SORT 33 (2) July-December 2009, 115-146

Testing for the existence of clusters

Claudio Fuentes and George Casella*

University of Florida

Abstract

Detecting and determining clusters present in a certain sample has been an important concern, among researchers from different fields, for a long time. In particular, assessing whether the clusters are statistically significant, is a question that has been asked by a number of experimenters. Recently, this question arose again in a study in maize genetics, where determining the significance of clusters is crucial as a primary step in the identification of a genome-wide collection of mutants that may affect the kernel composition.

Although several efforts have been made in this direction, not much has been done with the aim of developing an actual hypothesis test in order to assess the significance of clusters. In this paper, we propose a new methodology that allows the examination of the hypothesis test $H_0 : \kappa = 1$ *vs.* $H_1 : \kappa = k$, where κ denotes the number of clusters present in a certain population. Our procedure, based on Bayesian tools, permits us to obtain closed form expressions for the posterior probabilities corresponding to the null hypothesis. From here, we calibrate our results by estimating the frequentist null distribution of the posterior probabilities in order to obtain the *p*-values associated with the observed posterior probabilities. In most cases, actual evaluation of the posterior probabilities is computationally intensive and several algorithms have been discussed in the literature. Here, we propose a simple estimation procedure, based on MCMC techniques, that permits an efficient and easily implementable evaluation of the test. Finally, we present simulation studies that support our conclusions, and we apply our method to the analysis of NIR spectroscopy data coming from the genetic study that motivated this work.

MSC: 62F15, 62F03.

Keywords: Hierarchical models, Bayesian inference, frequentist calibration, Monte Carlo methods, p-values.

1 Introduction

In recent years, researchers have been working on the identification of genes that cause dosage-dependent changes in seed weight or composition of cereal grains. More

* *Address for correspondence:* Claudio Fuentes (cfuentes@stat.ufl.edu) and George Casella (casella@stat.ufl.edu), Department of Statistics, University of Florida, Gainesville, FL 32611.

Received: May 2009

precisely, on the identification of a genome-wide collection of mutants with quantitative effects on the seed. Since the major seeds constituents (protein, oil, starch, cellulose and water) have multiple near infrared absorption bands, the use of single-kernel *Near Infrared Reflectance* (NIR) spectroscopy has become a standard (non-destructive) technique to collect the data.

This technology provides an information-rich spectrum allowing multiple chemicals and structures to be detected and quantified, and therefore, detecting and determining well differentiated clusters from the NIR spectra should identify kernels with differing composition. In particular, when applied to a genetic screen, these clusters would correspond to mutants that separate into groups according to Mendelian frequencies.

The presence of a genetic factor that gives rise to distinct clusters can be verified through inheritance tests. But calibrations for all possible chemical changes within a kernel are costly and time consuming. In consequence, a statistic expressing true presence or absence of clusters would greatly facilitate the analysis of complex data sets and is needed as a primary step in the search and identification of composition mutants.

Hence, the problem we need to solve is to determine whether it is meaningful (in some sense) to partition a set of observations into different groups, and if so, how many of them. This problem is not new in statistics and several solutions have been proposed, going back to Hartigan's Rule (Hartigan, 1975), with more recent contributions from Tibshirani, Walther and Hastie (2001) and Sugar and James (2003). These methods tend to be distance based, and use measures (such as the gap statistic, or measures borrowed from information theory) to assess if clusters are far enough apart to be declared different.

Other methods focus on validity or repeatability of clusters, such as Auffermann, Ngan and Hu (2002), who use the bootstrap on Fisher's linear discriminant function in order to test for two clusters, but go no further. The bootstrap has also been used by Kerr and Churchill (2001) to assess stability of clusters, not directly testing significance but rather seeing if there are groups of genes that remain together. Other cluster detection methods are more *ad hoc*; for example Bolshakova, Azuaje and Cunningham (2005) look at a variety of deterministic clustering algorithms and validity measures in order to look for relevant clusters.

In a more Bayesian or hierarchical setting, McCullaugh and Yang (2006) specify priors on the parameters in the context of a Gaussian mixture model and make use of a Dirichlet process to assess the number of clusters. Fraley and Raftery (2002) also consider the use of mixture models to cluster the data but assess the significance of them using the BIC criterion. Other efforts consider the use of probability models for partitions of a set of n elements using a predictive approach and also make use of BIC to select the optimal partition (see Quintana, 2004). Pritchard, Stephens and Donnelly (2000) consider a Bayesian model and put a prior on the (unknown) number of clusters to compute posterior probabilities but do not go any further.

More recently, Booth, Casella and Hobert (2008) consider a different approach to cluster multivariate data, based on a multi-level linear mixed model. Their methodology

is fundamentally different from others in that they explicitly include the partition of the data (and not only the number of clusters) as a parameter. Then, making use of MCMC techniques they can obtain the posterior distribution of this parameter and use it to cluster the data. Nevertheless, none of these approaches attempt to develop a test to assess the significance of clusters.

The approach we propose here is slightly different, and exploits a Bayesian model selection methodology (making use of Bayes factors) to derive an explicit hypothesis test for the existence of clusters. In addition, our procedure is not distance based and hence avoids the use of a metric to determine the clusters. Also, our model parameterizes the partitions themselves and not only the number of clusters. This way, the evidence for clusters is not determined according to the "proximity" of the observations and the test takes full advantage of the probability structure considered to model the data and the space of partitions.

In Sections 2 and 3, we explain how to construct the hypothesis test in the Bayesian framework and implement our methodology to analyze the NIR spectroscopy data coming from the study mentioned above. Later in Section 4, we discuss a method to calibrate the procedure in order to simplify the interpretation of the results and facilitate the decision making. In Section 5 we present simulation studies that validate our conclusions and allow us to implement our calibration in data analysis, which we do in Section 6. Finally, in Section 7 we discuss the more relevant aspects of our method and possible extensions for future research.

2 Testing for clusters

Let us denote the data by the n-tuple $\mathbf{Y} = (Y_1, \ldots, Y_n)$, where each coordinate Y_i $(1 \leq i \leq n)$ is a p-vector of responses. Also, let $j = 1, \ldots, k$ be the number of clusters, such that the j-th cluster contains n_j elements of $\mathbf{Y}$. Since each Y_i $(1 \leq i \leq n)$ can be only in one cluster, we have $n_1 + \ldots + n_k = n$, $n_j > 0$ $(1 \leq j \leq k)$.

For instance, if $n = 6$ and $k = 3$ we might have the clusters

$$\begin{array}{ccc} \{Y_1, Y_3\} & \{Y_4\} & \{Y_2, Y_5, Y_6\} \\ n_1 = 2 & n_2 = 1 & n_3 = 3 \end{array}$$

and

$$\begin{array}{ccc} \{Y_1, Y_3\} & \{Y_2, Y_4\} & \{Y_5, Y_6\} \\ n_1 = 2 & n_2 = 2 & n_3 = 2 \end{array}$$

It is clear that several partitions are possible, even for fixed values of n and k. For this reason, we assume the existence an unknown parameter κ which determines the number

of clusters and a parameter ω (depending on κ) that determines the *partition* of $\mathbf{Y}$ into κ (non-empty) clusters. We immediately observe that, given $\kappa = k$, the number of all possible partitions of n objects into k clusters is given by $S(n,k)$, the *Stirling number of the second kind* (Gould, 1960). We will denote this set of partitions by $\mathscr{S}_{n,k}$.

Next, for any fixed partition $\omega \in \mathscr{S}_{n,k}$, we will denote by $Y_1^{(j)}, Y_2^{(j)}, \ldots, Y_{n_j}^{(j)}$ the n_j vectors of responses that are allocated in cluster j, where (to simplify the notation), we consider the responses to be ordered within a cluster. For instance, for the third cluster $\{Y_2, Y_5, Y_6\}$ in the first example on the previous page, we have $Y_1^{(3)} = Y_2$, $Y_2^{(3)} = Y_5$ and $Y_3^{(3)} = Y_6$. Notice that this notation implicitly determines a certain order for the observations within a cluster. This will not be problematic for our purposes, since later (in Section 2.2) we will assume that the observations are *iid* within a cluster.

Finally, to describe the elements of the vector $Y_\ell^{(j)}$ (the ℓ-th vector of responses in cluster j) we will write

$$Y_\ell^{(j)} = (y_{\ell 1}^{(j)}, \ldots, y_{\ell p}^{(j)})^\mathsf{T}$$

where $\ell = 1, \ldots, n_j$ and $j = 1, \ldots, k$.

2.1 Bayesian hypothesis testing

Given a set of observations, $Y_1, \ldots, Y_n$, our aim is to construct a framework to test the hypothesis

$$H_0 : \kappa = 1 \text{ vs. } H_1 : \kappa > 1.$$

Of course, the alternative hypothesis above implies that $\kappa = k$ (for some integer k), and we will concentrate on the simpler problem of testing

$$H_0 : \kappa = 1 \text{ vs. } H_1 : \kappa = k, \tag{1}$$

for some given k. This way, we have a *simple null vs. simple alternative* test and we can look at it as a model selection problem where we try to identify the model with the highest probability.

At this point, we take a Bayesian approach, and compute the Bayes factor associated with the hypothesis in (1), that is

$$BF_{10} = \frac{m(\mathbf{Y}|\, \kappa = k)}{m(\mathbf{Y}|\, \kappa = 1)}, \tag{2}$$

where $m(\mathbf{Y}|\, \kappa = k)$ denotes the distribution of the data $\mathbf{Y}$, given that we have exactly k clusters.

Observe that conditioning on $\kappa = k$ in (2) involves considering all the possible partitions $\omega \in \mathscr{S}_{n,k}$ that generate k clusters. We can rewrite the Bayes factor in terms of the partitions ω as

$$BF_{10} = \sum_{\omega \in \mathscr{S}_{n,k}} \frac{m(\mathbf{Y}|\ \omega)}{m(\mathbf{Y}|\ \omega_1)} \frac{\pi(\omega)}{\pi(\omega_1)}, \tag{3}$$

where ω_1 denotes the only existing cluster when $\kappa = 1$ and $\pi(\omega)$, $\pi(\omega_1)$ denote prior probabilities for the partitions ω and ω_1 respectively.

It follows by considering the extra assumption that $P(\kappa = k) = P(\kappa = 1) = 1/2$ (that is, assuming the hypotheses being tested are equally likely), that we can determine the posterior probability of H_0 as

$$P(H_0|\mathbf{Y}) = \frac{1}{1 + BF_{10}}. \tag{4}$$

This quantity, which is typically used as a model comparison criteria, will provide evidence against H_0 whenever $P(H_0|\mathbf{Y})$ is small.

2.2 Model and distribution assumptions

For any given partition $\omega \in \mathscr{S}_{n,k}$, we assume that all the observations in cluster j follow a $N(\boldsymbol{\mu}_j, \Sigma_j)$ distribution, that is,

$$Y_\ell^{(j)} \sim N(\boldsymbol{\mu}_j, \Sigma_j),$$

for $\ell = 1, \ldots, n_j$ and $j = 1, \ldots, k$. Then, the likelihood function of the sample is

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \omega | Y_1, \ldots, Y_n) = \prod_{j=1}^{k} \prod_{\ell=1}^{n_j} N(Y_\ell^{(j)} \mid \boldsymbol{\mu}_j, \Sigma_j),$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)$ and $\boldsymbol{\Sigma} = (\Sigma_1, \ldots, \Sigma_k)$.

