ISSN: 1696-2281
SORT 33 (2) July - December (2009)

SORT

Statistics and Operations Research Transactions

Sponsoring institutions

Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Institut d'Estadística de Catalunya

Supporting institution

Spanish Region of the International Biometric Society

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Published on paper
bearing the certificate
of sustainable forest
management

SORT

Volume 33
Number 2
July-December 2009
ISSN: 1696-2281

Invited article *(with discussion)*

Articles

Information for authors and subscribers

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

Testing for the existence of clusters

Claudio Fuentes and George Casella*

University of Florida

Abstract

Detecting and determining clusters present in a certain sample has been an important concern, among researchers from different fields, for a long time. In particular, assessing whether the clusters are statistically significant, is a question that has been asked by a number of experimenters. Recently, this question arose again in a study in maize genetics, where determining the significance of clusters is crucial as a primary step in the identification of a genome-wide collection of mutants that may affect the kernel composition.
Although several efforts have been made in this direction, not much has been done with the aim of developing an actual hypothesis test in order to assess the significance of clusters. In this paper, we propose a new methodology that allows the examination of the hypothesis test $\boldsymbol{H}_0 : \kappa = 1$ ***vs.*** $\boldsymbol{H}_1 : \kappa = \boldsymbol{k}$, where κ denotes the number of clusters present in a certain population. Our procedure, based on Bayesian tools, permits us to obtain closed form expressions for the posterior probabilities corresponding to the null hypothesis. From here, we calibrate our results by estimating the frequentist null distribution of the posterior probabilities in order to obtain the ***p***-values associated with the observed posterior probabilities. In most cases, actual evaluation of the posterior probabilities is computationally intensive and several algorithms have been discussed in the literature. Here, we propose a simple estimation procedure, based on MCMC techniques, that permits an efficient and easily implementable evaluation of the test. Finally, we present simulation studies that support our conclusions, and we apply our method to the analysis of NIR spectroscopy data coming from the genetic study that motivated this work.

MSC: 62F15, 62F03.

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

1 Introduction

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

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

Received: May 2009

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

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

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

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

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

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

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

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

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

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

2 Testing for clusters

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

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

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

and

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

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

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

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

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

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

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

2.1 Bayesian hypothesis testing

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.2 Model and distribution assumptions

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

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

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

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

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

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

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

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

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

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

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

2.3 Prior on the partitions

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

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

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

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

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

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

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

2.4 Estimation of the Bayes factor

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

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

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

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

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

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

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

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

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

The final Metropolis-Hastings algorithm is:

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

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

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

Notice that the described algorithm has stationary distribution g.

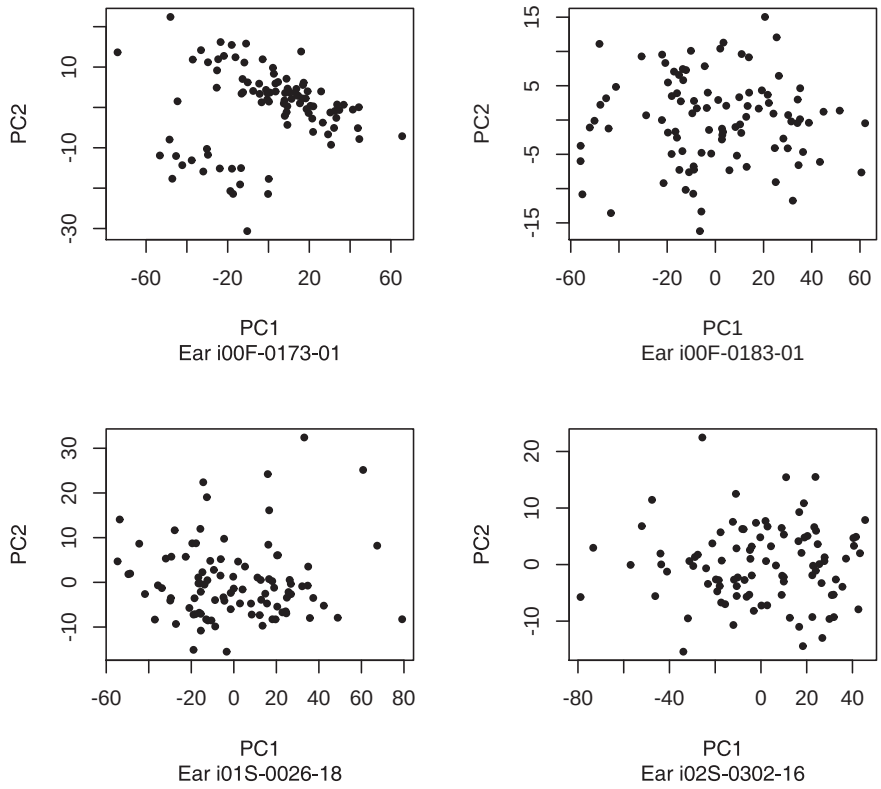


Figure 1: *Scatter-plots for the first two principal components of the NIR spectra for data sets with labels i00F-0173-01, i00F-0183-01, i01S-0026-18 and i02S-0302-16.*

3 Application to NIR spectroscopy data

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

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

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

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

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

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

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

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

4 Frequentist calibration

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

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

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

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

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

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

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

and

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

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

4.1 Bayes factor under the null distribution

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where

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

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

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

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

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

4.2 Estimation of the null distribution

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

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

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

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

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

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

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

5 Simulation studies

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

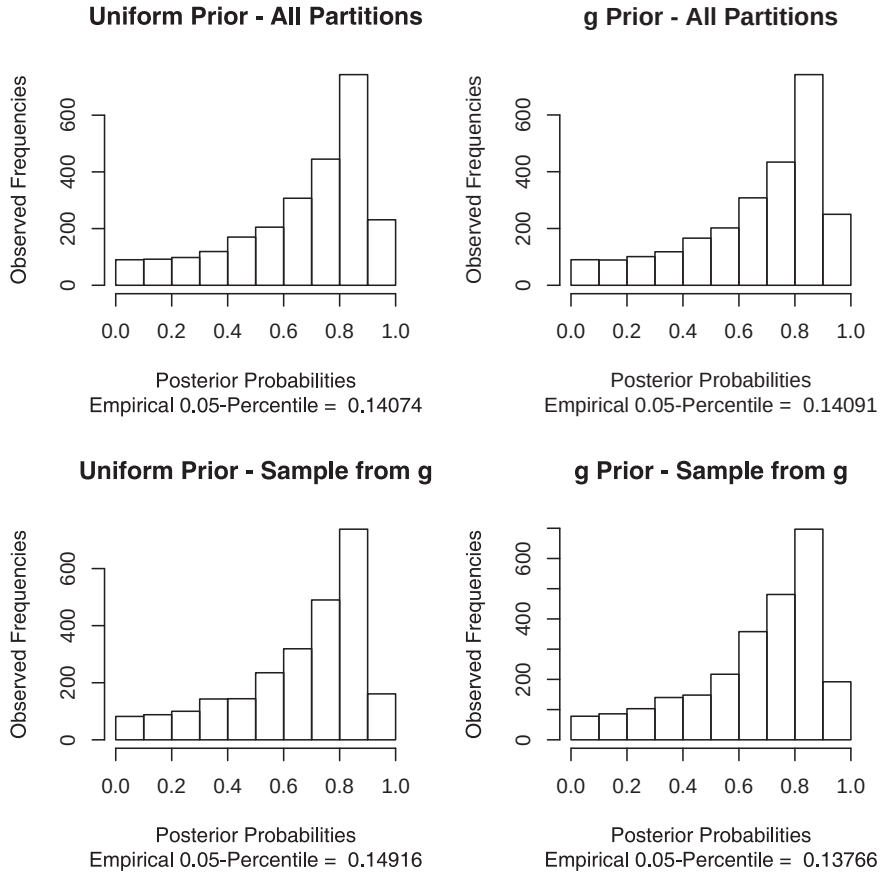


Figure 2: *Histograms of the null posterior probabilities for $n = 50$ and $k = 3$. The top row is the exact calculation, based on enumerating all partitions. The corresponding histograms in the bottom row are based on 2500 simulations of samples of size 52, representing* 25% *of all partitions.*

5.1 Goodness of the approximation

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

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

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

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

5.2 Error of approximation

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

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

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

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

5.3 Minimum cluster size

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

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

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

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

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

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

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

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

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

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

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

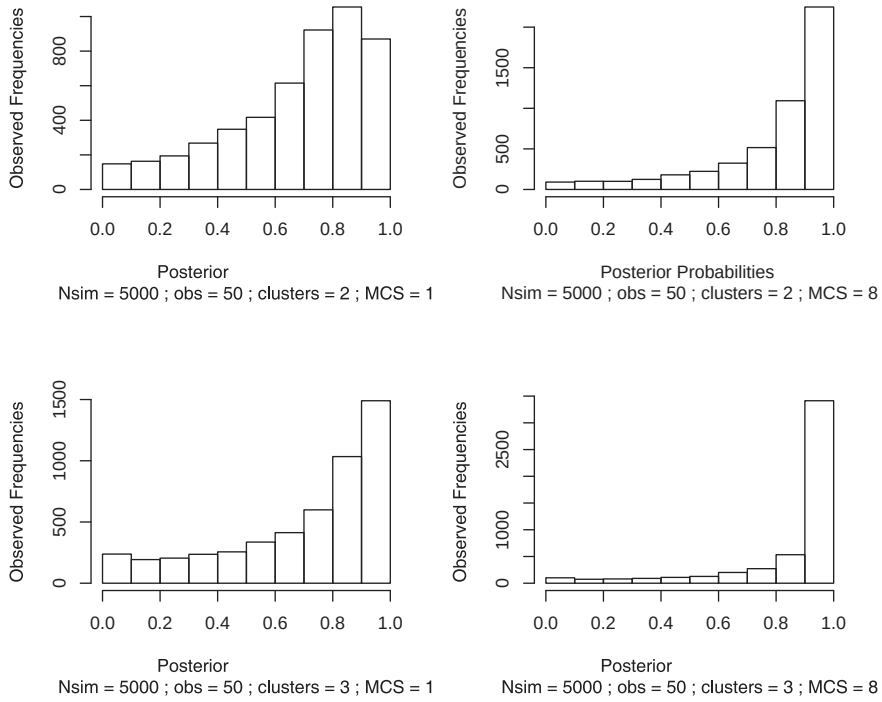


Figure 3: *Histograms of the null posterior probabilities for $n = 50$ and $k = 2$ (top row) or $k = 3$ (bottom row) clusters based on 5000 simulations. The minimum cluster size is set equal to 1 observation (left column) and 15% of the observations (right column).*

5.4 Power of the procedure

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

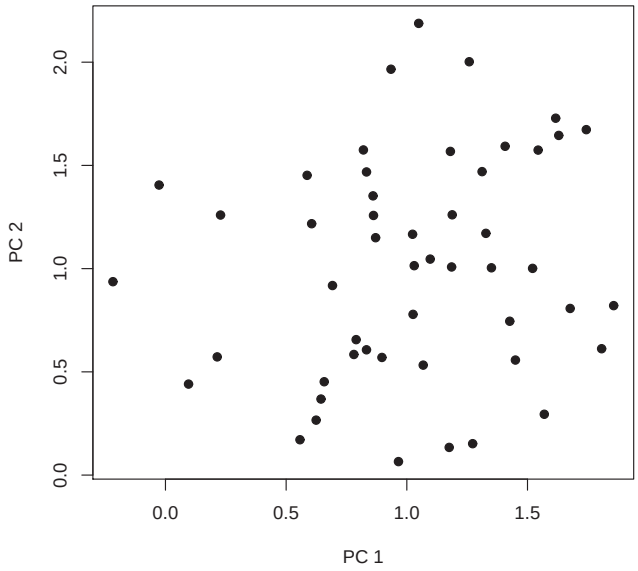


Figure 4: *Scatter-plot of 50 observations generated from a bivariate normal distribution with mean* $\boldsymbol{\mu} = (1,-1)^{\mathsf{T}}$ *and variance-covariance matrix* $\Sigma = diag(1/4, 1/4)$.

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

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

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

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

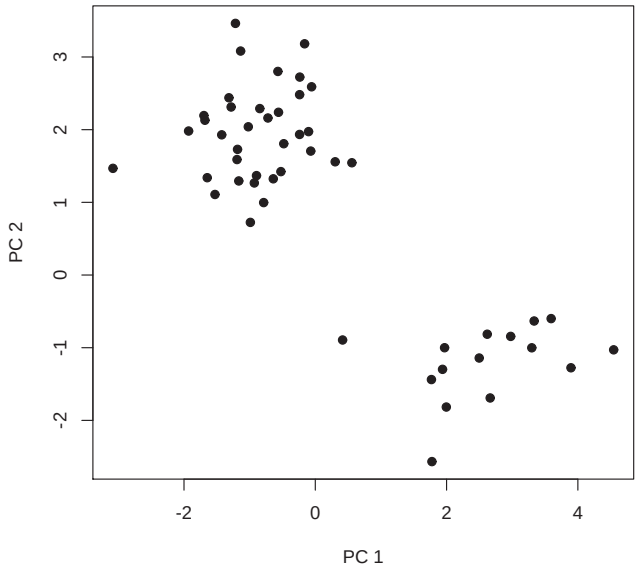


Figure 5: *Scatter-plot of 50 observations generated from two bivariate normal distributions with different means.*

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

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

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

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

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

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

6 Application of the calibration procedure

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

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

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

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

7 Discussion

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

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

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

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

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

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

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

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

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

Acknowledgments

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

References

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

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

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

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

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

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

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

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

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

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

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

Girón, F. J., Martínez, M. L., Moreno, E. and Torres, F. (2006). Objective testing procedures in linear models: calibration of the *p*-values. *Scandinavian Journal of Statistics*, 33, 765-784.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A Generating a random partition

A.1 An example

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

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

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

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

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

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

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

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

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

A.2 Derivation in the general case

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A.3 Partitions of an integer

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

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

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

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

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

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

Thus, the recursive relation is given by

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

To complete the specification we define

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

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

B Proofs of results in Section 4.1

B.1 Proof of Proposition 1

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

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

where

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

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

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

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

B.2 Proof of Lemma 1

Proof. We have

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

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

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

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

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

C Derivation of the marginal distribution

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

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

First, observe that

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

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

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

where

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

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

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

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

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

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

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

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

M. J. Bayarri
Universitat de València, Spain

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

References in the discussion

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

Acknowledgements

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

Adolfo Álvarez and Daniel Peña
Universidad Carlos III de Madrid

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

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

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

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

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

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

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

Rejoinder

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

Outliers and minimum cluster size

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

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

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

Model comparison and calibration

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

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

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

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

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

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

Final comments

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

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

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

SORT 33 (2) July-December 2009, 159-170

Nonparametric estimation of the expected accumulated reward for semi-Markov chains

Guglielmo D'Amico

Abstract

In this paper a nonparametric estimator of the expected value of a discounted semi-Markov reward chain is proposed. Its asymptotic properties are established and as a consequence of the asymptotic normality the confidence sets are obtained. An application in quality of life modelling is described.

MSC: 60K15, 62M09, 62G20.

Keywords: Discrete semi-Markov processes, empirical estimators, asymptotic properties, quality of life.

1 Introduction

Homogeneous semi-Markov chains (HSMC) have been recognized as a flexible and efficient tool in the modelling of stochastic systems. Recent results and applications are retrievable in Barbu, Boussemart and Limnios (2004) and Janssen and Manca (2007).

The idea to link rewards to the occupancy of a semi-Markov state led to the construction of semi-Markov reward processes. These processes have been analyzed and applied by many authors; see Howard (1971), De Dominicis and Manca (1986), Limnios and Oprişan (2001), Khorshidian and Soltani (2002), Janssen and Manca (2006), Stenberg, Manca and Silvestrov (2006, 2007) and Janssen and Manca (2007).

Corresponding author: Guglielmo D'Amico. Drug Sciences Department, "G.D'Annunzio" University, Chieti (Italy). via dei Vestini 30, 66013 Chieti, Italy. g.damico@unich.it
Received: November 2008
Accepted: October 2009

The inferential problems related to reward processes are seldom considered. Gardiner, Luo, Bradley, Sirbu and Given (2006) considered an estimator of the expected accumulated reward for non-homogeneous Markov reward processes with deterministic reward functions. D'Amico (2009) proposed Markov reward processes, with stochastic rate and impulse rewards, to study accumulated measure of the quality of life. In that paper the asymptotic properties of the nonparametric estimator of the higher order moments of the reward process have been established.

In this paper we face the nonparametric inference problems related to a discrete time semi-Markov reward process. We define an estimator of the expected accumulated reward and we prove that it is uniformly strongly consistent, and if properly centralized and normalized that it converges in distribution to a normal random variable. The goal is achieved developing the techniques of estimation for HSMC presented in Barbu and Limnios (2006).

The paper is divided in this way: first, the semi-Markov reward model is briefly depicted and the definition of the functional to which we are interested is given. Next, the asymptotic properties of the nonparametric estimator of the expected accumulated reward process are assessed. Finally, the practical usefulness of the results is shown by exposing a possible application to measure the quality of life.

2 The semi-Markov reward model

Homogeneous semi-Markov chains are a generalization of discrete time Markov chains allowing the times between transitions to occur at random times distributed according to any kind of distribution function which may depend on the current and the next state.

Let us consider a finite set of states $E = \{1, 2, \ldots, S\}$ in which the system can be into and a complete probability space (Ω, F, P) on which we define the following random variables:

$$X_n : \Omega \to E, \;\; T_n : \Omega \to \mathbb{N}. \tag{2.1}$$

They denote the state occupied at the *n*-th transition and the time of the *n*-th transition respectively.

Suppose that the process $(X_n, T_n)_{n\in\mathbb{N}}$ is a discrete time homogeneous Markov renewal process of kernel $\mathbf{q} = (q_{ij}(t))$; see Barbu *et al.* (2004). Elements of the kernel represent the following probabilities

$$q_{ij}(t) = P[X_{n+1} = j, T_{n+1} - T_n = t | X_n = i]. \tag{2.2}$$

From these quantities it is possible to define

$$Q_{ij}(t) = P[X_{n+1} = j, T_{n+1} - T_n \leq t | X_n = i] = \sum_{\tau=1}^{t} q_{ij}(\tau), \tag{2.3}$$

the probability to join, with next transition, state j within time t given the starting, at time zero, from the state i.

The process $\{X_n\}$ is a Markov chain with state space E and transition probability matrix $\mathbf{P} = \mathbf{Q}(\infty)$. We shall refer to it as the embedded Markov chain.

The unconditional waiting time distribution function in state i is

$$H_i(t) = P[T_{n+1} - T_n \leq t | X_n = i] = \sum_{j \in E} Q_{ij}(t). \tag{2.4}$$

Now it is possible to define the conditional cumulative distribution functions of the waiting time in each state, given the state subsequently occupied:

$$G_{ij}(t) = P\{T_{n+1} - T_n \leq t | X_n = i, X_{n+1} = j\} = \frac{1}{p_{ij}} \sum_{s=1}^{t} q_{ij}(s) \cdot 1_{\{p_{ij} \neq 0\}} + 1_{\{p_{ij}=0\}} \tag{2.5}$$

Define $\{N(t)\}$ by $N(t) = \sup\{n : T_n \leq t\}$ $\forall t \in \mathbb{N}$. The discrete time process $Z = (Z(t), t \in \mathbb{N})$ defined by $Z(t) = X_{N(t)}$ is a semi-Markov process of kernel $\mathbf{q}$. It represents, for each waiting time, the state occupied by the process X_n.

We define, $\forall i, j \in E$, and $t \in \mathbb{N}$, the semi-Markov transition probabilities:

$$\phi_{ij}(t) = P[X_{N(t)} = j | X_0 = i]. \tag{2.6}$$

They are obtained by solving the system of equations:

$$\phi_{ij}(t) = \delta_{ij}(1 - H_i(t)) + \sum_{k \in E} \sum_{\tau=1}^{t} q_{ik}(\tau) \phi_{kj}(t - \tau). \tag{2.7}$$

Algorithms to solve equations (2.7) are well known, see for example Janssen and Manca (2007).

To introduce a reward structure, we consider the score function $g : E \to \mathbb{R}$. This function assigns a reward (score) $g(j)$ when the process visits state $j \in E$. Define $\{Y(t)\}$ by $Y(t) = \sum_{s=1}^{t} d^s g(Z(s))$. It represents the discounted accumulated semi-Markov reward process. The quantity $d \in [0,1]$ is a discount factor introduced to compare present scores with future scores. The process $Y(t)$ is of interest, for example, in insurance mathematics see e.g. Stenberg *et al.* (2006, 2007) as well as in quality of life measurement – see D'Amico (2009) in which $Z(t)$ is considered to be a finite and ergodic Markov chain and g a stochastic score function.

The expected value of $Y(t)$ is of interest to synthesize the process behaviour. Let us denote $M_i(t) = E[Y(t)|X_0 = i]$. Then it results that

$$\begin{aligned} M_i(t) &= E[\sum_{s=1}^{t} d^s g(Z(s))|X_0 = i] \\ &= \sum_{s=1}^{t} d^s E[g(Z(s))|X_0 = i] = \sum_{s=1}^{t} d^s \sum_{j\in E} g(j)\phi_{ij}(s) \\ &= \sum_{s=1}^{t} \sum_{j\in E} d^s g(j)\phi_{ij}(s). \end{aligned} \tag{2.8}$$

$M_i(t)$ represents the functional we wish to estimate.

3 The estimation of the expected accumulated reward

Let us suppose now that we have a right-censored history of the HSMC until the observation time L:

$$H(L) = \{X_0, T_1, X_1, T_2, X_2, \ldots, T_{N(L)}, X_{N(L)}, u_L\} \tag{3.1}$$

where $N(L) = \max\{n \in \mathbb{N}|T_n \leq L\}$ and $u_L = L - T_{N(L)}$.

Following the line of research in Barbu and Limnios (2006, 2008), to estimate the semi-Markov kernel, we use the empirical estimator:

$$\hat{q}_{ij}(k,L) = \frac{\sum_{n=1}^{N(L)} 1_{\{X_{n-1}=i, X_n=j, T_n - T_{n-1}=k\}}}{\sum_{n=1}^{N(L)} 1_{\{X_n=i\}}} \tag{3.2}$$

To estimate the functional (2.8) we propose the estimator

$$\hat{M}_i(t;L) = \sum_{s=1}^{t} \sum_{j\in E} d^s g(j)\hat{\phi}_{ij}(s;L). \tag{3.3}$$

Estimator $\hat{\phi}_{ij}(s;L)$ is the (i,j)-th element of the transition probability matrix $\hat{\mathbf{\Phi}}$ which satisfies the matrix equation $\hat{\mathbf{\Phi}}(t) = \mathbf{I} - \hat{\mathbf{H}}(t) + \hat{\mathbf{q}} * \hat{\mathbf{\Phi}}(t)$, where $*$ denotes the matrix convolution product – see Barbu and Limnios (2006) for more details.

The following asymptotic property holds true:

Proposition 3.1 *For all $i \in E$ and $\theta \in \mathbb{N}$ the estimator $\hat{M}_i(\theta,L)$ is uniformly strongly consistent, that is*

$$\max_{i\in E}\max_{0\le\theta\le L}|\hat{M}_i(\theta,L)-M_i(\theta)| \xrightarrow{a.s.} 0 \text{ as } L\to\infty. \tag{3.4}$$

Proof. We use the following inequalities:

$$\begin{aligned}
&\max_{i\in E}\max_{0\le\theta\le L}|\hat{M}_i(\theta,L)-M_i(\theta)| = \max_{i\in E}\max_{0\le\theta\le L}\Big|\sum_{s=1}^{\theta}\sum_{j\in E} d^s g(j)\big(\hat{\phi}_{ij}(s,L)-\phi_{ij}(s)\big)\Big| \\
&\le \max_{i\in E}\max_{0\le\theta\le L}\sum_{s=1}^{\theta}\sum_{j\in E} d^s g(j)|\hat{\phi}_{ij}(s,L)-\phi_{ij}(s)| \\
&\le \sum_{s=1}^{\theta}\sum_{j\in E} d^s g(j)\max_{i\in E}\max_{0\le\theta\le L}|\hat{\phi}_{ij}(s,L)-\phi_{ij}(s)|
\end{aligned} \tag{3.5}$$

and this last quantity goes to zero almost surely as a consequence of the uniform strongly consistency of the estimators $\hat{\phi}_{ij}(s,L)$ given in Barbu and Limnios (2006). ■

To prove the asymptotic normality of estimator $\hat{M}_i(t;L)$ we need to introduce the following variables:

$$q_{ij}^{(n)}(t) = P[X_n = j, T_n = t | X_0 = i], \tag{3.6}$$

$$\psi_{ij}(t) \doteq \sum_{n=0}^{t} q_{ij}^{(n)}(t), \tag{3.7}$$

$$\Psi_{ij}(t) \doteq E_i[N_j(t)] = \sum_{n=0}^{t} Q_{ij}^{(n)}(t). \tag{3.8}$$

Finally, with μ_{ii} and μ_{ii}^* we shall denote the mean recurrence time of state i for the Markov renewal process $(X_n,T_n)_{n\in\mathbb{N}}$ and the mean recurrence time of state i for the embedded Markov chain $(X_n)_{n\in\mathbb{N}}$, respectively.

The following theorem describes the asymptotic normality of the estimator $\hat{M}_i(t;L)$.

Theorem 3.2 *For any fixed time $h \in \mathbb{N}$ and state $i \in E$, it results that*

$$\sqrt{L}(\hat{M}_i(h,L)-M_i(h)) \xrightarrow{d} N(0,\sigma^2_{M_i}(h)) \text{ as } L\longrightarrow\infty \tag{3.9}$$

where

$$\sigma^2_{M_i}(h) = \frac{\mu_{ii}^*}{\mu_{ii}}\sum_{m\in E}\frac{\mu_{mm}^2}{\mu_{mm}^*}\Big\{\sum_{r\in E}\sum_{s=1}^{h} d^{2s}\big[C_{imr}-g(m)\Psi_{im}\big]^2 * q_{mr}(s) - \Big(\sum_{s=1}^{h} d^s D_{im}(s)\Big)^2\Big\} \tag{3.10}$$

and

$$C_{imr} = \sum_{j\in E} g(j)[1-H_j] * \psi_{im} * \psi_{rj} \tag{3.11}$$

$$D_{im}(s) = \sum_{l\in E} \big(C_{ml} * q_{ml}(s) - g(m)\psi_{im} * Q_{ml}(s)\big) \tag{3.12}$$

Proof.

$$\sqrt{L}\big(\hat{M}_i(h,L) - M_i(h)\big) = \sqrt{L}\Big(\sum_{s=1}^{h}\sum_{j\in E} d^s g(j)(\hat{\phi}_{ij}(s,L) - \phi_{ij}(s))\Big) \tag{3.13}$$

In Barbu and Limnios (2006) it was proved that $\sqrt{L}(\hat{\phi}_{ij}(k,L) - \phi_{ij}(k))$ has the same asymptotic behaviour as

$$\sqrt{L}\Big\{\sum_{n=1}^{m}\sum_{u=1}^{m}[(1-H_j) * \psi_{in} * \psi_{uj} * \Delta q_{nu}](k) - \sum_{u=1}^{m} \psi_{ij} * \Delta Q_{ju}(k)\Big\}, \tag{3.14}$$

where $\Delta q_{ij}(k) \doteq \hat{q}_{ij}(k,L) - q_{ij}(k)$ and $\Delta Q_{ij}(k) \doteq \hat{Q}_{ij}(k,L) - Q_{ij}(k)$.

Applying this result to our functional we obtain that $\sqrt{L}\big(\hat{M}_i(h,L) - M_i(h)\big)$ has the same asymptotic distribution as

$$\begin{aligned}&\sqrt{L}\sum_{s=1}^{h}\sum_{v\in E}\sum_{l\in E}\Big[\sum_{j\in E} d^s(g(j)(1-H_j) * \psi_{iv} * \psi_{lj}) * \Delta q_{vl}\Big](s)\\ &-\sqrt{L}\sum_{s=1}^{h}\sum_{l\in E}\sum_{j\in E} d^s g(j)\psi_{ij} * \Delta Q_{jl}(\theta)\end{aligned} \tag{3.15}$$

Let us denote $C_{ivl} = \sum_{j\in E} g(j)[1-H_j] * \psi_{iv} * \psi_{lj}$, then by substitution of the kernel estimator (3.2) in formula (3.15) we get

$$\begin{aligned}&= \frac{1}{\sqrt{L}}\sum_{n=1}^{N(L)}\Bigg(\sum_{v\in E}\sum_{l\in E}\frac{L}{N_n(L)}\Big\{\sum_{s=1}^{h}\big[d^s C_{vl} * \big(1_{\{X_{n-1}=v,X_n=l,T_n-T_{n-1}=\cdot\}}\\ &- q_{vl}(\cdot)1_{\{X_{n-1}=v\}}\big)(s) - d^s g(v)\psi_{iv} * \big(1_{\{X_{n-1}=v,X_n=l,T_n-T_{n-1}\le\cdot\}}\\ &- Q_{vl}(\cdot)1_{\{X_{n-1}=v\}}\big)(s)\big]\Big\}\Bigg) = \frac{1}{\sqrt{L}}\sum_{n=1}^{N(L)} f(X_{n-1},X_n,T_n-T_{n-1})\end{aligned} \tag{3.16}$$

where the function $f : E\times E\times\mathbb{N} \longrightarrow \mathbb{R}$ is defined as follows:

$$
\begin{aligned}
f(m,r,z) \equiv & \sum_{v\in E}\sum_{l\in E}\frac{L}{N_n(L)}\Big\{\sum_{s=1}^{h}\Big[d^s C_{ivl} * \big(1_{\{X_{n-1}=v,X_n=l,T_n-T_{n-1}=\cdot\}} - q_{vl}(\cdot) \\
& \cdot 1_{\{X_{n-1}=v\}}\big)(s) - d^s g(v)\psi_{iv} * \big(1_{\{X_{n-1}=v,X_n=l,T_n-T_{n-1}\le\cdot\}} - Q_{vl}(\cdot)1_{\{X_{n-1}=v\}}\big)(s)\Big]\Big\} \\
= & \frac{L}{N_n(L)}\Big\{\sum_{s=1}^{h}\Big[d^s C_{imr} * (1_{\{z=\cdot\}}(s) - \sum_{l\in E} d^s C_{ml} * q_{ml}(s)) \\
& - d^s g(m)\psi_{im} * 1_{\{z\le\cdot\}}(s)\sum_{l=1}^{s} d^s g(m)\psi_{im} * Q_{ml}(s)\Big]\Big\} \\
= & \frac{L}{N_n(L)}\Big\{\sum_{s=1}^{h} d^s\Big[C_{imr} * 1_{\{z=\cdot\}} - g(m)\psi_{im} * 1_{\{z\le\cdot\}}(s) \\
& - \sum_{l\in E}\big(C_{iml} * q_{ml}(s) - g(m)\psi_{im} * Q_{ml}(s)\big)\Big]\Big\}
\end{aligned}
\tag{3.17}
$$

Pyke and Schauffele (1964) provide a central limit theorem for expressions of the type (3.16). Then, its application, as suggested by Barbu and Limnios (2006) for reliability indicators, will give us the asymptotic variance of $\hat{M}_i(t)$. The application of this theorem requires the computation of several quantities marked below in bold.

