Bayesian Analysis (2014) **9**, Number 3, pp. 613–658

Cluster Analysis, Model Selection, and Prior Distributions on Models

George Casella * Elías Moreno † and F. Javier Girón ‡

Abstract. Clustering is an important and challenging statistical problem for which there is an extensive literature. Modeling approaches include mixture models and product partition models. Here we develop a product partition model and a Bayesian model selection procedure based on Bayes factors from intrinsic priors. We also find that the choice of the prior on model space is of utmost importance, almost overshadowing the other parts of the clustering problem, and we examine the behavior of the model posterior probabilities based on different model space priors. We find, somewhat surprisingly, that procedures based on the often-used uniform prior (in which all models are given the same prior probability) lead to inconsistent model selection procedures. We examine other priors, and find that the Ewens-Pitman prior and a new prior, *the hierarchical uniform prior*, lead to consistent model selection procedures and have other desirable properties. Lastly, we compare the procedures on a range of examples.

Keywords: Bayesian model selection, Consistency, Hierarchical models, Intrinsic priors, Product partition models, Stochastic search

1 Introduction

Suppose that Y is an observable random variable with sampling distribution in a class of parametric densities $\mathfrak{F} = \{f(y|\theta), \theta \in \Theta\}$, where Θ is in the space $\mathbb{R}^k$, $k \geq 1$, and we observe a sample of n independent data $\mathbf{y} = (y_1, y_2, ..., y_n)$ coming from the densities in the class $\mathfrak{F}$. An interesting problem for its wide range of applications, and theoretical challenges, arises when we look at the sample as being split into clusters, where all of the observations in a cluster come from the same sample density $f(y|\theta)$, and the parameter θ of the density changes across clusters. The clustering problem consists of making inference on the number of clusters in the sample and the location of the sample observations into these clusters.

Assuming that a family of sampling models $\mathfrak{F}$ has been chosen, two difficulties appear in the clustering problem. A first one is the assessment of the prior distribution for the discrete parameters, the number of clusters and the partitions of the sample into these clusters. There is also the assessment of the prior for the (usually) continuous parameters θ of the densities of the partitions. Because of the typical lack of substantive prior information on these parameters, we often do not have enough information for a precise formulation of the priors. This leads us to propose the use of objective priors.

*Distinguished Professor, Department of Statistics, University of Florida, Gainesville, FL 32611.
†Professor, Department of Statistics, University of Granada, 18071, Granada, Spain. emoreno@ugr.es
‡Professor, Department of Statistics, University of Málaga. Málaga, Spain. fj_giron@uma.es

 DOI:10.1214/14-BA869

A second difficulty is that of computing the very many posterior model probabilities of the sampling models involved, even when the sample size is moderate. For instance, for a sample size as small as 20, the number of possible models (partitions) is as large as 51724158235372, which makes it infeasible to compute all of the posterior model probabilities, although fortunately only a small number of models will have nonnegligible posterior probabilities. Therefore, there is a need to develop an efficient stochastic search algorithm for computing posterior probabilities for those models having nonnegligible posterior probabilities.

1.1 Background

The literature on clustering is enormous. There are many ways to approach the problem, and here we take a model-based approach that can be divided into approaches based on *mixture models* (see, for instance, Fraley and Raftery (2002) , Fraley and Raftery (2007) and references therein), and approaches based on *product partition models* (PPMs). The latter is a model-based approach introduced by Hartigan (1990) and Barry and Hartigan (1992) (see Booth et al. 2008, and references therein). In this paper the sampling models for clustering are also product partition models constructed with normal regression models, and objective priors for models and for model parameters will be introduced.

In recent clustering approaches, Dirichlet distributions and PPMs are used as a prior for the clusters in the hierarchical Bayesian framework. Quintana and Iglesias (2003) propose an algorithm for the explicit construction of clusters based on PPM-type priors for partitions of experimental units. They noted that the proposed model is quite flexible because PPMs can be used to express a wide variety of prior distributions on partitions.

Lau and Green (2007) propose a general formulation for Bayesian model-based clustering that is optimal with respect to a specified loss function. They compare the new method to some recently discussed methods involving stochastic search or hierarchical clustering under the Dirichlet process by maximizing the posterior probability. In a somewhat different approach, Booth et al. (2008) propose a stochastic search algorithm driven by a mixture of two Metropolis-Hastings algorithms, one for small scale changes to individual objects and another for large scale moves involving entire clusters.

Here, we find that the choice of the prior over the model space (or equivalently over the partitions), is of utmost importance, almost overshadowing the other parts of the clustering problem. We examine a number of priors, the Ewens-Pitman, the Jensen-Liu, the typical default uniform prior, where every model has the same prior probability, and a new prior, the *hierarchical uniform prior*. We find that the Jensen-Liu and the uniform prior lead to inconsistent Bayesian procedures, while the Ewens-Pitman and the *hierarchical uniform* prior do not suffer from this disadvantage. For the continuous parameters we use intrinsic priors, and combine all pieces to evaluate the models using their posterior probabilities.

Table 1: Clusters Found in the Galaxy Data

Roeder (1990)	at least 3, no more than 7 (Confidence set)
Richardson and Green (1997)	6 has highest posterior probability
Roeder and Wasserman (1997)	The posterior clearly supports three groups
Lau and Green (2007)	Optimal number of clusters is three
Wang and Dunson (2011)	Five clusters

1.2 An Example

As an example of our concerns, we revisit the famous galaxy data of Postman et al. (1986), first analyzed by Roeder (1990). The data consist of measurements of velocities, in 10^3 km/sec, of 82 galaxies from a survey of the Corona Borealis region. Many researchers have looked at this data set, looking for clusters. Table 1 gives a summary of some findings. All of the analyses are in consensus that there are at least 3, and no more than 7 clusters. This is also in accord with the belief of the astronomers, as Postman et al. (1986) says "The unfilled survey alone samples the galaxy distribution more directly associated with the Cor Bor supercluster (the six rich Abell clusters ..."

As a simple approach to finding clusters in the galaxy data, we do the following:

1. Evaluate a partition using the Bayesian information criterion (BIC).
2. Use a uniform prior distribution on the set of models.
3. Run a stochastic search to minimize BIC.

This seemingly reasonable procedure found 11 clusters in the data, in fact, the top ten models all contained 11 clusters (Table 2 and Figure 1). This is totally opposite to the consensus analyses, as BIC put too many clusters in the interior velocities. The question is: what caused BIC to find so many clusters, especially because BIC is known to favor small models? The answer that we found, and the culprit that causes the shift to too many clusters, is the uniform prior on model space.

1.3 Summary

The rest of the paper is organized as follows. In Section 2 we describe the structure of the clustering problem and a description of the densities of the PPMs approach taken here. In Section 3 we provide priors for the discrete parameters, the partitions and the number of clusters, and Section 4 develops intrinsic priors for the continuous parameters, illustrated with the case of linear models. The model priors are evaluated in Section 5, where we obtain our consistency and inconsistency results. We describe

Table 2: Results of the stochastic search using BIC to evaluate the partitions with a uniform prior on model space. It shows the exp of the BIC value for the top five partitions, each with 11 clusters.

Rank	# Clusters	$\exp(-BIC) \times 10^{-71}$
1	11	49.53
2	11	33.13
3	11	20.19
4	11	15.48
5	11	14.39

our search algorithm in Section 6, a variation of the biased random walk of Booth et al. (2008), and in Section 7 we illustrate the performance of our procedure on both real and simulated data. Finally, Section 8 contains a concluding discussion, and there is a technical appendix with proofs of theorems.

2 Partitions, Configurations and Posterior Probabilities

We begin by formally describing the structure of the clustering problem, defining the terms that we will be using in the subsequent analyses. We start with a sample $\mathbf{y} = (y_1, y_2, \ldots y_n)$. For given p, we define a *partition* of the sample into p clusters by the vector $\mathbf{r}_p = (r_1^{(p)}, ..., r_n^{(p)})$, where $r_i^{(p)}$, $i = 1, \ldots n$, is an integer between 1 and p denoting the cluster to which y_i is assigned.

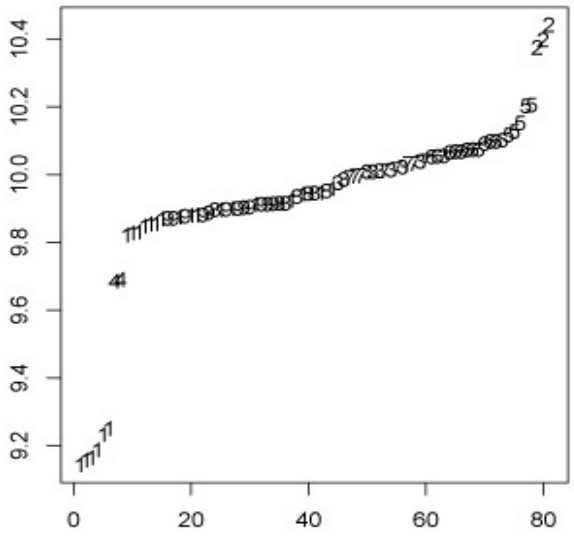
We next look at an example to clarify the definitions, and we will refer to Figure 2. Let $\mathfrak{R}_p$ represent the set of partitions of the sample into p clusters, which we call the *cluster class*. The number of partitions in $\mathfrak{R}_p$ is given by the *Stirling number of the second kind*, $\mathcal{S}(n, p)$. The four cluster classes of Figure 2 each have a Stirling number of partitions. The total number of partitions of the sample $\mathfrak{R} = \cup_p \mathfrak{R}_p$ is given by the *Bell number*, $\mathcal{B}_n = \sum_{p=1}^{n} \mathcal{S}(n, p)$, which in the case of Figure 2 is 15.

For the partitions in $\mathfrak{R}_p$ let n_i be the number of components of the sample located in the *ith* cluster for $i = 1, 2, ..., p$. Since the labels of the clusters are irrelevant, the number of ordered partitions of the sample of size n into p clusters $S(n, p)$ can be written as

$$S(n,p) = \sum_{\substack{n_1+\ldots+n_p=n \\ 1\le n_1\le\ldots\le n_p}} \binom{n}{n_1 \cdots n_p} \frac{1}{R(n_1, ..., n_p)}, \tag{1}$$

where $\binom{n}{n_1 \cdots n_p}$ is the multinomial coefficient and $R(n_1, ..., n_p) = \prod_{i=1}^{n} [\sum_{j=1}^{p} I(n_j = i)]!$ corrects the count by considering the redundant strings corresponding to the vector $(n_1, ..., n_p)$. For instance, for the vector $(n_1, ..., n_p)$ such that $n_1 = ... = n_{p-4} < n_{p-3} = n_{p-2} < n_{p-1} = n_p$, we have that $R(n_1, ..., n_p) = (p-4)!2!2!$.

Figure 1: Clustering velocities of 82 galaxies, where the x axes are the galaxy and the y axes its velocity.



Denoting the set of partitions for a fixed vector $(n_1, ..., n_p)$ by $\mathfrak{R}_{p;n_1,...,n_p}$, we can express the class $\mathfrak{R}_p$ as

$$\mathfrak{R}_p = \cup_{\substack{n_1+...+n_p=n \\ 1\le n_1\le...\le n_p}} \mathfrak{R}_{p;n_1,...,n_p},$$

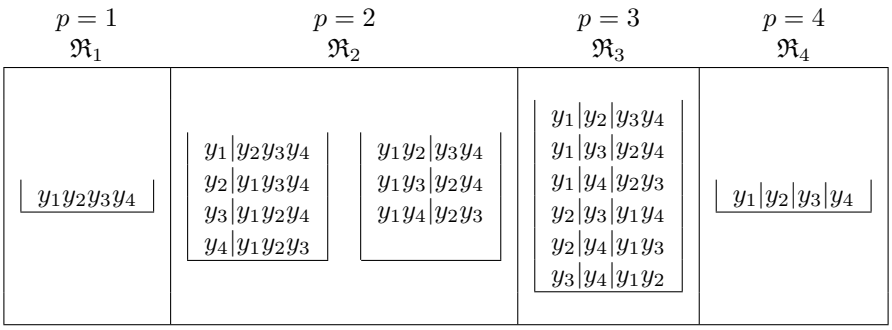
where $\mathfrak{R}_{p;n_1,...,n_p}$ denotes a *configuration class*, that is, the class of partitions in $\mathfrak{R}_p$ that have the same configuration $(n_1, ..., n_p)$. The number of partitions in a *configuration class* is given by the corresponding term in (1), that is

$$\text{Number of partitions in } \mathfrak{R}_{p;n_1,...,n_p} = \binom{n}{n_1 \cdots n_p} \frac{1}{R(n_1, ..., n_p)}. \tag{2}$$

As we see in Figure 2, a cluster class $\mathfrak{R}_p$ can have more than one configuration class; for $p = 2$ there are two configuration classes. In general, the number of configuration classes within $\mathfrak{R}_p$ is the number of ways the integer n can be partitioned into p parts, which we denote by $b(n, p)$. This number does not seem to have a closed form expression as a function of p and n. However, it can be shown (see, for instance, http://mathworld.wolfram.com/PartitionFunctionP.html) that $b(n, p)$ satisfies the recursive equation

$$b(n, p) = b(n - 1, p - 1) + b(n - p, p), \ 1 \le p \le n, \tag{3}$$

Figure 2: The structure of the clustering problem for $n = 4$. There are 15 possible partitions (models), the Bell number for $n = 4$. In each of the cluster classes, $p = 1, 2, 3, 4$ there are $1, 7, 6, 1$ partitions, the Stirling numbers of the second kind. Within the cluster class for $p = 2$ there are two configuration classes, corresponding to the configurations $y|yyy$ and $yy|yy$. The number of partitions in each configuration class is given in (2), and the number of configuration classes in each cluster class is $b(n, p)$ of (3).



with $b(n, 1) = 1$, and $b(n, n) = 1$. The number $b(n, p)$ can be large, even for moderate values of n and p, for instance $b(80, 35) = 89037$. However, it is much smaller than $S(80, 35)$, which has 82 digits.

Given a partition $\mathbf{r}_p = (r_1^{(p)}, \ldots, r_n^{(p)})$ the sampling density of the data $\mathbf{y}$ conditional on a given partition $\mathbf{r}_p$ is

$$f(\mathbf{y}|p, \mathbf{r}_p, \theta_p) = \prod_{j=1}^{p} \prod_{i:r_i=j} f(y_i|\theta_j) = \prod_{i=1}^{n} f(y_i|\theta_{r_i^{(p)}}), \tag{4}$$

where $\theta_p = (\theta_{r_1^{(p)}}, ..., \theta_{r_{p'}^{(p)}})$ is an unknown parameter and the components $\theta_{r_i^{(p)}}$ indicate the density in $\mathfrak{F}$ where the data y_i comes from. We will suppress the superscript (p) in $r_i^{(p)}$ if there is no confusion, and the likelihood will be simply written as $\prod_{i=1}^{n} f(y_i|\theta_{r_i})$.

The partition $\mathbf{r}_1 = (1, 1, ..., 1)$ corresponds to the case where there is only one cluster in the sample. Its corresponding likelihood function is given by $f(\mathbf{y}|1, \mathbf{r}_1, \theta) = \prod_{i=1}^{n} f(y_i|\theta),\ \theta \in \Theta$.

To complete the specification of the models, we need a prior distribution π for the parameters $p, \mathbf{r}_p, \theta_p$ so the generic Bayesian model $M_{\mathbf{r}_p}$ is given as $\{f(\mathbf{y}|p, \mathbf{r}_p, \theta_p), \pi(p, \mathbf{r}_p, \theta_p|n)\}$. A natural decomposition of this prior distribution is $\pi(p, \mathbf{r}_p, \theta_p|n) = \pi(\theta_p|p, \mathbf{r}_p, n)\ \pi(p, \mathbf{r}_p|n)$. The posterior probability of the configuration $\mathbf{r}_p$ in the class $\mathfrak{R}_p$, conditional on p clusters, is given by

$$\pi(\mathbf{r}_p|\mathbf{y}, p, n) = \frac{\pi(\mathbf{r}_p|p, n)m(\mathbf{y}|\mathbf{r}_p, p, n)}{\sum_{\mathbf{r}_p \in \mathfrak{R}_p} \pi(\mathbf{r}_p|p, n)m(\mathbf{y}|\mathbf{r}_p, p, n)}, \text{if } \mathbf{r}_p \in \mathfrak{R}_p,$$

where $m(\mathbf{y}|\mathbf{r}_p, p, n) = \int f(\mathbf{y}|p, \mathbf{r}_p, \theta_p)\pi(\theta_p|p, \mathbf{r}_p, n)d\theta_p$, is the marginal of the data under the partition $(p, \mathbf{r}_p)$. Without loss of generality, since $f(\mathbf{y}|1, \mathbf{r}_1, \theta)$ is nested in $f(\mathbf{y}|p, \mathbf{r}_p, \theta_p)$ it is convenient to write this posterior probability as

$$\pi(\mathbf{r}_p|\mathbf{y},p, n) = \frac{\pi(\mathbf{r}_p|p, n)\ B_{\mathbf{r}_p\mathbf{r}_1}(\mathbf{y})}{\sum_{\mathbf{r}_p\in\Re_p}\pi(\mathbf{r}_p|p, n)B_{\mathbf{r}_p\mathbf{r}_1}(\mathbf{y})}, \quad \text{if } \mathbf{r}_p \in \Re_p,$$

where $B_{\mathbf{r}_p,\mathbf{r}_1}(\mathbf{y}) = m(\mathbf{y}|\mathbf{r}_p, n)/m(\mathbf{y}|\mathbf{r}_1, n)$ represents the Bayes factor for comparing model $M_{\mathbf{r}_p}$ against the model for only one cluster, $M_{\mathbf{r}_1}$.

