

A Predictive View of Bayesian Clustering

Fernando A. Quintana*

September, 2004

Abstract

This work considers probability models for partitions of a set of n elements using a predictive approach, i.e., models that are specified in terms of the conditional probability of either joining an already existing cluster or forming a new one. The inherent structure can be motivated by resorting to hierarchical models of either parametric or nonparametric nature. Parametric examples include the product partition models (PPMs) and the model-based approach of Dasgupta and Raftery (1998), while nonparametric alternatives include the Dirichlet Process, and more generally, the Species Sampling Models (SSMs). Under exchangeability, PPMs and SSMs induce the same type of partition structure. The methods are discussed in the context of outlier detection in normal linear regression models and of (univariate) density estimation.

KEY WORDS: Density estimation; Dirichlet process; EM algorithm; Model-based clustering; Outlier detection; Product partition models; Species sampling models.

*Departamento de Estadística, Pontificia Universidad Católica de Chile, Avenida Vicuña Mackenna 4860, Santiago, CHILE. e-mail: quintana@mat.puc.cl. Partially supported by grant FONDECYT 1020712.

1 Introduction

This article reviews and explores the connections between several types of models that induce probability distributions on the space of partitions of a finite set of objects $S = \{1, \dots, n\}$, say. A fundamental application of such models is given by clustering algorithms, which attempt to find out a partition of S into a not necessarily predetermined number of nonempty subsets. This means that the elements of a partition are viewed as clusters. Many methods have been proposed in the statistical literature. An alternative that has proven to be particularly successful in applications, compared to other heuristic choices, assumes a mixture model with a probably unknown number of components. Typically, each mixture component depends on a parameter vector which could have common and idiosyncratic components. Parameter estimation is usually done via the EM algorithm. Once a number of components has been selected, clusters may be formed based on the posterior probability of component membership for each data point. This setup has been followed by many authors, in particular by McLachlan and Basford (1988), Banfield and Raftery (1993), Dasgupta and Raftery (1998), McLachlan and Peel (2000) and Fraley and Raftery (2002). For a concrete demonstration of the capabilities of this approach, see Yeung et al. (2001). Following Dasgupta and Raftery (1998), this method will be referred to as *model-based clustering* (MBC).

An alternative model-based approach is given by the product partition models (PPM) introduced by Barry and Hartigan(1992, 1993). Here, experimental units in a given group are thought of as sampled from a common distribution, and *a priori*, groups are formed according to a product distribution. Using this construction, Quintana and Iglesias (2003) proposed an algorithm for the selection of a single partition according to some specific decision problem of interest such as outlier detection.

In this article the discussion is centered on a predictive formulation of probability models for partitions. In other words, the focus is on models that are specified through the sequence of conditional probabilities of joining an already existing cluster or starting a new one. In some applications, this approach may be advantageous for prior

elicitation purposes. And of course, it is entirely equivalent to defining directly the corresponding joint probabilities. Several families of models are available. These are reviewed, with special interest in studying the various existing connections. Under the predictive view, particular attention is given to the partition structures related to the class of *species sampling models* (SSMs) described in Pitman (1996) and in Ishwaran and James (2003a). The resulting structures are quite general. In fact, they will be seen to encompass the MBC approach and, to a large extent, the class of PPMs. The flexibility of SSMs is highlighted by the fact that under some mild conditions they can be represented by means of a random probability measure (RPM). Therefore, one may consider quite flexible classes of models for partition structures under either a parametric or nonparametric approach, depending on the specific needs or interests. A practical advantage of this duality is that algorithms originally devised for one context can be readily adapted to the other.

The rest of the article is organized as follows. Section 2 reviews the several available models, with emphasis in the corresponding predictive specification of probability models. Several novel connections between partition structures are pointed out. Section 3 discusses hierarchical models built on partition structures, also describing various simulation schemes that can be used for posterior inference. This includes a general discussion of the sequential importance sampling and Gibbs sampling algorithms. Section 4 is devoted to clustering methods, particularly the algorithms by Dasgupta and Raftery (1998) and Quintana and Iglesias (2003). Sections 5 and 6 consider applications of the main ideas to the problems of outlier identification in normal linear regression models, and of univariate density estimation, presenting in each case a comparison of the results under the various approaches described earlier. Some final remarks are stated in Section 7.

2 A Predictive Approach to Probability Models on Partitions

Let $S = \{1, \dots, n\}$ represent a collection of n objects. In later Sections, the elements of S will represent the set of indexes associated to experimental units. Let $\rho = (S_1, \dots, S_k)$ denote a partition of S into k nonempty disjoint sets (clusters). Each set S_i in ρ consists of $n_i \geq 1$ elements of S , with $\sum_{i=1}^k n_i = n$. Without loss of generality it will be assumed that S_1, \dots, S_k are “sorted in ascending order” which means that $\min\{s : s \in S_1\} < \min\{s : s \in S_2\} < \dots < \min\{s : s \in S_k\}$. Let $s_i = j$ if $i \in S_j$ denote the cluster memberships, which can be used to give an alternative representation of ρ . By construction it follows that $s_1 = 1$. An additional “no gap” restriction will be assumed, that is, $s_i = \ell > 1$ for some $i \in S$ implies that there exist $i_1, \dots, i_{\ell-1} \in S$ such that $s_{i_j} = j$ for $j \leq \ell - 1$.

Defining probability models on the set \mathcal{S} of partitions of S is equivalent to specifying a distribution on the vector of cluster memberships: $p(s_1, \dots, s_n)$. Since $P(s_1 = 1)$ implies

$$p(s_1, s_2, \dots, s_n) = \prod_{j=2}^n p(s_j | s_{j-1}, \dots, s_2, s_1),$$

the definition of $p(s_1, s_2, \dots, s_n)$ can be done in terms of the conditional probabilities $p(s_j | s_{j-1}, \dots, s_2, s_1)$. Each specific choice will reflect a predictive view of the way clusters are formed. Indeed, such probability models describe how the elements of S are successively assigned to either existing clusters or to start a new one.

Before turning to specific examples it is necessary to introduce the following notation. For $1 \leq \ell \leq n$, let s_1, s_2, \dots, s_ℓ be a sequence of cluster memberships that are sorted in ascending order as explained earlier. Let $k_\ell = \max\{s_1, \dots, s_\ell\}$ be the number of clusters that have been formed among elements $1, \dots, \ell$, and let $m_{1,\ell}, \dots, m_{k_\ell,\ell}$ be the corresponding cluster sizes, i.e., $m_{i,\ell} = \sum_{j=1}^{\ell} I\{s_j = i\}$ for $1 \leq i \leq \ell$, where $\sum_{i=1}^{k_\ell} m_{i,\ell} = \ell$ for all $1 \leq \ell \leq n$.

Multiple choices are available for specifying the conditional probabilities. These are reviewed next.

2.1 Dirichlet Process Partitioning

Consider a model of the type

$$X_1, \dots, X_n | F \sim F, \quad F \sim \mathcal{D}(cF_0),$$

where \mathcal{D} represents the Dirichlet process (DP) introduced by Ferguson (1973) with base measure F_0 and weight parameter $c > 0$. Since F is a.s. discrete, there can be ties among the X_i 's, which can be also seen from the Polya urn representation of Blackwell and MacQueen (1973). Thus, a partition of S can be formed by defining clusters by means of the equivalence classes under the relation $i \sim j$ if and only if $X_i = X_j$. Specifically, let $s_1 = 1$, $X_1^* = X_1$ and for $j > 1$ let $s_j = s_i$ if $X_j = X_i$ for some $1 \leq i \leq j - 1$ and $s_j = k_{j-1} + 1$ if $X_j \notin \{X_1, \dots, X_{j-1}\}$, in which case denote $X_{s_j}^* = X_j$. In words, X_1^*, X_2^*, \dots represent the *cluster locations* or set of different (unique) values observed in the sample X_1, \dots, X_n . The above representation implies that the complete sample is in one-to-one correspondence with the set of locations and memberships $\{s_j\}$. Such representations are also useful from a computational viewpoint (Bush and MacEachern 1996, MacEachern and Müller 1998). It then follows that

$$P(s_{\ell+1} = i | s_\ell, \dots, s_1) = \begin{cases} \frac{m_{i,\ell}}{c+\ell} & \text{if } 1 \leq i \leq k_\ell \\ \frac{c}{c+\ell} & \text{if } i = k_\ell + 1. \end{cases} \quad (1)$$

A colorful description of this partition structure and its corresponding predictive rule is given by the *Chinese restaurant process* presented in Arratia, Barbour and Tavaré (1992). Here, customers enter a restaurant sequentially and sit one after the other. Initially all tables are folded up, so that one table is opened when the first customer enters. After customers $1, \dots, \ell$ are seated, customer $\ell + 1$ will either choose an empty table with probability $c/(c + \ell)$ for some $c > 0$, or an occupied table with probability proportional to the number of occupants at the given table. Other important properties of Dirichlet processes can be found in Blackwell and MacQueen (1973), Korwar and Hollander (1973), Antoniak (1974), Lo (1984), Diaconis and Freedman (1986), Rolin (1992), Diaconis and Kemperman (1996), Cifarelli and Melilli (2000) and references therein.