Let

$$
\begin{aligned}
A_{imr} \doteq & \sum_{z=1}^{\infty} f(m,r,z)q_{mr}(z) \\
= & \Big(\frac{L}{N_n(L)}\Big)\Big\{\sum_{s=1}^{h} d^s\Big[C_{imr} * q_{mr}(s) - g(m)\psi_{im} * Q_{mr}(s) \\
& - \sum_{l\in E}\big(C_{iml} * q_{ml}(s) - g(m)\psi_{im} * Q_{ml}(s)\big)p_{mr}\Big]\Big\}
\end{aligned}
\tag{3.18}
$$

consequently we have

$$
A_{im} \doteq \sum_{r\in E} A_{imr} = 0 \tag{3.19}
$$

Let

$$
\begin{aligned}
B_{imr} \doteq & \sum_{z=1}^{\infty} f^2(m,r,z)q_{mr}(z) \\
= & \sum_{z=1}^{\infty}\Big(\frac{L}{N_m(L)}\Big)^2\Big\{\sum_{s=1}^{h} d^s\Big[C_{imr} * 1_{\{z=\cdot\}}(s) - g(m)\psi_{im} * 1_{\{z\le\cdot\}}(s) \\
& - \sum_{l\in E}\big(C_{iml} * q_{ml}(s) - g(m)\psi_{im} * Q_{ml}(s)\big)\Big]\Big\}^2 q_{mr}(z)
\end{aligned}
\tag{3.20}
$$

Denoting $D_{im}(s) = \sum_{l\in E}[C_{iml} * q_{ml}(s) - g(m)\psi_{im} * Q_{ml}(s)]$ and developing the square we get

$$B_{imr} = \Big(\frac{L}{N_m(L)}\Big)^2 \Bigg\{ \sum_{z=1}^{\infty} \Big[\Big(\sum_{s=1}^{h} d^s \big[C_{imr} * 1_{\{z=\cdot\}}(s) - g(m)\psi_{im} * 1_{\{z\le\cdot\}}(s) \big] \Big)^2 + \Big(\sum_{s=1}^{h} d^s D_{im}(s) \Big)^2 - 2 \Big(\sum_{a=1}^{h} d^a D_{im}(a) \Big) \Big(\sum_{s=1}^{h} d^s \big[C_{imr} * 1_{\{z=\cdot\}}(s) - g(m)\psi_{im} * 1_{\{z\le\cdot\}}(s) \big] \Big) \Big] q_{mr}(z) \Bigg\} \tag{3.21}$$

Then

$$B_{imr} = \Big(\frac{L}{N_m(L)}\Big)^2 \Bigg\{ \sum_{s=1}^{h} d^s \big[C_{imr}^2 * q_{mr}(s) + g(m)\Psi_{im}^2 * q_{mr}(s) - 2g(m) C_{imr}\Psi_{im} * q_{mr}(s) \big] + \Big(\sum_{s=1}^{h} d^s D_{im}(s) \Big)^2 p_{mr} - 2 \Big(\sum_{a=1}^{h} d^a D_{im}(a) \Big) \Big(\sum_{s=1}^{h} d^s \big[C_{imr} * q_{mr}(s) - g(m)\psi_{im} * Q_{ml}(s) \big] \Big) \Bigg\} \tag{3.22}$$

$$= \Big(\frac{L}{N_m(L)}\Big)^2 \Big(\sum_{s=1}^{h} d^{2s} \big[C_{imr} - g(m)\Psi_{im} \big]^2 * q_{mr}(s) + \Big(\sum_{s=1}^{h} d^s D_{im}(s) \Big)^2 p_{mr} - 2 \Big(\sum_{a=1}^{h} d^a D_{im}(a) \Big) \Big(\sum_{s=1}^{h} d^s \big[C_{imr} * q_{mr}(s) - g(m)\psi_{im} * Q_{ml}(s) \big] \Big) \tag{3.23}$$

Now let us compute

$$B_{im} \doteq \sum_{r\in E} B_{imr} = \Big(\frac{L}{N_m(L)}\Big)^2 \Bigg\{ \Big(\sum_{r\in E} \sum_{s=1}^{h} d^{2s} \big[C_{imr} - g(m)\Psi_{im} \big]^2 * q_{mr}(s) \Big) + \Big(\sum_{s=1}^{h} d^s D_{im}(s) \Big)^2 - 2 \Big(\sum_{a=1}^{h} d^a D_{im}(a) \Big) \Big(\sum_{s=1}^{h} d^s D_{im}(s) \Big) \Bigg\} \tag{3.24}$$

$$= \Big(\frac{L}{N_m(L)}\Big)^2 \Bigg\{ \Big(\sum_{r\in E} \sum_{s=1}^{h} d^{2s} \big[C_{imr} - g(m)\Psi_{im} \big]^2 * q_{mr}(s) \Big) - \Big(\sum_{s=1}^{h} d^s D_{im}(s) \Big)^2 \Bigg\}$$

Since $A_{ij} = 0$, $m_i \doteq \sum_{j=1}^{s} A_{ij} \frac{\mu_{ii}^*}{\mu_{jj}^*} = 0$ and then $m_f \doteq \frac{m_i}{\mu_{ii}} = 0$. Consequently in the Pyke-Schaufele's central limit theorem

$$\begin{aligned}\sigma_{\mathbf{i}}{}^2 &\doteq \sum_{m\in E} B_{im}\frac{\mu^*_{ii}}{\mu^*_{mm}}\\ &= \mu^*_{ii}\sum_{m\in E}\left(\frac{L}{N_m(L)}\right)^2\frac{1}{\mu^*_{mm}}\left\{\left(\sum_{r\in E}\sum_{s=1}^{h} d^{2s}\left[C_{imr}-g(m)\Psi_{im}\right]^2 * q_{mr}(s)\right)\right.\\ &\left.-\left(\sum_{s=1}^{h} d^s D_{im}(s)\right)^2\right\}\end{aligned} \tag{3.25}$$

Moreover $B_f \doteq \frac{\sigma_i^2}{\mu_{ii}}$ and since $\frac{N_m(L)}{L} \xrightarrow{a.s.} \frac{1}{\mu_{mm}}$ as $L \longrightarrow \infty$, then we get

$$\begin{aligned}\sigma^2_{(M)}(h) &= B_f\\ &= \frac{\mu^*_{ii}}{\mu_{ii}}\sum_{m\in E}\frac{\mu^2_{mm}}{\mu^*_{mm}}\left\{\sum_{r\in E}\sum_{s=1}^{h} d^{2s}[C_{imr}-g(m)\Psi_{im}]^2 * q_{mr}(s)-\left(\sum_{s=1}^{h} d^s D_{im}(s)\right)^2\right\}\end{aligned} \tag{3.26}$$

■

Note that at this time it is an easy task to construct the confidence intervals for this estimate, in fact at first we have to estimate the variance $\sigma^2_M(h)$ by replacing in expression (3.26) each element with its corresponding estimator then, since $\hat{\sigma}^2_M(h)$ is a consistent estimator, the resulting confidence interval for $M_i(t)$ still has asymptotic level $100(1-\alpha)\%$ and is given by:

$$\hat{M}_i(t) - z_{\frac{\alpha}{2}} \times \frac{\hat{\sigma}_M(t)}{\sqrt{L}} \leq M_i(t) \leq \hat{M}_i(t) + z_{\frac{\alpha}{2}} \times \frac{\hat{\sigma}_M(t)}{\sqrt{L}} \tag{3.27}$$

4 Application in quality of life estimation

The results can be applied in order to solve many real life problems which require semi-Markov processes, such as disability insurance models, see D'Amico, Guillén and Manca (2009). Here, we discuss a possible application in the modelling and estimation of the quality of life evolution of a person.

One of the most recent approaches in the quality of life modelling and estimation is to assume that the observed quality of life of a person is, at any time, a discrete variable which can be assessed through a self-rated questionnaire or by an interviewer, see Limnios, Mesbah and Sadek (2004).

The use of Markov chains to describe the longitudinal process of the quality of life of a person has been suggested by Chen and Sen (2001, 2004), Limnios *et al.* (2004)

and more recently by D'Amico (2009). In particular in D'Amico (2009), Markov reward processes have been proposed to study accumulated measure of the quality of life of a person.

As already stated, semi-Markov chains have sojourn time distributions (2.5) of any type, this is why they are more appropriate to applications than the Markov chains. For this reason we suppose that, at any time, the quality of life of a person is described by a discrete variable (state) and that its evolution in time is described by a HSMC.

Following D'Amico (2009), we give the following definition:

Definition 4.1 *The accumulated quality of life index at time t is*

$$AIQL_i(t) = Y_i(t) \tag{4.1}$$

where $Y_i(t)$ is the discounted accumulated semi-Markov reward process given that $X_0 = i$.

The functional (2.8) represents the expected value of the accumulated quality of life index and is an important indicator for comparing different quality of life policies.

To illustrate the results obtained in the previous section we adopt a simulation strategy. In general we do not know the true form of the semi-Markov kernel which should be estimated via historical data. Unfortunately we are not in possession of real data, so we assume that data are generated by the unknown kernel $\widetilde{Q}$ identified by the following embedded Markov chain:

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0.70 & 0.30 & 0.00 \\ 0.50 & 0.00 & 0.50 \\ 0.00 & 0.35 & 0.65 \end{pmatrix} \end{array}$$

and the following conditional waiting time distribution functions:

$$\begin{aligned} &G_{1,1}(\cdot) = cdf(Lognormal)(4,2); G_{1,2}(\cdot) = cdf(Lognormal)(2,1) \\ &G_{2,1}(\cdot) = cdf(Exponential)(3); G_{2,3}(\cdot) = cdf(Exponential)(5.8) \\ &G_{3,2}(\cdot) = cdf(Lognormal)(4,0.5); G_{3,3}(\cdot) = cdf(Exponential)(2) \end{aligned}$$

Thus, when the process is in state $i = 1$, the next state is sampled from the probability distribution $(0.70, 0.30, 0.00)$. If, for example, the state $j = 2$ is selected then a waiting time in state $i = 1$ has to be sampled from the distribution $G_{1,1}(\cdot)$ which is a Lognormal with parameters $(4,2)$. At this time a new state is sample from the distribution $(0.50, 0.00, 0.50)$ and so on. We construct a trajectory of length $L = 2000$ of the semi-Markov process generated by the assumed kernel $\widetilde{Q}$. From this trajectory we estimate the quantity of interest by using the proposed estimators.

In Figure 1 we show the estimation of the transition probabilities (dashed lines) with starting state $i = 1$ and we compare them with those obtained assuming as true the kernel $\widetilde{Q}$ (continuous lines). The lower rigt hand plot shows the estimation of the expected value of the accumulated quality of life (dashed line) and the true value calculated by using kernel $\widetilde{Q}$ (continuous line) assumed to generate the data.

Finally, notice that it could be possible and interesting to construct estimators of the higher order moments of a semi-Markov chain with rewards. To this end, we shall estimate the renewal type equations established by Stenberg *et al.* (2006, 2007) since no explicit expression, as simple and manageable as formula (2.8), exists for higher order moments.

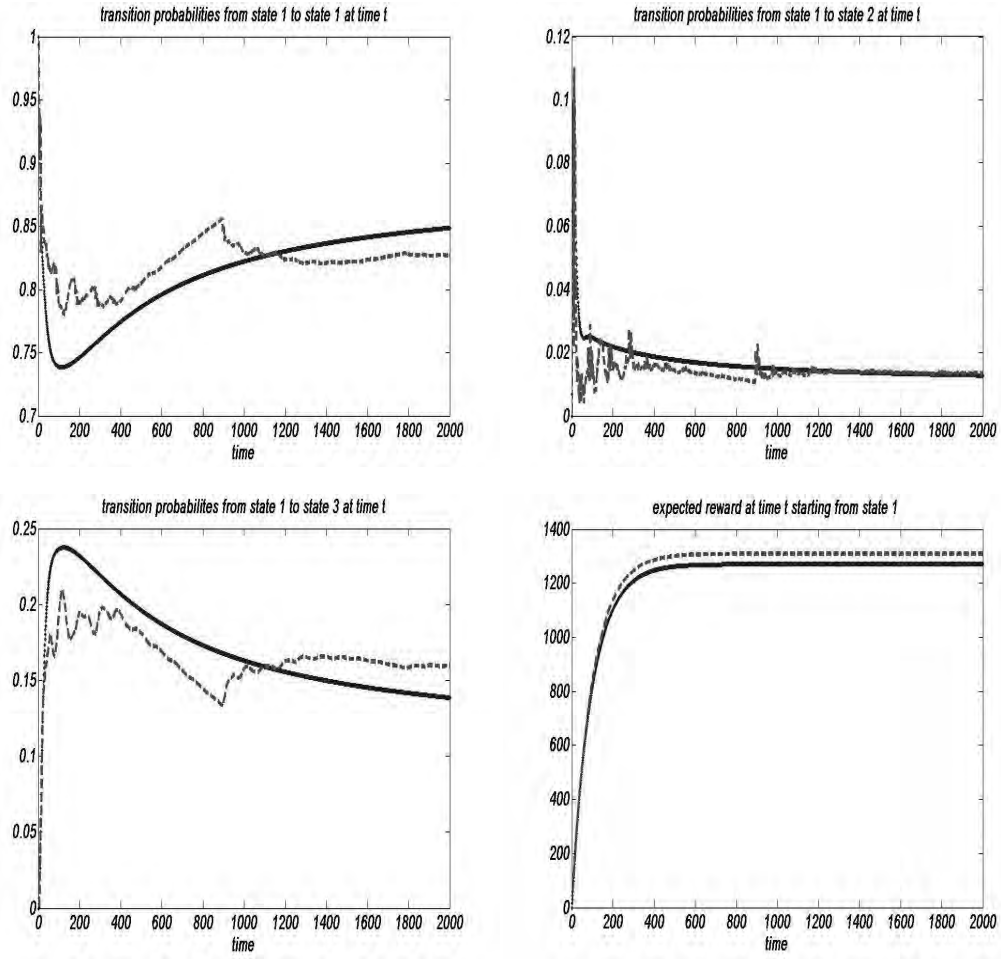


Figure 1: *Comparison between true and estimated values.*

Acknowledgments

The author would like to thank Profs. Nikolaos Limnios and Vlad Barbu for discussions and suggestions he had during his period as visiting researcher at the Université de Technologie de Compiègne (France). There, the author had the possibility of study the paper Barbu and Limnios (2006) before its publication.

References

Barbu, V., Boussemart, M. and Limnios, N. (2004). Discrete time semi-Markov model for reliability and survival analysis. *Communications in Statistics: Theory and Methods*, 33, 2833-2868.

Barbu, V. and Limnios, N. (2006). Nonparametric estimation for discrete time semi-Markov processes with applications in reliability. *Journal of Nonparametric Statistics*, 18, 483-498.

Barbu, V. and Limnios, N. (2008). *Semi-Markov Chains and Hidden Semi-Markov Models Toward Applications*. Springer-Verlag, New York Inc.

Chen, P. L. and Sen, P. K. (2004). Quality adjusted survival estimation with periodic observation. *Biometrics*, 57, 868-874.

Chen, P. L. and Sen, P. K. (2004). Quality adjusted survival estimation with periodic observation: a multistate survival analysis approach. *Communications in Statistics: Theory and Methods*, 33, 1327-1340.

D'Amico, G. (2009). Measuring the quality of life through Markov reward processes: analysis and inference. *Environmetrics*, to appear DOI: 10.1002/env.993.

D'Amico, G., Guillén, M. and Manca, R. (2009). Full backward non-homogeneous semi-Markov processes for disability insurance models: A Catalunya real data application. *Insurance: Mathematics and Economics*, 45, 173-179.

De Dominicis, R. and Manca, R. (1986). Some new results on the transient behaviour of semi-Markov reward processes. *Methods of Operations Research*, 53, 387-397.

Gardiner, J. C., Luo, Z., Bradley, C. J., Sirbu, C. M. and Given, C. W. (2006). A dynamic model for estimating changes in health status and costs. *Statistics in Medicine*, 25, 3648-3667.

Howard, R. (1971). *Dynamic Probabilistic Systems*, vol II. Wiley, New York.

Janssen, J. and Manca, R. (2006). *Applied Semi-Markov Processes*. Springer, New York.

Janssen, J. and Manca, R. (2007). *Semi-Markov Risk Models for Finance, Insurance and Reliability*. Springer, New York.

Khorshidian, K. and Soltani, A. R. (2002). Asymptotic behaviour of multivariate reward processes with nonlinear reward functions. *Bulletin of the Iranian Mathematical Journal*, 28, 1-17.

Limnios, N., Mesbah, M. and Sadek, A. (2004). A new index for longitudinal quality of life: modelling and estimation. *Environmetrics*, 15, 483-490.

Limnios, N. and Oprişan, G. (2001). *Semi-Markov Processes and Reliability*. Birkhäuser, Boston.

Pyke, R. and Schaufele, R. (1964). Limit theorems for Markov renewal processes. *Annals of Mathematical Statistics*, 35, 1746-1764.

Stenberg, F., Manca, R. and Silvestrov, D. (2006). Semi-Markov reward models for insurance. *Theory of Stochastic Processes*, 12, 239-254.

Stenberg, F., Manca, R. and Silvestrov, D. (2007). An algorithmic approach to discrete time non-homogeneous backward semi-Markov reward processes with an application to disability insurance. *Methodology and Computing in Applied Probability*, 9, 497-519.

SORT 33 (2) July-December 2009, 171-192

Estimation in the Birnbaum-Saunders distribution based on scale-mixture of normals and the EM-algorithm

N. Balakrishnan[1], Víctor Leiva[2], Antonio Sanhueza[3] and Filidor Vilca[4]

Abstract

Scale mixtures of normal (SMN) distributions are used for modeling symmetric data. Members of this family have appealing properties such as robust estimates, easy number generation, and efficient computation of the ML estimates via the EM-algorithm. The Birnbaum-Saunders (BS) distribution is a positively skewed model that is related to the normal distribution and has received considerable attention. We introduce a type of BS distributions based on SMN models, produce a lifetime analysis, develop the EM-algorithm for ML estimation of parameters, and illustrate the obtained results with real data showing the robustness of the estimation procedure.

MSC: Primary 65C10; Secondary 60E05.

Keywords: Birnbaum-Saunders distribution, EM-algorithm, kurtosis, maximum likelihood methods, robust estimation, scale mixtures of normal distributions.

1 Introduction

The family of scale mixtures of normal (SMN) distributions has attracted considerable attention; see, for example, Kelker (1970), Efron and Olshen (1978), Lange and Sinheimer (1993), Gneiting (1997), Taylor and Verbyla (2004), Walker and Gutiérrez-Peña (2007), and Lachos and Vilca (2007). This family provides flexible thick-tailed

Corresponding author: Víctor Leiva, Departamento de Estadística, Universidad de Valparaíso, Casilla 5030, Valparaíso, Chile. Email: victor.leiva@uv.cl; victor.leiva@yahoo.com
[1] Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada
[2] Department of Statistics, CIMFAV, Universidad de Valparaíso, Valparaíso, Chile
[3] Department of Mathematics and Statistics, Universidad de La Frontera, Temuco, Chile
[4] Department of Statistics, Universidade Estadual de Campinas, São Paulo, Brazil
Received: January 2009
Accepted: September 2009

distributions that are often used for robust estimation of parameters of a symmetric distribution; see Lange *et al.* (1989) and Lucas (1997). However, in many practical data involving variates such as lifetimes, pollutant concentrations, and family incomes, it is quite common to find skewed, heavy-tailed data. For this reason, it is necessary to have flexible distributions with good properties for fitting such kind of data. Distributions available with these characteristics are not abundant in the literature.

The Birnbaum-Saunders (BS) distribution is a positively skewed model with non-negative support that has also received considerable attention in the last two decades. This is primarily due to its derivation that is based on physical consideration, its attractive properties, and its close relationship to the normal distribution. These aspects of the BS model render it as an alternative to the normal model for data with non-negative support and positive skewness. For more details about various developments on the BS distribution, one may refer to Birnbaum and Saunders (1969a), Johnson *et al.* (1995, pp. 651-663), and Sanhueza *et al.* (2008).

Exploiting the relationship between the BS and normal distributions, it is possible to obtain a general class of BS distributions based on SMN models, which we call scale-mixture Birnbaum-Saunders (SBS) distributions. The three main reasons for developing this class of distributions are the following: (i) the use of the SBS specification to model observable data enables us to make robust estimation of parameters in a similar way to that of the SMS specification, which is not possible with the BS distribution or any other well-known compatible model such as the lognormal distribution, (ii) the theoretical arguments established in the genesis of the BS distribution can be transferred to the SBS one and thus it is an appropriate model for describing different phenomena that present accumulation of some type under stress, and (iii) SBS distributions allow us to efficiently compute the maximum likelihood (ML) estimates of the model parameters by using the EM-algorithm, which is not possible with the classical BS distribution; moreover, the estimation process proposed in this paper generalizes the one developed earlier by Birnbaum and Saunders (1969b). For more details about the EM-algorithm, see Dempster *et al.* (1977).

The rest of this paper is organized as follows. In Section 2, we introduce the SBS distributions and find their probability density function (pdf). In Section 3, we provide some properties, moments, conditional distributions, and some transformations of SBS models. In Section 4, we analyze some particular cases of these distributions. In Section 5, we produce a lifetime analysis mainly based on the failure rate function of SBS distributions. In Section 6, we describe the ML method for estimating the parameters of SBS models by means of the EM-algorithm. In Section 7, we provide an illustrative example that shows the usefulness of the SBS distributions for fitting three real data sets that are frequently utilized in the literature of this topic. Diagnostic and relative change procedures are used in this example, which show the inherent robustness of the estimation method based on SBS distributions. In addition, we discuss some aspects related to a computational implementation in R code for the results obtained in this paper. Finally, in Section 8, we draw some conclusions.

2 Scale-mixture Birnbaum-Saunders distributions

SMN models are related to the normal distribution through the stochastic representation

$$Y = \mu + \sqrt{g(U)}X, \tag{1}$$

where $X \sim \mathrm{N}(0,\sigma^2)$, U is a positive random variable (r.v.) independent of X with cumulative distribution function (cdf) $H(\cdot)$ indexed by a scalar or vector parameter and $g(\cdot)$ is a positive function. Note that when $g(U) = 1/U$ in equation (1), the distribution of Y reduces to the normal/independent distribution discussed by Lange and Sinheimer (1993). Similarly, when $g(U) = U$ in equation (1), the distribution of Y reduces to the SMN distribution studied by Fernandez and Steel (1999).

An r.v. Y has a SMN distribution with location and scale parameters, $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, respectively, iff its pdf is of the form

$$\phi_{\mathrm{SMN}}(y) = \int_0^\infty \phi\left(y|\mu, g(u)\,\sigma^2\right) dH(u), \tag{2}$$

where $\phi\left(\cdot\,|\mu, g(\cdot)\,\sigma^2\right)$ is the pdf of the normal distribution with mean μ and variance $g(\cdot)\,\sigma^2$ and $H(\cdot)$ is the cdf of U introduced in equation (1). For an r.v. Y with pdf given as in equation (2), the notation $Y \sim \mathrm{SMN}(\mu,\sigma^2;H)$ is used. Now, when $\mu = 0$ and $\sigma^2 = 1$, we use the simpler notation $Y \sim \mathrm{SMN}(H)$.

The BS distribution is related to the normal model through the stochastic representation

$$T = \frac{\beta}{4}\left[\alpha Z + \sqrt{\{\alpha Z\}^2 + 4}\right]^2, \tag{3}$$

where $Z \sim \mathrm{N}(0,1)$, $\alpha > 0$ and $\beta > 0$. Thus, if an r.v. T has the BS distribution with shape and scale parameters, α and β, respectively, then the notation $T \sim \mathrm{BS}(\alpha,\beta)$ is used in this case. From equation (3), the r.v. Z can be stochastically represented in terms of T as

$$Z = \frac{1}{\alpha}\left[\sqrt{\frac{T}{\beta}} - \sqrt{\frac{\beta}{T}}\right]. \tag{4}$$

In an analogous way, if the stochastic representation

$$T = \frac{\beta}{4}\left[\alpha\sqrt{g(U)}Z + \sqrt{\left\{\alpha\sqrt{g(U)}Z\right\}^2 + 4}\right]^2 \tag{5}$$

is considered, where $Y = \sqrt{g(U)}\,Z \sim \mathrm{SMN}(H)$, with $Z \sim \mathrm{N}(0,1)$, then the r.v. T follows a SBS distribution, which is denoted by $T \sim \mathrm{SBS}(\alpha,\beta;H)$. The stochastic representation

given in equation (5) is useful for generating random numbers, deriving moments and implementing the EM-algorithm for ML estimation in SBS models, which is shown in the following sections.

Theorem 1 *Let $T \sim \mathrm{SBS}(\alpha,\beta;H)$. Then, the pdf of T is*

$$f_T(t) = \phi_{\mathrm{SMN}}(a(t))A(t), \quad t>0, \alpha>0, \beta>0, \tag{6}$$

where $\phi_{\mathrm{SMN}}(\cdot)$ is the pdf given in equation (2) with $\mu=0$ and $\sigma^2=1$, $a(t) = \left[\sqrt{t/\beta}-\sqrt{\beta/t}\right]/\alpha$, and $A(t) = t^{-3/2}[t+\beta]/[2\alpha\beta^{1/2}]$ is the derivative of $a(t)$ with respect to t.

Proof. The required result is directly obtained from the stochastic representation given in equation (5) and the change-of-variable method. ■

3 Properties, moments, conditional distributions, and transformations of SBS models

The following theorem provides some properties of SBS distributions.

Theorem 2 *Let $T \sim \mathrm{SBS}(\alpha,\beta;H)$. Then,*

(i) *$cT \sim \mathrm{SBS}(\alpha, c\beta; H)$, with $c>0$;*
(ii) *$1/T \sim \mathrm{SBS}(\alpha, 1/\beta; H)$.*

Proof. Parts (i) and (ii) are directly obtained from the change-of-variable method. ■

Remark 1 Part (i) of Theorem 2 indicates that the SBS distributions belong to the scale family, while Part (ii) demonstrates that these distributions are closed under reciprocation; see Saunders (1974). In addition, Part (i) allows us to obtain a one-parameter SBS distribution by $\alpha T/\beta \sim \mathrm{SBS}(\alpha,\alpha;H)$.

The following theorem allows us to compute the moments of SBS distributions.

Theorem 3 *Let $T \sim \mathrm{SBS}(\alpha,\beta;H)$. If the r.v. $g(U)$ given in equation (1) has finite moments of all order, then the k-th moment of T is given by*

$$\mathbb{E}[T^k] = \beta^k \sum_{i=0}^{k} \binom{2k}{2i} \sum_{j=0}^{i} \binom{i}{j} \omega_{k+j-i} \left[\frac{\alpha}{2}\right]^{2[k+j-i]}, \quad k=1,2,\ldots,$$

where $\omega_r = \mathbb{E}[\{g(U)\}^r]$.

Proof. The required result is obtained from the stochastic representation given in equation (5) and by repeated application of the binomial theorem. ■

Remark 2 By using Theorem 2, the negative moments of T can be obtained by the fact that β/T and T/β have the same distribution. Consequently, we get $\mathbb{E}[T^{-k}] = \mathbb{E}[T^k]/\beta^{2k}$, for $k = 1, 2, \ldots$.

Corollary 1 *Let $T \sim \mathrm{SBS}(\alpha, \beta; H)$. Then, the mean, variance, and the coefficients of variation (CV), skewness (CS) and kurtosis (CK) of T are given by*

$$\mathbb{E}[T] = \frac{\beta}{2}[2 + \omega_1\alpha^2], \quad \mathrm{Var}[T] = \frac{\beta^2\alpha^2}{4}\left[\omega_1 + \{2\omega_2 - \omega_1^2\}\alpha^2\right],$$

$$\gamma[T] = \frac{\alpha\left[4\omega_1 + \{2\omega_2 - \omega_1^2\}\alpha^2\right]^{1/2}}{2 + \omega_1\alpha^2},$$

$$\alpha_3[T] = \frac{4\alpha\left[\{3\omega_2 - 3\omega_1^2\} + \frac{1}{2}\{2\omega_3 - 3\omega_1\omega_2 + \omega_1^3\}\alpha^2\right]}{\left[4\omega_1 + \{2\omega_2 - \omega_1^2\}\alpha^2\right]^{3/2}}, \quad \textit{and}$$

$$\alpha_4[T] = \frac{16\omega_2 + \{32\omega_3 - 48\omega_1\omega_2 + 24\omega_1^3\}\alpha^2 + \{8\omega_4 - 16\omega_1\omega_3 + 12\omega_1^2\omega_2 - \omega_1^4\}\alpha^4}{\left[4\omega_1 + \{2\omega_2 - \omega_1^2\}\alpha^2\right]^2},$$

respectively.

