows exponentially with sample size. As for any clustering method, exact optimization is not feasible for sample sizes much larger than 15, and we adopt a randomized search method for this task; see Appendix A.

3. MODELING PRELIMINARIES

3.1 Basic Setup

The general setup embodied in Tables 1 and 2 is conducive to clustering. For concreteness, we consider the data array induced by sampling the first five rows and six columns of Table 2:

$$y_{[5]} = \begin{matrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{matrix} \begin{pmatrix} A & C & G & T & A & C \\ C & C & G & A & A & A \\ A & G & G & G & G & A \\ A & A & G & C & A & T \\ A & A & G & T & A & A \end{pmatrix}. \tag{1}$$

The columns in (1) correspond to chromosomal sites, which act as the basis for comparing different species: species with the same nucleotide at a given site are genetically similar and, thus, more likely to belong to the same cluster. We are prompted to view (1) column-by-column:

$$\mathbf{y}_{[5]} = (ACAAA, CCGAA, GGGGG, TAGCT, AAGAA, CAATA),$$

where we regard each component as a repeated measurement from a process that depends on the underlying clustering of the sample.

Abstracting from this illustration, we assume an infinite population of units labeled by the positive integers $\mathbb{N} = \{1, 2, \ldots\}$, with each unit $i = 1, 2, \ldots$ associated with a sequence $y_i = y_i^1 y_i^2 \ldots$ of indefinite length in a response space $\mathcal{Y}$. In general, the response space can be countable or uncountable, finite or infinite, but we focus on the case where $\mathcal{Y}$ is finite of order $k = 1, 2, \ldots$ For example, $\mathcal{Y} = \{\text{C, D}\}$ in Table 1 corresponds to $k = 2$, and $\mathcal{Y} = \{\text{A, C, G, T}\}$ in Table 2 corresponds to $k = 4$. The gene expression dataset from Section 2.3 corresponds to an uncountable response space. We represent the data as a $\mathcal{Y}$-valued array

$$\mathbf{y} = \begin{matrix} y_1 \\ y_2 \\ y_3 \\ \vdots \end{matrix} \begin{pmatrix} y_1^1 & y_1^2 & \cdots \\ y_2^1 & y_2^2 & \cdots \\ y_3^1 & y_3^2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}, \tag{2}$$

whose restriction to an observation of length-M sequences for a sample $[n] = \{1, \ldots, n\}$ is denoted by

$$\mathbf{y}_{[n]} = \begin{matrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{matrix} \begin{pmatrix} y_1^1 & y_1^2 & \cdots & y_1^M \\ y_2^1 & y_2^2 & \cdots & y_2^M \\ \vdots & \vdots & \ddots & \vdots \\ y_n^1 & y_n^2 & \cdots & y_n^M \end{pmatrix}. \tag{3}$$

3.2 Partition Data

A *partition*, or *clustering*, of $[n] = \{1, \ldots, n\}$ is a collection of nonempty, disjoint subsets $\pi = \{b_1, \ldots, b_r\}$ for which $\bigcup_{j=1}^{r} b_j = [n]$. The subsets $b_1, \ldots, b_r$ in π are called *blocks*, *clusters*, or *classes*. For example, the five partitions of $\{1, 2, 3\}$ are

$$\{\{1,2,3\}\}, \quad \{\{1\},\{2,3\}\}, \quad \{\{1,2\},\{3\}\}, \quad \{\{1,3\},\{2\}\}, \quad \{\{1\},\{2\},\{3\}\}.$$

We write $\mathcal{P}_{[n]}$ to denote the set of all partitions of $[n]$, for $n = 1, 2, \ldots$, and $\mathcal{P}_{\mathbb{N}}$ to denote the space of partitions of $\mathbb{N}$.

When $\mathcal{Y}$ is finite, the columns of $\mathbf{y}$ in (2) induce a sequence $\pi^1, \pi^2, \ldots$ of partitions through

$$i \text{ and } i' \text{ are in the same block of } \pi^j \text{ if and only if } y_i^j = y_{i'}^j. \tag{4}$$

For example, the array $\mathbf{y}_{[5]}$ in (1) induces the partition sequence

$$\{\{1,3,4,5\},\{2\}\}, \quad \{\{1,2\},\{3\},\{4,5\}\}, \quad \{\{1,2,3,4,5\}\}, \quad \{\{1,5\},\{2\},\{3\},\{4\}\}, \quad \{\{1,2,4,5\},\{3\}\}, \quad \{\{1\},\{2,3,5\},\{4\}\}. \tag{5}$$

The partition sequence induced by $\mathbf{y}$ through (4) preserves some, but not all, structure of the data; however, for many problems, this reduction retains the essential information in $\mathbf{y}_{[n]}$, and we adopt the following assumption throughout the article.

Assumption 1. The partition sequence $\pi^1, \pi^2, \ldots$ induced by (4) is sufficient for clustering from the data array in (2).

Assumption 1 is justified in many applications we consider. For example, in Table 1, responses are labeled according to whether or not each justice's ruling coincides with the majority of other justices. However, we hope that justices rule without regard to whether they will be in the majority—naive as that may be (Irons 2006). For the data in Table 2, there are known biological relationships between compatible bases (A, T) and (C, G), but the clustering task only considers similarities and differences between species. Again, Assumption 1 seems appropriate both logically and empirically, based on the outcomes of Section 7.3. We stress that we can drop Assumption 1 with a slight modification to our model, but we streamline the presentation and only mention this in Section 8.

3.3 Data Transformations

Assumption 1 asserts that the process underlying (2) is invariant under one-to-one transformations of the response space. Two other statistically relevant transformations are relabeling and subsampling, which entail operations on the sampled units and are, therefore, extrinsic to the data-generating process.

3.3.1 Relabeling. For bookkeeping, we assign sampled units arbitrary labels in $[n] = \{1, \ldots, n\}$, where n is the sample size. Any such labeling is equally valid for our purposes, and so relabeling units by any permutation $\sigma : [n] \to [n]$ does not alter the information in the data. Under relabeling units by σ, the data array in (3) becomes

$$\mathbf{y}_{[n]}^{\sigma} = \begin{matrix} y_{\sigma(1)} \\ y_{\sigma(2)} \\ \vdots \\ y_{\sigma(n)} \end{matrix} \begin{pmatrix} y_{\sigma(1)}^1 & y_{\sigma(1)}^2 & \cdots & y_{\sigma(1)}^M \\ y_{\sigma(2)}^1 & y_{\sigma(2)}^2 & \cdots & y_{\sigma(2)}^M \\ \vdots & \vdots & \ddots & \vdots \\ y_{\sigma(n)}^1 & y_{\sigma(n)}^2 & \cdots & y_{\sigma(n)}^M \end{pmatrix},$$

which induces a sequence of relabeled partitions $\pi^{1\sigma}, \ldots, \pi^{M\sigma}$, where

$$i \text{ and } i' \text{ are in the same block of } \pi^{\sigma} \quad \text{if and only if} \quad \sigma^{-1}(i) \text{ and } \sigma^{-1}(i') \text{ are in the same block of } \pi. \tag{6}$$

For example, the relabeling of $\mathbf{y}_{[5]}$ in (1) by

$$\sigma = \begin{pmatrix} 1\,2\,3\,4\,5 \\ 2\,3\,1\,5\,4 \end{pmatrix}$$

is

$$\begin{matrix} y_3 \\ y_1 \\ y_2 \\ y_5 \\ y_4 \end{matrix} \begin{pmatrix} A & G & G & G & G & A \\ A & C & G & T & A & C \\ C & C & G & A & A & A \\ A & A & G & T & A & A \\ A & A & G & C & A & T \end{pmatrix},$$

which has associated partition sequence

$$\{\{1,2,4,5\},\{3\}\}, \quad \{\{1\},\{2,3\},\{4,5\}\}, \quad \{\{1,2,3,4,5\}\},$$
$$\{\{1\},\{2,4\},\{3\},\{5\}\}, \quad \{\{1\},\{2,3,4,5\}\}, \quad \{\{1,3,4\},\{2\},\{5\}\}.$$

Invariance of the data-generating process to relabeling underlies the important statistical concepts of exchangeability and label equivariance, see Section 5.1.1.

3.3.2 Subsampling. Though the population is infinite, we only observe a finite sample of n units. Subsampling $[m] \subseteq [n]$, $m \leq n$, transforms the data array by *restriction* to its first m rows,

$$\mathbf{y}_{[n]} \mapsto \mathbf{y}_{[m]} = \begin{matrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{matrix} \begin{pmatrix} y_1^1 & y_1^2 & \cdots \\ y_2^1 & y_2^2 & \cdots \\ \vdots & \vdots & \ddots \\ y_m^1 & y_m^2 & \cdots \end{pmatrix},$$

and restricts the partition sequence to partitions of $[m]$, $\pi_{[m]}^1, \pi_{[m]}^2, \ldots$, where

$$\pi_{[m]} = \{b_1 \cap [m], b_2 \cap [m], \ldots\} \setminus \{\emptyset\} \tag{7}$$

is the restriction of $\pi_{[n]} = \{b_1, b_2, \ldots\}$ to $[m]$ by deleting elements not in $[m]$. For example, the restriction of $\mathbf{y}_{[5]}$ in (1) under subsampling $[3] \subset [5]$ is

$$\mathbf{y}_{[3]} = \begin{matrix} y_1 \\ y_2 \\ y_3 \end{matrix} \begin{pmatrix} A & C & G & T & A & C \\ C & C & G & A & A & A \\ A & G & G & G & G & A \end{pmatrix},$$

with associated partition sequence

$$\{\{1,3\},\{2\}\}, \quad \{\{1,2\},\{3\}\}, \quad \{\{1,2,3\}\},$$
$$\{\{1\},\{2\},\{3\}\}, \quad \{\{1,2\},\{3\}\}, \quad \{\{1\},\{2,3\}\}.$$

In principle, a stochastic process on infinite arrays generates the data. In practice, the model is specified by the finite-dimensional distributions of the sampled data. These two notions coincide only if the finite-dimensional distributions are *sampling consistent* with respect to the above subsampling operation; see Section 5.1.2.