2.2 Species Sampling Models

Consider now inference on the number of distinct species in a certain large population of various species, where each species is assumed to have a different preassigned tag. Assume X_1, \dots, X_n are drawn from such population, so that X_ℓ represents the tag that corresponds to the ℓ th sampled individual. Using the same notation as earlier, let X_j^* denote the j th recorded species. Write $\mathbf{m}_\ell = (m_{1,\ell}, \dots, m_{k_\ell,\ell})$. The sequence $\{X_n\}$ is called a *species sampling sequence* if it is exchangeable and satisfies

$$X_1 \sim F_0 \tag{2}$$

$$\begin{aligned} X_{\ell+1} &\sim F_\ell \stackrel{\text{def}}{=} P(X_{\ell+1} | X_1, \dots, X_\ell, k_\ell = k) \\ &= \sum_{j=1}^k p_j(\mathbf{m}_\ell) \delta_{X_j^*} + p_{k+1}(\mathbf{m}_\ell) F_0, \end{aligned} \tag{3}$$

where F_0 is a non-atomic distribution function on the appropriate space, δ_x is a point-mass at x , $p_j \geq 0$, and $\sum_{j=1}^{k_\ell+1} p_j(\mathbf{m}_\ell) = 1$ for all possible values \mathbf{m}_ℓ and k_ℓ . Here the conditional probabilities $p(s_j | s_{j-1}, \dots, s_2, s_1)$ are expressed by the p_j functions, also called *predictive probability functions* (PPF) by Pitman (1996). Note that the PPFs depend on the cluster membership vectors only through the cluster sizes \mathbf{m}_ℓ . It can be shown (Pitman 1996) that the induced probability model for partitions under SSMs is of the form

$$P(\rho = (S_1, \dots, S_k)) = p(|S_1|, \dots, |S_k|), \tag{4}$$

where $|S_i|$ is the number of elements in S_i and the p function at the right-hand side of (4) is a symmetric function. In words, $P(\rho)$ depends on the specific partition only indirectly through the sizes $|S_j|$ of the partitioning subsets S_j . Pitman (1996) refers to the right-hand side of (4) as the *exchangeable partition probability function* (EPPF), also establishing the relationship between this and the PPF. Interestingly, the exchangeability of $\{X_n\}$ turns out to be equivalent to the existence of an EPPF p that determines the PPF (Pitman 1996).

An interesting fact concerning such sequences is that F_n given in (3) converges

almost surely in the total variation norm to a random distribution represented as

$$F = \sum_{j=1}^{\infty} w_j \delta_{X_j^*} + \left(1 - \sum_{j=1}^{\infty} w_j\right) F_0, \quad (5)$$

where w_j is the almost sure limit of $m_{j,\ell}/\ell$ as $\ell \rightarrow \infty$, with $P(\sum_{j=1}^{\infty} w_j \leq 1) = 1$, the $\{X_j^*\}$ are a sample from F_0 independent of the $\{w_j\}$, and the sequence $\{X_n\}$ is a sample from F (Pitman 1996). Any setup with a random probability measure F of the form (5) and a sample $\{X_n\}$ from F is called a *species sampling model* (SSM). In addition, if $P(w_1 > 0) = 1$ or equivalently $P(\lim_{\ell \rightarrow \infty} k_\ell/\ell = 0) = 1$ then $P(\sum_{j=1}^{\infty} w_j = 1) = 1$ in which case the model is said to be proper (Pitman 1996). Intuitively, the w_j weights in (5) represent the (limit) fraction of sampled individuals that belong to the species with tag equal to X_j^* , $j \geq 1$. An important special case of SSM is given by the DP. In this case the model is proper and w_1, w_2, \dots follow a *stick-breaking process*, defined as $w_1 = U_1$, $w_j = U_j \prod_{i=1}^{j-1} (1 - U_i)$ for $j \geq 2$, where U_1, U_2, \dots are i.i.d. with Beta(1, c) distribution (Sethuraman 1994).

The RPM F given in (5) is quite general, including not only the DP as a special case, but also the Dirichlet-multinomial process of Muliere and Secchi (1995), and the stick-breaking priors of Ishwaran and James (2001), among others. As in the DP case, a graphic description of the clustering structure is given by the *generalized Chinese restaurant process* (Lo et al. 1998, Brunner et al. 2001, Ishwaran and James 2003a), where the seating probabilities are defined in terms of the PPFs. Further properties of SSMs can be found in Pitman (1996). Nonparametric Bayesian inference for hierarchical models using SSMs is discussed in Ishwaran and James (2003a). Related extensions of stick-breaking priors can be found in Ishwaran and James (2003b).

2.3 Product Partition Models

An alternative way of defining probability models on partition structures is given by the class of *product partition models* (PPMs) as in Hartigan (1990) or Crowley (1997). A PPM is defined as follows. For any partition $\rho = \{S_1, \dots, S_k\}$ of S and

data X_1, \dots, X_n , it is assumed that

$$p(X_1, \dots, X_n | \rho) = \prod_{i=1}^k p_{S_i}(\mathbf{X}_{S_i}), \quad (6)$$

where $\mathbf{X}_{S_i} = (X_j : j \in S_i)$ and $p_{S_i}(\mathbf{X}_{S_i})$ depends only on S_i and not on other subsets in the partition. The partition ρ is in turn assigned a prior probability model

$$P(\rho = \{S_1, \dots, S_k\}) = \mathcal{K} \prod_{i=1}^k c(S_i), \quad (7)$$

where for $A \subset \{1, \dots, n\}$, $c(A) \geq 0$ is called the *cohesion* of subset A , and \mathcal{K} is a normalizing constant, so that the sum over all partitions is 1. It follows that the posterior distribution under (6)–(7) is again a PPM with cohesions given by $c(S_i)p_{S_i}(\mathbf{X}_{S_i})$.

The partition structures corresponding to PPMs and SSMs are not equivalent in general, but they do have a considerable intersection. For instance, the DP induces a marginal prior distribution on partitions that can be expressed as (7) with $c(S_i) = (|S_i| - 1)! \times c$, which is closely related to the *Ewens sampling formula* (Ewens 1972). Such a connection between the DP and PPMs was pointed out by Quintana and Iglesias (2003). More generally, if the cohesions depend on S_i only through its size $|S_i|$ then the product form (7) becomes a special case of (4). In contrast, $p(|S_1|, \dots, |S_k|) \propto (|S_1| + \dots + |S_k|)$ in (4) is not compatible with the product form (7). In any case, from (7) the predictive probabilities for PPMs are given by

$$p(s_{j+1} = i | s_j, \dots, s_2, s_1) \propto \begin{cases} \frac{c(S_i(j) \cup \{i\})}{c(S_i(j))} & \text{if } 1 \leq i \leq k_j \\ c(\{i\}) & \text{if } i = k_j + 1, \end{cases} \quad (8)$$

where $(S_1(j), \dots, S_{k_j}(j))$ is the partition associated to the first j elements of S , as determined by the memberships s_1, \dots, s_j . See additional applications and discussion of PPMs, particularly in the context of *change point problems*, in Barry and Hartigan (1992, 1993), Loschi and Cruz (2002) and Cruz et al. (2003).

2.4 Model-Based Clustering (MBC)

This approach was originally proposed by Banfield and Raftery (1993), based on work by Murtagh and Raftery (1984), and later extended by Dasgupta and Raftery (1998). They consider independent responses X_1, \dots, X_n and a mixture model

$$p(X_1, \dots, X_n | K, \boldsymbol{\theta}) = \prod_{i=1}^n \sum_{j=1}^K \tau_{K,j} f_j(X_i | \theta_j), \quad (9)$$

where K represents the number of components in the mixture, $\tau_{K,j}$ is the weight given to the j th element, with $\tau_{K,j} \geq 0$ and $\sum_{j=1}^K \tau_{K,j} = 1$ for all K , and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ are component-specific parameters of possibly varying dimensions. For fixed K , the probability that a given data point falls into category j is then $\tau_{K,j}$ and all data points are classified independently. The mixture (9) can be equivalently expressed as a hierarchical model

$$X_1, \dots, X_n | K, \boldsymbol{\theta}, e_1, \dots, e_n \sim p(X_i | \boldsymbol{\theta}, e_i = j) = f_j(X_i | \theta_j), \quad P(e_i = j) = \tau_{K,j}.$$

Using this alternative formulation, clusters are defined via the latent indicators e_i . It also follows that the probability of any cluster assignment is of multinomial type, so clusters are not constrained to be nonempty and no order relation exists between them. However, it is possible to reparametrize (e_1, \dots, e_n) to the ascending order and no gaps format described earlier, by simply reinterpreting the equivalence relation already stated for data points now in terms of mixture component indicators. The conditional predictive probabilities, under the restriction to K components, are thus given by

$$P_K(s_{j+1} = i | s_j, \dots, s_1) \propto \begin{cases} \sum_{(\nu_1, \dots, \nu_{k_j}) \in \mathcal{J}(K, k_j)} \tau_{K, \nu_i}^{n_i+1} \prod_{\ell=1, \ell \neq i}^{k_j} \tau_{K, \nu_\ell}^{n_\ell} & \text{if } 1 \leq i \leq k_j \\ \sum_{(\nu_1, \dots, \nu_{k_j}) \in \mathcal{J}(K, k_j)} (1 - \sum_{t=1}^{k_j} \tau_{K, \nu_t}) \prod_{\ell=1}^{k_j} \tau_{K, \nu_\ell}^{n_\ell} & \text{if } i = k_j + 1 \\ 0 & \text{otherwise,} \end{cases}$$

where for $k_j \leq K$

$$\mathcal{J}(K, k_j) = \{(\nu_1, \dots, \nu_{k_j}) \subset \{1, \dots, K\} : \nu_a \neq \nu_b \text{ for all } a \neq b \in \{1, \dots, k_j\}\},$$

and with $\mathcal{J}(K, k_j) = \emptyset$ if $k_j > K$. Optionally, some authors consider putting a prior distribution on K , so that $P(K = k) = p_k$ (e.g. Richardson and Green 1997, Stephens 2000), which implies (marginally) that

$$P(s_{j+1} = i | s_j, \dots, s_1) = \sum_{k=1}^n P_k(s_{j+1} = i | s_j, \dots, s_1) p_k.$$

3 Hierarchical Modeling

The previous section discusses how different partition structures can be described either conditionally or jointly. While the latter explicitly defines probability models on partitions, it will be argued that the former is the appropriate one for posterior simulation.