In order to complete the specification of the model, we will assume that $\Sigma_j = \text{diag}\,(\sigma_{1j}^2, \ldots, \sigma_{pj}^2)$. Then, for $r = 1, \ldots, p$ and $j = 1, \ldots, k$, we consider the prior distributions

$$\begin{aligned} \boldsymbol{\mu}_j &\sim N(\boldsymbol{\mu}_0^{(j)}, \tau^2 \Sigma_j), \\ \sigma_{rj}^2 &\sim IG(a, b) \end{aligned} \tag{5}$$

where $\boldsymbol{\mu}_0^{(j)} = (\mu_{01}^{(j)}, \ldots, \mu_{0p}^{(j)})'$ and $IG(a,b)$ denotes an inverted gamma distribution with parameters a and b. In this framework, we can compute the marginal distribution of the data $\mathbf{Y}$, given the partition ω. We obtain (see Appendix C for details)

$$
\begin{aligned}
m(\mathbf{Y}|\,\omega) &= \int L(\boldsymbol{\mu},\Sigma,\omega|Y_1,\ldots,Y_n)N(\boldsymbol{\mu}|\boldsymbol{\mu}_0^{(j)},\tau^2\Sigma_j)\pi(\Sigma_j)d\boldsymbol{\mu}_j d\Sigma_j \\
&= \left(\frac{2}{b}\right)^{pka} \frac{1}{\pi^{np/2}\Gamma(a)^{pk}} \qquad (6) \\
&\times \left[\prod_{j=1}^{k} \frac{\Gamma\left(\frac{n_j}{2}+a\right)^p}{(n_j\tau^2+1)^{p/2}}\right] \left[\prod_{j=1}^{k}\prod_{r=1}^{p} \frac{1}{\left(n_j s_{rj}^2 + n_j \frac{(\bar{y}_r^{(j)}-\mu_{0r}^{(j)})^2}{n_j\tau^2+1} + \frac{2}{b}\right)^{n_j/2+a}}\right].
\end{aligned}
$$

where $s_{rj}^2 = \sum_{i=1}^{n_j}(y_{ir}^{(j)} - \bar{y}_r^{(j)})^2/n_j$ and $\bar{y}_r^{(j)} = \sum_{i=1}^{n_j} y_{ir}^{(j)}/n_j$.

Of course, the expression in (6) depends on the values of the hyperparameters $\boldsymbol{\mu}_0^{(j)}$, a and b. We will address the setting of a and b in Section 3. The values of $\boldsymbol{\mu}_0^{(j)}$ can either reflect true prior information, or specify a submodel. In the absence of this kind of information, a default empirical choice is to set $\boldsymbol{\mu}_0^{(j)}$ to be equal to the sample means $\bar{y}^{(j)}$. As we will discuss later, this constraint will have no further impact other than to simplify our calculations, and will not affect the generality of the results we will show in the coming sections.

2.3 Prior on the partitions

When $\kappa = 1$ there is only one cluster (of size n), and thus we take $\pi(\omega_1) = 1$. For $\pi(\omega)$ we have many choices, but here we will mention only three. First, if we spread the prior mass uniformly in the set of all partitions into k clusters, then the number of such partitions is $S(n,k)$, and hence we take $\pi_U(\omega) = 1/S(n,k)$. An alternative prior is the marginal distribution of the number of clusters in a Dirichlet process (Pitman, 1996)

$$\pi_D(\omega) = \frac{\Gamma(m)m^k}{\Gamma(n+m)}\prod_{j=1}^{k}\Gamma(n_j),$$

where m is a parameter to be specified. This is a prior on all of partition space, and since we are restricting our calculations to a fixed k, this prior is essentially proportional to $\prod_{j=1}^{k}\Gamma(n_j)$. In contrast to the uniform prior, this prior will have the effect of favouring partitions with more *balanced* clusters, in the sense that partitions that allocate fewer observations in some clusters and concentrate the rest in another cluster will have lower probabilities.

We observe that none of the priors discussed above present a simple alternative if we are interested in sampling partitions from $\mathscr{S}_{n,k}$. Consequently, we will end this section discussing a strategy to generate random partitions according to a certain distribution g, suggested by Jim Pitman (personal communication).

In order to obtain a random partition of n objects into k clusters we use the following strategy: We take a vector of length n with $n-k$ 0's and k 1's, putting a 1 in the first position. Then we randomly generate a permutation of the remaining $n-1$ elements to distribute the $k-1$ 1's in the last $n-1$ places. If each 1 indicates the start of a cluster, we have generated a string to represent the clusters. For example, if $n=5$ and $k=3$, the string 11001 corresponds to the partition of five objects into clusters of size 1, 3 and 1. Finally, we randomly permute the Y vector, and place the Y_i's in the generated string. Although not immediately obvious (see Appendix A.2 for details), the probability of the generated partition ω is given by

$$g(\omega) = \frac{k!}{\binom{n-1}{k-1}\binom{n}{n_1\, n_2\, \ldots\, n_k}}. \tag{7}$$

In addition, we can easily modify the strategy to generate partitions with a minimum cluster size.

2.4 Estimation of the Bayes factor

So far, we have developed all the theoretical framework we require in order to compute the Bayes factors and the corresponding posterior probabilities $P(H_0|\mathbf{Y})$. However, we notice that the sum in (3) is indexed over the set of *all possible partitions*, which introduces two practical difficulties: first, the number of summands involved in the calculation is typically large, even if the number of observations and clusters is relatively small. For instance, if $n=48$ and $k=2$, we have $S(48,2) = 140,737,488,355,327$. Second, to compute the sum we need to list the partitions, which is not a trivial task.

This difficulty can be overcome by MCMC estimation of the Bayes factor. Several algorithms have been proposed and discussed in the literature. See, for example, Steele, Raftery and Emond (2006) and Ventura (2002). But here we will take a simpler approach which has (empirically) proven to work well within the extent of the application problem we are intending to solve.

Let π and g be distributions on the partition space $\mathscr{S}_{n,k}$. Suppose that π is the prior of interest and we can sample $\omega^{(1)},\ldots,\omega^{(M)}$ from g. If M is large enough, we can estimate the value of the Bayes factor through the importance sampling sum

$$\begin{aligned} BF_{10} &= \sum_{\omega\in\mathscr{S}_{n,k}} \left[\frac{m(\mathbf{Y}|\,\omega)}{m(\mathbf{Y}|\,\omega_1)}\right]\pi(\omega) = \sum_{\omega\in\mathscr{S}_{n,k}} \left[\frac{m(\mathbf{Y}|\,\omega)}{m(\mathbf{Y}|\,\omega_1)}\right]\frac{\pi(\omega)}{g(\omega)}g(\omega) \\ &\approx \frac{1}{M}\sum_{i=1}^{M}\left[\frac{m(\mathbf{Y}|\,\omega^{(i)})}{m(\mathbf{Y}|\,\omega_1)}\frac{\pi(\omega^{(i)})}{g(\omega^{(i)})}\right] \approx \frac{\sum_{i=1}^{M}\left[\frac{m(\mathbf{Y}|\,\omega^{(i)})}{m(\mathbf{Y}|\,\omega_1)}\frac{\pi(\omega^{(i)})}{g(\omega^{(i)})}\right]}{\sum_{i=1}^{M}\frac{\pi(\omega^{(i)})}{g(\omega^{(i)})}}, \end{aligned} \tag{8}$$

where the last expression in (8), while possibly biased, is proven to reduce the mean squared error (see Casella and Robert, 1998, and Van Dijk and Kloeck, 1984). Notice that if we consider g as the prior of interest in the first place, then importance sampling is not needed, and we just compute the Monte Carlo sum.

A first approach to calculate the Monte Carlo sum in (8) would be to sample from g. Although this is reasonable, the convergence is slow as the space of partitions is large, and the algorithm spends much time in areas of low probability. A better strategy is to direct the sampling to areas of high probability, where most of the contribution to the sum will lie. This can be accomplished with a Metropolis-Hastings modification which we now describe.

HYBRID RANDOM WALK ALGORITHM It is possible to incorporate a random walk component when generating the partitions, so that the search algorithm remains in areas of high probability. This way, the algorithm will allow a more accurate calculation of the Monte Carlo sum, and will maintain the correct stationary distribution.

To this end, we generate M partitions $(\omega^{(1)},\ldots,\omega^{(M)})$ according to a Metropolis-Hastings algorithm, which is a mixture of the following two steps:

- *Independent draw:* Draw candidate ω' from g.
- *Random walk:* At iteration t, obtain candidate ω' by choosing one observation at random from $\omega^{(t)}$, and moving it to one of the other $k-1$ clusters with equal probability.

The final Metropolis-Hastings algorithm is:

1. Draw candidate ω' from g.
2. At iteration t
   (a) With probability a, draw candidate ω' from the random walk starting from $\omega^{(t)}$, and with probability $1-a$ draw candidate ω' independently from g.
   (b) Compute the Metropolis-Hastings ratio

$$MH = \frac{g(\omega')}{\frac{a}{n(k-1)} + (1-a)g(\omega')} \times \frac{\frac{a}{n(k-1)} + (1-a)g(\omega^{(t)})}{g(\omega^{(t)})}$$

   (c) With probability $\min(1,MH)$ set $\omega^{(t+1)} = \omega'$, otherwise set $\omega^{(t+1)} = \omega^{(t)}$

Notice that the described algorithm has stationary distribution g.

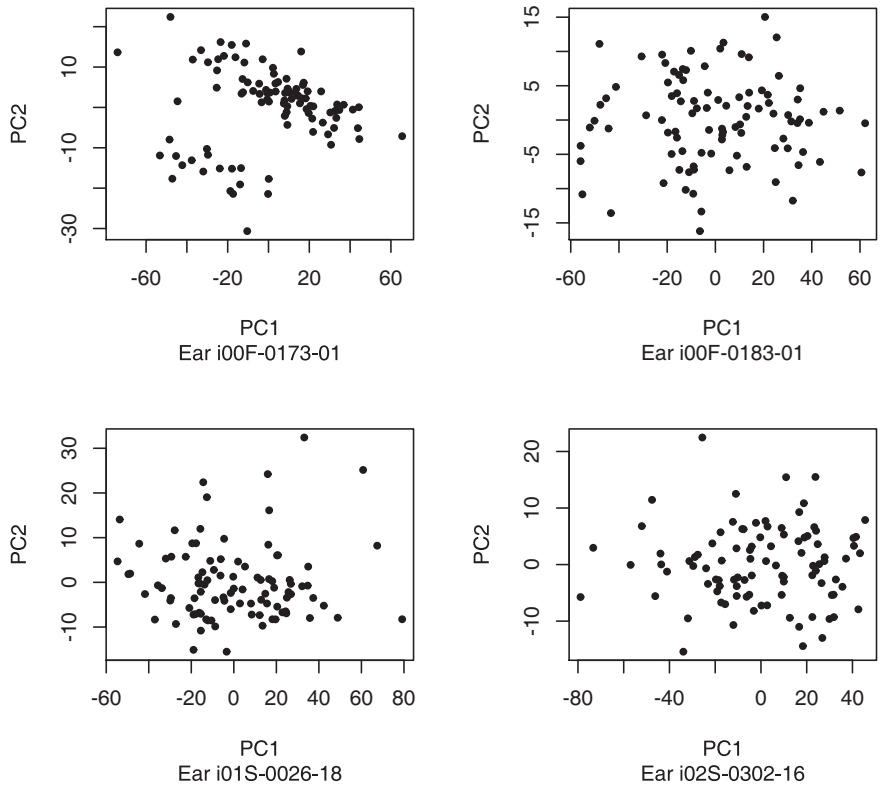

Figure 1: *Scatter-plots for the first two principal components of the NIR spectra for data sets with labels i00F-0173-01, i00F-0183-01, i01S-0026-18 and i02S-0302-16.*

3 Application to NIR spectroscopy data

In order to illustrate our procedure, we consider four data sets coming from the study described in Section 1. All of them consist of 96 vectors of dimension two, where the coordinates of the vectors corresponds to the first 2 principal components obtained from the NIR spectra collected from 96 kernels coming from a single ear of maize. The data sets under consideration are labeled i00F-0173-01, i00F-0183-01, i01S-0026-18 and i02S-0302-16 and researchers are interested in finding evidence for the existence of 2, 3 or 4 clusters in each one of them. Scatter-plots for the corresponding data sets appear in Figure 1.

Before we compute the posterior probabilities, let us recall that the expression in (6), and consequently, the Bayes factor, depends on the values of the hyperparameters a and b of the inverted gamma distribution. Since we are not interested in making any prior assumption about the tightness of the clusters, we prefer to consider a prior with high variability. We take $a = 2.01$ and $b = (a-1)^{-1} \approx 0.990099$.

Table 1: *Posterior probabilities for the hypothesis tests H_0: no clusters vs. H_1: κ =2, 3 and 4 clusters, with minimum cluster size 15% of the total number of observations.*

Label	n	$P(H_0\|k=2)$	$P(H_0\|k=3)$	$P(H_0\|k=4)$
i00F-0173-01	96	0.0219493	0.2790514	0.9900416
i00F-0183-01	96	0.2245521	0.9679710	0.9992699
i01S-0026-18	96	0.0393423	0.3610524	0.8909811
i02S-0302-16	96	0.6429479	0.9031773	0.9960509

Table 1 shows the posterior probabilities computed when testing for $k = 2$, 3 and 4 clusters. The values were obtained after 1000000 iterations, which seems to be an adequate number to ensure convergence of the Monte Carlo sum based on simulations. In addition, a constraint setting the minimum cluster size equal to 15% of the total number of observations was considered. This restriction is imposed because clusters of smaller sizes are not meaningful to the researchers in the context of the experiment.