3.4 Partition Models and the Data-Generating Process

Under Assumption 1, we model the data as a sequence of random partitions of $\mathbb{N}$, denoted $\Pi_{\mathbb{N}} = (\Pi^1, \Pi^2, \ldots)$. In general, this sequence can exhibit arbitrary dependence, but we assume $\Pi^1, \Pi^2, \ldots$ are independent and identically distributed (iid) throughout.

In the iid case, a *partition model* parameterized by Θ is a family $P_\Theta = \{P^\theta\}_{\theta\in\Theta}$ of probability distributions on partitions of $\mathbb{N}$. Each P^θ determines a collection of finite-dimensional distributions $(P_n^\theta)_{n=1,2,\ldots}$, where P_n^θ is the measure P^θ induces on partitions of $[n]$ by subsampling $[n] \subset \mathbb{N}$,

$$P_n^\theta(\pi) := P^\theta(\{\pi^* \in \mathcal{P}_\mathbb{N} : \pi^*_{[n]} = \pi\}), \quad \pi \in \mathcal{P}_{[n]}. \tag{8}$$

The collection $(P_n^\theta)_{n=1,2,\ldots}$ in (8) is *sampling consistent* by default: for every $n = 1, 2, \ldots$ and $m = 1, \ldots, n$,

$$P_m^\theta(\pi) = \sum_{\pi^* \in \mathcal{P}_{[n]} : \pi^*_{[m]} = \pi} P_n^\theta(\pi^*), \quad \text{for all } \pi \in \mathcal{P}_{[m]}. \tag{9}$$

In words, sampling consistency means that P_m^θ is the marginal distribution of P_n^θ under subsampling $[m] \subseteq [n]$, for every $m \leq n$. Because of sampling consistency, we can interpret the observed sequence as a realization of an unobserved *data-generating process* on the entire population. Lack of consistency prohibits this interpretation, as parameters estimated from a finite sample are meaningless for the population and unjustifiable for out-of-sample inference. If, in addition to (9), the family $(P_n^\theta)_{n\in\mathbb{N};\theta\in\Theta}$ satisfies

$$P_n^\theta(\pi^\sigma) = P_n^\theta(\pi), \quad \pi \in \mathcal{P}_{[n]}, \tag{10}$$

for all permutations $\sigma : [n] \to [n]$, for all $n \in \mathbb{N}$, then P_Θ is *exchangeable*, and the distribution of $\Pi_n \sim P_n^\theta$ does not depend on arbitrary labeling of the units.

While any one of these properties is often easy to impose, their combination is far more difficult. The lack of sampling consistency in some widely used models is sometimes swept under the rug, as our next example demonstrates.

3.4.1 Product Partition Models. With $c(b) > 0$ denoting the *cohesion* of each $b \subset \mathbb{N}$, the *product partition model* (PPM) with cohesion function c is the family of finite-dimensional distributions $(P_n^c)_{n\in\mathbb{N}}$ defined by

$$P_n^c(\pi) \propto \prod_{b\in\pi} c(b), \quad \pi \in \mathcal{P}_{[n]}. \tag{11}$$

Note that a product partition model is not exchangeable unless $c(b) = c_{\#b}$ depends on $b \subset \mathbb{N}$ only through its cardinality $\#b$. Exchangeable product partition models specialize the class of *Gibbs partitions*; see Pitman (2006), chap. 1 and 2.

For concreteness, let $c(b) = 1$, which corresponds to the uniform distribution and is, therefore, exchangeable. Then

$$P_2^c(\{\{1\},\{2\}\}) = 1/2, \tag{12}$$

but

$$P_3^c(\{\{1,3\},\{2\}\}) + P_3^c(\{\{1\},\{2,3\}\}) + P_3^c(\{\{1\},\{2\},\{3\}\})$$
$$= 1/5 + 1/5 + 1/5 = 3/5 \neq 1/2, \tag{13}$$

so this family is not sampling consistent. Together, the three partitions in (13) correspond to the event

$$\{1 \text{ and } 2 \text{ not in the same block}\},$$

which does not depend on the sample size. But (12) and (13) show that the probability of this event varies with sample size, indicating an inherent flaw in models that lack consistency.

Despite this issue, product partition models enjoy widespread use, for example, Quintana and Iglesias (2003, sec. 5.3) used the uniform distribution in Bayesian clustering. Park

and Dunson (2010, sec. 2) even claimed that PPMs are sampling consistent. In fact, the marginal distributions of product partition models not only fail to be consistent but also need not be of product partition type. For example, the uniform distribution on $\mathcal{P}_{[5]}$ induces the marginal distribution $P^1_{4,5}$ on $\mathcal{P}_{[4]}$, where $P^1_{4,5}(\pi) = (\#\pi + 1)/52$ and $\#\pi$ denotes the number of blocks of π. Now, suppose $P^1_{4,5}$ comes from an exchangeable product partition model with cohesions $c^*(j)$ for $j = 1, 2, \ldots$ Without loss of generality, we can arbitrarily assign $c^*(1) = 1$, so that $P^1_{4,5}(\{\{1\}, \{2\}, \{3\}, \{4\}\}) = 5/52 \propto c^*(1)^4$ immediately implies the normalizing constant $Z = 52/5$. Now, $P^1_{4,5}(\{\{1, 2\}, \{3\}, \{4\}\}) = c^*(1)^2 c^*(2)/Z = 4/52$ implies $c^*(2) = 4/5$. But these imply

$$\frac{3}{52} = P^1_{4,5}(\{\{1, 2\}, \{3, 4\}\}) = \frac{c^*(2)^2}{Z} = \frac{(4/5)^2}{52/5} = \frac{16/5}{52},$$

a contradiction.

3.4.2 Ewens Process. The product partition model is exchangeable and consistent only for the family of cohesion functions satisfying $c(b) = \theta(\#b - 1)!$ (Kerov 2005), for $\theta > 0$, which yields the celebrated *Ewens process* $P_\Theta = (P^\theta_n)_{n\in\mathbb{N};\,\theta>0}$, with

$$P^\theta_n(\pi) = \frac{\theta^{\#\pi}}{\theta^{\uparrow n}} \prod_{b\in\pi} (\#b - 1)!, \quad \pi \in \mathcal{P}_{[n]}, \tag{14}$$

where $\theta^{\uparrow n} = \theta(\theta + 1)\ldots(\theta + n - 1)$ (Ewens 1972). The Ewens process also arises from Dirichlet process priors (Ferguson 1973). Therefore, despite concerns with the product partition model highlighted in Section 3.4.1, many applications of product partition models avoid these issues by defaulting to the Dirichlet process prior. We discuss the Ewens process further in Section 4.1.

4. THE THREE-PARAMETER CLUSTER MODEL

For fixed $n = 1, 2, \ldots$, $\alpha > 0$, $k = 1, 2, \ldots$, and $B_{[n]} \in \mathcal{P}_{[n]}$, we observe the combinatorial identity

$$\prod_{b\in B_{[n]}} (k\alpha)^{\uparrow \#b} = \sum_{\pi\in\mathcal{P}_{[n]}} k^{\downarrow\#\pi} \prod_{b\in B_{[n]}} \prod_{b'\in\pi} \alpha^{\uparrow\#(b\cap b')}, \tag{15}$$

where $k^{\downarrow j} = k(k-1)\cdots(k-j+1)$, $\alpha^{\uparrow j} = \alpha(\alpha+1)\cdots(\alpha+j-1)$, and $\alpha^{\uparrow 0} = 1$, see Crane (2013). From (15), we derive a probability distribution on partitions of $[n]$ by normalization:

$$P^{\alpha,k,B_{[n]}}_n(\pi) = k^{\downarrow\#\pi} \prod_{b\in B_{[n]}} \frac{\prod_{b'\in\pi} \alpha^{\uparrow\#(b\cap b')}}{(k\alpha)^{\uparrow\#b}}, \quad \pi \in \mathcal{P}_{[n]}\,. \tag{16}$$

We call (16) the *three-parameter cluster distribution* with *mutation rate* α and *clustering parameter* $B_{[n]}$.

Definition 1 (The three-parameter cluster model). For $\Theta = (0, \infty) \times \mathbb{N} \times \mathcal{P}_{\mathbb{N}}$, the (α, k, B)*-three-parameter cluster model* is the family of distributions $(P^\theta_n)_{n\in\mathbb{N};\,\theta\in\Theta}$, where P^θ_n is given by (16) for $\theta = (\alpha, k, B_{[n]})$.

Though derived naively from the combinatorial identity (15), the three-parameter model possesses nice theoretical and empirical properties that suit it to clustering applications. We develop these properties systematically in Sections 5–7. First, we highlight its close connection to the Ewens process, which explains why we expect this family to occur in partition modeling.

4.1 Relation to the Ewens Process

The Ewens–Pitman distribution (Pitman 2006) is a two-parameter extension of the Ewens distribution (Ewens 1972), which originally appeared in the study of neutral allele sampling. When $B_{[n]} = \mathbf{1}_{[n]}$ is the one-block partition of $[n]$, (16) reduces to

$$P^{\alpha,k,\mathbf{1}_{[n]}}_n(\pi) = \frac{k^{\downarrow\#\pi}}{(k\alpha)^{\uparrow n}} \prod_{b\in\pi} \alpha^{\uparrow\#b}, \quad \pi \in \mathcal{P}_{[n]}, \tag{17}$$

which coincides with the Ewens–Pitman$(-\alpha, k\alpha)$ distribution. Thus, the Ewens–Pitman distribution acts as a null model for the three-parameter cluster model. Taking $\alpha \to 0$ and $k \to \infty$ such that $k\alpha \to \theta > 0$ recovers the Ewens distribution (14) in the limit.

Remark 1. Identity (15) is a special case of the more general identity,

$$\prod_{b\in B_{[n]}} (k\alpha_b)^{\uparrow\#b} = \sum_{\pi\in\mathcal{P}_{[n]}} k^{\downarrow\#\pi} \prod_{b\in B_{[n]}} \prod_{b'\in\pi} \alpha_b^{\uparrow\#(b\cap b')}, \tag{18}$$

where $(\alpha_b,\ b \in B_{[n]})$ are class specific mutation rates in the obvious extension of (16),

$$P^{\alpha,k,B_{[n]}}_n(\pi) = k^{\downarrow\#\pi} \prod_{b\in B_{[n]}} \frac{\prod_{b'\in\pi} \alpha_b^{\uparrow\#(b\cap b')}}{(k\alpha_b)^{\uparrow\#b}}, \quad \pi \in \mathcal{P}_{[n]}\,. \tag{19}$$

We only develop the model in (16), but all essential theoretical properties carry over to this more general setting.