Remark 3 The dimensionless ratios $\gamma[T]$, $\alpha_3[T]$, and $\alpha_4[T]$ are functionally independent of the scale parameter β, with the skewness and kurtosis being basically controlled by the shape parameter α.

The following theorem and its corollary provide conditional distributions that are used in Section 5 to implement the EM-algorithm for the ML estimation of the parameters of SBS models.

Theorem 4 *Let $T \sim \mathrm{SBS}(\alpha, \beta; H)$. Then, the r.v. T given $U = u$, which is denoted by $T|(U = u)$, follows the classical BS distribution with parameters $\sqrt{g(u)}\,\alpha$ and β, i.e., $T|(U = u) \sim \mathrm{BS}\left(\sqrt{g(u)}\,\alpha, \beta\right)$.*

Proof. By using equation (5) and given $U = u$, we have $T = \beta\left[\alpha_u Z + \sqrt{\{\alpha_u Z\}^2 + 4}\right]^2/4$, where $\alpha_u = \alpha\sqrt{g(u)}$, which establishes the required result. ∎

Corollary 2 *Let $T \sim \mathrm{SBS}(\alpha, \beta; H)$. Then:*

(i) *The pdf of the r.v. $U|(T = t)$ is given by*

$$h_{U|T}(u|t) = \frac{\phi\left(a(t)|0, g(u)\right) h_U(u)}{\phi_{\mathrm{SMN}}\left(a(t)\right)}, \quad u > 0;$$

(ii) *The moments of the r.v. $g(U)|(T = t)$ are given by*

$$\mathbb{E}\left[\{g(U)\}^s \middle| (T=t)\right] = \frac{1}{\phi_{\mathrm{SMN}}\big(a(t)\big)} \int_0^\infty [g(u)]^{s-\frac{1}{2}}\, \phi\left(\frac{a(t)}{\sqrt{g(u)}}\right) dH(u), \quad s \in \mathbb{R}.$$

Next, we present some transformations related to SBS distributions.

Theorem 5 *Let $T \sim \mathrm{SBS}(\alpha,\beta;H)$. Then:*

(i) *The pdf of the r.v. $V = T^\eta$, with $\eta > 0$, is given by*

$$f_V(v) = \phi_{\mathrm{SMN}}\left(\frac{1}{\alpha}\left[\left\{\frac{v}{\delta}\right\}^{1/\sigma} - \left\{\frac{\delta}{v}\right\}^{1/\sigma}\right]\right) \frac{1}{\alpha\sigma v}\left[\left\{\frac{v}{\delta}\right\}^{1/\sigma} + \left\{\frac{\delta}{v}\right\}^{1/\sigma}\right], \quad v > 0,$$

where $\delta = \beta^\eta$ and $\sigma = 2\eta$;

(ii) *The pdf of the r.v. $V = \log(T)$ is given by*

$$f_V(v) = \phi_{\mathrm{SMN}}\left(\frac{2}{\alpha}\sinh\left(\frac{v-\rho}{2}\right)\right)\frac{1}{\alpha}\cosh\left(\frac{v-\rho}{2}\right), \quad -\infty < v < \infty,$$

where $\rho = \log(\beta)$;

(iii) *The pdf of the r.v.*

$$V = \left[\frac{1}{\alpha}\left\{\sqrt{\frac{T}{\beta}} - \sqrt{\frac{\beta}{T}}\right\}\right]^k$$

is given by

$$f_V(v) = \begin{cases} \frac{1}{k} v^{\frac{1}{k}-1}\, \phi_{\mathrm{SMN}}\big(v^{\frac{1}{k}}\big), \ -\infty < v < \infty, & \text{if } k \text{ is odd,} \\ \frac{2}{k} v^{\frac{1}{k}-1}\, \phi_{\mathrm{SMN}}\big(v^{\frac{1}{k}}\big), \ v > 0, & \text{if } k \text{ is even.} \end{cases}$$

Proof. Parts (i), (ii) and (iii) are proved by using the change-of-variable method. ■

Remark 4 The density given in Theorem 5(i) corresponds to the pdf of an extension of the SBS family, which we call the three-parameter SBS distributions, denoted by $T \sim \mathrm{SBS}(\alpha,\delta,\sigma;H)$. Note that $\sigma = 2$ produces the SBS family. Similarly, the density given in Theorem 5(ii) can be seen as the pdf of an extension of the sinh-normal distribution introduced by Rieck and Nedelman (1991).

Corollary 3 *Let $T \sim \mathrm{SBS}(\alpha,\beta;H)$ and $V = \big[\sqrt{T/\beta} - \sqrt{\beta/T}\big]/\alpha$. Then:*

(i) *The pdf of $V_1 = |V|$ is $f_{V_1}(v_1) = 2\phi_{\mathrm{SMN}}(v_1)$, for $v_1 > 0$;*

(ii) *The pdf of $V_2 = V^2$ is $f_{V_2}(v_2) = \phi_{\mathrm{SMN}}(v_2)/\sqrt{v_2}$, for $v_2 > 0$;*

(iii) *The pdf of $V_3 = \exp(V)$ is $f_{V_3}(v_3) = \phi_{\mathrm{SMN}}\big(\log(v_3)\big)/v_3$, for $v_3 > 0$.*

Remark 5 From Corollary 3, we see that the random variables V_1, V_2, and V_3 follow the half-symmetric, generalized chi-square with one degree of freedom (d.f.) and log-symmetric distributions, respectively. For more details on these distributions, one may refer to Fang *et al.* (1990).

4 Special cases of the SBS family

In this section, some special cases of the SBS family are considered, which are based on the contaminated normal, slash and t models. These are obtained from the stochastic representation given in equation (5), with $g(U) = 1/U$ and U having a known pdf. In addition, from Corollary 2, the conditional distribution of $U|(T = t)$ is also considered for all these special cases.

4.1 The contaminated normal BS distribution

As is well-known, contaminated normal models can be used for describing symmetric data with outlying observations, where one of the parameters represents the percentage of outliers, while the other one can be interpreted as a scale factor; see Little (1988). The contaminated normal distribution can be utilized for generating a BS distribution, which we call contaminated normal Birnbaum-Saunders (CN-BS) distribution. This model can be used for describing positively skewed non-negative data in the presence of atypical observations.

Consider the case when $T \sim \mathrm{SBS}(\alpha, \beta; H)$, with H being the cdf of the r.v. U, which has a pdf of the form

$$h_U(u) = \quad \mathbb{I}_{\{\gamma\}}(u) + [1 - \quad]\mathbb{I}_{\{1\}}(u), \quad 0 < \quad < 1, 0 < \gamma < 1, \tag{7}$$

where $\mathbb{I}_A(\cdot)$ denotes the indicator function of the set A. Then, from equations (2), (6) and (7), we have the pdf of the r.v. T to be

$$f_T(t) = \left[\quad \sqrt{\gamma}\,\phi\left(\sqrt{\gamma}\,a(t)\right) + [1 - \quad]\phi\left(a(t)\right)\right] \frac{t^{-3/2}[t+\beta]}{2\,\alpha\,\sqrt{\beta}}, \quad t > 0, \tag{8}$$

with $\alpha > 0, \beta > 0, 0 < \quad < 1$, and $0 < \gamma < 1$, where $\phi(\cdot)$ is the standard normal pdf and $a(t)$ is given as in equation (6). The model with pdf given as in equation (8) is the CN-BS distribution. In this case, the pdf of $U|(T = t)$ is given by

$$h_{U|T}(u|t) = \quad p(t,u)\,\mathbb{I}_{\{\gamma\}}(u) + [1 - \quad]\,p(t,u)\,\mathbb{I}_{\{1\}}(u), \tag{9}$$

where

$$p(t,u) = \frac{\sqrt{u}\exp\left(-\frac{u\,a(t)^2}{2}\right)}{\sqrt{\gamma}\exp\left(-\frac{\gamma\,a(t)^2}{2}\right) + [1- \quad]\exp\left(-\frac{a(t)^2}{2}\right)}.$$

Thus,

$$\mathbb{E}\big[U|(T=t)\big] = \frac{1- \quad + \quad \gamma^{3/2}\exp\left(\frac{[1-\gamma]a(t)^2}{2}\right)}{1- \quad + \quad \sqrt{\gamma}\exp\left(\frac{[1-\gamma]a(t)^2}{2}\right)}. \tag{10}$$

4.2 The slash Birnbaum-Saunders distribution

The slash distribution presents heavier tails than the normal one. In addition, when its shape parameter converges to infinity this distribution approaches the normal one. As in the case of the CN-BS distribution, the slash model can be utilized for generating a BS distribution, which we call slash Birnbaum-Saunders (SL-BS) distribution. A study that relates the BS and slash distributions has been done by Gómez *et al.* (2009).

Consider the case when $T \sim \text{SBS}(\alpha,\beta;H)$, with H being the cdf of the r.v. $U \sim \text{Beta}(\ ,1)$, which has a pdf of the form

$$h_U(u) = \quad u^{\ -1}, \quad 0<u<1, \quad >0. \tag{11}$$

Then, from equations (2), (6) and (11), we have the pdf of the r.v. T to be

$$f_T(t) = \left[\ \int_0^1 u^{\ -1}\phi\left(a(t)\big|0,\tfrac{1}{u}\right)du\right]\frac{t^{-3/2}[t+\beta]}{2\,\alpha\sqrt{\beta}}, \quad t>0,\alpha>0,\beta>0, \quad >0. \tag{12}$$

The model with pdf given as in equation (12) is the SL-BS distribution. In this case, $U|(T=t) \sim \text{Gamma}\left(1/2+\ \ ,a(t)^2/2\right)$ truncated at $[0,1]$. Thus,

$$\mathbb{E}\big[U|(T=t)\big] = \left[\frac{1+2}{a(t)^2}\right]\frac{P_1\left(\frac{3}{2}+\ \ ,\frac{a(t)^2}{2}\right)}{P_1\left(\frac{1}{2}+\ \ ,\frac{a(t)^2}{2}\right)}, \tag{13}$$

where $P_x(a,b)$ denotes the cdf of the Gamma distribution of parameters a and b evaluated at x according to the parameterization established in the pdf given in equation (14).

4.3 The Student-t BS distribution

The Student-*t* distribution with d.f., denoted by t , has been used as an alternative model to the normal one for obtaining qualitatively robust parameter estimates; see Lange *et al.* (1989) and Lucas (1997). Special cases of the t distribution are the Cauchy model, when $= 1$, and the normal model, when $\to \infty$. As in the case of the CN-BS and SL-BS distributions, the t model can be utilized for generating a BS distribution, which we call Student-*t* Birnbaum-Saunders (t -BS) distribution. The t -BS distribution can be used for obtaining qualitatively robust parameter estimates with respect to the BS distribution; see Balakrishnan *et al.* (2007), Leiva *et al.* (2008), and Barros *et al.* (2008).

Consider the case when $T \sim \mathrm{SBS}(\alpha, \beta; H)$, with H being the cdf of the r.v. $U \sim$ Gamma(/2, /2), which has a pdf of the form

$$h_U(u) = \frac{\left[\frac{}{2}\right]^{\frac{}{2}} u^{\frac{}{2}-1}}{\Gamma(\frac{}{2})} \exp\left(-\frac{u}{2}\right), \quad u > 0, \quad > 0. \tag{14}$$

Then, from equations (2), (6) and (14), we have the pdf of the r.v. T to be

$$f_T(t) = \frac{\Gamma\left(\frac{+1}{2}\right)}{\sqrt{\pi}\sqrt{}\,\Gamma\left(\frac{}{2}\right)} \left[1 + \frac{1}{\alpha^2}\left\{\frac{t}{\beta} + \frac{\beta}{t} - 2\right\}\right]^{-\frac{+1}{2}} \frac{t^{-3/2}[t+\beta]}{2\,\alpha\sqrt{\beta}}, \quad t > 0, \tag{15}$$

with $\alpha > 0$, $\beta > 0$, and > 0. The model with pdf given as in equation (15) is the t -BS distribution. In this case, we have $U|(T=t) \sim \mathrm{Gamma}\big([\ +1]/2, [\ + a(t)^2]/2\big)$. Thus,

$$\mathbb{E}\big[U|(T=t)\big] = \frac{+1}{+a(t)^2}.$$

5 Lifetime analysis in the SBS family

A useful indicator in lifetime analysis is the failure rate, which, for a non-negative r.v. T with pdf f_T and cdf F_T, is defined as $r_T(t) = f_T(t)/[1-F_T(t)]$, for $t > 0$, and $0 < F_T(t) < 1$. Although the distribution of T may be characterized equally in terms of the pdf or of the failure rate, according to Cox and Oakes (1984, pp. 24-28) and Balakrishnan *et al.* (2007), it is convenient to check the behaviour of the failure rate because distributions with densities whose shapes are similar could have failure rates with different shapes. If $r_T(t)$ is an increasing or decreasing function in t, then the distribution T belongs to the class of increasing failure rate (IFR) or decreasing failure rate (DFR) distributions, respectively. If $r_T(t) = \lambda > 0$, for $t > 0$, we have $F_T(t) = 1 - \exp(-\lambda t)$, and F_T is the exponential cdf with parameter λ. However, there are distributional families that have a non-monotone failure rate. In this case, an important value for lifetime analysis is the

change point of the failure rate of T (denoted by t_c), which is the value where the hazard changes its behaviour. Within the class of distributions with a non-monotone failure rate, we can identify ∩-or-∪ shapes. Particularly, for the ∩-shaped case, we also have two cases, when the failure rate is initially increasing until its change point and then: (i) decreases to zero, as in the case of the lognormal distribution or (ii) decreases until it becomes stabilized at a positive constant, as in the case of the classical BS distribution. For this reason, for distributional families with a non-monotone failure rate, their change point and their limiting behaviour are aspects that should be studied. For more details about life distributions and lifetime analysis, see Johnson *et al.* (1995, pp. 651-663), Marshall and Olkin (2007), and Saunders (2007).

As mentioned, the BS model belongs to the upside-down (or ∩-shaped) class and its failure rate approaches $1/[2\alpha^2\beta]$ as $t \to \infty$; see Chang and Tang (1993). A complete study of the change point of the BS failure rate can be found in Kundu *et al.* (2008) and Bebbington *et al.* (2008). Next, we give some results related to the failure rate of SBS distributions.

Theorem 6 *Let $T \sim \mathrm{SBS}(\alpha,\beta;H)$. Then, the failure rate of T is*

$$r_T(t) = \frac{\phi_{\mathrm{SMN}}\big(a(t)\big)A(t)}{\Phi_{\mathrm{SMN}}\big(-a(t)\big)}, \quad t>0,\ 0<\Phi_{\mathrm{SMN}}(\cdot)<1,$$

where $a(t)$ and $A(t)$ are given as in equation (6) and $\Phi_{\mathrm{SMN}}(\cdot)$ is the cdf of the SMN family.

Proof. It follows immediately direct from the definition of the failure rate and the SMN symmetry. ■

Theorem 7 *Let $T \sim \mathrm{SBS}(\alpha,\beta;H)$ and $r_T(\cdot)$ be its failure rate. Then,*

$$\lim_{t\to\infty} r_T(t) = \frac{1}{2\alpha^2\beta}\lim_{t\to\infty} W_{g,H}\left(a(t)^2\right), \tag{16}$$

where

$$W_{g,H}\left(a(t)^2\right) = \frac{\int_0^\infty g^{-3/2}(u)\exp\left(-\frac{a(t)^2}{2g(u)}\right)dH(u)}{\int_0^\infty g^{-1/2}(u)\exp\left(-\frac{a(t)^2}{2g(u)}\right)dH(u)}.$$

Proof. For $T \sim \mathrm{SBS}(\alpha,\beta;H)$, we have $f_T(t) = \phi_{\mathrm{SMN}}(a(t))A(t)$ and a function $f(\cdot)$ such that $\phi_{\mathrm{SMN}}(y) = f(y^2)$, for all $y \in \mathbb{R}$. In this case,

$$f(w) = \int_0^\infty \frac{1}{\sqrt{2\pi\, g(u)}}\exp\left(-\frac{w}{2g(u)}\right)dH(u), \quad w \geq 0. \tag{17}$$

Thus, $W_f(w) = f'(w)/f(w) = -W_{g,H}(w)/2$. As $g(\cdot)$ is a positive function, we have $W_{g,H}(w) \geq 0$. The proof of this theorem is similar to the one given in Theorem 4 of Leiva *et al.* (2008). ■

Theorem 8 *Let $T \sim \text{SBS}(\alpha, \beta; H)$ and that the distribution of U is unimodal. Then, the pdf of the T is unimodal and the mode, denoted by t_m, is obtained as solution of*

$$W_{g,H}\big(a(t_m)^2\big) = -\frac{\alpha^2\beta t_m[t_m+3\beta]}{[t_m-\beta][t_m+\beta]},$$

where $0 < t_m < \beta$.

Proof. From equation (17), we have that $f(w)$ is a monotonic non-increasing function for all $w > 0$, and so $\phi_{\text{SMN}}(\cdot)$ is a unimodal function. The rest of the proof follows by using Equation (8) of Leiva *et al.* (2008), replacing $-W_{g,H}(u)/2$ by $w_g(u)$, for $u > 0$. ■

Theorem 9 *The failure rate of SBS distributions is an upside-down function for all values of α and β.*

Proof. Following the same procedure as in Kundu *et al.* (2008), we can write the SBS failure rate as

$$r_T(t) = \frac{\phi_{\text{SMN}}(a(t))A(t)}{\Phi_{\text{SMN}}\big(-a(t)\big)}, \quad t > 0.$$

Thus, it is enough to prove that $\lim_{t\to 0} r_T(t) = 0$. As $f_T(t) = \phi_{\text{SMN}}(a(t))A(t)$, this can be expressed as

$$f_T(t) = \frac{1}{2\alpha\beta^{1/2}}\left[\int_0^\infty \frac{1}{\sqrt{2\pi g(u)}}\Delta_1(t,u)\,dH(u) + \beta\int_0^\infty \frac{1}{\sqrt{2\pi g(u)}}\Delta_2(t,u)\,dH(u)\right],$$

where $\Delta_1(t,u) = t^{-1/2}\exp\big(-a(t)^2/[2g(u)]\big)$ and $\Delta_2(t,u) = t^{-3/2}\exp\big(-a(t)^2/[2g(u)]\big)$. Then, following Kundu *et al.* (2008), we have that $\lim_{t\to 0}\Delta_1(t,u) = \lim_{t\to 0}\Delta_2(t,u) = 0$. Thus, since $\lim_{t\to 0} f_T(t) = 0$ and $\lim_{t\to 0}\Phi_{\text{SMN}}(-a(t)) = 1$, we have $\lim_{t\to 0} r_T(t) = 0$. ■

Remark 6 Note that Theorems 6 and 7 contain the expression of $W_{g,H}(\cdot)$. Next, we specify this expression for some particular cases (indicated in brackets) and obtain the limit of $r_T(t)$ as $t \to \infty$.

(i) [**CN-BS distribution**] Since

$$W_{g,H}\big(a(t)^2\big) = \frac{1 - \ + \ \gamma^{3/2}\exp\left(\frac{[1-\gamma]a(t)^2}{2}\right)}{1 - \ + \ \sqrt{\gamma}\exp\left(\frac{[1-\gamma]a(t)^2}{2}\right)},$$

then $\lim_{t\to\infty} r_T(t) = \gamma/[2\alpha^2\beta]$; note that if $\gamma = 1$, we have the case of the classical BS distribution;

(ii) [**SL-BS distribution**] Since

$$W_{g,H}\left(a(t)^2\right) = \left[\frac{1+2}{a(t)^2}\right] \frac{P_1\left(\frac{3}{2}+\ ,\frac{a(t)^2}{2}\right)}{P_1\left(\frac{1}{2}+\ ,\frac{a(t)^2}{2}\right)},$$

where $P_x(a,b)$ denotes the cdf of the Gamma distribution of parameters a and b evaluated at x according to the parameterization established in the pdf given in equation (14), then $\lim_{t\to\infty} r_T(t) = 0$; note that, in this case, the BS class has a failure rate similar to that of the lognormal distribution;

(iii) [t **-BS distribution**] Since

$$W_{g,H}\left(a(t)^2\right) = \frac{+1}{+a(t)^2},$$

then $\lim_{t\to\infty} r_T(t) = 1/[2\alpha^2\beta]$, if $\ \to\infty$, which corresponds to the case of the classical BS distribution; however, if $\ \to 0$, then $\lim_{t\to\infty} r_T(t) = 0$, which is also the case when $\ = 1$ corresponding to the Cauchy-BS distribution, such as occurs with the failure rate of the lognormal and SL-BS distributions.

Figure 1 shows different shapes of the failure rate of SBS distributions through which is possible compare their shapes to those of the classical BS model. This graphical analysis is coherent with the results given in this section.

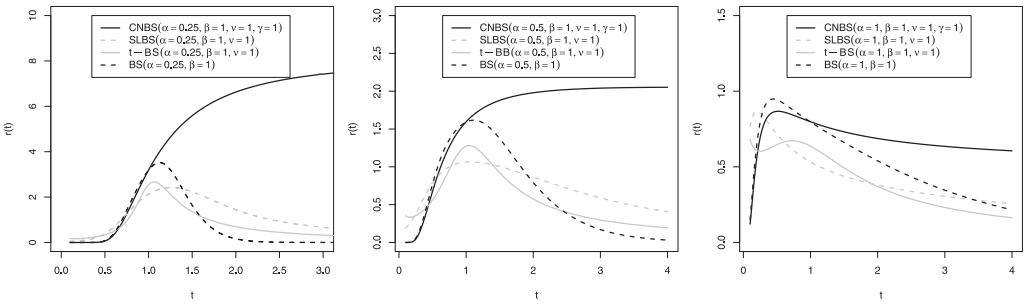


Figure 1: *Failure rate plots for the indicated distributions for some choices of the parameters.*

6 ML estimation via EM-algorithm in the SBS class

The EM-algorithm is a well-known tool for ML estimation when unobserved (or missing) data or latent variables are present while modeling. This algorithm enables the

computationally efficient determination of the ML estimates when iterative procedures are required. Specifically, let $\mathbf{t} = [t_1, \ldots, t_n]^\top$ and $\mathbf{u} = [u_1, \ldots, u_n]^\top$ denote observed and unobserved data, respectively. The complete data $\mathbf{t}_c = [\mathbf{t}^\top, \mathbf{u}^\top]^\top$ corresponds to the original data $\mathbf{t}$ augmented with $\mathbf{u}$. We now detail the implementation of the ML estimation of parameters of SBS distributions by using the EM-algorithm.

Let $T_1, \ldots, T_n$ be a random sample of size n, where $T_i \sim \mathrm{SBS}(\alpha, \beta; H)$, for $i = 1, \ldots, n$. Here, the parameter vector is $\boldsymbol{\theta} = [\alpha, \beta]^\top$, with $\boldsymbol{\theta} \in \boldsymbol{\Theta} \equiv \mathbb{R}^+ \times \mathbb{R}^+$. Let $\ell_c(\boldsymbol{\theta}\,|\mathbf{t}_c)$ and $Q(\boldsymbol{\theta}\,|\widehat{\boldsymbol{\theta}}) = \mathbb{E}[\ell_c(\boldsymbol{\theta}\,|\mathbf{t}_c)|\mathbf{t}, \widehat{\boldsymbol{\theta}}]$ denote the complete-data log-likelihood function and its expected value conditioned to the observed-data, respectively. Each iteration of the EM algorithm involves two steps, i.e., the expectation step (E-step) and the maximization step (M-step), which are defined by:

E-step. Compute $Q\big(\boldsymbol{\theta}\,|\widehat{\boldsymbol{\theta}}^{(r)}\big)$, for $r = 1, 2, \ldots$.

M-step. Find $\boldsymbol{\theta}^{(r+1)}$ such that $Q\big(\boldsymbol{\theta}^{(r+1)}\,|\widehat{\boldsymbol{\theta}}^{(r)}\big) = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q\big(\boldsymbol{\theta}\,|\widehat{\boldsymbol{\theta}}^{(r)}\big)$, for $r = 1, 2, \ldots$.

Note that, by using Theorem 4, the above setup can be written as

$$T_i|(U_i = u_i) \overset{\text{ind}}{\sim} \mathrm{BS}\big(\sqrt{g(u_i)}\,\alpha, \beta\big), \tag{18}$$

$$U_i \overset{\text{ind}}{\sim} h_U(u_i), \quad i = 1, \ldots, n. \tag{19}$$

We assume that the parameter vector that indexes the pdf $h_U(\cdot)$ is known. An optimal value of can then be chosen by using the Schwarz information criterion; see Spiegelhalter *et al.* (2002). Thus, under the hierarchical representation given in equations (18) and (19), it follows that the complete log-likelihood function associated with $\mathbf{t}_c = [\mathbf{t}^\top, \mathbf{u}^\top]^\top$ is given by

$$\ell_c(\boldsymbol{\theta}\,|\mathbf{t}_c) \propto -n\log(\alpha) - \frac{n}{2}\log(\beta) - \frac{1}{2\alpha^2}\sum_{i=1}^{n}\frac{1}{g(u_i)}\left[\frac{t_i}{\beta} + \frac{\beta}{t_i} - 2\right] + \sum_{i=1}^{n}\log(t_i + \beta). \tag{20}$$

Letting $\widehat{u}_i = \mathbb{E}[1/g(U_i)|t_i, \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}]$, for $i = 1, \ldots, n$, it follows that the conditional expectation of the complete log-likelihood function has the form

$$Q(\boldsymbol{\theta}\,|\widehat{\boldsymbol{\theta}}) \propto -n\log(\alpha) - \frac{n}{2}\log(\beta) - \frac{1}{2\alpha^2}\sum_{i=1}^{n}\widehat{u}_i\left[\frac{t_i}{\beta} + \frac{\beta}{t_i} - 2\right] + \sum_{i=1}^{n}\log(t_i + \beta). \tag{21}$$

We then have the EM-algorithm for the ML estimation of the parameters of the SBS distributions as follows:

E-step. Given $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, compute $\widehat{u}_i$, for $i = 1, \ldots, n$;

M-step. Update $\widehat{\boldsymbol{\theta}}$ by maximizing $Q(\boldsymbol{\theta}\,|\widehat{\boldsymbol{\theta}})$ in equation (21) over $\boldsymbol{\theta}$, which leads to the following expressions:

$$\widehat{\alpha}^2 = \frac{S_u}{\widehat{\beta}} + \frac{\widehat{\beta}}{R_u} - 2\bar{u} \quad \text{and} \quad \widehat{\beta}^2 - \widehat{\beta}\left[k(\widehat{\beta}) + 2\,\bar{u}R_u\right] + R_u\left[\bar{u}\,k(\widehat{\beta}) + S_u\right] = 0, \qquad (22)$$

where

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i, \quad S_u = \frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i t_i, \quad R_u = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{u}_i}{t_i}}, \quad \text{and} \quad k(x) = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x+t_i}}. \qquad (23)$$

Remark 7 Note that if $g(U) = 1$ in the EM-algorithm presented above (i.e., if the r.v. U is degenerate), then the M-step equations reduce to those when the BS distribution is used. Thus, the EM-algorithm here generalizes the results provided earlier by Birnbaum and Saunders (1969b). Moreover, the presented procedure provides an EM-algorithm for the t -BS distribution, which has been studied recently by Balakrishnan *et al.* (2007), Leiva *et al.* (2008), and Barros *et al.* (2008). Useful starting values necessary to implement this algorithm can be the ML estimates of the parameters of the BS distribution.

7 Illustrative numerical example

In this section, for the purpose of illustration, we analyze the data of Birnbaum and Saunders (1969b). These data correspond to fatigue life represented by cycles ($\times 10^{-3}$) until failure of aluminum specimens of type 6061-T6. These specimens were cut parallel to the direction of rolling and oscillating at 18 cycles per seconds. They were exposed to a pressure with maximum stress of 21,000 (Psi21), 26,000 (Psi26) and 31,000 (Psi31) pounds per square inch (psi) for $n = 101, 102$, and 101 specimens, respectively. All specimens were tested until failure.

We first present an exploratory data analysis. Table 1 provides a descriptive summary while Figure 2 shows the histograms and boxplots for Psi21, Psi26, and Psi31.

A careful look at Table 1 and Figure 2 reveals slightly positively skewed distributions with moderate kurtosis and some atypical observations, which can be potentially influential on the ML estimates of the parameters of the BS distribution. SBS distributions should consider the degrees of skewness and kurtosis present in the data. In addition, they also enable the estimation of the parameters of the model in a robust manner when atypical observations are present.

We now find the ML estimates of the parameters α and β of SBS distributions. Several authors have suggested to fix the parameter for the distribution of the r.v. U defined in equation (1) and assume it to be a known value or otherwise get information for it from the data. For instance, in the case of the Student-t distribution, the reason for doing it is that only when the parameter is fixed, the influence function is bounded, which allows us to obtain qualitatively robust estimators of parameters.

