Bayesian Classification

Lawrence J. Brunner[1]
Department of Statistics
The University of Toronto
Toronto, Canada

Albert Y. Lo[2]
Department of Information
and Systems Management
The University of Science and Technology
Clear Water Bay, Hong Kong

Research Report
Department of Statistics
University of Toronto

Version 12: Last modified Sept. 6, 1999

[1]Supported in part by Natural Sciences and Engineering Research Council of Canada Grant OGPIN 014

[2]Supported in part by Hong Kong Research Grants Council Competitive Earmarked Research Grant 674/95.

Abstract. The objective of a classification procedure is to identify subsets of observations that arise from the same probability distribution. Examples include cluster analysis, change point detection, discriminant analysis, outlier detection, and multiple comparisons. A distinction is made between random partition (mixture model) classification and fixed partition classification, and we show that for a Bayesian approach with natural models and priors, the former is a special case of the latter. We also show that in any Bayesian mixture problem where a Dirichlet prior is assumed for the unknown mixing distribution, the posterior distribution unavoidably depends upon the kind of classification analysis described here. For both fixed–partition and random partition classification problems, numerical difficulties are overcome by a Gibbs sampling method. The techniques are illustrated with data from a study of songbird diets.

1. Introduction.

In a classification problem, data obtained from each of *n* objects are used as a basis for deciding which of the objects are somehow similar. An "object" may be a person, an ant colony, a species, an experimental treatment, or any other entity capable of generating data. An object may potentially contribute more than one observation, and those observations may or may not be assumed independent. Our approach will be Bayesian. Therefore it will necessarily be based on a well–defined probability model for the data — unlike, for example, Draper and Smith's (1981) discussion of outlier detection, and many but not all of the versions of cluster analysis in the collection edited by Arabie, Hubert and De Soete (1996). Indeed, our view of the classification problem is that the objects we are trying to classify may have generated data from different probability distributions, and that our task is to group them into subsets so that objects are assigned to the same subset if and only if they have generated data from the same distribution.

Let X_i denote the observation (possibly vector valued) obtained from object i, i = 1, ..., n. We shall begin by studying the following model; then we will extend it. Given the vector $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$, let the observations $\mathbf{X} = (X_1, ..., X_n)$ be independent with respective densities $k(x_i|\phi_i)$, i = 1, ..., n. The functional form of k is known except for the value of ϕ_i. Our goal is Bayesian inference about a *partition* of the integers {1, ..., n}, consisting of a collection of subsets such that $\phi_i=\phi_j$ if and only if *i* and *j* belong to the same subset. That is, observations

with indices in the same subset come from the same distribution. This framework is general enough so that either directly or with straightforward extensions, it encompasses various Bayesian approaches to cluster analysis (Hartigan 1975 p. 113; Shapiro 1977; Binder 1978; Menzefricke 1981; Symons 1981; Titterington, Smith and Makov 1985; Lavine and West 1992; Grenander 1993 Ch. 14; Bernardo 1994; Bock 1996; Liu 1996), discriminant analysis (Geisser 1964, 1966, 1982; Bernardo 1988; Lavine and West 1992; Johnson and Kokolakis 1994; Johnson and Mouhab 1996), outlier detection (Box and Tiao 1968; Guttman 1973; Guttman, Dutter and Freeman 1978), and structural change in linear models and time series (Broemeling 1985, Ch. 7).

To give a bit more detail for a non−standard example, we remark that Menzefricke's (1981) Bayesian clustering of data sets is obtained by allowing the "observations" X_i to be vectors of i.i.d. random variables (so that the density $k(x_i|\phi_i)$ above is a product of marginals), one vector for each treatment. In this way, the problem of multiple comparisons emerges as one of classification. The "objects" are experimental treatments, and the statistician seeks to classify them into subsets that generate data from distributions with, for example, the same mean. In fact, any statistical procedure for deciding whether there is a relationship between some dependent variable and a categorical independent variable can be viewed a a type of classification analysis, for the question is really whether the values of the independent variable are generating data from a single probability distribution.

We also think it worthwhile to point out how a basic version of discriminant analysis fits our framework with no special modification. Here, the number of "groups" or subsets is known. There is a "training sample" of observations for which group membership is known, and an additional sample of observations for which group membership is unknown; these are to be classified. To obtain posterior classification probabilities using the methodology of the present paper, we simply assign prior probability zero to partitions with the wrong number of subsets, and also to partitions in which any two members of the training sample who belong to the same group appear in different subsets.

There would appear to be two major types of classification problem (Bock, 1996). *Fixed partition* classification makes no assumption about why the elements of ϕ might happen to be different; ϕ is

simply an unknown parameter. *Random partition* or mixture model classification assumes a two–stage sampling procedure in which $\phi_1, ..., \phi_n$ are first sampled independently from some unknown mixing distribution G, and then the X_i are independently sampled from $k(x_i|\phi_i)$, $i = 1, ..., n$. In the random partition model, the parameter is the probability distribution G, and $\boldsymbol{\phi}$ is a latent variable.

Random partition models are appealing in cases where the objects being classified have been sampled from some population, but the methods they require can be complicated. Not only is there an unobservable latent vector, but the parameter G is itself a probability distribution. In the absence of special restrictions (like assuming G to be supported on a finite number of points) the problem is non–parametric, and the natural candidate for a prior distribution on G will be some version of Ferguson's (1973) Dirichlet process prior, which is a stochastic process.

In this paper we show that for a Bayesian, random partition classification reduces to fixed partition classification with a special choice of prior. This means that although there is an important conceptual difference between fixed–partition and random–partition classification, the practical details of performing an analysis are the same –– except that for random–partition analysis, the prior distribution must have a particular form. Thus we are able to restrict our attention to the fixed partition version throughout most of the paper, and confine the discussion of stochastic processes to a single section. There, we will show how random partition classification is a special case of fixed partition classification, and also reveal a deep relationship between classification analysis and Bayesian inference for mixture models –– not just mixture models for classification analysis, but mixture models in general. We show that for any mixture problem where a Dirichlet prior is adopted for the unknown mixing distribution, the posterior distribution is based on a fixed partition classification analysis with a special kind of prior.

Here is the plan of the paper. In Section 2 we begin by describing a natural fixed–partition classification problem, adopt a fairly general prior, and compute the posterior distribution. Then we extend the model to allow for a covariate, and to let us deal in a natural way with change–point problems. We observe that consistency is obtained when each object being classified contributes a sequence of independent observations. This leads us to the view that in most classification problems, one should consider the sample size to be not the number of objects being classified, but the number of data values contributed by each object. From this perspective, most of the *ad hoc*

methods for cluster analysis are valiant but largely futile attempts to solve an n=1 problem.

Section 3 is concerned the problem of numerically evaluating posterior quantities for a fixed–partition classification analysis; this includes the mixture model version as a special case. Numerical evaluation truly is a problem, because the number of partitions of {1, ..., n}, grows so rapidly with the sample size *n* that even listing all of them is a practical impossibility for moderate sample sizes (see for example the account by Comtet 1974, Ch. 5). Our primary solution is to approximate posterior probabilities by means of a Gibbs sampler (Geman and Geman 1984, Gelfand and Smith 1990). That is, we construct a Markov chain whose limiting stationary distribution is the posterior distribution on partitions, and then use the strong law of large numbers for Markov chains (Chung 1967) to approximate any posterior quantity of interest. Since the Markov chain has a finite state space (the space of all partitions), checking conditions of convergence is easy, and in addition we benefit from a geometric rate of convergence and a central limit theorem.