5. PROPERTIES OF THE THREE-PARAMETER MODEL

We assume the response space $\mathcal{Y}$ is finite with cardinality $k = 1, 2, \ldots$, and thus the sequence of partitions induced by (4) lies in the subspace of partitions with k or fewer blocks. Since $k^{\downarrow j} = 0$ for $j > k$, (16) is supported on the subspace of partitions of $[n]$ with at most k blocks, as it should be. Also, for any $\alpha > 0$ and $k = 1, 2, \ldots$, either $P^{\alpha,k,\pi}_n(\pi) \neq P^{\alpha,k,\pi'}_n(\pi)$ or $P^{\alpha,k,\pi}_n(\pi') \neq P^{\alpha,k,\pi'}_n(\pi')$, for all $\pi \neq \pi' \in \mathcal{P}_{[n]}$, for all $n = 1, 2, \ldots$, and so the model is identifiable for the subparameter $B_{[n]} \in \mathcal{P}_{[n]}$. Since $B \in \mathcal{P}_{\mathbb{N}}$ is an infinite partition of the population, it is impossible to identify it based on a finite amount of data. In the following sections, we establish more nuanced properties.

5.1 Invariance Properties

5.1.1 Exchangeability and Equivariance Under Relabeling. By (6), any permutation $\sigma : [n] \to [n]$ induces a bijective relabeling of the sample and its associated partition sequence. Consequently, no information is lost under relabeling and inference should be unaffected.

In the three-parameter model, relabeling the sample transforms the parameter space $\Theta \mapsto \sigma\Theta$ by relabeling the partition parameter, $(\alpha, k, B) \mapsto (\alpha, k, B^\sigma)$. In general, *label equivariance* implies invariance of the model under the induced action $\Theta \mapsto \sigma\Theta$, for every permutation $\sigma : \mathbb{N} \to \mathbb{N}$ that fixes all but finitely many elements. For a partition model $P_\Theta =$

$(P_n^\theta)_{n\in\mathbb{N};\theta\in\Theta}$ to be label equivariant, it is sufficient that

$$P_n^{\sigma\theta}(\pi^\sigma) = P_n^\theta(\pi), \quad \pi \in \mathcal{P}_{[n]}, \quad \theta \in \Theta, \tag{20}$$

for all permutations $\sigma : [n] \to [n]$, for all $n \in \mathbb{N}$.

5.1.2 Sampling Consistency and Lack of Interference. For the three-parameter model $P^{\alpha,k,B}$, sampling consistency implies

$$P_m^{\alpha,k,B}(\pi) = \sum_{\pi'\in\mathcal{P}_{[n]}:\,\pi'_{[m]}=\pi} P_n^{\alpha,k,B}(\pi'), \quad \pi \in \mathcal{P}_{[m]}, \quad \text{for all } m \le n, \tag{21}$$

where B is any partition of $\mathbb{N}$. Lack of interference goes a step beyond sampling consistency by requiring

$$P_m^{\alpha,k,B_{[m]}}(\pi) = \sum_{\pi'\in\mathcal{P}_{[n]}:\,\pi'_{[m]}=\pi} P_n^{\alpha,k,B_{[n]}}(\pi'), \quad \pi \in \mathcal{P}_{[m]}, \quad \text{for all } m \le n, \tag{22}$$

where now the left-hand side depends only on $B_{[m]}$ and the right-hand side depends only on $B_{[n]}$. Condition (22) is critically important in cluster inference.

We assume the population segregates according to a partition B of $\mathbb{N}$. But since we only observe data for the sample $[n] \subset \mathbb{N}$, we can only hope to infer the restriction $B_{[n]}$. The data array $\mathbf{y}_{[n]}$ is sufficient for inference of $B_{[n]}$ if and only if its distribution depends on B only through $B_{[n]}$. Therefore, *lack of interference* for a partition model P_Θ corresponds to consistency under joint projection of the parameter space $\Theta \mapsto \Theta_{[n]}$ and the sample space $\mathbf{y}_\mathbb{N} \mapsto \mathbf{y}_{[n]}$, for each $n = 1, 2, \ldots$. The three-parameter model satisfies this stronger condition under the projection $(\alpha, k, B) \mapsto (\alpha, k, B_{[n]})$.

We summarize these properties in the following theorem, whose proof we delay to Section 8.

Theorem 1. The family of three-parameter cluster distributions determines a data-generating process for the population. In particular, for (α, k, B) ranging over $(0, \infty) \times \mathbb{N} \times \mathcal{P}_\mathbb{N}$, the family $P^{\alpha,k,B} = (P_n^{\alpha,k,B_{[n]}})_{n=1,2,\ldots}$ in (16) is label equivariant and sampling consistent.

5.2 Implications of Lack of Interference

Lack of interference vests our model with considerable inferential power, as we now discuss.

5.2.1 Missing Data. Missing data are common in the applications of Section 2. Less satisfactory approaches to handling it include filling in missing values or even removing measurements with missing observations altogether (Thurstone and Degan 1951; Sirovich 2003). Assuming missing observations occur at random, lack of interference allows us to fit our model to the observed dataset *as is*, without removing, cleaning, or otherwise defiling the data. Intuitively, lack of interference is related to sampling consistency, whose importance we demonstrated in Section 3.4.1. This aspect of our model features in each of the applications in Section 7.

5.2.2 Supervised Learning. In some applications, we observe a data sequence $\Pi_{[n]}$ for a sample $[n] = \{1, \ldots, n\}$ along with a partial clustering $B_{[m]}$, the restriction of the partition parameter to a subsample $[m] \subset [n]$. To infer the total clustering $B_{[n]}$ of the sample, supervised learning incorporates the information in the data $\Pi_{[n]}$ as well as the partial information $B_{[m]}$ about the clustering. In the three-parameter model, any partial observation $B_{[m]}$ determines a valid subparameter that consists of all partitions of $[n]$ that agree with $B_{[m]}$ under restriction. In particular, for inferring $B_{[n]}$, given $B_{[m]}$ and $\Pi_{[n]}$, we need only maximize the likelihood over the restricted parameter set

$$\{B^* \in \mathcal{P}_{[n]} : \; B^*_{[m]} = B_{[m]}\}.$$

5.3 Interpretation of Parameters

5.3.1 Number of Clusters. We have already mentioned that k is the cardinality of the response space, for example, $k = 2$ (concurrence and dissent) in Table 1 and $k = 4$ (A, C, G, T) in Table 2. We stress that k is *unrelated* to the number of clusters in the population. Whereas the model is only valid for sequences of partitions with k or fewer blocks, the clustering parameter B can be any partition. In general, the number of clusters of B should be left unspecified, unless a maximum number of clusters is known in advance. For example, in Section 2.3, the population separates into two types of leukemia, and so the inferred clustering should have at most two blocks. In general, any projective subsystem of $\{\mathcal{P}_{[n]}\}_{n=1,2,\ldots}$ determines a submodel of $P^{\alpha,k,B}$.

5.3.2 Clustering Parameter. The distribution in Table 3 assigns the most mass to the true clustering B and higher masses to partitions that more closely resemble B. For example, the partition $\pi = \{\{1, 2, 3, 4\}\}$, with probability $1/12$, respects the structure of B in that elements in the same block of B are in the same block of π. On the other hand, $\pi' = \{\{1\}, \{2, 4\}, \{3\}\}$, with probability $1/24$, does not respect the structure of B because it groups together elements 2 and 4, which are in different blocks of B. This example suggests the interpretation of B as a *center of mass* about which observations scatter.

5.3.3 Mutation Rate. Deviations from B are controlled by the magnitude of the *mutation rate* $\alpha > 0$. From (16), it is apparent that elements in different blocks of B behave independently of one another, that is, their probability of appearing in the same block of an (α, k, B)-partition is $1/k$. On the other hand, elements in the same cluster of B appear in the same block of an (α, k, B)-partition with probability $(\alpha + 1)/(k\alpha + 1) > 1/k$. The interpretation of α as a *variance* parameter stems from the

Table 3. Mass function for partitions of $\{1, 2, 3, 4\}$ under the (α, k, B)-model with $\alpha = 1$, $k = 3$, and $B = \{\{1, 2\}, \{3, 4\}\}$. The partition $\{\{1\}, \{2\}, \{3\}, \{4\}\}$ is assigned 0 mass because it exceeds the maximum of $k = 3$ blocks

π	$P_n^{\alpha,k,B}(\pi)$	π	$P_n^{\alpha,k,B}(\pi)$
$\{\{1, 2\}, \{3, 4\}\}$	$1/6$	$\{\{1, 3\}, \{2, 4\}\}$	$1/24$
$\{\{1, 2, 3, 4\}\}$	$1/12$	$\{\{1, 3\}, \{2\}, \{4\}\}$	$1/24$
$\{\{1, 2, 3\}, \{4\}\}$	$1/12$	$\{\{1, 4\}, \{2, 3\}\}$	$1/24$
$\{\{1, 2, 4\}, \{3\}\}$	$1/12$	$\{\{1\}, \{2, 3\}, \{4\}\}$	$1/24$
$\{\{1, 2\}, \{3\}, \{4\}\}$	$1/12$	$\{\{1, 4\}, \{2\}, \{3\}\}$	$1/24$
$\{\{1\}, \{2, 3, 4\}\}$	$1/12$	$\{\{1\}, \{2, 4\}, \{3\}\}$	$1/24$
$\{\{1\}, \{2\}, \{3, 4\}\}$	$1/12$	$\{\{1\}, \{2\}, \{3\}, \{4\}\}$	0
$\{\{1, 3, 4\}, \{2\}\}$	$1/12$		