We observe, for labels i00F-0173-01 and i01S-0026-18, low posterior probabilities for H_0 when testing for $\kappa = 2$ and 3, and high posterior probabilities when testing for $\kappa = 4$, indicating evidence for the existence of 2 and 3 clusters, but not for 4. Notice, however, that in both cases the evidence for clusters seems to be "strong" only when testing for $\kappa = 2$. The conclusion is fairly well supported by the respective scatter-plots.

For ear i00F-0183-01, we obtain a fairly low posterior probability for H_0 when testing for $\kappa = 2$ and very high posterior probabilities of the null for testing $\kappa = 3$ and 4. Thus, we obtain evidence for the existence of clusters when testing for 2 clusters, but the test seems to be conclusive (accept H_0) for the other two cases. Similarly, for ear i02S-0302-16 we obtain high posterior probabilities for H_0 in all the tests, but the results seem to be conclusive (H_0 is true) only when testing for 3 and 4 clusters. In addition, we observe that none of the scatter-plots for the last two data sets are very helpful to support the results of the test.

It follows that one of the practical difficulties in order to make a decision is that we do not have an error calibration for our procedure. Therefore we cannot properly measure the *strength* of the evidence against the null hypothesis and we cannot easily decide when the data provide enough evidence to make a conclusion, especially when we do not observe extreme values (close to 0 or 1) for our posterior probabilities. Hence, we need to develop a calibration procedure that facilitates the decision making of the researcher for our hypothesis test.

4 Frequentist calibration

In the previous sections, we have discussed how to produce posterior probabilities in order to measure evidence for the existence of clusters. Nevertheless, it is well know that these results need to be calibrated in order to establish the statistical significance of our findings. Specifically, we know that observing posterior probabilities below 0.5

suggests the presence of clusters in a certain data set, and the lower the better. But how low should the posterior probability be in order for the experimenter to make a good decision is unknown.

This problem is not new in Bayesian analysis and some solutions can be found in the literature. For instance, Jeffreys (1961) developed a scale to judge the evidence in favour of or against H_0 brought by the data, Bayarri and Berger (1998) developed an analog of the frequentist p-value in the Bayesian paradigm, and Girón, Martínez, Moreno and Torres (2006) calibrated intrinsic posterior probabilities to p-values. References and details of a number of these methods can be found in Robert (2001) and Ghosh, Delampady and Samanta (2006).

Here, we will solve the problem by determining the frequentist null distribution of $P(H_0|\mathbf{Y})$; that is, the distribution of $P(H_0|\mathbf{Y})$ as a function of the data $\mathbf{Y}$, when the null hypothesis is true. To this end, observe that we can rewrite the Bayes factor (3) in terms of the data $\mathbf{Y}$ as follows

$$BF_{10}(\mathbf{Y}) = \sum_{\omega \in \mathscr{S}_{n,k}} \lambda(\omega) T(\mathbf{Y}|\omega), \tag{9}$$

where for every $\omega \in \mathscr{S}_{n,k}$,

$$\lambda(\omega) = \left[\left(\frac{2}{b}\right)^{pa(k-1)} \frac{(n\tau^2+1)^{p/2}}{\Gamma(a)^{p(k-1)}\Gamma\left(\frac{n}{2}+a\right)^p} \prod_{j=1}^{k} \frac{\Gamma\left(\frac{n_j}{2}+a\right)^p}{(n_j\tau^2+1)^{p/2}} \right] \frac{\pi(\omega)}{\pi(\omega_1)}, \tag{10}$$

and

$$T(\mathbf{Y}|\omega) = \prod_{r=1}^{p} \left[\frac{\left(ns_r^2 + \frac{2}{b}\right)^{n/2+a}}{\prod_{j=1}^{k} \left(n_j s_{rj}^2 + \frac{2}{b}\right)^{n_j/2+a}} \right]. \tag{11}$$

Hence, the $\lambda(\omega)$'s capture the non-random terms of the Bayes factor and the $T(\mathbf{Y}|\omega)$'s absorb the data dependent portion.

4.1 Bayes factor under the null distribution

Let us consider first the one dimensional case ($p = 1$). Suppose we have $y_1, \ldots, y_n$ independent observations. Then, for a given partition ω, we have

$$\begin{array}{rcl} y_1^{(1)}, \ldots, y_{n_1}^{(1)} & \sim & iid\ N(\mu_1, \sigma_1^2) \\ y_1^{(2)}, \ldots, y_{n_2}^{(2)} & \sim & iid\ N(\mu_2, \sigma_2^2) \\ \vdots & \vdots & \vdots \\ y_1^{(k)}, \ldots, y_{n_k}^{(k)} & \sim & iid\ N(\mu_k, \sigma_k^2) \end{array} \qquad \begin{array}{l} \text{where } y_1^{(1)} + \ldots + y_{n_k}^{(k)} = y_1 + \ldots + y_n \\ \text{and } n = \sum_{j=1}^{k} n_j. \end{array}$$

When the null hypothesis is true, that is, there are no clusters in the data, we have $\mu_1 = \ldots = \mu_k = \mu$ and $\sigma_1^2 = \ldots = \sigma_k^2 = \sigma^2$, and we can prove (see Appendix B.1) the following result, which follows from Cochran's theorem (Kendall, Stuart, Ord and Arnold, 1998).

Proposition 1 *For $y_i^{(j)}$ as above, define*

$$u_j = \frac{n_j s_j^2}{\sigma^2} \ (j = 1, \ldots, k) \quad \textit{and} \quad u_{k+1} = \frac{\sum_{j=1}^k n_j (\bar{y}^{(j)} - \bar{y})^2}{\sigma^2}.$$

Then, under the null hypothesis, $u_1, \ldots, u_{k+1}$ are independent and

$$u_j \sim \chi^2_{n_j - 1} \quad (j = 1, \ldots, k), \quad u_{k+1} \sim \chi^2_{k-1}.$$

Observe that, for any given partition $\omega \in \mathscr{S}_{n,k}$, we can write

$$T(\mathbf{Y}|\omega) = \frac{1}{(\sigma^2)^{(k-1)a}} \frac{\left(\sum_{j=1}^{k+1} u_j + 2/b\sigma^2\right)^{\frac{n}{2}+a}}{\prod_{j=1}^k \left(u_j + 2/b\sigma^2\right)^{\frac{n_j}{2}+a}}, \tag{12}$$

where $(u_1, \ldots, u_{k+1})$ are defined as in Proposition 1. Hence, if the null hypothesis holds and if $V = (v_1, \ldots, v_{n-1})$ is a vector of independent and identically distributed χ_1^2 random variables, we have

$$u_j \stackrel{\mathscr{D}}{=} \sum_{i=n_{j-1}}^{n_j - 1} v_i \ (j = 1, \ldots, k) \quad \text{and} \quad \sum_{j=1}^{k+1} u_j \stackrel{\mathscr{D}}{=} \sum_{i=1}^{n-1} v_i,$$

where we take $n_0 = 1$. This result leads to the following proposition, whose proof is straightforward.

Proposition 2 *Let $v_1 \ldots, v_{n-1}$ be iid χ_1^2 random variables. Then, if the null hypothesis holds,*

$$T(\mathbf{Y}|\omega) \stackrel{\mathscr{D}}{=} \frac{1}{(\sigma^2)^{(k-1)a}} \frac{\left(\sum_{i=1}^{n-1} v_i + 2/b\sigma^2\right)^{\frac{n}{2}+a}}{\prod_{j=1}^k \left(\sum_{i=n_{j-1}}^{n_j - 1} v_i + 2/b\sigma^2\right)^{\frac{n_j}{2}+a}},$$

for every $\omega \in \mathscr{S}_{n,k}$.

An immediate consequence of the previous proposition is that the null distribution of $T(\mathbf{Y}|\omega)$ depends on the partition ω only through the cluster sizes $n_1, \ldots, n_k$. In other words, if ω_1 and ω_2 are two different partitions in $\mathscr{S}_{n,k}$, then the distributions of $T(\mathbf{Y}|\omega_1)$ and $T(\mathbf{Y}|\omega_2)$ will differ only if the corresponding cluster sizes differ for at

least one n_i. On the other hand, since all the priors discussed in Section 2.3 depend on the partitions ω only through the respective cluster sizes, we obtain that the same holds for the constants $\lambda(\omega)$ in (10).

It follows that, under the null hypothesis, we can group all the terms corresponding to partitions with the same cluster sizes in (9) into a single term whose multiplying constant, say ξ, will be the sum of the respective λ's. By doing this, it turns out that the number of different elements in the sum is determined by the number of ways that we can partition an integer n into k integers $n_1,\ldots,n_k$ such that $n = n_1 + \ldots + n_k$. Then, combining the previous propositions, we obtain the following lemma (see Appendix B.2 for a proof).

Lemma 1 *Let $\mathscr{P}_{n,k}$ be the set of all partitions of the integer n into exactly k terms and denote by ξ any of its elements. Then, under the conditions of Proposition 2*

$$BF_{10}(\mathbf{Y}|H_0) \stackrel{\mathscr{D}}{=} \sum_{\xi\in\mathscr{P}_{n,k}} \phi(\xi)T(V|\xi),$$

where

$$T(V|\xi) = \frac{1}{(\sigma^2)^{(k-1)a}} \frac{\left(\sum_{i=1}^{n-1} v_i + 2/b\sigma^2\right)^{\frac{n}{2}+a}}{\prod_{j=1}^{k}\left(\sum_{i=n_{j-1}}^{n_j-1} v_i + 2/b\sigma^2\right)^{\frac{n_j}{2}+a}}$$

and $\phi(\xi)$ is an appropriate normalizing constant for every $\xi \in \mathscr{P}_{n,k}$.

The previous results are important for two reasons: first, they provide us with a known probabilistic structure for each one of the components present in the Bayes factor, and second, they reduce the complexity of the problem allowing us to obtain the null distribution of the Bayes factor; we need to compute a sum with many fewer terms than what we have for the general case.

For example, if we consider $n = 70$ observations and $k = 4$ clusters, the total number of partitions of 70 elements into 4 clusters is greater than 5×10^{40}, whereas the number of ways of writing 70 as the sum of exactly 4 integers is given by $p(70,4) = 2484$. This remarkable difference is produced because the null hypothesis induces an *equivalence relation* in the space $\mathscr{S}_{n,k}$ of all partitions, where the *classes* are the partitions with the same number of elements.

In order to extend these results to the multidimensional case $(p > 1)$, we observe (from the assumptions of normality and independence in the model) that Propositions 1 and 2 remain valid componentwise. On the other hand, the diagonal structure of the variance-covariance matrices Σ_i under consideration induces independence between the coordinates of Y_i $(i = 1,\ldots,n)$ and consequently between the factors of the product in (11). Hence, no correlation is induced by our calculations and the generalization to higher dimensions proceeds in the obvious manner.

4.2 Estimation of the null distribution

In the light of the results obtained in the previous section, the derivation of the null distribution of the Bayes factor in closed form seems feasible. However some other difficulties add to the problem making it complicated (see Fuentes, 2008). But, at the same time, the very same results allow us to simulate observations from the null distribution of the posterior probabilities, which can be used to construct histograms or density estimates, depending on the interest.

If the null hypothesis is true we only care about the cluster sizes. Then, we can follow the same strategy pointed out in Section 2.4 to generate the partitions according to g, but without taking into account the permutations of the elements in the given partition. In other words, we need to correct the probabilities given by g, so that they do not take into account the number of redundant partitions that lead to the same cluster sizes.

It follows that the probabilities for the partitions ξ's are given by

$$g_0(\xi) = \frac{k!}{\mathscr{R}(n_1,\ldots,n_k)} \frac{1}{\binom{n-1}{k-1}}$$

where $\mathscr{R}$ is counting the number of redundant strings corresponding to partitions that give the same cluster sizes (see Appendix A.2).

It is easy to check that the null distribution of the posterior probabilities has a positive and continuous density on $(0,1)$. Then, based on the strong consistency of the empirical percentiles (see Sen and Singer, 1993), we can estimate the cutoff points for any given α-*level* by the corresponding α-th empirical percentile from our generated sample, or simply obtain an estimate of the p-value corresponding to our test statistics, depending on the interest.

Finally, we observe that the Bayes factor in (9) depends on the (unknown) value of σ^2, the variance of the observations under the null. Thus, in order to generate a sample from the null distribution, we can estimate σ^2 by the sample variance, or simply take it to be one if the original data set is centred and scaled.