The general idea of a Gibbs sampler on partitions comes from MacEachern (1994). He describes a related method for use in Bayesian density estimation, using a model that is a location mixture of normals. The most important difference between our procedure and MacEachern's, apart from technical details, is that ours applies to fixed–partition as well as to mixture model classification (strictly speaking, MacEachern's method applies to density estimation rather than classification analysis, but its use for the latter purpose is immediate). So, although other Markov chain Monte Carlo methods have been developed for mixture model classification analysis (Lavine and West's 1992 Gibbs sampler and Liu's 1996 sequential imputation scheme as well as MacEachern's Gibbs sampler), we believe that ours is the first that allows actual computation of *fixed–partition* Bayesian classification with no restrictions on the possible partitions or the number of objects being classified.

Section 4 is an application of the methods in this paper to clustering songbirds (using Gibbs sampling) and songbird species (exact calculation) according to the contents of their gizzards. In Section 5, we outline further applications and extensions of the method. In Section 6, we discuss the connection of the classification problem to Bayesian mixture models in general. We begin by

considering a mixture model classification analysis, in which a Dirichlet process prior (Ferguson 1973) has been adopted for the unknown mixing distribution. When we derive the posterior on partitions for this model, we see that it is a special case of the posterior for fixed–partition classification analysis. Thus we learn that for a Bayesian, mixture model classification analysis is equivalent to a fixed–partition classification analysis with a special form of prior distribution –– and that the form of that prior distribution is determined by the parameters of the Dirichlet process prior.

We then demonstrate a deep relationship between classification analysis and Bayesian inference for mixture models –– not just mixture models for classification analysis, but mixture models in general. We show that for any mixture problem where a Dirichlet prior is adopted for the unknown mixing distribution, the posterior distribution is based on a classification analysis that can be viewed as a fixed–partition analysis with a special choice of prior distribution. To give an idea of the range of applications involved, mixture models with Dirichlet priors have yielded non–parametric Bayes procedures for estimation of an arbitrary density (Escobar 1988, Escobar and West 1995, Ferguson 1983, Lo 1978 and 1984, MacEachern 1994, West, Escobar and Müller 1994) and inference about monotone hazard rates (Dykstra and Laud 1981, Lo and Weng 1989), as well as semi–parametric methods for decreasing densities (Brunner 1994, Brunner and Lo 1989), symmetric unimodal densities on the line (Brunner and Lo 1989), elliptically symmetric densities in $\mathbb{R}^k$ (Brunner 1989), rotationally symmetric densities on the sphere (Brunner and Lo 1994), and densities on the line that are unimodal but not necessarily symmetric (Brunner 1992). Our results show that all of these non–parametric methods depend, at least implicitly, on a garden variety parametric classification analysis.

2.1 Fixed-partition classification.

A classification of *n* objects corresponds to a partition of the set of integers $\{1, ..., n\}$. Such a partition P consists of *m* mutually exclusive and exhaustive subsets or "cells;" they will be denoted $C_1, ..., C_m$. The number of integers in cell C_i is n_i, $i = 1, ..., m$. When the dependence of m, C_i and n_i upon P need to be indicated explicitly, they will be written as m(P), $C_i(P)$ and $n_i(P)$ respectively.

Model 1. We will begin with the following model. Conditionally on value of $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$, the observations $\mathbf{X} = (X_1, ..., X_n)$ arise independently from distributions with densities $k(\cdot|\phi_i)$, $i = 1, ..., n$. The form of k is known exactly except for the parameter value ϕ_i; both X_i and ϕ_i can be vectors. Some of the observations may come from the same distribution. Let P be the partition of indices such that $\phi_i=\phi_j$ for $i\neq j$ if and only if *i* and *j* belong to the same cell. Denoting by θ_i the common parameter value of observations with indices in cell C_i and letting $\boldsymbol{\theta}_P = (\theta_1, ..., \theta_{m(P)})$ we have the model density

$$f(\mathbf{x}|P,\boldsymbol{\theta}_P) = \prod_{i=1}^{m(P)} \prod_{\ell \in C_i} k(x_\ell|\theta_i) \tag{2.1.1}$$

We construct a prior on the pair $(P,\boldsymbol{\theta})$ by first selecting a prior distribution $\pi(P)$, a completely general discrete distribution on the space of partitions. Conditionally on P, we adopt a fairly general prior on $\boldsymbol{\theta}_P$. That is, $d\pi(P,\boldsymbol{\theta}_P) = d\pi(\boldsymbol{\theta}_P|P)\pi(P) = d\pi(\theta_1, ..., \theta_{m(P)}|P)\ \pi(P)$. Note that the notation π is being used here for any prior or conditional prior distribution or density, with the arguments of the function providing information about what parameters are involved.

It is convenient to choose the conditional prior on $(\theta_1, ..., \theta_m)$ so that its components are independent; for maximum flexibility, the form of this prior may depend on the partition. That is,

$$d\pi(P,\boldsymbol{\theta}_P) = \pi(P)\ d\pi(\boldsymbol{\theta}_P)|P)=\pi(P)\ \prod_{i=1}^{m(P)}\ d\pi_i(\theta_i|P). \tag{2.1.2}$$

Our primary interest is in the marginal posterior distribution on partitions –– that is, on subset membership. It is given in Theorem 1a below; the proof is a simple calculation and appears in the Appendix.

Theorem 1a. For the model (2.1.1) and prior (2.1.2), the posterior distribution is specified by

$$\Pr(P|\mathbf{X}) \propto \pi(P)\ \prod_{i=1}^{m(P)} \int \prod_{\ell \in C_{i,P}} k(x_\ell|\theta)\ d\pi_{i,P}(\theta|P), \tag{2.1.3}$$

and given $(P,\mathbf{X})$, $\theta_1, ...,\theta_{m(P)}$ are independent with respective distributions

$$d\pi_i(\theta_i|\mathbf{X},P) \propto \prod_{\ell \in C_{i,P}} k(x_\ell|\theta_i)\ d\pi_{i,P}(\theta_i|P), \quad i=1,...,m(P)$$

This formulation is different from others such as Binder's (1978) not just in notation, but because both prior and posterior independence of the components of $\boldsymbol{\theta}_P$ appear explicitly. Prior and posterior independence seems to imply that independence between cells is the natural property, suggesting a model where independence holds only between cells. For example, there could be a time−series or spatial autocorrelation structure within cells, and one might choose to cluster only temporally or spatially adjacent observations. In this way, we obtain change point detection in time series as a type of classification analysis.

Model 2. Denote by $h(\mathbf{x}|\theta_i,C_{i,P})$ the joint density of observations with indices in cell *i* of partition P. The subset of indices $C_{i,P}$ serves to select the observations of which the density is a function, and may also select a vector of covariates that applies to those observations. In the i.i.d. case (Model 1), we would have $h(\mathbf{x}|\theta_i,C_{i,P}) = \prod_{\ell \in C_{i,P}} k(x_\ell|\theta_i)$. The model density is now

$$f(\mathbf{x}|P,\boldsymbol{\theta}_P) = \prod_{i=1}^{m(P)} h(\mathbf{x}|\theta_i,C_{i,P}). \tag{2.1.4}$$

Exactly the same calculations leading to Theorem 1a now yield

Theorem 1b. For the model (2.1.4) and prior (2.1.2), the posterior distribution is given by

$$\Pr(P|\mathbf{X}) \propto \pi(P) \prod_{i=1}^{m(P)} \int h(\mathbf{x}|\theta,C_{i,P})\, d\pi_{i,P}(\theta|P), \tag{2.1.5}$$

and given $(P,\mathbf{X})$, $\theta_1, ...,\theta_{m(P)}$ are independent with respective distributions

$$d\pi_i(\theta_i|\mathbf{X},P) \propto h(\mathbf{x}|\theta_i,C_{i,P})\, d\pi_{i,P}(\theta_i|P), \quad i=1,...,m(P)$$

Notice that the marginal posterior on partitions is just the prior, multiplied by an expression that looks like the product of marginal distributions of the data with the parameter integrated out —— marginal distributions in which a separate Bayesian analysis with model density $h(\cdot|\theta_i,C_{i,P})$ and prior $d\pi_{i,P}$ is being carried out on the data with indices in cell *i* of the partition, i=1,...,m(P). We will call these expressions *component marginals*, and denote the component marginal distribution by $dF(\mathbf{x}|C_{i,P})$, the *component posterior* distribution distribution by $dF(\theta_i|\mathbf{X},C_{i,P})$ and the *component predictive* density by $f(x|\mathbf{X},C_{i,P})$.