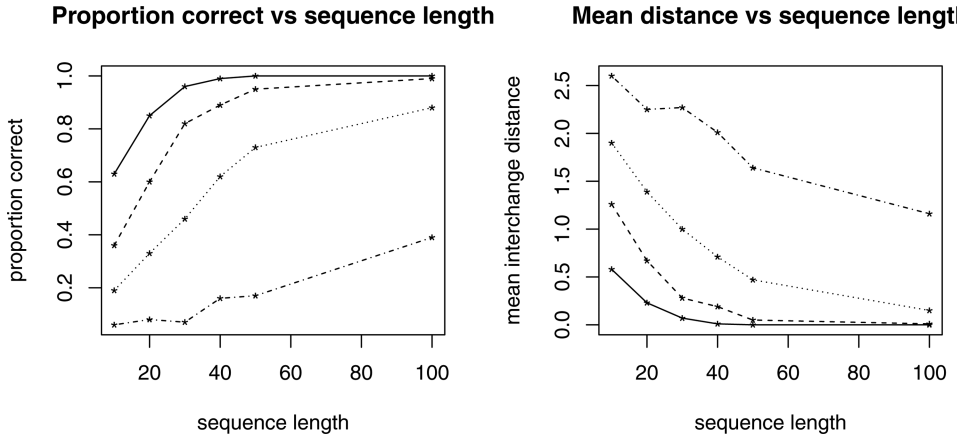

Figure 1. Plots of simulation results for the three-parameter model for a sample of size 8. We generated sequences of lengths 10, 20, 30, 40, 50, and 100 from the (α, k, B)-model with $k = 2$, $B = \{\{1, 2, 3, 4\}, \{5, 6, 7, 8\}\}$, and $\alpha = 1/2, 1, 2, 5$. *Proportion correct* is the proportion of times the maximum likelihood estimate of B equaled the true value in 100 iterations. *Average distance* is the mean interchange distance between the maximum likelihood estimate and the true value B for each α. The lines are ordered according to $\alpha = 1/2, 1, 2, 5$, from top to bottom in the left panel and bottom to top in the right panel.

observation that if i and j are in the same block of $B_{[n]}$, then

$$P_n^{\alpha,k,B_{[n]}}(\{i \text{ and } j \text{ in the same block}\}) \approx 1 \qquad \text{as } \alpha \to 0$$

and

$$P_n^{\alpha,k,B_{[n]}}(\{i \text{ and } j \text{ in the same block}\}) \approx 1/k \qquad \text{as } \alpha \to \infty.$$

This simple relationship between α and B explains the phenomenon in Figure 1, whereby longer sequences are needed to accurately estimate B when α is large. In the datasets we have studied, there is little noise: α is usually no larger than 2 and is often much smaller than 1. As a result, short sequences are sufficient to obtain a reliable clustering.

The interpretation of α as a mutation rate also agrees with the close connection between the three-parameter cluster model and the Ewens process, see Section 4.1. The parameter in the Ewens distribution corresponds to the mutation rate in a certain class of population genetics models (Tavaré 2001), with higher rate corresponding to more blocks, on average, in the observed random partition. In our model, the mutation rate quantifies how strongly the data obey the underlying clustering B, that is, the amount of noise in the data. As an illustration, let $k = 2$ and consider the values $\alpha = 1/10$, $\alpha = 1$, and $\alpha = 10$. For elements i and j in the same block of B, the probabilities that i and j occur in the same block of a random (α, k, B)-partition are $11/12$, $2/3$ and $11/21$, respectively.

5.4 Parameter Estimation

The log-likelihood of $(\alpha, B_{[n]})$ based on a partition sequence $\pi = (\pi^1, \ldots, \pi^M)$ of $[n]$ from the three-parameter model is

$$l_n(\alpha, B_{[n]}; \pi) = -\sum_{b \in B_{[n]}} \left(M \log(k\alpha)^{\uparrow \#b} - \sum_{j=1}^{M} \sum_{b' \in \pi^j} \log \alpha^{\uparrow \#(b \cap b')} \right) + \text{const}\,(\alpha, B_{[n]}), \tag{23}$$

where $\text{const}(\alpha, B_{[n]}) = \sum_{j=1}^{M} \log k^{\downarrow \#\pi^j}$ is constant with respect to $(\alpha, B_{[n]})$ and is henceforth omitted. Each term has the form $\log \alpha^{\uparrow j}$, $j = 0, 1, \ldots$, for which

$$\frac{d}{d\alpha} \log \alpha^{\uparrow j} = \sum_{i=0}^{j-1} \frac{d}{d\alpha} \log(\alpha + i) = \sum_{i=0}^{j-1} \frac{1}{\alpha + i} = s_j(\alpha);$$

and so the score function with respect to α is

$$\frac{\partial}{\partial \alpha} l_n(\alpha, B_{[n]}; \pi) = -\sum_{b \in B_{[n]}} \left(M s_{\#b}(\alpha) - \sum_{j=1}^{M} \sum_{b' \in \pi^j} s_{\#(b \cap b')}(\alpha) \right).$$

For fixed $B_{[n]}$, any gradient descent algorithm, for example, Gauss–Newton, finds the maximum of $l_n(\alpha, B_{[n]}; \pi)$ with respect to α, and consistency of the maximum likelihood estimator of α, with $B_{[n]}$ fixed at its true value, follows from standard estimation theory. On the other hand, optimizing (23) over the discrete state space of partitions is computationally challenging. The unrestricted parameter space of all partitions is enumerated by the Bell numbers, which grow exponentially with n (Stanley 2012). Even if the number of blocks of $B_{[n]}$ is bounded by $\tilde{k} = 1, 2, \ldots$, the search space is of order $\tilde{k}^n$, still too large for moderately sized samples. We mitigate this issue by implementing an efficient Markov chain search algorithm; see Appendix A.

In practice, we estimate $(\alpha, B_{[n]})$ by iterating between a gradient descent method for α, with $B_{[n]}$ fixed, and a randomized search algorithm for $B_{[n]}$, with α fixed. Though we do not prove consistency of the joint maximum likelihood estimator as sequence length increases, extensive computation suggests that the expectation of $l_n(\alpha, B_{[n]}; \Pi)$, where $\Pi \sim P_n^{\alpha_0,k,B_0}$, is maximized at the true parameter value (α_0, B_0), which implies the maximum likelihood estimator converges to the true value as sequence length increases. Our simulation results in Figure 1 and Table 4 further support this conclusion and also suggest that maximum likelihood estimation is efficient.

6. SIMULATION STUDY

We demonstrate key properties of the maximum likelihood estimator using simulated data. In Section 6.1, we show

Table 4. Summary of simulation results for joint estimation of (α, B) from the three-parameter model. We write (α_*, B_*) to denote the maximum likelihood estimate of (α, B). The column labeled "$\%B_*$" gives the percentage of simulations for which B_* equals the true value B. For α, we write $\hat{\alpha}_*$ to denote the mean of α_* with estimated standard error (s.e.) in parentheses

$k = 2$	$B = 12/345678$	$\alpha = 1/2$	$k = 4$	$B = 12/345678$	$\alpha = 1/2$
Length	$\%B_*$	$\hat{\alpha}_*$ (s.e.)	Length	$\%B_*$	$\hat{\alpha}_*$ (s.e.)
10	63	0.49 (0.25)	10	77	0.51 (0.22)
20	93	0.53 (0.20)	20	90	0.52 (0.14)
30	100	0.52 (0.12)	30	100	0.54 (0.13)
40	100	0.54 (0.13)	40	100	0.53 (0.11)
50	100	0.54 (0.12)	50	100	0.53 (0.10)
$k = 2$	$B = 12/345678$	$\alpha = 1$	$k = 4$	$B = 12/345678$	$\alpha = 1$
Length	$\%B_*$	$\hat{\alpha}_*$ (s.e.)	Length	$\%B_*$	$\hat{\alpha}_*$ (s.e.)
10	37	0.76 (0.27)	10	30	0.85 (0.23)
20	43	0.91 (0.28)	20	67	1.00 (0.18)
30	77	1.00 (0.24)	30	87	1.03 (0.16)
40	83	1.01 (0.21)	40	93	1.01 (0.18)
50	90	1.04 (0.19)	50	97	1.04 (0.15)
$k = 2$	$B = 1234/5678$	$\alpha = 1/2$	$k = 4$	$B = 1234/5678$	$\alpha = 1/2$
Length	$\%B_*$	$\hat{\alpha}_*$ (s.e.)	Length	$\%B_*$	$\hat{\alpha}_*$ (s.e.)
10	47	0.49 (0.24)	10	70	0.53 (0.25)
20	77	0.53 (0.18)	20	90	0.52 (0.17)
30	90	0.54 (0.15)	30	100	0.51 (0.12)
40	97	0.54 (0.14)	40	100	0.50 (0.10)
50	100	0.53 (0.10)	50	100	0.51 (0.09)
$k = 2$	$B = 1234/5678$	$\alpha = 1$	$k = 4$	$B = 1234/5678$	$\alpha = 1$
Length	$\%B_*$	$\hat{\alpha}_*$ (s.e.)	Length	$\%B_*$	$\hat{\alpha}_*$ (s.e.)
10	36	0.85 (0.29)	10	33	0.87 (0.26)
20	60	0.95 (0.22)	20	67	1.00 (0.18)
30	82	1.01 (0.20)	30	87	1.03 (0.16)
40	83	1.02 (0.17)	40	93	1.01 (0.18)
50	97	1.02 (0.16)	50	97	1.04 (0.15)

simulation results under different choices of (α, k, B) for a sample of size 8. In this case, we use exhaustive search to find joint maximum likelihood estimates for (α, B). Numerical computations illustrate asymptotic convergence to the true value. Since we fix the sample size throughout our simulation study, we write B, rather than $B_{[n]}$, to represent the clustering of the sample.

To measure precision of maximum likelihood estimates for B, we define the *interchange distance* between π and π' as the minimum number of elements that must be moved from one block to another to change π into π', equivalently π' into π. For example, the interchange distance between $\pi = \{\{1, 2, 3\}, \{4, 5\}, \{6\}\}$ and $\pi' = \{\{1, 4, 5\}, \{2, 3, 6\}\}$ is 2 since elements 1 and 6 must be moved to change π into π'. This choice of metric is motivated by similar distances for phylogenetic trees, for example, nearest-neighbor interchange (Sokal and Sneath 1963).

6.1 Asymptotic Consistency

Figure 1 displays simulation results for the exact maximum likelihood estimate of B using exhaustive search. Together, the two plots suggest that the distribution of the maximum likelihood estimator is centered on the true value, in the sense of interchange distance, and becomes tightly concentrated as sequence length increases. Also apparent is the interplay between mutation rate and accuracy of the partition estimate: increasing the mutation rate diminishes the role of the partition parameter, requiring more data to obtain an accurate estimate.