Table 1: *Descriptive statistics for the indicated data sets*

Data set	Mean	Median	StDev	CV	CS	CK	Range	Min.	Max.	n
Psi21	1400.84	1416.00	391.01	27.91%	0.14	−0.28	2070	370	2440	101
Psi26	397.88	400.00	62.32	15.66%	0.01	−0.21	327	233	560	102
Psi31	133.73	133.00	22.36	16.70%	0.33	0.97	142	70	212	101

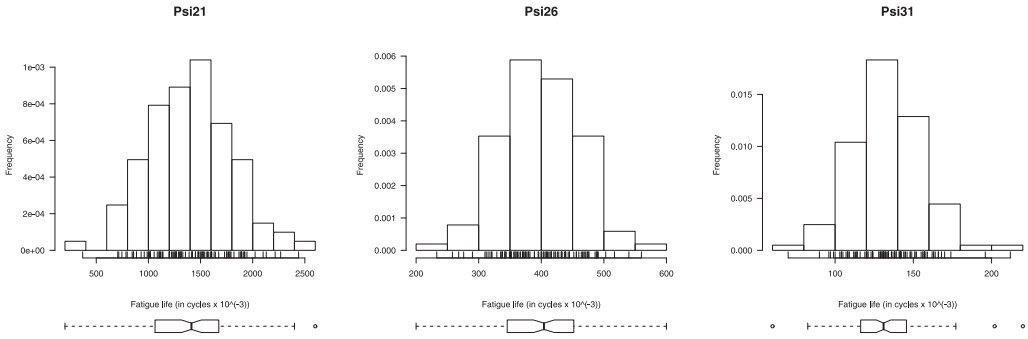


Figure 2: *Histograms and boxplots for the indicated data sets.*

For more details about these proposals and their justification, one can refer to Lucas (1997); see also Lange *et al.* (1989) and Leiva *et al.* (2008).

In order to select the best SBS model that fits the data, we have implemented the estimation procedure described in Section 6 in `R` code (`http://www.R-project.org`); see R Development Core Team (2008). As mentioned earlier, we use the ML estimates of α and β of the classical BS distribution as starting values in the numerical iterative procedure, which can be easily obtained from an `R` package called `bs` that is available from CRAN (`http://CRAN.R-project.org`); see Leiva *et al.* (2006). For ML estimation via EM-algorithm in SBS models, we have implemented the command `smnbsEstimation()`, which automatically chooses the distribution that best fits the data set among the CN-BS, SL-BS and t -BS distributions by maximizing the likelihood function. This command also computes the ML estimates of the parameters of SBS models separately. For instance, in the case of the t -BS distribution, the following algorithm can be used for estimating its parameters:

(A1) For = 1 to = 100 by 1:
- **(A1.1)** Determine the ML estimates of the parameters α and β of the t -BS model via the EM-algorithm proposed in Section 6 by beginning with the ML estimates of α and β of the BS distribution as starting values for the numerical procedure;
- **(A1.2)** Compute the likelihood function;

(A2) Choose the value of that maximizes the likelihood function and then establish as ML estimates of α and β those associated with that maximum likelihood function.

The real data sets used in this example are implemented in the `bs` package, which are called `psi21`, `psi26` and `psi31` and obtained by using the instructions `data(psi21)`, `data(psi26)`, and `data(psi31)`, respectively. Now, if the following commands are used

```
> smnbsEstimation(psi21)
> smnbsEstimation(psi26)
> smnbsEstimation(psi31)
```

then the distributions that best fit the Psi21, Psi26 and Psi31 data set are chosen among the CN-BS, SL-BS and t -BS models. In addition, the ML estimates of α and β of these models are computed. The results can be saved in R variables as follows:

```
> estimatespsi21 <- smnbsEstimation(psi21)
> estimatespsi26 <- smnbsEstimation(psi26)
> estimatespsi31 <- smnbsEstimation(psi31)
```

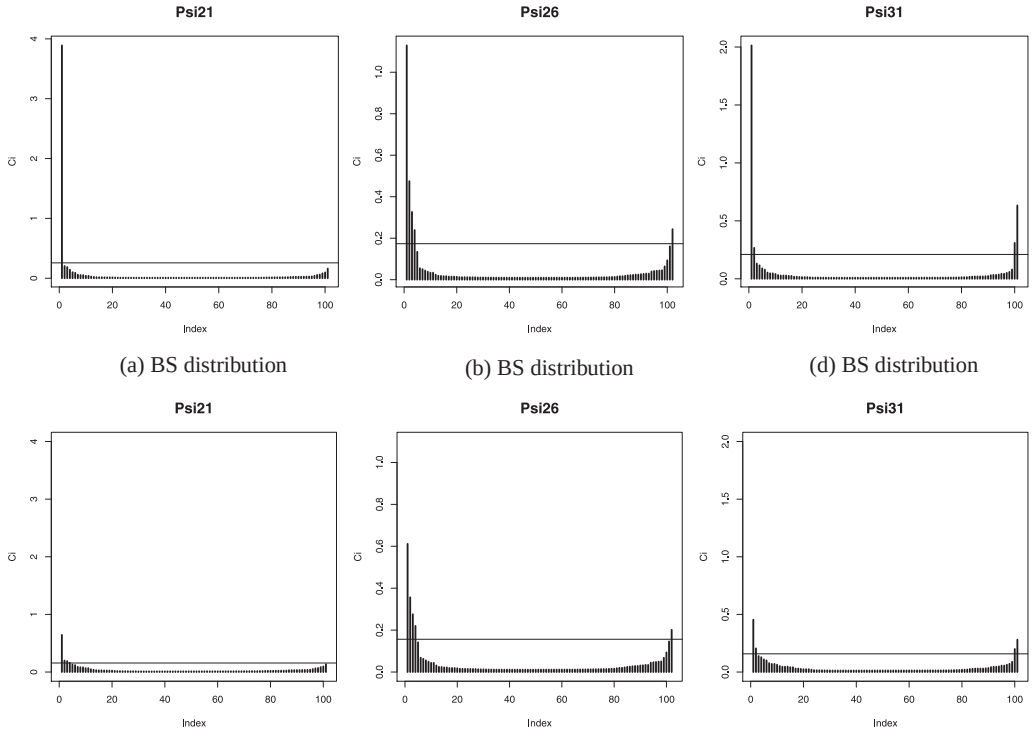
obtaining, respectively,

```
> smnbsEstimation(psi21)
$Best model
[1] "CN-BS"
$alpha
[1] 0.2737684
$beta
[1] 1356.624
$nu
[1] 0.02
$gamma
[1] 0.08
$logLikelihood
[1] -747.3013
> smnbsEstimation(psi26)
$Best model
[1] "BS-t"
$alpha
[1] 0.1534232
$beta
[1] 393.9034
$nu
[1] 18
$logLikelihood
[1] -567.4124
```

and

```
> smnbsEstimation(psi31)
$Best model
[1] "CN-BS"
$alpha
[1] 0.1465890
$beta
[1] 132.2940
$nu
[1] 0.06
$gamma
[1] 0.18
$logLikelihood
[1] -455.4714
```

From these results and within the three considered distributions, we can see that the CN-BS, t_{18}-BS, CN-BS distributions present the best fit to the Psi21, Psi26 and Psi31 data sets, respectively.



(a) BS distribution (b) BS distribution (d) BS distribution

(d) SBS distribution (e) SBS distribution (f) SBS distribution

Figure 3: *Influence index plots for the indicated data sets and models*

In order to show the inherent robustness of the estimation procedure based on Birnbaum-Saunders distributions from scale-mixture of normals, we carry out a brief diagnostic analysis based on local influence and relatives changes. For more details about local influence, see Cook (1986).

In Figure 3, we can observe the inherent robustness of the estimation procedure based on SBS distributions. Values of C_i, for $i = 1, \ldots, n$, –total local influence for the ith case– show a more pronounced potential influence of observations for the classical BS model. For more details about the local influence techniques in BS models, see Galea *et al.* (2004) and Leiva *et al.* (2007).

Table 2 presents the relative changes (RC), in percentage, of each parameter estimate, defined by $\mathrm{RC}_{\theta_j} = |[\hat{\theta}_j - \hat{\theta}_{j(\mathrm{I})}]/\hat{\theta}_j| \times 100\%$, for $j = 1, 2$, with $\theta_1 = \alpha$ and $\theta_2 = \beta$, where $\hat{\theta}_{j(\mathrm{I})}$ denotes the ML estimate of θ_j after the set I of cases has been removed. From this table, we note that the RC values are greater for the classical BS model than for the SBS models. Thus, all the specimens are retained in the analysis as they do not greatly affect the ML estimates under the SBS models.

Table 2: *RC (in %) for the indicated parameters, models and data sets*

		SBS		BS	
Data set	Dropped case(s)	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
S21	$\{1\}$	3.00	1.02	9.83	1.60
	$\{101\}$	1.41	0.56	1.48	0.61
	$\{1, 101\}$	4.44	0.49	11.5	0.99
S26	$\{1\}$	2.35	0.20	4.96	0.53
	$\{102\}$	1.66	0.22	1.96	0.35
	$\{1, 102\}$	4.00	0.02	7.02	0.18
S31	$\{1\}$	2.73	0.19	6.90	0.66
	$\{101\}$	2.25	0.18	3.56	0.48
	$\{1, 101\}$	4.94	0.00	10.7	0.18

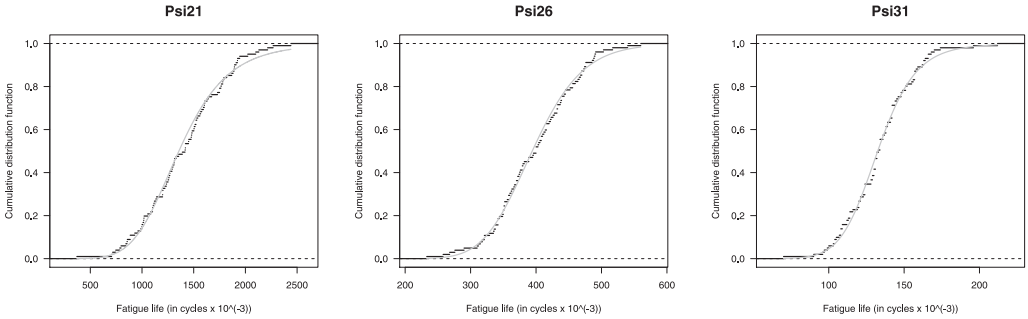


Figure 4: *Empirical cdf (in bold) against estimated SBS theoretical cdf (in gray) for the indicated data.*

In order to show how the SBS family fits the data, we use the invariance property of the ML estimators for obtaining the estimated SBS cdf, which is shown in Figure 4 on the empirical cdf of the data. In addition, the application of the Kolmogorov-Smirnov test provides the p-values 0.765, 0.974, and 0.799 for the Psi21, Psi26 and Psi31 data sets, respectively. These results suggest an excellent agreement between the SBS models and the data.

8 Concluding remarks

We have introduced a general class of Birnbaum-Saunders distributions based on scale mixtures of normal distributions. This class allows us to obtain qualitatively robust maximum likelihood estimates and efficiently compute these by using the EM-algorithm. Specifically, we have found the pdf, shown some properties, computed the moments, considered some transformations, and carried out a lifetime analysis based on the failure rate of scale-mixture Birnbaum-Saunders distributions. We have also presented some particular cases of these distributions based on the contaminated normal, slash and t models. In addition, we have implemented in R code different aspects pertaining to the considered distributions, including the mentioned EM-algorithm for determining the ML estimates of their parameters, which can make this model more attractive to users. Moreover, we have illustrated the results obtained for this class of distributions and discussed the computational implementation of them by using a numerical example with three different data sets, which display the flexibility, adequacy, and inherent robustness of the estimation procedure based on a Birnbaum-Saunders distribution from scale-mixture of normals.

8 Acknowledgments

The authors wish to thank the editors and referees for their helpful comments that aided in improving this article. This study was partially supported by a FAPESP grant from Brazil and FONDECYT 1080326, DIPUV 29-2006 and FONDECYT 1090265 grants from Chile.

References

Balakrishnan, N., Leiva, V., and López, J. (2007). Acceptance sampling plans from truncated life tests based on the generalized Birnbaum-Saunders distribution. *Communications in Statistics: Simulation and Computation*, 36, 643-656.

Barros, M., Paula, G. A., and Leiva, V. (2008). A new class of survival regression models with heavy-tailed errors: robustness and diagnostics. *Lifetime Data Analysis*, 14, 316-332.

Bebbington, M., Lai, C-D, and Zitikis, R. (2008). A proof of the shape of the Birnbaum-Saunders hazard rate function. *The Mathematical Scientist*, 33, 49-56.

Birnbaum, Z. W. and Saunders, S. C. (1969a). A new family of life distributions. *Journal of Applied Probability*, 6, 637-652.

Birnbaum, Z. W. and Saunders, S. C. (1969b). Estimation for a family of life distributions with applications to fatigue. *Journal of Applied Probability*, 6, 328-347.

Chang, D. S. and Tang, L. C. (1993). Reliability bounds and critical time for the Birnbaum-Saunders distribution. *IEEE Transactions on Reliability*, 42, 464-469.

Cook, R. D. (1986). Assessment of local influence (with discussion). *Journal of The Royal Statistical Society Series B–Statistical Methodology*, 48, 133-169.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* Chapman & Hall, London.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of The Royal Statistical Society Series B–Statistical Methodology*, 39, 1-38.

Efron, B. and Olshen, R. A. (1978). How broad is the class of normal scale mixtures. *Annals of Statistics*, 6, 1159-1164.

Fang, K. T., Kotz, S. and Ng. K. W. (1990). *Symmetric Multivariate and Related Distributions.* Chapman & Hall, London.

Fernandez, C. and Steel, M. (1999). Multivariate student *t* regression models: pitfalls and inference. *Biometrika*, 86, 153-167.

Kelker, D. (1970). Distribution theory of spherical distributions and a location scale parameter generalization. *Sankhyā A*, 32, 419-430.

Galea, M., Leiva, V. and Paula, G. A. (2004). Influence diagnostics in log-Birnbaum-Saunders regression models. *Journal of Applied Statistics*, 31, 1049-1064.

Gneiting, T. (1997). Normal scale mixtures and dual probability densities. *Journal of Statistical Computation and Simulation*, 59, 375-384.

Gómez, H. W., Olivares-Pacheco, J. F. and Bolfarine, H. (2008). An extension of the generalized Birnbaum-Saunders distribution. *Statistical and Probability Letters*, 79, 331-338.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995). *Continuous Univariate Distributions-Vol. 2*, Second edition. Wiley, New York.

Kundu, D., Kannan, N. and Balakrishnan, N. (2008). On the hazard function of Birnbaum-Saunders distribution and associated inference. *Computational Statistics and Data Analysis*, 52, 2692-2702.

Lachos, V. H. and Vilca, F. (2007). Skew-normal/independent distributions with applications. Technical report UNICAMP, Brazil. `http://www.unicamp.br/anuario/2007/IMECC/DE/DE--0040.html`

Lange, K. L., Little, J. A. and Taylor, M. G. J. (1989). Robust statistical modeling using the *t* distribution. *Journal of the American Statistical Association*, 84, 881-896.

Lange, K. and Sinsheimer, J. S. (1993). Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, 2, 175-198.

Leiva, V., Hernández, H. and Riquelme, M. (2006). A new package for the Birnbaum-Saunders distribution. *The R Journal (R News)*, 6, 35-40.

Leiva, V., Barros, M., Paula, G. A. and Galea, M. (2007). Influence diagnostics in log-Birnbaum-Saunders regression models with censored data. *Computational Statistics and Data Analysis*, 51, 5694-5707.

Leiva, V., Riquelme, M., Balakrishnan, N. and Sanhueza, A. (2008). Lifetime analysis based on the generalized Birnbaum-Saunders distribution. *Computational Statistics and Data Analysis*, 52, 2079-2097.

Little, R. J. A. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistics*, 37, 23-38.

Lucas, A. (1997). Robustness of the student *t* based M-estimator. *Communications in Statistics: Theory and Methods*, 26, 1165-1182.

Marshall, A. W. and Olkin, I. (2007). *Life Distributions*. Springer, New York.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org`

Rieck, J. R. and Nedelman, J. R. (1991). A log-linear model for the Birnbaum-Saunders distribution. *Technometrics*, 33, 51-60.

Sanhueza, A., Leiva, V. and Balakrishnan, N. (2008). The generalized Birnbaum-Saunders distribution and its theory, methodology and application. *Communications in Statistics: Theory and Methods*, 37, 645-670.

Saunders, S. C. (1974). A family of random variables closed under reciprocation. *Journal of the American Statistical Association*, 69, 533-539.

Saunders, S. C. (2007). *Reliability, Life Testing and Prediction of Services Lives*. Springer, New York.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of complexity and fit. *Journal of The Royal Statistical Society Series B–Statistical Methodology*, 64, 1-34.

Taylor, J. and Verbyla, A. (2004). Joint modeling of location and scale parameters of *t* distribution. *Statistical Modelling*, 4, 91-112.

Walker, S. G. and Gutiérrez-Peña, E. (2007). Bayesian parametric inference in a nonparametric framework. *Test*, 16, 188-197.

SORT 33 (2) July-December 2009, 193-212

Eliciting expert opinion for cost-effectiveness analysis: a flexible family of prior distributions

M. Martel, M. A. Negrín and F. J. Vázquez-Polo*

University of Las Palmas de Gran Canaria

Abstract

The Bayesian approach to statistics has been growing rapidly in popularity as an alternative to the classical approach in the economic evaluation of health technologies, due to the significant benefits it affords. One of the most important advantages of Bayesian methods is their incorporation of prior information. Thus, use is made of a greater amount of information, and so stronger results are obtained than with frequentist methods. However, since Stevens and O'Hagan (2002) showed that the elicitation of a prior distribution on the parameters of interest plays a crucial role in a Bayesian cost-effectiveness analysis, relatively few papers have addressed this issue.
In a cost-effectiveness analysis, the parameters of interest are the mean efficacy and mean cost of each treatment. The most common prior structure for these two parameters is the bivariate normal structure. In this paper, we study the use of a more general (and flexible) family of prior distributions for the parameters. In particular, we assume that the conditional densities of the parameters are all normal.
The model is validated using data of a real clinical trial. The posterior distributions have been simulated using Markov Chain Monte Carlo techniques.

MSC: 97K80, 62F15.

Keywords: Bayesian analysis, cost-effectiveness, prior information, elicitation, conditionally specified distributions.

1 Introduction

Spiegelhalter, Feedman and Parmar (1994) argued the use of Bayesian methodology as a formal basis for applying external evidence in cost-effectiveness analysis (CEA). Since then, many authors have discussed the advantages of this methodology versus the classical or frequentist approach (Briggs, 1999; Heitjan, Moskowitz and William, 1999; Fry-

**Address for correspondence:* Department of Quantitative Methods, Faculty of Economics, University of Las Palmas de Gran Canaria, E-35017 Las Palmas de Gran Canaria. Canary Islands, Spain. {mmartel or mnegrin or fjvpolo}@dmc.ulpgc.es

Received: April 2009

Accepted: September 2009

back, Chinnis and Ulvila, 2001; O'Hagan, Stevens and Montmartin, 2001; Vázquez-Polo and Negrín, 2004; among others).

The incorporation of prior information allows Bayesian methods to access more information and so to produce stronger inferences. Stevens and O'Hagan (2002) discuss the advantages of incorporating prior information in cost-effectiveness analysis of clinical trial data, exploring mechanisms to safeguard scientific rigour in the use of prior information. Since it has become available, a number of different techniques have been developed to elicit or extract prior information from experts (O'Hagan, Buck, Daneshkhah, Eiser, Garthwaite, Jenkinson, Oakley and Rakow, 2006). However, few studies have addressed these techniques in the area of health economics (Fenwick, Palmer, Claxton, Sculpher, Abrams and Sutton, 2006; Smith and Marshall, 2006; Leal, Wordsworth, Legood and Blair, 2007). In the present paper we aim to collaborate in promoting the use of Bayesian analysis by proposing a general and flexible method to incorporate prior information by means of conditionally specified distributions.

CEA is a form of economic evaluation that examines both the costs and health outcomes of alternative health technologies or treatments. The most prevalent measures for the comparison of treatments are the incremental cost-effectiveness ratio (ICER), the incremental net benefit (INB) and the cost-effectiveness acceptability curve (CEAC).

The ICER is defined by:

$$ICER = \frac{\gamma_1 - \gamma_0}{\mu_1 - \mu_0} = \frac{\Delta\gamma}{\Delta\mu}, \tag{1}$$

where γ_j and μ_j are the average cost and effectiveness under treatment j (1, new; and 0, for the current or control treatment), respectively.

The INB of treatment 1 versus treatment 0 is defined as

$$INB(R_c) = R_c \cdot \Delta\mu - \Delta\gamma, \tag{2}$$

for each R_c, which is interpreted by O'Hagan and Stevens (2001) as the cost that decision-makers are willing to accept in order to increase the effectiveness of the treatment applied by one unit. Thus, analyzing whether the alternative treatment is more cost effective than the control treatment is equivalent to determining whether $INB(R_c)$ is positive.

In practice, it is not a simple matter for the decision-maker to determine a single R_c, and so a CEAC is constructed. This curve provides a graphical representation of the probability of the alternative treatment being preferred ($\Pr(INB(R_c) > 0)$) for each value R_c.

We focus on the normal case. Classical cost-effectiveness analysis and most published Bayesian studies assume normality in the distributions of cost and effectiveness (Willan and O'Brien, 1996; Laska, Meisner and Siegel, 1997; Stinnett and Mullahy, 1998; Tambour, Zethraeus and Johannesson, 1998; Heitjan *et al.*, 1999; Briggs, 1999).

Although efficacy outcome data can be binary and patient cost data are likely to be right skewed, the central limit theorem guarantees for sufficiently large sample sizes that the means will be normally distributed. Löthgren and Zethraeus (2000) affirm that "the normal distribution result is valid whether or not the individual cost and effect distributions are normal. The more skewed and non-normal the individual distribution is, the larger sample sizes are needed for the normal distribution approximation to be valid".

A Bayesian analysis of the normal case was examined by O'Hagan *et al.* (2001), who considered the patient level data $\{x_{ij} : i = 1, 2, \ldots, n_j; j = 0, 1\}$ from a clinical trial, where $x_{ij} = (e_{ij}, c_{ij})$ consists of an effectiveness measures e_{ij} and an associated cost c_{ij}. The index j is used to denote the treatment and n_j denotes the sample size for each treatment j.

We denote by $f(x_{ij}|\mu_j, \gamma_j, \Sigma_j)$ the parametric distribution generating data x_{ij} from treatment j. The parameters of this function are the mean cost (γ_j), the mean efficacy (μ_j) and the variance-covariance matrix Σ_j. Then the likelihood is:

$$\ell(\bar{x}|\mu_0, \gamma_0, \Sigma_0; \mu_1, \gamma_1, \Sigma_1) = \prod_{j=0}^{1} \prod_{i=1}^{n_j} f(x_{ij}|\mu_j, \gamma_j, \Sigma_j). \tag{3}$$

It is assumed that $f(x_{ij}|\mu_j, \gamma_j, \Sigma_j)$ is a bivariate normal distribution for each treatment j

$$f(x_{ij}|\alpha_j, \Sigma_j) = (2\pi|\Sigma_j|)^{-1/2} \exp\left\{-\frac{1}{2}(x_{ij} - \alpha_j)^{\mathsf{T}} \Sigma_j^{-1} (x_{ij} - \alpha_j)\right\}, \tag{4}$$

where $\alpha_j = (\mu_j, \gamma_j)$, i.e. the mean effectiveness and cost for treatment j, respectively.

A Bayesian analysis of model (4) requires the specification of a prior distribution on the parameters. Quantifying the expert's opinion as a probability distribution is a difficult task, and the method presented is intended to help the expert perform the task in a way that is as easy, rigorous and computationally allowable as possible.

A convenient class of prior distributions is a general conditional-conjugate prior. Specifically, O'Hagan *et al.* (2001) assume a bivariate normal distribution for α_j and an inverse Wishart prior distribution for the variance matrices Σ_j.

Although the bivariate normal prior distribution is general and convenient it does present some limitations. For example, the correlation between variables is independent of the values of the variables and it is a unimodal distribution. In this paper, we study the use of a more general family of prior distributions for the parameters of interest. In particular, we assume that the conditional density of μ_j for a given γ_j and the conditional density of γ_j for a given μ_j are both normal. This assumption is manifestly different from one of classical bivariate normality with its familiar elliptical contours. The utility of conditionally specified priors has been explored in other areas, such as the analysis of insurance claims (Sarabia and Gómez-Déniz, 2008; Sarabia, Castillo, Gómez-Déniz and Vázquez-Polo, 2005).

The paper is organized as follows: Section 2 presents the normal case of cost-effectiveness analysis with prior distributions based on a conditional specification. In Section 3 some examples are given to show that the methodology is readily applicable. We use a practical application with real data from a clinical trial, comparing two alternative treatments for asymptomatic HIV patients. Markov Chain Monte Carlo (MCMC) procedures are used to simulate the posterior distribution. Section 4 presents a discussion of the results obtained and some conclusions are then drawn.

2 Bayesian cost-effectiveness analysis with prior distributions based on conditional specification

Our basic prior formulation for model (4) assumes that the joint distribution factorizes as

$$\pi(\alpha_0, \alpha_1, \Sigma_0, \Sigma_1) = \pi(\alpha_0) \cdot \pi(\alpha_1) \cdot \pi(\Sigma_0) \cdot \pi(\Sigma_1). \tag{5}$$

That is, we assume independence between treatments and between the means (α_j) and the variance matrices (Σ_j). Inverse Wishart prior distributions are assumed for the variance matrices Σ_0 and Σ_1. Specifically, we take $\Sigma_j \sim IW(A_j, f_j)$ the prior density of which is

$$\pi(\Sigma_j) \propto |\Sigma_j|^{(-f_j+3)/2} \exp\left\{tr(\Sigma_j^{-1}A_j)/2\right\},$$

over the space of positive-definite 2×2 matrices. Thus f_j is the prior degrees of freedom parameter and the prior expectation of Σ_j is $(f_j - 3)^{-1}A_j$, provided $f_j > 3$.

It is reasonable to assume a prior normal distribution of μ_j for a given γ_j and of γ_j for a given μ_j. A bivariate normal distribution was proposed by O'Hagan *et al.* (2001), but that is only a particular case with normal conditionals.

Castillo and Galambos (1989) showed the specification of the class of all bivariate densities with normal conditionals. We seek to obtain all joint densities $\pi(\mu, \gamma)$ such that every conditional density of μ given γ is normal with mean $\delta_1(\gamma)$ and variance $\sigma_1^2(\gamma)$ (which may depend on γ) and every conditional density of γ given μ with mean $\delta_2(\mu)$ and variance $\sigma_2^2(\mu)$ (which may depend on μ).

The above authors found that all the bivariate densities with normal conditionals are those of the form

$$\pi(\mu, \gamma) = \exp\left\{ \left(1, \mu, \mu^2\right) \begin{pmatrix} m_{00} & m_{01} & m_{02} \\ m_{10} & m_{11} & m_{12} \\ m_{20} & m_{21} & m_{22} \end{pmatrix} \begin{pmatrix} 1 \\ \gamma \\ \gamma^2 \end{pmatrix} \right\}. \tag{6}$$

The conditional expectations and variances are:

$$\begin{aligned}
\mathrm{E}[\mu|\gamma] &= -\frac{m_{10}+m_{11}\cdot\gamma+m_{12}\cdot\gamma^2}{2(m_{20}+m_{21}\cdot\gamma+m_{22}\cdot\gamma^2)},\\
\mathrm{Var}[\mu|\gamma] &= -\frac{1}{2(m_{20}+m_{21}\cdot\gamma+m_{22}\cdot\gamma^2)},\\
\mathrm{E}[\gamma|\mu] &= -\frac{m_{01}+m_{11}\cdot\mu+m_{21}\cdot\mu^2}{2(m_{02}+m_{12}\cdot\mu+m_{22}\cdot\mu^2)},\\
\mathrm{Var}[\gamma|\mu] &= -\frac{1}{2(m_{02}+m_{12}\cdot\mu+m_{22}\cdot\mu^2)}.
\end{aligned} \tag{7}$$

The distribution with density of the form (6) is an eight-parameter family of densities. The coefficient m_{00} is a normalizing constant that is determined by the other coefficients m and the requirement that the density should integrate to 1. Additionally, some restrictions on the coefficients m should be considered to ensure a positive value for the variances. This point is well illustrated by Arnold, Castillo and Sarabia (2001b).

This family of prior densities is very flexible and contains, as particular cases, many other distributions similar to that proposed in Bayesian literature for cost-effectiveness analysis. Thus, this family represents a significant extension to the usual priors considered. Its interest is twofold. Firstly, due to its conditioned-conjugancy property, it is very easy to simulate MCMC samples from posterior densities using Gibbs sampling. The practical application of the procedure presented in this paper is in accordance with Winkler (2001), as regards ease of use and ready acceptance, bearing in mind that the factors of expertise and prior knowledge can be incorporated into the computations (Malakoff, 1999). Secondly, this class of prior distributions contains a huge catalogue of highlighted prior densities (Spiegelhalter*et al.*, 1994 and Spiegelhalter, Myles, Jones and Abrams, 2000a). For instance, if we are willing to accept improper priors, then conditions for the above parameters (m_{00} among others) are not required. Sceptical priors about treatment effects are also easily elicited by making $E[\mu|\gamma]$ equal to zero and allowing a high degree of spread using the expression of the variances.

Thus, we encounter a great variety of distributions for different values of the m parameters. Some of these distributions are markedly different from classical bivariate normal densities. We now show the values of the m parameters for some particular cases.