2.2 Consistency

Each observation X_j may be a vector of length *N*, so that $h(\cdot|\theta_i,C_{i,P})$ is the joint density of a matrix with *N* rows and $n_{i,P}$ columns. In particular, the rows may be independent, so that each of the component data sets consists of *N* i.i.d. observations of dimension $n_{i,P}$, and $h(\cdot|\theta_i,C_{i,P})$ is a product of marginal densities for each i = 1, ..., m(P). In this case, a result of Doob (1949) ensures that the marginal posterior distribution on partitions will become concentrated on the true partition with probability one as *N* (not *n*) tends to infinity, for almost all parameter values. The discreteness of P means that "almost" all partitions means all partitions.

If we do not have *N* i.i.d. copies of the experiment, but instead some other model admitting of a consistent estimate of the true partition, then the Bayes procedure is also consistent. But it is important to repeat that this consistency is not obtained by letting the number of objects *n* become large. Rather, it is a large amount of information from *each object* that allows accurate determination of the true classification structure. In the absence of drastic *a priori* restrictions on the set of partitions with positive prior probability, increasing the number of objects results only in a geometric expansion of the parameter space, and posterior opinion about the classification of individual objects remains diffuse. Thus for example, ordinary cluster analysis (Bayesian or not) with no restriction on the number of clusters is properly viewed as an N=1 problem.

3. A Gibbs sampler for Bayesian classification

For small numbers of objects, the number of partitions is manageable. Detailed, substantively based priors on the space of partitions are possible, and the normalizing constant of (2.1.5) may be obtained explicitly (with more or less effort, depending on the component model and prior). But there are 115,975 partitions of 10 objects, and around 1.38 billion partitions of 15 objects (see Comtet 1974 for more detail). To resolve this difficulty we construct a Gibbs sampler (Geman and Geman 1984, Gelfand and Smith 1990) that may be used to obtain Markov chain Monte Carlo

approximations of posterior probabilities. The general idea for this method (not the details) comes from MacEachern (1994), who suggested a Gibbs sampler on partitions for Bayesian kernel density estimation. Like MacEachern, we define a Markov chain whose state space is the set of partitions, and whose stationary distribution is the posterior. We benefit from an exponential rate of convergence as well as a Strong Law of Large Numbers and a Central Limit Theorem (Chung 1967).

The procedure to be described here differs from MacEachern's in four important respects. First, MacEachern's n observations are i.i.d., while ours need only be independent between clusters. Second, our approach applies to a general component model and prior, and is not limited to the normal case. Third and perhaps most importantly, our method applies to fixed partition classification (which includes random partition classification as a special case), while MacEachern is solving a mixture model problem; we believe his solution is limited to random partition classification. A fourth apparent difference is that when we do specify our method to mixture models, the formulas are much simpler. This may be because while we use a Dirichlet process prior on the mixing distribution, MacEachern assumes a mixture of Dirichlet processes. We also mention that Lavine and West (1992) describe a Gibbs sampler for classification, but theirs is also the solution to a mixture model problem, the number of categories is assumed known in advance, and it is based on a Markov process with continuous state space; this makes convergence of the Gibbs sampler much harder to check in particular cases.

Our Gibbs sampler proceeds as follows. Start with an initial partition, to be denoted P_0. Remove the integer 1, and then put it back, possibly where it was before and possibly in another location, according to a well–chosen conditional probability distribution. By "location," we mean either into one of the existing subsets of the partition, or into a subset –– possibly a new subset –– by itself. With the resulting partition, perform the same kind of removal and replacement operation on the integer 2. Continue, successively removing and replacing the integers 3, ..., n. This n–step procedure will be called a *Gibbs cycle*. The partition at the end of one cycle will be denoted P_1, and continuing for M cycles yields P_1, ..., P_M. The conditional probabilities governing replacement of integers will be defined so that P_1, ... is the realization of a Markov chain whose unique stationary distribution is the posterior distribution on partitions. We may then use the

Strong Law of Large Numbers for Markov Chains (Chung 1967) to approximate the posterior with $E(b(P)|\mathbf{X}) \approx \frac{1}{M} \sum_{j=1}^{M} b(P_j)$.

The proposition below gives the probability distribution that governs how the Gibbs sampler moves from one partition to another. Let P be a partition of the integers {1, ..., n}, let $P^{(i)}$ denote P with the integer *i* removed, and let $P'_{i,w}$ be the partition formed by placing the integer *i* into cell C_w of $P^{(i)}$, w = 0, ..., $m(P^{(i)})$. Placing *i* in cell zero means that a new cell is formed, whose only occupant is the integer *i*. We will write the cells of $P'_{i,w}$ as C'_j, j = 1, ..., m(P′). The conditional distribution of $P'_{i,w}$ given $P^{(i)}$ is denoted $\pi(P'_{i,w}|P^{(i)})$; the notation π is used because this distribution depends only upon the prior on partitions.

Proposition. $\Pr(P'_{i,w} \mid \mathbf{X}, P^{(i)}) \propto \pi(P'_{i,w}|P^{(i)}) \prod_{j=1}^{m(P'_{i,w})} \int h(\mathbf{x}|\theta_j, C_{j,P'_{i,w}}) \, d\pi_{j,P'_{i,w}}(\theta|P'_{i,w})$. (3.1)

Note that the normalizing constant is given by summing from w = 0 to $m(P'_{i,w})$. The proof appears in the Appendix.

Computing expression (3.1) is not a problem, but in the important case where there is no covariate and the component prior is the same for all partitions, it assumes an even more convenient and suggestive form. This will be the case, for example, in mixture model classification, as we shall see later.

Theorem 2. Assume the notation of the Proposition above, a common component prior π_0 for each cell of every partition, and also that the component model density $h(\mathbf{x}|\theta, C_{j,P})$ depends on $C_{j,P}$ only through $\{x_k : k \in C_{i,P}\}$. Then $\Pr(P'_{i,w} \mid \mathbf{X}, P^{(i)}) \propto \pi(P'_{i,w}|P^{(i)}) \, f(x_i|\mathbf{X}, C_{w,P^{(i)}})$, where $f(x_i|\mathbf{X}, C_{w,P^{(i)}})$ is the component predictive density based on data with indices in cell *w* of $P^{(i)}$.

Again, the proof may be found in the Appendix.

To use the Proposition or Theorem 2 in applications, we must have $\pi(P'_{i,w}|P^{(i)})$. When all possible classifications are equally likely *a priori,* so that π(P) is a discrete uniform distribution, $\pi(P'_{i,w}|P^{(i)})$ is uniform too. That is,

$$\pi(P'_{i,w} \,|P^{(i)}) = \frac{1}{m(P^{(i)}) + 1} \quad \text{for } w = 0, ..., m(P^{(i)}). \tag{3.2}$$

For random partition classification, $\pi(P)$ assumes a special form we call the Pólya prior, and this implies

$$\pi(P'_{i,w} \,|P^{(i)}) = \begin{cases} \frac{c}{c + n - 1} & \text{for } w = 0 \\ \frac{n_w^{(i)}}{c + n - 1} & \text{for } w = 1, \cdots, m(P^{(i)}) \,, \end{cases} \tag{3.3}$$

See Section 6 for a discussion, including definition of the constant *c*.

4. Example: Songbird diets

We will illustrate the methods described in this paper with a subset of the data from an unpublished study of the contents of songbird gizzards; we would like to thank Bill McMartin of the University of Toronto Department of Forestry for permission to use the data. For each bird, we note the species, whether or not the gizzard contained insect prey of category I, and whether or not the gizzard contained insect prey of category II. The exact nature of the bird and prey species are deliberately not given here.