Table 4 shows simulation results for the joint estimation of (α, B) under different choices of α, k, and B. The table supports the conclusion that the maximum likelihood estimator for (α, B) is jointly consistent as sequence length increases, regardless of the values of α, k, and B. It also suggests that accuracy of the maximum likelihood estimator increases with k, which is intuitive since small values of k dampen the signal that any one component of the data sequence can emit. Figure 1 and Table 4 are indicative of what we observe in much more extensive simulations over a range of values for (α, k, B).

6.2 Estimation Using Random Search

For sample sizes larger than 15, or so, the space of partitions is too large to maximize the likelihood by exhaustive search. In the modern era, our model's viability hinges on its capacity to handle larger sample sizes. For this task, we search the space of partitions using a Markov chain algorithm, which we describe in Appendix A.

Effectiveness of the randomized search algorithm depends crucially on its ability to search the entire space of partitions without getting stuck at local maxima or in low-likelihood regions of the state space. To accomplish these tasks, our algorithm iterates between local and global moves, where global moves ensure that the whole space is explored efficiently and local moves look for incremental improvements by perturbing only one element at a time. Numerical tests reveal that our algorithms consistently find partitions in a neighborhood of the true

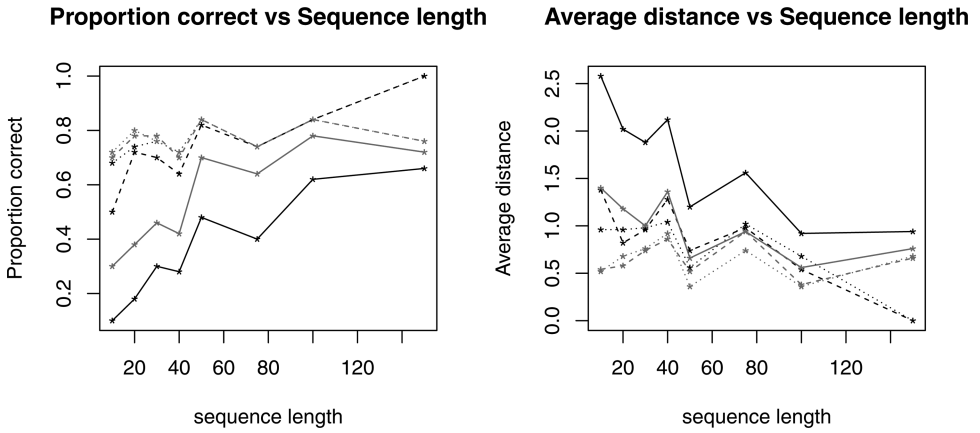

Figure 2. Performance of three-parameter model (black) and interchange algorithm (red) on partition sequence data generated for a sample of size $n = 8$ from the Gauss–Ewens process with $\lambda = 1$ (solid), $\lambda = 5$ (dotted), and $\lambda = 10$ (dashed).

clustering. In Section 7.2, we demonstrate this algorithm on a political science dataset with sample size 100.

6.3 Robustness to Model Misspecification

To test robustness to model misspecification, we simulated partition sequence data from a different model entirely. The Gauss–Ewens cluster process (McCullagh and Yang 2008) is a parametric Bayesian model for clustering from continuous response data. Under this model, the clustering parameter B is a random draw from the Ewens process with parameter $\theta > 0$. Given B, the response vectors are multivariate normal with mean vector μ and covariance matrix Σ_B. In our simulations, we take

$$\Sigma_B(i, i') = \begin{cases} \delta_{ii'} + \lambda, & i \text{ and } i' \text{ are in the same block of } B \\ \delta_{ii'}, & \text{otherwise,} \end{cases}$$

where $\delta_{ii'} = \mathbf{1}\{i = i'\}$ and $\lambda > 0$. In words, Σ_B models dependence so that observations from individuals in the same block of B are positively correlated and observations from individuals in different blocks of B are independent. Small values of λ indicate less correlation and, thus, more noise with respect to the underlying clustering.

To generate partition data, we first simulate a sequence of independent, identically distributed vectors $(Y^1, \ldots, Y^M)$ from the Gauss–Ewens process, where $Y^j = (Y_1^j, \ldots, Y_n^j)$ is multivariate normal with mean vector μ and covariance Σ_B, for $B \sim \text{Ewens}(1)$. From $(Y^1, \ldots, Y^M)$, we obtain a partition sequence $(\Pi^1, \ldots, \Pi^M)$, where Π^j is obtained by applying the k-means algorithm (Lloyd 1982) with $k = 2$ to the random vector Y^j, for each $j = 1, \ldots, M$. Clearly, the distribution of this sequence is not in our model. Nevertheless, Figure 2 shows that maximum likelihood estimation from our model is reasonably robust to this misspecification.

Because of their computational efficiency, distance-based methods are the most competitive with our approach. We compare the performance of our model to a distance-based method based on minimizing the interchange distance between the inferred clustering B and the partition sequence $(\Pi^1, \ldots, \Pi^M)$. Let $d_I(\pi, \pi')$ denote the interchange distance between π and π', as defined at the beginning of Section 6. The *minimum interchange estimate* of B from $(\pi^1, \ldots, \pi^M)$ is

$$\hat{B} = \arg\min_{B \in \mathcal{P}_{[n]}} \sum_{j=1}^{M} d_I(B, \pi^j), \tag{24}$$

which resembles the standard criterion in ordinary least-square regression.

Figure 2 compares performance of our method and the minimum interchange estimate for data simulated from the Gauss–Ewens process, as described above. Further analysis in Section 7.4 suggests that our method is more reliable than the distance-based approach on both simulated and real datasets.

6.4 Comparison for Gene Expression Dataset

As we highlight in Section 1, most clustering methods have been designed for continuous response data. Since our method is not designed for continuous data, comparing it to other prevailing approaches, such as support vector machines and k-nearest neighbor, is not fair. The next example demonstrates that our method, though at a disadvantage, is reasonably competitive with these leading methods.

Using Golub's gene expression dataset from Section 2.3, Yang, Miescke, and McCullagh (2012) compared permanental point process, support vector machine, discriminant analysis, and k-nearest neighbor approaches for a supervised learning task. They first unclassified a random subsample of tissues, and then they reassign classes according to each method. Performance is measured by the average number of wrongly classified tissues. The best case performance among these methods wrongly classifies approximately 0.5 per iteration, which corresponds to an error rate of about 2.08%; see Figure 3 in Yang, Miescke, and McCullagh (2012).

To compare our method to these approaches, we process the gene expression dataset as in Section 6.3 to obtain a sequence of partitions with $k = 2$ classes. We then perform the same supervised learning task using the three-parameter model, with α treated as a nuisance parameter. Whereas the best case performance of the above methods used sequences of length 200, we used only 50 genes and obtained an error rate of about 2.20%.

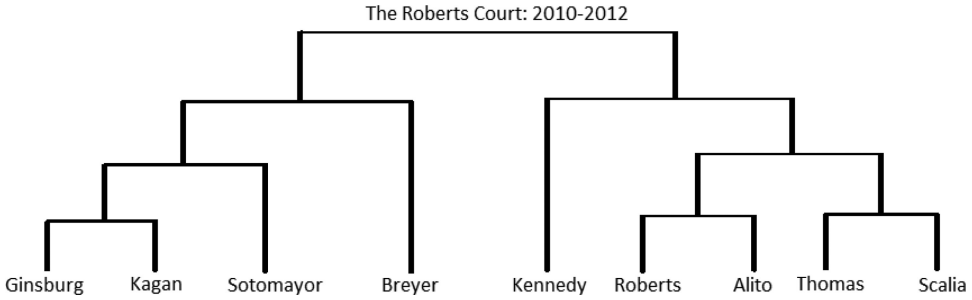

Figure 3. Hierarchical clustering of U.S. Supreme Court for 2010–2012. Edge lengths are not drawn to scale.

7. REAL DATA APPLICATIONS

We fit the three-parameter model to three real datasets. The Supreme Court example in Section 7.1 involves a relatively small sequence of partitions of nine elements for which exact maximum likelihood estimation is possible. The Senate example in Section 7.2 is similar to the Supreme Court data, but for a much larger sample of 100 senators; obtaining an estimate in this case requires randomized search of the state space. In Section 7.3, we cluster species into higher order taxa based on mitochondrial DNA sequence data.

7.1 Supreme Court Alignment, 2010–2012

We analyze voting alignment data for the most recent configuration of the U.S. Supreme Court under Chief Justice John Roberts, which has been seated during judicial terms 2010, 2011, and 2012. We apply the (α, k, B)-model to all U.S. Supreme Court rulings during this period. These data were obtained from the Washington University of St. Louis Supreme Court Database (*http://scdb.wustl.edu/index.php*), which contains voting data for all Supreme Court cases since 1946. For each judicial term, the data resembles Table 1 and has roughly 80 columns, one for each case.

7.1.1 Detecting Factions in the Court. Though Supreme Court justices do not declare political or philosophical affiliations, conventional wisdom segregates the present Court into Liberal, Conservative, and Swing categories, with Justices Breyer, Ginsburg, Kagan, and Sotomayor in the Liberal wing, Justices Alito, Roberts, Scalia, and Thomas in the Conservative wing, and Justice Kennedy as the lone Swing vote. (Kennedy's voting patterns have fluctuated since his appointment to the Court. He replaced Originalist nominee Robert Bork, who was never confirmed (Toobin 2008).) Within these categories, further subdivisions are sometimes noted, particularly among Conservatives, with Justices Roberts and Alito considered Mainstream Conservatives and Justices Scalia and Thomas considered Originalists, the disciples of Robert Bork (1997).

Although cases occur in temporal succession, we assume no temporal dependence in Supreme Court voting. Justices are appointed to lifetime terms, and so their rulings should reflect only their judicial philosophy and the unobserved nuances of each case. Under this assumption, we model the data for each term as an independent, identically distributed sequence of random partitions from the (α, k, B)-model, with $k = 2$ corresponding to the two outcomes for each case, concurrence (C) and dissent (D).

