TESTING FOR THE EXISTENCE OF CLUSTERS WITH APPLICATIONS TO NIR SPECTROSCOPY DATA

By

CLAUDIO FUENTES

A THESIS PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN STATISTICS

UNIVERSITY OF FLORIDA

2008

To my parents, who have been there at each step.

ACKNOWLEDGMENTS

I would like to acknowledge many people who gave me support and advice while writing this thesis. First, I want to thank my advisor, Distinguished Professor Dr. George Casella, who gave me the opportunity to work as his research assistant and to get involved with the project I present here. His wealth of knowledge was an invaluable source of help. It would not have been possible without his constant feedback and guidance.

I also want to thank my graduate committee members, Professor Dr. Clyde Schoolfield and Professor Dr. Mark Settles, for their additional assistance in my research. They provided a foundation for me to develop the correct methodologies to address specific issues within cluster analysis I discuss here.

I would like to give a special acknowledgment to Professor Dr. Alvaro Cofre for his constant support and advice. His teaching helped me to discover the beauty of mathematics. His example served as an inspiration for how I instruct in my own classroom. His encouragement brought me where I am today.

I am indebted to Vikneswaran Gopal for all his help and useful hints in the implementation of the codes to run the simulations included in this work.

Finally and most importantly, I would like to thank Alejandra for her unconditional love, and my parents, Jorge and Edith, for their love, encouragement and support in all that I do.

TABLE OF CONTENTS

APPENDIX

LIST OF TABLES

LIST OF FIGURES

Abstract of Thesis Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Master of Science in Statistics

TESTING FOR THE EXISTENCE OF CLUSTERS
WITH APPLICATIONS TO NIR SPECTROSCOPY DATA

By

Claudio Fuentes

August 2008

Chair: George Casella
Major: Statistics

The detection and determination of clusters has been of special interest among researchers from different fields for a long time. Different efforts have been made in cluster analysis, but most of them determine the clusters depending on the distance between the observations. Although these methods have been proven to work well, they are usually too sensitive to the metric that defines the distance and they lack statistical procedures that facilitate decision making.

In this paper we develop a procedure that permits testing for clusters using Bayesian tools. Specifically we study the hypothesis test $H_0 : \kappa = 1$ *vs.* $H_1 : \kappa > 1$, were κ denotes the number of clusters in a certain population. This problem can be solved by looking at the Bayes factors as in a model selection problem, and making decisions according to the posterior probabilities $P(H_0|data)$. Since the procedure is entirely data dependent, we calibrate our results by estimating the frequentist null distribution of the posterior probabilities. Hence we can establish appropriate cutoff points to reject the null hypothesis at any desired α-level in the usual sense.

While our setting allows (in theory) for explicit calculation of the Bayes factors and the posterior probabilities, our method is computationally too intensive in the most part. To overcome this difficulty we propose an estimation procedure based on MCMC techniques. Finally, we present simulation studies that validate our conclusions, and we apply our method to NIR spectroscopy data coming from a genetic study in maize.

CHAPTER 1
INTRODUCTION

The analysis of clusters has become one of the most important problems in statistics today, due to the numerous applications in different fields such as machine learning, image analysis and genetics, among others. For this reason, cluster analysis has received special attention from researchers in the recent years and different methods have been developed for its study.

The fundamental problem in cluster analysis is simple to state: Given a set of n distinguishable objects, we wish to distribute the objects into groups or clusters in such a way that the objects within each group are similar while the groups themselves are different. The problem, as stated, presents two immediate difficulties: first, given a set of observations we need to determine if it is meaningful to partition the observations into different groups and if so, how many of them. In other words, we need to assess the significance of the clusters present in the data. Second, we need to explicitly determine the clusters.

In the present work we will provide a solution for the first of these issues and we will present an application to a clustering problem coming from the field of genetics. In the next two sections we provide more details about the genetic problem that motivates this research and briefly introduce the methodology we propose.

1.1 Clustering Problem in NIR Spectroscopy Data

Cereal grains have become one of the largest sources of food for humans (Pan, 2000). Then, increasing the yield of cereals and improving grain quality or nutritive value is essential for food security and minimizing agriculture's environmental impact.

To this end, understanding and determining the important genetic factors that affect kernel composition is crucial. In this context different efforts have been made, however, the genetic variation in amount, quality of protein, carbohydrates and oil make this problem difficult, which calls for new strategies and the development of new techniques.

For instance, the Illinois long-term selections for High-Protein, Low-Protein, High-Oil and Low-Oil lines were based on biochemical phenotypes that are caused by changes in the internal structure of the kernel (Dudley et al, 1974). Although these lines were used in commercial development of high-oil varieties, they showed significant yield losses relative to standard hybrid varieties (Lambert et al, 1998).

Recently, Settles et al. have been working on the identification of genes that cause dosage-dependent changes in seed weight or composition. More precisely, on the identification of a genome-wide collection of mutants with quantitative effects on the seed. Their approach requires the use of non-destructive technologies to measure composition and seed weights from single kernels in order to track segregation of these quantitative traits. They use single-kernel *Near Infrared Reflectance* (NIR) spectroscopy to collect the data.

The NIR spectroscopy uses light absorption between 800 and 2500 nm to detect chemical constituents. This technique is used to quantify a broad variety of organic molecules and is useful for the study of maize seeds (see Williams and Norris, 2001 for a review).

The major seeds constituents (protein, oil, starch, cellulose and water) have multiple near infrared absorption bands, since bonds between N-H, C-H and O-H absorb in the near infrared. Hence, NIR provides an information-rich spectrum allowing multiple chemicals and structures to be detected and quantified.

The drawback of this technique is that the spectrum is complex, making it difficult to interpret and identify specific peaks that correspond to individual kernel constituents. Therefore, an empirical calibration of the procedure is needed for a specific NIR spectrometer to interpret the spectra and predict kernel composition.

Multiple NIR calibrations have been developed for maize kernels (Orman and Schumann, 1991; Weinstock et al., 2006; Baye et al, 2006; Janni et al., 2008). However, each of these calibrations are specific to an individual spectrometer and spectra collection

apparatus. In addition, calibrations limit NIR to predict the chemicals that are measured analytically. Thus, NIR has the capacity to identify variation in constituents for which there is no calibration.

In this context, the problem that motivates the present thesis is the determination of novel constituent variation.

Detecting and identifying well differentiated clusters from NIR spectra should identify kernels with differing composition. 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.

Then, the detection and identification of clusters is an important primary step for the identification of composition mutants, because calibrations for all possible chemical changes within a kernel are costly and time consuming. In addition there are likely to be novel composition variants for which is not possible to develop a calibration.

As we mentioned above, the direct analysis of the NIR spectra is difficult. Therefore, in order to look for clusters some data processing is needed.

For each kernel the absorption for nearly 800 wavelengths is measured and between 60 and 70 kernels are considered from each ear of maize. To overcome the dimensionality problem, typically a *Principal Components Analysis* (PCA) is performed, looking for the directions that capture most of the variability in the data.

Figure 1-1 shows the scatterplots for the first 5 principal components of the NIR spectra for label 4S520701.

Although visual inspection of the scatterplots suggests the presence of clusters in some of them, we notice that is difficult to make a conclusion about the existence of clusters in the data. Therefore, an objective measurement of the degree of evidence for the presence of clusters is necessary. Such a measurement can be implemented as a decision

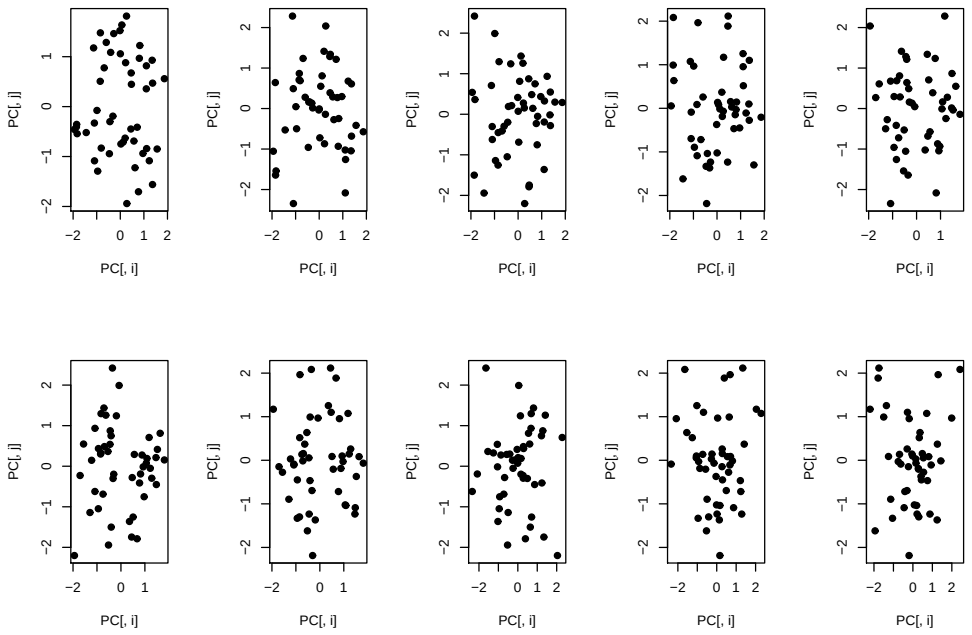

Figure 1-1. Scatterplots of the first five principal components of the NIR data for label 4S520701.

criteria to avoid the subjectivity of the visual inspection, which, in addition, is not even possible when the number of principal components exceeds three.

Hence, a statistic expressing true presence or absence of clusters would greatly facilitate the analysis of complex data sets and is required.

1.2 Cluster Analysis Methods and Our Methodology

Although there are many methods for clustering data, there are relatively few methods for assessing whether the clusters are statistically significant, a question that has been asked by a number of experimenters.

The idea of testing for significant clusters has a long history, going back to Hartigan's Rule (Hartigan 1975), with more recent contributions from Tibshirani et al. (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 et al. (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 adhoc; for example Bolshakova et al. (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. Pritchard et al. (2000) consider a Bayesian model and put a prior on the (unknown) number of clusters to compute posterior probabilities.

More recently, Booth et al. (2007) 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 and evaluate it.

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 the following chapters we explain how to construct the test in the Bayesian framework and we develop a calibration procedure to take into account the variability of the method and facilitate the interpretation of the results.

CHAPTER 2
THEORETICAL FRAMEWORK

In this chapter we will discuss the theoretical foundation that allows us to test for the number of clusters in a Bayesian framework and we will determine the model and distributions needed to perform the analysis in a general setting.

2.1 Notation Setup and Preliminaries

We will 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. Let $j = 1, \ldots, k$ be the number of clusters, such that the j-th one contains n_j elements of $\mathbf{Y}$. Since each Y_i $(1 \leq i \leq n)$ can be only in one cluster, we have $n_1 + \ldots + n_k = n$.