For this problem, the parameter ϕ_i is a vector of four quantities representing the probabilities of observing the combinations of prey species. The density $k(.|\phi_i)$ is multinomial, and we will adopt a conjugate component prior that is Dirichlet. Based on the idea that these major food sources should not both be absent from the gizzards of very many birds, and also upon considerations of habitat overlap of the prey species, the prior Dirichlet parameters were chosen to be $\alpha_1 = 1$ and $\alpha_2 = \alpha_3 = \alpha_4 = 4$; the parameter α_1 corresponds to the cell where both prey species are absent. For mixture model classification, this component prior is the base probability distribution of the Dirichlet process prior.

4.1 Classifying birds

To illustrate the Gibbs sampling method, we will pretend that we do not know species, and classify individual birds. We will perform a mixture model classification, because it seems reasonable to the think of the bird (the object) being sampled first, and then the prey being selected conditionally on the characteristics of the bird and its habitat. Because we do not anticipate a large number of meaningful clusters, we will let $c = 1$ in the Pólya prior on partitions (6.3).

Starting with a uniformly chosen partition of the integers {1, ..., 283}, we warmed up the Gibbs sampler for 1,000 cycles, and then simulated a Markov chain segment of 5,000 partitions one Gibbs cycle apart. The estimated posterior probability of a partition is simply the number of times the partition was visited, divided by 5000. Table 4.1 shows the estimated posterior probabilities for several potentially interesting partitions.

Table 4.1: Estimated posterior probabilities for several partitions of songbirds, based on diet

Partition	Estimated Posterior Probability
All birds in one subset	0.0002
One subset for each species	0.0008
Two subsets: Birds of species (1 & 3) vs. (2)	0.0034
Each bird in a separate subset	0.0000

The low probabilities in Table 4.1 are entirely typical of both real and simulated data. In simulation studies, we repeatedly have found that when individual observations are being classified, the posterior probabilities of all the partitions tended to be quite low; this includes the true partition we are trying to identify (in a simulation study we know which observations come from identical distributions). Only when we fix the number of objects being classified and increase the number of independent observations from each object does the posterior probability of the correct partition become large. This phenomenon illustrates the discussion of consistency in Section 2.2.

4.2 Classifying bird species

Another approach to this data set is to classify songbird species rather than individual birds, and to view the samples obtained from each species to be homogeneous *a priori*. In this case there are only five partitions of the three objects we seek to classify, and the calculation can be done with a calculator or with interactive software. Table 4.2.1 shows all the data we need to perform the analysis. Also, a serious prior on partitions is easy to evoke. Based on substantive considerations, species One and Three were expected to have similar diets, while Species Two was expected to be different. This partition was assigned prior probability 0.35, and to "keep an open mind," 0.35 was also assigned to the partition with only one subset (representing no relationship between species and diet). The remaining 0.30 prior probability was divided evenly among the remaining three partitions.

Table 4.2.1: Species and presence of prey category for a sample of 283 songbirds.
Cell contents are counts (and percentages of column total).

Prey Category		Songbird Species		
I	II	1	2	3
Absent	Absent	2 (1.90)	1 (0.93)	2 (2.86)
Absent	Present	43 (40.95)	7 (6.48)	21 (30.00)
Present	Absent	8 (7.62)	44 (40.74)	12 (17.14)
Present	Present	52 (49.52)	56 (51.85)	35 (50.00)

Table 4.2.2 shows the results of the analysis, both for the substantive prior described above, and for a more "unbiased" one involving a uniform prior on partitions and a uniform component

Dirichlet prior with $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$. Both analyses group species One and Three together, separately from Two; this is consistent with what one would conclude from Table 4.2.1 using other methods.

Table 4.2.2: Posterior distribution on partitions of songbird species

Partition of Bird Species	Posterior Probability	
	Substantive Prior	Uniform Prior
(1 2 3)	2.8×10^{-54}	1.5×10^{-59}
(1 2) (3)	5.1×10^{-11}	1.6×10^{-11}
(1 3) (2)	0.99932	0.986673
(1) (2 3)	0.000068	0.000105
(1) (2) (3)	3.1×10^{-8}	0.013223

It is worth observing that the classification of species is essentially a Bayesian multiple comparison procedure; it is equivalent to Menzefricke's "Bayesian clustering of data sets" for multinomial data. Classification of species was more successful than classification of individual birds (in that a clearer picture emerged) because there was a generous amount of data for each of the objects being classified.

5. Additional Remarks.

It would also be useful to expand the model by including a parameter that is assumed to be the same for each object being classified, and making all the other prior distributions (as well as the component model *h*) conditional on this parameter. Such an extended model would be useful in outlier analysis, where there might be a vector of regression coefficients that is the same for each case, and we wish to classify cases according to their dispersion. The challenge here is to extend the Gibbs sampling method to allow a parameter that is the same over all partitions, without losing the simplicity of a Markov chain with finite state space.

The Gibbs sampler on partitions is a method of moving randomly from one partition of the data to another, based on the data values and the prior distribution. With slight modification, it becomes a

method of moving deterministically from partition to partition, so that the posterior probability (or the expected value of a utility function) increases at each step. Thus one obtains a numerical method that is capable of finding at least a local maximum —– for example the posterior mode, or the local maximum of some utility function. It is similar in spirit to a numerical search suggested by Bernardo (1994) and Grenander (1993), though it is based on steepest ascent rather than simulated annealing.

A potential difficulty is that there may be many local maxima; this can happen with optimization on discrete spaces in general. One solution (as described, for example, by Geman and Geman 1984) is simulated annealing, where a deterministic search rule is metaphorically "heated" so that it melts into a random process that will not remain stuck at a local maximum, but will not stay at a global maximum either. The process is then gradually "cooled" back to a deterministic search. The function that governs cooling is called an annealing schedule. Geman and Geman prove that if the annealing schedule is right, this procedure will converge to the global maximum. Unfortunately, their theorem prescribes a schedule that is too gradual to be actually computed for most problems, so that one must try a faster schedule (or better, several faster schedules) and hope for the best. This is essentially Bernardo's (1994) simulated annealing method for cluster analysis.

Our preference is to carry out the process in two phases. We suggest first running a simple Gibbs sampling program that moves randomly from partition to partition without cooling, but remembers where it happened to encounter partitions with high posterior probability. These partitions (along with others that are chosen from a uniform distribution on partitions) can later be used as starting points for a number of fully deterministic searches based on steepest ascent in the space of partitions.

6. Mixture models and classification

The material in this section is of a more technical nature than the rest of the paper, and complements some developments in Lo, Brunner and Chan (1998). Readers who are interested primarily in applications may just note that random partition classification with a Dirichlet prior on the mixing distribution is a special case of the fixed–partition classification with a prior given by

Theorem 3 below. The rest of this section may then be skipped.

6.1 Random partition classification. Here is the random partition model. Conditionally on an unknown probability distribution G, the data $\mathbf{X} = (X_1, ..., X_n)$ are assumed to come from a distribution with density

$$f(\mathbf{x}|G) = \prod_{i=1}^{n} \int k(x_i|\phi_i)\, dG(\phi_i) \,. \tag{6.1.1}$$

This density can be interpreted in terms of the following two–stage sampling procedure. First, the *n* objects, or equivalently, $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$ are sampled independently from the distribution G, and then conditionally on $\boldsymbol{\phi}$ and G, the X_i are sampled from a distribution with density $k(\cdot|\phi_i)$, independently for i = 1, ..., n. Note that the conditional distribution of $\mathbf{X}$ given $\boldsymbol{\phi}$ and G depends on $\boldsymbol{\phi}$ but not G. This is a kind of Markov structure. Exploiting the Markov structure and re–parameterizing from $\boldsymbol{\phi}$ to $(P,\boldsymbol{\theta}_P)$ as in Section 2, we have

$$f(\mathbf{x}|\boldsymbol{\phi},G) = f(\mathbf{x}|\boldsymbol{\phi}) = f(\mathbf{x}|P,\boldsymbol{\theta}_P) = \prod_{i=1}^{m(P)} \prod_{\ell \in C_i} k(x_\ell\,|\theta_i(P)). \tag{6.1.2}$$

This is the same as our Model 1 for fixed partition classification.