From a modeling perspective, great flexibility is obtained when the partition structure is placed at the parameter level, rather than at the observation level. This leads to the following class of models:

$$X_1, \dots, X_n | \theta_1^*, \dots, \theta_k^*, \rho, \boldsymbol{\psi} \stackrel{ind}{\sim} p(X_i | \theta_i, \boldsymbol{\psi}), \quad \theta_i = \theta_{s_i}^* \quad (10)$$

$$\theta_1^*, \dots, \theta_k^* | \rho, \boldsymbol{\psi} \stackrel{iid}{\sim} f_0(\cdot | \boldsymbol{\psi}) \quad (11)$$

$$\rho \sim \text{a conditionally specified distribution}, \quad (12)$$

where $f_0(\cdot | \boldsymbol{\psi})$ is a known density and $\boldsymbol{\psi} \sim p(\boldsymbol{\psi})$ is a hyperparameter vector that is *a priori* independent of ρ . When the conditionally specified distribution ρ is of the product form (7), model (10)–(12) is usually referred to as *parametric* PPM in the sense that clusters are defined by a common parameter value rather than by matching data values. But (10)–(12) is more general than parametric PPMs, also including the mixture models discussed in Section 2.4.

A nonparametric alternative to (10)–(12) is

$$X_1, \dots, X_n | \theta_1, \dots, \theta_n, \boldsymbol{\psi} \stackrel{ind}{\sim} p(X_i | \theta_i, \boldsymbol{\psi}) \quad (13)$$

$$\theta_1, \dots, \theta_n | F, \boldsymbol{\psi} \stackrel{iid}{\sim} F(\cdot | \boldsymbol{\psi}) \quad (14)$$

$$F \sim SSM(F_0, p), \quad (15)$$

where $SSM(F_0, p)$ refers to the RPM associated to a species sampling model with EPPF p and non-atomic probability measure $F_0(\cdot|\boldsymbol{\psi})$ having density $f_0(\cdot|\boldsymbol{\psi})$. A typical assumption in this case is that, *a priori*, $\boldsymbol{\psi} \sim p(\boldsymbol{\psi})$ is independent of the probability model for F .

Equations (13)–(15) define a fairly general class of nonparametric models, and the discrete nature of the SSM prior allows for a variable number of clusters in a natural way. This suggests a connection between the parametric and nonparametric classes stated above. In the case of mixture models, such connection has been hinted at in Richardson and Green (1997) and in Stephens (2000). An explicit relationship in terms of PPMs and DPs was described in Quintana and Iglesias (2003). A more general connection follows from the discussion in Section 2.3: if $\theta_1, \dots, \theta_n$ are exchangeable then for every nonparametric model (13)–(15) there is a unique parametric model (10)–(12) that has the same induced prior distribution on partitions of the form (4). In fact, the parametric model is obtained from the nonparametric one by integrating out the RPM F , i.e. by marginalizing over F . Exchangeability holds, for instance, for a large portion of PPMs, namely, those where the cohesion functions $c(S_i)$ depend on S_i only through $|S_i|$. Hence, within a wide range of models there are alternative nonparametric and parametric formulations implying the same model on the partitions. The choice between these is therefore entirely dependent on the modeling interests or needs.

To fit the parametric or nonparametric models it is typically necessary to resort to simulation-based methods. It is interesting to point out that the relationship between partition structures of both types of model has turned out to be fruitful to adapt algorithms originally devised for one case to the other. Such an example concerning the DP and PPMs can be found in Quintana and Iglesias (2003). Concretely, the Gibbs sampling algorithm of Bush and MacEachern (1996) and MacEachern and Müller (1998) was adapted to the case of PPMs. More generally, the predictive formulation (rather than coping with the joint structure based on the EPPF) turns out to be useful for computational purposes. This is clearly seen in Ishwaran and

James (2003a) who develop sequential importance sampling (SIS) (Kong et al. 1994) and the Gibbs sampling for nonparametric models using general SSMs.

Implementing the SIS algorithm requires drawing under either model from a distribution like

$$p(\theta_{\ell+1}|\theta_1, \dots, \theta_\ell, X_1, \dots, X_{\ell+1}, \boldsymbol{\psi}) \propto \sum_{i=1}^{k_\ell} p(X_{\ell+1}|\theta_i^*, \boldsymbol{\psi})p(s_{\ell+1} = i|s_1, \dots, s_\ell)\delta_{\theta_i^*}(\theta_{\ell+1}) \\ + p(X_{\ell+1}|\theta_{\ell+1}, \boldsymbol{\psi})p(s_{\ell+1} = k_\ell + 1|s_1, \dots, s_\ell)f_0(\theta_{\ell+1}|\boldsymbol{\psi}). \quad (16)$$

When the likelihoods $p(X_i|\theta_i, \boldsymbol{\psi})$ and $f_0(\theta_i|\boldsymbol{\psi})$ are conjugate, evaluation of (16) is usually straightforward. A collapsed version of the algorithm can be derived as well, by integrating out the locations first and then imputing only memberships. The implied predictive structure generated by doing so can be described in terms of *generalized weighted Chinese restaurant processes* (Ishwaran and James 2003a). The same basic story applies, but now the seating probability for a new table is proportional to

$$\int p(X_{\ell+1}|\theta_{\ell+1}, \boldsymbol{\psi})p(s_{\ell+1} = k_\ell + 1|s_1, \dots, s_\ell)f_0(\theta_{\ell+1}|\boldsymbol{\psi}) d\theta_{\ell+1},$$

while the probability of joining the i th (already formed) table is proportional to

$$\frac{\int \prod_{j \in S_i(\ell) \cup \{\ell+1\}} p(X_j|\theta_i^*, \boldsymbol{\psi})f_0(\theta_i^*|\boldsymbol{\psi}) d\theta_i^*}{\int \prod_{j \in S_i(\ell)} p(X_j|\theta_i^*, \boldsymbol{\psi})f_0(\theta_i^*|\boldsymbol{\psi}) d\theta_i^*} \times p(s_{\ell+1} = i|s_1, \dots, s_\ell).$$

In other words, the prior predictive probabilities are *a posteriori* weighted by quantities that depend of the marginal distribution of the data for people seated in a table (old or new), conditional on hyperparameters. However, it is important to point out that practical difficulties arise in the nonconjugate case, due to the need of drawing from distributions that are proportional to forms like $p(X_{\ell+1}|\theta_{\ell+1}, \boldsymbol{\psi})f_0(\theta_{\ell+1}|\boldsymbol{\psi})$. In contrast, the non-exchangeable conjugate case requires no special step when dealing with model (10)–(12). See additional details about the SIS algorithm for nonparametric problems in Liu (1996), MacEachern, Clyde and Liu (1999), Quintana and Newton (2000), Ishwaran and James (2003a) and references therein.

Likewise, the Gibbs sampling can be implemented by drawing from distributions similar to (16). Specifically, let $\boldsymbol{\theta}_{-\ell}$ denote the vector $\boldsymbol{\theta}$ without the ℓ th coordinate,

and let the k_n^- be the number of elements in the partition that corresponds to $\boldsymbol{\theta}_{-\ell}$. Two cases arise. If $s_\ell = i$ and $m_{i,n} > 1$ then $(\theta_1^*, \dots, \theta_{k_n}^*) \equiv (\theta_1^{*-}, \dots, \theta_{k_n}^{*-})$, $m_{i,n}^- = m_{i,n} - 1$, and $m_{j,n}^- = m_{j,n}$ for $j \neq i$. In contrast, if $m_{i,n} = 1$ then $k_n^- = k_n - 1$ and one cluster disappears, so define $(\theta_j^{*-}, m_{j,n}^-) = (\theta_j^*, m_{j,n})$ for $1 \leq j \leq i - 1$ and $(\theta_j^{*-}, m_{j,n}^-) = (\theta_{j+1}^*, m_{j+1,n})$ for $i \leq j \leq k_n^-$. Then

$$p(\theta_\ell | \boldsymbol{\theta}_{-\ell}, X_1, \dots, X_n, \boldsymbol{\psi}) \propto \sum_{i=1}^{k_n^-} p(X_\ell | \theta_i^{*-}, \boldsymbol{\psi}) h(\rho^-, \ell, i) \delta_{\theta_i^*}(\theta_{\ell+1}) \\ + p(X_{\ell+1} | \theta_{\ell+1}, \boldsymbol{\psi}) h(\rho^-, \ell, k_n^- + 1) f_0(\theta_{\ell+1} | \boldsymbol{\psi}),$$

where

$$h(\rho^-, \ell, i) = \begin{cases} \frac{P(\rho^-(S_1^-, \dots, S_{i-1}^-, S_i^- \cup \{\ell\}, S_{i+1}^-, \dots, S_{k_n^-}^-))}{P(\rho^-(S_1^-, \dots, S_{k_n^-}^-))} & \text{if } 1 \leq i \leq k_n^- \\ \frac{P(\rho^-(S_1^-, \dots, S_{k_n^-}^-, \{\ell\}))}{P(\rho^-(S_1^-, \dots, S_{k_n^-}^-))} & \text{if } i = k_n^- + 1. \end{cases} \quad (17)$$

Care must be taken to ensure that no gaps are originated, and that the clusters are sorted in ascending order. In the exchangeable case, (17) equals $P(s_n = i | s_{n-1}, \dots, s_1)$ which shows the usefulness of the predictive approach. In the non-exchangeable case it is possible to adapt the ‘‘full and empty’’ clusters algorithm by MacEachern and Müller (1998). The basic idea is to carry along the clusters that are currently imputed together with a number of empty clusters that are to be used if needed. By doing so, the evaluation of integrals like $\int p(X_i | \theta_i, \boldsymbol{\psi}) f_0(\theta_i | \boldsymbol{\psi})$ is completely avoided. A somewhat different algorithm based on the introduction of latent variables to facilitate Gibbs sampling is given in Walker and Damien (1996), which is in turn related to ideas presented in Damien, Wakefield and Walker (1997). Computational approaches based on Metropolis-Hastings moves for the case of DPs can be found in Neal (2000) and in Jain and Neal (2004).