In the class of all models $\Re$ the posterior probability of model $M_{\mathbf{r}_p}$ is given by

$$\pi(p, \mathbf{r}_p|\mathbf{y}) = \frac{\pi(p, \mathbf{r}_p|n)\ B_{\mathbf{r}_p\mathbf{r}_1}(\mathbf{y})}{\sum_{p=1}^{n}\sum_{\mathbf{r}_p\in\Re_p}\pi(p, \mathbf{r}_p|n)\ B_{\mathbf{r}_p\mathbf{r}_1}(\mathbf{y})}, \text{ if } \mathbf{r}_p \in \Re. \tag{5}$$

The posterior probability of p clusters is

$$\pi(p|\mathbf{y}) = \frac{\sum_{\mathbf{r}_p\in\Re_p}\pi(p, \mathbf{r}_p|n)\ B_{\mathbf{r}_p\mathbf{r}_1}(\mathbf{y})}{\sum_{p=1}^{n}\sum_{\mathbf{r}_p\in\Re_p}\pi(p, \mathbf{r}_p|n)B_{\mathbf{r}_p\mathbf{r}_1}(\mathbf{y})}, \ 1 \le p \le n. \tag{6}$$

We note that $B_{\mathbf{r}_1\mathbf{r}_1}(\mathbf{y}) = 1$. The advantage of writing these posterior probabilities in terms of Bayes factors for nested models is that there are available objective priors for the model parameters to define them.

In passing we note that using only these Bayes factors we can also compare the non-nested models contained in the class $\Re$. Indeed, the Bayes factor for comparing models $M_{\mathbf{r}_p}$ and $M_{\mathbf{r}_q}$, for arbitrary p and q, is given by $B_{\mathbf{r}_p\mathbf{r}_q}(\mathbf{y}) = B_{\mathbf{r}_p\mathbf{r}_1}(\mathbf{y})/B_{\mathbf{r}_q\mathbf{r}_1}(\mathbf{y})$.

3 Priors on the Discrete Parameters: The Number of Clusters and Partitions

In this section we concentrate on priors over the models, $\pi(p, \mathbf{r}_p|n)$, which can be factorized as $\pi(p, \mathbf{r}_p|n) = \pi(\mathbf{r}_p|p, n)\pi(p|n)$. In this representation it will be seen that the prior distribution on the partitions $\pi(\mathbf{r}_p|p, n)$ plays a much more relevant role than the prior distribution on the number of clusters $\pi(p|n)$. Although both factors depend on n, the former is much more sensitive to n than the latter, as the size of the cluster classes grows exponentially with n. We consider here four priors on $(p, \mathbf{r}_p)$ which are motivated by reasonable but different assumptions.

3.1 The Uniform Prior

The first prior one may use, in the absence of information about the models, is the *uniform prior*, which gives the same probability to every model, that is,

$$\pi^U(p, \mathbf{r}_p|n) = \frac{1}{\mathcal{B}_n}, \tag{7}$$

where $\mathcal{B}_n$ is the Bell number. As we will see, this seemingly innocuous choice can have unforseen consequences.

3.2 The Ewens-Pitman Prior

The Dirichlet random process provides a distribution for $(p, \mathbf{r}_p)$ given by

$$\pi^{EP}(p, \mathbf{r}_p|\lambda, n) = \frac{\Gamma(\lambda)}{\Gamma(n+\lambda)}\lambda^p \prod_{j=1}^{p} \Gamma(n_j),\ p = 1, ..., n,\ \mathbf{r}_p \in \mathfrak{R}_p, \tag{8}$$

where λ is an unknown positive hyperparameter which has to be assessed. This prior has been used extensively (Crowley 1997; Quintana and Iglesias 2003; Booth et al. 2008; Jensen and Liu 2008; McCullagh and Yang 2006). A detailed scheme to derive the prior (8) is presented in McCullagh and Yang (2006). They also note that the limit, as $p \to \infty$, is the Ewens process (Ewens 1972; Ishwaran and Zarepour 2002; Pitman 1996), also called the Chinese restaurant process (Aldous 1985; Pitman 1996).[1]

We note that the difference among the partitions $\mathbf{r}_p$ in $\mathfrak{R}_{p;n_1,...,n_p}$ is simply the different ways we assign n_i components of the sample of size n to the density labeled by θ_i, for $i = 1, ..., p$. This implies that *a priori* these exchangeable partitions should have the same probability, and from the Ewens-Pitman prior it follows that the distribution $\pi^{EP}(\mathbf{r}_p|\mathfrak{R}_{p;n_1,...,n_p}, n)$ is uniform.

3.3 The Jensen-Liu Prior

An alternative prior is that given by Jensen and Liu (2008),

$$\pi^{JL}(p, \mathbf{r}_p|\lambda, n) \propto \lambda^{p-1}(\lambda + p)\prod_{i=1}^{p}(\lambda + i)^{-n_i}, \tag{9}$$

where λ is an unknown positive hyperparameter which has to be assessed. According to the authors "it favors equal allocations of observations, that is, the prior probability that a new observation is placed in any one of the existing clusters is uniform." We note that the distribution $\pi^{JL}(\mathbf{r}_p|\mathfrak{R}_{p;n_1,...,n_p}, n)$ is uniform.

3.4 The Hierarchical Uniform Prior (HUP)

One of our goals is to develop an objective prior for models, and to do so we consider the structure of the cluster problem as described in Section 2. Thus, we first split the entire set of partitions by conditioning on the number of clusters, and also split these sets into subsets of partitions having the same configurations. To carry out our prior specification, we start from the decomposition

$$\pi(p, \mathbf{r}_p|n) = \pi(\mathbf{r}_p|\mathfrak{R}_{p;n_1,...,n_p}, n)\pi(\mathfrak{R}_{p;n_1,...,n_p}|p, n)\pi(p|n).$$

[1] We have referred to this distribution in a variety of ways, and each way has received criticism from some quarter. We believe that *Ewens-Pitman* allocates the correct degree of recognition.

Since the partitions $\mathbf{r}_p$ in $\mathfrak{R}_{p;n_1,\ldots,n_p}$ are exchangeable, from (2) we have that

$$\pi^{HU}(\mathbf{r}_p|\mathfrak{R}_{p;n_1,\ldots,n_p},n)=\binom{n}{n_1\cdots n_p}^{-1}R(n_1,\ldots,n_p),\ \mathbf{r}_p\in\mathfrak{R}_{p;n_1,\ldots,n_p}.$$

Next, we make the assumption that the sets of partitions $\mathfrak{R}_{p;n_1,\ldots,n_p}$ in $\mathfrak{R}_p$ obtained as the vector $(n_1,\ldots,n_p)$ varies are *a priori* equally likely. The reason is that the likelihoods of the partitions in $\mathfrak{R}_p$ are of the same complexity, that is contain p different sampling models, regardless of the configuration class $\mathfrak{R}_{p;n_1,\ldots,n_p}$. Then, recalling (3), it follows that

$$\pi^{HU}(\mathfrak{R}_{p;n_1,\ldots,n_p}|p,n)=b(n,p)^{-1}. \tag{10}$$

To complete the specification of the prior $\pi(p,\mathbf{r}_p|n)$ we need to construct $\pi(p|n)$. For doing that we observe that when analyzing a cluster problem of a sample of size n the extreme case of having n clusters should be given *a priori* a smaller probability than that given to any other case. Extending this argument for any n, it might be reasonable that the prior distribution on the number of clusters p be, in a smooth way, a decreasing function of p. A candidate for $\pi(p|n)$ might then be a truncated Poisson distribution $\mathcal{P}(p|\lambda)$, where λ is an unknown parameter. Assuming the default improper Jeffreys distribution for λ, $\pi^J(\lambda)\propto\lambda^{-1/2}$, the marginal distribution for p is given by $\int\mathcal{P}(p|\lambda)\pi^J(\lambda)d\lambda$ which we now truncate to the set $\{1,\ldots,n\}$. This prior is rather close to the uniform prior in the sense that it has a very flat tail, and consequently for large n it will dilute the prior probabilities in $\{1,\ldots,n\}$. A way to derive a prior for p with thinner tail than the above one is obtained by replacing the Jeffreys prior $\pi^J(\lambda)$ with the intrinsic prior $\pi^I(\lambda|\lambda_0=1)$ constructed by testing the Poisson null hypothesis $H_0:\lambda=\lambda_0$ versus $H_1:\lambda\in R^+$. This prior, given by Moreno (2005), is

$$\pi^I(\lambda|\lambda_0=1)=\frac{\lambda^{-1/2}}{\Gamma(1/2)}e^{-(\lambda+1)}{}_0F_1(1/2,\lambda),$$

where ${}_0F_1(\frac{1}{2},\lambda)$ denotes the confluent hypergeometric function. The reason for taking $\lambda_0=1$ is that the one cluster model is the reference model throughout the analysis. The resulting marginal intrinsic distribution for p is

$$\pi^I(p|n)=\frac{m^I(p)}{\sum_{p=1}^n m^I(p)},\ p=1,\ldots,n,\quad m^I(p)=\int_0^\infty\frac{\lambda^p\ e^{-\lambda}}{p!}\ \pi^I(\lambda|\lambda_0=1)\ d\lambda. \tag{11}$$

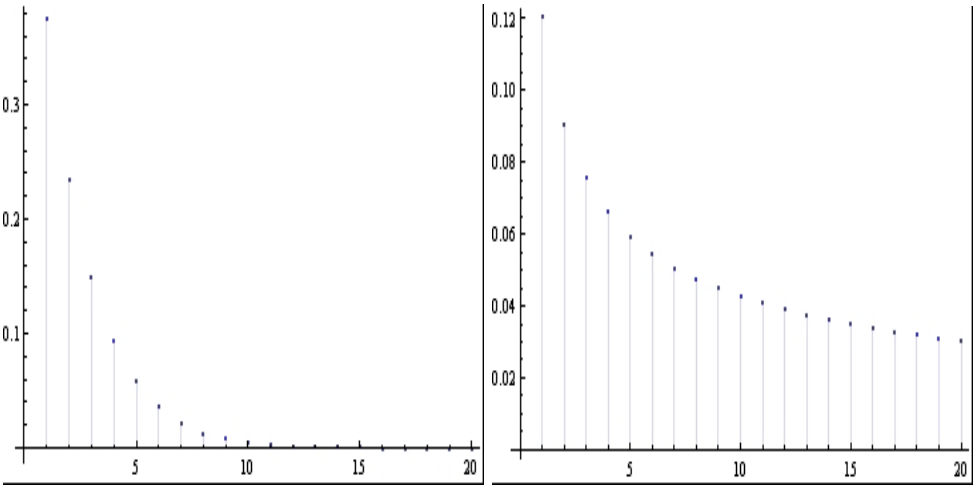
This prior does not incorporate any subjective prior information on p. If meaningful prior information is available we certainly should take the prior accordingly.

Figure 3 shows a plot of the Poisson-Intrinsic prior $m^I(p|n=20)$ along with the Poisson-Jeffreys prior. There we see that the drop in prior probabilities, as we move from one cluster, is very flat for the Poisson-Jeffreys prior, but steeper for the Poisson-Intrinsic prior.

Then, using (11) together with the development above, we obtain the prior distribution for $(p,\mathbf{r}_p)$, for $\mathbf{r}_p\in\mathfrak{R}_p$ and $p=1,\ldots,n$,

$$\pi^{HUP}(p,\mathbf{r}_p|n)=\binom{n}{n_1\cdots n_p}^{-1}R(n_1,\ldots,n_p)\ b(n,p)^{-1}\frac{m^I(p)}{\sum_{p=1}^n m^I(p)}.$$

Figure 3: Prior probabilities from the Poisson-Intrinsic prior of (11) (left panel), and the Poisson-Jeffreys prior (right panel), $n = 20$.



3.5 Comparisons

Both the Ewens-Pitman and Jensen-Liu priors require specification of the hyperparameter λ. For Ewens-Pitman, it is known that the expected prior number of clusters is given by

$$E(p|\lambda, n) = \sum_{p=1}^{n} \frac{\lambda}{\lambda + p - 1}.$$

Thus, large values of λ will lead to a larger number of clusters, and the value of the prior probabilities is quite sensitive to the choice of λ, so that the selection of this hyperparameter is important. (In Quintana and Iglesias (2003) the distribution (8) was ruled out due to its sensitivity to λ; see their Table 2, page 570.) Rather than presenting a lengthy sensitivity analysis of these priors, we just want to illustrate their performance with a small numerical example; further comparisons are in Section 5. Table 3 shows prior probability specifications for two configuration classes for $n = 10$.

A priori it does not seem to us that we would have any reason to assign different probabilities to the configurations $\{1, 3, 6\}$ and $\{2, 3, 5\}$ as they are of the same complexity, i.e., the likelihood function of any partition having configuration either $\{1, 3, 6\}$ or $\{2, 3, 5\}$ contains only three different densities, but this is what is done by the Uniform, the Jensen-Liu and the Ewens-Pitman priors as Table 3 illustrates for $\lambda = 1$. The fact that there are 840 partitions corresponding to the configuration $\{1, 3, 6\}$, and 2525 partitions corresponding to $\{2, 3, 5\}$ only explains the numbers for the uniform prior in Table 3. Moreover, the preferences reverse, with Ewens-Pitman and Jensen-Liu favoring

Table 3: Prior probabilities for exchangeable partition sets in $\mathfrak{R}_3$ for $n = 10$. The designation $\{1, 3, 6\}$ refers to having three clusters with $1, 3$ and 6 observations; $\{2, 3, 5\}$ has $2, 3$ and 5 observations. (Both Ewens-Pitman and Jensen-Liu have $\lambda = 1$.)

	Prior Probabilities			
Configuration	Ewens-Pitman	Hierarchical Uniform	Uniform	Jensen-Liu
$\{1, 3, 6\}$	0.17	0.14	0.09	0.15
$\{2, 3, 5\}$	0.10	0.14	0.27	0.08

$\{1, 3, 6\}$, and the uniform prior favoring $\{2, 3, 5\}$.

An important difference between the HUP and the other priors is that the HUP is the only one that assignes equal probability to the configuration classes contained in $\mathfrak{R}_p$. We note that the four priors assume for $\pi(\mathbf{r}_p|\mathfrak{R}_{p;n_1,\ldots,n_p}, n)$ a uniform distribution.

4 Intrinsic Priors for the Continuous Parameters θ_p

As we noted in Section 3, for computing the Bayes factors $B_{\mathbf{r}_p,\mathbf{r}_1}(\mathbf{y})$, $\mathbf{r}_p \in \mathfrak{R}$, prior distributions for the continuous parameters θ and θ_p are needed. The usual objective choices are the reference priors $\pi^N(\theta)$ and $\pi^N(\theta_p)$ associated with the sampling models $f(\mathbf{y}|1, \mathbf{r}_1, \theta)$ and $f(\mathbf{y}|p, \mathbf{r}_p, \theta_p)$, respectively (Berger et al. 2009). However, these priors are typically improper, and while this is not an inconvenience for estimating θ and θ_p, it is a serious problem for model comparison, as they leave the Bayes factor $B_{\mathbf{r}_p,\mathbf{r}_1}(\mathbf{y})$ defined only up to an arbitrary multiplicative constant.