The maximum likelihood estimates for (α, B) from the 2010, 2011, and 2012 terms are remarkably consistent. With α constant across clusters as in (16), we obtain estimates of 0.17, 0.23, and 0.18 for terms 2010, 2011, and 2012, respectively. Also, if we restrict B to the space of partitions with at most two blocks, we estimate the same clustering for each term:

$$\begin{aligned} \{\text{Liberal}\}: &\quad \text{Breyer, Ginsburg, Kagan,} \\ &\quad \text{Sotomayor and} \\ \{\text{Conservative, Swing}\}: &\quad \text{Alito, Roberts, Scalia,} \\ &\quad \text{Thomas, Kennedy.} \end{aligned}$$

This clustering reflects the observation that Justice Kennedy has sided with Conservatives more often than Liberals in recent years. Also, the estimate of $\alpha \approx 0.2$ manifests the fact that justices rarely break rank—many cases are decided 9-0 or 5-4, with Liberals and Conservatives on opposite sides and Justice Kennedy breaking the tie. By fitting the model with cluster-specific mutation rates as in (19), we find little discrepancy between the two groups: $\alpha_{\text{Liberal}} = 0.15, 0.26, 0.10$ and $\alpha_{\text{Conservative, Swing}} = 0.18, 0.21, 0.26$ for 2010, 2011, and 2012, respectively.

If we allow any clustering of the Court, we observe only one change to the estimated partition: our estimate for the 2010 term separates Kagan from the other three Liberal justices. Incidently, 2010 was Kagan's first year on the Court. As such, she was recused from several cases and, quite possibly, had not yet aligned herself strongly with the Liberal wing.

7.1.2 Detecting Subclusters. By restricting our attention to cases for which justices in each cluster were not unanimous, we can investigate further subclustering within each of the factions $\{\text{Liberal}\}$ and $\{\text{Conservative, Swing}\}$. We have already discussed the conventional subpartitioning of Conservatives into Originalist and Mainstream wings; however, we know of no clear distinction among the Liberals. Two Liberals, Kagan and Sotomayor, are new to the Court, and the nuances of their respective philosophies are not yet well understood. Table 5 summarizes estimates obtained by restricting the data to these two clusters.

For the aggregate data from terms 2010–2012, the estimates in Table 5 partition the $\{\text{Conservative, Swing}\}$ block into Swing (Kennedy), Originalist (Scalia, Thomas), and Mainstream Conservative (Roberts, Alito) subblocks. The estimate $\alpha \approx 0.90$ reflects that this wing of the Court is harder to predict than the Court at-large, on the event that it does not vote in unison.

Table 5. Table of joint maximum likelihood estimates of (α, B) for Supreme Court dataset restricted only to the subsets {Conservative, Swing} and {Liberal}, respectively

Conservatives		
Term	$\hat{\alpha}$	$\hat{B}$
2010	0.29	{Kennedy}, {Roberts, Alito} {Scalia}, {Thomas}
2011	0.55	{Kennedy}, {Roberts, Alito}, {Scalia, Thomas}
2012	1.19	{Kennedy}, {Roberts, Alito, Scalia}, {Thomas}
2010–2012	0.90	{Kennedy}, {Roberts, Alito}, {Scalia, Thomas}
Liberals		
Term	$\hat{\alpha}$	$\hat{B}$
2010	0.50	{Sotomayor}, {Ginsburg, Kagan}, {Breyer}
2011	–	{Sotomayor}, {Ginsburg}, {Kagan}, {Breyer}
2012	1.00	{Sotomayor, Ginsburg, Kagan}, {Breyer}
2010–2012	1.55	{Sotomayor}, {Ginsburg, Kagan}, {Breyer}

Liberals are even more fickle. Over 2010–2012, the maximum likelihood estimate for the Liberal wing has three clusters, with Kagan and Ginsburg together in a block and the other two alone as singletons, but the estimated mutation rate of 1.55 suggests that even this alliance is weak. In 2011, the Liberals had no systematic alignment, and in 2012 Breyer was the lone wolf in the Liberal wing. Breyer is the most moderate of the four Liberal justices and the anomalous estimate in 2012 reflects that he was the lone dissent from the Liberal wing in six cases that term.

Figure 3 shows a complete hierarchical clustering of the Court obtained by recursively iterating the three-parameter model in subclusters. This tree coincides with the commonly held ideological hierarchy of the Court: Scalia and Thomas on the right, Kennedy in the center leaning right, Breyer on the left leaning center, and Ginsburg on the far-left (Toobin 2008).

7.2 Voting Alignment in the 107th U.S. Senate

We obtained voting data for the 107th U.S. Senate (*ftp://voteview.com/sen107kh.ord*), which was seated from January 3, 2001, to January 3, 2003, a significant period in recent U.S. history that includes the September 11th terrorist attacks as well as a transfer of power from Democrat Bill Clinton to Republican George W. Bush. With this transfer of power in the White House came a corresponding transfer in the Senate, which was split evenly between 50 Republicans and 50 Democrats for much of the term, with the tie-breaking vote in the hands of the President of the Senate, Republican Vice President Dick Cheney.

On each amendment, each senator votes *yea* or *nay*, or abstains. Senators can abstain from voting due to absence or for political reasons. Though senators are not forced to vote on any given initiative, they are expected to represent their constituency. As a result, few senators abstain from more than a handful of votes in a session, and we treat these missing votes as occurring completely at random.

7.2.1 Clustering in the Senate. Especially given the even split among Republican and Democratic senators, we expect to observe a strong clustering of senators into two voting blocs, with possible defection of Senator Jim Jeffords, who switched parties from Republican to Independent in the middle of the session. The Senate consists of 100 individuals, and so even a search of all clusterings with at most two blocks entails a set with $2^{99} \approx 6.3 \times 10^{29}$ elements. Initializing at the state with all senators in one block, we ran the cut-and-paste/cocktail algorithm (see Section 8) for approximately 7500 steps using R. After 230 min, the Markov chain converged to within interchange distance 3 of the eventual maximum likelihood estimate found by performing an extensive local search and gradient descent in the vicinity of this partition. The maximum likelihood estimate obtained from this search divides the Senate into Republicans and Democrats, but with Independent Jim Jeffords in the Democratic cluster and Democrat Zell Miller in the Republican cluster. The estimate for α is 0.36, signifying a high degree of party loyalty. When we allow party-specific mutation rates, we estimate $\alpha_{\text{Republican}} = 0.37$ and $\alpha_{\text{Democratic}} = 0.35$, indicating that the two parties exhibit identical behavior.

Given what is known about the 107th Senate, this clustering is reasonable. The first run through the algorithm displaced Democrats Diane Feinstein, Tom Carper, and Zell Miller in the Republican cluster, Republican Jesse Helms in the Democratic cluster, and Independent Jim Jeffords in the Democratic cluster. Local search corrected each of these anomalies except for Zell Miller and Jim Jeffords. Prior to our analysis, we were aware of Jeffords's party change, and so his alignment with the Democrats was expected. We were not, however, aware of Zell Miller's close connection with Republicans. Further investigation revealed Miller as an outcast in the Democratic party at this time. A cursory overview of writings and discussions of the 107th Senate revealed no significant differences between our estimated partition and the true alignment of senators. Our inferred clustering seems to have detected meaningful structure.

7.2.2 Classifying Unlabeled Units. A common inference problem entails determining the class membership of newly observed units, given a clustering for an observed sample as well as a response for the new units. In Section 5.2.2, we discussed that supervised learning from the three-parameter model is identical to maximum likelihood inference on a restriction of the parameter space. To test our model's ability for this task, we randomly unlabeled a proportion of senators and attempted to reconstruct the complete classification by maximum likelihood. Our method had no difficulty recovering the full maximum likelihood clustering. Since the algorithm was capable of finding the maximum likelihood clustering with all senators unclassified, it is no surprise that it performed even better when the parameter space is restricted to a smaller set.

7.3 Inferring Higher Order Taxa from Mitochondrial DNA

The University of Montreal maintains a database of 1898 complete mitochondrial DNA (mtDNA) sequences for 1283 different organisms (http://www.bch.umontreal.ca/ogmp/projects/other/mt-list.html). We took sequences from the database for the 15 species summarized in Table 6. We chose these species to ensure significant diversity as well as reader familiarity. Before analysis, we aligned sequences using ClustalOmega (http://www.ebi.ac.uk/Tools/msa/clustalo/), a widely used alignment tool.

Because of the biological process of recombination, the induced partition sequence tends to have the same partition along contiguous portions of the chromosome. Such occurrences are

Table 6. List of organisms studied in Section 7.3. The rightmost column gives the phylum, class, and order of each organism. All species are in the animal kingdom

Genus and species (common name)	Higher order taxa
Xenopeltis unicolor (snake)	Chordata, Reptilia, Squamata
Crocodylus niloticus (African crocodile)	Chordata, Reptilia, Crocodilia
Tinca tinca (doctor fish)	Chordata, Actinopterygii, Cypriniformes
Iguana iguana (iguana)	Chordata, Reptilia, Squamata
Homo sapiens (human)	Chordata, Mammalia, Primates
Canis familiaris (dog)	Chordata, Mammalia, Carnivora
Ursus americanus (black bear)	Chordata, Mammalia, Carnivora
Bos taurus (cow)	Chordata, Mammalia, Artiodactyla
Sus scrofa (wild boar)	Chordata, Mammalia, Artiodactyla
Rhinoceros unicornis (rhinoceros)	Chordata, Mammalia, Perissodactyla
Echinococcus granulosus (tapeworm)	Platyhelminthes, Cestoda, Cyclophyllidea
Taenia asiatica (Asian tapeworm)	Platyhelminthes, Cestoda, Cyclophyllidea
Tigriopus californicus (crustacean)	Arthropoda, Maxillopoda, Harpacticoida
Daphnia pulex (water flea)	Arthropoda, Branchiopoda, Cladocera
Drosophila melanogaster (fruit fly)	Arthropoda, Insecta, Diptera

likely due to a single event of recombination, and so we reduced partitions appearing more than once in succession to a single observation in the data sequence.