For instance, if $n = 6$ and $k = 2$, we might have the clusters

$$\begin{array}{cc} \{Y_1, Y_3, Y_4, Y_6\} & \{Y_2, Y_5\} \\ j = 1 & j = 2 \\ n_1 = 4 & n_2 = 2 \end{array}$$

Clearly, other configurations are also possible. For example, if $n = 6$ and $k = 3$ we might have

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

or

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

For this reason, we assume that there exists an unknown parameter κ which determines the number of clusters and a parameter ω (depending on κ) which determines the *partition* of $\mathbf{Y}$ into κ (non-empty) clusters.

Given $\kappa = k$, define $\mathcal{S}_{n,k}$ to be the set of all partitions of n objects into k clusters. Standard combinatorial results tell us that the number of elements in $\mathcal{S}_{n,k}$ is $S(n, k)$, the

Stirling number of the second kind, given by the formula (Gould, 1960)

$$S(n,k) = \frac{1}{k!}\sum_{j=1}^{k}(-1)^{k-j}\binom{k}{j}j^n.$$

Now, for any fixed partition $\omega \in \mathcal{S}_{n,k}$, 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, with $Y_i^{(j)} = Y_{\lambda(i,j)}$, where $\lambda(i,j)$ is the i-th smallest index among the vectors allocated in cluster j. This way, we can fully characterize the observations in each cluster. For instance, in our second example, for cluster $j = 3$ we have $Y_1^{(3)} = Y_2$, $Y_2^{(3)} = Y_5$ and $Y_3^{(3)} = Y_6$.

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

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

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

2.2 Hypothesis Testing and Bayesian Analysis

Our aim is to test the hypothesis

$$H_0: \text{ No clusters } \textit{ vs. } H_1: \text{ At least two clusters.} \tag{2–1}$$

Since the number of clusters must be a positive integer (notice that "no clusters" is the same as "one cluster"), we can state the hypothesis (2–1) in terms of the parameter κ as follows

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

Notice that the alternative hypothesis, H_1, has a complex structure. For this reason, we will consider first the more simple hypothesis

$$H_0 : \kappa = 1 \textit{ vs. } H_1 : \kappa = k, \tag{2–2}$$

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 (2–2), that is

$$BF_{10} = \frac{m(\mathbf{Y}|\ \kappa = k)}{m(\mathbf{Y}|\ \kappa = 1)}, \tag{2–3}$$

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

On the other hand, by definition

$$BF_{10} = \frac{P(H_1|\mathbf{Y})}{P(H_0|\mathbf{Y})} \Big/ \frac{P(H_1)}{P(H_0)},$$

where it is customary to make the prior assumption that the hypotheses being tested are equally likely.[1] Hence, assuming that $P(\kappa = k) = P(\kappa = 1) = 1/2$, we can compute the posterior probability of H_0 as

$$P(H_0|\mathbf{Y}) = \frac{1}{1 + BF_{10}}. \tag{2–4}$$

This quantity can be used as a model comparison criteria. For our hypothesis setting, small values of $P(H_0|\mathbf{Y})$ will provide evidence against H_0.

Since we will use the posterior probability as a measurement of the strength of the evidence against the null hypothesis, we will refer to this quantity as the *Bayesian P-value*[2] throughout this paper.

[1] This assumption is typically used as a way to reflect the ignorance prior to the experiment.

[2] The term *Bayesian P-value* is often referred to another concept in the Bayesian literature, but any of those definitions will not be included in the present paper to avoid confusion.

Finally, observe that conditioning on $\kappa = k$ in (2–3), involves considering all the possible partitions $\omega \in \mathcal{S}_{n,k}$ that generate k clusters. Then, if we define ω_1 to be the only existing cluster when $\kappa = 1$, using the *Law of Total Probability* we can rewrite the Bayes factor in terms of the partitions ω as

$$BF_{10} = \sum_{\omega \in \mathcal{S}_{n,k}} \frac{m(\mathbf{Y}|\ \omega)}{m(\mathbf{Y}|\ \omega_1)} \frac{\pi(\omega)}{\pi(\omega_1)}, \tag{2–5}$$

where $\pi(\omega)$ and $\pi(\omega_1)$ denote probabilities for the partitions ω and ω_1 respectively.

In the next section, we will specify the distributions that will allow us to obtain an expression for the Bayes factor in closed form.

2.3 Model and Distributions

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

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

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 $\boldsymbol{\Sigma} = (\Sigma_1, \ldots, \Sigma_k)$.

From the Bayesian perspective, in order to complete the specification of the model, we need to put prior distributions on $\boldsymbol{\mu}_j$ and Σ_j, for $j = 1, \ldots, k$. We consider

$$\begin{aligned} \boldsymbol{\mu}_j &\sim N(\boldsymbol{\mu}_0^{(j)}, \tau^2 \Sigma_j), \\ \Sigma_j &\sim \pi(\Sigma_j), \end{aligned}$$

where $\boldsymbol{\mu}_0^{(j)} = (\mu_{01}^{(j)}, \ldots, \mu_{0p}^{(j)})'$. For Σ_j, we could put a Wishart prior, but we will take a simpler approach and assume that $\Sigma_j = \text{diag}(\sigma_{1j}^2, \ldots, \sigma_{pj}^2)$, with

$$\sigma_{rj}^2 \sim IG(a, b) = \frac{1}{\Gamma(a) b^a} \frac{1}{(\sigma_{rj}^2)^{a+1}} e^{-1/b\sigma_{rj}^2},$$

where $r = 1, \ldots, p$ and $j = 1, \ldots, k$.

Since we are interested in the marginal distribution of the data, $\mathbf{Y}$, given the partition ω_k, we need to evaluate

$$m(\mathbf{Y}|\,\omega) = \int L(\boldsymbol{\mu}, \Sigma, \omega | Y_1, \ldots, Y_n) N(\boldsymbol{\mu} | \boldsymbol{\mu}_0^{(j)}, \tau^2 \Sigma_j) \pi(\Sigma_j) d\boldsymbol{\mu}_j d\Sigma_j.$$

Completing squares, we first integrate with respect to $\boldsymbol{\mu}_j$ to obtain

$$\begin{aligned} &\int \prod_{\ell=1}^{n_j} N(Y_\ell^{(j)} \,|\boldsymbol{\mu}_j, \Sigma_j) N(\boldsymbol{\mu} | \boldsymbol{\mu}_0^{(j)}, \tau^2 \Sigma_j)\, d\boldsymbol{\mu}_j \\ &= \left(\frac{1}{2\pi}\right)^{p n_j/2} \frac{1}{|\Sigma|^{n_j/2}} \left(\frac{2\pi}{n_j\tau^2 + 1}\right)^{p/2} \\ &\quad \times \exp\left\{ -\frac{1}{2} \left(\sum_{\ell=1}^{n_j} (Y_\ell^{(j)} - \bar{Y}^{(j)})' \Sigma^{-1} ((Y_\ell^{(j)} - \bar{Y}^{(j)})) + \frac{n_j}{n_j\tau^2 + 1} (\bar{Y}^{(j)} - \boldsymbol{\mu}_0^{(j)})' \Sigma^{-1} (\bar{Y}^{(j)} - \boldsymbol{\mu}_0^{(j)}) \right) \right\}, \end{aligned} \tag{2–6}$$

where $\bar{Y}^{(j)} = \sum_{\ell=1}^{n_j} Y_\ell^{(j)} / n_j$.

Then, since Σ_j is a diagonal matrix, the expression in braces in (2–6) simplifies to

$$\sum_{r=1}^{p} \frac{-1}{2\sigma_{rj}^2} \left(n_j s_{rj}^2 + \frac{n_j}{n_j\tau^2 + 1} (\bar{y}_r^{(j)} - \mu_{0r}^{(j)})^2 \right),$$

where

$$\bar{y}_r^{(j)} = \frac{1}{n_j} \sum_{\ell=1}^{n_j} y_{\ell r}^{(j)} \quad , \quad s_{rj}^2 = \frac{1}{n_j} \sum_{\ell=1}^{n_j} (y_{\ell r}^{(j)} - \bar{y}_r^{(j)})^2.$$

We finally obtain (see the appendix for details)

$$\begin{aligned} m(\mathbf{Y}|\,\omega) &= \left(\frac{2}{b}\right)^{pka} \frac{1}{\pi^{np/2} \Gamma(a)^{pk}} \\ &\quad \times \left[\prod_{j=1}^{k} \frac{\Gamma\left(\frac{n_j}{2} + a\right)^p}{(n_j\tau^2 + 1)^{p/2}} \right] \left[\prod_{j=1}^{k} \prod_{r=1}^{p} \frac{1}{\left(n_j s_{rj}^2 + n_j \frac{(\bar{y}_r^{(j)} - \mu_{0r}^{(j)})^2}{n_j\tau^2 + 1} + \frac{2}{b} \right)^{n_j/2 + a}} \right]. \end{aligned} \tag{2–7}$$

To set the prior parameters, recall that the mean and the variance of the Inverse Gamma distribution are $[b(a-1)]^{-1}$ and $[b^2(a-1)^2(a-2)]^{-1}$ respectively. We take

$$a = 2.01, \quad b = \frac{1}{a-1} \approx 0.990099,$$

this way, we obtain prior distributions for the σ_{rj}^2 with mean 1, and a large variance, reflecting lack of information on the variability within the clusters, prior to the experiment.

For simplicity, we set $\boldsymbol{\mu}_0^{(j)}$ to be equal to the sample means $\bar{y}^{(j)}$. Although choosing prior parameters once the data is observed is arguable, it has the advantage of centering the clusters where we observe the sample means.

In general, $\boldsymbol{\mu}_0^{(j)}$ can be taken to be of the form $X\beta$ where $\beta_{q\times 1}$ $(q < p)$ is a vector of parameter and $X_{p\times q}$ is a matrix of covariates associated to the experiment. In later development, we will model this part of the specification.

Having determined the distributions that specify the model, we can compute the ratio $m(\mathbf{Y}|\,\omega)/m(\mathbf{Y}|\,\omega_1)$ in (2–5), for every partition $\omega \in \mathcal{S}_{n,k}$. We obtain

$$\begin{aligned}\frac{m(\mathbf{Y}|\,\omega)}{m(\mathbf{Y}|\,\omega_1)} &= \left(\frac{2}{b}\right)^{pa(k-1)} \frac{(n\tau^2+1)^{p/2}}{\Gamma(a)^{p(k-1)}\Gamma\left(\frac{n}{2}+a\right)^p} \prod_{j=1}^{k} \frac{\Gamma\left(\frac{n_j}{2}+a\right)^p}{(n_j\tau^2+1)^{p/2}} \\ &\times \prod_{r=1}^{p} \left[\frac{\left(ns_r^2+\frac{2}{b}\right)^{n/2+a}}{\prod_{j=1}^{k}\left(n_j s_{rj}^2+\frac{2}{b}\right)^{n_j/2+a}}\right],\end{aligned} \tag{2–8}$$

where

$$s_r^2 = \frac{1}{n}\sum_{i=1}^{n}(y_{ir}-\bar{y}_r)^2\ , \quad \bar{y}_r = \frac{1}{n}\sum_{i=1}^{n} y_{ir},$$

and

$$s_{rj}^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(y_{ir}^{(j)}-\bar{y}_r^{(j)})^2\ , \quad \bar{y}_r^{(j)} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ir}^{(j)}.$$

Hence, in order to obtain a closed expression for the summands of the Bayes factor, we only need to explicitly determine the ratios $\pi(\omega)/\pi(\omega_1)$. To this end, we have to specify priors on the partitions space.