- **Independence:** If we assume prior independence between the mean of the effectiveness (μ) and the mean of the costs (γ) for a given treatment, the conditional distributions do not depend on the other parameter, and the conditional expectations and variances will be of the form:

$$\begin{aligned}
\mathrm{E}[\mu|\gamma] &= \mathrm{E}[\mu] = -\frac{m_{10}}{2\cdot m_{20}},\\
\mathrm{Var}[\mu|\gamma] &= \mathrm{Var}[\mu] = -\frac{1}{2\cdot m_{20}},\\
\mathrm{E}[\gamma|\mu] &= \mathrm{E}[\gamma] = -\frac{m_{01}}{2\cdot m_{02}},\\
\mathrm{Var}[\gamma|\mu] &= \mathrm{Var}[\gamma] = -\frac{1}{2\cdot m_{02}}.
\end{aligned} \tag{8}$$

Thus, the conditions for independence are that the m's satisfy the following conditions:

$$m_{11} = m_{12} = m_{21} = m_{22} = 0, \quad m_{20} < 0, \quad m_{02} < 0. \tag{9}$$

- **Bivariate normal distribution**

  Another important case of bivariate distribution with normal conditional is that of the bivariate normal distribution. The bivariate normal prior distribution in Bayesian CEA was proposed by O'Hagan *et al.* (2001) and it is included as a particular case of the conditionally specified prior.
  For the terms μ and γ this can be expressed as

$$\pi(\mu,\gamma|\delta_\mu,\delta_\gamma,\sigma_\mu,\sigma_\gamma,\rho) = \frac{1}{2\pi\sigma_\mu\sigma_\gamma\sqrt{1-\rho^2}}\exp\left\{\frac{Q}{2(1-\rho^2)}\right\},$$

  where σ_μ and σ_γ are the expectations of the mean effectiveness and mean cost respectively, σ_μ and σ_γ are the standard deviation, ρ is the Spearman rho correlation coefficient, and Q is the quadratic expression

$$Q = \frac{(\mu-\delta_\mu)^2}{\sigma_\mu^2} - \frac{2\rho(\mu-\delta_\mu)(\gamma-\delta_\gamma)}{\sigma_\mu\sigma_\gamma} + \frac{(\gamma-\delta_\gamma)^2}{\sigma_\gamma^2}.$$

  The conditional distributions are normal with mean and variance

$$\begin{aligned}
\mathrm{E}(\mu|\gamma) &= \delta_\mu + \frac{\rho\sigma_\mu}{\sigma_\gamma}(\gamma-\delta_\gamma),\\
\mathrm{Var}(\mu|\gamma) &= \sigma_\mu^2(1-\rho^2),\\
\mathrm{E}(\gamma|\mu) &= \delta_\gamma + \frac{\rho\sigma_\gamma}{\sigma_\mu}(\mu-\delta_\mu),\\
\mathrm{Var}(\gamma|\mu) &= \sigma_\gamma^2(1-\rho^2).
\end{aligned} \tag{10}$$

The prior information can be elicited from expressions (7) and (10). Thus, the condition for the bivariate normal distribution is that the m's satisfy the following conditions (Arnold *et al.*, 2001a,b).

$$m_{12} = m_{21} = m_{22} = 0, m_{20} < 0, m_{02} < 0 \quad \text{and} \quad m_{11}^2 < 4m_{02}m_{20}. \tag{11}$$

Of course the use of conditional normal distributions is not the only way to elicit a bivariate normal distribution. In this sense Sarmanov (1966) and Ting Lee (1996) propose a family of bivariate distributions that can be elicited taking into account the marginal distributions.

- **A more general case:**

The improvement obtained from the use of conditionally specified priors is the wide range of prior information that may be elicited. For example, there are some combinations of m's that have non-normal marginal densities. In particular, bimodal or even trimodal densities may be encountered. These distributions must satisfy the conditions for integrability of (6) (Gelman and Meng, 1991, Arnold *et al.*, 2000, Arnold, Castillo and Sarabia, 2001a).

$$m_{22} < 0, \quad 4m_{22}m_{02} > m_{12}^2, \quad 4m_{22}m_{20} > m_{21}^2. \tag{12}$$

However, there is a price to be paid for the flexibility of our prior structure, namely that there are eight hyperparameters to assess. Given the difficulties of eliciting a high-dimensional joint probability distribution, we concentrate on eliciting some important summaries of the distribution, such as means and variances. We recommend the method for matching conditional moments proposed by Arnold, Castillo and Sarabia (1998). For a conditionally specified prior such as (6-7), we can try to match conditional moments, whose approximate values will be supplied by expert opinion. In our analysis, at least eight conditional moments are needed to determine all the hyperparameters. However, this might not be enough to determine the prior information and so it is preferable for the expert to supply more than eight conditional moments. We recognize that it is unlikely that such prior values will be consistent and what we propose is to select a prior of the form (6) that will have conditional moments that are minimally disparate from those provided a priori by the experts.

Let us assume that prior assessed values for the conditional means and variances of the effectiveness and cost are obtained for several different given values of the cost and effectiveness, respectively.

$$\begin{aligned} \mathrm{E}[\mu|\gamma_{p_1}] &= e_{p_1} \quad \forall p_1 = 1,2,\ldots,P_1. \\ \mathrm{Var}[\mu|\gamma_{p_2}] &= var(e)_{p_2} \quad \forall p_2 = 1,2,\ldots,P_2. \\ \mathrm{E}[\gamma|\mu_{p_3}] &= c_{p_3} \quad \forall p_3 = 1,2,\ldots,P_3. \\ \mathrm{Var}[\gamma|\mu_{p_4}] &= var(c)_{p_4} \quad \forall p_4 = 1,2,\ldots,P_4. \end{aligned} \tag{13}$$

where $P_1 + P_2 + P_3 + P_4 \geq 8$.

A unique solution for this system of equations is unlikely to be possible for any choice of the eight hyperparameters. A possible solution is to allow any deviance between the prior conditional moment and the knowledge of the expert. We define as the objective function the sum of the squared deviances (Arnold, Castillo and Sarabia, 1999). The hyperparameters are obtained by minimizing the objective function subject to constraints (12). A LINGO® code containing the procedure used in this article is available from the authors upon request. The prior distribution obtained must be checked by the experts so as not to obtain local minima in the optimization.

The choice of subjective priors is thus a difficult one, and requires the expert to take into account both psychological and behavioural aspects in order to obtain a coherent prior distribution (Baranski and Petrusic, 1994; Yates, 1990; Yaniv, Yates and Smith, 1991; among others). On the one hand, psychological studies have shown how well subjects make estimates and how different techniques of elicitation may produce different responses (Winkler, 1967; Staël von Holstein, 1970. An excellent review of this question was performed by Hogarth, 1975). Furthermore, many pioneering empirical studies (Kahneman and Tversky, 1972; Chesley, 1978; among others) have shown that training and maturity help an expert quantify prior probabilities.

Systematic methods of elicitation are presented in Kadane, Dickey, Winkler, Smith and Peters (1980), Garthwaite, Kadane and O'Hagan (2005) and a recent review of the question was provided by O'Hagan *et al.* (2006). We present a plausible alternative procedure from which it may be realistic to expect the elicitation of the (conditioned) prior mean and variance or other quantities; the specification based conditioned distribution theory may then be used to obtain a full specification of the prior distribution. Inspired by Berger (1994), we propose to use a class of plausible priors to ensure that as many reasonable priors as possible are included. Such a class does not require a strong mathematical training to be elicited and the priors are computationally manageable.

One practical situation where this more general prior distribution can be useful is that of the bimodal case. A bimodal distribution (or in general multimodal distribution) typically indicates that the distribution is in fact the sum of two or more different distributions, each with a single notable peak. Suppose that the treatment involves some risk, and there is a probability that complications

may appear. In that case, the effectiveness could be lower and the costs higher (McIntosh, Ramsey, Berry and Urban, 2001; Viviane and Barkun, 2008). If it is possible to distinguish and to record which patients suffer complications during the study, it would be plausible to propose as the likelihood of the data a mixture of bivariate normal distributions where the weight of each distribution is the probability of complications (Negrín and Vázquez-Polo, 2006). However it is not often easy to define a complication. Two possible solutions would be either to fix an arbitrary threshold cost (or effectiveness) to define the complication, or to approximate the probability of a complication using finite-mixture distributions (Diebolt and Robert, 1994). Conditionally specified distribution can be useful when the presence of complications is not clearly delimited. In this case the prior information can be modelled by a bimodal bivariate distribution, using conditionally specified prior distributions.

3 An example with real data

The data used in this section were obtained from a real clinical trial developed in 1999 in which a comparison was made between various highly active antiretroviral treatment protocols applied to asymptomatic HIV patients (Pinto, López, Badía, Corna and Benavides, 2000).

We only considered the direct costs (of drugs, medical visits and diagnostic tests), and as the effectiveness we considered the improvement in the quality of life, measured using the visual analogue scale (VAS) of the EQ-5D instrument (Brooks, 1996). In particular, we used the variation in the VAS by the end of the study. Cost and effectiveness values were recorded six months after the beginning of the study.

In this exercise, two three-way treatment protocols were compared. The first of these (d4T + 3TC + IND) combined the drugs estavudine (d4T), lamivudine (3TC) and indinavir (IND); the second treatment protocol (d4T + ddl + IND) combined estavudine (d4T), didanosine (ddl) and indinavir (IND).

Table 1 summarizes the statistical data. The d4T + ddl + IND treatment was more costly than the d4T + 3TC + IND treatment, by an average of 164.82 euros. The d4T + ddl + IND treatment was on average more effective, with an improvement in the patients' quality of life of 4.94 units, while those who were given the d4T + 3TC + IND treatment only experienced a VAS improvement of 4.56 units.

Table 1: *Statistical summary of costs (in thousands of euros) and effectiveness (change in VAS).*

	d4T + 3TC + IND		d4T + ddl + IND	
Statistical measure	Cost	Change in VAS	Cost	Change in VAS
Mean	7.142	4.56	7.307	4.94
s.d.	0.001573	15.17	0.001720	13.98
n	$n_0 = 268$		$n_1 = 93$	

For a fully Bayesian analysis, priors for the parameters of interest must be specified. Prior information was obtained from three experts who participated in the study and reflects the reasoning behind the design of the trial. The elicitation method was implemented in an interactive computer program. The computer displays assessment questions and the expert types in answers that reflect his opinion. At any point in the elicitation process, the expert can review the coherence of his probability judgments. Prior distributions derived from experts' consensus are displayed graphically to be reviewed.

Our Bayesian experiment requires the elicitation of normal distributions. A univariate normal distribution is characterized by two parameters, the mean and the variance. The mean (which coincides with the median) and the first and third quartile were requested of the experts in an elicitation process to obtain the prior mean and variance of the parameters of interest. Kadane and Wolfson (1998) suggest that the expert is only comfortable providing the mean and quartiles. Normal distributions were fitted to similar fitting procedures using percentile judgements in Cooke and Slijkhuis (2003), Denham and Mengersen (2007) and Kennedy, Anderson, O'Hagan, Lomas, Woodward, Gosling and Heinemeyer (2008).

- Independence

  The first analysis shows the independence case. For the purpose of this analysis, we took the design of the study to imply prior expectations for the parameters of interest. The experts' expectations show an average of 4.5 units of effectiveness for the control treatment (d4T + 3TC + IND), with a prior variance of 2.25. For the same treatment, the design anticipates an average cost of 5000 euros, with a variance of 4. The value of the m parameters is calculated directly, in the knowledge of the prior mean and variance of effectiveness and cost. For this prior information, the values are:

$$m_{01} = 1.25, m_{02} = -0.125, m_{10} = 2, m_{11} = 0,$$
$$m_{12} = 0, m_{20} = -0.2222, m_{21} = 0, m_{22} = 0.$$

  The elicitation process is very similar for the new treatment (d4T + ddl + IND). In this case, the experts considered this treatment to be less effective, with an average of 4 units of effectiveness and a prior variance of 2.5. They also expected it to be more expensive, with a prior mean cost of 6000 euros, and a variance of 6.25. The values of the m parameters for this treatment are

$$m_{01} = 0.96, m_{02} = -0.08, m_{10} = 1.6, m_{11} = 0,$$
$$m_{12} = 0, m_{20} = -0.2, m_{21} = 0, m_{22} = 0.$$

  We assume a diffuse prior distribution for the variance-covariance matrix Σ_i. Under the assumption of noninformative priors, we set $A_0 = A_1 = \text{diag}(1,1)$, $f_0 = f_1 = 2$,

where diag(a_i) is the $n \times n$ diagonal matrix with a_i elements. This assumption is repeated in the following analysis.

Figure 1 shows the contour plot of the joint distribution of the prior information of effectiveness and cost for each treatment, and the contour plot of the joint distribution of the prior incremental effectiveness and cost between treatments.

The posterior distribution was simulated using WinBUGS (Spiegelhalter, Thomas and Best, 2000b). A total of 100000 iterations were carried out (after a burn-in period of 50000 simulations). Convergence was evaluated for all parameters using several tests provided within the WinBUGS Convergence Diagnostics and Output Analysis software (CODA). The constant m_{00} is not required to ensure convergence.

Table 2 shows the posterior analysis for the independence case. The posterior incremental effectiveness is estimated at -0.02928 units with a standard deviation of 1.328. The incremental cost is estimated at 0.162 units.

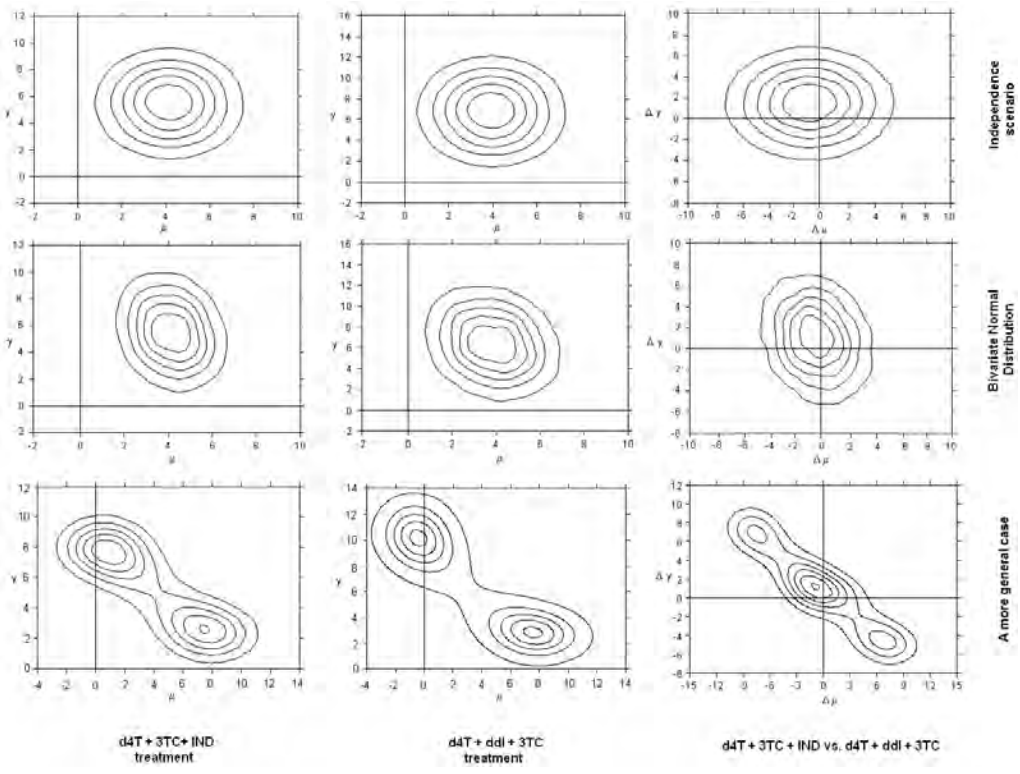


Figure 1: *Contour plots of the joint prior distribution of μ and γ.*

Table 2: *Posterior moments: mean and standard deviation.*

	Independence	Bivariate-Normal distribution	Bimodal case
μ_0	4.540 (0.7911)	4.45 (0.7862)	3.359 (0.777)
γ_0	7.137 (0.09542)	7.137 (0.0952)	7.128 (0.09661)
μ_1	4.507 (1.069)	4.422 (1.06)	2.152 (0.8683)
γ_1	7.302 (0.1784)	7.301 (0.1784)	7.293 (0.1824)
$\Delta\mu$	-0.02956 (1.328)	-0.02397 (1.318)	-1.207 (1.165)
$\Delta\gamma$	0.1628 (0.2028)	0.1613 (0.2023)	0.1643 (0.2063)

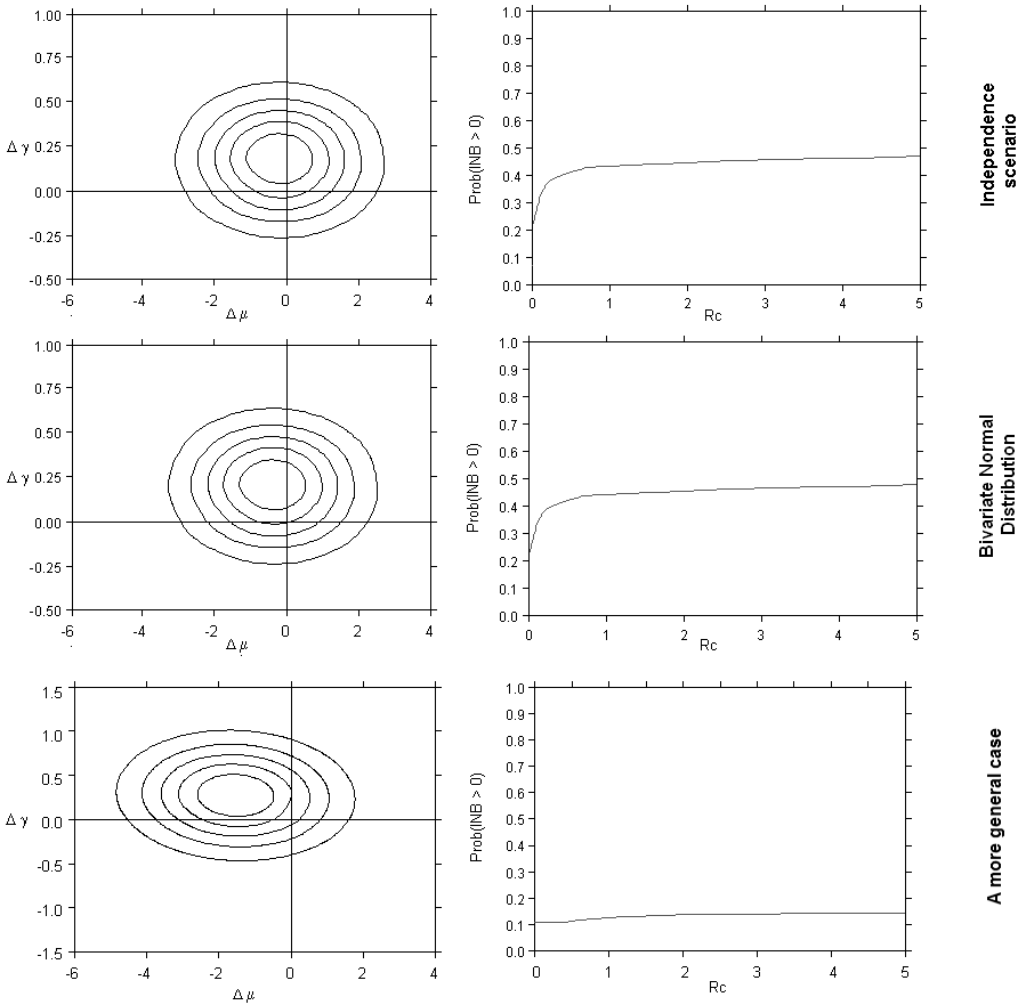


Figure 2: *Contour plots of the joint posterior distribution of μ and γ and the cost-effectiveness acceptability curves.*

Figure 2 shows the cost-effectiveness plane and the cost-effectiveness acceptability curve. It is apparent that the treatment (d4T + ddl + IND) will never be preferable to the treatment (d4T + 3TC + IND), as the probability of a positive INB is always below 50%.

- Bivariate normal distribution:

  The previous analysis was repeated under the assumption of a correlation between cost and effectiveness for each treatment. We asked the experts to assess the correlation directly by specifying a value between -1 and 1. Although many researchers have suggested that the direct assessment of moments is a poor method of quantifying opinion, Clemen, Fischer and Winkler (2000) found that this method performed best for eliciting a correlation. The experts considered a prior correlation of $\rho = -0.2$. By incorporating this information into the prior information described in the previous subsection, we calculated the following prior parameters:

$$m_{01} = 1.6146, m_{02} = -0.1302, m_{10} = 2.4306, m_{11} = -0.0694,$$
$$m_{12} = 0, m_{20} = -0.2315, m_{21} = 0, m_{22} = 0,$$

  for the (d4T + 3TC + IND) treatment, and

$$m_{01} = 1.2108, m_{02} = -0.0833, m_{10} = 1.9829, m_{11} = -0.0527,$$
$$m_{12} = 0, m_{20} = -0.2083, m_{21} = 0, m_{22} = 0,$$

  for the (d4T + ddl + IND) treatment.

  Figure 1 includes the joint distribution of the prior information on the effectiveness and cost for each treatment, and the joint distribution of the prior incremental effectiveness and cost between treatments. It also shows the negative correlation between effectiveness and cost. Figure 2 shows the cost-effectiveness plane and the cost-effectiveness acceptability curve. The results are similar to those reached in the independence case. The treatment (d4T + 3TC + IND) is always preferred for any willingness to pay.

- A more general case:

  This example shows a prior bimodal density for the effectiveness and cost of each treatment. The experts agreed that effectiveness and cost depend on the presence of complications during the treatment, mainly due to the existence of concomitant illnesses. In this paper, the dependence between effectiveness and cost is specified by a conditional probability among the elicitation variables of interest. We asked the experts for the conditional median and first and third quartiles to elicit the mean and the variance of the normal distributions.

If there were no complications during the study, the mean effectiveness of the (d4T + 3TC + IND) treatment would be close to 8 units, and the mean cost would be about 2000 euros. The mean effectiveness decreases to 1, and the mean cost increases to 8000 euros with the presence of complications.

The presence of complications has more costly consequences for the (d4T + ddl + 3TC) treatment. The mean cost increases to 10000 euros and the mean effectiveness is reduced to 0. Under favourable conditions, the mean effectiveness is about 8 units and the mean cost is about 3000 euros.

This prior information was elicited through a conditionally specified prior distribution, compiling information about the conditional moments. Descriptions of the conditions were given to the experts in written form. An example of a verbal statement of a conditional event (Garthwaite and Al-Awadhi, 2001) is

- Suppose that a large number of patients are examined and their average cost is 2000 euros. What is your median estimate of their effectiveness?
- Consider the situation in which we know that your median value is true. In the light of this, assess your quartiles for effectiveness.

Table 3 shows the conditional moments employed in the elicitation process.

Table 3: *Prior conditional moments.*

Moment	Condition	(d4T+3TC+IND)	(d4T+ddl+IND)
E($\mu\|\gamma$)	$\gamma = 2$	8	8
	$\gamma = 5$	4	3.5
	$\gamma = 8$	2	0
Var($\mu\|\gamma$)	$\gamma = 2$	2.5	2.5
	$\gamma = 5$	2.25	2
	$\gamma = 8$	1.75	1.5
E($\gamma\|\mu$)	$\mu = 0$	8	10
	$\mu = 4$	5	6
	$\mu = 8$	2	3
Var($\gamma\|\mu$)	$\mu = 0$	6	4.5
	$\mu = 4$	4.5	7.5
	$\mu = 8$	2.5	7

By using this prior information we can calculate the values of the hyperparameters, applying them to the optimization problem explained in the previous section:

$$m_{01} = 9.3931, m_{02} = -0.6198, m_{10} = 8.1442, m_{11} = -1.8114,$$
$$m_{12} = 0.1012, m_{20} = -0.5241, m_{21} = 0.1412, m_{22} = -0.0147$$

for the (d4T + 3TC + IND) treatment, and

$$m_{01} = 7.5074, m_{02} = -0.4014, m_{10} = 67.0746, m_{11} = -1.0390,$$
$$m_{12} = 0.0231, m_{20} = -0.4792, m_{21} = 0.1209, m_{22} = -0.0165$$

for the (d4T + ddl + IND) treatment.

Figure 1 shows the joint distribution of the prior information on the effectiveness and cost of each treatment, together with the joint distribution of the prior incremental effectiveness and cost between treatments. There was found to be a bimodal joint distribution for cost and effectiveness. This joint distribution, and the marginal distributions of effectiveness and cost were shown to the experts to assess the adequacy of the elicitation.

It is important to note that the mean of the marginal distributions of effectiveness and cost for both treatments coincides with the prior mean elicited in the "independent" section. However, this more general model opens up a wide range of possibilities for incorporating different prior beliefs far removed from those of the conventional bivariate normal distribution.

Figure 2 shows the measures used to take decisions. The analysis shows a preference for the treatment (d4T + 3TC + IND) for all the scenarios. In fact, the CEAC is always lower than the critical value 50%. It is important to point out that, although we have considered similar prior means of effectiveness and costs, the uncertainty about the right decision is different for the independence scenario, the bivariate normal distribution and the bimodal prior distribution. If we considered a willingness to pay of 5 euros, the probability of preferring the (d4T + 3TC + IND) for the first two scenarios is only 52%. This probability increases to 85% for the more general case. This is due to the fact that the latter model includes in the analysis the prior information that any complication arising during the treatment would have more important consequences with the (d4T + ddl + 3TC) treatment than with the control treatment.

4 Conclusions and discussion

The Bayesian approach allows the incorporation of prior information. In a fully Bayesian analysis, the procedures used to elicit expert opinion are an active research issue. This paper studies the use of a general family of prior distributions for the mean of the effectiveness and cost. In particular, we assume that the conditional density of the mean effectiveness for a given mean cost and the conditional density of the mean cost for a given mean effectiveness are both normal.

The improvement gained over the use of conditionally specified priors is the wide range of prior information that may be elicited. Prior information from more than one source, or different structures of effectiveness and costs depending on whether complications occur, are some cases whereby a conventional bivariate prior distribution is not enough to specify the prior information. Conventional cases, such as the independence case and bivariate prior information, are included as particular cases of this more general analysis. The posterior distribution is easily simulated using MCMC techniques (Gelman, Carlin, Stern and Rubin, 1995; Gilks, Richardson and Spiegelhalter, 1996; Gamerman and Lopes, 2006).

The practical application shows the sensitivity of the results to the prior distribution. The more general case, in which experts provide a bimodal prior distribution for mean effectiveness and mean cost, the probability of preference for the conventional treatment (d4T + 3TC + IND) varies from 85% to 89% in a willingness to pay range of $R_c \in (0,5)$. For the more conventional prior structure, bivariate or independent normal distributions, this probability varies from 65% to 52% for the same range of R_c.

However, this methodology also present some disadvantages. Psychological research has shown that conditional assessments can be affected by biases such as conservatism (Edwards and Phillips, 1964) and, intuitively, making assessments conditionally on hypothetical data is a more difficult task than making unconditional assessments. Besides, the large number of parameters present in high-dimensional conditionally specified priors is the source of their flexibility but, in practice, poses elicitation problems. In this context, the sensitivity analysis may play a crucial role (Stevens and O'Hagan, 2002). In our opinion, conditionally specified priors are not a panacea but certainly, for many classical data analysis situations, they offer a manageable and more flexible alternative to the usual, rather restrictive, priors.

Acknowledgments

We wish to thank the referees for their comments. Research partially funded by SEJ2006-12685 (MEC, Spain) and ECO2009-14152 (Ministerio de Ciencia e Innovación, Spain). We thank Xavier Badía for allowing us to use his data archive. All responsibility for the further analysis of the data, of course, is ours alone.

References

Arnold, B., Castillo, E. and Sarabia, J. (1998). Bayesian analysis for classical distributions using conditionally specified priors. *Sankhya-The Indian Journal of Statistics*, 60, 228-245.

Arnold, B., Castillo, E. and Sarabia, J. M. (1999). *C*onditional Specification of Statistical Models. Springer-Verlag, New York.

Arnold, B., Castillo, E. and Sarabia, J. (2001a). Bayesian inference using conditionally specified priors. *Handbook of Applied Econometrics and Statistical Inference*. Marcel Dekker, New York.

Arnold, B., Castillo, E. and Sarabia, J. (2001b). Conditionally specified distributions: an introduction. *Statistical Science*, 16, 249-274.

Arnold, B., Castillo, E., Sarabia, J. and González-Vega, L. (2000). Multiple modes in densitites with normal conditionals. *Statistics and Probability Letters*, 49, 355-363.

Baranski, J. V. and Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception and Psychophysics*, 55, 412-428.

Berger, J. O. (1994). An overview of robust Bayesian analysis (with discussion). *Test*, 3, 5-124.

Briggs, A. H. (1999). A Bayesian approach to stochastic cost-effectiveness analysis. *Health Economics*, 8, 257-261.

Brooks, R. and EuroQol group (1996). EuroQol: The current state of play. *Health Policy*, 37, 53-72.

Castillo, E. and Galambos, J. (1989). Conditional distributions and the bivariate normal distribution. *Metrika*, 36, 209-214.

Chesley, G. R. (1978). Subjective probability elicitation techniques: A performance comparison. *Journal of Accounting Research*, 16, 225-241.

Clemen, R. T., Fischer, G. W. and Winkler, L. R. (2000). Assessing dependence: Some experimental results. *Management Science*, 46, 1100-1115.

Cooke, R. M. and Slijkhuis, K. A. (2003). *Expert judgement in the uncertainty analysis of dike ring failure frequency.* In W. R. Blischke and D. N. Prabhakar Murthy (Eds.). *Case Studies in Reliability and Maintenance* (331-352). Wiley, Chichester.