5 Simulation studies

The computation of the posterior probabilities and their null distribution are both based on MCMC techniques. Therefore, before we use our procedure for the analysis of real data we need to determine the quality of our estimates. To this end, we considered several simulations to study the convergence of the Monte Carlo sums and the error rates, among others. In this section we present the results of some simulations to illustrate the main features of our method.

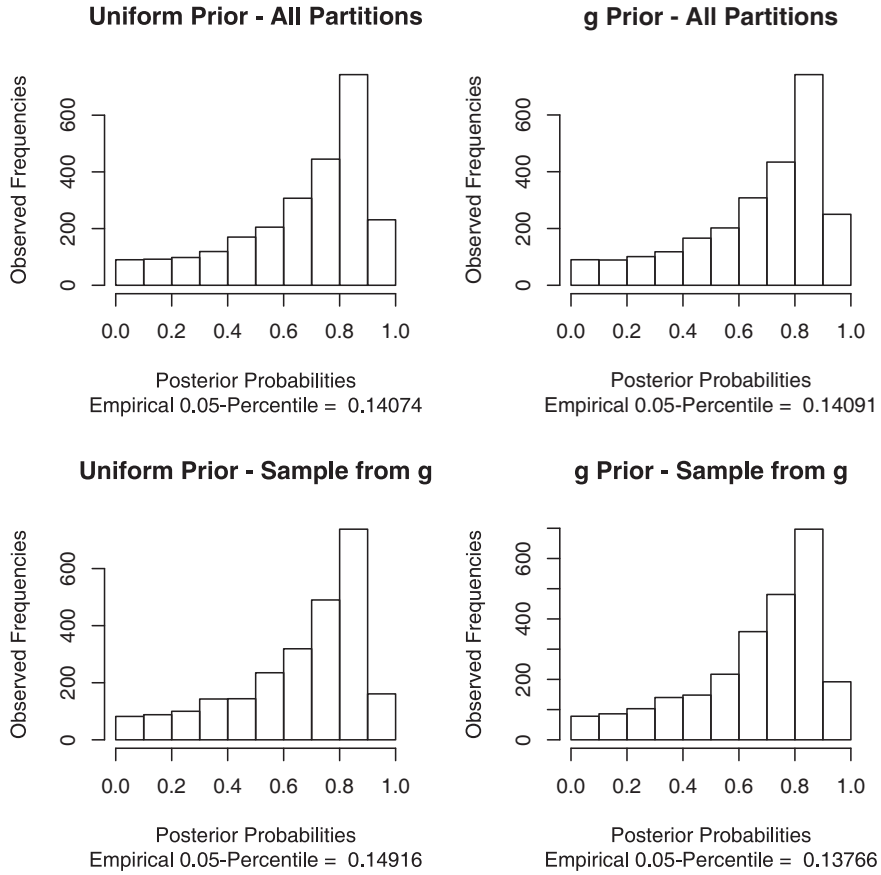

Figure 2: *Histograms of the null posterior probabilities for $n = 50$ and $k = 3$. The top row is the exact calculation, based on enumerating all partitions. The corresponding histograms in the bottom row are based on 2500 simulations of samples of size 52, representing 25% of all partitions.*

5.1 Goodness of the approximation

The null distribution of the posterior probabilities is unknown and therefore, determining how good is our approximation is not trivial. To address this problem, we apply our procedure to $n = 10$, 25 and 50, and $k = 2$, 3, and 4. These quantities, although arbitrary, allow us to list all the partitions. Then we compute the exact Bayes factor and our estimate of the Bayes factor for the same generated data. Proceeding this way, we obtain simulations of the posterior probabilities and we compare the histograms and the 0.05 percentile.

In Figure 2 we see the results for $n = 50$ and $k = 3$. For this case, the number of elements in the partitions space is $p(50,3) = 208$. In the top row all partitions were used, allowing us to obtain the exact Bayes factor. For the corresponding histograms in the bottom row, 2500 samples of size 52 (25% of the total number of elements in the

partition space) were drawn from g_0 to compute the approximations. We considered the uniform prior (first column) and the g_0 prior (second column) in our calculations.

We observe that the histograms are virtually identical and that the differences between the empirical 0.05-percentiles is less than 0.012 in all cases. The results are similar in all the cases we studied, indicating that our method is fairly accurate in approximating the null distribution of the posterior probabilities and suggesting that the selection of the prior for the partition space has little effect in the calculations.

5.2 Error of approximation

The cutoff points for the α-level tests will be ultimately determined by the corresponding empirical α-percentiles. Hence, we estimate the variability of the procedure by computing the standard error associated with replications of the experiment.

Table 2 presents the results corresponding to 6 simulation studies for the case $n = 50$, $k = 3$. The empirical 0.05 percentiles are obtained based on 2500 simulations, sampling 52 out of 208 partitions per iteration. The obtained standard error is less than 0.003 which is fairly small considering the number of simulations per repetition. Similar results are obtained when changing the values of n and k, indicating that convergence of the empirical α-percentile is reached moderately fast.

Table 2: *For $\alpha = .05$, six replications of the posterior probabilities percentile (2500 simulations) for 50 observations and 3 clusters. The number of considered partitions per iteration is 52.*

α-level	α-percentile
0.05	0.15416 0.16448 0.17061 0.17005 0.17298 0.16455
Mean SE	0.16614 0.00277

5.3 Minimum cluster size

When the null hypothesis is true, there are no clusters. For this reason, one might expect the histograms of the posterior probabilities to be very skewed to the left with most of the observations falling in the vicinity of 1. However, looking at our simulations, we observe that a considerable number of observations fall below 0.5.

Table 3: *For α-level 0.05 and minimum cluster size 1, cutoff points based on 5000 simulations. The number in parenthesis is to the standard error based on 6 repetitions.*

Clusters	Observations		
	50	60	70
2	0.15261 (0.00198)	0.18647 (0.00161)	0.20709 (0.00230)
3	0.09782 (0.00153)	0.13556 (0.00311)	0.16973 (0.00141)
4	0.05454 (0.00034)	0.09268 (0.00095)	0.13836 (0.00118)

In Table 3 we show the results corresponding to the empirical 0.05 percentile for n =50, 60, 70 and k =2, 3, 4. The values are obtained as the average of 6 repetitions of 5000 simulations each. In parenthesis we report the respective standard errors. We observe not only that the cutoff points for the 0.05-level test are fairly small, but also the following pattern. For every n, the value of the cutoff points decreases as the number of clusters increases, that is, about 5% of the generated posterior probabilities are located closer to zero as the k increases.

The general behaviour is that for fixed n, as k increases the histograms, while still skewed to the left, tend to spread more mass to smaller values resulting in fatter tails instead of the expected thin tails.

The most likely explanation for this phenomena is that the number of elements that constitutes a cluster is not defined. Therefore, our procedure tends to consider as their own clusters observations that deviate from the *overall behaviour*. Since these deviations fall randomly in different directions, it is generally difficult to cluster all of them in one group and allocate the rest in another for the case $k = 2$, but this problem simplifies as we consider more clusters to separate the observations.

This conjecture is not proven in this work, but is supported by our simulations. If we predefined the minimum number of observations that determine a cluster, then we observe the previous behaviour changes the direction, that is, for every n, the value of the cutoff points increases as the number of clusters increases. We believe the reason for this change in the behaviour is that once the minimum cluster size is determined, we cannot consider as a cluster a few observations that deviate from the general pattern, unless they match with the minimum cluster size (MCS) required. In other words, by introducing this new parameter in the model, we are reducing our possibilities of finding clusters by chance.

Table 4: *For α-level 0.05 and minimum cluster size equal to 15% of the observations, cutoff points based on 5000 simulations. The number in parenthesis is to the standard error based on 6 repetitions.*

Clusters	Observations		
	50	60	70
2	0.25523 (0.00381)	0.31751 (0.00322)	0.36356 (0.00313)
3	0.29041 (0.00208)	0.39503 (0.00294)	0.51345 (0.00511)
4	0.29198 (0.00226)	0.51517 (0.00210)	0.68352 (0.00243)

Table 4 shows the results of simulations obtained under the same conditions we described above, but setting the minimum cluster size equal to the 15% of the observations. We can see how the introduction of the minimum cluster size as a new parameter, reverses the pattern observed in Table 3 and also increases the value of the empirical 0.05-percentiles.

In general, as the minimum cluster size increases (and therefore the probability of finding clusters by chance decreases) the histograms become more skewed to the left and tend to concentrate more mass near one, as we can observe in Figure 3. In other words, the extra restriction provides more intuitive results for the simulated posterior probabilities.

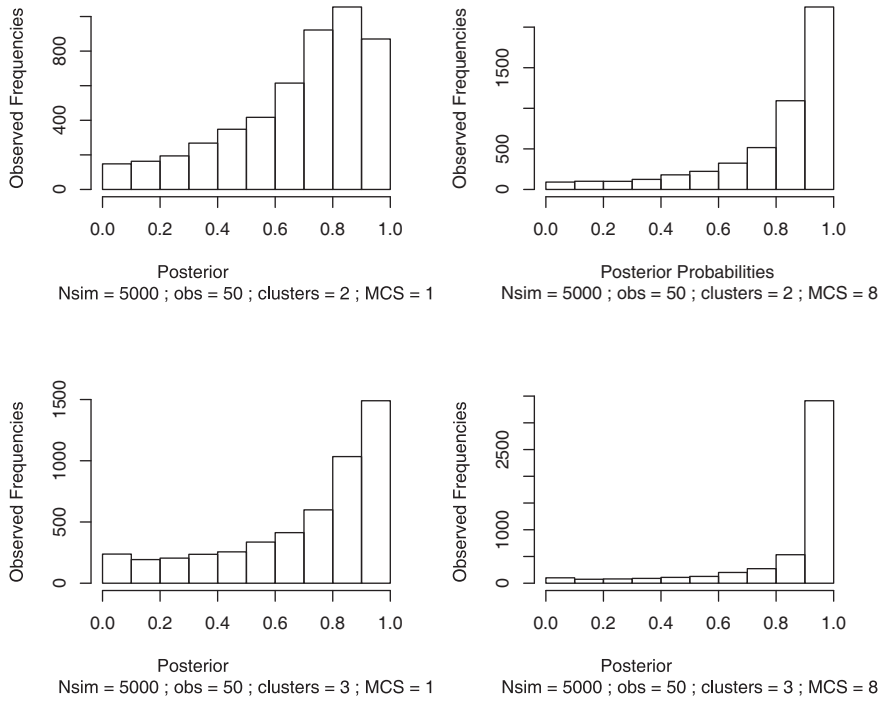

Figure 3: *Histograms of the null posterior probabilities for $n = 50$ and $k = 2$ (top row) or $k = 3$ (bottom row) clusters based on 5000 simulations. The minimum cluster size is set equal to 1 observation (left column) and 15% of the observations (right column).*

5.4 Power of the procedure

Finally, we need to assess the reliability of the posterior probabilities in detecting clusters. In particular, we need to check the behaviour of the posterior probabilities in the most extreme cases, that is, when there are no clusters (that is, the null hypothesis is true) and when there are at least two clusters in the data.

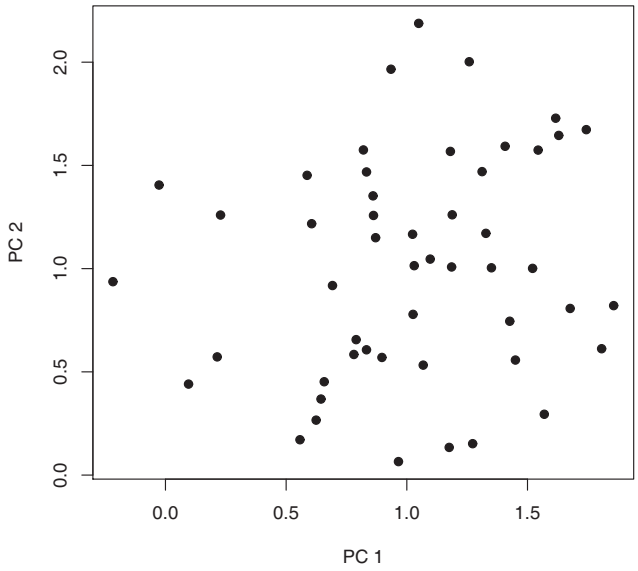

Figure 4: *Scatter-plot of 50 observations generated from a bivariate normal distribution with mean* $\boldsymbol{\mu} = (1,-1)^{\mathsf{T}}$ *and variance-covariance matrix* $\Sigma = diag(1/4, 1/4)$.