2.4 Prior on the Partitions

Since there is only one cluster of size 1, we take $\pi(\omega_1) = 1$. For $\pi(\omega)$ we have a number of choices, but here we will mention only two.

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 $\pi_S(\omega) = 1/S(n,k)$.

An alternative prior is the marginal distribution of the number of clusters in a Dirichlet process (Pitman, 1996)

$$\pi(\omega) = \frac{\Gamma(m)m^k}{\Gamma(n+m)} \prod_{j=1}^{k} \Gamma(n_j), \tag{2–9}$$

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 favoring partitions with large clusters.

2.5 Estimation of the Bayes Factor

Up to this point, we should be able to compute the Bayes factor in (2–5) and to determine the posterior probabilities $P(H_0|\mathbf{Y})$ in closed form. However, notice that the sum we need to compute is over the set of *all possible partitions* which has as many elements as it is determined by the *Stirling number of the second kind*. This introduces two new difficulties to the problem:

1. The number of summands involved in the calculation is often too large, even if the number of observations and clusters is relatively small. For instance, if $n = 48$ and $\kappa = 2$, we have $S(48,2) = 140,737,488,355,327$.

2. In order to compute the sum, we need to list *all the possible partitions* which is extremely difficult.

To overcome these difficulties, we will consider an indirect approach and estimate the value of the Bayes Factor using Monte Carlo techniques.

Let π and g be distributions on the partition space $\mathcal{S}_{n,k}$. 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 with the importance sampling sum

$$\begin{aligned} BF_{10} &= \sum_{\omega \in \mathcal{S}_{n,k}} \left[\frac{m(\mathbf{Y}|\,\omega)}{m(\mathbf{Y}|\,\omega_1)} \right] \pi(\omega) \\ &= \sum_{\omega \in \mathcal{S}_{n,k}} \left[\frac{m(\mathbf{Y}|\,\omega)}{m(\mathbf{Y}|\,\omega_1)} \right] \frac{\pi(\omega)}{g(\omega)} g(\omega) \\ &\approx \frac{1}{M} \sum_{i=1}^{M} \left[\frac{m(\mathbf{Y}|\,\omega^{(i)})}{m(\mathbf{Y}|\,\omega_1)} \frac{\pi(\omega^{(i)})}{g(\omega^{(i)})} \right] \\ &\approx \frac{\sum_{i=1}^{M} \left[\frac{m(\mathbf{Y}|\,\omega^{(i)})}{m(\mathbf{Y}|\,\omega_1)} \frac{\pi(\omega^{(i)})}{g(\omega^{(i)})} \right]}{\sum_{i=1}^{M} \frac{\pi(\omega^{(i)})}{g(\omega^{(i)})}}, \end{aligned} \tag{2–10}$$

where the expression in (2–10), 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 the importance sampling is not needed and we just compute the Monte Carlo sum.

To generate a random partition of n objects into k clusters we use the following strategy, suggested by Jim Pitman (personal communication). We use a vector of length n with $n-k$ 0's and k 1's, one of them in the first position. Then, we randomly generate a permutation of the the remaining $n-1$ elements to distribute (uniformly) the $k-1$ 1's in the last $n-1$ places.

Each 1 represents the start of a cluster. For example, if $n = 5$ and $k = 3$, the vector 11001 corresponds to the partition of five objects into clusters of size $1, 3, 1$. Once the partition has been generated, we randomly permute the $\mathbf{Y}$ vector, and place the Y_i's in the given partition. Although not immediately obvious (see the appendix for details), the probability of this partition ω is given by

$$g(\omega) = \frac{k!}{\binom{n-1}{k-1} \binom{n}{n_1\ n_2\ \ldots\ n_k}} \tag{2–11}$$

We observe that we can easily modify our strategy and the distribution g in order to generate partitions that corresponds to a minimum cluster size.

Finally, we can estimate the Bayes factor according to the following algorithm:

For $i = 1, \ldots, M$

1. Generate a random string of 1's and 0's as above.
2. Generate a random permutation of $Y_1, \ldots, Y_n$.
3. Cluster the permuted $Y_1, \ldots, Y_n$ according to the random string.
4. Calculate the ratio of marginals from (2–8).
5. Approximate the Bayes factor using (2–10).

Hybrid random walk algorithm. It is possible to slightly modify the algorithm described above, incorporating a random walk component. This way, the search algorithm will remain in areas of high probability to obtain a more accurate calculation of the Monte Carlo sum, and will maintain the correct stationary distribution.

We generate M partitions $(\omega^{(1)}, \ldots, \omega^{(M)})$ according to a Metropolis-Hastings algorithm, which combines two steps, an independent draw from the distribution g and a random walk, into a single algorithm. These steps are described below:

- *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 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 $\omega^{(t+1)} = \omega^{(t)}$

This is a Metropolis-Hastings algorithm with stationary distribution g.

CHAPTER 3
CALIBRATION PROBLEM

We have constructed a tool that allows us to test the hypothesis in (2–2), using posterior probabilities as a measurement of the strength of the evidence against the null. However, we notice that the *Bayesian P-value* is not a *P-value* in the usual sense.

Our Bayesian P-value measures (based on the data) how likely is the null hypothesis to be true in opposition to the alternative being tested and, in this sense, it is a valid measurement of evidence against the null. But, in contrast to the standard P-value, there is no error calibration for our procedure. For instance we lack a reference distribution, such as the null distribution, which determines the variability of the statistic when the null hypothesis is true.

Hence, one of the main drawbacks of our procedure is that we can not easily set up cutoff points to determine when we have strong evidence against the null hypothesis. All we know is that posterior probabilities below 0.5 suggests the presence of clusters and the lower the better. But how low should the Bayesian P-value be in order for the experimenter to make a good decision?

Another difficulty is how to compare different outcomes from an experiment. Since the variability of the posterior probabilities under the null hypothesis is unknown, it is hard to tell if the difference between two different outcomes is due to the actual sensitivity of the test in detecting clusters or if it is purely due to the random variation associated to the experiment.

These problems are not new in Bayesian analysis and some solutions have been presented in the literature. For instance, Jeffreys (1961) developed a scale to judge the evidence in favor of or against H_0 brought by the data. The scale goes as follows:

- If $\log_{10} BF_{10}$ varies between 0 an 0.5, the evidence against H_0 is *poor*.
- If it is between 0.5 and 1, it is *substantial*.
- If it is between 1 and 2, it is *strong*.

- If it is above 2 it is *decisive*.

Obviously, this scaling of the Bayes factor gives some indication of the strength of the evidence, but the precise bounds separating one strength from another are a matter of convention and they can be arbitrarily changed (Kass and Ratery, 1995).

Some efforts have been made in order to establish appropriate bounds on Bayes factor in order to obtain a less subjective measurement of the evidence against the null. Other efforts attempt to develop the analog of the frequentist P-value in the Bayesian paradigm. For references and details of these methods see Robert (2001) and Ghosh et al. (2006).

Here, we will address 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. Since the posterior probability is in one to one correspondence with the Bayes factor, we observe that determining the null distribution of the Bayes factor is an equivalent problem.

First, let us rewrite the Bayes factor (2–5) in terms of the data, $\mathbf{Y}$, as

$$BF_{10}(\mathbf{Y}) = \sum_{\omega \in \mathcal{S}_{n,k}} \lambda(\omega) T(\mathbf{Y}|\omega), \tag{3–1}$$

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

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

and

$$T(\mathbf{Y}|\omega) = \prod_{r=1}^{p} \left[\frac{\left(n s_r^2 + \frac{2}{b}\right)^{n/2+a}}{\prod_{j=1}^{k} \left(n_j s_{rj}^2 + \frac{2}{b}\right)^{n_j/2+a}} \right]. \tag{3–2}$$

Then, 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.

In the following sections we will discuss the behavior under the null hypothesis of the components $T(\mathbf{Y}|\omega)$ and from there we will consider the difficulties and alternatives in obtaining the null distribution of the Bayes factor.

3.1 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

$$\begin{aligned} y_1^{(1)}, \ldots, y_{n_1}^{(1)} &\sim iid\ N(\mu_1, \sigma_1^2) \\ y_1^{(2)}, \ldots, y_{n_2}^{(2)} &\sim iid\ N(\mu_2, \sigma_2^2) \\ \vdots \quad &\vdots \quad \vdots \\ y_1^{(k)}, \ldots, y_{n_k}^{(k)} &\sim iid\ N(\mu_k, \sigma_k^2) \end{aligned}$$

where $y_1^{(1)} + \ldots + y_{n_k}^{(k)} = y_1 + \ldots + y_n$ and $n = \sum_{j=1}^{k} n_j$.

When the null hypothesis is true, *i.e.*, there are no clusters in the data, we have $\mu_1 = \ldots = \mu_k = \mu$; $\sigma_1^2 = \ldots = \sigma_k^2 = \sigma^2$ and we can prove the following lemma, which is essentially contained in Cochrane's theorem.

Proposition 1. *Let*

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

Then, under the conditions described above, $u_1, \ldots, u_{k+1}$ are independent and

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

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. □

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

$$T(\mathbf{Y}|\omega) = \frac{1}{(\sigma^2)^{(k-1)a}} \frac{\left(\sum_{j=1}^{k+1} u_j + 2/b\sigma^2\right)^{\frac{n}{2}+a}}{\prod_{j=1}^{k} \left(u_j + 2/b\sigma^2\right)^{\frac{n_j}{2}+a}}, \tag{3–3}$$

where $(u_1, \ldots, u_{k+1})$ are defined as in Proposition 1. Hence, if the null hypothesis holds and if $\mathbf{X} = (x_1, \ldots, x_{n-1})$ is a vector of *independent and identically distributed* χ_1^2 random variables, we have

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

where we take $n_0 = 1$. This result leads to the following proposition.

Proposition 2. *Under the null hypothesis*

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

for every $\omega \in \mathcal{S}_{n,k}$, where $x_1 \ldots, x_{n-1}$ are iid χ_1^2 random variables.

We notice that the null distribution of $T(\mathbf{Y}|\omega)$ depends on the partition ω only through the cluster sizes $n_1, \ldots, n_k$. This, in addition to the previous two propositions, brings us to the following result.