To perform a classification analysis, we require the marginal posterior distribution of P. Now if we had a prior distribution on the pair $(P,\boldsymbol{\theta}_P)$, we could obtain the posterior exactly as in the fixed–partition case. However, in random–partition classification the parameter is not the pair $(P,\boldsymbol{\theta}_P)$; it is the mixing distribution G. Therefore we will proceed by deriving the marginal prior distribution of $\boldsymbol{\phi}$, in the process re–parameterizing to obtain the marginal distribution of $(P,\boldsymbol{\theta}_P)$; this will serve as our "prior" on partitions. The conditional distribution of P given $\mathbf{X}$ is then the posterior distribution on partitions for a mixture model classification.

Our prior distribution on the mixing distribution G will be a Dirichlet process distribution (Ferguson 1973), which reduces to an ordinary multinomial Dirichlet if G is supported on a finite number of points. The prior parameter of the Dirichlet process is a finite measure which we will write as

$c\pi_0$, where c is a positive constant and π_0 is a probability distribution with the same domain as G; π_0 is the prior expected value of G. If G were supported by two points ϕ_1 and ϕ_2, the Dirichlet process distribution would be a beta distribution with parameters $c\pi_0(\phi_1)$ and $c\pi_0(\phi_2)$.

Theorem 3. Assume the model (6.1.1), and let the prior on G be a Dirichlet process with parameter $c\pi_0$, where c is a positive constant and π_0 is a probability distribution defined on the same domain as G. Then re–expressing the vector $\boldsymbol{\phi}$ as the pair $(P,\boldsymbol{\theta}_P)$, the (prior) distribution of the pair $(P,\boldsymbol{\theta}_P)$ is as follows. The marginal distribution of the partition P is given by

$$\pi(P) = \frac{c^{m(P)}\,\Gamma(c)\prod_{i=1}^{m(P)}(n_i(P)-1)!}{\Gamma(c+n)}\,, \tag{6.1.3}$$

and given P, $\theta_1, ..., \theta_{m(P)}$ are independent with common distribution π_0.

The proof of Theorem 3 appears in the Appendix. From the details given there, one can see that the marginal distribution of the pair $(P,\boldsymbol{\theta}_P)$ may be understood in terms of Blackwell and MacQueen's (1973) Pólya urn scheme, as follows. One begins by selecting a point ϕ_1 from a "continuum of colors" according to the probability distribution π_0. A marble with unit weight and color ϕ_1 is placed at that point. The distribution on the continuum of colors is then re–normalized so that the original distribution π_0 has weight $c/(c+1)$, and the point ϕ_1 has weight $1/(c+1)$. Next, we draw the color ϕ_2 from the new distribution $(c\pi_0+\delta_{\phi_1})/(c+1)$, place a second marble there, draw a third point ϕ_3 from $(c\pi_0+\delta_{\phi_1}+\delta_{\phi_2})/(c+2)$, and so on.

After n draws from the extended Pólya urn, we have marbles numbered 1, ..., n; their colors are $\phi_1, ..., \phi_n$ respectively. Marbles whose numbers appear in the same cell of the partition P are all of the same color, and the vector $\boldsymbol{\theta}_P$ contains a list of the unique colors. This experiment determines a marginal distribution on the set of partitions, given by $\pi(P)$ in (6.1.3). Therefore we will call it the *Pólya prior* on partitions.

An immediate consequence of Theorem 3 is that, performing the same calculations as in the fixed partition case, the conditional distribution of P given $\mathbf{X}$ is given by (2.1.3), setting $\pi_{i,P} \equiv \pi_0$, and

letting $\pi(P)$ be specified by (6.1.3). That is, for the random partition classification model we have

$$\Pr(P|\mathbf{X}) \propto \left(c^{m(P)} \prod_{i=1}^{m(P)} (n_i(P) - 1)! \right) \left(\prod_{i=1}^{m(P)} \int \prod_{\ell \in C_i(P)} k(x_\ell \mid \theta)\, d\pi_0(\theta) \right) \qquad (6.1.4)$$

In this way, we discover that for a Bayesian, random–partition classification is a special case of fixed–partition classification, one in which the Dirichlet prior on the mixing distribution dictates a specific form for both the prior on partitions and the component priors. Notice how this mixture modelling dictates that the components of $\boldsymbol{\theta}_P$ be identically distributed as well as independent.

6.2 A general Bayesian mixture model. We now show that classification plays a crucial role in any Bayesian mixture problem where a Dirichlet prior has been assumed for the mixing distribution. Our first hint of this comes from Lo's (1984, Theorem 2) formula for Bayesian kernel density estimation. There, except for differences in notation, Lo's formula (2.4) has exactly the form (6.1.4) above, and Lo's W(**P**) is exactly the posterior distribution on partitions for a mixture model classification problem. So we see that the general formula for Bayesian classification has been hidden in plain view for well over a decade.

Assume the mixture model (6.1.1) and a Dirichlet process prior on the mixing distribution G. Again, the parameter in this problem is G, and we will express its posterior distribution in terms of $E(h(G)|\mathbf{X})$, where h is a non–negative function. Notice that in our treatment of mixture model classification analysis we did not pursue the posterior distribution of G. Instead, we integrated G out and re–parameterized to obtain a marginal posterior on partitions.

Theorem 4 expresses the posterior distribution of G as a mixture of Dirichlet processes. It is equivalent to the following four–step recipe for simulating a value of h(G). Of course this is just a conceptual recipe, since step three calls for the simulation of G from a Dirichlet process, something that is only possible when the Dirichlet process is an ordinary Dirichlet distribution. Still, the following procedure captures the idea of Theorem 4.

1. Sample a partition of {1, ..., n} from the posterior distribution of a random–partition

classification analysis; that is, from expression (6.1.4) for $\Pr(P|\mathbf{X})$.

2. Given the partition P from step 1, sample $\theta_1, ..., \theta_{m(P)}$ independently from the component posterior distributions $dF(\theta_i|\mathbf{X}_{i,P}) \propto \prod_{\ell \in C_i(P)} k(x_\ell|\theta_i)\ d\pi_0(\theta_i),\ i = 1, ..., m(P)$.

3. Given the partition P from step one and the vector $\boldsymbol{\theta}_P$ from step 2, sample G from a Dirichlet process distribution with parameter measure $c\pi_0 + \sum_{i=1}^{m(P)} n_i(P)\,\delta_{\theta_i}$, where δ_{θ_i} is a point mass (Dirac probability) measure at θ_i.

4. Compute h(G).

Theorem 4. Let $\mathbf{X} = (X_1, ..., X_n)$ be drawn from the distribution with mixture density (6.1.1). Let the prior distribution on G be a Dirichlet process with parameter measure $c\pi_0$, written $dP(G|c\pi_0)$, where c is a positive constant and π_0 is a probability distribution with the same domain as G. Then for any non–negative function h,

$$E[h(G)|\mathbf{X}] =$$

$$\sum_P \Big[\int \cdots \int \Big\{ \int h(G)\ dP(G|c\pi_0 + \textstyle\sum_{i=1}^{m(P)} n_i(P)\,\delta_{\theta_i}) \Big\} \prod_{i=1}^{m(P)} dF(\theta_i|\mathbf{X}_{i,P}) \Big]\ \Pr(P|\mathbf{X}), \qquad (6.2.1)$$

where $\Pr(P|\mathbf{X})$ is given by (6.1.4), and other expressions are defined in steps 1 through 4 of the recipe above.