4 Clustering

Each of the models we have described provides a certain posterior distribution on partitions of S . Posterior inference in these models can then be used to provide

clustering procedures that select a single partition according to some given criterion. The MBC approach considers the mixture model (9), and model fitting is based on the EM algorithm, which implies maximum likelihood estimation of the θ parameters. Therefore, MBC involves selecting *both* the model and the number of components. To do so, the number of clusters is treated as fixed and known, although different values are tried and later compared using a combination of Bayes factors and the Bayesian Information Criterion (BIC) proposed by Schwarz (1978). Thus, the partition that maximizes the likelihood is determined for several or all the values of K . The selected partition is the one that corresponds to the number of mixture components that attains the highest BIC value. See further details in Fraley and Raftery (1998, 2002).

Unfortunately, there does not seem to be an easy way to extend the MBC methodology to the more general forms of (10)–(12) or (13)–(15). Nevertheless, alternative procedures can be used to construct clusters of experimental units, for instance, by finding the MAP partition that corresponds to the chosen prior clustering structure. Doing so, however, leads to several practical problems. First of all, even for moderate values of n , the number of partitions is quite large, typically resulting in a very small probability at the posterior mode. Therefore, locating the posterior mode may not be really meaningful. And secondly, one is typically interested in some additional decision problem such as parameter estimation, hypothesis testing, etc.

The work by Quintana and Iglesias (2003), considers a formal decision theoretic framework to group experimental units *and* to solve the decision problem of interest in the context of PPMs. The basic idea was to consider a loss function of the form $\kappa l_1 + (1 - \kappa)|\rho|$, where $0 < \kappa < 1$ is a cost-complexity parameter and l_1 is related to the specific decision problem of interest. This choice implies a trade-off (controlled by κ) between optimally solving the decision problem (i.e., minimizing l_1) and simplicity in the partition structure (i.e., small number of clusters). The same basic idea can be extended to the more general models discussed in this work. In particular, the algorithm by Quintana and Iglesias (2003) can be applied to models like (10) – (12) or (13) – (15) with no special change.

5 Data Illustration: Finding Outliers in Linear Regression Models

Consider a standard linear regression model $y_i \sim N(x_i^T \theta, \sigma^2)$ for which the interest is on parameter estimation and outlier detection. The idea of using a finite mixture model to identify outliers has been discussed by several authors. Earlier references include Aitkin and Tunnicliffe Wilson (1980), and Jorgensen (1990). More recently, a number of approaches have been discussed in the statistical literature. Scott (2001) studies minimum distance estimation, rather than using the EM algorithm, in the context of an incomplete mixture model, applying the method to the identification of clusters of outliers and “good” data points (non-outliers), as well as parameter estimation. The number of components needs to be specified beforehand though. Hennig (2002, 2003) considers a method based on fixed point clusters (FPC), i.e. subsets of data points with the property that every point in the FPC is a non-outlier with respect to its own set of parameter estimators, thus splitting the data set in two parts. FPCs need not be unique, and different FPCs may overlap. Hennig (2003) presents a fixed point-type iterative algorithm that finds “substantial” FPCs as defined according to least squares estimators.

We illustrate here the methods described in earlier Sections in the context of the above problem. To do so, two strategies are applied as discussed in the upcoming two subsections.

5.1 Model-Based Clustering for Regression Models

First, we discuss application of the MBC method to this problem. Assume that for a given K

$$p(y_1, \dots, y_n) \equiv l(\{y_i\} | \{\tau_{K,j}\}, \{\theta_j\}, \beta) = \prod_{i=1}^n \sum_{j=1}^K \tau_{K,j} \frac{1}{\sigma} \phi \left(\frac{y_i - \theta_j - x_i^T \beta}{\sigma} \right), \quad (18)$$

i.e., a mixture of regression models with common covariates and regression coefficients β but different intercepts $\{\theta_j\}$. The main idea of this modeling strategy is that

outlying points may form a separated cluster from the main body of data points. In fact, several clusters can be formed with different subsets of outliers, which are represented by the various components of the mixture (18). These atypical points are then identified by assessing the magnitude and sign of the corresponding residual. This is an extension of the model discussed in Quintana and Iglesias (2003).

The computational strategy followed here is readily derived from the method explained in Dasgupta and Raftery (1998). For each fixed K the mixture density (18) is maximized by means of the EM algorithm (Dempster, Laird and Rubin 1977). The trick consists of introducing latent cluster indicators $Z_{ij} = 1$ if the i th observation belongs to the j th cluster and $Z_{ij} = 0$ otherwise. The complete data log-likelihood then becomes

$$l(\{y_i\}, \{Z_{ij}\} | \{\tau_{K,j}\}, \{\theta_j\}, \beta) = \sum_{i=1}^n \sum_{j=1}^K Z_{ij} \left(\log(\tau_{K,j}) - \frac{1}{2\sigma^2} (y_i - \theta_j - x_i^T \beta)^2 - \frac{\log(\sigma^2)}{2} \right).$$

The E step then consists of computing, for $i = 1, \dots, n$ and $j = 1, \dots, K$,

$$\hat{Z}_{ij} = E(Z_{ij} | \{y_i\}, \{\theta_j\}, \beta, \{\tau_{K,j}\}) = \frac{\tau_{K,j} \exp\{-\frac{1}{2\sigma^2}(y_i - \theta_j - x_i^t \beta)^2\}}{\sum_{\ell=1}^K \tau_{K,\ell} \exp\{-\frac{1}{2\sigma^2}(y_i - \theta_\ell - x_i^t \beta)^2\}},$$

i.e., the estimated posterior probability that the i th observation belongs to the j th cluster. In the M step the expected complete data likelihood

$$l^*(\{y_i\}, \{\hat{Z}_{ij}\} | \{\tau_{K,j}\}, \{\theta_j\}, \beta) = \sum_{i=1}^n \sum_{j=1}^K \hat{Z}_{ij} \left(\log(\tau_{K,j}) - \frac{1}{2\sigma^2} (y_i - \theta_j - x_i^T \beta)^2 - \frac{\log(\sigma^2)}{2} \right)$$

is maximized. This reduces to computing $\hat{\tau}_{K,j} = \frac{1}{n} \sum_{i=1}^n \hat{Z}_{ij}$ and solving the weighted least squares problem

$$\min_{\boldsymbol{\eta}} \sum_{i=1}^n \sum_{j=1}^K (y_i - \theta_j - x_i^T \beta)^2 \hat{Z}_{ij},$$

where $\boldsymbol{\eta} = (\theta_1, \dots, \theta_k, \beta)$. This leads to values $\{\hat{\theta}_j\}$ and $\hat{\beta}$, from which $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^K \hat{Z}_{ij} (y_i - \hat{\theta}_j - x_i^T \hat{\beta})^2$ is readily computed. The two steps just described are

repeated until convergence is achieved. Based on a suggestion made by Dasgupta and Raftery (1998), a convenient starting partition for the iterative procedure is obtained by applying an agglomerative clustering algorithm with K clusters to the residuals of the regression model $y_i \sim N(\theta + x_i^T \beta, \sigma^2)$. An alternative initialization method, which works well in practice for problems of this type, is the k -means algorithm (Hartigan and Wong 1979), also applied to the regression residuals.

For each K a partition is thus obtained simply by identifying, after convergence, the cluster component with highest \hat{Z}_{ij} value for each i . To compare all such partitions, Dasgupta and Raftery (1998) propose computing the value of

$$\text{BIC} = 2l(\{y_i\}|\{\hat{\tau}_{K,j}\}, \{\hat{\theta}_j\}, \hat{\beta}) - (2K + p) \log(n), \quad (19)$$

where p is the number of covariates in the design matrix \mathbf{X} (which excludes the intercept term) and $2K + p$ is the total number of parameters to be estimated in (18). The number of mixture components (and partition) with highest BIC value is then selected as the algorithm output. See further details about MBC in Fraley and Raftery (2002), and more generally, about the EM algorithm applied to mixtures in Redner and Walker (1984) and in McLachlan and Peel (2000).

5.2 Dirichlet Process Based Clustering

Assume now the following model

$$\begin{aligned} y_1, \dots, y_n | \theta_1^*, \dots, \theta_k^*, \beta, \rho, \sigma^2 &\stackrel{ind}{\sim} N(\theta_i + x_i^T \beta, \sigma^2), & \theta_i = \theta_{s_i}^* \\ \theta_1^*, \dots, \theta_k^* | \rho, \beta, \sigma^2 &\stackrel{iid}{\sim} N(\theta_0, \tau_0^2 \sigma^2) \\ \beta &\sim N(\beta_0, \sigma^2 \mathbf{B}_0) \\ \sigma^2 &\sim IG(\nu_0, \lambda_0) \\ \rho &\sim \text{product distribution with } c(S_i) = c \times (|S_i| - 1)!, \end{aligned} \quad (20)$$

where *a priori*, ρ and σ^2 , are independent, $\nu_0, \lambda_0, \tau_0^2, c > 0$ are known constants and \mathbf{B}_0 is a known positive definite matrix. This extends the model analyzed in Quintana

and Iglesias (2003) to the multivariate case. As in the model discussed in Section 5.1, clusters are related to the regression residuals. Note that the prior predictive probabilities (8) coincide in this case with those stated in (1) and obtained for DP priors. However, only a parametric version based on PPMs is assumed here, as inferences on the distribution of the intercept coefficients $\{\theta_i\}$ is not sought after in this specific application. The above framework can be argued to be a reasonable choice for the purpose of identifying outlying data points, as the prior structure favors the formation of a relatively small number of clusters.