Lemma 1. *Let $\mathcal{P}_{n,k}$ 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*

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

where

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

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

Proof. We have

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

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

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

$$\begin{aligned} BF_{10}(\mathbf{Y}) &\stackrel{d}{=} \sum_{\xi \in \mathcal{P}_{n,k}} T(\mathbf{X}|\xi) \sum_{\Lambda(\xi)} \lambda(\omega) \\ &= \sum_{\xi \in \mathcal{P}_{n,k}} \phi(\xi) T(\mathbf{X}|\xi), \end{aligned}$$

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

The importance of the previous results lies mainly on two key consequences. First, they provide with a known probabilistic structure for each one of the components present in the Bayes factor. This structure will be fundamental in order to obtain the null distribution of the Bayes factor and consequently in obtaining the desired posterior probabilities.

Second, they reduce the complexity of the problem, for now, in order 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, then the total number of partitions of 70 elements into 4 clusters is given by $S(70, 4) = 58074856870647580265837480687605796566705$, whereas the number of ways of writing 70 as the sum of exactly 4 integers is given by $p(70, 4) = 2484$. The reason for such a remarkable difference is because, under the null hypothesis, we do not take into account which elements go into a specific cluster and we only distinguish the partitions according to how many elements are in the clusters. In other words, we are not taking account of permutations of elements between clusters, but only the number of elements that form the clusters.

Finally, we need to extend these results to the multidimensional case ($p > 1$).

From the assumptions of normality and independence in the model we have 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 (3–2). Hence, no correlation is induced by our calculations and the generalization to higher dimensions proceeds in the obvious manner.

Unfortunately, even with these results the determination of the exact null distribution of the Bayes factor is difficult. In the following sections we will briefly discuss some of the problems that prevent us from obtaining the null distribution in closed form and some of the alternatives that can be considered to estimate it.

3.2 Difficulties in Determining the Exact Null Distribution of the Bayes Factor

Let us recall the Bartlett's statistic for the homogeneity of variances (Glaser, 1980)

$$L = \frac{\prod_{j=1}^{k} (s_j^2)^{\nu_j/\nu}}{\sum_{j=1}^{k} \frac{\nu_j}{\nu} s_j^2}, \tag{3–4}$$

where

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

and

$$\nu_j = n_j - 1 \ , \quad \nu = \sum_{j=1}^{k}\nu_j = n - k.$$

The L statistic in (3–4) has a known distribution and the similarities with $T(\mathbf{Y}|\omega)$ (for $p = 1$) suggest that the null distribution of the Bayes factor (or at least its moments) can be determined analytically. However, the derivation of the null distribution of L is based on the same techniques used to determine the asymptotic distribution of the likelihood ratio test (Jorgensen, 1993) which we can not be implemented in deriving the null distribution of T, due to some important differences, namely:

1. The shifting term $2/b$ in T has no equivalent in the statistic L.
2. The exponents in L satisfy $\sum_{j=1}^{k}\nu_j = \nu$, while in T,

$$\sum_{j=1}^{k}\frac{n_j}{2} + a = \frac{n}{2} + ka \neq \frac{n}{2} + a.$$

3. The denominator of L corresponds to the pooled variance, while in T,

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

On the other hand, suppose we can determine the exact null distribution of $T(\mathbf{Y}|\omega)$ for any given partition $\omega \in \mathcal{S}_{n,k}$. Then, in order to determine the null distribution of the Bayes factor we will have to consider the sum, which is the sum required to list all possible partitions of the integer n into exactly k integers. This last problem becomes extremely difficult as n and k increase.

Here, we consider two alternatives to address the problem. When n and k are fairly small (in the sense that we can list the partitions) we can determine the null distribution using saddlepoint approximations. Or we can determine the null distribution through simulations using Monte Carlo techniques.

3.3 Saddlepoint Approximation

When n and k are relatively small, in the sense that it is possible to list all the elements $\xi \in \mathcal{P}_{n,k}$, we can use a saddlepoint approximation to obtain an estimate of the tail probabilities of our statistic $P(H_0|\mathbf{Y})$ assuming the null hypothesis to be true. In other words, we can obtain the frequentist *P-value* associated with our observed posterior probability.

The result in Lemma 1 permits us to obtain an approximation for

$$P(\mathbf{X}) = \frac{1}{1 + BF_{10}(\mathbf{X})} \tag{3–5}$$

where $\mathbf{X} = (x_1, \ldots, x_{n-1})$ is a vector of *iid* χ^2_1 random variables.

For any fixed value of σ^2, we proceed as follows (Easton *et al.*, 1986). First we compute

$$\tilde{R}(T) = \kappa_1 T + \frac{(k+1)\kappa_2 T^2}{2} + \frac{\kappa_3 \sigma^3 (k+1)^2 T^3}{6} + \frac{\kappa_4 \sigma^4 (k+1)^3 T^4}{24},$$

where κ_i, correspond to the i-th cumulant of (3–5).

Then, we estimate the tail probability of $P(\mathbf{X})$ by

$$P(P(\mathbf{X}) > a) \approx \int_{T_l}^{\infty} \left[\frac{(k+1)\tilde{R}''(t)}{2\pi}\right]^{1/2} exp\{(k+1)(\tilde{R}(t) - t\tilde{R}'(T_0))\}dt,$$

where $\tilde{R}(T_l) = a$.

Thus, in order to approximate the tail probabilities of (3–5) we only need to compute $\kappa_1, \kappa_2, \kappa_3$ and κ_4, which can be obtained from the equations

$$\begin{aligned}
\kappa_1 &= \mu_1 \\
\kappa_2 &= \mu_2 - \mu_1^2 \\
\kappa_3 &= \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3 \\
\kappa_4 &= \mu_4 - 4\mu_1\mu_3 - 3\mu_2^2 + 12\mu_2\mu_1^2 - 6\mu_1^4,
\end{aligned}$$

where μ_i is the i-th moment of $P(\mathbf{X})$.

Finally, the values of μ_1, μ_2, μ_3 and μ_4 can be obtained through integration of a Taylor expansion of $P(\mathbf{X})$, using the fact that the terms in the Taylor polynomial are powers of χ^2 random variables.

Clearly, this procedure involves many calculations, and is only applicable when $\mathcal{P}_{n,k}$ is relatively small. Therefore, this technique while interesting from a theoretical perspective, is of little practical use.

For this reason, we will not explore further properties or consequences related with the saddlepoint technique, but rather, we will discuss an alternative which is relatively easy to implement and that has no restriction on the size of $\mathcal{P}_{n,k}$.

3.4 MCMC and Simulation Alternative

The results in Section 3.1 allow us to simulate observations from the null distribution of the posterior probabilities. Then, we can use these generated values to either construct histograms or do density estimation, depending on the interest.

Here we will focus in the generation of histograms, which is enough to capture the main features of the null distribution. The histograms will also permit us (via asymptotic results) to obtain cutoff points for the posterior probabilities at any *frequentist* α*-level* we want.

Again, the main difficulty lies on the calculation of the sum in the Bayes factor. As we have already pointed out, this requires us to list all the partitions in $\mathcal{P}_{n,k}$. Nevertheless we can proceed in a similar way as in section 2.5.

Since when the null hypothesis is true we only care about the cluster sizes, we follow the same strategy to generate the partitions according to g, but we do not take into account the permutations of the elements in the given partition. Thus, 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

$$g_0(\xi) = \frac{k!}{\mathcal{R}(n_1, n_2, \ldots, n_k)} \frac{1}{\binom{n-1}{k-1}}$$

where $\mathcal{R}$ is counting the number of partitions that give the same cluster sizes (see the appendix for a definition).

Then, we can generate values of the Bayes factor $B_{10}(\mathbf{X})$ as follows. Sample $\xi^{(1)}, \ldots, \xi^{(M)}$ from g_0 and compute the importance sampling sum

$$\begin{aligned} B_{10}(\mathbf{X}) &= \sum_{\xi \in \mathcal{P}_{n,k}} \phi(\xi) T(\mathbf{X}|\xi) \pi(\xi) \\ &\approx \frac{\sum_{i=1}^{M} \frac{\phi(\xi) T(\mathbf{X}|\xi) \pi(\xi^{(i)})}{g_0(\xi^{(i)})}}{\sum_{i=1}^{M} \frac{\pi(\xi^{(i)})}{g_0(\xi^{(i)})}}. \end{aligned} \tag{3–6}$$

In order to construct the histogram, we need to simulate $B_{10}(\mathbf{X})$ several times. To obtain N simulations, the final algorithm is:

For $\ell = 1, \ldots, N$ repeat:

For $i = 1, \ldots, M$

1. Generate a random string of 1's and 0's according to g_0.
2. Generate $x_1, \ldots, x_{n-1}$ *iid* χ^2_1 random variables.
3. Compute $\sum_{i=1}^{n-1} x_i$ and $\sum_{i=n_{j-1}}^{n_j - 1} x_i$ according to the random string.
4. Compute $T(\mathbf{X}|\xi)$.
5. Approximate the Bayes factor using (3–6).

Finally, to estimate the cutoff points for any given α-*level* we use the following result (Sen and Singer 1993)

Theorem 1 (Convergence of the empirical percentiles)**.** *Let $X_1, \ldots, X_n$ be a random sample from a distribution F for which its p-th percentile ξ_p is uniquely determined. Then $X_{(p)}$ converges a.s. to ξ_p, where $X_{(p)}$ denotes the p-th empirical percentile.*

Hence, we estimate the cutoff points for any α level by the corresponding α-th empirical percentile from our generated sample.

Observe that the Bayes factor in (3–1) depends on the values of σ^2, the common variance of the observations under the null. Thus, in order to generate a sample from the null distribution, we need to set a value for σ^2.

We will overcome this difficulty by estimating σ^2 by the sample variance. More precisely, we will center and re-scale our observations in the standard way.

CHAPTER 4
SIMULATION STUDIES

The computation of the posterior probabilities and the null distribution of the Bayesian P-value we present in this thesis are both based on Monte Carlo techniques. Therefore, before we use our procedure for data analysis we need to calibrate it and study its behavior. In particular, we need to put special attention to the convergence issues and the error rate of the method. In this section we present some simulation results to exhibit the main features of our technique.

4.1 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 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 4-1 we see the results for $n = 50$ and $k = 3$ based on 2500 simulations. For this case, the number of elements in the partitions space is $p(50, 3) = 208$. All of them were considered to obtain the exact Bayes factor (first row) and samples of size 52 (25% of the total number of elements in the partition space) were drawn from g_0 to compute the approximations (second row). We also 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 the 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 election of the prior for the partition space has little effect in the calculations.