Fortunately, the sampling model $f(\mathbf{y}|1, \mathbf{r}_1, \theta)$ is nested in $f(\mathbf{y}|p, \mathbf{r}_p, \theta_p)$ and then the reference priors can be converted into the so-called intrinsic priors (Berger and Pericchi 1996b; Moreno 1997; Moreno et al. 1998), for which not only is the Bayes factor well-defined, but also the intrinsic prior for θ_p concentrates probability mass around θ, a desirable condition for model comparison (known as the *Savage continuity condition*). Furthermore, the intrinsic priors have been proved to behave well in a wide variety of problems (Berger and Pericchi 1996a; Berger and Mortera 1999; Kim and Sun 2000; Casella and Moreno 2005, 2009; Girón et al. 2006; Moreno 2005; Moreno et al. 2010; Casella et al. 2009, among others). The intrinsic prior for the parameter θ_p, conditional on an arbitrary but fixed point θ, is given by

$$\pi^I(\theta_p|\theta) = \pi^N(\theta_p|\mathbf{r}_p) E_{\mathbf{y}(\ell_p)|\theta_p} \frac{f(\mathbf{y}|1, \mathbf{r}_1, \theta)}{\int f(\mathbf{y}(\ell_p)|p, \mathbf{r}_p, \theta_p)\pi^N(\theta_p)d\theta_p},$$

where the expectation is taken with respect to the sampling distribution $f(\mathbf{y}(\ell_p)|p, \mathbf{r}_p, \theta_p)$ with $\mathbf{y}(\ell_p)$ a vector of dimension $\ell_p = kp + 1$. Here ℓ_p denotes the minimal training sample size needed for estimating θ_p with the prior $\pi^N(\theta_p)$. It can be easily checked that $\pi^I(\theta_p|\theta)$ is a probability distribution. The unconditional intrinsic prior for θ_p is given by $\pi^I(\theta_p) = \int \pi^I(\theta_p|\theta)\pi^N(\theta)d\theta$,and the pair $\left(\pi^N(\theta), \pi^I(\theta_p)\right)$ is the intrinsic prior

for comparing models $M_{\mathbf{r}_1}$ and $M_{\mathbf{r}_p}$. We note that they are improper priors whose moments typically do not exist, which seems to be a reasonable property for objective priors, although they are well-calibrated priors in the sense that both depend on a unique arbitrary multiplicative constant, the arbitrary constant inherited from $\pi^N(\theta)$, which cancels out in the ratio. Therefore, the Bayes factor for intrinsic priors

$$B_{\mathbf{r}_p\mathbf{r}_1}(\mathbf{y}) = \frac{\int f(\mathbf{y}|p, \mathbf{r}_p, \theta_p)\pi^I(\theta_p)d\theta_p}{\int f(\mathbf{y}|1, \mathbf{r}_1, \theta)\pi^N(\theta)d\theta}, \quad \mathbf{r}_p \in \mathfrak{R},$$

is free of arbitrary constants, needs neither subjective input nor actual (data-dependent) training samples, and is completely automatic.

4.1 The Case of Linear Models

We now consider the case where the class of parametric sample densities, $\mathfrak{F}$, is the class of linear models with k regressors. For example, suppose that the sample $(y_1, ..., y_n)$ follows a normal linear model

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim N_n(\varepsilon|0, \tau^2\mathbf{I}_n),$$

where $\mathbf{X}$ is an $n \times k$ design matrix of full rank, β is a vector of regression coefficients with dimension k, and τ^2 is the common variance of the error terms. This is the sampling model for one cluster in the sample and, in the notation of the preceding section, we have

$$f(\mathbf{y}|1, \mathbf{r}_1, \beta, \tau) = N_n(\mathbf{y}|\mathbf{X}\beta, \tau^2\mathbf{I}_n).$$

The reference prior for the parameters of this model is $\pi^N(\beta, \tau) = c/\tau$, where c is an arbitrary positive constant. Without loss of generality, suppose we split the sample into p clusters, where one cluster is formed with the first n_1 components of the sample, a second cluster is formed with the second n_2 components, and so on. These clusters correspond to the partition of the sample as $\mathbf{y}' = (\mathbf{y}_1', ..., \mathbf{y}_p')$ and the corresponding partition of the design matrix as

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \cdots \\ \mathbf{X}_p \end{pmatrix},$$

where $\mathbf{X}_i$ has dimensions $n_i \times k$. Then, the sampling model for p clusters is given by

$$f(\mathbf{y}|p, \mathbf{r}_p, \beta_1, ..., \beta_p, \sigma_p) = \prod_{i=1}^{p} N_n(\mathbf{y}_i|\mathbf{X}_i\beta_i, \sigma_p^2\mathbf{I}_{n_i}),$$

assuming that σ_p^2 is the common variance of the p clusters model.

We note that the model for p clusters can be written as the linear model $\mathbf{y} = \mathbf{V}\gamma + \eta$, where $\mathbf{V}$ is the following lower triangular $n \times kp$ design matrix

$$\mathbf{V} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{X}_2 & \mathbf{X}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{X}_p & \mathbf{X}_p & \cdots & \mathbf{X}_p \end{pmatrix},$$

γ being the $k \times p$ vector $\gamma' = (\gamma_1', \ldots, \gamma_p')$, where $\gamma_1 = \beta_1, \gamma_2 = \beta_2 - \beta_1, \ldots, \gamma_p = \beta_p - \beta_{p-1}$ are k dimensional vectors, and the random vector η is distributed as $N_n(\eta|\mathbf{0}, \sigma_p^2\mathbf{I}_n)$. Thus, the sampling model for p clusters is now given by $f(\mathbf{y}|p, \mathbf{r}_p, \gamma_1, \ldots, \gamma_p, \sigma_p) = N_n(\mathbf{y}|\mathbf{V}\gamma, \sigma_p^2\mathbf{I}_n)$. Since $f(\mathbf{y}|1, \mathbf{r}_1, \beta, \tau)$ is nested in $f(\mathbf{y}|p, \mathbf{r}_p, \gamma_1, \ldots, \gamma_p, \sigma_p)$, by simply making $\beta_1 = \beta_2 = \cdots = \beta$ or, equivalently, $\gamma_1 = \beta, \gamma_2 = \cdots = \gamma_p = 0$, direct application of the standard intrinsic methodology gives the intrinsic prior for the parameters $(\gamma_1, \ldots, \gamma_p, \sigma_p)$ conditional on a fixed point (β, τ), as

$$\pi^I(\gamma_1, \ldots, \gamma_p, \sigma_p|\beta, \tau) = \frac{2}{\pi\tau(1 + \sigma_p^2/\tau^2)} N_{pk}(\gamma|(\mathbf{X}\beta, \mathbf{0}, \ldots, \mathbf{0})', (\sigma_p^2 + \tau^2)\mathbf{W}^{-1}),$$

where $\mathbf{W}^{-1} = n/(pk+1)(\mathbf{V}'\mathbf{V})^{-1}$. Note that the conditional intrinsic prior is centered at the null model (the one cluster model), and its covariance structure depends on the covariance matrix of the model of p clusters only. The unconditional intrinsic priors are given by the pair $(\pi^N(\beta, \tau), \pi^I(\gamma_1, \ldots, \gamma_p, \sigma_p))$, where

$$\pi^I(\gamma_1, \ldots, \gamma_p, \sigma_p) = \int \pi^I(\gamma_1, \ldots, \gamma_p, \sigma_p|\beta, \tau)\pi^N(\beta, \tau)\, d\gamma\, d\tau. \tag{12}$$

4.2 Bayes Factors for Intrinsic Priors

The objective intrinsic Bayesian model for one cluster is

$$M_{\mathbf{r}_1} : \{N_n(\mathbf{y}|\mathbf{X}\beta, \tau^2\mathbf{I}_n), \pi^N(\beta, \tau)\},$$

and for p clusters

$$M_{\mathbf{r}_p} : \{N_n(\mathbf{y}|\mathbf{V}\gamma, \sigma_p^2\mathbf{I}_n), \pi^I(\gamma_1, \ldots, \gamma_p, \sigma_p)\},$$

where $\pi^I(\gamma_1, \ldots, \gamma_p, \sigma_p)$ is given in (12). To compute the Bayes factor of model $M_{\mathbf{r}_p}$ versus model $M_{\mathbf{r}_1}$, we note that the residual sum of squares of a partition, with cluster sizes $(n_1, ..., n_p)$, is equal to $RSS_{n_1,\ldots,n_p} = \mathbf{y}'(\mathbf{I} - \mathbf{H}_\mathbf{V})\mathbf{y}$, where $\mathbf{H}_\mathbf{V} = \mathbf{V}(\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'$. Simple but cumbersome algebra shows that, due to the lower diagonal structure of the matrix $\mathbf{V}$, the residual sum of squares for that partition, with cluster sizes $(n_1, ..., n_p)$, can be written as

$$RSS_{n_1,\ldots,n_p} = \sum_{i=1}^{p} RSS_{n_i},$$

where RSS_{n_i} is the residual sum of squares from the regression in the i^{th} cluster. Some lengthy calculations render a quite simple form for the Bayes factor for intrinsic priors. The following theorem summarizes the result.

Theorem 1. *The Bayes factor for the model $M_{\mathbf{r}_p}$ versus model $M_{\mathbf{r}_1}$ is given by*

$$B_{\mathbf{r}_p\mathbf{r}_1}(\mathbf{y}) = \frac{2}{\pi}(pk+1)^{(p-1)k/2}\int_0^{\pi/2}\frac{\sin^{(p-1)k}\varphi\,(n+(pk+1)\sin^2\varphi)^{(n-pk)/2}}{(n\mathcal{R}_{\mathbf{r}_p}+(pk+1)\sin^2\varphi)^{(n-k)/2}}\,d\varphi \quad (13)$$

where the statistic $\mathcal{R}_{\mathbf{r}_p}$ is

$$\mathcal{R}_{\mathbf{r}_p} = \frac{RSS_{n_1}+\ldots+RSS_{n_p}}{RSS_n},$$

with $RSS_{n_i} = \mathbf{y}_i'(\mathbf{I}-\mathbf{H}_i)\mathbf{y}_i$, $i=1,...,p$, $RSS_n = \mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}$, and $\mathbf{H}_i$ and $\mathbf{H}$ the hat matrices associated with $\mathbf{X}_i$ and $\mathbf{X}$, respectively.

Proof. The marginal under model $\mathbf{r}_p$ is given by

$$m_{\mathbf{r}_p}(\mathbf{y}) = \int\left\{\int\left\{\int N_n(\mathbf{y}|\mathbf{X}\beta,\tau^2\mathbf{I}_n)\pi^I(\gamma_1,\ldots,\gamma_p,\sigma_p|\beta,\tau)\,d\gamma\right\}\pi^N(\beta,\tau)\,d\beta\right\}d\sigma_p\,d\tau.$$

The integral with respect to γ and β can be done in closed form, and a change of variables from (σ_p,τ) to polar coordinates gives the above expression for the Bayes factor. □

Substituting the Bayes factor for the intrinsic priors (13) in expressions (5) and (6), the intrinsic posterior probability of model $M_{\mathbf{r}_p}$ and the posterior probability of p clusters in the sample, are obtained. We want to note that with a large number of observations, the factor $\mathcal{R}_{\mathbf{r}_p}$ in (13) can get very close to zero, causing numerical problems in computation (the integral is returned as infinite). However, the transformation $t = \frac{pk+1}{\mathcal{R}_{\mathbf{r}_p}}\sin^2(\varphi)$ results in a representation that is numerically very stable, and allows for doing all computations on the log scale.

Lastly, there is one technicality to note. If a partition contains a value of n_i with $n_i < k$ then, of course, the regression cannot be fit in that cluster. We proceed by fitting the largest model feasible in that particular partition, with the limiting case being a cluster of size 1, to which we assign $\mathrm{Var}(Y)$ as the residual sum of squares. (One might consider assigning a residual sum of squares of zero to such a cluster, but this unduly rewards clusters of size 1.)

5 The Effect of the Prior on the Consistency of the Bayesian Procedure

It is well-known that for regular sampling models, the Bayesian model selection procedure is consistent when the dimension of the sampling model is fixed and comparisons are pairwise. In that case consistency of the Bayesian model selection procedure is inherited from the consistency of the Bayes factor, because the model prior does not play any role in the consistency of the procedure. However, when the dimension of the

model grows with the sample size, the model prior plays an important role for obtaining consistency, and this is the case in clustering. As we will see, the choice of the prior on model space is of major importance in determining the asymptotic behavior of a clustering procedure. Surprisingly, the actual choice of Bayes factor is of almost no consequence in determining consistency, as many Bayes factors have the same asymptotic representation.

In this section we look at the asymptotic behavior of Bayesian clustering procedures when using the four model priors of Section 3. Perhaps the most surprising result is that the uniform prior, which gives the same prior probability to every model, and the Jensen-Liu prior, that allocates a new observation uniformly in any one of the existing clusters, lead to inconsistent Bayesian procedures. Furthermore, this is the case when sampling from the simpler model, and the number of clusters is finite, a situation in which consistency is typically obtained for a Bayesian testing procedure. This behavior is explained by observing that, in clustering, the model prior depends on the sample size and, as the sample size tends to infinity, the speed of convergence of the prior to its limit compared with that of the Bayes factor is now crucial.

In what follows we will assume that the number of clusters p is bounded, that is, $p \leq T < \infty$. We could be more general and put a growth rate on the number of clusters (Moreno et al. 2010) but, in practice, assuming that p is bounded is certainly a realistic constraint.

For large samples, approximations of Bayes factors for intrinsic priors depend on the dimensions of the competing models and a *pseudodistance* between them. If M_i and M_j are arbitrary general normal linear models, the pseudodistance from M_i to M_j is defined as

$$\delta_{ij} = \frac{1}{\sigma_i^2}\beta_i' \frac{\mathbf{X}_i'(\mathbf{I}_n - \mathbf{H}_j)\mathbf{X}_i}{n}\beta_i.$$

Note that this pseudodistance is defined for every pair of models, not only nested models, and it is not symmetric. Some useful properties of δ_{pi} are the following: (a) The distance from any model M_i to itself is always 0, (b) if M_i is nested in M_j, then $\delta_{ij} = 0$, and (c) if model M_i is nested in M_j, then $\delta_{ki} \geq \delta_{kj}$ for any model M_k.

We start with the following lemma, where we recall that the singular class $\mathfrak{R}_{1;n}$ contains only the one cluster model $M_{\mathbf{r}_1}$. (We use the notation $[M]$ to denote the model that generated the sample.)

Lemma 1. *Suppose that model M_i, of dimension i, is nested in model M_j, of dimension j, and M_t is the true model. Under the sampling model M_t, as $n \to \infty$, the Bayes factor can be approximated by*

$$B_{\mathbf{r}_j\mathbf{r}_i}(\mathbf{y}) \approx \exp\left\{\left(\frac{i-j}{2}\right)\log\left(\frac{n}{j+1}\right)\right\}\left(\frac{1+\delta_{ti}}{1+\delta_{tj}}\right)^{n/2} \quad [M_t],$$

where δ_{ti} and δ_{tj} are the pseudodistances from the true model to models M_i and M_j, respectively.

In particular, when sampling from model $M_{\mathbf{r}_1}$, for large n, the Bayes factor $B_{\mathbf{r}_p\mathbf{r}_1}(\mathbf{y})$ can be approximated by

$$B_{\mathbf{r}_p\mathbf{r}_1}(\mathbf{y}) \approx \left(\frac{pk+1}{n}\right)^{k(p-1)/2} [M_{\mathbf{r}_1}].$$

Proof. The first expression immediately follows from Lemma 3 in Girón *et al.* (2010), and the second expression follows from noting that $\delta_{r_j r_1}$ is zero for all models $M_{\mathbf{r}_j}$. □

Lemma 1 shows that for large n, when sampling from $M_{\mathbf{r}_1}$, the Bayes factor $B_{\mathbf{r}_p\mathbf{r}_1}(\mathbf{y})$ is constant across partitions $\mathbf{r}_p$ in the class $\mathfrak{R}_p$. Moreover, this asymptotic result is not limited to intrinsic Bayes factors; for example we know that BIC is asymptotically equivalent to the intrinsic Bayes factor. Moreover, Casella et al. (2009) show that the approximation holds for a wide class of priors.

In the following two subsections we analyze the limiting behavior of the four priors of Section 3, and then examine their effect on the consistency of the resulting Bayes procedures.

5.1 Limiting Behavior of the Model Priors

For a fixed value of p, we denote by $\mathbf{Z}_n$ a p-dimensional random vector which takes values $(n_1, \ldots, n_p)$ on the set of integers satisfying the conditions $n_1 \leq \cdots \leq n_p$ and $n_1 + \cdots + n_p = n$, with distribution

$$\Pr(\mathbf{Z}_n = (n_1, \ldots, n_p)|\mathfrak{R}_p) = \Pr(M_{(n_1,\ldots,n_p)}|\mathfrak{R}_p),$$

which represents the prior probability of any model arising from the priors of Section 3 in the set of models $\mathfrak{R}_p$. The study of the limiting behavior of the posterior probability of the configuration classes cannot be done straightforwardly as the sample space where the random vector $\mathbf{Z}_n$ takes values changes with the sample size n. If we consider instead the common space of the simplex $\mathcal{S}_p$, all models can be regarded as points of this simplex. Thus, we define the random vector $\mathbf{Y}_n = \frac{1}{n}\mathbf{Z}_n$, and study the limiting behavior of $\mathbf{Y}_n$ for the different priors. Note that the sample space of $\mathbf{Y}_n$ is $\mathcal{S}_p$ for all n, and $\Pr(\mathbf{Y}_n = \left(\frac{n_1}{n}, \ldots, \frac{n_p}{n}\right)|\mathfrak{R}_p) = \Pr(\mathbf{Z}_n = (n_1, \ldots n_p)|\mathfrak{R}_p)$.

Theorem 2.