Let $\boldsymbol{\eta} = (\boldsymbol{\theta}, \beta, \sigma^2)$, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ is a vector of regression intercepts, and $\boldsymbol{\eta}_\rho$ is the version of $\boldsymbol{\eta}$ that corresponds to the specific partition ρ . For instance, $\boldsymbol{\theta}_\rho = (\theta_{\rho,1}, \dots, \theta_{\rho,n})$, where for every subset $S_j \in \rho$ the $\theta_{\rho,i}$ with $i \in S_j$ are all identical. The algorithm in Quintana and Iglesias (2003) is next used to select the partition according to the loss function

$$L(M_\rho, \boldsymbol{\eta}_\rho, \boldsymbol{\eta}) = \frac{\kappa_1}{n} \|\boldsymbol{\theta} - \boldsymbol{\theta}_\rho\|^2 + \kappa_2 (\sigma^2 - \sigma_\rho^2)^2 + \frac{\kappa_3 \|\beta - \beta_\rho\|^2}{p} + (1 - \kappa_1 - \kappa_2 - \kappa_3) |\rho|, \quad (21)$$

where M_ρ represents the model obtained when fixing ρ in (20), $|\rho|$ is the number of subsets in ρ , and $\|\cdot\|$ is the Euclidean norm in the appropriate space. Here, κ_i , $i = 1, 2, 3$ are constants that control the relative weights of the various components of the loss function (21), with $0 < \kappa_i < 1$ and $\kappa_1 + \kappa_2 + \kappa_3 < 1$. Compared to the setup used by Quintana and Iglesias (2003), this loss function allows for possibly different weights given to the components related to regression coefficients β and variance of errors σ^2 .

The algorithm uses the Bayes estimate $\hat{\boldsymbol{\eta}}_B = E(\boldsymbol{\eta}|\mathbf{y})$ of $\boldsymbol{\eta}$, which is required as an input, where $\mathbf{y} = (y_1, \dots, y_n)$ represents the observed responses. It is readily shown that the optimal partition minimizes

$$SC_{\kappa_1, \kappa_2, \kappa_3}(\rho) = \frac{\kappa_1}{n} \|\hat{\boldsymbol{\theta}}_B - \hat{\boldsymbol{\theta}}_\rho\|^2 + \kappa_2 (\hat{\sigma}_B^2 - \hat{\sigma}_\rho^2)^2 + \frac{\kappa_3 \|\hat{\beta}_B - \hat{\beta}_\rho\|^2}{p} + (1 - \kappa_1 - \kappa_2 - \kappa_3) |\rho|,$$

where $\hat{\boldsymbol{\eta}}_\rho = (\hat{\boldsymbol{\theta}}_\rho, \hat{\sigma}_\rho^2, \hat{\beta}_\rho) = E(\boldsymbol{\eta}|\rho, \mathbf{y})$ is the Bayes estimate when the partition ρ is fixed in model (20). Starting with $\rho = \{\{1, \dots, n\}\}$, the most outlying element is

detached from the initial group and the several partitions obtained by adding this element to already existing clusters or by forming a new cluster (i.e. a singleton) are assessed. The partition with the lowest value of $SC_{\kappa_1, \kappa_2, \kappa_3}(\rho)$ is selected at this stage, and the process continues if such a partition improves the one selected from the previous stage (in the $SC_{\kappa_1, \kappa_2, \kappa_3}(\rho)$ sense). The process continues until no further improvement is possible. See further details in Quintana and Iglesias (2003).

Finally, it is interesting to point out that the $(1 - \kappa_1 - \kappa_2 - \kappa_3)$ factor in (21) implies that the optimal partition will have a limited (small to moderate) number of clusters, depending on the magnitude of this factor. In that sense, the above algorithm is particularly useful to detect partitions where only a small portion of the data are “atypical”, as is usually the case in outlier detection in the context of linear regression models.

5.3 Numerical Results

In this section the models and methods discussed in Sections 5.1 and 5.2 will be applied to two datasets. The analysis is carried out to the extent needed for an adequate comparison of outlier detection in regression models via clustering algorithms as described earlier. The first dataset is referred to as the *regression dilemma* and presented in Hocking (1996). In turn, this is a modification of the example discussed in Hocking and Pendleton (1983). We consider here two versions of this problem, i.e., the dataset with and without the 27th observation (for a complete discussion and analysis of these data, as well as the reason why two versions are considered, see Hocking 1996). The second example consists of data on personal savings ratio (defined as the aggregate personal saving divided by disposable income) for the 1960-1970 period on 50 countries. The data are available on-line in the R system by typing `help(LifeCycleSavings)` at the R prompt, and have been extensively analyzed in Belsley, Kuh and Welsch (1980).

The methods discussed in Section 5.1 and 5.2 were then applied to these two examples, using $c = 1$ in all cases. For model (20) the hyperparameters were chosen

to be $\beta_0 = 0$, $\theta_0 = 0$, $\mathbf{B}_0 = 10000 \times \mathbf{I}$, $\nu_0 = 2.01$, $\lambda_0 = 1.01$, and $\tau_0^2 = 10000$ thus reflecting non-informative prior distributions. Following the discussion in Quintana and Iglesias (2003), the weights in (21) should be chosen such that κ_1 and $1 - \kappa_1 - \kappa_2 - \kappa_3$ are small compared to κ_2 and κ_3 . For the purpose of this clustering application, this is accomplished by setting $\kappa_1 = 0.001$, $\kappa_2 = 0.098$ and $\kappa_3 = 0.9$, which gives the largest weight to the most important decision problem, i.e., estimation of β , while assigning weight 0.001 to the model complexity term.

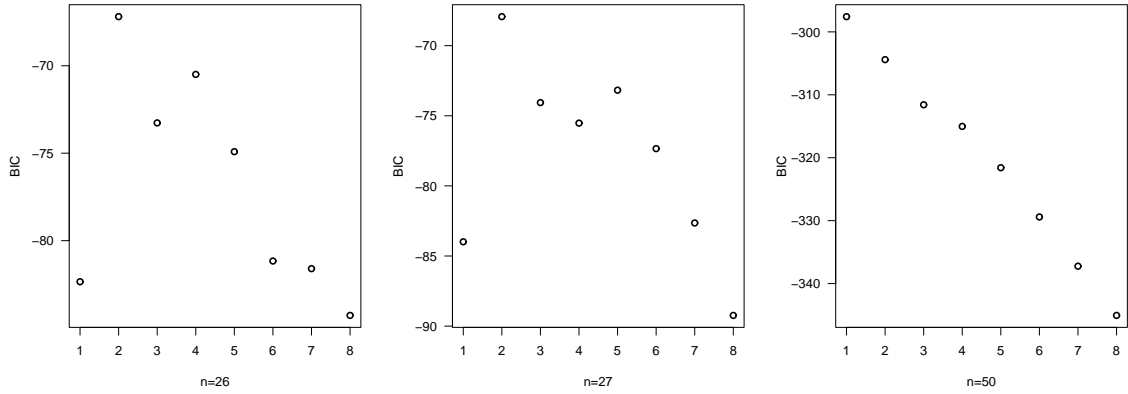


Figure 1: Values of the BIC criterion (19) for the dilemma (26 and 27 data points) and for the savings ratio datasets. Horizontal axes represent the number of mixture components K in (18), chosen in this case to range from 1 through 8.

Figure 1 shows the values of BIC computed according to (19) for the two versions of the dilemma dataset and for the savings ratio example. In both versions of the dilemma example, the selected number of clusters is 2, where one cluster consists only of observation 17 and the other cluster groups the remaining observation. This has a clear interpretation: observation 17 is an outlier, regardless of the presence or absence of the last data point. On the other hand, for the savings ratio example the model selected corresponds to only 1 cluster, which means that no outlier is detected, i.e., all data points fall into the same category.

Figure 2 shows the posterior means of the intercept coefficients θ_i , $i = 1, \dots, n$ that are obtained after fitting model (20) to the examples, using standard DP Gibbs

sampling, e.g. as described in Section 3. Because the outlier candidates are chosen according to the values of these posterior means, the plots illustrate how the partitions to be analyzed will be formed. Of course, the final decision depends also on the remaining components of (21). In this case, there is a clear indication that observation 17 is an outlier for the dilemma dataset. Also, when including observation 27, all posterior means have a small decrease which is of about the same magnitude in all cases except for observation 24. This is consistent with the findings in Hocking (1996). For the second example, observation 46 appears as the farthest from the main body of data points.

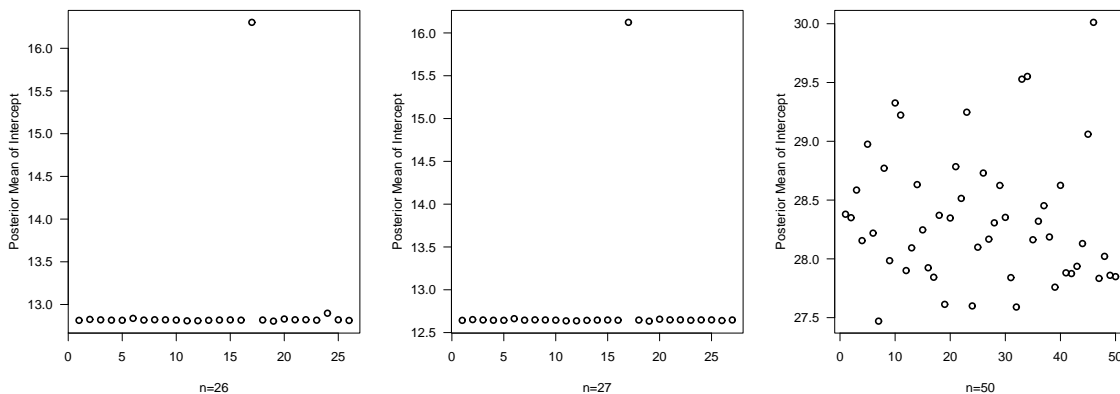


Figure 2: *Posterior means of intercept coefficients for the dilemma (26 and 27 data points) and for the savings ratio datasets. Horizontal axes represent the observation index.*

When applying the Algorithm to the examples, the results shown in Table 1 are obtained. For the dilemma dataset (with or without the 27th observation) the conclusion is the same as in the case of the MBC method. However, for the savings ratio problem the Algorithm declares observation 46 to be an outlier, which is different from the results of MBC. It is worth noting that the EM iterations for MBC when fixing the number of clusters at 2 converged to the partition where one of the subsets is $\{10, 11, 23, 33, 34, 46\}$ (as a comparison, $SC_{\kappa_1, \kappa_2, \kappa_3}$ for this partition is 1.4592). These are exactly the 6 most outlying points in the sense of distance to the overall mean

of intercepts. However, the increased likelihood with respect to the model with a single cluster does not compensate for the model complexity penalty reflected in the definition of BIC. In contrast, the flexibility inherent to model (20) and the step-by-step nature of the algorithm allow the user to try simpler models that in some examples may perform better than the MLE with respect to the loss function adopted.