Denham, R. and Mengersen, K. (2007). Geographically assisted elicitation of expert opinion for regression models. Bayesian Analysis. *Journal of the Royal Statistical Society, B*, 2, 99-136.

Diebolt, J. and Robert, C. P. (1994). Estimation of Finite Mixture Distributions through Bayesian Sampling. *Journal of the Royal Statistical Society, B*, 56, 363-375.

Edwards, W. and Phillips, L. D. (1964). *Man as transducer for probabilities in Bayesian command and control systems.* In G. L. Bryan and M. W. Shelley (Eds.). *Human Judgements and Optimality*. Wiley, New York.

Fenwick, E., Palmer, S., Claxton, K., Sculpher, M., Abrams, K. and Sutton, A. (2006). An iterative Bayesian approach to health technology assessment: application to a policy of preoperative optimization for patients undergoing major elective surgery. *Medical Decision Making*, 26, 480-496.

Fryback, D., Chinnis, J. and Ulvila, J. (2001). Bayesian cost-effectiveness analysis. An example using the GUSTO trial. *International Journal of Technology Assessment in Health Care*, 17, 83-97.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall, London.

Garthwaite, P. H. and Al-Awadhi, S. A. (2001). Non-conjugate prior distribution assessment for multivariate normal sampling. *Journal of the Royal Statistical Society, Series B*, 63, 95-110.

Garthwaite, P. H., Kadane, J. B. and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100, 680-701.

Gelman, A. and Meng, X. (1991). A note on bivariate distributions that are conditionally normal. *American Statistics*, 45, 125-126.

Gelman, R., Carlin, J., Stern, H. S. and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Heitjan, D. (1997). Bayesian interim analysis of phase II cancer trials. *Statistics in Medicine*, 16, 1791-1802.

Heitjan, D., Moskowitz, A. and William, W. (1999). Bayesian estimation of cost-effectiveness ratios from clinical trials. *Health Economics*, 8, 191-201.

Hogarth, R. M. (1975). Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association*, 70, 271-294.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S. and Peters, S. C. (1980). Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, 75, 845-854.

Kadane, J. B. and Wolfson, L. J. (1998). Experiences in elicitation. *The Statistician*, 47, 3-19.

Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430-454.

Kennedy, M., Anderson, C., O'Hagan, A., Lomas, M., Woodward, I., Gosling, J. P. and Heinemeyer, A. (2008). Quantifying uncertainty in the biospheric carbon flux for England and Wales. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171, 109-135.

Laska, E., Meisner, M. and Siegel, C. (1997). Statistical inference for cost-effectiveness ratios. *Health Economics*, 6, 229-242.

Leal, J., Wordsworth, S., Legood, R. and Blair, E. (2007). Eliciting expert opinion for economic models: An applied example. *Value in Health*, 10, 195-203.

Löthgren, M. and Zethraeus, N. (2000). Definition, interpretation and calculation of cost-effectiveness acceptability curves. *Health Economics*, 9, 623-630.

Malakoff, D. (1999). Bayes offers a new way to make sense of numbers. *Science*, 286, 1460-1464.

McIntosh, M. W., Ramsey, S. D., Berry, K. and Urban, N. (2001). Parameter solicitation for planning cost effectiveness studies with dichotomous outcomes. *Health Economics*, 10, 53-66.

Negrín, M. and Vázquez-Polo, F. J. (2006). Bayesian cost-effectiveness analysis with two measures of effectiveness: the cost-effectiveness acceptability plane. *Health Economics*, 15, 363-372.

O'Hagan, A. and Stevens, J. (2001). A framework for cost-effectiveness analysis from clinical trial data. *Health Economics*, 10, 303-315.

O'Hagan, A., Stevens, J. and Montmartin, J. (2001). Bayesian cost-effectiveness analysis from clinical trial data. *Statistics in Medicine*, 20, 733-753.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, J., Oakley, J. and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley and Sons, Chicester.

Pinto, J., López, C., Badía, X., Corna, A. and Benavides, A. (2000). Análisis coste-efectividad del tratamiento antirretroviral de gran actividad en pacientes infectados por el VIH asintomáticos. *Medicina Clínica*, 114, 62-67.

Saltelli, A., Chan, K. and Scott, E. M. (2000). *Sensitivity Analysis*. Wiley, Chichester.

Sarabia, J. M. and Gómez-Déniz, E. (2008). Construction of multivariate distributions: a review of some recent results. *SORT*, 32, 3-36.

Sarabia, J. M., Castillo, E., Gómez-Déniz, E. and Vázquez-Polo, F. J. (2005). A class of conjugate priors for log-normal claims based on conditional specification. *Journal of Risk and Insurance*, 72, 479-495.

Sarmanov, O. V. (1966). Generalized normal correlation and two-dimensional Frechet classes. *Doklady (Soviet Mathemathics)*, 168, 596-599.

Smith, M. K. and Marshall, S. (2006) A Bayesian design and analysis for dose-response using informative prior information. *Journal of Biopharmaceutical Statistics*, 16, 695-709.

Spiegelhalter, D., Feedman, L. and Parmar, M. (1994). Bayesian approaches to randomized trials (with discussion). *Journal of the Royal Statistical Society, Series A*, 157, 357-416.

Spiegelhalter, D. J., Myles, J. P., Jones, D. R. and Abrams, K. R. (2000a). Bayesian methods in health technology assessment: a review. *Health Technol Assess*, 4.

Spiegelhalter, D., Thomas, A. and Best, N. (2000b). WinBUGS version 1.3. User manual, Biostatistics Unit, Cambridge Medical Research Council.

Staël von Holstein, C. A. S. (1970). Measurement of subjective probability. *Acta Psychologica*, 34, 146-159.

Stevens, J. and O'Hagan, A. (2002). Incorporation of genuine prior information in cost-effectiveness analysis of clinical trial data. *International Journal of Technology Assessment in Health Care*, 18, 782-790.

Stinnett, A. and Mullahy, J. (1998). Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making*, 18, S65-S80.

Tambour, M., Zethraeus, N. and Johannesson, M. (1998). A note on confidence intervals in cost-effectiveness analysis. *International Journal of Technology Assessment in Health Care*, 14, 467-471.

Ting-Lee, M. L. (1996). Properties and applications of the Sarmanov family of bivariate distributions. *Communications Statistics: Theory and Methods*, 25, 1207-1222.

Vázquez-Polo, F. J. and Negrín, M. (2004). Incorporating patients' characteristics in cost-effectiveness studies with clinical trial data: a flexible Bayesian approach. *SORT*, 28, 87-108.

Viviane, A. and Barkun, N. A. (2008). Estimates of costs of hospital stay for variceal and nonvariceal upper gastrointestinal bleeding in the United States. *Value in Health*, 11, 1-3.

Willan, A. and O'Brien, B. (1996). Confidence intervals for cost-effectiveness ratios: an application of Fielle's theorem. *Health Economics*, 5, 297-305.

Winkler, R. L. (1967). The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association*, 62, 776-800.

Winkler, R. L. (2001). Why Bayesian analysis hasn't caught on in healthcare decision making. *International Journal of Technology Assessment in Health Care*, 17, 56-66.

Yaniv, I., Yates, J. F. and Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, 110, 611-617.

Yates, J. F. (1990). *Judgment and decision making*. Prentice-Hall, Englewood Cliffs.

SORT 33 (2) July-December 2009, 213-232

How much Fisher information is contained in record values and their concomitants in the presence of inter-record times?

Morteza Amini and Jafar Ahmadi*

Abstract

It is shown that, although the distribution of inter-record time does not depend on the parent distribution, Fisher information increases when inter-record times are included. The general results concern different classes of bivariate distributions and propose a comparison study of the Fisher information. This study is done in situations in which the univariate counterpart of the underlying bivariate family belongs to a general continuous parametric family and its well-known subclasses such as location-scale and shape families, exponential family and proportional (reversed) hazard model. We derived some explicit formulas for the additional information of record time given records and their concomitants (bivariate records) for some classes of bivariate distributions. Some common distributions are considered as examples for illustrations and are classified according to this criterion. A simulation study and a real data example from bivariate normal distribution are considered to study the relative efficiencies of estimator based on bivariate record values and inter-record times with respect to the corresponding estimator based on iid sample of the same size and bivariate records only.

MSC: 62G30; 62B10; 62G32; 62B05.

Keywords: Bivariate family, hazard rate function, reversed hazard rate, location and scale families, proportional (reversed) hazard model, shape family.

1 Introduction

Let $\{(X_i, Y_i), i \geq 1\}$ be a sequence of bivariate random variables from a continuous distribution with the real valued parameter θ. Let $\{R_n, n \geq 1\}$ be the sequence of record values in the sequence of X's. Then the Y-variable associated with the X-value which is

* *Address for correspondence:* Department of Statistics, School of Mathematical Sciences, Ferdowsi University of Mashhad, P.O. Box 91775-1159, Mashhad, Iran. E-mail addresses: mort.amini@gmail.com (Morteza Amini), ahmadi-j@um.ac.ir (J. Ahmadi).

Received: June 2009
Accepted: October 2009

quantified as the nth record is called the concomitant of the nth record and is denoted by $R_{[n]}$. The most important use of concomitants of record values arises in experiments in which a specified characteristic's measurements of an individual are made sequentially and only values that exceed or fall below the current extreme value are recorded. So the only observations are bivariate record values, i.e., records and their concomitants. Such situations often occur in industrial stress, life time experiments, sporting matches, weather data recording and some other experimental fields.

Under certain regularity conditions, the Fisher information about the real parameter θ contained in a random variable X with density $f(x;\theta)$ is defined by (see, for example, Lehmann, 1989, p. 115), $I_X(\theta) = E\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right)^2 = -E\left(\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\right)$. The Fisher information plays an important role in statistical estimation and inference through the information (Cramér-Rao) inequality and its association with the asymptotic properties, especially the asymptotic variance of the maximum likelihood estimators. It can also be used to compute the variance of the estimator whose variance is equal to Cramér-Rao lower bound, i.e., $\delta(X)$, $\mathrm{Var}(\delta(X)) = (\frac{\partial}{\partial \theta}\mathrm{E}\delta(X))^2 / I_X(\theta)$. Abo-Eleneen and Nagaraja (2002) investigated some properties of Fisher information in an order statistic and its concomitant. Recently, Nagaraja and Abo-Eleneen (2008) considered bivariate censored samples and evaluated the Fisher information contained in a collection of order statistics and their concomitans.

Several authors have considered the amount of Fisher information in record data and have discussed its applications in inference [see, for example, Ahmadi and Arghami (2001, 2003), Hofmann and Nagaraja (2003), Balakrishnan and Stepanov (2005) and references therein]. However, the treatment of Fisher information contained in the bivariate record values is very limited. The question "How much information is contained in records and their concomitants about a specified parameter?" was addressed by Amini and Ahmadi (2007, 2008).

The time at which a record appears is called *record time*. There is no information, in record times themselves, about the sampling distribution, since for a continuous sampling distribution F, the joint distribution of record times does not depend on F (see, Arnold *et al.*, 1998, Section 2.5). Nevertheless, there is crucial information about F in the joint distribution of record times and record values. Actually, in the process of obtaining the bivariate record values, one usually observes the record times. So, it is worthwhile to use them, since they provide meaningful additional information. Ahmadi and Arghami (2003) and Hofmann (2004) presented some comparison results of Fisher information in univariate record values and record times with the Fisher information contained in the same number of random univariate observations. The aim of this paper is to investigate the amount of Fisher information in bivariate record values in the presence of inter-record times in some well-known bivariate classes of distributions. We have especially focused on the increment of Fisher information by considering inter-record times. We also study some estimation results based on bivariate record values and inter-record times.

The rest of paper is organized as follows. Section 2 contains some preliminaries and introduction to some classes of univariate and bivariate distributions. In Section 3, we establish some general results to compare the amount of the Fisher information contained in a set of the first n bivariate record values and inter-record times with a bivariate random sample of same size from the parent distribution. For each result, we give some examples for illustration. In Section 4, a simulation study and a real data example from bivariate normal distribution are also presented.

2 Preliminaries

Let $\{(X_i,Y_i), i \geq 1\}$ be a sequence of iid bivariate random variables with an absolutely continuous cumulative distribution function (cdf) $F_{X,Y}(x,y;\theta)$, where θ is a real valued parameter. The marginal probability density function (pdf) and cdf of X are denoted by $f_X(x;\theta)$ and $F_X(x;\theta)$, respectively. Furthermore, $h_X(x;\theta) = f_X(x;\theta)/\bar{F}_X(x;\theta)$ and $\tilde{h}_X(x;\theta) = f_X(x;\theta)/F_X(x;\theta)$ are the hazard rate and the reversed hazard rate functions of X, respectively, where $\bar{F}_X(x;\theta) = 1 - F_X(x;\theta)$.

The sequence of bivariate record values is defined as $(R_n, R_{[n]}) = (X_{T_n}, Y_{T_n})$, $n \geq 1$, where $T_1 = 1$ with probability one and for $n \geq 2$, $T_n = \min\{j : j > T_{n-1},\ X_j > X_{T_{n-1}}\}$.

An analogous definition deals with lower records and their concomitants. In this paper, we assume that the data available for study are records (upper or lower), inter-record times and their concomitants. Such data may be rewritten as $(R_1,\Delta_1,R_{[1]}), (R_2,\Delta_2,R_{[2]}), \ldots, (R_n,\Delta_n,R_{[n]})$, where $\Delta_i = T_{i+1} - T_i - 1$, $i = 1,2,\ldots,n-1$, $\Delta_n = 0$ are the number of trials needed to obtain new records. Let us denote

$$\mathbf{R}_n = (R_1,\ldots,R_n),\ \boldsymbol{\Delta}_n = (\Delta_1,\ldots,\Delta_n), \mathbf{C}_n = (R_{[1]},\ldots,R_{[n]}).$$

Suppose the observed data is $(r_1,\delta_1,s_1),\ldots,(r_n,\delta_n,s_n)$, then the joint pdf of the first n upper records and inter-record times is (see Arnold *et al.*, 1998, p. 169)

$$f_{(\mathbf{R}_n,\boldsymbol{\Delta}_n)}(\mathbf{r}_n,\boldsymbol{\delta}_n;\theta) = \prod_{i=1}^{n} f_X(r_i;\theta)\{F_X(r_i;\theta)\}^{\delta_i}. \tag{1}$$

Using (1) the joint pdf of records, inter-record times and their concomitants is given by

$$f_{(\mathbf{R}_n,\boldsymbol{\Delta}_n,\mathbf{C}_n)}(\mathbf{r}_n,\boldsymbol{\delta}_n,\mathbf{s}_n;\theta) = \prod_{i=1}^{n} f_{X,Y}(r_i,s_i;\theta)\{F_X(r_i;\theta)\}^{\delta_i}. \tag{2}$$

So, the conditional probability mass function of $\boldsymbol{\Delta}_n$ given $(\mathbf{R}_n,\mathbf{C}_n)$ is given by

$$f_{(\boldsymbol{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n)}(\boldsymbol{\delta}_n|\mathbf{r}_n,\mathbf{s}_n;\theta) = \prod_{i=1}^{n-1}[F_X(r_i;\theta)]^{\delta_i}\bar{F}_X(r_i;\theta). \tag{3}$$

In order to perform a comparison study, first let us consider some classes of univariate and bivariate distributions as follows:

$$\mathscr{F} = \{ f_{X,Y} : f_{Y|X} \text{ is free of } \theta \},$$

$$\mathscr{B} = \{ f_{X,Y} : f_{X,Y}(x,y;\theta) = a(\theta)b(x,y)\exp\{c(\theta)d(x,y)\}, a(\theta) > 0, b(x,y) > 0 \},$$

$$\mathscr{K} = \{ f_{X,Y} : f_{Y|X}(y|x) \text{ is in the form of } f_{X,Y} \text{ in } \mathscr{B} \},$$

$$\mathscr{C}_1 = \{ F_X : \bar{F}_X(x;\theta) = (\bar{G}(x))^{\alpha(\theta)} \},$$

$$\mathscr{C}_2 = \{ F_X : F_X(x;\theta) = (H(x))^{\beta(\theta)} \},$$

$$\mathscr{D}_i = \{ f_{X,Y} \in \mathscr{F} : F_X \in \mathscr{C}_i \},\ i = 1,2,$$

$$\mathscr{E}_i = \{ f_{X,Y} \in \mathscr{K} : F_X \in \mathscr{C}_i, \text{ with } c(\theta) = \alpha(\theta)I_1(i) + \beta(\theta)I_2(i) \},\ i = 1,2,$$

$$\mathscr{G} = \{ f_{X,Y} \in \mathscr{F} : f_X \in \mathscr{E} \},$$

$$\mathscr{H} = \{ f_{X,Y} \in \mathscr{K} : f_X \in \mathscr{E} \},$$

$$\mathscr{L}_{\mathscr{B}} = \{ f_{X,Y} \in \mathscr{B} :\ F_X(x;\theta) = F_0(x-\theta),\ \theta \in \mathbb{R} \text{ or } F_X(x;\theta) = F_1(\theta x),\ \theta > 0 \},$$

$$\mathscr{S}_{\mathscr{B}} = \{ f_{X,Y} \in \mathscr{B} :\ F_X(x;\theta) = F_1(x^{\theta}),\ \theta > 0,\ x > 0 \},$$

$$\mathscr{L}_{\mathscr{K}} = \{ f_{X,Y} \in \mathscr{K} :\ F_X(x;\theta) = F_0(x-\theta),\ \theta \in \mathbb{R} \text{ or } F_X(x;\theta) = F_1(\theta x),\ \theta > 0 \}$$

and

$$\mathscr{S}_{\mathscr{K}} = \{ f_{X,Y} \in \mathscr{K} :\ F_X(x;\theta) = F_1(x^{\theta}),\ \theta > 0,\ x > 0 \},$$

where $\alpha(\theta)$ and $\beta(\theta)$ are real positive functions, $G(x)$ and $H(x)$ are arbitrary continuous cdf's, free of θ, $\bar{G}(x) = 1 - G(x)$, $\mathscr{E}$ in $\mathscr{G}$ and $\mathscr{H}$ stands for the well-known exponential family, F_0 and F_1 are arbitrary cdf's, free of θ ($F_i(t) = F_X(t;i),\ i = 0,1$) and $\bar{F}_i(x) = 1 - F_i(x),\ i = 0,1$. Let $h_i(x)$ and $\tilde{h}_i(x),\ i = 0,1$ stand for the standard hazard rate and the reversed hazard rate functions of a random variable with pdf f_i and cdf $F_i,\ i = 0,1$, respectively.

Indeed, $\mathscr{C}_1$ and $\mathscr{C}_2$ stand for two well-known families of distributions in life-time experiments literature, the proportional hazard model and proportional reversed hazard model, respectively (see for example Lawless, 2003). Classes $\mathscr{B}$, $\mathscr{D}_1$ and $\mathscr{D}_2$ include several well-known distributions (see Amini and Ahmadi, 2008). We should emphasize that, although in the two classes $\mathscr{D}_1$ and $\mathscr{D}_2$, $f_{Y|X}$ is free of θ, f_Y may depend on it. In

fact by considering a single (X,Y), one would find X a sufficient statistic for θ. Since $\mathscr{C}_1$ and $\mathscr{C}_2$ are both subsets of $\mathscr{E}$, $\mathscr{D}_1$ and $\mathscr{D}_2$ are both subsets of $\mathscr{G}$.

It is clear that $\mathscr{L}_{\mathscr{B}} \subset \mathscr{B}$, $\mathscr{S}_{\mathscr{B}} \subset \mathscr{B}$ and $\mathscr{D}_i \subset \mathscr{G} \subset \mathscr{H} \subset \mathscr{B}, i = 1,2$. Note that in the functional form of $\mathscr{B}$, one may let $d(x,y,\eta) = 0$ and $a(\theta,\eta) = 1$ to obtain a form of $f_{Y|X}(y|x)$ that is free of θ.

We shall note that, although we have used bivariate upper records and times to obtain the results of this paper, corresponding results for bivariate lower records are derived and classified in Table 8.

The hazard rate function and the reversed hazard rate functions are important characteristics for the analysis of reliability data. A random variable X is said to be Increasing Hazard Rate (Reversed Hazard Rate), Decreasing Hazard Rate (Decreasing Reversed Hazard Rate) or Constant Hazard Rate (Constant Reversed Hazard Rate), and is denoted by IHR (IRHR), DHR (DRHR) or CHR (CRHR), if its hazard rate (reversed hazard rate) function is increasing, decreasing or constant, respectively.

3 Main results

Since reparameterizing $\theta = z(\gamma)$, for a differentiable $z(.)$, transforms the Fisher information of any data to $(\frac{\partial}{\partial\gamma}z(\gamma))^2 I_X(z(\gamma))$ (see Lehmann, 1989), we may assume throughout that $c(\theta) = \theta$. To prove the main results of this paper, we need the following lemma. The proof is easy and hence is omitted.

Lemma 1 *The pdf $f_{X,Y}(x,y;\theta)$ belongs to $\mathscr{B}$ with natural parameter θ ($c(\theta) = \theta$) if and only if $\frac{\partial^2}{\partial\theta^2}\log f_{X,Y}(x,y;\theta)$ does not depend on x and y.*

Note: Obviously, we have

$$I_{\mathbf{R}_n,\mathbf{\Delta}_n,\mathbf{C}_n}(\theta) = I_{\mathbf{R}_n,\mathbf{C}_n}(\theta) + I_{\mathbf{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\theta), \tag{4}$$

where $I_{\mathbf{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\theta) = I_{\mathbf{\Delta}_n|\mathbf{R}_n}(\theta)$ is indeed $\mathrm{E}_{\mathbf{R}_n}(I_{\mathbf{\Delta}_n|\mathbf{R}_n}(\theta))$. Hereafter, we will use the notation $I_{\mathbf{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\theta)$ instead of $\mathrm{E}_{\mathbf{R}_n}(I_{\mathbf{\Delta}_n|\mathbf{R}_n}(\theta))$.

Proposition 1 *Let $\{(X_i,Y_i), i \geq 1\}$ be a sequence of iid bivariate random variables with pdf $f_{X,Y}(x,y;\theta)$, then $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) \geq I_{\mathbf{R}_n,\mathbf{C}_n}(\theta)$, with equality while F_X is free of θ and the increment of Fisher information by considering inter-record times is given by*

$$I_{\mathbf{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\theta) = -\sum_{i=1}^{n-1}\mathrm{E}\left[\frac{F_X(R_i;\theta)}{F_X(R_i;\theta)}\frac{\partial^2}{\partial\theta^2}\log F_X(R_i;\theta) + \frac{\partial^2}{\partial\theta^2}\log\bar{F}_X(R_i;\theta)\right].$$

So $I_{\mathbf{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\theta) = 0$ when F_X is free of θ.

Proof From (4), we conclude that $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) \geq I_{\mathbf{R}_n,\mathbf{C}_n}(\theta)$. Using (3) and the fact that $\mathrm{E}(\delta_i|R_i) = F_X(R_i;\theta)/\bar{F}_X(R_i;\theta)$ along with the definition of Fisher information, the proof is complete. ■

The univariate case of Proposition 1 is obtained by Hofmann (2004).

Example 1 (Farlie-Gumbel-Morgenstern family of distributions) *Let*

$$f_{X,Y}(x,y;\theta) = f_X(x)\,f_Y(y)[1+\theta(1-2F_X(x))(1-2F_Y(y))], -1<\theta<1.$$

Amini and Ahmadi (2007) showed that for this family $I_{\mathbf{R}_n,\mathbf{C}_n}(\theta) > nI_{(X,Y)}(\theta)$. *However, since* F_X *is free of* θ, *Proposition 1 yields that* $I_{\mathbf{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\theta) = 0$. *So* $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) > nI_{(X,Y)}(\theta)$.

Theorem 1 *Suppose* $f_{X,Y}(x,y;\theta)$ *belongs to* $\mathscr{K}$ *and let* $l(x;\theta) = \dfrac{\partial^2}{\partial\theta^2}\log f_X(x;\theta)$. *Then*

(i) if $l(x;\theta)$ *is decreasing in* x *and* $F_X(x;\theta)$ *is strictly log-concave or log-linear in* θ, *then* $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) > nI_{(X,Y)}(\theta)$;

(ii) if $l(x;\theta)$ *is increasing in* x *and* $F_X(x;\theta)$ *is strictly log-convex or log-linear in* θ, *then* $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) < nI_{(X,Y)}(\theta)$.

Proof

(i) Equation (2) yields

$$\frac{\partial^2}{\partial\theta^2}\log f_{(\mathbf{R}_n,\mathbf{\Delta}_n,\mathbf{C}_n)}(\mathbf{r}_n,\boldsymbol{\delta}_n,\mathbf{s}_n;\theta) = \sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f_{X,Y}(r_i,s_i;\theta) + \sum_{i=1}^{n-1}\delta_i\frac{\partial^2}{\partial\theta^2}\log F_X(r_i;\theta). \tag{5}$$

The second term of the right hand side of (5) is non-positive by assumption. On the other hand

$$\frac{\partial^2}{\partial\theta^2}\log f_{X,Y}(r_i,s_i;\theta) = \frac{\partial^2}{\partial\theta^2}\log f_X(r_i;\theta) + \frac{\partial^2}{\partial\theta^2}\log f_{Y|X}(s_i|r_i;\theta). \tag{6}$$

By assumptions and in the view of Lemma 1 the second term on the right hand side of (6) does not depend on r_i and s_i. Noting that record values are stochastically ordered, i.e., $R_i <_{st} R_{i+1}$, we have $\mathrm{E}(l(R_i;\theta)) > \mathrm{E}(l(R_{i+1};\theta))$ for each i, since $l(x;\theta)$ is decreasing in x. Thus the proof is complete using the definition of Fisher information.

(ii) The proof is similar to that of part (i). ■

Remark 1 *One can easily see that*

$$\frac{\partial^2}{\partial\theta^2}\log F_X(x;\theta)=\frac{L(x;\theta)}{F_X^2(x;\theta)},$$

where $L(x;\theta)=F_X(x;\theta)\partial^2/\partial\theta^2 F_X(x;\theta)-(\partial/\partial\theta F_X(x;\theta))^2$. *So* $F_X(x;\theta)$ *is strictly log-concave, log-linear or strictly log-convex if and only if* $L(x;\theta)$ *is negative, zero or positive. This approach is used in the next illustrative examples.*

Remark 2 *For the case of lower records, their concomitants and inter-record times the result of the Theorem 1 holds by considering* $\bar{F}_X$ *instead of* F_X *and replacing increasing by decreasing and vice versa.*

Example 2 Bivariate normal with a known correlation r ***and*** $\mu_{\mathbf{X}}=r^{-1}\mu_{\mathbf{Y}}=\sigma_{\mathbf{X}}=\sigma_{\mathbf{Y}}=\theta$. *This family does not belong to class* $\mathscr{B}$. *However, the distribution of* Y *given* $X=x$ *is normal with mean* rx *and variance* $\theta^2(1-r^2)$ *which is a member of* $\mathscr{B}$. *So, this family is a member of* $\mathscr{K}$. *Taking* $\alpha=\theta^{-1}$, $l(x;\alpha)=-\alpha^{-4}/2-\alpha^{-3}x/4$ *which is decreasing in* x. *Also*

$$L(x;\alpha)=\frac{1}{2\pi}e^{-(1/2)(\alpha x-1)^2}\left[(\alpha x-1)\int_{\alpha x-1}^{\infty}e^{-(1/2)u^2}du-e^{-(1/2)(\alpha x-1)^2}\right],$$

it can be shown that the expression in the bracket on the right hand side of the above equation is negative (see Ahmadi and Arghami, 2001). Hence $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta)>nI_{(X,Y)}(\theta)$ *by Theorem 1.*

Theorem 2 *Let* $f_{X,Y}(x,y;\theta)$ *belong to* $\mathscr{B}$. *Then* $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta)$ *is less than, equal to or greater than* $nI_{(X,Y)}(\theta)$ *if* $F_X(x;\theta)$ *is strictly log-convex, log-linear and strictly log-concave, respectively in* θ.

Proof By Lemma 1, the first term on the right hand side of (5) does not depend on r_i's and s_i's, and it's expected value equals $-nI_{(X,Y)}(\theta)$. This completes the proof. ■

Some illustrative examples of Theorem 2 are bivariate normal with a known correlation r, $\mu_{\mathbf{X}}=\theta$ $\mu_{\mathbf{Y}}=\mathbf{0}$ and $\sigma_{\mathbf{X}}=\sigma_{\mathbf{Y}}=\mathbf{1}$, Arnold and Strauss's bivariate exponential (Arnold and Strauss, 1988 [See also Amini and Ahmadi, 2008]), Mardia's bivariate Pareto distribution with the joint pdf

$$f_{X,Y}(x,y;\theta)=\theta(\theta+1)(1+x+y)^{-(\theta+2)},\ x,\ y,\ \theta>0,$$

McKay's bivariate gamma distribution and Bilateral bivariate Pareto distribution. The results of these examples are summarized in Table 8 and the last two are presented below.