The simulations of the null distribution indicate that even when the null hypothesis holds, there is still a fair chance of detecting clusters. This probability decreases when we incorporate the minimum cluster size as a parameter. Hence, we need to check the performance of our test statistic when analyzing data sets with no clusters. To this end, we generated several data sets, each one from a single multivariate normal distribution, so all the observations within a data set have the same mean and variance-covariance matrix, and consequently, they form a unique cluster.

The posterior probabilities were obtained after 1000000 iterations, considering a minimum cluster size equal to 15% of the total number of observations. The results for $k = 2,3$ and 4 clusters are listed in Table 5. In Figure 4 we show the scatter-plot corresponding to 50 observations from a bivariate normal distribution with mean $\boldsymbol{\mu} = (1,-1)^{\mathsf{T}}$ and variance-covariance matrix $\Sigma = \text{diag}(1/4, 1/4)$. We observe that the posterior probabilities are fairly high when testing for 3 and 4 clusters, showing very weak evidence for the presence of clusters. The smaller value is obtained for testing two clusters, but still the corresponding posterior probability is too high to be declared significant according to our calibrations (see Table 4). Other simulations agree with these results.

Table 5: *Posterior probabilities after 1000000 iterations for the observations in Figure 4. The minimum cluster size is equal to 15% of the observations.*

k	2	3	4
$P(H_0)$	0.44249	0.68144	0.93203

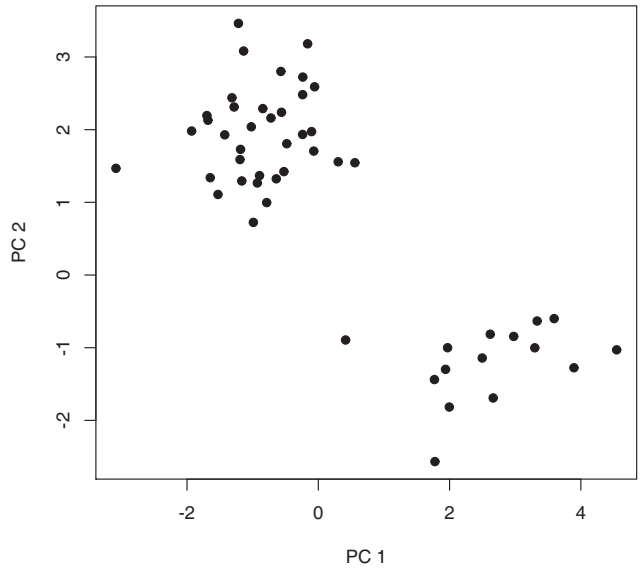

Figure 5: *Scatter-plot of 50 observations generated from two bivariate normal distributions with different means.*

The other case we need to consider is when we have at least two clusters. We generated data sets composed from observations coming from multivariate normal distributions with different means, depending on the number of clusters we wanted to test. This way, we purposely created clusters in the data sets to be tested. In Figure 5 we show the scatter-plot corresponding to 50 observations. Of them, 35 were generated from a bivariate normal distribution with mean $\boldsymbol{\mu} = (-1,2)^{\mathsf{T}}$ and variance-covariance matrix $\Sigma = \text{diag}(1/2,1/2)$, and the remaining 15 were generated from a bivariate normal distribution with mean $\boldsymbol{\mu} = (3,-1)^{\mathsf{T}}$ and same variance-covariance matrix Σ. By construction, we have two clusters in the data, which can be easily noticed. The posterior probabilities obtained after 1000000 iterations are listed in Table 6. The minimum cluster size is 15% of the total number of observations.

The posterior probabilities obtained for the data are very small, indicating strong evidence against the null hypothesis in all of the tests. Observe that although the data set was constructed with two clusters, we still have strong evidence for the existence of three or four clusters. This happens because we have two very distinguishable groups, and each group is fairly easy to separate in other two groups. Similar results are observed in other simulations indicating that our procedure for detecting clusters is fairly accurate.

Table 6: *Posterior probabilities after 1000000 iterations for the observations in Figure 5. The minimum cluster size is equal to 15% of the observations.*

k	2	3	4
$P(H_0)$	1.09×10^{-4}	1.50×10^{-5}	3.50×10^{-5}

Table 7: *Posterior probabilities and 0.05-percentiles corresponding to the hypotheses tests H_0: no clusters vs. H_1: κ =2, 3 and 4 clusters, with minimum cluster size equal to 15% of the total number of observations.*

Label	n	$P(H_0\|k=2)$	$P(H_0\|k=3)$	$P(H_0\|k=4)$
i00F-0173-01	96	0.0219493	0.2790514	0.9900416
i00F-0183-01	96	0.2245521	0.9679710	0.9992699
i01S-0026-18	96	0.0393423	0.3610524	0.8909811
i02S-0302-16	96	0.6429479	0.9031773	0.9960509
0.05-ptle.		0.5110465	0.8153481	0.9313412

6 Application of the calibration procedure

We now illustrate how the calibration procedure discussed in Section 4 can be used in the analysis of the NIR spectroscopy data.

In Table 7 we find the posterior probabilities for the data considered in Section 3 plus the 0.05-percentile of the null distribution for the respective tests. When comparing the results with the respective 0.05-percentiles we obtain, for labels i00F-0173-01 and i01S-0026-18, strong evidence for the existence of 2 and 3 clusters in the sense that we reject the null hypothesis at level 0.05. For ear i00F-0183-01 we find significant evidence for the existence of clusters only when testing for $\kappa = 2$. Finally, for ear i00F-0183-01 we do not find significant evidence for the existence of clusters in any of the tests. In this case, the smallest posterior probability for H_0 is 0.643 (obtained when $k = 2$), which is above the corresponding cutoff point at level 0.05. Then, we can conclude that the null hypothesis is true in all the cases.

Notice that while we can reject the null hypothesis for more than one test, the obtained posterior probabilities look quite different. For instance, in label i01S-0026-18 the posterior probability corresponding to $k = 2$ is much smaller than the posterior probabilities for $k = 3$, one might think that data is providing more evidence in favour of 2 clusters than 3 clusters. However, the posterior probabilities should not be compared for different values of k, because their respective null distributions correspond to different probability spaces and may differ (see Figure 3). In particular, the α-percentiles will differ as we can see in Tables 3 and 4.

While in this case it is difficult to decide about the number of clusters, the data set clearly indicates strong evidence for the existence of clusters and demands special attention from the researcher.

7 Discussion

We have proposed a method for testing for clusters based on Bayesian model selection. Our method does not test directly the more general hypothesis H_0 : *No clusters* vs. H_1 : *At*

least two clusters, but it does provide an accurate notion of the cluster structure present in the data, by considering the simpler hypotheses $H_0 : \kappa = 1$ *vs.* $H_1 : \kappa = k$, where κ denotes the number of clusters. In practical applications, the researchers are often interested in testing for a specific number of clusters and, in that sense, our procedure provides a desirable answer.

In addition, the frequentist calibration discussed in Section 4 greatly facilitates decision making, providing interpretable results when assessing the significance of clusters is required. Furthermore, our calibration procedure brings up an interesting feature (due to the skewness of the frequentist null distribution), namely, that posterior probabilities for the null hypothesis may provide strong evidence against H_0 (there is no clusters), even if their values are apparently large. This suggests that special attention should be put on decision making and in the calibration mechanism when comparing any two models using Bayes factors.

Simulation studies validate the performance of the test, showing that the posterior probabilities give small values when there are true clusters in the data, and large values (relative to the calibrated scale) when there are no clusters in the data. Extreme observations and outliers may affect the values of the posterior probabilities and consequently the conclusions of the test. However, the introduction of a minimum cluster size (MCS) as a parameter in the model corrects this problem and produces more meaningful results. This parameter, rather than being exogenous to the model, is naturally incorporated in the procedure, for experimenters typically are not interested in clusters defined by very few observations.

Regarding the convergence of the estimators, simulations also show that when computing the posterior probabilities (the test statistic) about 1000000 iterations are needed to reach convergence of the Monte Carlo sum. This is a small number of partitions considering the size of the partition space. On the other hand, when estimating the null distribution of the posterior probabilities, about 5000 iterations are necessary to reach convergence of the α-percentiles, where each one of them should be calculated using a number of partitions equal to approximately 25% of the size of the number of partitions of the integer n.

Before we conclude this paper, we would like to make a few comments and remarks on some aspects for future research.

In our formulation, we considered a specific set of priors: a *normal* prior for the cluster means and an *inverted gamma* prior for the cluster variances. Although the selection of these types of priors is justified, a natural question is how robust is our test to the selection of the prior. We have only begun to study this matter, looking at the effect of the inverted gamma hyperparameters in the model. We have observed that different choices of the parameters a and b for the inverted gamma prior produce different results for the posterior probabilities, but the corresponding null distributions change accordingly and therefore the significance of the tests remain the same. In other words, the conclusions of our procedure do not change substantially if a different set of parameters is considered for the inverted gamma prior.

As we pointed out in the introduction, we have presented a testing procedure and a calibration method to determine the strength of the evidence when detecting clusters, but we have not identified the clusters. Some modifications to the algorithms allow us (for a fixed k) to conduct a stochastic search in the space of partitions $\mathscr{S}_{n,k}$ to find the optimal partition that determines the clusters. Such modifications suggest that, under our procedure, two observations will not be allocated in different clusters just because they are far apart. Hence, our procedure is not dependent on any distance (in the metric sense) in looking for evidence for clusters, but only uses the probabilistic model defined for the observations and the partition space. This feature is particularly interesting, because some points that fall far away from the mean of their respective "true" cluster only by chance will be declared to be in the wrong clusters under some distance based methods. In fact, it is a property of our algorithm that the "optimal" clusters need not be convex.

Finally, we have implemented the cluster test in the R package `bayesclust`, which can be downloaded free from http://cran.r-project.org/.

Acknowledgments

The authors would like to acknowledge Dr. Mark Settles for introducing the problem that motivated this paper in a very practical situation and for all his interesting questions throughout the research. Also we would like to thank Vikneswaran Gopal for his invaluable help in the improvement and implementation of the coding required in this work, as well as for his dedicated efforts in the construction of an R package to facilitate the interface with the users. This research was partially supported by NSF-DBI grant 0606607.

References

Andrews, G. (1976). *The Theory of Partitions*. Addison-Wesley, Reading MA.

Auffermann, W. F., Ngan, S. C. and Hu, X. (2002). Cluster significance testing using the bootstrap. *NeuroImage*, 17, 583-591.

Bayarri, M. J. and Berger, J. (1998). Quantifying surprise in the data and model verification (with discussion). *Bayesian Statistics 6*, J. M. Bernardo, *et al.*, eds., 53-82, Oxford University Press, Oxford.

Bolshakova, N., Azuaje, F. and Cunningham, P. (2005). An integrated tool for microarray data clustering and cluster validity assessment. *Bioinformatics*, 21, 451-455.

Bona, M. (2004). *Combinatorics of Permutations*. Chapman & Hall/CRC, London.

Booth, J. G., Casella, G. and Hobert, J. P. (2008). Clustering using objective functions and stochastic search. *Journal of Royal Statistical Society, Series B*, 70, 119-140.

Casella, G. and Robert, C. (1998). Post-processing accept-reject samples: recycling and rescaling. *Journal of the Computational and Graphical Statistics*, 7, 139-157.

Easton, G. S. and Rochetti, R. (1986). General saddlepoint approximations with applications to L statistics. *Journal of the American Statistical Association*, 81, 420-423.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97, 611-631.

Fuentes, C. (2008). *Testing for the Existence of Clusters with Applications to NIR Spectroscopy Data*. Master Thesis, University of Florida, Florida.

Ghosh J. K., Delampady, M. and Samanta, T. (2006). *An Introduction to Bayesian Analysis: Theory and Methods*. Springer, New York.

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, 765-784.

Gould, H. W. (1960). Stirling number representation problems. *Proceedings of the American Mathematical Society*, 11, 447-451.

Glaser, R. E. (1980). A characterization of Bartlett's statistic involving incomplete beta functions. *Biometrika*, 67, 53-58.

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

Jeffreys, H. (1961). *Theory of Probability*. Third Edition. Oxford University Press, Oxford.

Kendall, M., Stuart, A., Ord, J. K. and Arnold, S. (1999). *Kendall's Advanced Theory of Statistics, Volume 2A: Classical Inference and the Linear Model*. Hodder Arnold, 6th Edition, London.

Kerr, M. K. and Churchill, G. A. (2001). Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 8961-8965.