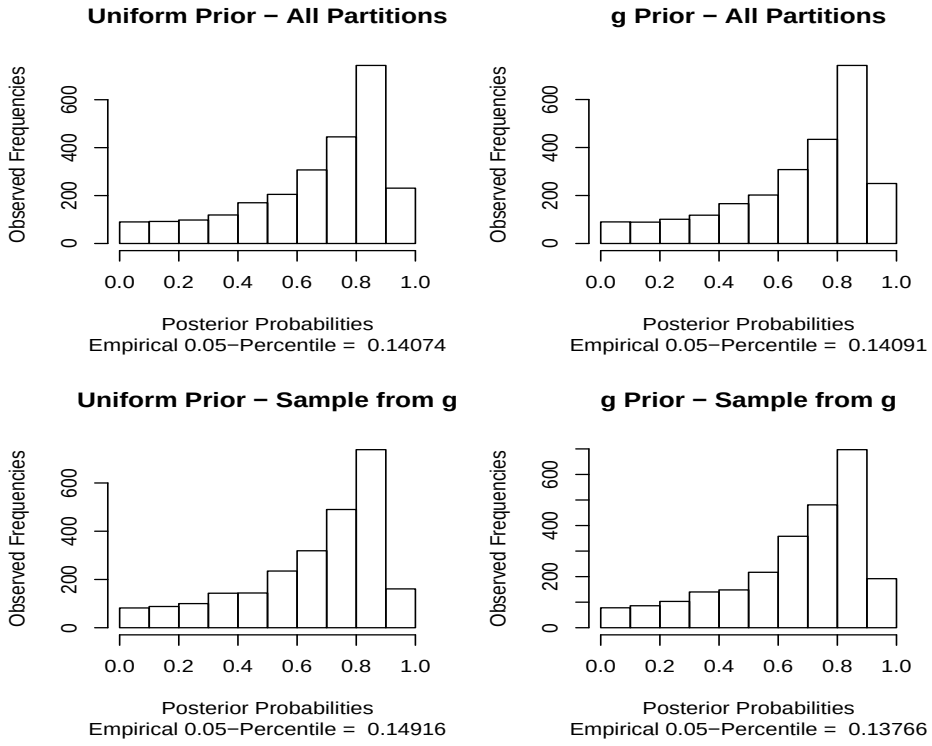

Figure 4-1. Histograms of the null posterior probabilities for $n = 50$ and $k = 3$ clusters based on 2500 simulations.

4.2 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 4-1 presents the results corresponding to 6 replications 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 changing the values of n and k, indicating that convergence of the empirical α-percentile is reached moderately fast.

4.3 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

Table 4-1. Posterior probabilities for 2500 simulations of 50 observations and 3 clusters. The number of considered partitions per iteration is 52.

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

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.

In Table 4-2 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 decrease 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 behavior is that for fixed n, as k increases the histograms, while still skewed to the left, tends to spread more mass to smaller values resulting in fatter tails instead of the expected thin tails.

Table 4-2. Cutoff points for minimum cluster size 1 and α-level 0.05, based on 5000 simulations. The number in parenthesis correspond to the standard errors after 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)

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 behavior*. Since these deviations fall randomly in different directions, it is general 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 determines a cluster, then we observe the previous behavior changes the direction, *i.e.*, 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 behavior is that once the minimum cluster size is determined, we can not 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 cluster by chance.

Table 4-3 shows the results of simulations obtained under the same conditions we described above, but setting the MCS equal to the 15% of the observations. We can see how the introduction of the MCS as a new parameter, reverses the pattern observed in Table 4-2 and also increases the value of the empirical 0.05-percentiles.

Table 4-3. Cutoff points for minimum cluster size 15% of the observations and α-level 0.05, based on 5000 simulations.

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)

In general, as the MCS 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 than spread in the tails as we can observe in Figure 4-2. In other words, the extra restriction provides more intuitive results for the simulated posterior probabilities.

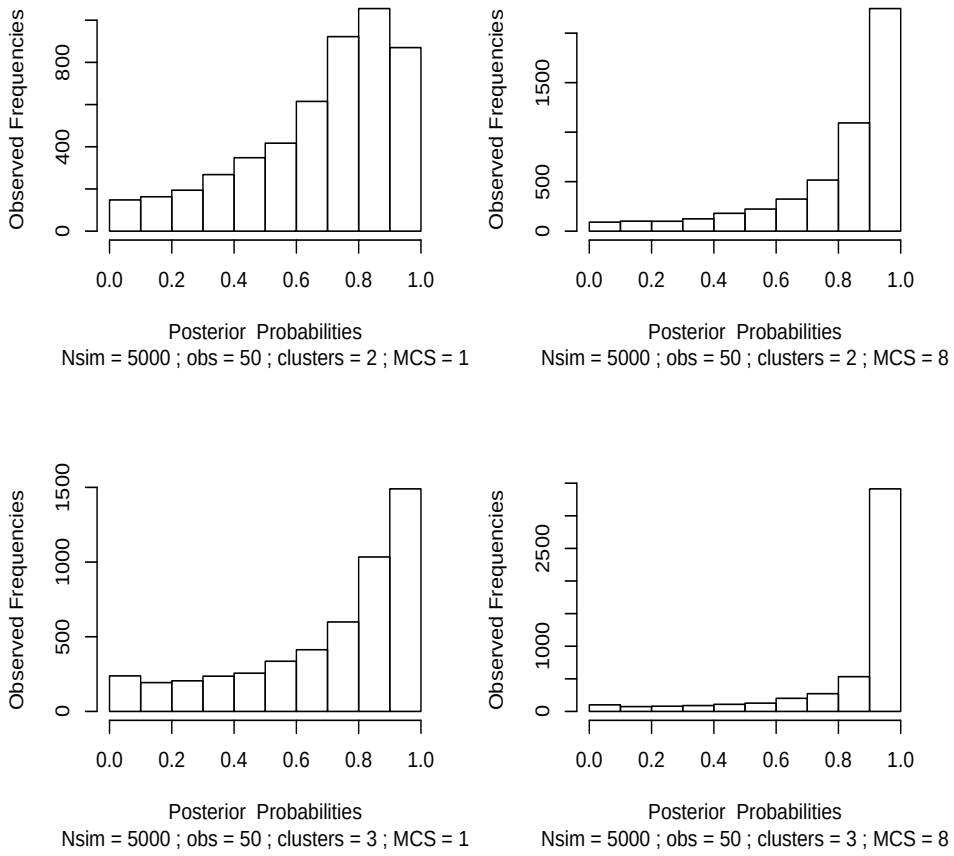

Figure 4-2. Histograms of the null posterior probabilities for $n = 50$ and $k = 2$ and 3 clusters based on 5000 simulations. The minimum cluster size is set equal to 1 observation and 15% of the observations in each case.

4.4 Optimal Number of Partitions

We need to determine the *optimal number of partitions* (ONP) to be considered in our simulations, that is, how many elements from the space of partitions need to be sampled in order to obtain good estimates of our posterior probabilities. We separate the problem in two cases: the ONP for the *regular* posterior probabilities, *i.e.*, the posterior probabilities we compute from the data, and the ONP for the posterior probabilities under the null hypothesis.

ONP for the general case. For fixed n and k, recall that the number of partitions is given by the Stirling number $S(n, k)$. As we discussed earlier, this number is usually

large which makes the exploration of the partition space difficult, even under a probabilistic scheme and then, the ONP has to be determined upon convergence of the posterior probabilities.

To this end, we sequentially increase the number of sampled partitions and recomputed the posterior probability according to the Monte Carlo sum until we reach convergence. That is, until the *oscillation* of the sequence of Bayesian P-values stabilizes. Of course, the number of partitions needed to reach convergence will vary as we change n and k. For our purposes, with $50 \leq n \leq 100$ and $k = 2, 3, 4$, our simulations indicates that we need over 200000 sampled partitions.

ONP under the null hypothesis. Under the null hypothesis the total number of elements in the partition space reduces considerably. but we need to simulate a large number of posterior probabilities in order to construct the histograms and compute the empirical α-percentiles.

To assess the ONP in this case, we consider again $n = 10, 25, 50$ and $k = 2, 3, 4$. Then, for a fixed set of generated observations we computed the empirical α-percentile obtained by computing the posterior probabilities considering all partitions and we compare with the results obtained by sampling a number of partitions equivalent to 25%, 50% and 100% of the total number of partitions given by $p(n, k)$. Since our sample procedure allows repetitions, we also considered the effect of *oversampling*, that is, sampling from the space of partitions more than $p(n, k)$ (the total number of elements in the partition space) times.

In Table 4-4 we show the results for 50 observations and 3 clusters where 2500 simulations were considered. We observe that the discrepancy in the results is less than 0.005, which is due not only to the number of partitions but also to the considered number of simulations. Similar results are obtained for the other cases suggesting that sampling as many partitions as about 25% percent of the total number of elements in the partition space is sufficient to obtain reasonable results.

Table 4-4. Posterior probabilities for 1000 simulations of 50 observations and 3 clusters.

Part/Iteration	α-level	α-percentile
all		0.14341
52		0.14852
104	0.05	0.14139
208		0.14368
500		0.14229

4.5 Sensitivity of the Procedure

Finally, we need to assess how reliable our Bayesian P-value is in detecting clusters. In particular, we need to check the behavior of the posterior probabilities in the most extreme cases, that is, when there are no clusters (*i.e.* the null hypothesis is true) and when there are at least two clusters in the data.

The simulations of the null distribution indicate that even when the null hypothesis holds, there is still a fair chance of detecting clusters which reduces when we incorporate the minimum cluster size as a parameter. Hence, we need to check the performance of our Bayesian P-value when analyzing data sets with no clusters.

To this end, we generated several data sets, each one from a single Multivariate Normal distribution. Then, all the observations within a data set have the same mean and variance-covariance matrix, and consequently, they form a unique cluster.

In Figure 4-3 we show the scatter-plot corresponding to 50 observations from a bivariate Normal distribution with mean $\boldsymbol{\mu} = (1, -1)'$ and variance-covariance matrix $\Sigma = \text{diag}(1/4, 1/4)$.

The posterior probabilities were obtained after 500000 iterations considering a MCS of 20% of the observations. The results for k =2, 3 and 4 clusters are listed in Table 4-5

Table 4-5. Posterior probabilities after 500000 iterations for the observations in fig. 4-3. The MCS is 20% of the observations.