Example 3 McKay's bivariate gamma distribution *(McKay, 1934). Suppose* (X,Y) *has the joint pdf*

$$f_{X,Y}(x,y;\theta) = \frac{\theta^{a+b}}{\Gamma(a)\Gamma(b)} x^{a-1}(y-x)^{b-1}e^{-\theta y},\ y>x>0,\ \theta>0, \tag{7}$$

where a and b are known positive real numbers and $\Gamma(.)$ *is the well-known gamma function.*
This family is a member of $\mathscr{B}$ *and the marginal distribution of X is gamma with parameters a and* θ*. Hence* $L(x;\theta) = \frac{\theta^{2a-2}x^a}{\Gamma(a)^2}e^{-\theta x}\left\{(a-1-\theta x)\int_0^x y^{a-1}\exp(-\theta y)\,dy - x^a e^{-\theta x}\right\}$*, which is negative (see Ahmadi and Arghami, 2003). Therefore, applying Theorem 2,* $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) > nI_{(X,Y)}(\theta)$*.*

Example 4 Bilateral bivariate Pareto distribution. *This family has the joint pdf (for example, see De Groot, 1970)*

$$f_{X,Y}(x,y;\theta) = \theta(\theta+1)(a-b)^{\theta}(y-x)^{-(\theta+2)},\ x<b<a<y,\ \theta>1, \tag{8}$$

where the two quantities a and b are known positive real numbers.
This is again a member of $\mathscr{B}$*, and the marginal pdf of X is given by* $f_X(x;\theta) = \frac{\theta(a-b)^{\theta}}{(a-x)^{\theta+1}}$*.*
We obtain $L(x;\theta) = 0$*. Hence applying Theorem 2,* $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) = nI_{(X,Y)}(\theta)$*.*

Theorem 3 ***(Location or scale marginal in*** $\mathscr{B}$***)*** *Let* $f_{X,Y}(x,y;\theta,\eta)$ *belong to* $\mathscr{L}_{\mathscr{B}}$*, then:*

(i) $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta)$ *is less than, equal to or greater than* $nI_{(X,Y)}(\theta)$ *if X, is IRHR, CRHR or DRHR, respectively;*

(ii) the increment of Fisher information by considering inter-record times is equal to

$$I_{\mathbf{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\theta) = \sum_{i=1}^{n-1} E\left[\frac{h_0^2(R_i-\theta)}{F_0(R_i-\theta)}\right], \tag{9}$$

for a location marginal and equals

$$I_{\mathbf{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\theta) = \sum_{i=1}^{n-1} E\left[\frac{R_i^2 h_1^2(\theta R_i)}{F_1(\theta R_i)}\right], \tag{10}$$

for the scale marginal.

Proof

(i) One can easily show that for both location and scale families, $\frac{\partial^2}{\partial\theta^2}\log F_X(x;\theta)$ and $\frac{\partial^2}{\partial x^2}\log F_X(x;\theta)$ have the same sign, that is, convexity, linearity and concavity of

$\log F_X(x;\theta)$, in x is similar to that in θ. On the other hand, $F_X(x;\theta)$ is strictly log-convex, log-linear or strictly log-concave in x if and only if the reversed hazard rate function, $\tilde{h}_X(x;\theta)$, is increasing, constant or decreasing in x, respectively. So the results of part(i) follow from Theorem 2 and Remark 2.

(ii) Use Proposition 1. Note that for location and scale families, $\frac{\partial^2}{\partial\theta^2}\log F_X(x;\theta)$ is equal to $\frac{\partial^2}{\partial x^2}\log F_X(x;\theta)$ and $x^2\frac{\partial^2}{\partial x^2}\log F_X(x;\theta)$, respectively. Also $\frac{\partial^2}{\partial x^2}\log F_X(x;\theta)$ equals $\frac{\partial}{\partial x}\log\tilde{h}_X(x;\theta)$ and $\frac{\partial^2}{\partial x^2}\log\bar{F}_X(x;\theta)$ equals $\frac{\partial}{\partial x}\log h_X(x;\theta)$. So

$$\begin{aligned} I_{\boldsymbol{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\theta) &= \sum_{i=1}^{n-1}\mathrm{E}\left(h_0'(R_i-\theta)-\frac{\tilde{h}_0'(R_i-\theta)F_0(R_i-\theta)}{\bar{F}_0(R_i-\theta)}\right) \\ &= \sum_{i=1}^{n-1}E\left[\frac{h_0^2(R_i-\theta)}{F_0(R_i-\theta)}\right], \end{aligned}$$

for a location marginal and

$$\begin{aligned} I_{\boldsymbol{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\theta) &= \sum_{i=1}^{n-1}E\left[R_i^2\left(h_1'(\theta R_i)-\frac{\tilde{h}_1'(\theta R_i)F_1(\theta R_i)}{\bar{F}_1(\theta R_i)}\right)\right] \\ &= \sum_{i=1}^{n-1}E\left[\frac{R_i^2h_1^2(\theta R_i)}{F_1(\theta R_i)}\right], \end{aligned}$$

for the scale marginal. ■

Example 5 Bivariate normal with known correlation r, $\mu_{\mathbf{X}}=\theta$ $\mu_{\mathbf{Y}}=\mathbf{0}$ ***and*** $\sigma_{\mathbf{X}}=\sigma_{\mathbf{Y}}=\mathbf{1}$. *The considered bivariate normal family belongs to* $\mathscr{L}_{\mathscr{B}}$, *and the normal distribution is DRHR. Hence Theorem 3 also yields that* $I_{\mathbf{R}_n,\mathbf{C}_n,\boldsymbol{\Delta}_n}(\theta) > nI_{(X,Y)}(\theta)$. *Table 1 shows the values of* $I_{\boldsymbol{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\theta)$ *for* $n=2,3,5,7,10$ *of the normal distribution.*

Example 6 *(Continuation of Example 3) This family belongs to* $\mathscr{L}_{\mathscr{B}}$ *and the distribution of* X *is DRHR. Therefore, applying Theorem 3,* $I_{\mathbf{R}_n,\mathbf{C}_n,\boldsymbol{\Delta}_n}(\theta) > nI_{(X,Y)}(\theta)$. *Table 2 shows the values of* $\theta^2 I_{\boldsymbol{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\theta)$ *for different values of n and a. As can be seen, these values increase as the shape parameter a increases.*

Example 7 Bivariate gamma exponential distribution (i). *Suppose that*

$$f_{X,Y}(x,y;\theta)=\theta d x\exp\{-(\theta x+dxy)\}\ \ x>0,\ y>0,\ \theta>0, \tag{11}$$

where d is a known positive real number. This family belongs to $\mathscr{L}_{\mathscr{B}}$, *and the exponential distribution is DRHR. Therefore Theorem 3 yields that* $I_{\mathbf{R}_n,\mathbf{C}_n,\boldsymbol{\Delta}_n}(\theta) > nI_{(X,Y)}(\theta)$. *The values of* $\theta^2 I_{\boldsymbol{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\theta)$ *in Table 2 with* $a=1$ *are the corresponding Fisher information for the exponential distribution.*

Table 1: *The values of $I_{\mathbf{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(0)$ for $n = 2,3,5,7,10$ from standard normal distribution.*

n	2	3	5	7	10
$I_{\mathbf{\Delta}_n\|\mathbf{R}_n,\mathbf{C}_n}(0)$	1.6718	4.7961	15.7557	33.5634	73.9717

Corollary 1 ***(Shape marginal in $\mathscr{B}$)*** *Let $f_{X,Y}(x,y;\theta)$ belong to $\mathscr{S}_{\mathscr{B}}$. Then $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta)$ is less than, equal to or greater than $nI_{(X,Y)}(\theta)$, if $k(x) = x\tilde{h}_1(x)$ is increasing, constant or decreasing in x.*

Proof As in the proof of Theorem 2, and since we have in shape family $F_X(x;\theta) = F_1(x^\theta)$, taking $\gamma = x^\theta$ it follows that

$$\frac{\partial^2}{\partial\theta^2}\log F_1(x^\theta) = x^\theta(\log x)^2[\frac{\partial}{\partial\gamma}\gamma\tilde{h}_1(\gamma)].$$

Since $x > 0$, this gives us the result. ■

Example 8 Sub-class of $\mathscr{H}$ with power distribution marginal. *In order to illustrate the result of Corollary 1, a sub-class of $\mathscr{S}$ with $F_X(x) = x^\theta$, $x > 0,\ \theta > 0$ is concerned. Hence, $f_{Y|X}(y|x)$ must have the functional form of $\mathscr{B}$. So this class is also a sub-class of $\mathscr{H}$ with power distribution marginal. For power distribution, $k(x) = x\tilde{h}_1(x) = 1,\ x > 0$, which is constant. So $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) = nI_{(X,Y)}(\theta)$ by Corollary 1. An example of such bivariate distributions can be*

$$f_{X,Y}(x,y;\theta) = \theta^2 x^{-1}\exp\{\theta(\log x + x - y)\},\ 0 < x < y,\ \theta > 0.$$

Corollary 2 ***(Location or scale marginal in $\mathscr{K}$)*** *Let $f_{X,Y}(x,y;\theta)$ belong to $\mathscr{L}_{\mathscr{K}}$. Then $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta)$ is less (greater) than $nI_{(X,Y)}(\theta)$ if X is IRHR (DRHR), or CRHR and $l(x;\theta)$ is increasing (decreasing) in x.*

Proof The proof is similar to that of Theorems 1 and 3. ■

Table 2: *The values of $\theta^2 I_{\mathbf{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\theta)$ for $n = 3(2)7,10$ and $a = 0.5,1,2$ of gamma distribution.*

	a		
n	0.5	1	2
3	5.2036	8.8980	15.4526
5	27.0356	41.6880	66.4591
7	78.9683	114.1098	172.0214
10	245.0912	332.3383	473.1286

Example 9 *(Continuation of Example 2) The considered bivariate normal family belongs to $\mathscr{L}_{\mathscr{K}}$ with respect to parameter α, $l(x;\theta)$ is decreasing in x and the normal distribution is DRHR. Hence $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) > nI_{(X,Y)}(\theta)$ by Corollary 2.*

Corollary 3 ***(Shape marginal in $\mathscr{K}$)*** *Let $f_{X,Y}(x,y;\theta,\eta)$ belong to $\mathscr{S}_{\mathscr{K}}$. Then $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta)$ is less (greater) than $nI_{(X,Y)}(\theta)$, if $k(x) = x\tilde{h}_1(x)$ is increasing (decreasing) or constant and $l(x;\theta)$ is increasing (decreasing) in x.*

Proof The proof is similar to Theorem 1 and Corollary 1. ■

Corollary 4 *Let $\{(X_i,Y_i),\ i \geq 1\}$ be distributed as the family*

$$\{f_{X,Y}(x,y;\theta) \in \mathscr{B}; F_X \text{ is free of } \theta\},$$

then $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) = nI_{(X,Y)}(\theta)$.

Proof It is deduced from Theorem 2, since $L(x;\theta) = 0$. ■

Example 10 ***Bivariate gamma exponential distribution (ii).*** *Consider the joint pdf*

$$f_{X,Y}(x,y;\theta) = \frac{a^b\theta}{\Gamma(b)}x^b \exp\{-(ax+\theta xy)\}\ \ x>0,\ y>0,\ \theta>0, \tag{12}$$

where a and b are known positive real numbers. This family is a member of $\mathscr{B}$ and F_X is free of θ. Therefore by Corollary 4, $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) = nI_{(X,Y)}(\theta)$.

Remark 3 *For the case of lower records, their concomitants and inter-record times the results of Theorem 3 and Corollaries 1, 2 and 3 are reversed. For example in Corollary 2 $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta)$ is less than (greater than) $nI_{(X,Y)}(\theta)$, if X is DHR (IHR) or CHR and $l(x;\theta)$ is decreasing (increasing) in x. Note that in this case, we consider the standard hazard rate function in location and scale families, i.e., $h_i(x)$, $i = 0,1$.*

Theorem 4 *Let $f_{X,Y}(x,y;\theta)$ belong to $\mathscr{E}_1$, then:*

(i) $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) > nI_{(X,Y)}(\theta)$;

(ii) the increment of Fisher information by considering inter-record times is equal to

$$I_{\mathbf{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\theta) = \left(\frac{\alpha'(\theta)}{\alpha(\theta)}\right)^2 \sum_{i=1}^{n-1} i(i+1)\xi(i+2), \tag{13}$$

where $\xi(.)$ is the Riemann Zeta function.

Table 3: *The values of* $(\alpha(\theta)/\alpha'(\theta))^2 I_{\mathbf{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\alpha)$ *for* $n = 2(3)14$ *in* (13).

n	2	5	8	11	14
$(\alpha(\theta)/\alpha'(\theta))^2 I_{\mathbf{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\alpha)$	2.404	41.688	170.222	442.365	912.397

Proof

(i) The class $\mathscr{E}_1$ is a subclass of $\mathscr{B}$. Assuming $\alpha(\theta) = \alpha$, we have

$$L(x;\alpha) = -(\log \bar{G}(x))^2 \bar{G}(x)^{\alpha},$$

which is clearly negative. Hence, the result follows from Theorem 2.

(ii) Using Proposition 1, we have

$$\begin{aligned} I_{\mathbf{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\alpha) &= \sum_{i=1}^{n-1} \mathrm{E}\left(\frac{(\log \bar{G}(R_i))^2}{1-\bar{G}(R_i)^{\alpha}}\right) \\ &= \alpha^{-2} \sum_{i=1}^{n-1} \mathrm{E}\left(\frac{(\log \bar{F}(R_i))^2}{1-F(R_i)}\right) \\ &= \alpha^{-2} \sum_{i=1}^{n-1} \frac{1}{(i-1)!} \int_0^1 \frac{(-\log v)^{i+1}}{1-v}\, dv. \end{aligned}$$

Expanding the term $1/(1-v)$, we get

$$\begin{aligned} I_{\mathbf{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\alpha) &= \alpha^{-2} \sum_{i=1}^{n-1} \frac{1}{(i-1)!} \sum_{j=1}^{\infty} \int_0^1 v^{j-1} (-\log v)^{i+1}\, dv \\ &= \alpha^{-2} \sum_{i=1}^{n-1} i(i+1)\xi(i+2). \end{aligned}$$

■

Table 3 shows the values of $(\alpha(\theta)/\alpha'(\theta))^2 I_{\mathbf{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\alpha)$ for $n = 2(3)14$ in class $\mathscr{E}_1$.

Example 11 *(Continuation of Example 7) The distribution of* X *is exponential with parameter* θ*, which belongs to* $\mathscr{C}_1$*. Also the conditional distribution of* Y *given* $X = x$ *is free of* θ*. Hence, this family is a member of* $\mathscr{E}_1$*. Therefore Corollary 4 yields that* $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) > nI_{(X,Y)}(\theta)$.

Theorem 5 *Let* $f_{X,Y}(x,y;\theta,\eta)$ *belong to* $\mathscr{E}_2$*, then:*

(i) $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) = nI_{(X,Y)}(\theta)$*;*

(ii) the increment of Fisher information by considering inter-record times is equal to

$$I_{\mathbf{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\theta) = \left(\frac{\beta'(\theta)}{\beta(\theta)}\right)^2 \sum_{i=1}^{n-1} \varphi(i), \tag{14}$$

where

$$\varphi(i)=\sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\frac{1}{rs}\left[\frac{1}{(r+s-1)^i}-\frac{1}{(r+s)^i}\right]. \tag{15}$$

Proof

(i) The class $\mathscr{E}_2$ is a subclass of $\mathscr{B}$ with $c(\theta)=\beta(\theta)$. We may assume without loss of generality that $\beta(\theta)=\beta$. The result follows from Theorem 2, since $L(x;\beta)=0$.

(ii) Using Proposition 1, we have

$$\begin{aligned}I_{\boldsymbol{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\beta)&=\sum_{i=1}^{n-1}\mathrm{E}\left(\frac{H(R_i)^{\beta}(\log H(R_i))^2}{(1-H(R_i)^{\beta})^2}\right)\\&=\beta^{-2}\sum_{i=1}^{n-1}\mathrm{E}\left(\frac{F(R_i)(\log F(R_i))^2}{(1-F(R_i))^2}\right)\\&=\beta^{-2}\sum_{i=1}^{n-1}\frac{1}{(i-1)!}\int_0^1\frac{v}{(1-v)^2}(\log v)^2(-\log(1-v))^{i-1}\,dv.\end{aligned}$$

Expanding $\log(v)$ we have

$$\begin{aligned}I_{\boldsymbol{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\beta)&=\beta^{-2}\sum_{i=1}^{n-1}\frac{1}{(i-1)!}\sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\frac{1}{rs}\int_0^1 v(1-v)^{r+s-2}(-\log(1-v))^{i-1}\,dv\\&=\beta^{-2}\sum_{i=1}^{n-1}\sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\frac{1}{rs}\left[\frac{1}{(r+s-1)^i}-\frac{1}{(r+s)^i}\right].\end{aligned}$$

■

Table 4 shows the values of $\varphi(i)$ for $i=1(1)7$, which are calculated to 4 decimal places using the R package. These values tend very quickly to 1 as i increases, such that they are approximately equal to one, for $i\geq 7$. Hence using these values is a proper approach to calculate $I_{\boldsymbol{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\beta)$. Table 5 shows these values for $n=2(3)14$ in class $\mathscr{E}_2$.

Table 4: *The values of $\varphi(i)$ in (15) for $i=1(1)7$.*

i	1	2	3	4	5	6	7
$\varphi(i)$	0.8857	0.9772	0.9943	0.9984	0.9995	0.9999	1.0000

Table 5: *The values of $(\beta(\theta)/\beta'(\theta))^2 I_{\boldsymbol{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\beta)$ in (14) for $n=2(3)14$.*

n	2	5	8	11	14
$(\beta(\theta)/\beta'(\theta))^2 I_{\boldsymbol{\Delta}_n|\mathbf{R}_n,\mathbf{C}_n}(\beta)$	0.8857	3.8608	6.8602	9.8602	12.8602

Example 12 *(Continuation of Example 4). We have* $F_X(x;\theta) = \left(\frac{a-b}{a-x}\right)^{\theta}$. *Hence* F_X *belongs to* $\mathscr{C}_2$ *and therefore* $f_{X,Y}(x,y;\theta) \in \mathscr{E}_2$. *Thus, using Theorem 5,* $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) = nI_{(X,Y)}(\theta)$.

Theorem 6 *Let* $f_{X,Y}(x,y;\theta)$ *belong to* $\mathscr{F}$ *or* $\mathscr{K}$. *Then* $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta)$ *is less than, equal to or greater than* $nI_{(X,Y)}(\theta)$ *if and only if* $I_{\mathbf{R}_n,\mathbf{\Delta}_n}(\theta)$ *is less than, equal to or greater than* $nI_X(\theta)$, *respectively.*

Proof From equations (1) and (2)

$$I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) = I_{\mathbf{R}_n,\mathbf{\Delta}_n}(\theta) - \mathrm{E}\left[\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f_{Y|X}(R_{[i]}|R_i;\theta)\right].$$

The expectation above is equal to zero and $n\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f_{Y|X}(Y|X;\theta)\right]$ in $\mathscr{F}$ and $\mathscr{K}$, respectively. Hence, in both classes

$$I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) - nI_{(X,Y)}(\theta) = I_{\mathbf{R}_n,\mathbf{\Delta}_n}(\theta) - nI_X(\theta). \qquad \blacksquare$$

A result similar to Theorem 6 holds for lower records.

4 Estimation

To illustrate the applications of comparison study of Fisher information, discussed in previous section, we present a simulation study and a real data example.

4.1 A simulation study

In order to compare the performance of estimators based on bivariate records and inter-record times with corresponding estimators based on other types of data, consider a bivariate normal distribution. For simplicity, we may assume that the only unknown parameter in this model is $\theta = \mathrm{E}(X)$, i.e.,

$$f_{X,Y}(x,y;\theta) = \frac{1}{2\pi\sqrt{1-r^2}} \exp\left\{\frac{\left[(x-\theta)^2 + y^2 - 2r(x-\theta)y\right]}{-2(1-r^2)}\right\}, \qquad (16)$$

$$x,\ y \in \mathbb{R},\ \theta \in \mathbb{R}.$$

The likelihood equation for deriving the MLE of θ based on bivariate record values and inter-record times ($\hat{\theta}_{RCT}$, if exists) is as follows:

$$\sum_{i=1}^{n} R_i - n\theta - r\sum_{i=1}^{n} R_{[i]} - (1-r^2)\sum_{i=1}^{n} \delta_i \tilde{h}_0(R_i-\theta) = 0. \quad (17)$$

In this case, $\hat{\theta}_{RCT}$ has no explicit form and the values of this estimator have to be derived by numerical methods.

Now, the following criteria are interesting:

(a) Relative efficiency (RE) of estimator based on bivariate record values and inter-record times with respect to estimator based on bivariate record values only.

(b) RE of estimator based on bivariate record values and inter-record times with respect to estimator based on an independent bivariate random sample of the same size.

For deriving the RE of case **(a)**, we may consider the likelihood equation for deriving the MLE of θ based on bivariate record values only ($\hat{\theta}_{RC}$, if exists) as follows

$$\sum_{i=1}^{n} R_i - n\theta - r\sum_{i=1}^{n} R_{[i]} - (1-r^2)\sum_{i=1}^{n-1} h_0(R_i-\theta) = 0. \quad (18)$$

Again, the values of $\hat{\theta}_{RC}$ have to be derived by numerical methods. For deriving the RE of case **(b)**, note that the MLE of θ based on an iid sample of size n from this bivariate family equals

$$\hat{\theta}_{IID} = n^{-1}[\sum_{i=1}^{n} X_i - r\sum_{i=1}^{n} Y_i],$$

which is an unbiased estimator of θ with a variance equal to $(1-r^2)/n$.

Table 6: ***(a)*** *$RE(\hat{\theta}_{RCT}, \hat{\theta}_{RC})$ in bivariate normal distribution for different values of r and n.*

	r								
n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3	3.052	2.799	2.633	2.431	2.207	1.909	1.494	1.175	1.035
5	6.583	6.088	5.517	4.756	3.707	2.975	2.062	1.372	1.044
7	12.291	11.325	9.506	7.469	5.735	4.137	2.777	1.624	1.057
10	25.743	22.500	18.092	13.622	9.527	6.489	3.821	2.060	2.052

Table 7: ***(b)*** *$RE(\hat{\theta}_{RCT}, \hat{\theta}_{IID})$ in bivariate normal distribution for different values of r and n.*

	r								
n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3	1.680	1.630	1.594	1.596	1.550	1.625	1.727	1.918	2.767
5	2.568	2.557	2.425	2.420	2.307	2.216	2.172	2.244	2.810
7	3.716	3.605	3.563	3.446	3.240	3.132	2.813	2.694	2.884
10	5.784	5.865	5.573	5.382	4.950	4.439	3.872	3.448	3.142

Tables 6 and 7 show the simulated values of RE of $\hat{\theta}_{RCT}$ based on the first n bivariate upper records and inter-record times relative to $\hat{\theta}_{RC}$ and $\hat{\theta}_{IID}$, respectively, which are derived using 100,000 iterations generated by the R package. The minimum number of iterations is used to derive the root of equations (17) and (18) to 3 decimal places. Also the default method of finding the roots of equations in the R package is considered. As one can see in Figure 1, $\mathrm{MSE}(\hat{\theta}_{RCT})$ decreases as n or r increases. The simulated values showed that $\mathrm{MSE}(\hat{\theta}_{RCT})$ has similar values for positive and negative values of r. Also, since θ is a location parameter, the values of $\mathrm{MSE}(\hat{\theta}_{RCT})$ does not depend on θ. The values of $\mathrm{RE}(\hat{\theta}_{RCT}, \hat{\theta}_{RC})$ and $\mathrm{RE}(\hat{\theta}_{RCT}, \hat{\theta}_{IID})$ increase as n increases. The values of Table 7 seem to have a minimum point when r increases and the value of r for which $\mathrm{RE}(\hat{\theta}_{RCT}, \hat{\theta}_{IID})$ is minimum, tends to 1 by increasing n.

These values show that, in this example, the estimator of $\theta = \mathrm{E}(X)$ based on bivariate record values and inter-record times is more efficient than the corresponding estimator based on bivariate record values only and the estimator based on an iid bivariate sample of the same size. The result of Fisher information comparison for the parameter $\theta = \mathrm{E}(X)$ in this model and the fact that considering inter-record times causes an increment of Fisher information, uphold these estimation results.

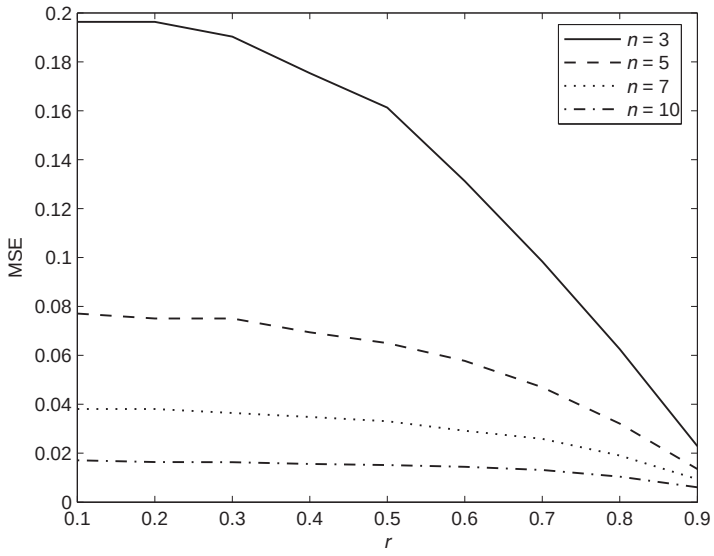


Figure 1: *$\mathrm{MSE}(\hat{\theta}_{RCT})$ in bivariate normal distribution for different values of r and n.*

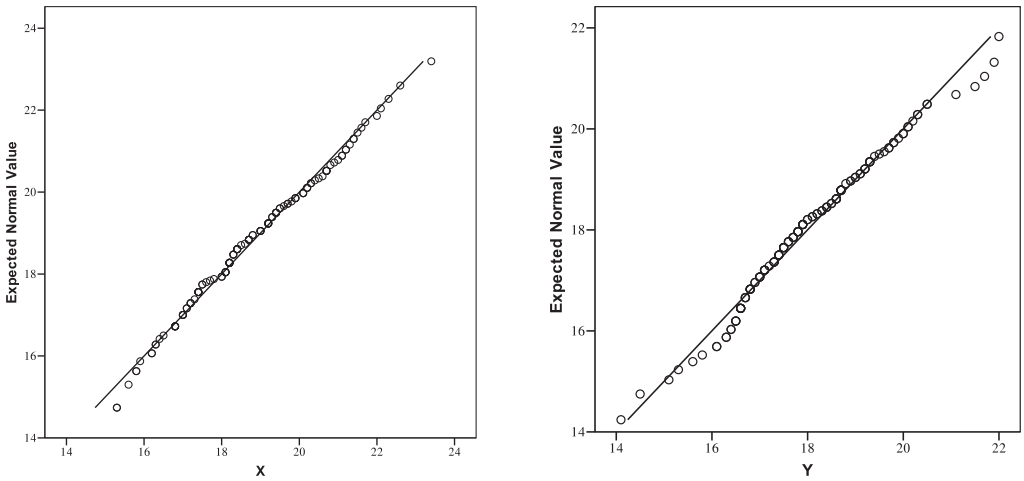
4.2 A real data example

As a real data example, we have considered 130 observations of temperatures at Neuenburg, Switzerland, on July (X) and August (Y), during 1864-1993 (Arnold *et al.*,

1998, p. 278). For these data, bivariate record values and inter-record times are given as follows:

Year	1864	1865	1869	1870	1881	1904	1911	1928	1983
i	1	2	3	4	5	6	7	8	9
Records (July), R_i	19.0	20.1	21.0	21.4	21.7	22.0	22.1	22.6	23.4
Concomitants (August), $R_{[i]}$	17.3	16.7	17.5	16.1	18.5	19.5	21.7	20.1	19.6
Inter-record times, Δ_i	0	3	0	10	22	6	16	54	0

In order to check the normality of the marginal distributions of X and Y, the corresponding Q-Q plots are drawn as follows.



The values of the Mardia test statistics (Mardia, 1974) are obtained as $V_1^* = 8.36 \times 10^{-141}$ and $V_2^* = -0.289$. Since the null hypothesis is rejected for large values of V_1^* and $|V_2^*|$, this indicates that the bivariate normal model provides a good fit to the above data.

Maximum likelihood estimates of the parameters, based on bivariate record values and also based on bivariate record values and inter-record times, are obtained by solving likelihood equations of bivariate normal distribution numerically as follows:

Parameter (θ)	μ_1	μ_2	σ_1^2	σ_2^2	ρ
Bivariate records	20.35	17.36	0.89	2.67	0.60
Bivariate records and times	20.12	17.21	1.32	2.82	0.63
Complete sample ($n = 130$)	18.79	18.04	2.89	2.15	0.31
$I_{\mathbf{\Delta}_n \mid \mathbf{R}_n, \mathbf{C}_n}(\hat{\theta})$	58.64	0	168.06	0	0

The complete sample estimators and the values of $I_{\mathbf{\Delta}_n \mid \mathbf{R}_n, \mathbf{C}_n}(\theta)$ (estimated values if unknown) are also given. As we can see, larger values of $I_{\mathbf{\Delta}_n \mid \mathbf{R}_n, \mathbf{C}_n}(\theta)$ cause a larger difference of the estimate based on bivariate records and complete sample estimates,

with respect to the corresponding difference of the estimate based on bivariate records and times.

5 Concluding remarks

In this paper, we have considered the problem of studying Fisher information in bivariate records in the present of inter-record times. Although, there is no information in record times themselves about the sampling distribution, the joint distribution of records and record times depends on it. We have seen that they provide significant additional information (see Table 8). For various cases an explicit formula for the increment of the Fisher information in the presence of inter-record times have obtained. Some general results have established to compare the amount of Fisher information in bivariate records and inter-record times with a random sample. Several classes of common univariate and bivariate families of distributions have been taken into account and some examples have been given in each cases to explain the results. The results of Section 4 show that the estimator on the basis of bivariate record values including inter-record times is more efficient than the corresponding estimator based on iid sample of the same size and the estimator based on bivariate records only. These results agree with the facts that $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) > nI_{(X,Y)}(\theta)$ and $I_{\mathbf{R}_n,\mathbf{C}_n,\mathbf{\Delta}_n}(\theta) > I_{\mathbf{R}_n,\mathbf{C}_n}(\theta)$ (when F_X depends on θ) for bivariate normal distribution.