Dataset	n	$\hat{\rho}$	$SC_{\kappa_1, \kappa_2, \kappa_3}(\hat{\rho})$
Dilemma	26	$\{S - \{17\}, \{17\}\}$	0.00214
Dilemma	27	$\{S - \{17\}, \{17\}\}$	0.00200
Savings Ratio	50	$\{S - \{46\}, \{46\}\}$	0.02736

Table 1: *Best partitions determined by the Algorithm for the dilemma (26 and 27 data points) and for the savings ratio datasets, with $S = \{1, \dots, n\}$.*

The weight parameters κ_i , $i = 1, 2, 3$ do not affect the way the outlier candidates are picked, but they do play a key role in the final partition selected. It is then natural to wonder how sensitive the results are to the selection of these values. Such assessment necessarily depends on the specific application. For the savings ratio example, the same partitions are obtained if we moved them to $\kappa_1 = 0.05$, $\kappa_2 = 0.1$ and $\kappa_3 = 0.8$ (data not shown), which suggests the procedure is robust to this important component of the decision problem.

As a final comparison, the above calculations were repeated using now cohesions given by $c(S_i) = c$. This implies

$$p(s_{j+1} = i | s_j, \dots, s_2, s_1) = \begin{cases} \frac{1}{c+k_j} & \text{if } 1 \leq i \leq k_j \\ \frac{c}{c+k_j} & \text{if } i = k_j + 1. \end{cases}$$

Compared to the DP-style predictive probabilities, this structure favors the formation of a large number of small clusters. Although the posterior distributions for some specific parameters are indeed somewhat different (data not shown), we obtained the same partitions reported in Table 1.

6 Data Illustration: Density Estimation

Consider now a more traditional application of mixture models, namely, the estimation of a (univariate) density function. In other words, the purpose of this application is the determination of the appropriate number of mixture components as well as the estimation of their corresponding parameters.

More generally, the problem of density estimation arises when data y_1, \dots, y_n are considered to be sampled from a certain distribution G but G is itself assumed unknown. Such problems (and their multiple variations) arise in nearly all branches of Statistics, as evidenced by the numerous references available. Nonparametric methods are specially well suited for density estimation, kernel-based methods being a popular choice. References can be found, among others, in Silverman (1986), Scott (1992), Devroye and Lugos (2001) and from a nonparametric Bayesian viewpoint, in Müller and Quintana (2004).

The focus of this application will be on mixtures of normals, but other extensions such as mixtures of t distributions (e.g. Stephens 2000) can be easily implemented. As in the case of the example discussed in Section 5, the main motivation behind this illustration is to compare the algorithms described in Section 4 and the partitions they select.

6.1 Model-Based Clustering for Density Estimation

Assume that for a given K

$$p(y_1, \dots, y_n) \equiv l(\{y_i\} | \{\tau_{K,j}\}, \{\mu_j\}, \{\sigma_j^2\}) = \prod_{i=1}^n \sum_{j=1}^K \tau_{K,j} \frac{1}{\sigma_j} \phi\left(\frac{y_i - \mu_j}{\sigma_j}\right), \quad (22)$$

i.e., a mixture of normal densities with component-specific means and variances. This is a special case of the model discussed in Dasgupta and Raftery (1998). Thinking intuitively in terms of histograms, the idea is to locate mixture components at places where the observed data lump together.

To fit (22) the EM algorithm is a convenient choice. As in Section 5, the key idea

is the introduction of latent cluster indicators $Z_{ij} = 1$ if the i th observation comes from the j th normal component and $Z_{ij} = 0$ otherwise. Thus, for a fixed K , the E step is accomplished by computing

$$\hat{Z}_{ij} = \frac{\tau_{K,j} \exp\{-\frac{1}{2\sigma_j^2}(y_i - \mu_j)^2\}}{\sum_{\ell=1}^K \tau_{K,\ell} \exp\{-\frac{1}{2\sigma_\ell^2}(y_i - \mu_\ell)^2\}}, \quad i = 1, \dots, n; j = 1, \dots, K$$

while the M step consists of evaluating

$$\hat{\tau}_{K,j} = \frac{1}{n} \sum_{i=1}^n \hat{Z}_{ij}, \quad \hat{\mu}_j = \frac{\sum_{i=1}^n \hat{Z}_{ij} y_i}{\sum_{i=1}^n \hat{Z}_{ij}}, \quad \text{and} \quad \hat{\sigma}_j^2 = \frac{\sum_{i=1}^n \hat{Z}_{ij} (y_i - \hat{\mu}_j)^2}{\sum_{i=1}^n \hat{Z}_{ij}},$$

for $j = 1, \dots, K$. These steps are iterated until convergence is reached. Finally, the number of mixtures components is obtained by determining K for which the highest BIC is attained. This implies equal number of estimated means $\hat{\mu}_j$ and variances $\hat{\sigma}_j^2$, together with the clustering constructed by identifying, for each observation, the index ℓ maximizing \hat{Z}_{ij} for $j = 1, \dots, \ell$.

The calculations just described can be implemented using the MCLUST package by Fraley and Raftery (2002b), which can be freely downloaded from the following URL address: <http://www.stat.washington.edu/mclust>. In particular, this package can be readily loaded into the R system.

6.2 Dirichlet Process Model for Density Estimation

Consider now the nonparametric counterpart of the model discussed in Section 6.1. Concretely, consider the model discussed in Escobar and West (1995):

$$\begin{aligned} y_i | \mu_i, V_i &\stackrel{ind}{\sim} N(\mu_i, V_i) \\ (\mu_1, V_1), \dots, (\mu_n, V_n) | F &\stackrel{iid}{\sim} F \\ F | m, \tau &\sim DP(c, F_0(\mu, V; m, \tau)), \end{aligned} \tag{23}$$

where the baseline distribution $F_0(\mu, V; m, \tau)$ corresponds to a pair (μ, V) such that $\mu | \tau, V \sim N(m, \tau V)$, $V^{-1} \sim \Gamma(b/2, B/2)$, and *a priori* $m \sim N(a, A)$ and $\tau^{-1} \sim \Gamma(w/2, W/2)$ are independent, with known hyperparameters a, A, b, B, w and W .

It can be shown that the Bayes solution for the density estimation problem is given by the predictive density for a new observation $p(y_{n+1}|y_1, \dots, y_n)$ which can be computed using the method discussed in Escobar and West (1995), but incorporating the fix described in Bush and MacEachern (1996). The nonparametric formulation is useful, among other things, because it provides a simple interpretation to quantities such as the posterior predictive distribution, namely

$$p((\mu_{n+1}, V_{n+1}) \in H|y_1, \dots, y_n) = E(F(H)|y_1, \dots, y_n),$$

for all Borel subsets H in the appropriate space. This expression is implicitly used when evaluating $p(y_{n+1}|y_1, \dots, y_n)$. See further details in Escobar and West (1995).

The decision theoretic formulation of the clustering problem is similar to that of Section 5. Write $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$, $\mathbf{V} = (V_1, \dots, V_n)$ and $\boldsymbol{\eta} = (\boldsymbol{\mu}, \mathbf{V}, m, \tau)$. Also, $\boldsymbol{\eta}_\rho$ denotes the version of $\boldsymbol{\eta}$ that corresponds to the parametric model that results after conditioning on the given partition ρ . The algorithm in Quintana and Iglesias (2003) requires the specification of a loss function, given in this case by

$$L(M_\rho, \boldsymbol{\eta}_\rho, \boldsymbol{\eta}) = \frac{\kappa_1}{n} \|\boldsymbol{\mu} - \boldsymbol{\mu}_\rho\|^2 + \frac{\kappa_2}{n} \|\mathbf{V} - \mathbf{V}_\rho\|^2 + \kappa_3(m - m_\rho)^2 + \kappa_4(\tau - \tau_\rho)^2 + (1 - \kappa_1 - \kappa_2 - \kappa_3 - \kappa_4)|\rho|, \quad (24)$$

where M_ρ represents the model obtained when fixing ρ in (23), and the constants $0 < \kappa_i < 1$ in (24) are constrained by $\sum_{i=1}^4 \kappa_i < 1$.