Lavine, M. and Shervish, M. (1999). Bayes factors: what they are and what they are not. *American Statistician*, 53, 119-122.

McCullaugh, P. and Yang, J. (2006). How many clusters?. *Technical Report, Department of Statistics*. University of Chicago, Chicago.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. *Statistics, Probability and Game Theory*. IMS Lecture Notes Monograph Series, 30, 245-267, Institute of Mathematical Statistics, Hayward, CA.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155, 945-959.

Quintana, F. A. (2004). A predictive view of bayesian clustering. *Journal of Statistical Planning and Inference*, 136, 2407-2429.

Robert, C. P. (2001). *The Bayesian Choice*. Second Edition. Springer-Verlag, New York.

Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman & Hall, New York-London.

Steele, R., Raftery, A. E. and Emond, M. J. (2003). Computing normalizing constants for finite mixture models via incremental mixture importance sampling (IMIS). *Journal of Computational and Graphical Statistics*, 15, 712-734.

Sugar, C. and James, G. (2003). Finding the number of clusters in a data set: an information theoretic approach. *Journal of the American Statistical Association*, 98, 750-763.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, Series B*, 63, 411-423.

Van Dijk, H. and Kloeck, T. (1984). Experiments with some alternatives for simple importance sampling in Monte Carlo integration. *Bayesian Statistics 4*, J. Bernardo, M. DeGroot, D. Lindley and A. Smith Eds. North-Holland, Amsterdam.

Ventura, V. (2002). Non-parametric bootstrap recycling. *Statistics and Computing*, 12, 261-273.

A Generating a random partition

A.1 An example

Establishing (7) is not difficult, but some care must be taken in counting partitions, especially with respect to ordered versus unordered partitions. To be very clear, we start with an example. Suppose that $n = 8$ and $k = 4$, which is a small set of partitions, but big enough to be interesting. We know that the number of partitions of 8 objects into k cells, with no empty cell, is the Stirling number of the second kind, $\mathcal{S}_{8,4} = 1701$.

The strategy outlined in Section 2.3 will, for this case, generate $\binom{7}{3} = 35$ partitions. The only possible cluster sizes for $n = 8$ and $k = 4$ are

Partition	Number of $0-1$ Strings
$\{(1),(1),(1),(5)\}$	$4 = \binom{4}{1\,3}$
$\{(1),(1),(2),(4)\}$	$12 = \binom{4}{1\,1\,2}$
$\{(1),(1),(3),(3)\}$	$6 = \binom{4}{2\,2}$
$\{(1),(2),(2),(3)\}$	$12 = \binom{4}{1\,1\,2}$
$\{(2),(2),(2),(2)\}$	$1 = \binom{4}{4}$
Total	35

To actually count the number of 0-1 strings that correspond to a partition, we must account for redundancies. For example, the partition $\{(1),(1),(1),(5)\}$ arises from the four strings 11110000, 11100001, 11000011, and 10000111. This can be calculated by noting that there are 3 redundant clusters (each with one object), which tells us that the number of 0-1 strings corresponding to $\{(1),(1),(1),(5)\}$ is the multinomial coefficient $\binom{4}{1\,3}$.

Now that we can generate and count the 0-1 strings, we next need to make the correspondence with the $\mathcal{S}_{8,4} = 1701$ partitions in the population. To do this, note, for example, that corresponding to the partition $\{(1),(1),(1),(5)\}$ are $\binom{8}{1\,1\,1\,5}$ ordered arrangements in the population, and $\binom{8}{1\,1\,1\,5}/(1!\,3!)$ unordered arrangements. Thus, the probability of any partition of Y into the clusters $\{(1),(1),(1),(5)\}$ is given by

$$P(\{(1),(1),(1),(5)\}) = \frac{\binom{4}{1\,3}}{\binom{7}{3}} \times \frac{1!\,3!}{\binom{8}{1\,1\,1\,5}} = \frac{4!}{\binom{7}{3}\binom{8}{1\,1\,1\,5}}.$$

Lastly, notice that when we count the unordered arrangements, we obtain

$$\frac{\binom{8}{1\,1\,1\,5}}{1!\,3!} + \frac{\binom{8}{1\,1\,2\,4}}{1!\,1!\,2!} + \frac{\binom{8}{1\,1\,3\,3}}{2!\,2!} + \frac{\binom{8}{1\,2\,2\,3}}{1!\,1!\,2!} + \frac{\binom{8}{2\,2\,2\,2}}{4!} = 1701, \tag{13}$$

which is $\mathscr{S}_{8,4}$, the Stirling number of the second kind (and giving us an alternative representation of this number).

A.2 Derivation in the general case

It should now be clear how to derive the probability of the generation scheme in the general case. To ease notation we define the following function $\mathscr{R}$, which counts redundancies. For a partition $n_1, n_2, \ldots, n_k$, with $\sum_j n_j = n$, define

$$\mathscr{R}(n_1, n_2, \ldots, n_k) = \prod_{i=1}^{n} \left[\sum_{j=1}^{k} I(n_j = i) \right] !, \tag{14}$$

where $I(\cdot)$ is the indicator function. The function $\mathscr{R}$ counts the redundant strings, and allows us to efficiently calculate g, for example,

$$\mathscr{R}(1,1,1,5) = 1!3!.$$

With this notation, we see that the 0-1 generation scheme gives us a partition with probability

$$\frac{k!}{\mathscr{R}(n_1, n_2, \ldots, n_k)} \times \frac{1}{\binom{n-1}{k-1}}. \tag{15}$$

We note in passing that since this is a probability distribution on the ordered partitions, we have the identity

$$\sum_{\substack{n_1 + \ldots + n_k = n \\ n_1 \leq n_2 \leq \ldots \leq n_k}} \frac{1}{\mathscr{R}(n_1, n_2, \ldots, n_k)} = \frac{1}{k!} \binom{n-1}{k-1}. \tag{16}$$

Now, for each $n_1, n_2, \ldots, n_k$ the number of ways of partitioning n objects is

$$\frac{\binom{n}{n_1\ n_2\ \ldots\ n_k}}{\mathscr{R}(n_1, n_2, \ldots, n_k)}. \tag{17}$$

Multiplying (15) and (17) results in the probability of a partition ω being given by,

$$g(\omega) = \frac{k!}{\binom{n-1}{k-1} \binom{n}{n_1\ n_2\ \ldots\ n_k}}. \tag{18}$$

Note that this is a fully normalized probability distribution on the set of all partitions of n objects into k nonempty clusters, as

$$\sum_{\omega\in\mathscr{P}_k} g(\omega) = \sum_{n_1+\ldots+n_k=n} \frac{k!}{\binom{n-1}{k-1}\binom{n}{n_1\, n_2\, \ldots\, n_k}}$$
$$= \sum_{\substack{n_1+\ldots+n_k=n \\ n_1\le n_2\le\ldots\le n_k}} \sum_{\omega\in\mathscr{P}_{n_1,n_2,\ldots,n_k}} \frac{k!}{\binom{n-1}{k-1}\binom{n}{n_1\, n_2\, \ldots\, n_k}},$$

where $\mathscr{P}_{n_1,n_2,\ldots,n_k}$ is the subset of $\mathscr{P}_k$ with cluster sizes $(n_1,n_2,\ldots,n_k)$. As the summand is invariant to the inner sum, we can write

$$\sum_{\omega\in\mathscr{P}_k} g(\omega) = \sum_{\substack{n_1+\ldots+n_k=n \\ n_1\le n_2\le\ldots\le n_k}} \frac{k!}{\binom{n-1}{k-1}\binom{n}{n_1\, n_2\, \ldots\, n_k}} \sum_{\omega\in\mathscr{P}_{n_1,n_2,\ldots,n_k}} 1$$
$$= \sum_{\substack{n_1+\ldots+n_k=n \\ n_1\le n_2\le\ldots\le n_k}} \frac{k!}{\binom{n-1}{k-1}\binom{n}{n_1\, n_2\, \ldots\, n_k}} \frac{\binom{n}{n_1\, n_2\, \ldots\, n_k}}{\mathscr{R}(n_1,n_2,\ldots,n_k)},$$

which follows from (17). Canceling terms and applying (16) shows that $\sum_{\omega\in\mathscr{P}_k} g(\omega) = 1$.

Observe that the expression in (18) does not depend on the function $\mathscr{R}$ defined in (14) and therefore, it may seem that the introduction of such function was completely unnecessary. However, notice that the introduction of the function $\mathscr{R}$ serves our purposes in two respects: first, it keeps the derivation of the probability mass function g in a natural framework. Second, it permits us to obtain a simple expression for the distribution g_0 over the space $\mathscr{P}_{n,k}$ used in Section 4.2.

Finally, from (13) and (17) we obtain that an alternative representation of the Stirling number of the second kind is

$$\sum_{\substack{n_1+\ldots+n_k=n \\ n_1\le n_2\le\ldots\le n_k}} \frac{\binom{n}{n_1\, n_2\, \ldots\, n_k}}{\mathscr{R}(n_1,n_2,\ldots,n_k)} = \mathscr{S}_{n,k}.$$

In practical applications, experimenters are less interested in partitions with small cluster sizes, and a useful variation of this generation scheme incorporates that restriction. If m is the minimum number of objects in a cluster, we can generate partitions corresponding to this minimum specification with the following variation of the algorithm of Section 2.3.

For minimum cluster size m, start with k blocks of the form $[10\ldots0]$, which consist of one 1 and $m-1$ zeros. Place one block at the beginning of the string, then randomly allocate the remaining $k-1$ blocks and $n-mk$ zeros. As before, each 1 signifies the beginning of a cluster, but now each cluster will have at least m objects. An argument similar to that leading to (18) will show that under the present generation scheme, the probability of a partition with at least m objects in each cluster is

$$g_m(\omega) = \frac{k!}{\binom{n-mk+k-1}{k-1}\binom{n}{n_1\, n_2\, \ldots\, n_k}}, \tag{19}$$

which is a normalized probability distribution on the set of all partitions with minimum cluster size m.

A.3 Partitions of an integer

Let us consider first the following general definition (Bona, 2004):

Definition 1 *Let $a_1 \geq a_2 \geq \ldots \geq a_m \geq 1$ be integers so that $a_1 + a_2 + \ldots + a_m = n$. Then the array $a = (a_1, a_2, \ldots, a_m)$ is called a partition of the integer n, and the numbers a_i $(i = 1, \ldots, m)$ are called the parts of the partition a. The number of all partition of n is denoted by $p(n)$.*

For example, the integer 5 has seven partitions, namely (5), (4,1), (3,2), (3,1,1), (2,2,1), (2,1,1,1) and (1,1,1,1,1). Therefore, $p(5) = 7$.

Here we are interested in the particular case of partitions of an integer n into exactly k parts, that is, the arrays of exactly k (positive) integers such that their sum is equal to n. In our example, for $n = 5$ and $k = 3$ we have the partitions (2,2,1) and (3,1,1). In addition, if we denote by $p(n,k)$ the number of partitions of n into exactly k terms, we obtain that $p(5,3) = 2$.

Our problem is to determine $p(n,k)$ for any values n and k. Although we cannot obtain an explicit formula to compute $p(n,k)$, we can obtain a recursive relation by noticing:

- If one of the terms in a partition is 1, then the rest corresponds to a partition of $n-1$ into $k-1$ terms.
- If none of the terms in the partition is 1, then we can subtract 1 from each term and obtain a partition of $n-k$ into k parts.

Thus, the recursive relation is given by

$$p(n,k) = p(n-1,k-1) + p(n-k,k). \tag{20}$$

To complete the specification we define

$$\begin{aligned} p(n,k) &= 0 \quad \text{, for } n < k \\ p(n,n) &= 1 \quad \text{, for } n \geq 0 \\ p(n,0) &= 0 \quad \text{, for } n \geq 1. \end{aligned}$$

Hence, we can compute the recursive relation (20) for any (n,k). For further references and results on partitions of integers see Andrews (1976).