κ	$P(H_0)$
2	0.59895
3	0.72213
4	0.91140

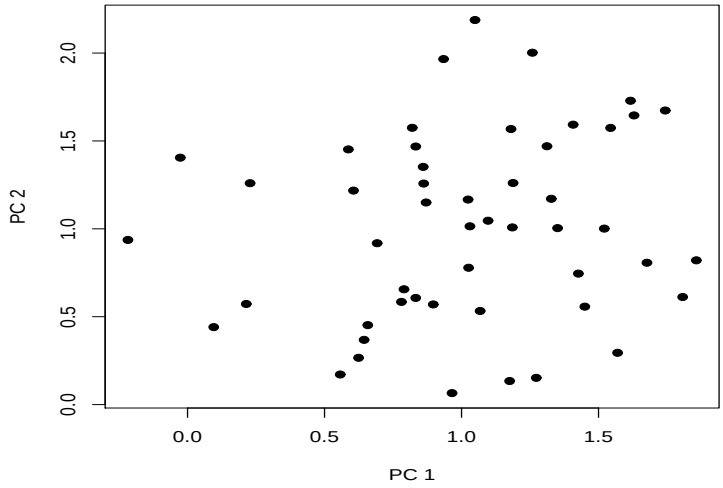

Figure 4-3. Scatter-plot of 50 observations generated from a bivariate Normal distribution with mean $\boldsymbol{\mu} = (1, -1)'$ and variance-covariance matrix $\Sigma = \text{diag}(1/4, 1/4)$

We observe that the posterior probabilities are fairly high in all the cases, showing very weak evidence for the presence of cluster. Notice that the smaller value is obtained for testing two clusters, but still the posterior probability is too high to be considered significant according to our calibrations. Other simulations agree with this results.

The other case we need to consider is when we have at least two clusters. We generate data sets composed from observations coming from multivariate Normal distributions with different means, depending on the number of clusters we want to test. This way, we purposely created clusters in the data sets to be tested.

In Figure 4-4 we show the scatter-plot corresponding to 50 observations. Of them, 35 were generated from a bivariate Normal distribution with mean $\boldsymbol{\mu} = (-1, 2)'$ and variance-covariance matrix $\Sigma = \text{diag}(1/2, 1/2)$, and the remaining 15 were generated from a bivariate Normal distribution with mean $\boldsymbol{\mu} = (3, -1)'$ 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 500000 iterations are listed in Table 4-6. The MCS is 20% of the total number of observations.

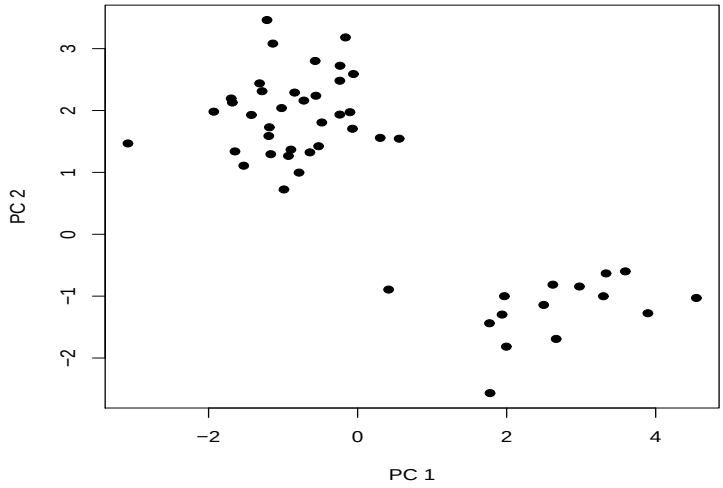

Figure 4-4. Scatter-plot of 50 observations generated from two bivariate Normal distributions with different means.

Table 4-6. Posterior probabilities after 500000 iterations for the observations in fig. 4-4. The MCS is 20% of the observations.

κ	$P(H_0)$
2	2.04E-05
3	3.14E-05
4	0.00669

The posterior probabilities obtained for the data are very small, indicating strong evidence against the null hypothesis in all the test. 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. Nevertheless, the strongest evidence points to the existence of two clusters which is the number. Similar results are observed in other simulations indicating that our procedure for detecting clusters is fairly accurate.

CHAPTER 5
ANALYSIS OF THE NIR DATA

As we briefly explained in the introduction, the data comes from the following experiment: from each ear of maize, a sample of 60 to 100 kernels is taken. For each kernel a NIR spectrum is collected and the absorption for about 800 wavelengths is recorded. Then, in order to simplify the structure of the data and facilitate study of it, a principal components analysis is carried out to obtain the directions that capture most of the variability in the data.

Next, the number of principal components considered in the analysis is determined so that they explain most of the variability in the data with the least number of vectors. In the context of this experiment, generally the first two principal components capture over 95% of the variability of the data, which is satisfactory for the researcher. This way, the NIR spectroscopy output is compressed to 60-100 vectors of dimension 2.

Figure 5-1 shows the scatter-plots for the first 2 principal components for 4 ears of maize, obtained using the methodology described above.

Finally, we apply our procedure and compute the posterior probabilities to the treated data. We test for $k = 2$, 3 and 4 clusters, which are the interpretable number of clusters in the context of the experiment.

Table 5-1 shows the posterior probabilities for the data in Figure 5-1. The values were obtained after 500000 iterations. Also, the minimum cluster size is set to 20% of the total number of observations. This quantity defines a meaningful cluster size for the experimenter and prevent the test for finding clusters due to extremes observations.

Table 5-1. Test of the hypotheses H_0: no clusters vs. H_1: k clusters, with minimum cluster size 20% of the total number of observations.

Label	n	$P(H_0\|\kappa = 2)$	$P(H_0\|\kappa = 3)$	$P(H_0\|\kappa = 4)$
i00F-0173-01	96	0.009098	0.840622	0.976126
i00F-0183-01	96	0.877506	0.980882	0.999744
i01S-0026-18	96	0.000266	0.392694	0.044762
i02S-0302-16	96	0.499539	0.972905	0.997341

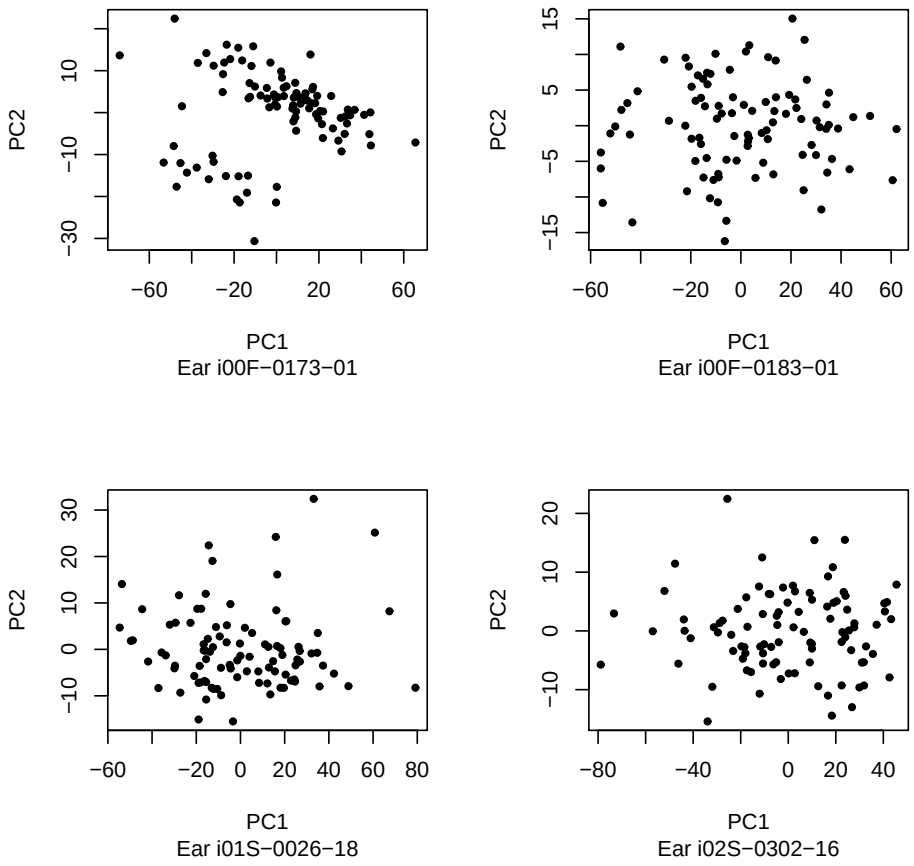

Figure 5-1. Scatter-plots for the first two principal components of the NIR spectra for labels i00F-0173-01, i00F-0183-01, i01S-0026-18 and i02S-0302-16

In Figure 5-2 we show the convergence of the procedure for ear i00F-0173-01.

When comparing these results with the values in Table 4-3 we see that for the data set i00F-0173-01 we have strong evidence for the existence of 2 clusters but not for 3 or 4. This conclusion is clearly supported by the respective scatter-plot.

For ear i00F-0183-01 we do not find evidence of the existence of clusters and we conclude that the null hypothesis is true. The conclusion is also fairly obvious by visual examination of the scatter-plot. The same conclusion holds for ear i02S-0302-16, although is not immediately obvious by inspection of the scatter-plot. In this case, the smallest Bayesian P-value is 0.499539 which is above of the corresponding cutoff point at level 0.05.

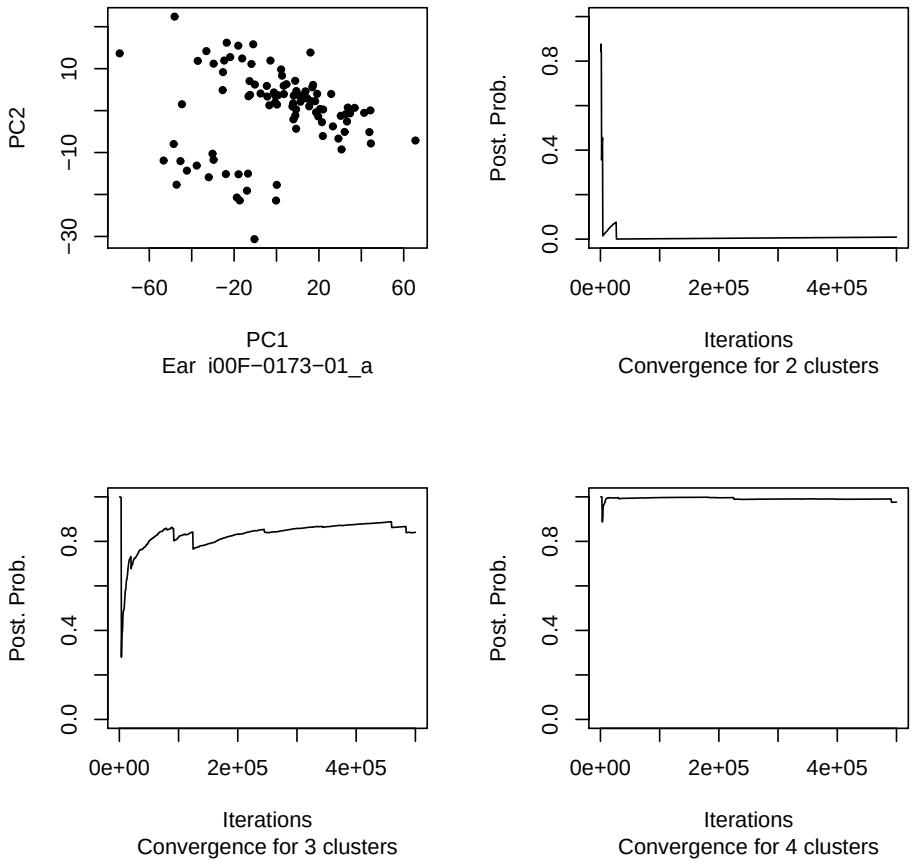

Figure 5-2. Convergence of the posterior probabilities for testing $k = 2$, 3 and 4 clusters.