(*a*) *For fixed p and the hierarchical uniform prior, $\mathbf{Y}_n = \frac{1}{n}\mathbf{Z}_n$ converges in distribution to a uniform prior on the simplex $\mathcal{S}_p$.*

(*b*) *For fixed p and λ and the Ewens-Pitman prior, $\mathbf{Y}_n = \frac{1}{n}\mathbf{Z}_n$ converges in distribution to an improper prior distribution whose density is proportional to*

$$\frac{1}{x_1 \cdot x_2 \cdots (1 - x_1 - \cdots - x_{p-1})}, \tag{14}$$

restricted to the simplex $\mathcal{S}_p$. Note that this limiting function does not depend on λ.

(c) *For fixed p and the uniform prior, $\mathbf{Y}_n = \frac{1}{n}\mathbf{Z}_n$ converges in distribution, and in probability, to a degenerate distribution concentrated at the vertex $(\frac{1}{p}, \ldots, \frac{1}{p})$ of the simplex $\mathcal{S}_p$.*

(d) *For fixed p and λ and the Jensen-Liu prior, $\mathbf{Y}_n = \frac{1}{n}\mathbf{Z}_n$ converges in distribution, and in probability, to a degenerate distribution concentrated at the interior point $\left(\frac{\lambda_T^H}{p(\lambda+T)}, \ldots, \frac{\lambda_T^H}{p(\lambda+1)}\right)$ of the simplex $\mathcal{S}_p$, where λ_p^H denotes the harmonic mean of $\lambda+1, \ldots, \lambda+p$.*

Proof. The proof is given in Appendix A.2 □

Thus, of all four priors, only the HUP converges to a proper distribution. The Ewens-Pitman prior converges to an improper Dirichlet distribution with parameters equal to zero although, when sampling from model $M_{\mathbf{r}_1}$, the limit of the posterior distribution is proper. What is, perhaps, most distressing is the limiting behavior of the uniform and Jensen-Liu priors, which degenerate to a point and, thus, could have undo influence in the clustering algorithm.

5.2 Consistency of Bayes Procedures

We first state two theorems about the consistency of the Bayes procedures when sampling from the one cluster model, for the four priors considered in the paper.

Theorem 3. *Suppose that $p \leq T$ and we use either the uniform prior or the Jensen-Liu prior on the class of all partitions $\mathfrak{R} = \cup_{p=1}^{T} \mathfrak{R}_p$. Then, when sampling from $M_{\mathbf{r}_1}$ the Bayes procedure is inconsistent in both the cluster classes and the configuration classes. Moreover, in the probability space generated by the cluster classes $\{\mathfrak{R}_p,\ p = 1, ..., T\}$, the asymptotic posterior distribution of $\mathfrak{R}_p$, $p = 1, ..., T$, is*

$$\lim_{n\to\infty} [M_{\mathbf{r}_1}] Pr(\mathfrak{R}_p|\mathbf{y}) = \begin{cases} 1, & \text{if} \quad p = T, \\ 0, & \text{if} \quad p \leq T-1; \end{cases}$$

thus the largest model is chosen with probability one.

Proof. The proof is given in Appendix A.3 □

The implications of this theorem are quite interesting, and help explain some of what we had observed in looking at examples (illustrated in Section 7). With priors like the uniform, the answers from the cluster algorithm tend to be partitions with a large number of clusters, and a small number of subjects per cluster.

The situation for the Ewens-Pitman prior and the hierarchical uniform prior is different.

Theorem 4. *Suppose that we use either the Ewens-Pitman prior or the hierarchical uniform prior on the class of all partitions* $\mathfrak{R} = \cup_{p=1}^{T} \mathfrak{R}_p$. *Then, when sampling from* $M_{\mathbf{r}_1}$, *the Bayesian procedure is consistent. That is, in the probability space generated by the cluster classes* $\{\mathfrak{R}_p,\ p = 1, ..., T\}$,

$$\lim_{n\to\infty} [M_{\mathbf{r}_1}] Pr(\mathfrak{R}_1|\mathbf{y}) = 1,$$

so the correct model is chosen with probability 1.

Proof. The proof is given in Appendix A.4 □

Thus, both of these priors exhibit good asymptotic behavior. Consider the rate of convergence of the posterior probability of $\mathfrak{R}_1$ to one. For the Ewens-Pitman prior the convergence rate is $O\left(\frac{\log n}{n^{p-1}}\right)$, while for the HUP it is $O\left(\frac{1}{n^{p-1}}\right)$. This means that the convergence rate is faster when using the HUP than with the Ewens-Pitman prior. The difference is the presence of the $\log n$ term. We also note that the rate with the Ewens-Pitman prior also depends on the value of λ; the larger this value the slower the consistency under the null.

Unfortunately, the Bayesian procedures for clustering are inconsistent when sampling from an alternative model $M_{\mathbf{r}_t} \neq M_{\mathbf{r}_1}$. This comes from the fact that model $M_{\mathbf{r}_t}$ may not be distinguishable from model $M_{\mathbf{r}'_t}$, assuming that both partitions $\mathbf{r}_t$ and $\mathbf{r}'_t$ belong to the same configuration class $\mathfrak{R}_{t;n_1\cdots n_t}$. To show this assertion let us consider the simple sampling model $N(y|\theta, \sigma^2)$ with no regressors.

Theorem 5. *For the normal model with no regressors* $N(y|\theta, \sigma^2)$, *when sampling from the cluster model* $M_{\mathbf{r}_t} \neq M_{\mathbf{r}_1}$, *the Bayesian procedure for any exchangeable prior over the models is inconsistent.*

Proof. The proof is given in Appendix A.5 □

To further analyze consistency in clustering we realize that, as the sample size tends to infinity, the way we allocate the components of the sample in the clusters is not an issue, and hence we need only to consider the proportions of the sample in the clusters. Consequently, as n tends to infinity, the notion of consistency is now specific to the class of partitions having a limiting configuration $(n_1/n, ..., n_p/n) \to (\nu_1, \ldots, \nu_p)$, as $n \to \infty$. The interpretation of model $M_{\nu_1,\ldots,\nu_p}$ is that observations are assigned to clusters in the proportion $\nu_1, \ldots, \nu_p$.

This also implies that, as $n \to \infty$, the model $M_{\mathbf{r}_p}$ is not distinguishable from $M_{\mathbf{r}'_p}$, assuming that both partitions $\mathbf{r}_p$ and $\mathbf{r}'_p$ belong to the same configuration class $\mathfrak{R}_{p;n_1,\ldots,n_p}$. Consequently, as n tends to infinity, consistency will be examined in the probability space generated by the configuration classes $\{\mathfrak{R}_{p;n_1,\ldots,n_p}, n_1 \leq ... \leq n_p,\ n_1 + ... + n_p = n,\ p = 1, 2, ...\}$, and in the probability space generated by the cluster

classes $\{\mathfrak{R}_p,\ p = 1, ..., T\}$. We also note that configuration classes, and their limits, can be identified with points in the simplex $\mathcal{S}_p = \{\nu = (\nu_1, \ldots . \nu_{p-1}); 0 \leq \nu_1 \leq \cdots \leq \nu_{p-1},$ and $\nu_1 + \cdots + \nu_{p-1} < 1\}$, and cluster classes $\mathfrak{R}_p$ with the simplex $\mathcal{S}_p$. This is illustrated in Figure 4, where we show the cluster classes and models for $p = 3$. The cluster class $\mathfrak{R}_3$ is the entire simplex, while $\mathfrak{R}_2$ and $\mathfrak{R}_1$ are an edge and a vertex, respectively. The model $M_{1/3,1/3,1/3}$, the extreme right vertex of the simplex, is the limit of the uniform prior. Of course, this is extendable to higher dimensional simplexes.

For simplicity, we restrict our analysis to the case of at most two clusters, i.e. $T = 2$.

Let M_ν denote a two cluster model with a proportion ν of the data in the first cluster and a proportion $1 - \nu$ in the second cluster, where $\nu \in \mathcal{S}_2 = (0, 1/2]$. This means that the observations in the first cluster come from a normal model $N(\theta_1, \tau^2)$, and the rest from the normal model $N(\theta_2, \tau^2)$, where $\theta_1 \neq \theta_2$. Thus, the two cluster sampling model is

$$M_\nu = \begin{cases} \prod_{j=1}^{n_1} N(y_j|\theta_1, \tau^2), & \text{if} \quad n_1/n \leq \nu, \\ \prod_{j=n_1+1}^{n} N(y_j|\theta_2, \tau^2), & \text{if} \quad n_1/n > \nu. \end{cases}$$

Note that the one cluster model is a particular case of M_ν when $\nu = 0$.

The pseudodistance from model M_ν to model $M_{\nu'}$, with $\nu, \nu' \in [0, 1/2]$, turns out to be

$$\delta_{\nu\nu'} = \begin{cases} d\,|\nu - \nu'|\ \frac{\nu}{\nu'} & \text{if} \quad \nu' \geq \nu, \\ d\,|\nu - \nu'|\ \frac{1-\nu}{1-\nu'}, & \text{if} \quad \nu' \leq \nu, \end{cases}$$

where $d = (\theta_1 - \theta_2)^2/\tau^2$ is the Kullback-Leibler distance between the sampling models $N(y|\theta_1, \tau^2)$ and $N(y|\theta_2, \tau^2)$.

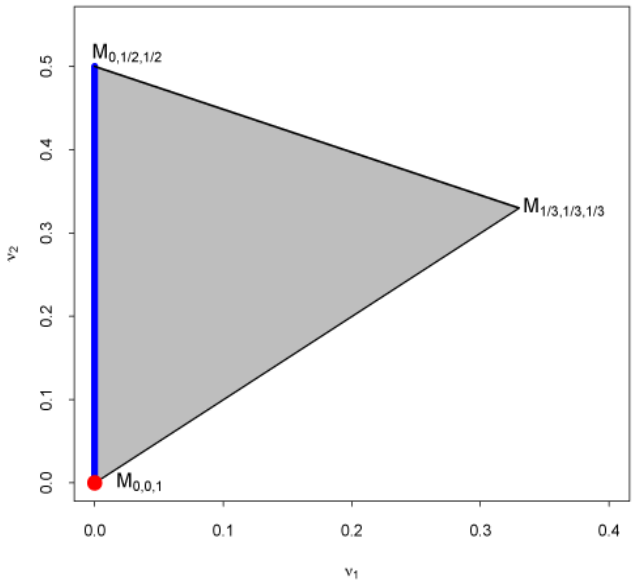
The next theorem shows that, even for sensible priors, posterior probability consistency for the configuration classes does not hold.

Theorem 6. *When sampling from model M_ν, with $\nu \in (0, 1/2]$, posterior consistency of the configuration classes, for the Ewens-Pitman, hierarchical uniform, uniform and Jensen-Liu priors, does not hold. Furthermore, for the Ewens-Pitman and hierarchical uniform priors there is a strictly positive upperbound of the posterior probability of any configuration class (including the true one) strictly less than 1. For the uniform prior and the Jensen-Liu prior the posterior probability of the true configuration class goes to 0.*

Proof. The proof is given in Appendix A.6 □

We note that we have really proved that the posterior probability is concentrated around the true sampling model, for the Ewens-Pitman and the hierarchical uniform priors, even though no model —not even the true one— has posterior probability equal to 1. This suggests a weaker form of consistency based on the convergence in probability to one of the posterior distribution of the true configuration class.

Figure 4: The simplex for $p = 3$ is in grey. The point in red represents the cluster class $\mathfrak{R}_1$, that is, the one cluster model, while the blue line represents $\mathfrak{R}_2$. The rest of the simplex represents the cluster class $\mathfrak{R}_3$. Note that cluster class $\mathfrak{R}_1$ is a vertex (extreme point) of the simplex, the cluster class $\mathfrak{R}_2$ is an edge, and the model which assigns an equal proportion to the three cluster configuration is also an extreme point of the simplex.



To prove the next theorem, we adapt the notation of subsection 5.1 to the case where $p = 2$. In particular, as $n_2 = n - n_1$, Z_n is now a random variable which takes the values n_1 such that $1 \leq n_1 \leq \lfloor n/2 \rfloor$,

$$\Pr(Z_n = n_1) = \Pr(\mathfrak{R}_{2;n_1,n_2}|\mathbf{y}).$$

The random variable $Y_n = \frac{1}{n} Z_n$ now takes values on the simplex $\mathcal{S}_2$, with distribution

$$\Pr\left(Y_n = \frac{n_1}{n}\right) = \Pr(\mathfrak{R}_{2;n_1,n_2}|\mathbf{y}),$$

and its asymptotic behavior for the four different priors will be considered in Theorem 7.

Theorem 7. *When sampling from any two cluster model M_ν, with $\nu \in (0, 1/2]$, the Bayes procedures for the Ewens-Pitman prior and the hierarchical uniform priors are*

"weakly" consistent, that is the posterior distribution of the random variable Y_n converges in distribution —or in probability— to the value ν. Further, the Bayes procedures for the uniform and the Jensen-Liu priors are not weakly consistent.

Proof. The proof is given in Appendix A.7 □

Remark 1. *Although the asymptotic behavior of the Bayes procedure for the Ewens-Pitman and the hierarchical uniform priors is the same, for small sample sizes there can be substantial differences in the distributon of Y_n, specially when the true model M_ν is close to the one-cluster model, i.e. when ν is near 0. The reason is, as has been pointed out before, than the Ewens-Pitman prior slightly favors the one cluster model when the sample size is small.*

6 Search Algorithm

Here we develop a hybrid search algorithm, using a Metropolis-Hastings (MH) algorithm that has stationary distribution proportional to the Bayes factor times the prior odds. The hybrid algorithm is a mixture of a random walk and a jump. We use a random walk component to be able to explore locally, and the jump allows escape from regions with small Bayes factors.

In setting up the algorithm there is one immediate computational problem, that of calculating the correct probabilities for the MH ratio. In the random walk piece we solve this problem by using the *biased random walk* of Booth et al. (2008). Suppose, for example, that at iteration t we have the partition $r^{(t)}_{p^{(t)}}$. We now generate a candidate partition $r'_{p'}$ from a distribution G, and accept the move with probability

$$\min\left\{\frac{\pi(p,\mathbf{r}'_p|n)B_{\mathbf{r}_{p'}\mathbf{r}_1}}{\pi(p,\mathbf{r}^{(t)}_{p^{(t)}}|n)B_{\mathbf{r}^{(t)}_{p^{(t)}}\mathbf{r}_1}}\frac{G(r^{(t)}_{p^{(t)}}|r'_{p'})}{G(r'_{p'}|r^{(t)}_{p^{(t)}})},1\right\}$$

where $\pi(p,\mathbf{r}'_p|n)$ is the prior, and the Bayes factor $B_{\mathbf{r}_{p'},\mathbf{r}_1}$ is given in (13). The computational problem arises in calculating the ratio of candidate probabilities, which could entail summing over an enormous number of partitions. However, the biased random walk has the property that $G(x|y) = G(y|x)$, and thus these terms cancel from the MH ratio. (Details and properties of the biased random walk are discussed in Booth et al. (2008), so here we will just give a brief description.)

Biased Random Walk

With the current iteration at $r^{(t)}_{p^{(t)}}$, we generate a candidate $r'_{p'}$ as:

1. If $p = 1$, choose an observation at random from all n observations, and move the chosen observation to its own cluster. The new configuration is $r'_{p'}$.

2. If $p > 1$, choose an observation at random from all n observations.

   (a) If the object is a singleton cluster, move it to one of the $p-1$ other clusters with probability $1/(p-1)$.

   (b) If the object is not a singleton cluster, move it to one of the $p-1$ other clusters, or to its own (new) cluster, each with probability $1/p$.

This is the biased random walk which, although similar to a nearest neighbor random walk, has the property that the probability of the move $r^{(t)}_{p^{(t)}} \to r'_{p'}$ is the same as the probability of the move $r'_{p'} \to r^{(t)}_{p^{(t)}}$, eliminating the necessity for calculating these probabilities in the Metropolis-Hastings algorithm.

Unfortunately, the mixing from the biased random walk is too slow for clustering large, or even medium, data sets. If we are in a good portion of the space then the random walk will explore that portion, but we also need to be able to escape from areas with small Bayes factors. To do so we have a second piece in the search algorithm, a jump based on sampling from the Ewens-Pitman distribution of (8). We can sample from this distribution using a number of algorithms (see Neal 2000, for example), but we will draw our candidate using the algorithm of Kyung et al. (2010), which has been shown to mix better than many of its competitors. (Note that using the Ewens-Pitman distribution to drive a search has nothing to do with the choice of model space priors.)