B Proofs of results in Section 4.1

B.1 Proof of Proposition 1

Proof. Let $y_1,\ldots,y_n \sim iid\ N(\mu,\sigma^2)$. For a given partition of the data into k clusters, the following decomposition holds

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{j=1}^{k} n_j(\bar{y}^{(j)}-\bar{y})^2 + \sum_{j=1}^{k} n_j s_j^2$$

where

$$s_j^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(y_i^{(j)}-\bar{y}^{(j)})^2\,, \quad j=1,\ldots,k.$$

Standard calculations show that $\bar{y}^{(j)}$ and s_j^2 are independent for all $j=1,\ldots,k$. On the other hand, s_j^2 is independent of $\bar{y}^{(i)}$ (for $i\neq j$), because none of the observations in s_j^2 are used to compute $\bar{y}^{(i)}$. Hence, for any $j=1,\ldots,k$, s_j^2 is independent of $\{\bar{y}^{(i)}\}_{i=1}^k$. Finally, noticing that

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{k} n_i\bar{y}^{(i)}$$

we obtain that s_j^2 and $\sum_{j=1}^k n_j(\bar{y}^{(j)}-\bar{y})^2$ are independent for $j=1,\ldots,k$. Since s_i^2 and s_j^2 are clearly independent for $i\neq j$, the result follows. ■

B.2 Proof of Lemma 1

Proof. We have

$$\begin{aligned} BF_{10}(\mathbf{Y}) &= \sum_{\omega\in\mathscr{S}_{n,k}} \lambda(\omega)T(\mathbf{Y}|\omega) \\ &\overset{\mathscr{D}}{=} \sum_{\omega\in\mathscr{S}_{n,k}} \lambda(\omega)T(V|\xi) \qquad \text{, by Proposition 2} \\ &= \sum_{\xi\in\mathscr{P}_{n,k}}\sum_{\Lambda(\xi)} \lambda(\omega)T(V|\xi), \end{aligned}$$

where $\Lambda(\xi)=\{\omega : \omega \text{ has clusters of size determined by } \xi\}$.

Since $T(V|\xi)$ depends on the partitions ω only through the clusters size, we obtain

$$BF_{10}(\mathbf{Y}) \stackrel{\mathscr{D}}{=} \sum_{\xi \in \mathscr{P}_{n,k}} T(V|\xi) \sum_{\Lambda(\xi)} \lambda(\omega)$$
$$= \sum_{\xi \in \mathscr{P}_{n,k}} \phi(\xi) T(V|\xi),$$

where $\phi(\xi) = \sum_{\Lambda(\xi)} \lambda(\omega)$. ∎

C Derivation of the marginal distribution

Under the model formulation of Section 2.2, the marginal distribution of the data $\mathbf{Y}$ given a partition ω is

$$m(\mathbf{Y}|\,\omega) = \int\int \prod_{j=1}^{k} \prod_{\ell=1}^{n_j} N(Y_\ell^{(j)}|\boldsymbol{\mu}_j, \Sigma_j) N(\boldsymbol{\mu}|\boldsymbol{\mu}_0^{(j)}, \tau^2\Sigma_j) \pi(\Sigma_j) d\boldsymbol{\mu}_j d\Sigma_j.$$

First, observe that

$$\prod_{j=1}^{k} \prod_{\ell=1}^{n_j} N(Y_\ell^{(j)}|\boldsymbol{\mu}_j, \Sigma_j) N(\boldsymbol{\mu}|\boldsymbol{\mu}_0^{(j)}, \tau^2\Sigma_j) \tag{21}$$
$$\propto \exp\left\{ -\frac{1}{2} \left[\sum_{\ell=1}^{n_j} (Y_\ell^{(j)} - \boldsymbol{\mu}_j)^\mathsf{T} \Sigma_j^{-1} (Y_\ell^{(j)} - \boldsymbol{\mu}_j) + (\boldsymbol{\mu}_j - \boldsymbol{\mu}_0^{(j)})^\mathsf{T} [\tau^2\Sigma_j]^{-1} (\boldsymbol{\mu}_j - \boldsymbol{\mu}_0^{(j)}) \right] \right\}.$$

Completing the square in the exponent shows that (21) is proportional to

$$\exp\left\{ -\frac{1}{2} \left[\sum_{\ell=1}^{n_j} (Y_\ell^{(j)} - \bar{Y}^{(j)})^\mathsf{T} \Sigma_j^{-1} (Y_\ell^{(j)} - \bar{Y}^{(j)}) \right. \right. + \frac{n_j\tau^2 + 1}{\tau^2} \left(\boldsymbol{\mu}_j - \delta(\bar{Y}^{(j)}) \right)^\mathsf{T} \Sigma^{-1} \left(\boldsymbol{\mu}_j - \delta(\bar{Y}^{(j)}) \right)$$
$$\left. \left. + \frac{n_j}{n_j\tau^2 + 1} (\bar{Y}^{(j)} - \boldsymbol{\mu}_0^{(j)})^\mathsf{T} \Sigma^{-1} (\bar{Y}^{(j)} - \boldsymbol{\mu}_0^{(j)}) \right] \right\},$$

where

$$\delta(\bar{Y}^{(j)}) = \frac{\tau^2}{n_j\tau^2 + 1} \left[n_j \bar{Y}^{(j)} + \frac{1}{\tau^2} \boldsymbol{\mu}_0^{(j)} \right] \quad \text{and} \quad \bar{Y}^{(j)} = \frac{1}{n_j} \sum_{\ell=1}^{n_j} Y_\ell^{(j)}.$$

Integrating with respect to $\boldsymbol{\mu}_j$, we obtain

$$\int \prod_{\ell=1}^{n_j} N(Y_\ell^{(j)}|\boldsymbol{\mu}_j, \Sigma_j) N(\boldsymbol{\mu}|\boldsymbol{\mu}_0^{(j)}, \tau^2\Sigma_j)\, d\boldsymbol{\mu}_j = \left(\frac{1}{2\pi} \right)^{pn_j/2} \frac{1}{|\Sigma|^{n_j/2}} \left(\frac{2\pi}{n_j\tau^2 + 1} \right)^{p/2}$$
$$\times \exp\left\{ -\frac{1}{2} \left(\sum_{\ell=1}^{n_j} (Y_\ell^{(j)} - \bar{Y}^{(j)})^\mathsf{T} \Sigma^{-1} ((Y_\ell^{(j)} - \bar{Y}^{(j)})) + \frac{n_j}{n_j\tau^2 + 1} (\bar{Y}^{(j)} - \boldsymbol{\mu}_0^{(j)})^\mathsf{T} \Sigma^{-1} (\bar{Y}^{(j)} - \boldsymbol{\mu}_0^{(j)}) \right) \right\} \times \pi(\Sigma_j). \tag{22}$$

Under the assumption $\Sigma_j = \text{diag}(\sigma^2_{1j}, \ldots, \sigma^2_{pj})$ and considering the priors $\sigma^2_{rj} \sim IG(a,b)$, the expression in braces in (22) simplifies to

$$\sum_{r=1}^{p} \frac{-1}{2\sigma^2_{rj}} \left(n_j s^2_{rj} + \frac{n_j}{n_j\tau^2+1} (\bar{y}^{(j)}_r - \mu^{(j)}_{0r})^2 \right),$$

where $\bar{y}^{(j)}_r = \sum_{\ell=1}^{n_j} y^{(j)}_{\ell r}/n_j$ and $s^2_{rj} = \sum_{\ell=1}^{n_j} (y^{(j)}_{\ell r} - \bar{y}^{(j)}_r)^2/n_j$, for $r = 1, \ldots, p$ and $j = 1, \ldots, k$. Lastly, we note that the integral with respect to σ^2_{rj} is the kernel of a gamma distribution, and a standard calculation yields

$$m(\mathbf{Y}|\ \omega_k) = \left(\frac{2}{b}\right)^{pka} \frac{1}{\pi^{np/2}\Gamma(a)^{pk}} \times \left[\prod_{j=1}^{k} \frac{\Gamma\left(\frac{n_j}{2}+a\right)^p}{(n_j\tau^2+1)^{p/2}} \right] \left[\prod_{j=1}^{k} \prod_{r=1}^{p} \frac{1}{\left(n_j s^2_{rj} + n_j \frac{(\bar{y}^j_r - \mu^j_{0r})^2}{n_j\tau^2+1} + \frac{2}{b} \right)^{n_j/2+a}} \right].$$

Discussion of "Testing for the existence of clusters" by María Jesús Bayarri, Adolfo Álvarez and Daniel Peña

M. J. Bayarri
Universitat de València, Spain

It is a pleasure for me to comment on the paper by Fuentes and Casella, and I thank the editors for their kind invitation. This paper deals with (Bayesian) testing of H_0 : no clusters versus H_1: exactly k clusters, both hypotheses being composite. Although this problem has been treated at large in the literature, the authors present an original parametrization of H_1 in terms of the possible partitions of the data into k clusters. It is more common in the literature to instead consider indicator latent variables z_{ij} which equals 1 if observation y_i is in cluster j, and 0 otherwise. Both arguments should be equivalent (for compatible priors), and I would have liked to see a discussion of the relative merits of each approach. The combinatorial arguments and results in the paper, required for dealing with the space of partitions are very elegant.

Conditional on a given partition, the authors use simple conjugate hierarchical priors allowing close form derivation of the (conditional) marginal likelihoods (conditional prior predictive) needed for computing the Bayes Factors. Closed form computations are often (as in the case of huge model spaces) highly desirable.

I have several concerns with respect to the prior used, but I will focus only on the ones that seem methodologically most dangerous to me. One is the prior used for the σ's, specially in the arbitrariness of the scale of the 'vague' inverted gamma prior of convenience used (see Section 3). It is well known that Bayesian model selection requires an *extremely* careful choice of the scale of any 'objective' prior used, and even if considering the variances to have a mean of 1 and a 'high variability' (as quantified by a large but arbritary variance of 100) might be innocuous for inference under a given model, it is not so for model selection.

Another concern is the "post-processed" restriction of cluster size to be at least 15% of the sample size. This restriction was not based on strong prior believes, but was adopted because of some undesirable features on the analysis under the prior without this constraint. These undesirable results might indicate that the prior was not good enough, and that a prior discouraging very small clusters should be used instead. Note that 'discouraging' is very different from 'truncating' the cluster size: with truncation, a cluster with a smaller sample size (even if arbitrarily close to the imposed minimum size) will never be discovered, even if overwhelmingly supported by the data.

Perhaps the methodology that worries me the most among the ones used in the paper is the solving of a well posed model selection problem by a set of, independently solved, comparisons among two models. Take the situation in Section 3. As stated, researchers are interested in finding whether there are no clusters $(\kappa = 1)$, or two

($\kappa = 2$), three ($\kappa = 3$) or four ($\kappa = 4$) clusters. That is, interest is in selecting one model (or hypothesis) among the 4 previous ones. Solving this model selection problem by solving 3 independent testing problems of $H_0 : \kappa = 1$ versus $H_1 : \kappa = k$ for $k = 2,3,4$ seems to me not only methodologically incorrect (whether a Bayesian or a frequentist procedure is used), but also it is not even useful as an approximation, since the correct Bayesian analysis uses exactly the same ingredients as this ad-hoc analysis.

Posing the problem as a model selection problem, one is faced with choosing between 4 models or hypothesis, $\mathscr{M}_k : \kappa = k$ for $k = 1,2,3,4$. One then assesses prior probabilities to the models, $p(\mathscr{M}_k) = \pi_k$ and derives the corresponding posteriors probabilities $p(\mathscr{M}_k|\mathbf{Y})$. Notice that in this formulation the models probabilities (both prior and posterior) add to 1, as they should. It is trivial to show that the posterior probabilities can be expressed in terms of the Bayes factors as

$$p(\mathscr{M}_k|\mathbf{Y}) = \left[\sum_{j=1}^{4} \frac{\pi_j}{\pi_k} B_{jk}\right]^{-1},$$

where B_{jk} is Bayes factor of model $\mathscr{M}_j$ to model $\mathscr{M}_k$. Also, since $B_{jk} = B_{j1} \times B_{1k} = B_{j1}/B_{k1}$ it is trivial to derive the correct Bayes posterior probabilities of each model from the Bayes factors already computed in the paper: no new inputs are needed (recall that H_0 in the paper is here $\mathscr{M}_1$). Indeed if, in the spirit of the paper, we consider all models equally likely a priori, the posterior probabilities are given by:

Label	$\Pr(\kappa = 1)$	$\Pr(\kappa = 2)$	$\Pr(\kappa = 3)$	$\Pr(\kappa = 4)$	$\Pr(\kappa \geq 2)$
i00F-0173-01	0.0208	0.9254	0.0536	0.0002	0.9792
i00F-0183-01	0.2228	0.7696	0.0074	0.0002	0.7772
i01S-0026-18	0.0366	0.8941	0.0648	0.0045	0.9634
i02S-0302-16	0.6001	0.3332	0.0643	0.0024	0.3999

These are probability distributions, their interpretation is clear and there is no need to calibrate anything. The last column gives the probability of at least two clusters, a probability that the authors could not compute (remember that I solely used the outputs from this paper, and nothing else, to produce the table above).