Denoting $\hat{\boldsymbol{\eta}}_B = E(\boldsymbol{\eta}|\mathbf{y})$ and $\hat{\boldsymbol{\eta}}_\rho = E(\boldsymbol{\eta}|\rho, \mathbf{y})$ the Bayes and conditional Bayes (given a partition ρ) estimates of $\boldsymbol{\eta}$, the clustering problem reduces to finding ρ^* minimizing

$$SC_{\kappa_1, \kappa_2, \kappa_3, \kappa_4}(\rho) = \frac{\kappa_1}{n} \|\hat{\boldsymbol{\mu}}_B - \hat{\boldsymbol{\mu}}_\rho\|^2 + \frac{\kappa_2}{n} \|\hat{\mathbf{V}}_B - \hat{\mathbf{V}}_\rho\|^2 + \kappa_3(\hat{m} - \hat{m}_\rho)^2 + \kappa_4(\hat{\tau} - \hat{\tau}_\rho)^2 + (1 - \kappa_1 - \kappa_2 - \kappa_3 - \kappa_4)|\rho|.$$

With these elements, the algorithm in Quintana and Iglesias (2003) works as described at the end of Section 5.2, with the obvious changes required to adapt it to this context.

6.3 Numerical Results

Consider now application of the methods discussed earlier to the well-known Galaxy dataset described in Roeder (1990) and later analyzed in several works, including Escobar and West (1995), Richardson and Green (1997), Stephens (2000), and Ishwaran and James (2003a). The data consists of $n = 82$ measured velocities (in 10^3 km/s), relative to our own galaxy, of galaxies from six well-separated conic sections of the space.

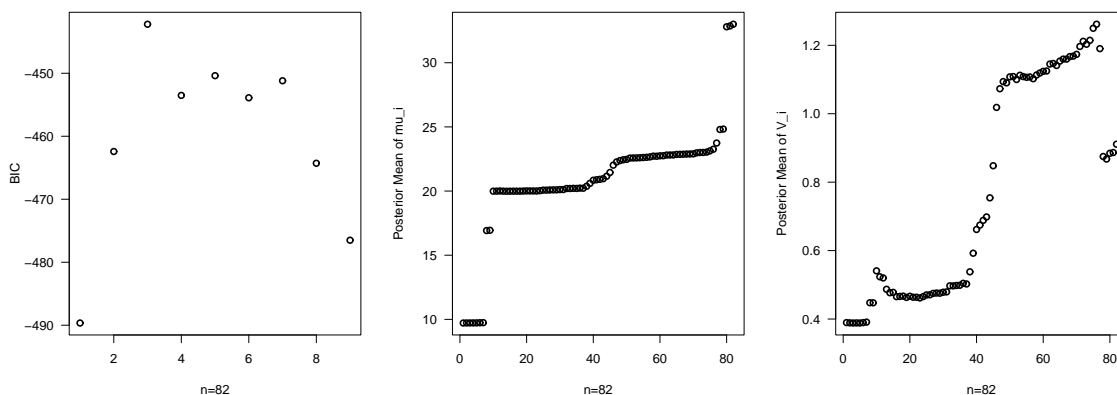


Figure 3: Values of the BIC criterion under model (22) and posterior means for μ_i and V_i under model (23) for the galaxy dataset. Horizontal axes represent number of mixture components K for the leftmost plot and observation indexes for the middle and rightmost plots.

Figure 3 shows the BIC values associated to the MBC approach for up to 9 clusters, obtained using the MCLUST package. The selected number of components (clusters) is $K = 3$, with estimated means 9.710, 21.404 and 33.044, variances 0.1785, 4.8567 and 0.8496, and with corresponding weights 0.085, 0.879 and 0.036. This suggests a large cluster that groups the central portion of the data, and one lesser cluster to each side of this. For comparison, Figure 3 also shows the estimated posterior means of μ_i and V_i that result from fitting model (23) to the same dataset. It is convenient to recall that according to the Algorithm in Quintana and Iglesias (2003), these are precisely the quantities that determine potential candidates to be separated from the

main group when forming different clusters.

Figure 4 shows the histogram of the observed data together with the fitted 3-component mixture density resulting from the MBC approach (dashed line). The graph also includes the predictive density $p(y_{n+1}|y_1, \dots, y_n)$ (solid line) obtained after fitting model (23) to the same dataset. It is interesting to observe that, according to this last curve, the large central cluster appears to be composed of three smaller ones but with centers that are very close to each other, which probably explains the results of the MBC approach.

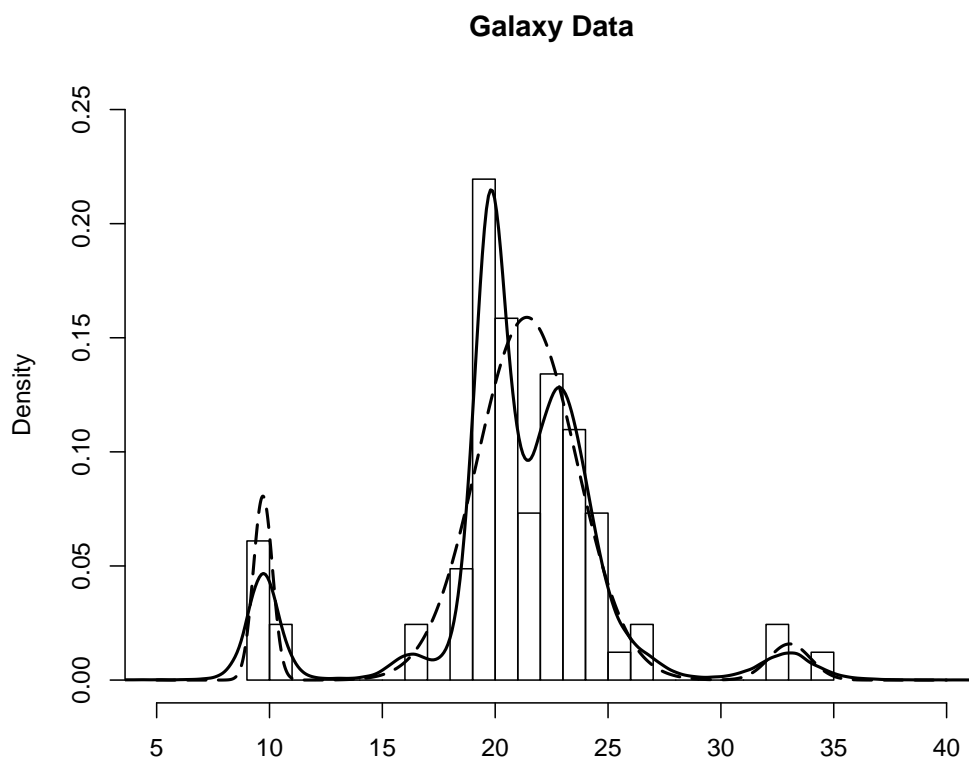


Figure 4: *Estimated density of the Galaxy Dataset, using nonparametric model (solid line) and mixture of normals (dashed line). A histogram of the observed data is included to facilitate comparison*

Table 2 shows the clusters resulting from applying the algorithm and the MBC approach to the galaxy dataset. Two applications of the algorithm are shown (lines 2 and 3 in Table 2). In the first case the cost-complexity parameters are chosen to emphasize the main target of the modeling via (23), i.e. the density estimation. Indeed, the κ_i constants give more weight to accuracy in estimating means and variances for each data point, while downweighting the role of m and τ and putting little restriction to the formation of clusters. Doing so, the optimal partition consists of 4 clusters. Looking at the predictive density in Figure 4 the fourth added cluster represents the observed data immediately to the left of the central cluster. In the second case, the parameter estimation components of (24) are given the same relative weights as before, but the model complexity part is now more relevant. Doing so results in the same partition found when using the MBC methodology.

Method	$(\kappa_1, \kappa_2, \kappa_3, \kappa_4)$	$\hat{\rho}$	$SC_{\kappa_1, \kappa_2, \kappa_3, \kappa_4}(\hat{\rho})$
MBC	–	$\{\{1 - 7\}, \{8 - 79\}, \{80 - 82\}\}$	–
Algorithm	$(\frac{10}{23}, \frac{10}{23}, \frac{1}{23}, \frac{1}{23})$	$\{\{1 - 7\}, \{8 - 9\}, \{10 - 79\}, \{80 - 82\}\}$	1.9563
Algorithm	$(\frac{10}{32}, \frac{10}{32}, \frac{1}{32}, \frac{1}{32})$	$\{\{1 - 7\}, \{8 - 79\}, \{80 - 82\}\}$	2.3581

Table 2: *Best partitions determined using the Algorithm and the MBC approach for the galaxy dataset.*

It is worth noting that other similar combinations of values for the κ_i coefficients lead to one of these two partitions. The case where $\kappa_1 = 1$ and $\kappa_i = 0$ if $i \geq 2$ leads to a partition with 5 clusters (data not shown) which is visually consistent with the histogram shape in Figure 4, but represents a totally unrealistic approach to the density estimation problem viewed from a decision theoretic standpoint.

7 Discussion

This paper explores probability models for partition structures that can be defined in a predictive (conditional) fashion. It argues that the predictive formulation may be

useful for the purpose of both, prior elicitation and posterior simulation. The models can be motivated either from a parametric or nonparametric viewpoint and for a wide range of cases the induced probability distributions on partitions are identical.

As an illustration of the potential advantages of the various models discussed, two substantially different applications were considered. The first application concerns outlier detection in linear regression models, considering two different datasets. The basic idea consists of constructing clusters of data points according to how much they deviate from the overall linear trend, i.e., based on the residuals. Two algorithms were applied in this context. The first one is an extension of the model-based clustering method of Dasgupta and Raftery (1998) while the second one extends the proposal in Quintana and Iglesias (2003). The underlying model for the former has a specific formulation that is easier to understand unconditionally, while the latter allows for a much wider variety of model specifications. In fact the two methods have a different target, which explains the differences found in the results of the second dataset described in Section 5.3. From a computational viewpoint, the MBC method is straightforward to implement and very fast to produce results. In contrast, the extension of the algorithm by Quintana and Iglesias (2003) is relatively simple to implement but it involves more computing than MBC, because partitions need to be assessed one by one until convergence. Nevertheless, in outlier detection problems, the number of partitions to be examined is typically small, which makes this algorithm an attractive alternative. Finally, it is necessary to point out that neither algorithm discussed here has been designed to produce diagnostics other than outlier detection. Methods for dealing with problems like masking or swamping or for deriving other types of diagnostics in the specific context of linear regression models are beyond the scope of this work.