Jumping with the Ewens-Pitman Distribution

We use the Ewens-Pitman distribution to generate a random jump because it is easy to calculate the Metropolis-Hastings correction. With the current iteration at $r^{(t)}_{p^{(t)}}$, we generate a candidate $r'_{p'}$ as follows: Start with $\mathbf{n}_p = (n_1, \ldots n_p)$ obtained from $r^{(t)}_{p^{(t)}}$, and draw $\mathbf{q}$ from the Dirichlet distribution

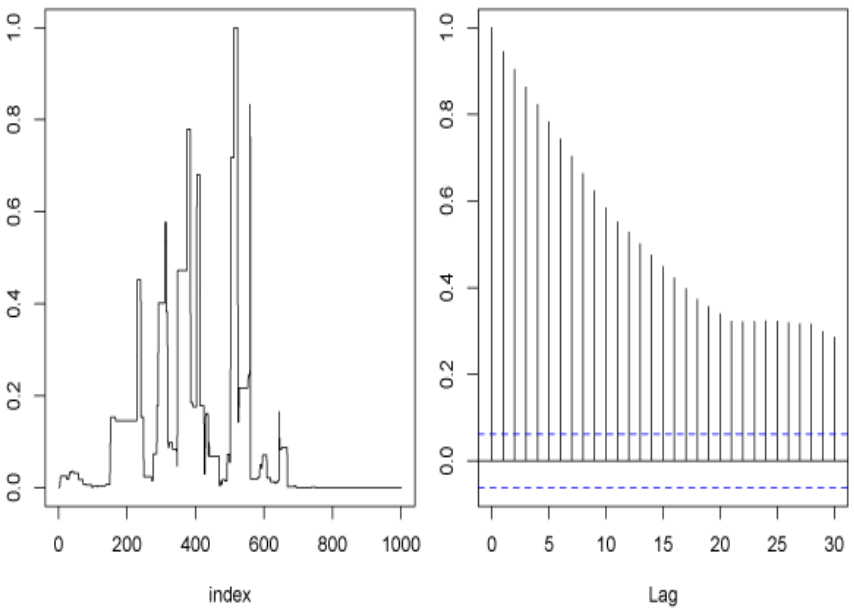
$$\mathbf{q} \sim f(\mathbf{q}|\mathbf{n}_p) = \frac{\Gamma(2n)}{\prod_{j=1}^{p} \Gamma(n_j + 1)} \prod_{j=1}^{p} q_j^{n_j}. \tag{15}$$

Given $\mathbf{q}$, we draw $\mathbf{n}'_{p'}$ from

$$\mathbf{n}'_{p'} \sim P(\mathbf{n}'_{p'}|\mathbf{q}) \sim \frac{\frac{\Gamma(n)}{\Gamma(n+\lambda)} \lambda^{p'} \prod_{j=1}^{p'} \Gamma(n'_j) \binom{n}{n'_1 \cdots n'_{p'}} \prod_{j=1}^{p'} q_j^{n'_j}}{\sum_{\mathbf{n}} \frac{\Gamma(n)}{\Gamma(n+\lambda)} \lambda^{p} \prod_{j=1}^{p} \Gamma(n_j) \binom{n}{n_1 \cdots n_p} \prod_{j=1}^{p} q_j^{n_j}}. \tag{16}$$

It has been established (see Kyung et al. (2010)) that the Ewens-Pitman distribution (8) is the stationary distribution of the transition kernel $K(\mathbf{n}_p, \mathbf{n}'_{p'}) = \int_{\mathbf{q}} P(\mathbf{n}'_{p'}|\mathbf{q}) f(\mathbf{q}|\mathbf{n}_p)\, d\mathbf{q}$. Moreover, to sample from this distribution we can generate a candidate according to the multinomial distribution $\binom{n}{n_1 \cdots n_p} \prod_{j=1}^{p} q_j^{n_j}$ and apply an MH step with ratio $\frac{K(\mathbf{n}_p, \mathbf{n}'_{p'})}{K(\mathbf{n}'_{p'}, \mathbf{n}_p)} = \lambda^{p'-p} \frac{\prod_{j=1}^{p'} \Gamma(n'_j)}{\prod_{j=1}^{p} \Gamma(n_j)}$.

Figure 5: Sample path of MCMC algorithm. The left panel shows 1000 Bayes factors (scaled by the maximum) and the right panel shows the corresponding autocorrelations. The acceptance rate of the Metropolis-Hastings algorithm was 22%.



Finally, we take our search to be a mixture of the biased random walk and the jump, choosing the biased random walk with probability a, set by the user. In our searches we have taken $a = .75$.

Figure 5 shows a sample of the Metropolis-Hastings search used for the first Galaxy example. The acceptance rate was 22%, in line with the recommendations of Roberts et al. (1997).

7 Examples

In this section we give a number of examples to illustrate the working of the clustering algorithm, and examine the effect of the choice of the prior on model space. We start with the well known Galaxy data (Roeder 1990) as a benchmark. We then look at some simulated examples, where we check that the HUP intrinsic Bayesian procedure gives the highest weight to the correct model, and then we look at the effectiveness of the search algorithm. We then apply our method to two data sets, and provide comparisons

with other algorithms.

7.1 Galaxy Data

The Galaxy data consists of 82 observations on the velocity (km/second) of 82 galaxies in the Corona Borealis Region. It is well accepted that there are between 5 and 7 clusters in the data (Richardson and Green 1997; Jasra et al. 2005). Using an intercept-only model, we ran our algorithm with HUP and uniform weights.

The results are displayed in Figure 6, where we see that the search with HUP weights gave five clusters in the top partition with Bayes factor=6.2×10^{10}, while the second partition, which differs by the switching of one galaxy, has Bayes factor=4.9×10^{10}. The top 25 partitions in the search all had 5 clusters. In contrast, the unweighted uniform search found partitions with 9 clusters, which by consensus is too many clusters. From the results in Section 5, this performance is expected.

7.2 Evaluating the Models

Next we evaluate the ability of the HUP intrinsic Bayes procedure to find the best model, regardless of the search algorithm. We look at two examples, for $n = 7$ and $n = 9$, where we can enumerate all of the models. We generate the data from the model

$$Y_{ij} \sim N(\mu_i, 1), \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, k, \quad \mu_i = i, \tag{17}$$

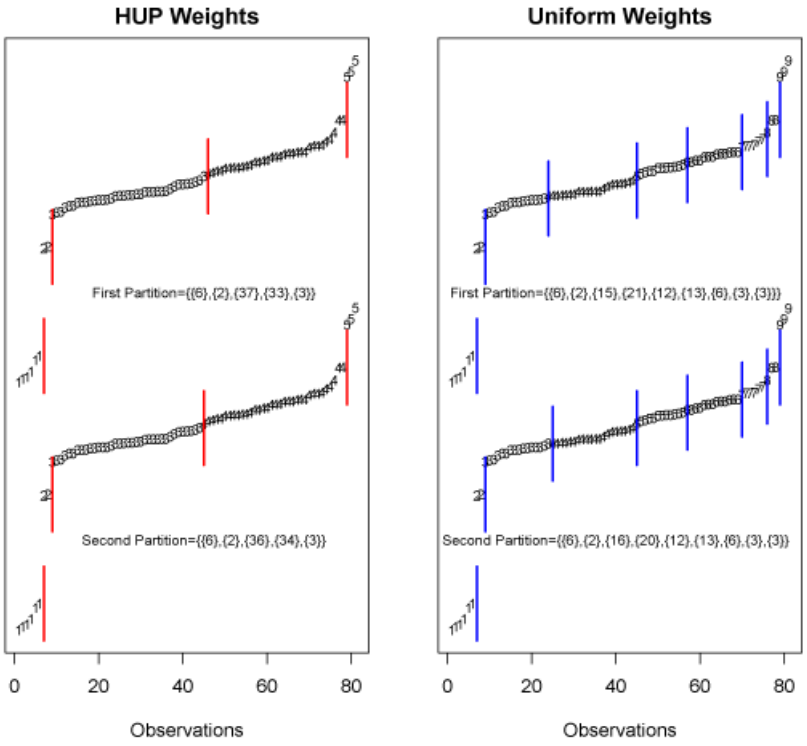
where $\sum_i n_i = n$, $n = 7$ or 9.

For a particular configuration of n and n_i we generated 25 data sets. For each data set we calculated all of the Bayes factors. For $n = 7$ there are 877 distinct partitions, and for $n = 9$ there are $21,147$ distinct partitions. For each of the 25 data sets we checked if the posterior odds of a model using the uniform prior, the Ewens-Pitman prior with $\lambda = 1$, and the HUP prior, was in the top 10 models

The results are shown in Table 4, where we see excellent performance of the Bayes factors with HUP weights. The Ewens-Pitman weight does well except when there are many small clusters, such as $(1, 1, 1, 2, 2)$, being dominated by the HUP weights. However, for these cases the unweighted Bayes factor does the best, reflecting its bias toward partitions with a large number of clusters. However, these are the only cases where the unweighted Bayes factor does better than the HUP weights.

The comparison with K-means shows that K-means has a very difficult time in identifying the true model. Since K-means only returns one partition, we could not calculate percentiles, but instead gave it ten tries to identify the true model. We also helped out K-means by starting with the correct number of clusters, that is, telling it to find a partition with the same number of clusters as the true cluster. Even with this help its performance was well below that of the Bayes factors.

Figure 6: Galaxy data (Roeder 1990), 82 observations. The left panel shows the top two clusters from the search algorithm with HUP weights on the Bayes factors, with red lines delimiting the clusters. The right panel shows the top two clusters from the search algorithm with unweighted Bayes factors, with blue lines delimiting the clusters. In each panel the top and bottom clusters differ by one observation marking a shift point in the long middle string. The stochastic search was run for 50,000 iterations.



7.3 Simulated Regression Data

Next we give an example of the search algorithm using data simulated from the regression model

$$Y_i \sim \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n_i, \quad \varepsilon_i \sim N(0,1) \tag{18}$$

with

$$\begin{aligned} \text{First } n_i \text{ observations} &: \beta_0 = -1, \beta_1 = -2 \\ \text{Second } n_i \text{ observations} &: \beta_0 = 2, \beta_1 = 0 \\ \text{Third } n_i \text{ observations} &: \beta_0 = -1, \beta_1 = 3 \end{aligned}$$

and the x_i are generated uniformly in $(0, 10)$. We actually did a large number of simulations, using data sets of different sizes and different configurations. Here we only present a typical result; the other simulations were very similar.

Table 4: For $n = 7$ and $n = 9$, 25 data sets were generated from the indicated configuration class and Bayes factors were calculated for every partition of $n = 7$ or $n = 9$ observations. For the posterior odds columns, the number is the average percentile of the posterior odds for the true model. For K-means we calculated the proportion of times that K-means found the true partition out of ten tries. K-means was always started at the number of clusters in the true partition.

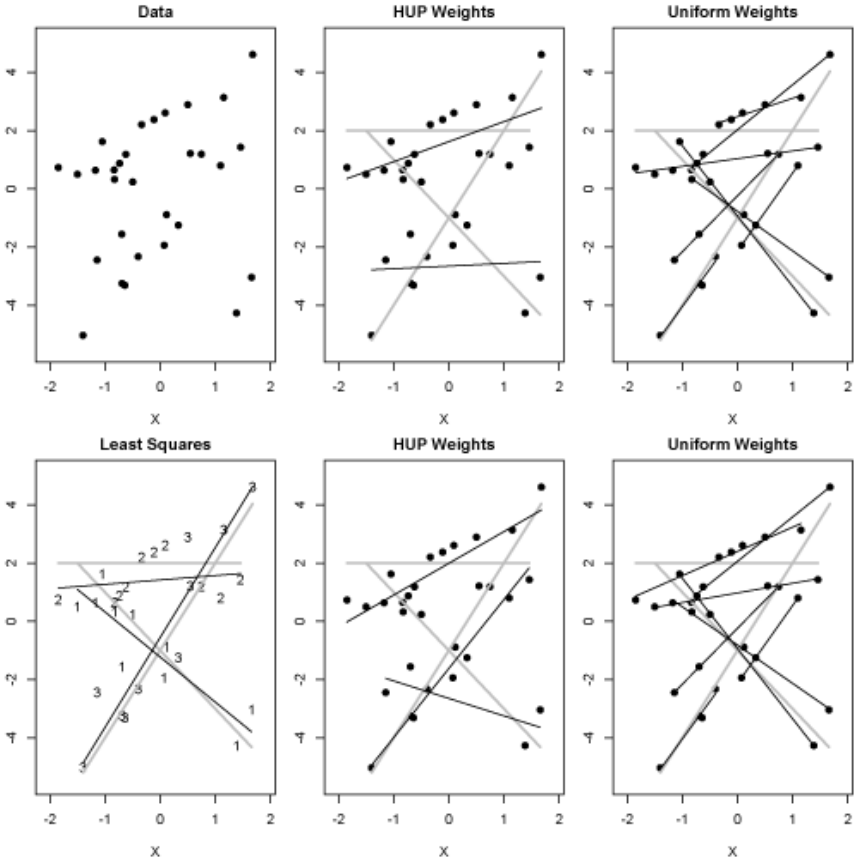
		Posterior Odds			
	Configuration Class	Uniform	Ewens-Pitman	HUP	K-means
$n = 7$	(7)	0.942	0.999	0.999	– – –
	$(3, 4)$	0.919	0.971	0.945	0.440
	$(2, 2, 3)$	0.908	0.826	0.919	0.470
	$(1, 2, 2, 2)$	0.787	0.446	0.713	0.390
	$(1, 1, 1, 2, 2)$	0.426	0.131	0.402	0.450
$n = 9$	(9)	0.949	1.000	1.000	– – –
	$(3, 3, 3)$	0.927	0.922	0.846	0.140
	$(2, 3, 4)$	0.902	0.933	0.909	0.270
	$(2, 2, 2, 3)$	0.967	0.809	0.939	0.160
	$(1, 2, 3, 3)$	0.875	0.755	0.734	0.210
	$(1, 2, 2, 2, 2)$	0.944	0.469	0.891	0.130

Our example has $n_i = 10$, and we use the data shown in Figure 7. The algorithm was run for $50,000$ iterations and representative partitions, from the top 25 using HUP weights, and the top 25 using unweighted Bayes factors, are shown in the Figure.

First note that when looking at the data without clusters identified (upper left panel) it is quite difficult to discern what the true clusters might be. The lower left panel shows that least squares, with knowledge of the true clusters, does reasonably well. In this light, the performance of the Bayes factor with HUP weights is quite remarkable. The top 25 partitions all had either 2 or 3 clusters, and the two partitions that we show are representative. The three cluster partition, in particular, does a very good job of recovering the underlying structure.

Similar to what we saw with the galaxy data, this example shows that searching large data sets without prior weights on the partitions leads to finding partitions with too many clusters. The rightmost panels in Figure 7 are representative of the top 25 partitions from the unweighted Bayes factor search, all of which had seven clusters. The underlying structure is not recovered. As mentioned before, one reason why we are doing cluster analysis is to find partitions with a small number of meaningful clusters. As of now it seems that the best way to accomplish this is to have HUP weights on the

Figure 7: Simulated data ($n = 30$) from model (18). The upper leftmost panel shows the data without cluster labels. The lower leftmost panel shows the cluster-identified data along with the true models (grey) and the least squares fit based on knowing the cluster membership. The middle panels show typical results from the search with HUP weights, and the right panels show typical results from unweighted searches. The stochastic search was run for 50,000 iterations.

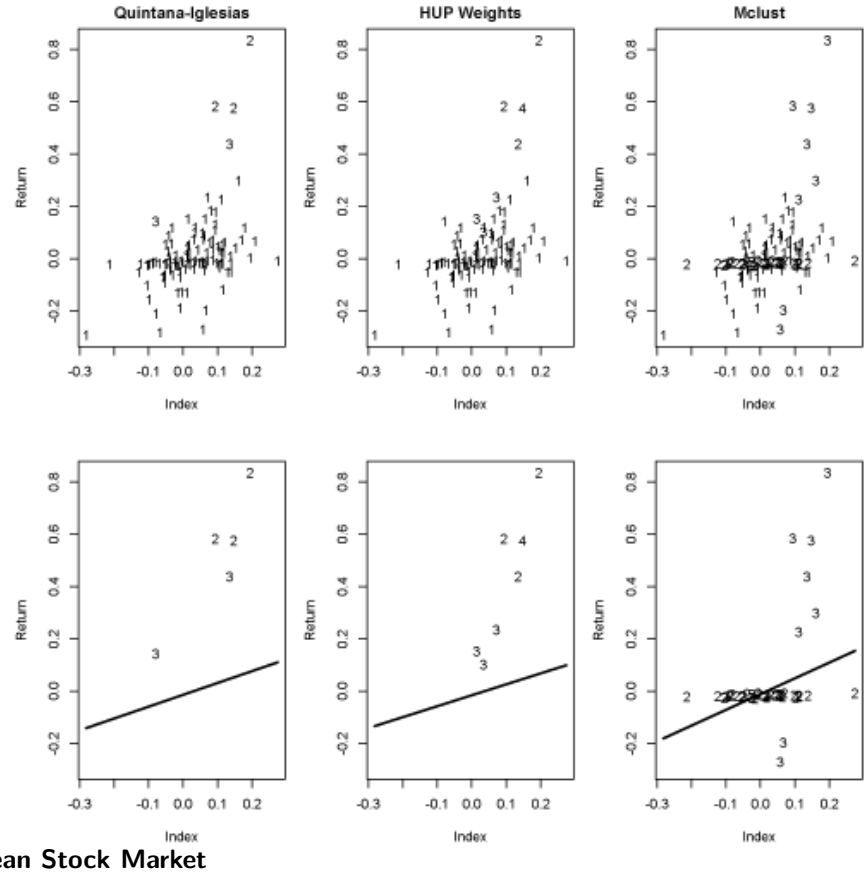


partitions.