For the 15 species in Table 6, we fit the three-parameter model to the resulting de-duplicated mtDNA sequence data. The resulting estimate of B is the partition of species into the phyla {Chordata, Arthropoda} and {Platyhelminthes}, with an estimated mutation rate of 0.23. This estimate was obtained using the stochastic search split-and-merge algorithm, beginning with the initial state of all species in the same cluster. The algorithm found this partition in only 25 iterations, which took about 25 min using R on a laptop computer. To guard against finding only a local maximum, we ran the algorithm several more times, each time initialized in a random starting state from the Ewens–Pitman distribution. We also computed the log-likelihood for other intuitive choices of B and a range of α values; see Figure 4.

7.3.1 Inferring Hierarchical Structure. According to evolutionary theory, the species in Table 6 are related by a phylogenetic tree, with species partitioned into singleton clusters at one end and a most recent common ancestor at the other. The outcome of the previous section suggests that evolution occurred by an initial branching into the classes {Chordata, Arthropoda} and {Platyhelminthes}. Using the three-parameter model, we searched for further hierarchical structure by restricting the mtDNA dataset to the sequences labeled by those species in phyla Chordata and Arthropoda. Using randomized search from both the cut-and-paste algorithm and the split-and-merge algorithm, we obtain a maximum likelihood estimate that partitions these species into {Chordata} and {Arthropoda}, as expected, with a mutation rate of 0.40; see the right panel in Figure 4.

By recursively iterating this procedure in subclusters, we can obtain an estimated phylogenetic tree for these species. Figure 3 shows a similar analysis for the Supreme Court dataset.

7.4 Comparison to Competing Approaches

We compared our estimates in the above examples with those obtained by the minimum interchange algorithm. The two approaches agree for the Supreme Court and Senate datasets, but disagree for the mtDNA dataset. For the mtDNA dataset, our optimal clustering into {Chordata, Arthropoda} and

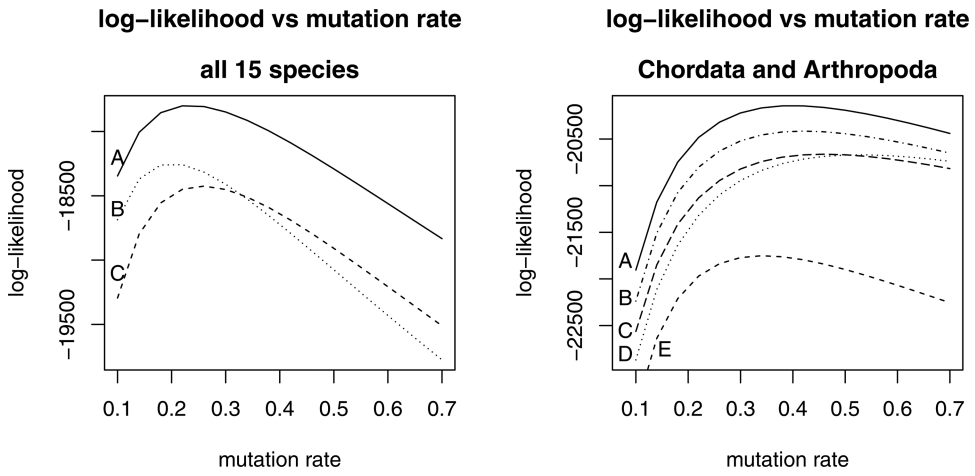

Figure 4. Left: Plot of log-likelihood versus mutation rate for fixed partition parameter on all 15 species: A: {Chordata, Arthropoda}, {Platyhelminthes}; B: {Chordata}, {Arthropoda}, {Platyhelminthes}; C: {Chordata}, {Arthropoda, Platyhelminthes}. The maximum occurs at a mutation rate of around 0.23 for the clustering {Chordata, Arthropoda}, {Platyhelminthes}. Right: Plot of log-likelihood verses mutation rate for fixed partition parameter on 13 species in phyla Chordata and Arthropoda: A: {Chordata}, {Arthropoda}; B: {Chordata, *T. calif.*}, {*D. pulex*, *D. melan.*}; C: {Chordata, *D. pulex*}, {*T. calif.*, *D. melan.*}; D: {Chordata, Arthropoda}; E: {Chordata-nonmammal}, {Chordata-mammal}, {Arthropoda}. The maximum occurs around 0.40 for {Chordata}, {Arthropoda}.

{Platyhelminthes} is among the several clusterings with lowest interchange score, but the minimum interchange clustering separates species 13 (copepod) as a singleton apart from the rest, which has no biological justification. This observation, paired with the outperformance on the simulated data in Figure 2, suggests that our model is more reliable for general clustering applications. In particular, the simulated data and mtDNA datasets have more noise than the Supreme Court and Senate datasets. Since our approach models noise through the mutation rate, it is better equipped. These observations underscore the risk of algorithmic methods: there is often no good way to diagnose their performance on novel datasets. On the other hand, our model endows a straightforward interpretation to the inferred parameters, which affords the statistician additional discretion in new applications.

8. DISCUSSION

We have proposed a parametric model for clustering from categorical data sequences. Under Assumption 1, the data reduce to a sequence of partitions of the sample, from which we infer an overall clustering by maximum likelihood estimation. The model has three parameters that are easy to interpret in terms of the induced properties of the partition sequence, see Section 5.3. By allowing for cluster-specific mutation rates, the model is equipped for situations in which intracluster rates differ substantially. For large sample sizes, we use a Markov chain Monte Carlo search method that is provably efficient.

For situations where Assumption 1 fails, we can modify our model as follows. For each column in the data array (2), we define *labeled classes* $C_l^j = \{i : y_i^j = l\}$, the set of units with response $l = 1, \ldots, k$ in column j, for each $j = 1, 2, \ldots$ These classes then determine a *labeled partition* $\tilde{\pi}^j = (C_1^j, \ldots, C_k^j)$, where subsets can now be empty. With B representing the overall clustering and $\alpha_1, \ldots, \alpha_k > 0$ denoting class-specific mutation rates, the distribution in (16) generalizes to

$$P_n^{\alpha_1,\ldots,\alpha_k,k,B}(\tilde{\pi}) = \prod_{b\in B} \frac{\prod_{j=1}^k \alpha_j^{\uparrow \#(b\cap C_j)}}{(\alpha_1+\cdots+\alpha_k)^{\uparrow \#b}}, \tag{25}$$

where $\tilde{\pi} = (C_1, \ldots, C_k)$ is now a k-tuple of labeled classes. This model retains the key properties of the (α, k, B)-model, particularly label equivariance and sampling consistency.

In some applications, for example, bioinformatics (Ewens and Grant 2005), the columns in (2) are dependent. In this case, we can extend our model to a family of finite-order reversible Markov chains, called the $(\alpha, k, B; r)$*-three-parameter Markovian cluster model*, where $r \geq 1$ is the order of the Markov chain. For example, with $r = 1$, the transition probability from π to π' is

$$p_n^{\alpha,k,B_{[n]}}(\pi,\pi') = k^{\downarrow \#\pi'} \prod_{b^*\in B_{[n]}} \prod_{b\in\pi} \frac{\prod_{b'\in\pi'} (\alpha/k)^{\uparrow \#(b^*\cap b\cap b')}}{\alpha^{\uparrow \#(b^*\cap b)}}, \tag{26}$$

which is reversible with respect to the (α, k, B)-three-parameter distribution in (16). The Markov model retains the label equivariance, sampling consistency, and reversibility properties of the iid model. In general, although our model easily extends to (25) and (26), there are practical issues regarding missing observations and parameter inference that require further investigation. In the applications of Section 7, these generalizations offer no new insights.

APPENDIX A: RANDOMIZED SEARCH ALGORITHMS

When the sample size is moderate to large, the space of partitions is too big to search exhaustively during maximum likelihood estimation. In this case, we use a randomized search algorithm for approximate inference. Our search algorithms iterate between local- and global-move Markov chains on the space of partitions. When the number of clusters in B is unspecified, the global-move chain evolves by split-and-merge (Pitman 2002); when the number of clusters is known and finite ($\tilde{k} < \infty$), the global-move chain evolves by cut-and-paste (Crane 2014).

For $(\tilde{\alpha}, \tilde{\theta})$ satisfying either

(i) $\tilde{\alpha} < 0$ and $\tilde{\theta} = -\tilde{k}\tilde{\alpha}$, for some $\tilde{k} = 1, 2, \ldots$, or
(ii) $0 \leq \tilde{\alpha} \leq 1$ and $\tilde{\theta} > \tilde{\alpha}$,

the Ewens–Pitman$(\tilde{\alpha}, \tilde{\theta})$-distribution on partitions of $[n]$ is

$$P_n^{\tilde{\alpha},\tilde{\theta}}(\pi) = \frac{(\tilde{\theta}/\tilde{\alpha})^{\uparrow \#\pi}}{\tilde{\theta}^{\uparrow n}} \prod_{b\in\pi} -(-\tilde{\alpha})^{\uparrow \#b}, \quad \pi \in \mathcal{P}_{[n]}. \tag{A.1}$$

Under (i), the Ewens–Pitman family is restricted to partitions with at most $\tilde{k}$ blocks. Under (ii), the Ewens–Pitman family is supported by all partitions. In our randomized search algorithms, we choose $(\tilde{\alpha}, \tilde{\theta})$ to satisfy (i) or (ii), depending on the parameter space of B in our model.

A.1 Global Search: Cut-and-Paste Algorithm

For $(\tilde{\alpha}, \tilde{\theta})$ satisfying (i), the Ewens cut-and-paste chain with parameter $(\tilde{\alpha}, \tilde{k})$ evolves on partitions of $[n]$ with at most $\tilde{k}$ blocks. Let $\pi = \{b_1, \ldots, b_r\}$, $r = 1, \ldots, \tilde{k}$, be the current state.

(a) Independently, each block b_i is partitioned into $\tilde{\pi}^i$ according to the Ewens–Pitman law with parameter $(-\tilde{\alpha}/\tilde{k}, \tilde{\alpha})$.
(b) For each $i = 1, \ldots, r$, the blocks of $\tilde{\pi}^i$ are labeled uniformly without replacement in $\{1, \ldots, \tilde{k}\}$.
(c) The next state π' is obtained by aggregating blocks in (b) with the same label and then removing the labels.

The most attractive feature of the cut-and-paste chain is that it assigns strictly positive probability to any transition in the search space. Furthermore, it converges to its stationary distribution in $O(\log n)$ steps (Crane and Lalley 2012), where n is the sample size; hence, this chain searches the parameter space exponentially quickly with respect to sample size. Importantly, the parameters $(\tilde{\alpha}, \tilde{k})$ have no logical relationship to the parameters (α, k, B) in the three-parameter model.