The second application discussed in this paper concerns density estimation. Here, clusters are represented by different mixture components, which can be naturally interpreted under either MBC or DP-based nonparametric models. The results from Section 6 highlight that the flexibility underlying nonparametric models such as (23)

can be fruitful in a variety of scenarios. The algorithm by Quintana and Iglesias (2003) can also be used here to find a reasonably optimal partition according to (24). The flexibility does not come for free though, as using the algorithm turns out to be computationally more expensive than MBC, requiring over 30 iterations until convergence for the galaxy dataset. The application also reveals that the resulting partition and number of clusters may be, in some cases, sensitive to different choices of the κ_i parameters. Unfortunately, there does not seem to be a general rule for choosing such constants, but they rather depend on the specific decision problem to be solved.

The above discussion suggests that both algorithms may be used in tandem. Indeed, MBC may be quite fast and efficient in providing a reasonable partition in a wide variety of problems. However, if the interest is on finding the partition that minimizes loss functions such as (21) or (24), the MBC solution may not be the most appropriate, simply because it was not designed to solve decision problems of this type. In such cases, the algorithm by Quintana and Iglesias (2003) may be extended to consider the MBC partition as a starting point and proceed in the search of other potentially better partitions in the sense of minimizing expected loss.

Finally, the applications considered in this work are just examples of the general models and methods described in Sections 2 through 4. In fact, these can be applied to many more fields of interest, though the discussion of implementation details is necessarily problem-specific.

Acknowledgments

The author is indebted to Professors Joseph Kadane and Peter Müller for their valuable suggestions, and to two referees for their comments that helped to improve the final version of this manuscript.

References

- Aitkin, M. and Tunnicliffe Wilson, G. (1980). Mixture models, outliers, and the EM algorithm, *Technometrics* **22**: 325–331.
- Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *The Annals of Statistics* **2**: 1152–1174.
- Arratia, R., Barbour, A. D. and Tavaré, S. (1992). Poisson process approximations for the Ewens sampling formula, *Annals of Applied Probability* **2**: 519–535.
- Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics* **49**: 803–821.
- Barry, D. and Hartigan, J. A. (1992). Product partition models for change point problems, *The Annals of Statistics* **20**: 260–279.
- Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems, *Journal of the American Statistical Association* **88**: 309–319.
- Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics: Identifying influential data and sources of collinearity*, New York: John Wiley & Sons.
- Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes, *The Annals of Statistics* **1**: 353–355.
- Brunner, L., Chan, A., James, L. and Lo, A. Y. (2001). Weighted Chinese restaurant processes and Bayesian mixture models, Unpublished manuscript.
- Bush, C. A. and MacEachern, S. N. (1996). A semiparametric Bayesian model for randomised block designs, *Biometrika* **83**: 275–285.
- Cifarelli, D. M. and Melilli, E. (2000). Some New Results for Dirichlet Priors, *The Annals of Statistics* **28**: 1390–1413.

- Crowley, E. M. (1997). Product partition models for normal means, *Journal of the American Statistical Association* **92**: 192–198.
- Cruz, F., Loschi, R., Iglesias, P. and Arellano-Valle, R. (2003). A Gibbs Sampling Scheme to Product Partition Model: An Application to Change-Point Problems, *Computers & Operations Research* **30**: 463–482.
- Damien, P., Wakefield, J. C. and Walker, S. G. (1997). Gibbs sampling for Bayesian non-conjugate and hierarchical models using auxiliary variables, *Journal of the Royal Statistical Society, Series B* **61**: 331–344.
- Dasgupta, A. and Raftery, A. E. (1998). Detecting Features in Spatial Point Processes With Clutter via Model-Based Clustering, *Journal of the American Statistical Association* **93**: 294–302.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society Series B* **39**: 1–37.
- Devroye, L. and Lugos, G. (2001). *Combinatorial Methods in Density Estimation*, New York: Springer-Verlag.
- Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates (with discussion), *The Annals of Statistics* **14**: 1–67.
- Diaconis, P. and Kemperman, J. (1996). Some new tools for Dirichlet priors, in J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds), *Bayesian Statistics 5. Proceedings of the Fourth Valencia International Meeting*, Oxford University Press, pp. 97–106.
- Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association* **90**: 577–588.
- Ewens, W. (1972). The sampling theory of selectively neutral alleles, *Theoretical Population Biology* **3**: 87–112.

- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems, *The Annals of Statistics* **1**: 209–230.
- Fraley, C. and Raftery, A. (1998). How many clusters? Which clustering methods? Answers via model-based cluster analysis, *Computer Journal* **41**: 578–588.
- Fraley, C. and Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation, *Journal of the American Statistical Association* **97**: 611–631.
- Fraley, C. and Raftery, A. E. (2002b). MCLUST: Software for model-based clustering, density estimation and discriminant analysis, *Technical report*, Department of Statistics, University of Washington.
- Hartigan, J. A. (1990). Partition models, *Communications in Statistics, Part A – Theory and Methods* **19**: 2745–2756.
- Hartigan, J. A. and Wong, M. A. (1979). A k-means clustering algorithm, *Applied Statistics* **28**: 100–108.
- Hennig, C. (2002). Fixed point clusters for linear regression: computation and comparison, *Journal of Classification* **19**: 249–276.
- Hennig, C. (2003). Clusters, Outliers, and Regression: Fixed Point Clusters, *Journal of Multivariate Analysis* **86**: 183–212.
- Hocking, R. R. (1996). *Methods and applications of linear models: Regression and the analysis of variance*, New York: John Wiley & Sons.
- Hocking, R. R. and Pendleton, O. J. (1983). The regression dilemma, *Communications in Statistics Series A* **12**: 497–527.
- Ishwaran, H. and James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors, *Journal of the American Statistical Association* **96**: 161–173.

- Ishwaran, H. and James, L. F. (2003a). Generalized weighted Chinese restaurant processes for species sampling mixture models, *Statistica Sinica* **13**: 1211–1235.
- Ishwaran, H. and James, L. F. (2003b). Some further developments for Stick-Breaking priors: finite and infinite clustering and classification, *Sankhya, Series A* **65**: 577–592.
- Jain, S. and Neal, R. M. (2004). A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model, *Journal of Computational and Graphical and Statistics* **13**: 158–182.
- Jorgensen, M. (1990). Influence-based diagnostics for finite mixture models, *Biometrics* **46**: 1047–1058.
- Kong, A., Liu, J. S. and Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems, *Journal of the American Statistical Association* **89**: 278–288.
- Korwar, R. M. and Hollander, M. (1973). Contributions to the theory of Dirichlet processes, *The Annals of Probability* **1**: 705–711.
- Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations, *The 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, *Technical report*, Hong Kong University of Science and Technology, Department of Information and Systems Management.
- Loschi, R. and Cruz, F. (2002). An analysis of the influence of some prior specifications in the identification of change points via product partition model, *Computational Statistics and Data Analysis* **39**: 477–501.

- MacEachern, S. N. and Müller, P. (1998). Estimating Mixture of Dirichlet Process Models, *Journal of Computational and Graphical Statistics* **7**(2): 223–338.
- MacEachern, S. N., Clyde, M. and Liu, J. S. (1999). Sequential Importance Sampling for Nonparametric Bayes Models: The Next Generation, *Canadian Journal of Statistics* **27**: 251–267.
- McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*, New York: Marcel Dekker.
- McLachlan, G. J. and Peel, D. (2000). *Finite Mixture models*, New York: Wiley.
- Muliere, P. and Secchi, P. (1995). A note on a proper Bayesian Bootstrap, *Technical Report 18*, Università degli Studi di Pavia, Dipartimento di Economia Politica e Metodi Quantitativi.
- Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian Data Analysis, *Statistical Science* **19**(1): 95–110.
- Murtagh, F. and Raftery, A. E. (1984). Fitting straight lines to point patterns, *Pattern Recognition* **17**: 479–483.
- Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models, *Journal of Computational and Graphical Statistics* **9**: 249–265.
- Pitman, J. (1996). Some Developments of the Blackwell-MacQueen Urn Scheme, in T. S. Ferguson, L. S. Shapley and J. B. MacQueen (eds), *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, Hayward, California: IMS Lecture Notes - Monograph Series, pp. 245–268.
- 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.

- Quintana, F. A. and Newton, M. A. (2000). Computational aspects of Nonparametric Bayesian analysis with applications to the modeling of multiple binary sequences, *Journal of Computational and Graphical Statistics* **9**(4): 711–737.
- Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review* **26**: 195–202.
- Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion), *Journal of the Royal Statistical Society, Series B* **59**: 731–792.
- Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies, *Journal of the American Statistical Association* **85**: 617–624.
- Rolin, J.-M. (1992). Some useful properties of the Dirichlet process, *Technical Report 9207*, Center for Operations Research & Econometrics, Université Catholique de Louvain.
- Schwarz, G. (1978). Estimating the Dimension of a Model, *The Annals of Statistics* **6**: 461–464.
- Scott, D. (2001). Parametric Statistical Modeling by Minimum Integrated Square Error, *Technometrics* **43**: 274–285.
- Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: Wiley.
- Sethuraman, J. (1994). A constructive definition of dirichlet priors, *Statistica Sinica* **4**: 639–650.
- Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*, London: Chapman & Hall/CRC.

- Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components - alternative to reversible jump methods, *Annals of Statistics* **28**: 40–74.
- Walker, S. G. and Damien, P. (1996). Sampling a Dirichlet process mixture model, *Technical report*, Business School, University of Michigan.
- Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. and Ruzzo, L. (2001). Model-Based Clustering and Data Transformations for Gene Expression Data, *Bioinformatics* **17**(10): 977–987.