7.4 Analyses and Comparisons

In this section we look at two different data sets, from Economics and Biostatistics. We see that in all cases the HUP Intrinsic Bayes procedure performs extremely well, not only giving reasonable answers in its own right, but also comparing favorably with other approaches.

Figure 8: Results from the analysis of the *Concha y Toro* data. From left to right the panels correspond to Quintana and Iglesias, Intrinsic Bayes with HUP weights, and Mclust, respectively. The top panels show scatterplots of the stock index against growth, with the plotting character corresponding to the cluster identifier. The lower panels show the regression fit to the largest cluster, plotting the remaining points as outliers.



Chilean Stock Market

Quintana and Iglesias (2003) (QI) analyze economic data pertaining to the winemaker *Concha y Toro*. This is simple linear regression data, using a model of the form

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where y = the *Concha y Toro* stock return, and x = a stock market index, similar to the US Dow-Jones Index. The data are fully described by QI.

QI use a version of their full PPM model set up for outlier detection, as they are interested in seeing if the *Concha y Toro* returns follow the market. The model they fit has a common slope parameter, and they use individual intercepts to create the clusters.

They use a PPM Gibbs sampler with the algorithm of Bush and MacEachern (1996). For a variety of parameter choices their analysis supports partitioning the data into a small number of clusters, three or four, where one cluster is very large (signifying the data without outliers) and the other clusters are very small, containing the outliers. The partition that they obtained, assuming normal errors, is displayed in Figure 8.

We ran the data with our algorithm using only default settings and HUP weights, that is, we did not tune the model for outliers. The results from that analysis are also shown in Figure 8. We found four clusters, one large one containing the data without outliers, and three others which can clearly be considered outliers. Thus, our default analysis gives results that were very similar to those of QI, with the exception that our algorithm was only set up to find clusters, not specifically to find outliers.

Finally, we also ran Mclust on the data, which also found three clusters. However, it did not find one large one and two small ones but rather two large ones and one small one. The larger cluster from Mclust had a slope similar to the large clusters found by QI and the HUP Bayes procedure, but the second largest cluster found by Mclust was not found by the others; both QI and HUP Bayes put those observations into the first cluster.

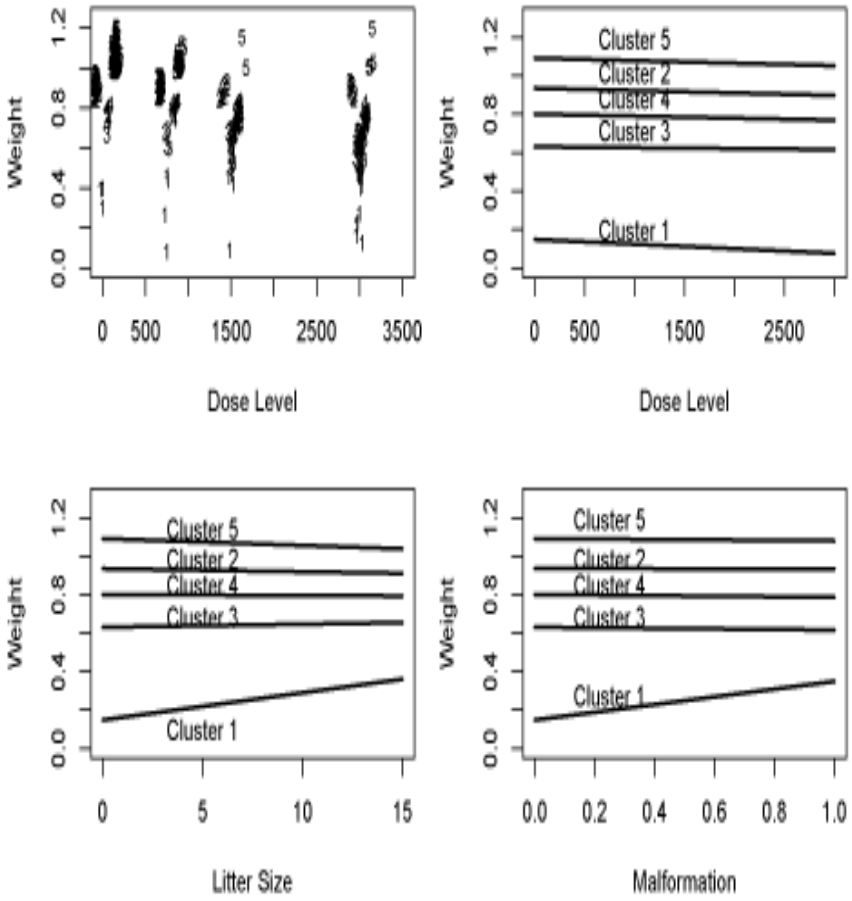
Developmental Toxicology Data

We look at the data analyzed by Dunson et al. (2003), and many other authors (see the references therein). It is data of a developmental toxicity study of ethylene glycol in mice conducted by the National Toxicology Program, and first reported in Price et al. (1985). During pregnancy, mice were exposed to four levels of ethylene glycol (0, 0.75, 1.5, and 3 mg/kg). The response of interest is the fetal weight of the babies. Other covariates were measured but here, like Dunson et al. (2003), we focus on two others (in addition to dose level): litter size, and a $0-1$ indicator of malformations.

After removing some observations with missing data, the remaining data set had $n = 1048$ observations. We ran a stochastic search for $50,000$ iterations using the Bayes factors with HUP weights; we did not use the unweighted Bayes factor due to its previous poor performance. The results are shown in Figure 9, which shows a partition that is typical of the top 10. All of the top 10 partitions had five clusters, and there was little difference among them.

The five clusters are quite interesting, showing that the partition aligns the mice on the intercepts, with fetal weight increasing as we move from Cluster 1 to Cluster 5 and with the slopes having little effect. The effect of dose level is similar in the clusters (parallel lines), decreasing the fetal weight slightly at higher doses. However, it is clear that in this partition the effect of the ethylene glycol dose is independent of the fetal weight. Also, the effect of litter size on fetal weight is also minimal in the partition. The only substantial slope effects are in Cluster 1 for the litter size and malformation. There we see an effect of increasing weight with litter size, and that zero malformations align with the lower fetal weight mice, and increased malformations are associated with higher fetal weights. However, in the other clusters, where all fetal weights are higher,

Figure 9: Results from the cluster analysis of the developmental toxicology data. Shown are the results of the partition with the highest posterior odds; the top 10 partitions all had five clusters. The upper left panel shows the data identified by cluster assignment, and the remaining panels show the slopes of the clusters for each of the three explanatory variables. The stochastic search was run for $50,000$ iterations. The cluster sizes are $(23, 341, 187, 301, 196)$. (For clarity, not all points are plotted in the clusters, and they are jiggled.)



there is no effect due to malformation.

8 Discussion

We have presented an objective Bayesian analysis for clustering based on product partition models using a model selection approach. Our major finding concerns the sensitivity of the procedure to the choice of the prior on model space.

We have seen that the "default" uniform prior on models and the Jensen-Liu prior lead not only to inconsistent procedures, but also to small sample performance that is not desirable. The Ewens-Pitman prior and the hierarchical uniform prior lead to Bayesian consistent procedures under the null and weakly consistent under the alternative. In the absence of prior information our preferred prior over the models is the hierarchical uniform prior that arises from a decomposition of the set of partitions of the sample in classes dictated by the number of clusters. Each class has also been split in subclasses where in each subclass we group all the partitions that only differ by a permutation of the components of the sample, and thus come from the same sampling model. This alleviates the difficulty of assigning a prior to the partitions inside the classes. A truncated Poisson-Intrinsic prior has been chosen for the number of clusters; it gives decreasing probability to partitions with higher numbers of clusters, and has performed well in our evaluations.

We also note the following about clustering priors:

1. Cluster analysis is useful, and will only result in useful inferences, when the answer contains a relatively small number of clusters. The prior should move us toward partitions with a small number of clusters, so the clusters themselves are large. Some advantages of the Ewens-Pitman (λ) prior in clustering are that it is easy to implement in Gibbs sampling schemes, and the expected number of clusters is known to be an increasing function of λ which permits to elicit the parameter assuming that meaningful prior information on the number of clusters is available.

2. Even if the truth is that there are a large number of clusters (say 500 observations have 70 true clusters) this results in a useless inference. In such a case it is better to find partitions with a small number of clusters that explain a large portion of the variability.

3. The limit results of Theorem 2 point out the behavior of the four priors considered. The fact that only the HUP converges to a proper prior tells us that it is correctly compensating for the increasing number of models, which the other priors do not do. Furthermore, even when the Ewens-Pitman prior is in the limit an improper distribution, the limit of its posterior distribution is proper.

   Other points that we would like to emphasize are:

4. The findings in our examples are consistent with the theory. The HUP produces partitions with a small number of clusters, while the unweighted Bayes factors almost always return a configuration with a large number of small clusters.

5. We can apply some of our previous results (Casella et al. 2009; Moreno et al. 2010) to show that our procedure is consistent for choosing between two nested cluster

models, as for instance $M_{\mathbf{r}_p}$ and $M_{\mathbf{r}_1}$, assuming that one of them is the true one. In such a pairwise comparison the prior on model space plays no role, as there are only two models. Pairwise consistency holds when the number of clusters grows at the rate $p = O(n^\alpha)$ for $\alpha < 1$, and for $\alpha = 1$ when the models $M_{\mathbf{r}_1}$ and $M_{\mathbf{r}_p}$ are not too close, that is when $\delta_{\mathbf{r}_p\mathbf{r}_1}$ is greater than a positive number which can be specified (Moreno et al. 2010).

6. Lemma 1 shows that for large n and when sampling from $M_{\mathbf{r}_1}$, the Bayes factor $B_{\mathbf{r}_p\mathbf{r}_1}(\mathbf{y})$ is constant across partitions $\mathbf{r}_p$ in the class $\mathfrak{R}_p$. This asymptotic result is not limited to intrinsic Bayes factors; for example we know that BIC is asymptotically equivalent to the intrinsic Bayes factor. Moreover, Casella et al. (2009) show that the approximation holds for a wide class of priors.

7. When sampling from a model $M_{\mathbf{r}_t}$ different from $M_{\mathbf{r}_1}$the asymptotic approximation of a generic Bayes factor $B_{\mathbf{r}_p\mathbf{r}_1}(\mathbf{y})$ is no longer constant across partitions in $\mathfrak{R}_p$ but depends on the pseudo-distances between the true model $M_{\mathbf{r}_t}$ to any $M_{\mathbf{r}_p}$, which in turn depends on the design matrices of the models. In this case consistency involves very stringent conditions on the design matrices and distances which have a difficult interpretation. However, a weaker form of consistency under the alternative is also satisfied for either the Ewens-Pitman or the HUP over the models.

References

Aldous, D. J. (1985). "Exchangeability and related topics." In Hennequin, P. (ed.), *École d'Été de Probabilités de Saint-Flour XIII â 1983*, volume 1117 of *Lecture Notes in Mathematics*, 1–198. Springer Berlin Heidelberg.
URL http://dx.doi.org/10.1007/BFb0099421 620

Barry, D. and Hartigan, J. A. (1992). "Product Partition Models for Change Point Problems." *The Annals of Statistics*, 20(1): 260–279.
URL http://dx.doi.org/10.1214/aos/1176348521 614

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). "The formal definition of reference priors." *The Annals of Statistics*, 37(2): 905–938.
URL http://dx.doi.org/10.1214/07-AOS587 623

Berger, J. O. and Mortera, J. (1999). "Default Bayes Factors for Nonnested Hypothesis Testing." *Journal of the American Statistical Association*, 94(446): 542–554.
URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10474149 623

Berger, J. O. and Pericchi, L. R. (1996a). "The intrinsic Bayes factor for linear models." In *Bayesian Statistics 5*, 23–42. New York: Oxford University Press. 623

— (1996b). "The Intrinsic Bayes Factor for Model Selection and Prediction." *Journal of the American Statistical Association*, 91(433): 109–122.

URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476668 623

Booth, J. G., Casella, G., and Hobert, J. P. (2008). "Clustering using objective functions and stochastic search." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1): 119–139.
URL http://dx.doi.org/10.1111/j.1467-9868.2007.00629.x 614, 616, 620, 633

Bush, C. A. and MacEachern, S. N. (1996). "A semiparametric Bayesian model for randomised block designs." *Biometrika*, 83(2): 275–285.
URL http://biomet.oxfordjournals.org/content/83/2/275.abstract 641

Casella, G., Girón, F. J., Martínez, M. L., and Moreno, E. (2009). "Consistency of Bayesian procedures for variable selection." *The Annals of Statistics*, 37(3): 1207–1228.
URL http://dx.doi.org/10.1214/08-AOS606 623, 628, 643, 644

Casella, G. and Moreno, E. (2005). "Intrinsic meta-analysis of contingency tables." *Statistics in Medicine*, 24(4): 583–604.
URL http://dx.doi.org/10.1002/sim.2038 623

— (2009). "Assessing Robustness of Intrinsic Tests of Independence in Two-Way Contingency Tables." *Journal of the American Statistical Association*, 104(487): 1261–1271.
URL http://www.tandfonline.com/doi/abs/10.1198/jasa.2009.tm08106 623

Crowley, E. M. (1997). "Product Partition Models for Normal Means." *Journal of the American Statistical Association*, 92(437): 192–198.
URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1997.10473616 620

Dunson, D. B., Chen, Z., and Harry, J. (2003). "A Bayesian Approach for Joint Modeling of Cluster Size and Subunit-Specific Outcomes." *Biometrics*, 59(3): 521–530.
URL http://dx.doi.org/10.1111/1541-0420.00062 641

Ewens, W. (1972). "The sampling theory of selectively neutral alleles." *Theoretical Population Biology*, 3(1): 87 – 112.
URL http://www.sciencedirect.com/science/article/pii/0040580972900354 620

Fraley, C. and Raftery, A. E. (2002). "Model-Based Clustering, Discriminant Analysis, and Density Estimation." *Journal of the American Statistical Association*, 97(458): 611–631.
URL http://www.tandfonline.com/doi/abs/10.1198/016214502760047131 614

— (2007). "Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering." *Journal of Classification*, 24(2): 155–181.
URL http://dx.doi.org/10.1007/s00357-007-0004-5 614

Girón, F. J., Martínez, M. L., Moreno, E., and Torres, F. (2006). "Objective Testing Procedures in Linear Models: Calibration of the p-values." *Scandinavian Journal of Statistics*, 33(4): 765–784.
URL http://dx.doi.org/10.1111/j.1467-9469.2006.00514.x 623

Hartigan, J. (1990). "Partition models." *Communications in Statistics - Theory and Methods*, 19(8): 2745–2756.
URL http://www.tandfonline.com/doi/abs/10.1080/03610929008830345 614

Ishwaran, H. and Zarepour, M. (2002). "Dirichlet Prior Sieves in Finite Normal Mixtures." *Statistica Sinica*, 12: 941–963. 620

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). "Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling." *Statistical Science*, 20(1): 50–67.
URL http://dx.doi.org/10.1214/088342305000000016 636

Jensen, S. T. and Liu, J. S. (2008). "Bayesian Clustering of Transcription Factor Binding Motifs." *Journal of the American Statistical Association*, 103(481): 188–200.
URL http://www.tandfonline.com/doi/abs/10.1198/016214507000000365 620

Kim, S. and Sun, D. (2000). "Intrinsic Priors for Model Selection Using an Encompassing Model with Applications to Censored Failure Time Data." *Lifetime Data Analysis*, 6(3): 251–269. 623

Kyung, M., Gill, J., and Casella, G. (2010). "Estimation in Dirichlet random effects models." *The Annals of Statistics*, 38(2): 979–1009.
URL http://dx.doi.org/10.1214/09-AOS731 634

Lau, J. W. and Green, P. J. (2007). "Bayesian Model-Based Clustering Procedures." *Journal of Computational and Graphical Statistics*, 16(3): 526–558.
URL http://www.tandfonline.com/doi/abs/10.1198/106186007X238855 614, 615

McCullagh, P. and Yang, J. (2006). "Stochastic classification models." In *Proceedings of the International Congress of Mathematicians*, 669–686. Madrid: Springer. 620

Moreno, E. (1997). "Bayes factors for intrinsic and fractional priors in nested models: Bayesian robustness." In *L_1-Statistical Procedures and Related Topics*, 257–270. Hayward, CA: Institute of Mathematical Statistics. 623

— (2005). "Objective Bayesian methods for one-sided testing." *Test*, 14(1): 181–198.
URL http://dx.doi.org/10.1007/BF02595402 621, 623

Moreno, E., Bertolino, F., and Racugno, W. (1998). "An Intrinsic Limiting Procedure for Model Selection and Hypotheses Testing." *Journal of the American Statistical Association*, 93(444): 1451–1460.
URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1998.10473805 623

Moreno, E., Girón, F. J., and Casella, G. (2010). "Consistency of objective Bayes factors as the model dimension grows." *The Annals of Statistics*, 38(4): 1937–1952.
URL http://dx.doi.org/10.1214/09-AOS754 623, 627, 643, 644

Neal, R. M. (2000). "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265.
URL http://amstat.tandfonline.com/doi/abs/10.1080/10618600.2000.10474879 634

Pitman, J. (1996). *Some developments of the Blackwell-MacQueen urn scheme*, volume Volume 30 of *Lecture Notes–Monograph Series*, 245–267. Hayward, CA: Institute of Mathematical Statistics.
URL http://dx.doi.org/10.1214/lnms/1215453576 620

Postman, M., Huchra, J. P., and Geller, M. J. (1986). "Probes of large-scale structure in the Corona Borealis region." *Astronomical Journal*, 92: 1238–1247. 615

Price, C. J., Kimmel, C. A., Tyl, R. W., and Marr, M. C. (1985). "The developmental toxicity of ethylene glycol in rats and mice." *Toxicology and Applied Pharmacology*, 81(1): 113 – 127.
URL http://www.sciencedirect.com/science/article/pii/0041008X85901267 641

Quintana, F. A. and Iglesias, P. L. (2003). "Bayesian clustering and product partition models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2): 557–574.
URL http://dx.doi.org/10.1111/1467-9868.00402 614, 620, 622, 640

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 (Statistical Methodology)*, 59(4): 731–792.
URL http://dx.doi.org/10.1111/1467-9868.00095 615, 636

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). "Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms." *The Annals of Applied Probability*, 7(1): pp. 110–120.
URL http://www.jstor.org/stable/2245134 635

Roeder, K. (1990). "Density Estimation with Confidence Sets Exemplified by Superclusters and Voids in the Galaxies." *Journal of the American Statistical Association*, 85(411): 617–624.
URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10474918 615, 635, 637

Roeder, K. and Wasserman, L. (1997). "Practical Bayesian Density Estimation Using Mixtures of Normals." *Journal of the American Statistical Association*, 92(439): 894–902.
URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1997.10474044 615

Wang, L. and Dunson, D. B. (2011). "Fast Bayesian Inference in Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 20(1): 196–216.
URL http://www.tandfonline.com/doi/abs/10.1198/jcgs.2010.07081 615

Appendix

Appendix A Technical Details

In this section we prove the asymptotic results in Theorems 2, 3, and 4. We begin with some preliminary lemmas that are needed, and then give detailed proofs of the theorems.