The proof of Theorem 4 is given in the Appendix. Again, the role of $\Pr(P|\mathbf{X})$ as the primary mixing distribution in the mixture posterior (6.2.1) shows that a wide range of non–parametric Bayesian methods based on mixtures depend on an ordinary parametric classification analysis with a special choice of prior; references are given in the introduction. In fact, when a Bayesian is unwilling to assume a specific form for the mixing distribution in a mixture model, there are really just two tractable choices for the prior. One is a Dirichlet or Dirichlet process prior, in which case Theorem 4 applies directly. The other is a mixture of Dirichlet or Dirichlet process distributions (for

example, Antoniak 1974, Escobar and West 1995, MacEachern 1994). In this more complex case the posterior distribution depends on a mixture of classification analyses.

6.3 Priors for mixture model classification. For any Bayesian mixture model problem with a Dirichlet prior, the parameters are the positive constant *c* and the base probability distribution π_0, which is the expected value of the prior. In choosing π_0, it should be helpful to recall that it functions as a component prior in the parametric analyses that are performed separately on the data in each cluster. If prior opinion is available about the value of ϕ in the kernel (i.e., component model) density $k(x|\phi)$, that opinion should be expressed by π_0. Even if such prior opinion is not available, it will often be desirable on grounds of convenience to choose π_0 as a conjugate prior for the density k, so that the marginal density of the observations in (6.1.4) may be obtained in closed form. This done, the parameters of π_0 may be selected so that in the mixture

$$f(x) = \int k(x|\phi)\, d\pi_0(\phi), \qquad (6.3.1)$$

the density f is as close as possible to one's best prior guess about the unconditional density of the data. Purists who object to choosing priors for convenience might consider setting f(x) equal to their best guess about the density of the data, and then solving (6.3.1) for π_0 as an integral equation; this is feasible in some cases, for example that of Brunner and Lo (1989).

We will now discuss the choice of *c*. When a set of data $X_1, ..., X_n$ are assumed to arise from an unknown discrete distribution F, and a Dirichlet process prior is directly adopted for F as in Ferguson's original (1973) case, *c* is clearly the weight given to prior opinion. But when a Dirichlet process prior is adopted for the mixing distribution in a mixture model, the meaning of *c* is less obvious. Its role has been studied by Antoniak (1974), who shows that the prior mean number of subsets (cells in the partition) is asymptotically $c \log(1+n/c)$ as $n \to \infty$, and by Korwar and Hollander (1973), who prove that for a fixed partition P, $m(P) / \log(n) \xrightarrow{a.s.} c$ as $n \to \infty$. Both these results suggest that larger values of *c* imply larger numbers of subsets, and so presumably smaller values of c would be used to represent a prior opinion that there will be just a few subsets.

This is right, but it can be shown much more directly. Looking at expression (6.1.3), consider the partition with just one cell, so that m(P) = 1 and $n_{1,P} = n$. It is very easy to see that this prior probability goes to one as $c \to 0$. At the other extreme, consider the partition with n cells, where m(P) = n and $n_{i,P} = 1$, i = 1, ..., m(P). A more lengthy but still elementary calculation using Stirling's formula and L'Hôpital's rule shows that the prior probability of this partition goes to one as $c \to \infty$. These simple results clarify the role of *c* in Bayesian mixture models generally, and show that it is quite useful to have an explicit formula for the Pólya prior on partitions.

As another application of Theorem 3, consider Lo's (1984, Theorem 2) explicit formula for Bayesian kernel density estimation. Recognize Lo's W(**P**) as our mixture model posterior Pr(P|**X**), note that prior probability one implies posterior probability one, and let $c \to 0$ as in Section 4 of this paper (our *c* is Lo's $\alpha[R]$). Then we see that if the parameter measure of the Dirichlet process prior is "deflated" in this manner, the Bayesian kernel density estimate reduces to a parametric predictive density in which π_0 (Lo's $\alpha/\alpha[R]$) serves as prior and the kernel *k* acts as the model. The solution to this deflation problem had eluded us for quite a while, but with Theorem 3 it is straightforward.

Theorem 4 suggests the following question, which is of theoretical interest. Since mixture model classification is just a special kind of fixed–partition classification, suppose that we were to replace the random–partition classification analysis in Theorem 4 (step one of the recipe) with another fixed–partition classification analysis. This is just a matter of replacing the Pólya prior on partitions, perhaps with a uniform prior. The result is obviously still a distribution on G. Our question is whether this distribution is still the posterior on the mixing distribution of a mixture model, and if so, whether it is possible to recover the prior. It is worth exploring whether priors obtained in this manner would be convenient to work with or easily interpretable.

On the topic of "other" priors for mixing distributions, note that our results on mixtures apply strictly to the case of a Dirichlet or Dirichlet process priors on unknown mixing distributions. In the literature on Bayesian density estimation, one tradition (Antoniak 1974, Escobar and West 1995, MacEachern 1994, West Müller and Escobar 1994) uses a prior that is a mixture of Dirichlet

processes. We believe that for a mixture model classification with this type of prior, the clustering phase of MacEachern's Gibbs sampler for density estimation (see also West et al. 1994) will prove to be the analogue of the Gibbs sampler suggested here. This remains to be checked.

References.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to nonparametric problems. *Annals of Statistics,* **2**, 1152–1174.

Arabie, P., Hubert, L. J. and De Soete, G. (1996). *Clustering and Classification.* World Scientific Publications: River Edge, N.J.

Bernardo, J. M. (1988). Bayesian linear probabilistic classification. In S. S. Gupta and J. O. Berger (Eds.) *Statistical decision theory and related topics IV*. New York: Springer–Verlag.

Bernardo, J. M. (1994). Optimal prediction with hierarchical models: Bayesian clustering. In P. R. Freeman and A. F. M. Smith (Eds.) *Aspects of uncertainty: A tribute to D. V. Lindley*. New York: Wiley.

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika*, **65**, 31–38.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Annals of Statistics* **1**, 353–355.

Bock, H–H. (1989) Probabilistic aspects in cluster analysis. In O. Opitz (Ed.), *Conceptual and numerical analysis of data*. Heidelberg: Springer –Verlag.

Bock, H–H. (1996), Probability models and hypothesis testing in partitioning cluster analysis. In P. Arabie, L. J. Hubert and G. De Soete (1996), Clustering and Classification. World Scientific Publications: River Edge, N.J.

Box, G. E. P. and Tiao, G. C. (1968), A Bayesian approach to some outlier problems. *Biometrika* **55**, 119–129.

Broemeling, L. D. (1985). *Bayesian analysis of linear models*. New York: Marcel Deckker.

Brunner, L. J. (1989). Bayes procedures for elliptically symmetric distributions. Unpublished Ph. D. Dissertation, State University of New York at Buffalo, Department of Statistics.

Brunner, L. J. (1992) Bayesian nonparametric methods for data from a unimodal density. *Statistics and Probability Letters* **14**, 195–199.

Brunner, L. J. and Lo, A. Y. (1989). Bayes methods for a symmetric and unimodal density and

its mode. *Annals of Statistics* **17**, 1550–1566.

Brunner, L. J. and Lo, A. Y. (1994). Bayes methods for directional data. *Canadian Journal of Statistics* **22**, 401–412.

Chung, K. L. (1967). Markov chains with stationary transition probabilities. Springer–Verlag.

Comtet, L. (1974). *Advanced combinatorics* (Revised and enlarged edition). Boston: D. Reidel.

Doob, J. (1949). Application of the theory of martingales. *Coll. Int. du CNRS*: Paris.

Draper, N. and Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: Wiley.

Dykstra, R.L. and Laud, P. (1981) A Bayesian nonparametric approach to reliability, *Annals of Statistics*, **9** 356–367.

Escobar, M. D. (1988). Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Unpublished Ph. D. Dissertation, Yale University, Department of Statistics.