In the data from ear i01S-0026-18, we find strong evidence for the presence of clusters in all the tests. The greater value among the posterior probabilities is 0.392694 corresponding to the test for 3 clusters. This value is still below the cutoff point at level 0.05 for the corresponding test and therefore provide strong evidence for the existence of 3 clusters in the data. On the other hand, we notice that the minimum posterior probability is reached in the test for 2 clusters. However these numbers should not be compared directly because of the differences among their respective null distribution of the posterior probabilities.

Visual inspection of the scatter-plot also suggests the existence of clusters, although it is not clear how many of them there are in the data.

While in this case is difficult to decide about the number of clusters, the data set clearly indicate the existence of them and demands special attention from the researcher, which is the main objective.

Finally, we point out that because of the multiple tests conducted in the context of the investigation, it is of the interest of the researcher to perform a False Discovery Rate analysis to determine the most significant tests. Our procedure can be easily extended to this type of analysis by increasing the number of simulations of the posterior probabilities. Sometimes the number of tests to be simultaneously perform are nearly 900, so in order to obtain the FDR α-percentiles we run 100000 simulations to obtain the posterior probabilities. In figure 5-3 we can see the histograms for $n = 65$ and $k = 2$ and 3 based on 100000 constructed with this purpose.

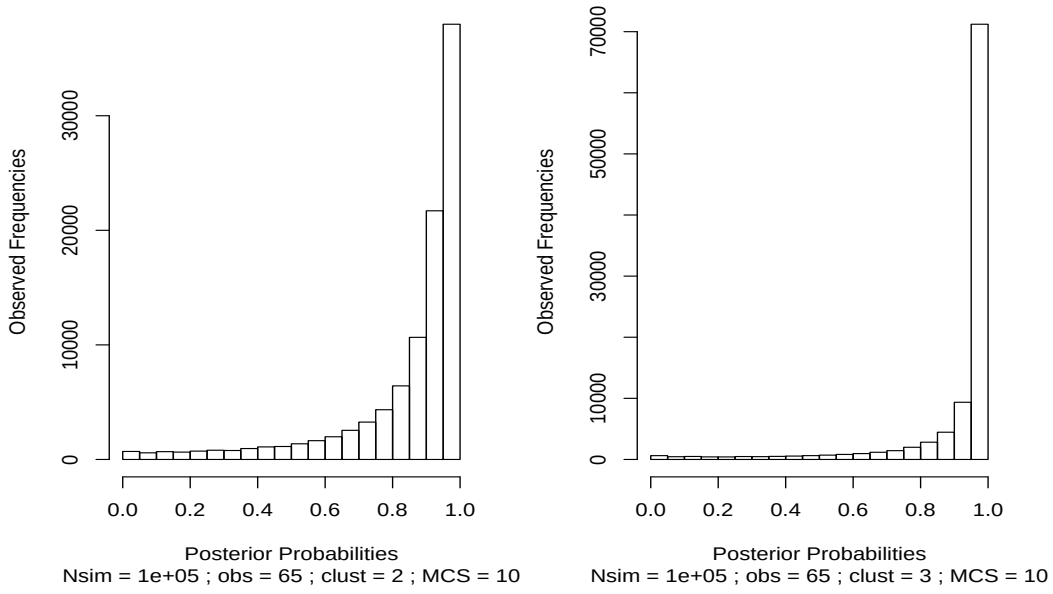

Figure 5-3. Histograms of the null posterior probabilities for $n = 65$ and $k = 2$ and 3 clusters based on 100000 simulations. The minimum cluster size is set equal to 15% of the observations in each case.

CHAPTER 6
CONCLUSIONS

We have proposed a method for testing for clusters based on Bayesian model selection. Although our method does not test directly the more general hypothesis H_0 : No clusters *vs.* H_1 : At least two clusters it provides an accurate notion of the cluster structure of the data and facilitates the decision making when evidence for the existence of clusters is needed.

Also, our method proved too be very sensitive in the detection of clusters isolating extreme observations or outliers as their own clusters. This situation can mislead the conclusions and therefore special attention have to be put in the parameters κ (the number of clusters) and MCS (the minimum cluster size) when performing the test.

While we developed a testing procedure and a calibration method to determine the strength of the evidence when detecting clusters, we have not identified the clusters. Some simulations suggest that, under our procedure, two observations will not necessarily be declared to be in different clusters, even if they seem to be far apart. Hence, our procedure is not taking into account any distance (in the metric sense) to look for evidence for clusters, but only take advantage of the probabilistic model defined for the observations and the partition space. This feature is particularly interesting, because some observations may fall far away from the mean of their respective "true" cluster only by chance. Such observations will be declared to be in the wrong clusters under any distance based method.

Some modifications to the algorithms allow a stochastic search in the space of partitions to find the optimal partition that determine the clusters for a fixed $\kappa = k$. Such modifications and methods to identify the optimal clusters are currently under research.

We finally observe that results from tests for different number of clusters should not be compared directly, because the values for the corresponding Bayesian P-values correspond to observations coming from different distributions and therefore, different probability spaces. Hence, another question for future research is how to compare the

results of tests for different number of clusters. Our methodology compare model pairwise and the apparent lack of coherence of the Bayes factors (Shervish, 1999) make this an interesting and difficult problem, particularly when determining the correct number of clusters present in the data is of interest.

APPENDIX A
ON THE PROBLEM OF PARTITIONS

A.1 Generating a Random Partition

Establishing (2–11) 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 2.5 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

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

which is $\mathcal{S}_{8,4}$, 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 $\mathcal{R}$, which counts redundancies. For a partition $n_1, n_2, \ldots, n_k$, with $\sum_j n_j = n$, define

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

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

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

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

$$\frac{k!}{\mathcal{R}(n_1, n_2, \ldots, n_k)} \times \frac{1}{\binom{n-1}{k-1}}. \tag{A–2}$$

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

$$\sum_{\substack{n_1+\ldots+n_k=n \\ n_1 \le n_2 \le \ldots \le n_k}} \frac{1}{\mathcal{R}(n_1, n_2, \ldots, n_k)} = \frac{1}{k!}\binom{n-1}{k-1}. \tag{A–3}$$

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

$$\frac{\binom{n}{n_1\ n_2\ \ldots\ n_k}}{\mathcal{R}(n_1, n_2, \ldots, n_k)}. \tag{A–4}$$

Multiplying (A–2) and (A–4) results in the probability of a partition ω being given by,

$$g(\omega) = \frac{k!}{\binom{n-1}{k-1}\binom{n}{n_1\ n_2\ \ldots\ n_k}}. \tag{A–5}$$

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

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

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

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

which follows from (A–4). Cancelling terms and applying (A–3) shows that $\sum_{\omega\in\mathcal{P}_k} g(\omega) = 1$.

As a final note, it follows from (A–1) and (A–4) that another representation of the Stirling Number of the Second Kind is.

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

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 2.5.

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 (A–5) will show that under the present generation scheme, the probability of a partition with at least m objects in each cluster is

$$g_m(\omega) = \frac{k!}{\binom{n-mk+k-1}{k-1}\binom{n}{n_1\ n_2\ \ldots\ n_k}}, \tag{A–6}$$

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 more 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 can not 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

$$p(n,k) = p(n-1,k-1) + p(n-k,k). \tag{A–7}$$

To complete the specification of the recursion, we define

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

The recursive relation (A–7) can be easily computed, for any (n,k), using R or another programming language. Table A-1 shows the number of partitions $p(n,k)$ for $n = 1,\ldots,6$.

Table A-1. Number of partitions $p(n,k)$ for $n = 1,\ldots,6$.

		k					
		1	2	3	4	5	6
	1	1					
	2	1	1				
	3	1	1	1			
n	4	1	2	1	1		
	5	1	2	2	1	1	
	6	1	3	3	2	1	1

The topic of integer partitions has been extensively researched from combinatorial, number theoretical and analytic aspects. For further references and results see Andrews, 1976.

APPENDIX B
DERIVATION OF THE MARGINAL DISTRIBUTION

Under the model formulation of Section 2.3, 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

$$\begin{aligned}
&\prod_{j=1}^{k}\prod_{\ell=1}^{n_j} N(Y_\ell^{(j)}|\boldsymbol{\mu}_j,\Sigma_j)N(\boldsymbol{\mu}|\boldsymbol{\mu}_0^{(j)},\tau^2\Sigma_j) \qquad \text{(B–1)}\\
&\quad\propto\quad \exp\left\{-\frac{1}{2}\left[\sum_{\ell=1}^{n_j}(Y_\ell^{(j)}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(Y_\ell^{(j)}-\boldsymbol{\mu}_j)+(\boldsymbol{\mu}_j-\boldsymbol{\mu}_0^{(j)})'[\tau^2\Sigma_j]^{-1}(\boldsymbol{\mu}_j-\boldsymbol{\mu}_0^{(j)})\right]\right\}.
\end{aligned}$$

Now,

$$\begin{aligned}
&\sum_{\ell=1}^{n_j}(Y_\ell^{(j)}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(Y_\ell^{(j)}-\boldsymbol{\mu}_j)+(\boldsymbol{\mu}_j-\boldsymbol{\mu}_0^{(j)})'[\tau^2\Sigma_j]^{-1}(\boldsymbol{\mu}_j-\boldsymbol{\mu}_0^{(j)})\\
&\quad=\quad \sum_{\ell=1}^{n_j}(Y_\ell^{(j)}-\bar{Y}^{(j)})'\Sigma_j^{-1}(Y_\ell^{(j)}-\bar{Y}^{(j)})+n_j(\bar{Y}^{(j)}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\bar{Y}^{(j)}-\boldsymbol{\mu}_j)\\
&\qquad +(\boldsymbol{\mu}_j-\boldsymbol{\mu}_0^{(j)})'\frac{1}{\tau^2}\Sigma_j^{-1}(\boldsymbol{\mu}_j-\boldsymbol{\mu}_0^{(j)})
\end{aligned}$$

and completing squares

$$\begin{aligned}
&n_j(\bar{Y}^{(j)}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\bar{Y}^{(j)}-\boldsymbol{\mu}_j)+(\boldsymbol{\mu}_j-\boldsymbol{\mu}_0^{(j)})'\frac{1}{\tau^2}\Sigma_j^{-1}(\boldsymbol{\mu}_j-\boldsymbol{\mu}_0^{(j)})\\
&\quad=\quad \frac{n_j\tau^2+1}{\tau^2}\left(\boldsymbol{\mu}_j-\delta(\bar{Y}^{(j)})\right)'\Sigma^{-1}\left(\boldsymbol{\mu}_j-\delta(\bar{Y}^{(j)})\right)+\frac{n_j}{n_j\tau^2+1}(\bar{Y}^{(j)}-\boldsymbol{\mu}_0^{(j)})'\Sigma^{-1}(\bar{Y}^{(j)}-\boldsymbol{\mu}_0^{(j)}),
\end{aligned}$$

where

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

Then, the double product in (B–1) is proportional to

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

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

$$\begin{aligned}&\int\prod_{\ell=1}^{n_j}N(Y_\ell^{(j)}\,|\boldsymbol{\mu}_j,\Sigma_j)N(\boldsymbol{\mu}|\boldsymbol{\mu}_0^{(j)},\tau^2\Sigma_j)\,d\boldsymbol{\mu}_j\\ &= \left(\frac{1}{2\pi}\right)^{pn_j/2}\frac{1}{|\Sigma|^{n_j/2}}\left(\frac{2\pi}{n_j\tau^2+1}\right)^{p/2}\\ &\quad\times\exp\left\{-\frac{1}{2}\left(\sum_{\ell=1}^{n_j}(Y_\ell^{(j)}-\bar{Y}^{(j)})'\Sigma^{-1}((Y_\ell^{(j)}-\bar{Y}^{(j)}))+\frac{n_j}{n_j\tau^2+1}(\bar{Y}^{(j)}-\boldsymbol{\mu}_0^{(j)})'\Sigma^{-1}(\bar{Y}^{(j)}-\boldsymbol{\mu}_0^{(j)})\right)\right\}\\ &\quad\times\pi(\Sigma_j).\end{aligned} \tag{B–2}$$