A.2 Global Search: Split-and-Merge Algorithm

For $(\tilde{\alpha}, \tilde{\theta})$ satisfying (ii) and $p \in (0, 1)$, the split-and-merge algorithm evolves on the unrestricted space of partitions by drawing $U \sim \text{Uniform}(0, 1)$ at each step. If $U < p$, then the operation *split* is performed, whereby a block of the current partition π is chosen randomly with probability proportional to its size and is partitioned according to the Ewens–Pitman$(\tilde{\alpha}, \tilde{\theta})$ distribution. If $U \geq p$, then two blocks of π are drawn uniformly without replacement and are *merged* into a single block.

A.3 Local Search: Cocktail Algorithm

For $(\tilde{\alpha}, \tilde{\theta})$ in the parameter space of the Ewens–Pitman family, the cocktail algorithm evolves on partitions by updating one element at a

time. Let $\pi \in \mathcal{P}_{[n]}$ be the current state of the chain. First, an element $u \in [n]$ is sampled uniformly at random and removed from π to obtain $\pi_{[n]\setminus u}$. Given $\pi_{[n]\setminus u}$, the removed element u is reinserted into $\pi_{[n]\setminus u}$ according to the seating rule of the $(\tilde{\alpha}, \tilde{\theta})$-Chinese restaurant process:

$$\mathrm{pr}(u \mapsto b \mid \pi_{[n]\setminus u}) \propto \begin{cases} \#b - \tilde{\alpha}, & b \in \pi_{[n]\setminus u} \\ \tilde{\theta} + \tilde{\alpha}\#\pi_{[n]\setminus u}, & b = \emptyset. \end{cases}$$

When $(\tilde{\alpha}, \tilde{\theta})$ satisfies (i), the cocktail chain is restricted to partitions with at most $\tilde{k}$ blocks; otherwise, it is unrestricted.

By iterating between the local and global chains, our search algorithm explores the partition space for local and global maxima. To effectively use this algorithm, we take a step in the global chain followed by a prespecified number of moves in the cocktail algorithm. We accept all moves in the global chain, and we accept moves in the cocktail chain according to the Metropolis–Hastings algorithm. This choice reflects our observation that local maxima often occur only a few steps away from partitions with low likelihood, and so rejecting global moves can be counter-productive to search.

APPENDIX B: PROOF OF THEOREM 1

To establish label equivariance for $P^{\alpha,k,B}$, we observe that the distribution of $\Pi_n \sim P_n^{\alpha,k,B_{[n]}}$ depends only on the sizes of the subsets $\{b \cap b' : b \in B_{[n]},\ b' \in \Pi_n\}$ and the block sizes of $B_{[n]}$ and Π_n, all of which are invariant under the joint transformation $\Pi_n \mapsto \Pi_n^\sigma$ and $(\alpha, k, B) \mapsto (\alpha, k, B^\sigma)$. Label equivariance follows readily from (20).

Sampling consistency and lack of interference are less obvious. Fix $\alpha > 0, k = 1, 2, \ldots$, and B a partition of $\mathbb{N}$. To verify (22) for the three-parameter model, we first observe that $B_{[n]}$ and $B_{[n+1]}$ coincide except for the unique block $b_* \in B_{[n]}$ into which $n+1$ is inserted to obtain $B_{[n+1]}$, with $b_* = \emptyset$ when $\{n+1\}$ is appended to $B_{[n]}$ as a singleton. Furthermore, any $\pi' \in \mathcal{P}_{[n+1]}$ for which $\pi'_{[n]} = \pi$ must coincide with π except in the block $b'_* \in \pi$ into which $n+1$ is inserted. Properties of the rising factorial imply

$$\prod_{b \in B_{[n+1]}} \prod_{b' \in \pi'} \alpha^{\uparrow \#(b \cap b')} = (\alpha + \#(b_* \cap b'_*)) \prod_{b \in B_{[n]}} \prod_{b' \in \pi} \alpha^{\uparrow \#(b \cap b')} \quad \text{and}$$

$$\prod_{b \in B_{[n+1]}} (k\alpha)^{\uparrow \#b} = (k\alpha + \#b_*) \prod_{b \in B_{[n]}} (k\alpha)^{\uparrow \#b},$$

from which (22) follows:

$$\sum_{\pi' \in D_{n,n+1}^{-1}(\pi)} P_{n+1}^{\alpha,k,B_{[n+1]}}(\pi') = P_n^{\alpha,k,B_{[n]}}(\pi)$$

$$\left(\frac{k - \#\pi}{k\alpha_{b_*} + \#b_*} \alpha_{b_*} + \sum_{b'_* \in \pi} \frac{\alpha_{b_*} + \#(b_* \cap b'_*)}{k\alpha_{b_*} + \#b_*} \right) = P_n^{\alpha,k,B_{[n]}}(\pi).$$

[Received June 2014. Revised September 2014.]

REFERENCES

Banfield, J. D., and Raftery, A. E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–821. [1]

Barry, D., and Hartigan, J. A. (1992), "Product Partition Models for Change Point Problems," *Annals of Statistics*, 20, 260–279. [1]

Binder, D. A. (1978), "Bayesian Cluster Analysis," *Biometrika*, 65, 31–38. [1]

Bork, R. H. (1997), *The Tempting of America*, New York: Free Press. [10]

Campbell, N. A., Reece, J. B., and Mitchell, L. G. (1999), *Biology* (5th ed.), Menlo Park, CA: Benjamin-Cummings. [2]

Crane, H. (2013), "Some Algebraic Identities for the α-Permanent," *Linear Algebra and Its Applications*, 439, 3445–3459. [5]

Crane, H. (2014), "The Cut-and-Paste Process," *Annals of Probability*, 42, 1952–1979. [13]

Crane, H., and Lalley, S. P. (2013), "Convergence Rates of Markov Chains on Spaces of Partitions," *Electronic Journal of Probability*, 18, 1–23. [13]

Crowley, E. M. (1997), "Product Partition Models for Normal Means," *Journal of the American Statistical Association*, 92, 192–198. [1]

Efron, B., and Thisted, R. (1976), "Estimating the Number of Unseen Species: How Many Words did Shakespeare Know?," *Biometrika*, 63, 435–447. [1]

——— (1987), "Did Shakespeare Write a Newly Discovered Poem?," *Biometrika*, 74, 445–455. [1]

Ewens, W. J. (1972), "The Sampling Theory of Selectively Neutral Alleles," *Theoretical Population Biology*, 3, 87–112. [1,5]

Ewens, W. J., and Grant, G. (2005), *Statistical Methods in Bioinformatics. Statistics for Biology and Health*, New York: Springer. [13]

Felsenstein, J. (2004), *Inferring Phylogenies*, Sunderland, MA: Sinauer Associates, Inc. [1]

Ferguson, T. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230. [5]

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caliguri, M. A., Bloomfield, C. D., and Lander, E. S. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531–537. [2]

Hartigan, J. A. (1990), "Partition Models," *Communications in Statistics—Theory and Methods*, 19, 2745–2756. [1]

Hjort, N. L. (1990), "Nonparametric Bayes Estimators Based on Beta Processes in Models for Life History Data," *Annals of Statistics*, 18, 1259–1294. [1]

Irons, P. (2006), *A People's History of the Supreme Court: The Men and Women Whose Cases and Decisions Have Shaped Our Constitution*, New York: Penguin Books. [3]

Kerov, S. (2005), "Coherent Random Allocations, and the Ewens–Pitman Formula," *Zapiski Nauchnykh Seminarov POMI*, 325, 127–145. [5]

Kingman, J. F. C. (1978), "Random Partitions in Population Genetics," *Proceedings of the Royal Society of London,* Series A, 361, 1–20. [1]

Lau, J. W., and Green, P. J. (2007), "Bayesian Model-Based Clustering Procedures," *Journal of Computational and Graphical Statistics*, 16, 526–558. [1]

Lloyd, S. P. (1982), "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, 28, 129–137. [1,9]

Lord, F. M. (1980), *Applications of Item Response Theory to Practical Testing Problems*, Mahwah, NJ: Erlbaum. [1]

McCullagh, P., and Yang, J. (2008), "How Many Clusters?," *Bayesian Analysis*, 3, 101–120. [1,9]

Park, J. H., and Dunson, D. B. (2010), "Bayesian Generalized Product Partition Model," *Statistica Sinica*, 20, 1203–1226. [5]

Pitman, J. (2006), *Combinatorial Stochastic Processes, Lecture Notes in Mathematics* (Vol. 1875), Berlin: Springer-Verlag. [4,5]

——— (2002), "Poisson-Dirichlet and GEM Invariant Distributions for Split-and-Merge Transformation of an Interval Partition," *Combinatorics, Probability and Computing*, 11, 501–514. [13]

Quintana, F. A., and Iglesias, P. L. (2003), "Bayesian Clustering and Product Partition Models," *Journal of the Royal Statistical Society,* Series B, 65, 557–574. [4]

Sirovich, L. (2003), "A Pattern Analysis of the Second Rehnquist U.S. Supreme Court," *PNAS*, 100, 7432–7473. [1,6]

Sokal, R. R., and Sneath, P. H. A. (1963), *Numerical Taxonomy*, San Francisco: W.H. Freeman. [8]

Stanley, R. (2012), *Enumerative Combinatorics (Cambridge Studies in Advanced Mathematics)* (Vol. 1, 2nd ed.), New York: Cambridge University Press. [7]

Tavaré, S. (2001), *Ancestral Inference in Population Genetics, Lecture Notes in Mathematics* (Vol. 1837), Berlin: Springer-Verlag. [7]

Thurstone, L. L., and Degan, J. W. (1951), "Factorial Study of the Supreme Court," *Proceedings of the National Academy of Sciences*, 37, 628–635. [1,6]

Toobin, J. (2008), *The Nine: Inside the Secret World of the Supreme Court*, New York: Anchor. [10,11]

Yang, J., Miescke, K., and McCullagh, P. (2012), "Classification Based on a Permanental Process With Cyclic Approximation," *Biometrika*, 99, 775–786. [1,9]