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

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

Ferguson, T. S. (1983), "Bayesian density estimation by mixtures of normal distributions," In M. H. Rizvi, J. S. Rustagi and D. Siegmund (Eds.) *Recent Advances in Statistics*. New York: Wiley.

Geisser, S. (1964). Posterior odds for multivariate normal classification. *Journal of the Royal Statistical Society, Series B*, **26**, 69–76.

Geisser, S. (1966). Predictive discrimination. In P. R. Krishnaiah (Ed.), *Multivariate Analysis*. New York: Academic Press.

Geisser, S. (1982). Bayesian discrimination. In P. R. Krishnaiah and L. N. Kanal (Eds.), *Handbook of Statistics, Vol. 2*. New York: North–Holland.

Gelfand, A. E. and Smith, A. F. M. (1990). "Sampling–based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 972–985.

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

Grenander, U. (1993). *General Pattern theory: A mathematical study of regular structures*. New York: Oxford University Press.

Guttman, I. (1973). Care and handling of univariate or multivariate outliers in detecting spuriosity —— a Bayesian approach. *Technometrics*, **15**, 723–738.

Guttman, I. Dutter, R. and Freeman, P. R. (1978). Care and handling of univariate outliers in the general linear model to detect spuriosity –– a Bayesian approach. *Technometrics*, **20**, 187–193.

Hartigan, J. A. (1975). *Clustering algorithms*. New York: Wiley.

Johnson, W. O. and Kokolakis, G. E. (1994). Bayesian classification based on multivariate binary data. *Journal of Statistical Planning and Inference*, **41**, 21–35.

Johnson, W. O. and Mouhab, A. (1996). A Bayesian decision theory approach to classification problems. *Journal of Multivariate Analysis*, **56**, 232–244.

Korwar, R. M. and Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *Annals of Probability* **1**, 705–711.

Lavine, M. and West, M. (1992). A Bayesian method for classification and discrimination. *Canadian Journal of Statistics,* **20**, 451–461.

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

Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics,* **12**, 351–357.

Lo, A. Y., Brunner, L. J. and Chan, A. T. (1998) Weighted Chinese restaurant processes and Bayesian mixture models. Research Report, Department of Information and Systems Management, Hong Kong University of Science and Technology.

Lo, A.Y. and Weng, C.S. (1989). On a class of Bayesian nonparametric estimates: II. Hazard rate estimates. *Ann. Instit. Statist. Math.,* **41**, 227–245.

Menzefricke, U. (1981). Bayesian clustering of data sets. *Communications in Statistics – Theory and Methods*, **1**, 65–77.

MacEachern, S.N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics.–Simulation,* **23** 727–741.

Shapiro, C. P. (1977). Classification by maximum posterior probability. *Annals of Statistics* **5**, 185–190.

Symons, M. J. (1981). Clustering criteria and multivariate normal mixtures. *Biometrics* **37**, 35–43.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York: Wiley.

West, M., Müller, P. and Escobar, M. D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In P. R. Freeman and A. F. M. Smith (Eds.) *Aspects of uncertainty: A tribute to D. V. Lindley*. New York: Wiley.

Appendix: Proofs

Proof of Theorem 1a.

$$\begin{aligned}
\Pr(P|\mathbf{X}) &\propto \pi(P) \int \prod_{i=1}^{m(P)} \prod_{\ell \in C_{i,P}} k(x_\ell|\theta_i)\, d\pi(\boldsymbol{\theta}_P|P) \\
&= \pi(P) \int \cdots \int \prod_{i=1}^{m(P)} \prod_{\ell \in C_{i,P}} k(x_\ell|\theta_i) \prod_{i=1}^{m(P)} d\pi_{i,P}(\theta_i|P) \\
&= \pi(P) \prod_{i=1}^{m(P)} \int \prod_{\ell \in C_{i,P}} k(x_\ell|\theta)\, d\pi_{i,P}(\theta|P)\ ,
\end{aligned}$$

and

$$\begin{aligned}
d\pi(\boldsymbol{\theta}_P|\mathbf{X},P) &\propto \prod_{i=1}^{m(P)} \prod_{\ell \in C_{i,P}} k(x_\ell|\theta_i)\, d\pi(\boldsymbol{\theta}_P|P) \\
&= \prod_{i=1}^{m(P)} \prod_{\ell \in C_{i,P}} k(x_\ell|\theta_i) \prod_{i=1}^{m(P)} d\pi_{i,P}(\theta_i|P) \\
&= \prod_{i=1}^{m(P)} \prod_{\ell \in C_{i,P}} k(x_\ell|\theta_i) d\pi_{i,P}(\theta_i|P) \qquad \blacksquare
\end{aligned}$$

Proof of the Proposition. We will write $P'_{i,w}$ as P', $j = 1, ..., m(P')$. Using Bayes' theorem,

$$\begin{aligned}
\Pr(P'|\,\mathbf{X},P^{(i)}) &\propto \pi(P'|P^{(i)})\, f(\mathbf{x}|P',P^{(i)}) = \pi(P'|P^{(i)})\, f(\mathbf{x}|P') \\
&= \pi(P'|P^{(i)}) \int f(\mathbf{x}|P',\boldsymbol{\theta}_{P'})\, d\pi(\boldsymbol{\theta}_{P'}|P') \\
&= \pi(P'|P^{(i)}) \int \cdots \int \prod_{j=1}^{m(P')} \int h(\mathbf{x}|\theta_j,C_{j,P'}) \prod_{j=1}^{m(P')} d\pi_{j,P'}(\theta_j|P') \\
&= \pi(P'|P^{(i)}) \prod_{j=1}^{m(P')} \int h(\mathbf{x}|\theta_j,C_{j,P'})\, d\pi_{j,P'}(\theta|P') \qquad \blacksquare
\end{aligned}$$

Proof of Theorem 2. Assume the notation of the proposition. We distinguish between the case $w = 0$, where the integer *i* is placed into a new cell by itself, and the case $w = 1, ..., m(P^{(i)})$, where i is placed into an existing cell of $P^{(i)}$. In the former case, the conditional independence of observations in different cells implies

$$\Pr(P'|\mathbf{X},P^{(i)}) \propto \pi(P'|P^{(i)}) \left[\int h(\mathbf{x}|\theta,C_{0,\,P'})\, d\pi_0(\theta) \right] \left[\prod_{j \neq w}^{m(P')-1} \int h(\mathbf{x}|\theta,C_{j,\,P'})\, d\pi_0(\theta) \right]$$

$$= \pi(P'|P^{(i)}) \left[\int h(\mathbf{x}|\theta,C_{0,\,P'})\, d\pi_0(\theta) \right] \left[\prod_{j=1}^{m(P^{(i)})} \int h(\mathbf{x}|\theta,C_{j,\,P^{(i)}})\, d\pi_0(\theta) \right]$$

$$\propto \pi(P'|P^{(i)}) \int h(\mathbf{x}|\theta,C_{0,\,P'})\, d\pi_0(\theta) = \pi(P'|P^{(i)})\, f(x_i|\mathbf{X},C_{w,P^{(i)}}) \ ,$$

where we recognize $h(\mathbf{x}|\theta,C_{0,\,P'})$ as just the marginal density of X_i conditional on θ, and the predictive density indicated is a (component) prior predictive. In the case where $w \neq 0$, $\Pr(P'|\mathbf{X},P^{(i)})$ is proportional to

$$\pi(P'|P^{(i)}) \prod_{j=1}^{m(P')} \int h(\mathbf{x}|\theta,C_{j,\,P'})\, d\pi_0(\theta|\,P')$$

$$= \pi(P'|P^{(i)}) \int h(\mathbf{x}|\theta,C_{w,\,P'})\, d\pi_0(\theta) \prod_{j \neq w}^{m(P')-1} \int h(\mathbf{x}|\theta,C_{j,\,P'})\, d\pi_0(\theta)$$