A.1 Preliminary Lemmas

Lemma 2. *For every $p = 1, \ldots,$ under the conditions $n_1 \leq \cdots \leq n_p$, and $\sum_{i=1}^{p} n_i = n$, the sum of the series*

$$\sum \binom{n}{n_1 \ldots n_p} (\lambda+1)^{-n_1} \ldots (\lambda+p)^{-n_p} \approx O\left(\frac{\lambda_p^H}{p}\right)^{-n},$$

where λ_p^H is the harmonic mean of $(\lambda+1)^{-1}, \ldots, (\lambda+p)^{-1}$.

Proof. To evaluate the sum, multiply and divide by $(\sum_{i=1}^{p} \frac{1}{\lambda+i})^n$. Sum the resulting multinomial to get

$$\sum \binom{n}{n_1 \cdots n_p} (\lambda+1)^{-n_1} \ldots (\lambda+p)^{-n_p} = \left(\sum_{i=1}^{p} \frac{1}{\lambda+i}\right)^n \approx O\left(\frac{\lambda_p^H}{p}\right)^{-n}.$$

□

Lemma 3. *For every p the sum*

$$S = \sum_{C_p} \frac{1}{n_1 \times \cdots \times n_p},$$

where $C_p = \{(n_1, ..., n_p) : n_1 \leq \cdots \leq n_p, \sum_{i=1}^{p} n_i = n\}$, is of order $O(n^{-1}(\log n)^{p-1})$.

Proof. Denote $x_i = n_i/n$, for $i = 1, \ldots, p-1$. Then we can write the sum S as

$$S = n^{-p} \sum \frac{1}{x_1 \cdot x_2 \cdots (1 - x_1 - x_2 - \cdots - x_{p-1})},$$

where the multiple sum is extended for $x_i = 1/n, \ldots, 1 - 1/n$ in steps of size $1/n$, for $i = 1, \ldots, p-1$. Now, the sum can be approximated, for large values of n, by the multiple integral

$$n^{p-1} \int_{1/n}^{1-1/n} \cdots \int_{1/n}^{1-1/n} \frac{1}{x_1 \cdot x_2 \cdots (1 - x_1 - x_2 - \cdots - x_{p-1})} \, dx_1 \ldots dx_{p-1} = n^{p-1}\mathbf{I},$$

where the factor n^{p-1} is the adjustment due to the $1/n$ step for each variable, and $\mathbf{I}$ is the integral.. Thus, the original sum is

$$\sum \frac{1}{n_1 \cdot \cdots \cdot n_p} \approx n^{-1}\mathbf{I}.$$

To evaluate the integral, consider the change of variables

$$\theta_1 \quad = \quad x_1, \quad \theta_i = \frac{x_i}{1 - \sum_{j=1}^{i-1} x_j}, \quad i = 2, \ldots, p-1.$$

As the Jacobean J of the variables $(x_1, \ldots, x_{p-1})$ with respect to the new variables $(\theta_1, \ldots, \theta_{p-1})$ is $J = \prod_{i=1}^{p-1} \left(1 - \sum_{j=1}^{i-1} x_j\right)$, the integral $\mathbf{I}$ in terms of the new variables can be written as

$$\mathbf{I} = \int_{\frac{1}{n}}^{1-\frac{1}{n}} \cdots \int_{\frac{1}{n-p+2}}^{1-\frac{1}{n-p+2}} \frac{1}{\theta_1 \cdot \theta_2 \cdot \cdots \cdot \theta_{p-1}} \, d\theta_1 \ldots d\theta_{p-1}$$

which is equal to

$$\mathbf{I} = \prod_{i=1}^{p-1} \int_{\frac{1}{n-i+1}}^{1-\frac{1}{n-i+1}} \frac{1}{\theta_i} \, d\theta_i = \prod_{i=1}^{p-1} \log(n-i) \sim O((\log n)^{p-1}).$$

□

A.2 Proof of Theorem 2

For part (a), the proof is immediate as the distribution of $\mathbf{Y}_n$ on the simplex $\mathcal{S}_p$ is a discrete uniform distribution on the points of the simplex of the form $\left(\frac{n_1}{n}, \ldots, \frac{n_p}{n}\right)$, where $n_1 \leq \cdots \leq n_p$ and $n_1 + \cdots + n_p = 1$. From the definition of the hierarchical uniform prior,

$$\Pr(\mathbf{Y}_n = (\frac{n_1}{n}, \ldots, \frac{n_p}{n}) | \mathfrak{R}_p) = \Pr(\mathfrak{R}_{p;n_1,\ldots,n_p} | \mathfrak{R}_p) = \frac{1}{b(n,p)} \approx \frac{p!}{n^{p-1}},$$

and this discrete uniform prior on the lattice

$$L_p = \{(\frac{n_1}{n}, \ldots, \frac{n_p}{n}), \quad n_1 \leq \cdots \leq n_p, \quad n_1 + \cdots + n_p = 1,$$

converges *in distribution* to a continuous uniform distribution on the simplex, which is a Dirichlet $\mathcal{D}(1, \ldots, 1)$ truncated at the simplex $\mathcal{S}_p$.

For part (b), for the Ewens-Pitman prior, it follows that the prior probability of $\mathfrak{R}_{p;n_1,\ldots,n_p}$, given p and λ, is

$$\Pr(\mathfrak{R}_{p;n_1 \ldots n_p} | \mathfrak{R}_p) \propto \binom{n}{n_1 \cdots n_p} \times \prod_{i=1}^{p} \Gamma(n_i) = \frac{1}{n_1 \cdot \cdots \cdot n_p}.$$

Thus, the distribution of $\mathbf{Y}_n$ on the simplex $\mathcal{S}_p$ is

$$\Pr\Big(\mathbf{Y}_n = \Big(\frac{n_1}{n}, \ldots, \frac{n_p}{n}\Big)\Big|\mathcal{S}_p\Big) \propto \frac{1}{\frac{n_1}{n} \cdot \ldots \cdot \frac{n_p}{n}}.$$

It is clear from the form of the probability mass function of the discrete random vector $\mathbf{Y}_n$, that the limiting distribution is given by the function (2).

For part (c), we can discard the redundancy term $R(n_1, \ldots, n_p)$ for large n, as it is of an order of magnitude much smaller than $\binom{n}{n_1 \ldots n_p}$. We then write the prior probability of $\mathfrak{R}_{p;n_1,\ldots,n_p}$ as

$$\Pr(\mathfrak{R}_{p;n_1,\ldots,n_p}|\mathfrak{R}_p) \propto \binom{n}{n_1 \cdots n_p}.$$

This implies that

$$\Pr(\mathbf{Z}_n = (n_1, \ldots, n_p)|\mathfrak{R}_p) \propto \binom{n}{n_1 \cdots n_p}.$$

But, because of the proportionality symbol, we can write the preceding as

$$\Pr(\mathbf{Z}_n = (n_1, \ldots, n_p)|\mathfrak{R}_p) \propto \binom{n}{n_1 \cdots n_p}\left(\frac{1}{p}\right)^n = \binom{n}{n_1 \cdots n_p}\left(\frac{1}{p}\right)^{n_1} \cdots \left(\frac{1}{p}\right)^{n_p},$$

and this means that the unrestricted $\mathbf{Z}_n$ follows a multinomial distribution with parameters n and $\boldsymbol{p} = (1/p, \ldots, 1/p)$.

For large n, because of the restriction $n_1 + \cdots + n_p = n$, this multinomial distribution can be approximated by a multivariate normal distribution with the same mean vector and covariance matrix, that is,

$$\mathbf{Z}_n \approx N_p(n(1/p, \ldots, 1/p)^t, \Sigma_p),$$

where the covariance matrix $\Sigma_p = (n/p^2)(p\mathbf{I} - \mathbf{J})$, where $\mathbf{J}$ is a matrix of ones. Thus, the distribution of $\mathbf{Y}_n$ can be approximated by the following multivariate normal distribution

$$\mathbf{Y}_n \approx N_p\Big((1/p, \ldots, 1/p)^t, \frac{1}{n^2}\Sigma_p\Big),$$

which converges to the degenerate random variable at mean vector $(\frac{1}{p}, \ldots, \frac{1}{p})$, as the covariance matrix of $\mathbf{Y}_n$ converges to the null matrix.

For part (d), from the expression of the Jensen-Liu prior (9), it follows that the prior probability of $\mathfrak{R}_{p;n_1,\ldots,n_p}$, given p and λ, can be written

$$\Pr(\mathbf{Z}_n = (n_1, \ldots, n_p)|\mathfrak{R}_p) = \Pr(\mathfrak{R}_{p;n_1\ldots n_p}|\mathfrak{R}_p) \propto \binom{n}{n_1 \cdots n_p}\prod_{i=1}^{p}\left(\frac{(\lambda+i)^{-1}}{\sum_{i'}(\lambda+i')^{-1}}\right)^{n_i}.$$

Thus, the distribution of $\mathbf{Z}_n$ is the following multinomial distribution

$$\mathbf{Z}_n \sim \mathcal{M}\left(n; \frac{(\lambda+1)^{-1}}{\sum_{i'}(\lambda+i')^{-1}}, \ldots, \frac{(\lambda+p)^{-1}}{\sum_{i'}(\lambda+i')^{-1}}\right),$$

and, for large n, the distribution of $\mathbf{Y}_n$ can be approximated by the following multivariate normal distribution

$$\mathbf{Y}_n \approx N_p\left(\frac{(\lambda+1)^{-1}}{\sum_{i'}(\lambda+i')^{-1}},\ldots,\frac{(\lambda+p)^{-1}}{\sum_{i'}(\lambda+i')^{-1}}\right)', \frac{1}{n^2}\Sigma_{p,\lambda}\Big),$$

which converges to a random variable degenerate at the mean vector as the covariance matrix of $\mathbf{Y}_n$ converges to the null matrix.

Note. The Jensen-Liu prior restricted to the simplex shows a similar behavior to the uniform prior on the set of all cluster models with at most T clusters, and both are inconsistent. Indeed, when λ goes to infinity, the Jensen-Liu prior converges to the Uniform prior on the set of all cluster models for all $p \leq T$.

A.3 Proof of Theorem 3

Uniform Prior First, we will prove that the posterior probability of $\mathfrak{R}_T$ converges to 1, when n goes to infinity. The uniform prior on the set of all models results in the prior distribution on the cluster classes $\Pr(\mathfrak{R}_p) \propto S_n^{(p)}$, which can be approximated for large n by $\Pr(\mathfrak{R}_p) \approx \frac{p^n}{p!}$.

Applying Bayes theorem, we have that, for $p = 1,\ldots,T$

$$\Pr(\mathfrak{R}_p|\mathbf{y}) \propto \Pr(\mathfrak{R}_p) \times B_{p\,1}(\mathbf{y}) \propto \frac{p^n}{p!}\left(\frac{k\,p+1}{n}\right)^{k(p-1)/2}.$$

Therefore, normalizing the Bayes factors, we have that as $n \to \infty$,

$$\Pr(\mathfrak{R}_p|\mathbf{y}) = \frac{\frac{p^n}{p!}\left(\frac{k\,p+1}{n}\right)^{k(p-1)/2}}{\sum_{p=1}^{T}\frac{p^n}{p!}\left(\frac{k\,p+1}{n}\right)^{k(p-1)/2}} \to \begin{cases} 0 & \text{for } p = 1,\ldots,T-1, \\ 1 & \text{for } p = T. \end{cases}$$

To show inconsistency in the configuration classes, we have to show that within $\mathfrak{R}_T$ the posterior distribution of the models does not converge in distribution to the degenerate distribution corresponding to the vertex $\mathfrak{R}_1$. In fact, from Theorem 2, under the uniform prior on the set of all models, this posterior distribution converges to the degenerate distribution on the equal size T clusters $\delta_{(\frac{1}{T},\ldots,\frac{1}{T})} \in \mathfrak{R}_T$. This completes the proof for the uniform distribution.

Jensen-Liu From (9) it is easy to see that the marginal prior of p or, equivalently $\mathfrak{R}_p$, is

$$\Pr(\mathfrak{R}_p) \propto \lambda^{p-1}(\lambda+p)\sum\binom{n}{n_1\ldots n_p}(\lambda+1)^{-n_1}\ldots(\lambda+p)^{-n_p}.$$

Using the asymptotic approximation of Lemma 2 the prior can be approximated by

$$\Pr(\mathfrak{R}_p) \propto \lambda^{p-1}(\lambda+p)\left(\frac{\lambda_p^H}{p}\right)^{-n}.$$

Recalling the expression for the Bayes factor given in Lemma 1, we have from Bayes Theorem, for $p = 1, \ldots, T$,

$$\Pr(\mathfrak{R}_p|\mathbf{y}) \propto \Pr(\mathfrak{R}_p) \times B_{p\,1}(\mathbf{y}) \propto \lambda^{p-1}(\lambda + p)\left(\frac{\lambda_p^H}{p}\right)^{-n}\left(\frac{k\,p+1}{n}\right)^{k(p-1)/2}.$$

The leading term of the posterior probability is $(\lambda_p^H/p)^{-n}$, which is an increasing function of p for every n. Normalizing the posterior probabilities we finally get

$$\Pr(\mathfrak{R}_p|\mathbf{y}) \to \begin{cases} 0 & \text{for } p = 1, \ldots, T-1, \\ 1 & \text{for } p = T. \end{cases}$$

It is also the case that within the cluster class $\mathfrak{R}_T$ the posterior distribution of the models does not converge in distribution to the degenerate distribution corresponding to the vertex $\mathfrak{R}_1$. In fact, from Theorem 2, this posterior distribution converges to the degenerate distribution of the model $M_{\left(\frac{\lambda_T^H}{p(\lambda+T)},\ldots,\frac{\lambda_T^H}{p(\lambda+1)}\right)} \in \mathfrak{R}_T$. This completes the proof.

Note that the limiting distribution in this case depends on the value of the paramenter λ of the Jensen-Liu prior. Further, as λ increases, the limiting distribution converges to the T-cluster model with equal size clusters.