Under the assumption $\Sigma_j=\mathrm{diag}(\sigma_{1j}^2,\ldots,\sigma_{pj}^2)$ and considering the priors

$$\sigma_{rj}^2\sim IG(a,b)=\frac{1}{\Gamma(a)b^a}\frac{1}{(\sigma_{rj}^2)^{a+1}}e^{-1/b\sigma_{rj}^2}$$

the expression in braces in (B–2) simplifies to

$$\sum_{r=1}^{p}\frac{-1}{2\sigma_{rj}^2}\left(n_js_{rj}^2+\frac{n_j}{n_j\tau^2+1}(\bar{y}_r^{(j)}-\mu_{0r}^{(j)})^2\right),$$

where

$$\bar{y}_r^{(j)}=\frac{1}{n_j}\sum_{\ell=1}^{n_j}y_{\ell r}^{(j)}\quad,\quad s_{rj}^2=\frac{1}{n_j}\sum_{\ell=1}^{n_j}(y_{\ell r}^{(j)}-\bar{y}_r^{(j)})^2$$

for $r=1,\ldots,p$ and $j=1,\ldots,k$.

It follows

$$\begin{aligned}m(\mathbf{Y}|\,\omega) &= \left(\frac{1}{2\pi}\right)^{np/2}\left(\frac{1}{\Gamma(a)b^a}\right)^{pk}\prod_{j=1}^{k}\left(\frac{1}{n_j\tau^2+1}\right)^{p/2}\\ &\quad\prod_{r=1}^{p}\int\frac{1}{(\sigma_{rj}^2)^{n_j/2+a+1}}\exp\left\{-\frac{1}{2\sigma_{rj}^2}\left(n_js_{rj}^2+\frac{n_j}{n_j\tau^2+1}(\bar{y}_r^{(j)}-\mu_{0r}^{(j)})^2+\frac{2}{b}\right)\right\}d\sigma_{rj}^2\end{aligned}$$

Finally, noticing that in the integral we have the kernel of a distribution

$$IG\left(\frac{n_j}{2}+a,\frac{2}{\left(n_j s_{rj}^2+\frac{n_j}{n_j\tau^2+1}(\bar{y}_r^{(j)}-\mu_{0r}^{(j)})^2+\frac{2}{b}\right)}\right)$$

we obtain

$$\begin{aligned} m(\mathbf{Y}|\,\omega_k) &= \left(\frac{2}{b}\right)^{pka}\frac{1}{\pi^{np/2}\Gamma(a)^{pk}} \\ &\quad\times\left[\prod_{j=1}^{k}\frac{\Gamma\left(\frac{n_j}{2}+a\right)^p}{(n_j\tau^2+1)^{p/2}}\right]\left[\prod_{j=1}^{k}\prod_{r=1}^{p}\frac{1}{\left(n_j s_{rj}^2+n_j\frac{(\bar{y}_r^j-\mu_{0r}^j)^2}{n_j\tau^2+1}+\frac{2}{b}\right)^{n_j/2+a}}\right]. \end{aligned}$$

APPENDIX C
BAYES FACTORS AND HYPOTHESIS TESTING

Suppose we want to test the hypotheses

$$H_0 : \theta \in \Theta_0 \ \textit{vs.}\ H_1 : \theta \in \Theta_1.$$

We may consider the following Bayesian setting. Let $\pi_0 = P(\theta \in \Theta_0)$ and $\pi_1 = P(\theta \in \Theta_1) = 1 - \pi_0$ be priors probabilities on H_0 and H_1. Also, let $q_0(\theta)$ and $q_1(\theta)$ be two probability density functions on Θ_0 and Θ_1 respectively. Then, the prior for θ is given by

$$\pi(\theta) = \pi_0 q_0(\theta) I(\theta \in \Theta_0) + \pi_1 q_1(\theta) I(\theta \in \Theta_1)$$

Then, if $f(x|\theta)$ is a pdf, the marginal distribution of the data is

$$\begin{aligned} m(x) &= \int f(x|\theta)\pi(\theta)d\theta \\ &= \pi_0 \int_{\Theta_0} f(x|\theta)q_0(\theta)d\theta + \pi_1 \int_{\Theta_1} f(x|\theta)q_1(\theta)d\theta \end{aligned}$$

and we obtain

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)} = \frac{\pi_0 f(x|\theta)q_0(\theta)I(\theta \in \Theta_0) + \pi_1 f(x|\theta)q_1(\theta)I(\theta \in \Theta_1)}{m(x)}.$$

It follows that the posterior odds are

$$\frac{P(\Theta_0|X)}{P(\Theta_1|X)} = \frac{\pi_0 \int_{\Theta_0} f(x|\theta)q_0(\theta)d\theta}{\pi_1 \int_{\Theta_1} f(x|\theta)q_1(\theta)d\theta}.$$

This quantity can be used as a measurement of the evidence against the null contained in the data.

We can also define the Bayes factor of H_0 relative to H_1 to be the ratio of the posterior probabilities of the null and alternative hypotheses over the ratio of the prior

probabilities of the null and alternative hypotheses, *i.e.*,

$$\begin{aligned} BF_{01} &= \frac{P(H_0|X)}{P(H_1|X)} \Big/ \frac{\pi_0}{\pi_1} \\ &= \frac{\int_{\Theta_0} f(x|\theta) q_0(\theta) d\theta}{\int_{\Theta_1} f(x|\theta) q_1(\theta) d\theta} \end{aligned}$$

Notice that:

- Equivalently we may define the Bayes factor of H_1 relative to H_0 and we obtain $BF_{10} = 1/BF_{01}$.
- Unlike the posterior odds, value of the Bayes factor does not depend on the values of π_0 and π_1.
- For the simple null vs. simple alternative case, *i.e.*, $H_0 : \theta = \theta_0$ *vs.* $H_1 : \theta = \theta_1$, we obtain

  $$BF_{01} = \frac{f(x|\theta_0)}{f(x|\theta_1)},$$

  the likelihood ratio.

For a complete review on Bayes factors and their use see Robert, 2001.

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.

Baye, T. M., Pearson, T. C. and Settles, A. M. (2006), "Development of a Calibration to Predict Maize Seed Composition Using Single Kernel Near-Infrared Spectroscopy," *Journal of Cereal Science*, 43, 236–243.

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*.

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.

Dudley, J. W., Lambert, R. J. and Alexander, D. E. (1974), "Seventy Generations of Selection for Oil and Protein Concetration in the Maize Kernel," *Seventy Geneations of Selection for Oil and Protein in Maize*, J. W. Dudley, Ed (Madison, W. I.: Crops Society of America), 181–212.

Easton, G. S. and Rochetti, R. (1986), "General Saddlepoint Approximations with Applications to L Statistics," *Journal of the American Statistical Association*, 81, 420–423.

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

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," *New York: Wiley*.

Janni, J., Weinstock, B. A., Hagen, L. and Wright, S. (2008), "Novel Near-Infrared Sampling Apparatus for Single Kernel Analysis of Oil Content in Maize," *Applied Spectroscopy*, 62, 423–426.

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

Jorgensen, B. (1993), "The Theory of Linear Models," *Chapman & Hall*, New York-London.

Kass, R. E. and Raftery, A. E. (1995), "Bayes Factor and Model Uncertainty," *Journal of the American Statistical Association*, 90, 773–795.

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.

Lambert, R. J., Alexander, D. E. and Han, Z. J. (1998), "A High Oil Pollinator Enhacement of Kernel Oil and Effects on Grain Yields of Maize Hybrids," *Agronomy Journal*, 90, 211–215.

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.*

Orman, B. A. and Schumann, R. A. (1991), "Comparison of Near-Infrared Spectroscopy Calibration Methods for the Prediction of Protein, Oil, and Starch in Maize Grain," *Journal of Agriculture and Food Chemistry*, 39, 883–886.

Pan, D. (2000), "Starch Synthesis in Maize," *Carbohydrate Reserves in Plants: Synthesis and Regulation*, A.K Gupta and N. Kaur, Eds (Amsterdam:Elsevier), 125–146.

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.

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.

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).

Weinstock, B. A., Janni, J., Hagen, L. and Wright, S. (2006), "Prediction of Oil and Oleic Acid Concentrations in Individual Corn (Zea mays L.) Kernels Using Near-Infrared Reflectance Hyperspectral Imaging and Multivariate Analysis," *Applied Spectroscopy*, 60, 9–16.

Williams, P. and Norris, K. (2001), "Near-Infrared Technology in the Agricultual and Food Industries, 2nd Edition," *American Association of Cereal Chemists, Inc*, Minn., USA.

BIOGRAPHICAL SKETCH

Claudio Fuentes was born in Santiago, Chile, in 1977. Upon graduation from high school, he enrolled as a student at the Pontificia Universidad Catolica de Chile, where he received the degree of Bachelor of Science in mathematics in 2001, and his Master in Statistics in December 2003. During his undergraduate and graduate education he was appointed as a teaching assistant in both the faculty of physics and the faculty of mathematics.

In August 2005, Fuentes entered the PhD. program in the Department of Statistics at the University of Florida. During his education there, he was appointed as a teaching assistant and graduate teaching instructor. He worked as a research assistant for his advisor, Distinguished Professor Dr. George Casella. In August 2008 he earned the degree of Master of Science in statistics with a thesis on Cluster Analysis. He currently is working toward a PhD. in statistics under Dr. Casella, with interest in inference on selected populations.