$$= \pi(P'|P^{(i)}) \int h(\mathbf{x}|\theta,C_{w,\,P'})\, d\pi_0(\theta) \prod_{j \neq w}^{m(P^{(i)})-1} \int h(\mathbf{x}|\theta,C_{j,\,P^{(i)}})\, d\pi_0(\theta)$$

$$\propto \pi(P'|P^{(i)}) \int h(\mathbf{x}|\theta,C_{w,\,P'})\, d\pi_0(\theta)$$

$$= \pi(P'|P^{(i)}) \int f(x_i|\{x_j:\, j \in C_{w,\,P^{(i)}}\},\theta)\, h(\mathbf{x}|\theta,C_{w,\,P^{(i)}})\, d\pi_0(\theta)$$

$$\propto \pi(P'|P^{(i)}) \frac{\int f(x_i|\{x_j:\, j \in C_{w,P^{(i)}}\},\theta)\, h(\mathbf{x}|\theta,C_{w,P^{(i)}})\, d\pi_0(\theta)}{\int h(\mathbf{x}|\theta,C_{w,P^{(i)}})\, d\pi_0(\theta)} = \pi(P'|P^{(i)})\, f(x_i|\mathbf{X},C_{w,P^{(i)}}) \ \blacksquare$$

Proof of Theorem 3. Let $h(\boldsymbol{\phi})$ be a non–negative function (not a component model density!), and define h' so that $h(\boldsymbol{\phi}) = h'(P,\boldsymbol{\theta}_P)$ whenever the pair $(P,\boldsymbol{\theta}_P)$ corresponds to $\boldsymbol{\phi}$. Then

$$E[h(\boldsymbol{\phi})] = E\{E[h(\boldsymbol{\phi})|G]\} = \int\int h(\boldsymbol{\phi})\, dF(\boldsymbol{\phi}|G)\, dP(G|c\pi_0)$$

$$= \int\int\cdots\int h(\boldsymbol{\phi}) \prod_{i=1}^{n} dG(\phi_i)\ dP(G|c\pi_0),$$

where $dP(G|c\pi_0)$ refers to the distribution of the Dirichlet process with parameter measure $c\pi_0$. Applying Lo's (1984) Lemma One to exchange order of integration *n* times, the expression

becomes

$$\int \cdots \int \int dP(G|c\pi_0 + \sum_{i=1}^{n} \delta_{\phi_i})\, h(\boldsymbol{\phi})\, \frac{d(c\pi_0 + \sum_{i=1}^{n-1}\delta_{\phi_i}))(\phi_n)}{c+n-1} \cdots \frac{d(c\pi_0 + \delta_{\phi_1})(\phi_2)}{c+1}\, d\pi_0(\phi_1)$$
$$= \frac{\Gamma(c)}{\Gamma(c+n)} \int \cdots \int h(\boldsymbol{\phi}) \prod_{i=1}^{n} d(c\pi_0 + \sum_{j=1}^{i-1} \delta_{\phi_j}(\phi_i),$$

where δ_{ϕ_i} indicates a point mass measure at ϕ_i. Lo's (1984) Lemma Two allows this to be expressed as a sum over partitions, automatically re–parameterizing from $\boldsymbol{\phi}$ to $(P,\boldsymbol{\theta}_P)$ so that we may write

$$E[h(\boldsymbol{\phi})] = \quad E[h'(P,\boldsymbol{\theta}_P)] = \frac{\Gamma(c)}{\Gamma(c+n)} \sum_{P} \left\{ \int \cdots \int h'(P,\boldsymbol{\theta}_P) \prod_{i=1}^{m(P)} (n_i(P)-1)!\; dc\pi_0(\theta_{i,P}) \right\}$$
$$= \sum_{P} \left\{ \int \cdots \int h'(P,\boldsymbol{\theta}_P) \prod_{i=1}^{m(P)} d\pi_0(\theta_{i,P}) \right\} \frac{c^{m(P)}\,\Gamma(c) \prod_{i=1}^{m(P)} (n_i(P)-1)!}{\Gamma(c+n)}$$

establishing the claim of the theorem. ∎

Proof of Theorem 4. Bayes' theorem states that

$$E(h(G)|\mathbf{X}) = \frac{\int h(G)\, f(\mathbf{x}|G)\, dP(G|c\pi_0)}{\int f(\mathbf{x}|G)\, dP(G|c\pi_0)}.$$

The numerator of this quantity is

$$\int \int \cdots \int h(G) \prod_{i=1}^{n} k(x_i|\phi_i) dG(\phi_i)\, dP(G|c\pi_0)$$

$$= \int \cdots \int \int h(G) \prod_{i=1}^{n} k(x_i|\phi_i)\, dP(G|c\pi_0 + \sum_{i=1}^{n} \delta_{\phi_i})$$
$$\frac{d(c\pi_0 + \sum_{i=1}^{n-1}\delta_{\phi_i}))(\phi_n)}{c+n-1} \cdots \frac{d(c\pi_0 + \delta_{\phi_1})(\phi_2)}{c+1}\, d\pi_0(\phi_1)$$

$$= \frac{\Gamma(c)}{\Gamma(c+n)} \int \cdots \int \int h(G) \prod_{i=1}^{n} k(x_i|\phi_i)\, dP(G|c\pi_0 + \Sigma_{i=1}^{n} \delta_{\phi_i}) \prod_{i=1}^{n} d(c\pi_0 + \Sigma_{j=1}^{i-1} \delta_{\phi_j})(\phi_i),$$

where the exchange of integration is accomplished by Lo's (1984) Lemma One. Following the combinatoric argument of Lo's Lemma Two, this may be written as

$$\frac{\Gamma(c)}{\Gamma(c+n)} \sum_{P} \Big[\prod_{i=1}^{m(P)} (n_i(P)-1)! \int \cdots \int \Big\{ \int h(G)\, dP(G|c\pi_0 + \Sigma_{i=1}^{m(P)} n_i(P)\, \delta_{\theta_i}) \Big\} \prod_{\ell \in C_i(P)} k(x_\ell|\theta_i)\, dc\pi_0(\theta_i) \Big].$$

Notice that the collapsing of $\boldsymbol{\phi}$ into $\boldsymbol{\theta}_P$ has changed the n–fold integral into an m(P)–fold integral. Multiplying and dividing by the product of normalizing constants for $dF(\theta_i|\mathbf{X}_{i,P})$ defined in step 2 of the "recipe," our expression for the numerator becomes

$$\frac{\Gamma(c)}{\Gamma(c+n)} \sum_{P} \Big[\Big(c^{m(P)} \prod_{i=1}^{m(P)} (n_i(P)-1)!\Big)\Big(\prod_{i=1}^{m(P)} \int \prod_{\ell \in C_{i,P}} k(x_\ell|\theta_i)\, d\pi_0(\theta_i)\Big)$$

$$\int \cdots \int \Big\{ \int h(G)\, dP(G|c\pi_0 + \Sigma_{i=1}^{m(P)} n_i(P)\, \delta_{\theta_i}) \Big\} \prod_{i=1}^{m(P)} dF(\theta_i|\mathbf{X}_{i,P}) \Big] . \quad (*)$$

The denominator of $E(h(G)|\mathbf{X})$ is just (*) with $h(G)\equiv 1$. Integrating probability distributions to one and cancelling ratios of gammas in numerator and denominator, we get a numerator equal to (*) without the gammas. The denominator is

$$\sum_{P} \Big(c^{m(P)} \prod_{i=1}^{m(P)} (n_i(P)-1)!\Big)\Big(\prod_{i=1}^{m(P)} \int \prod_{\ell \in C_i(P)} k(x_\ell|\theta_i)\, d\pi_0(\theta_i)\Big) .$$

We recognize this as the normalizing constant of $Pr(P|\mathbf{X})$, and the result follows. ∎