A.4 Proof of Theorem 4

Hierarchical Uniform Prior We will prove that the posterior distribution of $\mathfrak{R}_1$ converges to 1, when n goes to infinity. By Bayes theorem, and the fact that the prior over the cluster classes $\mathfrak{R}_1, \ldots, \mathfrak{R}_p$ is uniform, we have that, for $p = 1, \ldots, T$,

$$\Pr(\mathfrak{R}_p|\mathbf{y}) \propto Pr(\mathfrak{R}_p) \times B_{p\,1}(\mathbf{y}) \propto B_{p\,1}(\mathbf{y}) = \left(\frac{k\,p+1}{n}\right)^{k(p-1)/2}.$$

Therefore, normalizing the Bayes factors, we have that, as $n \to \infty$,

$$\Pr(\mathfrak{R}_p|\mathbf{y}) = \frac{\left(\frac{k\,p+1}{n}\right)^{k(p-1)/2}}{\sum_{p=1}^{T}\left(\frac{k\,p+1}{n}\right)^{k(p-1)/2}} \to \begin{cases} 0 & \text{for } p = 2, \ldots, T, \\ 1 & \text{for } p = 1. \end{cases}$$

Ewens-Pitman Prior From (8), the joint prior for p and $\mathbf{r}_p$, we can calculate the marginal prior of p or, equivalently $\mathfrak{R}_p$, as

$$\Pr(\mathfrak{R}_p) \propto \lambda^{p-1} n^{p-1} \sum \frac{1}{n_1 \cdot \ldots \cdot n_p}.$$

Again recalling Lemma 1, Bayes Theorem yields, for $p = 1, \ldots, T$,

$$\Pr(\mathfrak{R}_p|\mathbf{y}) \propto \Pr(\mathfrak{R}_p) \times B_{p\,1}(\mathbf{y}) \propto \lambda^{p-1} n^{p-1} \sum \frac{1}{n_1 \cdot \ldots \cdot n_p}\left(\frac{k\,p+1}{n}\right)^{k(p-1)/2}.$$

Using Lemma 3, we can approximate the posterior by

$$\Pr(\mathfrak{R}_p|\mathbf{y}) \approx C_p \lambda^{p-1} n^{-1} (\log n)^{p-1} \left(\frac{k\,p+1}{n}\right)^{k(p-1)/2} \approx K_{p,k,\lambda}\, n^{-1} (n^{-k/2} \log n)^{p-1},$$

where C_p is a finite positive constant depending on p and $K_{p,k,\lambda}$ is a positive constant depending on p, k and λ. It is now clear that, as $n \to \infty$,

$$\Pr(\mathfrak{R}_p|\mathbf{y}) \to \begin{cases} 0 & \text{for } p = 2, \ldots, T, \\ 1 & \text{for } p = 1. \end{cases}$$

A.5 Proof of Theorem 5

In fact, we prove that for any large but fixed sample size n, the posterior probability of the true model $M_{\mathbf{r}_t}$ given the data $\mathbf{y}$, $\Pr(M_{\mathbf{r}_t}|\mathbf{y}, n)$ can be made as small as desired by choosing n sufficiently large. In fact,

$$\Pr(M_{\mathbf{r}_t}|\mathbf{y}, n) = \frac{1}{1 + S_T(n)},\ [M_{\mathbf{r}_t}],$$

where

$$S_T(n) = \sum_{p=1}^{T} \sum_{\substack{n_1 \le \cdots \le n_p \\ n_1 + \cdots + n_p = n}} \sum_{\substack{\mathbf{r}_p \in \mathfrak{R}_{p;n_1,\ldots,n_p} \\ \mathbf{r}_p \neq \mathbf{r}_t}} B_{\mathbf{r}_p \mathbf{r}_t}(\mathbf{y})\ \frac{\pi(p, \mathbf{r}_p|n)}{\pi(t, \mathbf{r}_t|n)}.$$

Under $M_{\mathbf{r}_t}$, and using Lemma 1, the Bayes factor $B_{\mathbf{r}'_t \mathbf{r}_t}(\mathbf{y}) = B_{\mathbf{r}'_t \mathbf{r}_1}(\mathbf{y}) / B_{\mathbf{r}_t \mathbf{r}_1}(\mathbf{y})$ can be approximated for large values of n, by

$$B_{\mathbf{r}'_t \mathbf{r}_t}(\mathbf{y}) \approx \left(\frac{1 + \delta_{\mathbf{r}_t \mathbf{r}_t}}{1 + \delta_{\mathbf{r}_t \mathbf{r}'_t}}\right)^{n/2} = 1,$$

as the pseudodistance $\delta_{\mathbf{r}_t \mathbf{r}'_t}$ from the true model $M_{\mathbf{r}_t}$ to any other model $M_{\mathbf{r}'_t}$, in the configuration class $\mathfrak{R}_{t;n_1 \cdots n_t}$ is equal to 0. Further, for any exchangeable prior on the models $\pi(p, \mathbf{r}_p|n)$ we have that

$$\frac{\pi(t, \mathbf{r}'_t|n)}{\pi(t, \mathbf{r}_t|n)} = 1,\ \text{for } \mathbf{r}'_t, \mathbf{r}_t \in \mathfrak{R}_{t;n_1 \cdots n_t},$$

and hence

$$S_T(n) > \sum_{\substack{\mathbf{r}'_t \in \mathfrak{R}_{t;n_1 \cdots n_t} \\ \mathbf{r}'_t \neq \mathbf{r}_t}} B_{\mathbf{r}'_t \mathbf{r}_t}(\mathbf{y})\ \frac{\pi(t, \mathbf{r}'_t|\lambda, n)}{\pi(t, \mathbf{r}_t|\lambda, n)} = \binom{n}{n_1 \ldots n_t} - 1 > 0,\ [M_{\mathbf{r}_t}].$$

As $S_T(n)$ goes to infinity as n increases, the posterior probability of the true model goes to zero, and this proves the theorem. In fact, the posterior probability of any model in the configuration class $\mathfrak{R}_{t;n_1 \cdots n_t}$ goes to zero.

A.6 Proof of Theorem 6

First we prove that, when sampling from any model M_ν in $\mathfrak{R}_2$, the limit of the posterior probability of this set of configuration classes is 1, when n goes to infinity.

Note that the pseudodistance from any model in the configuration class $\mathfrak{R}_{p;n_1\cdots n_p}$ to any other model in the configuration class $\mathfrak{R}_{q;n_1'\cdots n_q'}$ does not depend on the particular models chosen within these classes, as the pseudodistance only depends on the corresponding design matrices, whose elements only consist of 0's and 1's, as there are no regressors. In our case, this fact implies that the Bayes factors only depend on the ratio n_1/n, and this greatly simplifies the computation of the posterior probabilities of the configuration classes.

From Lemma 1, we have that under the true model M_ν, the Bayes factor for comparing $M_{n1/n}$ and M_0, for large values of n, is approximated by

$$B_{\frac{n_1}{n}\,0} \approx \Big(\frac{3}{n}\Big)^{1/2}\Big(\frac{1+\delta_{\nu\,0}}{1+\delta_{\nu\,n_1/n}}\Big)^{n/2} \quad [M_\nu] \quad \text{for all} \quad n_1 = 1,\ldots,n-1.$$

Therefore, the posterior distribution of $\mathfrak{R}_1$ is

$$\Pr(\mathfrak{R}_1|\mathbf{y}_n) \approx \frac{1}{1+\sum_{n_1=1}^{\lfloor n/2\rfloor}\frac{\Pr(\mathfrak{R}_{2;n_1,n_2})}{\Pr(\mathfrak{R}_1)}\big(\frac{3}{n}\big)^{1/2}\Big(\frac{1+\delta_{\nu\,0}}{1+\delta_{\nu\,n_1/n}}\Big)^{n/2}}.$$

If we prove that the sum in the denominator

$$S(\tau,n) = \sum_{n_1=1}^{\lfloor n/2\rfloor}\frac{\Pr(\mathfrak{R}_{2;n_1,n_2})}{\Pr(\mathfrak{R}_{1;n})}\Big(\frac{3}{n}\Big)^{1/2}\Big(\frac{1+\delta_{\nu\,0}}{1+\delta_{\nu\,n_1/n}}\Big)^{n/2}$$

goes to infinity for the corresponding prior, then the posterior of $\mathfrak{R}_1$ goes to 0, and consequently the posterior of $\mathfrak{R}_2$ goes to 1, when n tends to infinity.

But as the quotient

$$\frac{1+\delta_{\nu\,0}}{1+\delta_{\nu\,n_1/n}}$$

is strictly greater than 1 for any ν and n_1/n in $(0,1/2]$, the last term of the sum $S(\tau,n)$ grows exponentialy with n. On the other hand, as the prior ratios

$$\frac{\Pr(\mathfrak{R}_{2;n_1,n_2})}{\Pr(\mathfrak{R}_{1;n})}$$

are of order $O(\log n)$, $O(n)$, $O(2^{n/2})$, and $O(2^{n/2})$ for the Ewens-Pitman, the hierarchical uniform, the Jensen-Liu and the uniform priors, respectively, the sum $S(\nu,n)$ goes to infinity at an exponential rate.

This result means that we can restrict the study of consistency to the configuration classses in $\mathfrak{R}_2$.

By Bayes theorem, the posterior probability of the configuration classes $\mathfrak{R}_{2;n_1,n_2}$, for any prior $\Pr(\mathfrak{R}_{2;n_1,n_2})$ and for large n, is approximated by

$$\Pr(\mathfrak{R}_{2;n_1,n_2}|\mathbf{y}) \propto \Pr(\mathfrak{R}_{2;n_1,n_2})(1+\delta_{\nu\, n_1/n})^{-n/2}.$$

We first prove the theorem for the hierarchical unform prior. In this case all prior probabilities are equal; thus the posterior is

$$\Pr^{HU}(\mathfrak{R}_{2;n_1,n_2}|\mathbf{y}) \propto (1+\delta_{\nu\, n_1/n})^{-n/2}.$$

Whithout loss of generality, assume that $n_1^* = \nu\, n$; then the "likelihood" $(1+\delta_{\nu\, n_1/n})^{-n/2}$ as a function of n_1, can be approximated, in a discrete neighborhood of n_1^* as follows:

$$\left(1+\delta_{\nu\, n_1/n}\right)^{-n/2} \approx \exp(-d|n_1^* - n_1|/2).$$

But this approximation is also valid for all values of $n_1 = 1, \ldots, \lfloor n/2 \rfloor$ as the terms outside the neighborhood decrease exponentially to 0.

Then, the posterior of the configuration classes can be approximated by

$$\Pr^{HU}(\mathfrak{R}_{2;n_1,n_2}|\mathbf{y}) \propto \frac{\exp(-d|n_1^* - n_1|/2))}{\sum_{n_1=1}^{\lfloor n/2 \rfloor} \exp(-d|n_1^* - n_1|/2)}.$$

Now, the sum of the denominator is finite as it can be decomposed into 1, when $n_1 = n_1^*$, and twice the sum of a geometric series whose ratio is $\exp[-d/2]$. Further, its limit, when n tends to infinity, is

$$1 + \frac{2\exp[-d/2]}{1-\exp[-d/2]}.$$

Thus, an upper bound $b(d)$ of the posterior probability of the configurations is given by the maximum of $\exp(-d|n_1^* - n_1|/2)$, which is attained at $n_1 = n_1^*$ and equals 1, divided by the preceding limit, that is

$$0 < b(d) = \frac{e^{d/2}-1}{e^{d/2}+1} < 1, \quad \text{for all} \quad d > 0.$$

The same proof serves for the Ewens-Pitman distribution by noting that the prior

$$\Pr^{EP}(\mathfrak{R}_{2;n_1,n_2}) \propto \frac{1}{n_1(n-n_1)}$$

does not influence the posterior for, in this case, the "likelihood" dominates the prior.

This last argument cannot be applied, however, either to the uniform or the Jensen-Liu priors, as these priors dominate the "likelihood". In fact, the second part of the theorem follows from the next one.

A.7 Proof of Theorem 7

For the EP and HU priors and for large values of n, the distribution of the discrete random variable Y_n,

$$\Pr^{HU}\Big(Y_n = \frac{n_1}{n}\Big) \propto (1 + \delta_{\nu\, n_1/n})^{-n/2}$$

and

$$\Pr^{EP}\Big(Y_n = \frac{n_1}{n}\Big) \propto \frac{1}{n_1(n-n_1)}(1 + \delta_{\nu\, n_1/n})^{-n/2}$$

can be approximated by continuous densities, making the change of variable $y = n_1/n$, of the form

$$f^{EP}(y|\nu, n) \propto y^{-1}(1-y)^{-1} \exp\Big(-\frac{1}{2} d\, n|y-\nu|\Big) \quad y, \nu \in (0, 1/2],$$

and

$$f^{HU}(y|\nu, n) \propto \exp\Big(-\frac{1}{2} d\, n|y-\nu|\Big) \quad y, \nu \in (0, 1/2],$$

respectively.

Both posterior densities are proper even though the asymptotic Ewens-Pitman prior is improper. Also note that for large n, due to the exponential nature of the common asymptotic Bayes factor, both densities have the same asymptotic behavior —the Bayes factor overwhelms the improper prior— and therefore, both densities can be approximated by a double exponential or Laplace distribution with location parameter ν and scale parameter $\frac{2}{dn}$, that is by a $La(y|\nu, \frac{2}{dn})$, truncated at the set $(0, 1/2]$. As the variance of the posterior distribution goes to 0, as n tends to infinity, the posterior distribution of Y_n converges in distribution to a degenerate distribution located at ν. This proves the first part of the therorem.

For the uniform prior, and for large values of n, using the asymptotic normal approximation of the prior in the proof of part c) of Theorem 2, when $p = 2$, the distribution of Y_n, can be approximated by the continuous density

$$f^{U}(y|\nu, n) \propto n(y|1/2, 1/4n) \exp\Big(-\frac{1}{2} d\, n|y-\nu|\Big) \quad y, \nu \in (0, 1/2],$$

where $n(y|\cdot, \cdot)$ denotes the density function of the corresponding normal distribution.

This posterior density turns out to be proper in the interval $(0, 1/2]$, but does not correspond to any known distribution. However, it can be shown that for some values of ν in a certain interval, which depends on d, the posterior can be approximated by a normal distribution centered at $\mu(\nu, d)$, a function which does not depend on n, and variance tending to 0 as n goes to infinity. This implies that the posterior of Y_n converges in probability to a degenerate distribution at $\mu(\nu, d)$.

The explicit expression of $\mu(\nu, d)$ is a complicated formula. In fact, it is the solution in y of the equation

$$\frac{d\,\nu^2}{y\,(d\,\nu(y-\nu) + x)} = 4 - 8y,$$

which is the real root of a polynomial equation of third degree.

If, for instance, we assume that $d = 1$, it can be shown that $\nu < \mu(\nu, d)$ for all $0 < \nu < 3/8$, and this implies that there cannot be consistency for this range of values of ν. Interestingly, if $\nu = 3/8$, the procedure is consistent as $\mu(\nu, 1) = \nu$.

The proof of weak inconsistency for the Jensen-Liu prior follows a similar pattern to that of the uniform prior. In fact, the asymptotic distribution of the Jensen-Liu prior, that follows from the proof of part d) in therorem 2 for $p = 2$, is the normal distribution

$$N\Big(y\Big|\frac{\lambda+2}{2\lambda+3}, \frac{(\lambda+1)(\lambda+2)}{n\,(2\lambda+3)^2}\Big);$$

thus, the continuous asymptotic approximation to the posterior of Y_n is

$$f^{JL}(y|\nu, n) \propto n\Big(y\Big|\frac{\lambda+2}{2\lambda+3}, \frac{(\lambda+1)(\lambda+2)}{n\,(2\lambda+3)^2}\Big) \exp\Big(-\frac{1}{2}d\,n|y-\nu|\Big) \quad y, \nu \in (0, 1/2].$$

This is a proper logconcave density in the interval $(0, 1/2]$ which can be approximated by a (truncated) normal distribution, the mean of which is $\rho(\nu, d, \lambda)$, a function that does not depend on n, and its variance goes to 0 as n tends to infinity. This means that the distribution of Y_n converges in probability to a degenerate distribution at $\rho(\nu, d, \lambda)$, which is the solution in y of the equation

$$\frac{d\,\nu^2}{y\,(d\,\nu(y-\nu)+x)} = \frac{(2\lambda+3)(2-3y+\lambda(2y-1)}{(\lambda+1)(\lambda+2)}.$$

As before, $\rho(\nu, d, \lambda)$ is the real solution of a polynomial cubic equation.

For most values of $\lambda \in (0, \infty)$, and of $d \in (0, \infty)$, the inequality $\nu < \rho(\nu, d, \lambda)$ holds, implying weak inconsistency. In particular, when $\lambda = 0$ and $d = 1$, then $\rho(\nu, d, \lambda) > 1/2$, so that we always have inconsistency for all values of ν. Also note that when $\lambda \to \infty$, $\lim_{\lambda\to\infty} \rho(\nu, d, \lambda) = \mu(\nu, d)$. This behavior is to be expected as the Jensen-Liu prior approaches the uniform prior when λ increases to infinity.

Finally, note from the preceding that the Bayes procedures for the uniform and the Jensen-Liu priors do place most probability (approaching 1 as n goes to infinity) in a neighborhood of $\mu(\nu, d)$ and $\rho(\nu, d, \lambda)$ respectively, and this implies that the probability in a neighborhood of the true model M_ν can be made as close to 0 as desired. This means that the upper bound for the posterior probability of the true model is 0, and this proves the last part of Theorem 7.

Acknowledgments

G. Casella was supported by National Science Foundation Grants DMS–0631632 and SES–0631588. E. Moreno and F.J. Girón are supported by Ministerio de Ciencia y Tecnología, Grant MTM2011-28945 and Junta de Andalucía Grant SEJ–02814.