Addressing the problem as a model selection problem makes the picture much clearer and provides appropriate measures of evidence for each of the models. Also, the conclusions are somewhat different from those in the paper. Thus for labels iOOF-0173-01 and i01S-0026-18 there is strong evidence *only* for 2 clusters, but not for 3 clusters as concluded in Section 6 (indeed the evidence is quite strong *against* existence of 3 clusters). For i00F-0183-01, chances of 2 cluster is about 77%, but there is a non-negligible probability (about 22.3%) that there are no clusters. Lastly, for label i02S-0302-16, the data entirely rules out again existence of 3 or 4 clusters, but while the odds are about 3 to 2 for "no cluster" against "2 clusters" (hence favouring no clusters), there seems to be considerable uncertainty for this data set, as appropriately reflected by the

posterior distribution. The differences in the conclusions between the model selection analysis just presented, and the ad-hoc analysis proposed by the authors (based instead on repeated independent tests of only 2 models) is not more dramatic because in this example the whole picture is well captured by deciding between "no cluster" or "2 clusters". If the data would have given clear indication of 3 clusters, the differences between the two analyses would have been much larger.

My last methodological piece of disagreement with the authors is in the calibration process. First, as said before, the probability distribution on the model space is the right (inferential) answer to this problem and, as a legitimate probability distribution, it does not need any calibration. A different issue is that of deciding thresholds for optimal decisions. To put the argument in the same footing as the developments in the paper, I'll restrict myself to discussing the simple testing of H_0 versus H_1 (although the correct formulation, would require an overall loss function). When this testing is explicitly addressed as a decision problem, a loss function is needed, and the optimal decision (often posed as 'accept' or 'reject' H_0) is the one minimizing the expected loss. Since the decisions rules in the paper are all in terms of posterior probabilities, the loss function implicitly considered for this testing is a $0-\ell_i$ loss, that is, the loss for a correct decision is 0 and the loss for *incorrectly* deciding that H_i is true is ℓ_i. If the two type of errors are considered equally bad, $\ell_1 = \ell_2$, and we have the ubiquitous $0-1$ loss. In general, the optimal decision is to reject H_0 if, in paper's notation,

$$BF_{10}(\mathbf{Y}) > \frac{\pi_0}{\pi_1}\frac{\ell_1}{\ell_0} \leftrightarrow \Pr(H_0|\mathbf{Y}) < \frac{\ell_0}{\ell_0+\ell_1}$$

Thus, the optimal thresholdings for posterior probabilities and for Bayes factors for this simple loss function are as given above, and no calibration is needed. The calibration developed in the paper is particularly dangerous for two reasons: 1) it is only based on one type of error (the error under the null); there is no decision procedure, whether frequentist (minimax) or Bayesian which does not take into account both type of errors, and 2) the cut-off point is data dependent, effectively requiring a loss function that depends on the data in inappropriate ways. This data-dependent decision rules can be shown to exhibit aberrant behaviour in terms of expected losses.

In our model checking work (see authors' references) we used p-values because, in contrast with the problem addressed by the authors, ours was not a well defined model selection scenario; in particular, there was no alternative model, only the null model was identified. Hence we had to resort to less than optimal procedures.

Last, I would like to briefly address one more worrisome issue, namely multiplicity. Multiplicity issues are clearly present in this paper because the same data are analyzed multiple times. Bayesian model selection (by which I refer to selecting one among a set of models, and not to the multiple individual testing given in this paper) can control for multiplicity through appropriate priors on the model space (see Scott and Berger, 2008). This is an important issue, but out of the scope of this discussion.

Let me finalize by congratulating the authors again for a provocative and interesting paper, and by thanking the editors for the invitation to comment on this paper.

References in the discussion

James G. Scott and James O. Berger (2008). *Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem*. Duke University. Department of Statistical Science. Discussion Paper 2008-10.

Acknowledgements

This research was supported in part by the Spanish Ministry of Education and Science, under grant MTM2007-61554.

Adolfo Álvarez and Daniel Peña
Universidad Carlos III de Madrid

The usual way to approach the cluster problem from the Bayesian point of view is to compute the posterior probability of the hypothesis $H_k = k$ clusters in the data, and to select the number of clusters by the maximum value of this probability. This procedure can also be approximated by using the BIC to select the number of clusters. The approach presented in this paper is new. The authors are able to compute the sampling distribution of $P(H_0|\mathbf{Y})$, when H_0 is true and $\mathbf{Y}$ is the data set. The method they use is very ingenious and the results obtained thought-provoking. Thus, we want to congratulate the authors for this interesting contribution to the cluster problem.

The paper can be extended in several directions. First, the assumption that the covariance matrices are diagonal is very restricted because the most interesting multivariate data sets do not have this property. In fact, the main reason to use multivariate methods is because we have data with a non-diagonal covariance matrix. On the other hand, if this hypothesis is relaxed and we allow for full covariance matrices the generalizations made in section 4.1 are not straightforward. Second, outliers are not taked into accout. If the data come from a heavy-tailed distribution, as if often the case, the null hypothesis that all the data come from a normal distribution should be rejected, and it may appear that the data are not homogeneous when in fact they have been generated by the same common distribution. Outliers could be incorporated by assuming groups of small size, even of size one, into the procedure. Third, the effect of the prior on the partitions is not clear, and the consequences of different priors need to be investigated. Fourth, it would be helpful to define a criterion to set the number of clusters, because, as presented in the article, it is not possible to compare the posterior probabilities obtained from different numbers of clusters.

A possible limitation of the proposed procedure is computation. For large data set with many possible clusters the procedure may be unfeasible. Assuming the set up of the paper, under the null we can solve the problem in one dimension and then it is shown in the paper that the Bayes factor for a given partition ω depends only on the cluster sizes. Thus, the number of terms in the sum in the computation of the Bayes factor is the number of ways we can split n into $n_1, \ldots, n_k$, where $\sum n_i = n$. This number can be huge for large n and moderate k.

The ideas presented in this paper can be extended to other problems. For instance, Peña, Rodriguez and Tiao (2004) proposed the SAR (splitting and recombine) cluster procedure based on a heterogeneity measure between each observation and the entire sample. The aim is to split the sample into homogeneous small groups, and then

recombine the observations to get the final data configuration. If the recombining process is done by groups instead of observations, then you need an hypothesis test to merge two groups. For this you can not use traditional tests like those of homogeneity of means and/or variances because the groups being tested are not independent since they are as built so that they are as homogeneous as possible. In this approach you can add groups one by one testing the hypothesis of two clusters at a time, avoiding the problem of comparing between several configuration clusters for the same data. Some calibration about the minimum cluster size should also be necessary in this case, and the ideas presented in this paper can also be useful in this context.

In closing, I want to congratulate the authors for this excellent work that I hope will stimulate further research in this important field.

Peña, D, Rodriguez, J. and Tiao, G. C. (2004). A general partition cluster algorithm. *Proceedings in Computational Statistics*. J. Antoch (editor). Phsyca-Verlag, New York, 371-380.

Rejoinder

We would like to thank Professors Álvarez, Bayarri, and Peña for their careful reading of our paper. Their valuable comments and detailed discussions have provided us with a better insight of the difficulties and alternatives related to the topic presented in our article. Here, we would like to take the opportunity to address some of their concerns.

Outliers and minimum cluster size

Let us start by commenting on the problem of outliers and minimum cluster size. As all discussants noted, this is a very important matter that needs to be carefully taken under consideration. We should first recall that our procedure allows for the existence of clusters of any size, including clusters of size 1 (there is no methodological constraint). The introduction of a minimum cluster size to constrain the space of partitions was done by explicit request of the researcher, for whom clusters of small size were meaningless in the context of the experiment. It is true, however, that we should not have been so strict.

The restriction we considered in the application completely dismisses the possible presence of outliers, since we force every single point to belong to some cluster. Proceeding this way may not be adequate for several reasons, and in this sense, the distinction between "discouraging" and "truncating" clusters of small size is quite relevant. This discouragement can be achieved by using a different prior, such as the marginal distribution of the number of clusters in a Dirichlet process, $\pi_D(\omega)$, mentioned in Section 2.3.

Having said that, we should also point out that the solution we provide in the paper allows us to develop a sampling strategy, which is needed in order to provide a solution. The choice of a more clever prior, that avoids setting a minimum cluster size, has to provide not only a satisfactory theoretical solution, but also admit the practical implementation of an estimation procedure, such as the one discussed in Section 2.4, in order to solve the problem. The space of possible partitions contemplated in the model is so big that avoiding the use of Monte Carlo techniques is unrealistic.

Model comparison and calibration

A different concern expressed in the discussions has to do with the limitation of the procedure when comparing the results of the tests for two (or more) different values of

κ (the number of clusters under the alternative hypothesis). The problem, as well noticed by the reviewers, relies on the fact that we did not consider any type of measure over the model space and, therefore, we can not reconcile the outputs from tests that allow a different number of clusters under the alternative.

One possible solution to this problem would consist of specifying a probability measure over the model space, as detailed by Prof. Bayarri. Although we did not discuss this in the paper, we did examine this problem and evaluated a number of alternatives. Unfortunately, this more general solution led us to two difficulties that we did not want to address at that time. One of them (the simpler one) was how to determine a valid criterion to set the maximum number of clusters in a general setting. Typically, in a real situation, the experts would have an idea of a reasonable upper bound for that number (in our application that number was 4), but this is not necessarily the case, in fact, this might be part of the research question.

Now, suppose we know the maximum number of clusters that may be present in the data. Then, a second (more difficult) problem was how to develop a calibration procedure for our method. Although the real need of a calibration is questioned by the discussants, this was an important concern for us for the following reason. In the context of the experiment that motivated the problem, the clusters would reflect the presence of certain mutant genes in the composition of the kernel that was measured. If we have two kernels with the same composition and same mutant genes, just due to the variability of the experiment the readings from the two kernels might differ, and so would the associated posterior probabilities. When we extend this situation to hundreds of kernels being analyzed (which was our case), it is desirable to have certain control on the error rate associated with the decision making. Determining the null distribution of the posterior probabilities allows us to have control on the type I error and, in addition, permits the analysis of multiple tests. We do this by implementing procedures that control the false discovery rate, procedures that are broadly accepted by practitioners. However, we were only successful in determining the null distribution of the posterior probabilities when testing against $\kappa = k$, arbitrary, but fixed.

The use of the frequentist calibration procedure presented in the paper has another attractive feature. One of the main concerns when using Bayes factors for model selection is that the results tend to be sensitive to the choice of the prior. In the context of our problem, changing the values of the parameters a and b of the inverted gamma distribution may greatly affect the values of the posterior probabilities. However, changing the values of such parameters will also affect the shape of the null distribution, and therefore, decision making based on the α-level test will be consistent in the sense that, given the data, the null hypothesis will be accepted or rejected regardless of the choice of the values for a and b. We did observe this phenomenon when simulating and testing the procedure and the reason for its occurrence is very simple. Once the values of the hyperparameters are set, the form of the posterior probabilities (our test statistic) is well determined, and the corresponding null distribution and α-percentile will be computed accordingly. It follows that the frequentist calibration provides us with

an objective decision rule in a very concrete sense. Having said all that, we agree that adding to the testing procedure the probabilistic structure over the model space would provide the experimenter with an additional piece of important information.

Final comments

Finally, we would like to call attention to some interesting comments that also offer possibilities for further investigation. Alvarez and Peña ask about relaxing the assumption that the covariance matrix for each cluster is diagonal. Considering a more flexible structure for such matrices would certainly make the procedure applicable in more diverse scenarios. Bayarri suggests an alternative to our calibration procedure, by introducing a loss ℓ_i when an incorrect decision is made. Although the selection of such a loss (which will determine the threshold for decision making) is no less arbitrary than determining cut-off points based on the α-level (the cut-off points depends purely on α and the null distribution, and not on the data), this alternative offers a new approach and new possibilities that should be explored.

The incorporation of the partition as a parameter in the model, although not new, is an idea less explored in the literature. This approach is different from the more common mixture model alternative, whose limitations are well discussed in Booth *et al.* 2008. On the other hand, our calibration procedure, which combines Bayesian and frequentist ideas on hypothesis testing and is, perhaps a bit controversial, offers a valid and interpretable answer to the problem.

We would like to end this rejoinder by thanking our discussants one more time for their thoughtful remarks. We truly hope that our work and their comments help to bring more interest and attention to this important subject.