Table 8: *Classification of some bivariate distributions based on information properties, by considering their marginal properties.*

Bivariate distribution	URC	URCT	LRC	LRCT	UR	URT	LR	LRT
Bivariate Normal,								
$\theta = \mathrm{E}(X)$ or $\mathrm{Var}(X)$	$<$	$>$	$<$	$>$	$<$	$>$	$<$	$>$
$\theta = \mathrm{E}(Y)$ or $\mathrm{Var}(Y)$	$=$	$=$	$=$	$=$	$=$	$=$	$=$	$=$
McKay's Biv. Gamma (7)								
$0 < a < 1$	$>$	$>$	$<$	$<$	$>$	$>$	$<$	$<$
$a = 1$	$=$	$>$	$<$	$=$	$=$	$>$	$<$	$=$
$a > 1$	$<$	$>$	$<$	$>$	$<$	$>$	$<$	$>$
Biv. Gamma exponential (11)	$=$	$>$	$<$	$=$	$=$	$>$	$<$	$=$
(12)	$=$	$=$	$=$	$=$	$=$	$=$	$=$	$=$
Bilateral Biv. Pareto (8)	$<$	$=$	$=$	$>$	$<$	$=$	$=$	$>$
Mardia Biv. Pareto	$=$	$>$	$<$	$=$	$=$	$>$	$<$	$=$
Arnold and Strauss's Bivariate Exponential [7]	$>$	$>$	$<$	$<$	$<$	$>$	$<$	$>$
Class $\mathscr{E}_1$	$=$	$>$	$<$	$=$	$=$	$>$	$<$	$=$
Class $\mathscr{E}_2$	$<$	$=$	$=$	$>$	$<$	$=$	$=$	$>$

Finally, some common bivariate distribution are classified in Table 8, according to the introduced criteria. The abbreviations URC (LRC), URCT (LRCT), UR (LR) and URT (LRT) are considered for upper (lower) records with their concomitants, upper (lower) records with their concomitants and inter-record times, upper (lower) records and upper (lower) records and inter-record times, respectively. The symbols "$>$", "$=$" and "$<$" mean that the Fisher information contained in the first n of the aforementioned statistics about θ is greater than, equal to and less than Fisher information contained in a random sample of size n from the parent bivariate distribution (or its X-marginal distribution for record statistics without concomitants). The results of the columns URC, LRC, UR and LR are given by Amini and Ahmadi (2008). The columns URT and LRT are the results of Theorem 6. From Table 8 we observe that there is a marked increase in the Fisher information by including inter-record times.

Acknowledgement

The authors would like to thank two referees for their constructive comments and suggestions which led to a considerable improved on an earlier version of this paper. Partial support from "Ordered and Spatial Data Center of Excellence of Ferdowsi University of Mashhad" is acknowledged.

References

Abo-Eleneen, Z. A. and Nagaraja, H. N. (2002). Fisher information in an order statistic and its concomitant. *Annals of the Institute of Statistical Mathematics*, 54, 667-680.

Ahmadi, J. and Arghami, N. R. (2001). On the Fisher information in record values. *Metrika*, 53, 195-206.

Ahmadi, J. and Arghami, N. R. (2003). Comparing the Fisher information in record values and i.i.d. observations. *Statistics*, 37, 435-441.

Amini, M. and Ahmadi, J. (2007). Fisher information in record values and their concomitants about the dependence and correlation parameters. *Statistics & Probability Letters*, 77, 964-972.

Amini, M. and Ahmadi, J. (2008). Comparing Fisher information in record values and their concomitants with random observations. *Statistics*, 42, 393-405.

Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1998). *Records*. John Wiley & Sons, New York.

Arnold, B. C. and Strauss, D. (1988). Bivariate distributions with exponential conditionals. *Journal of the American Statistical Association*, 83, 552-527.

Balakrishnan, N. and Stepanov, A. (2005). On the Fisher information in record data. *Statistics & Probability Letters*, 76, 537-545.

De Groot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

Hofmann, G. (2004). Comparing the Fisher information in record data and random observations. *Statistical Papers*, 45, 517-528.

Hofmann, G. and Nagaraja, H. N. (2003). Fisher information in record data. *Metrika*, 57, 177-193.

Lawless, J. L. (2003). *Statistical Models and Methods for Lifetime Data*. Second Edition. John Wiley, New York.

Lehmann, E. L. (1989). *Theory of Point Estimation*. John Wiley, New York.

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis for testing normality and robustness studies. *Sankhyā*, 58, 105-121.

McKay, A. T. (1934). Sampling from batches. *Journal of the Royal Statistical Society*, Supplement 1, 207-216.

Nagaraja, H. N. and Abo-Eleneen, Z. A. (2008). Fisher information in order statistics and their concomitants in bivariate censored samples. *Metrika*, 67, 327-347.

SORT 33 (2) July-December 2009, 233-248

Some improved two-stage shrinkage testimators for the mean of normal distribution

Zuhair Al-Hemyari*

Abstract

In this paper, we introduced some two-stage shrinkage testimators (TSST) for the mean μ when a prior estimate μ_0 of the mean μ is available from the past, by considering a feasible form of the shrinkage weight function which is used in both of the estimation stages with different quantities. The expressions for the bias, mean squared error, expected sample size and relative efficiency for the both cases when σ^2 known or unknown, are derived and studied. The discussion regarding the usefulness of these testimators under different situations is provided as conclusions from various numerical tables obtained from simulation results.

MSC: Primary 62F10, Secondary 62F99.

Keywords: Normal distribution, two-stage, shrinkage, preliminary test region, relative efficiency.

1 Introduction

1.1 TSST and Background

Let X be normally distributed with unknown mean μ and variance σ^2. Assume that prior information about μ is available in the form of an initial estimate μ_0 of μ. However, in certain situations the prior information is available only in the form of an initial guess value μ_0 of μ, then this guess may be utilized to improve the estimation procedure. For example, a bulb producer may know that the average life of his product may be close to 1000 hours. Here we may take $\mu_0 = 1000$. In such a situation it is natural to start

* *Address for correspondence:* Department of Mathematical and Physical Sciences, University of Nizwa, Oman, and on leave from Mathematics Department, College of Ibn Al-Haitham, Baghdad University, Iraq, e-mail addresses: alhemyari@unizwa.edu.om; drzuhair111@yahoo.com

Received: November 2008
Accepted: October 2009

with an estimator $\bar{X}$ of μ and modify it by moving it closer to μ_0, so that the resulting estimator, though perhaps biased, has a smaller mean squared error than that of $\bar{X}$ in some interval around μ_0. This method of constructing an estimator of θ that incorporates the prior information μ_0 leads to what is known as a shrunken estimator (see Thompson, 1968).

At the same time, it is an important aspect of estimation that one should be able to get an estimator quickly using minimum cost of experimentation. The cost of experimentation can be achieved by using any prior information available about μ and devising a two-stage shrunken estimator in which it is possible to obtain an estimator from a small first stage sample, and an additional second stage sample is required only if this estimator is not reliable (see Kambo, Handa and Al-Hemyari, 1991). The earliest work on two-stage estimation procedure is the paper by Katti (Katti, 1962). He developed a two-stage technique for the mean (μ) of a normal population when the variance (σ^2) is known. A number of other authors (see Al-Hemyari, 2009; Al-Hemyari and Al-Bayyati, 1981; Arnold and Al-Bayyati, 1970; Kambo *et al.*, 1992; Kambo *et al.*, 1991; Waiker, Ratnaparkhi, Schuurmann, 2001; Ratnaparkhi, Waiker, Schuurmann, 2001 and Waiker, Schuurmann and Raghunathan, 1984) have tried to develop new two-stage shrinkage testimators of the Katti type. The relevance of such types of TSST lies in the fact that, though perhaps they are biased, have smaller MSE than $\bar{X}$ in some interval around μ_0. A Two-stage shrinkage testimation (TSST) procedure is defined as follows. Let $X_{1i},\ i = 1,2,\ldots,n_1$ be a random sample of small size n_1 from $f(x|\mu)$. Compute the sample mean $\bar{X}_1$ and sample variance s^2 (unbiased estimator of σ^2, if σ^2 is unknown) based on n_1 observations. Construct a preliminary test region (R) in the space of μ, based on μ_0 and an appropriate criterion. If $\bar{X}_1 \in R$, shrink $\bar{X}_1$ towards μ_0 by shrinkage factor $0 \leq \varphi(\bar{X}_1) \leq 1$ and use the estimator $\varphi(\bar{X}_1)(\bar{X}_1 - \mu_0) + \mu_0$ for μ. But if $\bar{X}_1 \notin R$, obtain $X_{2i},\ i = 1,2,\ldots,n_2$ an additional sample of size $n_2 (= n - n_1)$, compute $\bar{X}_2$, and take the estimator of μ as the combined sample mean $\bar{X} = (n_1\bar{X}_1 + n_2\bar{X}_2)/(n_1 + n_2)$. Thus a two-stage shrinkage testimator of μ is given by:

$$\hat{\mu} = \{[\varphi(\bar{X}_1)(\bar{X}_1 - \mu_0) + \mu_0]I_R + [\bar{X}]I_{\bar{R}}\}, \tag{1}$$

where I_R and $I_{\bar{R}}$ are respectively the indicator functions of the acceptance region R and the rejection region $\bar{R}$.

1.2 The Modification

The TSST $\hat{\mu}$ is completely specified if the shrinkage weight factor $\varphi(\bar{X}_1)$ and the region R are specified. Consequently, the success of $\hat{\mu}$ depends upon the proper choice of $\varphi(\bar{X}_1)$ and R. Some choices for $\varphi(\bar{X}_1)$ and R are given in Al-Hemyari, 2009; Al-Hemyari and Al-Bayyati, 1981; Arnold and Al-Bayyati, 1970; Kambo *et al.*, 1992; Kambo *et al.*, 1991; Katti, 1962; Waiker *et al.*, 2001; Ratnaparkhi *et al.*, 2001 and Waiker *et al.*, 1984.

Other choices with different estimation problems are discussed in Al-Hemyari, Kurshid and Al-Gebori, 2009; Al-Hemyari and Al-Bayyati, 1981; Saxena and Singh, 2006 and Thompson, 1968. We proposed two-stage shrinkage testimators in this paper for the mean μ when σ^2 is known or unknown denoted by $\tilde{\mu}_i$, $i=1,2$, which are a modifications of $\hat{\mu}$ defined in (1). The proposed testimator takes the general form:

$$\tilde{\mu}_{=}\{[\varphi(\bar{X}_1)(\bar{X}_1-\mu_0)+\mu_0]I_R+[[(1-\varphi(\bar{X}_1)(\bar{X}-\mu_0)+\mu_0]]I_{\bar{R}}\}. \quad (2)$$

The main distinguishing feature of this type of TSST from conventional two stage shrinkage testimators is that, the pretest region rejects the prior estimate μ_0 only partially and even if $\bar{X}_1 \notin R$, μ_0, is given some weight though small in estimation of second stage. The expressions for the bias, mean squared error, expected sample size and relative efficiency of $\tilde{\mu}$ for the both cases when σ^2 known or unknown, are derived and studied theoretically and numerically. Comparisons with the earlier known results are made.

2 Formulation, assumptions and derivation of the proposed TSST with known σ^2

We define the general proposed estimator when σ^2 is known in this section. The bias, mean squared error, expected sample size, and relative efficiency expressions of the proposed testimator are derived. A suitable shrinkage function $\varphi(\bar{X}_1)$ is chosen, and finally some properties are also discussed.

2.1 The proposed testimator

Let X be normally distributed with unknown μ and known variance σ^2. Assume that a prior estimate μ_0 about μ is available from the past. The first proposed testimator is:

$$\tilde{\mu}_1=\{[\bar{X}_1-ae^{-n_1b(\bar{X}_1-\mu_0)^2/\sigma^2}(\bar{X}_1-\mu_0)]I_R+[[ae^{-n_1b(\bar{X}_1-\mu_0)^2/\sigma^2}(\bar{X}-\mu_0)+\mu_0]]I_{\bar{R}}\}. \quad (3)$$

R_1 is taking as the pretest region of size α for testing $H_0:\mu=\mu_0$ against $H_1:\mu\neq\mu_0$, where

$$R_1=[\mu_0-z_{\alpha/2}\sigma/\sqrt{n_1},\mu_0+z_{\alpha/2}\sigma/\sqrt{n_1}],\ \varphi(\bar{X}_1)=1-a\exp[-n_1b(\bar{X}_1-\mu_0)^2/\sigma^2], \quad (4)$$

$b\geq 0$, $0\leq a\leq 1$, and $z_{\alpha/2}$ is the upper $100(\alpha/2)$ percentile point of the standard normal distribution.

2.2 Bias ratio, MSE, Expected sample size and Relative Efficiency Expressions

It can be easily shown that the bias and mean squared error of $\tilde{\mu}_1$ are, respectively, given by:

$$\begin{aligned} B(\tilde{\mu}_1|\mu) &= (\sigma/\sqrt{n_1})\{J_1(a_1,b_1)+\lambda_1(J_0(a_1,b_1)-1)+a(2b+1)^{-3/2}e^{-b\lambda_1^2/(2b+1)}\times \\ &\times ((1/(1+f))-a(1(f+1)^{-1})[\sqrt{2b+1}J_1(a_2,b_2)+\lambda_1 J_0(a_2,b_2)]) \\ &- a\lambda_1\sqrt{f}(1+f)^{-1}\sqrt{2b+1}e^{-b\lambda_1^2/(2b+1)}(J_0(a_2,b_2)-1)\}, \end{aligned} \tag{5}$$

$$\begin{aligned} MSE(\tilde{\mu}_1|\mu) &= (\sigma^2/n)\{J_2(a_1,b_1)-2a(2b+1)^{-5/2}e^{-b\lambda_1^2/(2b+10)}[(2b+1)J_2(a_2,b_2) \\ &+ \lambda_1(1-2b)\sqrt{2b+1}J_1(a_2,b_2)-2b\lambda_1^2 J_0(a_2,b_2)]+a^2(1-(1+f)^{-2})\times \\ &\times (4b+1)^{-5/2}e^{-2b\lambda_1^2/(4b+1)}[(4b+1)J_2(a_3,b_3)+2\lambda_1\sqrt{4b+1}J_1(a_3,b_3) \\ &+ \lambda_1^2 J_0(a_3,b_3)]+\lambda_1^2(1-J_0(a,b))+a^2(1+f)^{-2}(4b+1)^{-5/2}((4b+1) \\ &+ \lambda_1^2)e^{-2b\lambda_1^2/(4b+1)}-2a\lambda_1(1+f)^{-1}(2b+1)^{-3/2}e^{-b\lambda_1^2/(2b+1)}\times \\ &\times [\lambda_1-\sqrt{2b+1}J_1(a_2,b_2)-\lambda_1 J_0(a_2,b_2)]+a^2f^2(1+\lambda_1^2)(1+f)^{-2}\times \\ &\times (4b+1)^{-1/2}e^{-2b\lambda_1^2/(4b+1)}(1-J_0(a,b))+2a^2\sqrt{f}(1+f)^{-2}\times \\ &\times (4b+1)^{-3/2}e^{-2b\lambda_1^2/(4b+1)}[\lambda_1+\sqrt{4b+1}J_1(a_2,b_2)+\lambda_1 J_0(a_2,b_2)] \\ &-2a\lambda_1^2\sqrt{f}(1+f)^{-1}(2b+1)^{-1/2}e^{-b\lambda_1^2/(2b+1)}J_0(a_2,b_2)\}, \end{aligned} \tag{6}$$

where

$$\begin{aligned} a_1 &= \lambda_1 - z_{\alpha/2}, & b_1 &= \lambda_1 + z_{\alpha/2}, \\ a_2 &= (\lambda_1^- z_{\alpha/2})/\sqrt{2b+1}, & b_2 &= (\lambda_2 + z_{\alpha/2})/\sqrt{2b+1}, \\ a_3 &= (\lambda_1 - z_{\alpha/2})/\sqrt{4b-1}, & b_3 &= (\lambda_1 + z_{\alpha/2})/\sqrt{4b-1}, \\ \lambda_1 &= \sqrt{n_1}(\mu-\mu_0)/\sigma, & f &= n_2/n_1, \end{aligned}$$

and

$$J_i(a_j,b_j) = \int_{a_j}^{b_j} \frac{1}{\sqrt{2\pi}} y^i e^{-y^2/2}dy, \quad i=0,1,2,\ j=1,2. \tag{7}$$

The expected sample size and the efficiency of $\tilde{\mu}_1$ relative to $\bar{X}$ are given respectively by:

$$E(n|\tilde{\mu}_1) = n_1[1 + f(1 - J_0(a_1, b_1), \tag{8}$$

$$Eff(\tilde{\mu}_1|\mu) = \sigma^2/E(n|\tilde{\mu}_1)MSE(\tilde{\mu}_1|\mu). \tag{9}$$

2.3 Selection of 'a'

It seems reasonable to select 'a' that minimizes the $MSE(\tilde{\mu}_1|\mu_0)$. Setting $((\partial/\partial a)$ $MSE(\tilde{\mu}_1|\mu_0))$ to zero, we get:

$$\begin{aligned} a = \bar{a}_1 = (1/n_1)[(2b+1)^{-3/2}J_2(a_2^*, b_2^*)/((4b+1)^{-3/2}((1+f)^{-2}(1 - J_2(a_3^*, b_3^*)) + \\ + J_2(a_3^*, b_3^*)) + \alpha f(1+f)^{-2}/\sqrt{4b+1})], \end{aligned} \tag{10}$$

where

$$\begin{aligned} a_2^* &= -z_{\alpha/2}/\sqrt{2b+1}, & b_2^* &= -a_2^*, \\ a_3^* &= -z_{\alpha/2}/\sqrt{4b-1}, & \text{and} \quad b_3^* &= -a_3^*. \end{aligned}$$

Since $(\partial^2/\partial a^2)MSE(\tilde{\mu}_1|\mu_0) \geq 0$. It follows that the minimizing value of $a \in [0,1]$ is given by:

$$\tilde{a} = \begin{cases} 0, & \text{if} \quad \bar{a}_1 \leq 0, \\ \bar{a}_1, & \text{if} \quad 0 \leq \bar{a}_1 \leq 1, \\ 1, & \text{if} \quad \bar{a}_1 \geq 1. \end{cases} \tag{11}$$

2.4 Some properties

i) Unbiasedness: If $\mu = \mu_0$, or $n_1 \to \infty$, the proposed testimator turns into the unbiased estimator, otherwise it is biased. Thus, we conclude the following: There does not exist, any unbiased estimator of μ in the class of testimators $\{\tilde{\mu} : 0 \leq \varphi(\bar{X}_1) \leq 1\}$ except the above undesirable cases.

ii) Minimum mean squared error estimator: It is not easy with the type of the proposed testimator to establish the minimum mean squared error biased estimator, i.e., $MSE(\tilde{\mu}|\mu) \leq MSE(\bar{X})$, for every $\varphi(\bar{X}_1)$ and every μ with strict inequality for at least one μ. But when $\mu = \mu_0$ the inequality holds, this means that by a proper choice of $\varphi(\bar{X}_1)$, the proposed TSST performs better (in the sense of smaller MSE) than $\bar{X}$ in the neighbourhood of μ_0. Also $Eff(\tilde{\mu}_1|\mu) \geq 1$ as $\lambda_1 \to \pm\infty$.

iii) Odd and even functions: It is easily seen that $B(\tilde{\mu}_1|\mu)$ is an odd function of λ_1, whereas $E(n|\tilde{\mu}_1)$, $MSE(\tilde{\mu}_1|\mu)$ and $Eff(\tilde{\mu}_1|\mu)$ are all even functions of λ_1.

iv) Consistent and dominant estimator: since

$$\lim_{n_1\to\infty} B(\tilde{\mu}_1|\mu) = 0 \quad \text{and} \quad \lim_{n_1\to\infty\infty} MSE(\tilde{\mu}_1|\mu) = 0,$$

$\tilde{\mu}_1$ is a consistent estimator of μ. Also $\tilde{\mu}_1$ dominates $\bar{X}$ in large n_1 and n_2 in the sense that

$$\lim_{n_1,n_2\to\infty} [MSE(\tilde{\mu}_1|\mu) - MSE(\bar{X})] \leq 0.$$

v) Special cases: It may be noted here, when $a = 0$, the equations (3), (5), (6), (8) & (9) agree with the result of Katti (Katti, 1962) also when $b = 0$, $(1-a) = k$, the same expressions agree with the result of Arnold and Al-Bayyati (Arnold and Al-Bayyati, 1970) when $b \to \infty$ and $a = 1$, the result agrees with the result of Kambo, Handa and Al-Hemyari (Kambo *et al.*, 1991), and when the second stage shrinkage function $(1-\varphi(\bar{X}_1)) = 1$, the result agrees with the result of Al-Hemyari (Al-Hemyari, 2009).

3 Formulation, assumptions and derivation of the proposed TSST with unknown σ^2

3.1 The proposed testimator

When σ^2 is unknown, it is estimated by

$$s^2 = \sum_{i=1}^{n_1} (X_i - \bar{X}_1)^2/(n_1 - 1).$$

Again taking region R_2 as the pretest region of size α for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ in the testimator $\tilde{\mu}_1$ defined in equation (3) and denoting the resulting estimator as $\tilde{\mu}_2$. The testimator $\tilde{\mu}_2$ employs R_2 given by:

$$\begin{aligned} R_2 &= [\mu_0 - t_{\alpha/2,n_1-1}s/\sqrt{n_1}, \mu_0 + t_{\alpha/2,n_1-1}s/\sqrt{n_1}], \\ \varphi(\bar{X}_1) &= 1 - a\exp[-n_1 b(\bar{X}_1 - \mu_0)^2/s^2], \end{aligned} \tag{12}$$

where $t_{\alpha/2,n_1-1}$ is the upper $100(\alpha/2)$ percentile point of the t distribution with $n_1 - 1$ degrees of freedom.

3.2 Bias ratio, MSE, expected sample size and relative efficiency expressions

The expressions for bias, *MSE* and expected sample size are given respectively by:

$$\begin{aligned} B(\tilde{\mu}_2|\mu) &= (\sigma/\sqrt{n_1})\int_{s^2=0}^{\infty}\{J_1(a_1,b_1)+\lambda_1(J_0(a_1,b_1)-1)+a(2b+1)^{-3/2}\times \\ &\times e^{-b\lambda_1^2/(2b+1)}((1/(1+f))-a(1+(f+1)^{-1})[\sqrt{2b+1}J_1(a_2,b_2) \\ &+ \lambda_1 J_0(a_2,b_2)])-a\lambda_1\sqrt{f}(1+f)^{-1}\sqrt{2b+1}e^{-b\lambda_1^2/(2b+1)}\times \\ &\times (J_0(a_2,b_2)-1)\}f(s_1^2|\sigma^2)ds_1^2, \end{aligned} \tag{13}$$

$$\begin{aligned} MSE(\tilde{\mu}_2|\mu) &= (\sigma^2/n)\int_{s_1^2=0}^{\infty}\{J_2(a_1,b_1)-2a(2b+1)^{-5/2}e^{-b\lambda_1^2/(2b+10)}\times \\ &\times [(2b+1)J_2(a_2,b_2)+\lambda_1(1-2b)\sqrt{2b+1}J_1(a_2,b_2) \\ &- 2b\lambda_1^2 J_0(a_2,b_2)]+a^2(1-(1+f)^{-2})(4b+1)^{-5/2}e^{-2b\lambda_1^2/(4b+1)}\times \\ &\times [(4b+1)J_2(a_3,b_3)+2\lambda_1\sqrt{4b+1}\,J_1(a_3,b_3)+\lambda_1^2 J_0(a_3,b_3)] \\ &+ \lambda_1^2(1-J_0(a,b))+a^2(1+f)^{-2}(4b+1)^{-5/2}((4b+1)+\lambda_1^2)\times \\ &\times e^{-2b\lambda_1^2/(4b+1)}-2a\lambda_1(1+f)^{-1}(2b+1)^{-3/2}e^{-b\lambda_1^2/(2b+1)}\times \\ &\times [\lambda_1-\sqrt{2b+1}J_1(a_2,b_2)-\lambda_1 J_0(a_2,b_2)]+a^2f^2(1+\lambda_1^2)(1+f)^{-2}\times \\ &\times (4b+1)^{-1/2}e^{-2b\lambda_1^2/(4b+1)}(1-J_0(a,b))+2a^2\sqrt{f}(1+f)^{-2}\times \\ &\times (4b+1)^{-3/2}e^{-2b\lambda_1^2/(4b+1)}[\lambda_1+\sqrt{4b+1}J_1(a_2,b_2)+\lambda_1 J_0(a_2,b_2)] \\ &-2a\lambda_1^2\sqrt{f}(1+f)^{-1}(2b+1)^{-1/2}e^{-b\lambda_1^2/(2b+1)}J_0(a_2,b_2)\}f(s_1^2|\sigma^2), \end{aligned} \tag{14}$$

and

$$E(n|\tilde{\mu}_2)=n_1\int_0^{\infty}[1+f(1-J_0(a_1^*,b_1^*))]f(s^2|\sigma^2)ds^2, \tag{15}$$

where $a_3^*=(\lambda_1-t_{\alpha/2,n_1-1}s_1/\sigma)/\sqrt{4b+1}$, $b_3^*=(\lambda_1+t_{\alpha/2,n_1}s_1/\sigma)/\sqrt{4b+1}$,
$a_2^*=(\lambda_1-t_{\alpha/2,n_1-1}xxs_1/\sigma)/\sqrt{2b+1}$, $b_2^*=(\lambda_1+t_{\alpha/2,n_1-1}s_1/\sigma)/\sqrt{2b+1}$,
$a_1^*=\lambda_1-t_{\alpha/2,n_1-1}s_1/\sigma$, $b_1^*=\lambda_1+t_{\alpha/2,n_1-1}$,

and $f(s^2|\sigma^2)$ is the p. d .f. of s^2. If $\mu=\mu_0$, the above expressions reduce to:

$$B(\tilde{\mu}_2|\mu_0)=0, \tag{16}$$

$$\begin{aligned} MSE(\tilde{\mu}_2|\mu_0) &= (\sigma^2/n_1)\{(1-\alpha)((1-2a(2b+1)^{-3/2}+a^2(4b+1)^{-3/2}\times \\ &\times (1-(1+f)^{-2}))+a^2(1+f)^{-2}(4b+1)^{-3/2}-2t_{\alpha/2,n_1-1}\times \\ &\times \Gamma(n_1/2)/g_0\sqrt{\pi(n_1-1)}+4at_{\alpha/2,n_1-1}\Gamma(n_1/2)/[g_2\sqrt{\pi(n_1-1)}\times \\ &\times (2b+1)+2a^2(1-(1+f)^{-2})t_{\alpha/2,n_1-1}\Gamma(n_1/2) \\ &\quad /[g_4\sqrt{\pi(n_1-1)}(4b+1)]+\alpha a^2 f\sqrt{4b+1}(1+f)^{-2}\}, \end{aligned} \tag{17}$$

where $g_m = [\Gamma((n_1-1)/2)(1+t^2_{\alpha/2,n_1-1}(mb+1)/(n_1-1))^{n_1/2}]$ $m = 0,2,4$ and

$$E(n|\tilde{\mu}_2) = n_1[1+\alpha f]. \tag{18}$$

The relative efficiency of $\tilde{\mu}_2$ defined by:

$$Eff(\tilde{\mu}_2|\mu_0) = \sigma^2/E(n|\tilde{\mu}_2)MSE(\tilde{\mu}_2|\mu_0) \tag{19}$$

Also, it is easily seen that

$$\lim_{n_1,n_2\to\infty} MSE(\tilde{\mu}_2|\mu) = 0 \quad \text{and} \quad \lim_{n_1,n_2\to\infty}[MSE(\tilde{\mu}_2|\mu_0) - MSE(\bar{X})] \leq 0.$$

3.3 Selection of 'a'

Proceeding in the manner as in the last section, we get the minimizing value of 'a' as follows:

$$\begin{aligned} a &= \bar{a}_2 = (1/n_1)[(1-\alpha)(2b+1)^{-3/2} - 2(t_{\alpha/2,n_1-1}\Gamma(n_1/2)/\sqrt{\pi(n_1-1)}(2b+1)g_2) \\ &\quad /[(4b+1)^{-3/2}((1-(1+f)^{-2})(1-\alpha)+(1+f)^{-2})+2(t_{\alpha/2,n_1-1}\Gamma(n_1/2) \\ &\quad /\sqrt{\pi(n_1-1)}(4b+1)^{-1}g_4(1-(1+f)^{-2})+\alpha f(1+f)^{-2}(4b+1)^{-1/2}). \end{aligned} \tag{20}$$

Since $(\partial^2/\partial a^2)MSE(\tilde{\mu}_2|\mu_0) \geq 0$. It follows that the minimizing value of $a \in [0,1]$ is given by:

$$\tilde{a} = \begin{cases} 0, & \text{if} \quad \bar{a}_2 \leq 0 \\ \bar{a}_2, & \text{if} \quad 0 \leq \bar{a}_2 \leq 1, \\ 1, & \text{if} \quad \bar{a}_2 \geq 1. \end{cases} \tag{21}$$

4 Examples

Example 1: Data were collected regarding weight, length and diameter of the Carp fish in Dokan lake (see Al-Hemyari and Al-Bayyati, 1981), where the estimation of the hunted quantity was calculated. In this example we will use the same data to illustrate how we can apply the proposed testimator $\tilde{\mu}_1$ as an estimator for the average length of the Carp fish. From the past data we had $\mu_0 = 33.314$, and $\sigma^2 = 13.814$. We draw a sample of size $n_1 = 5, 10$, $\bar{X}_1, R_1$ and $\tilde{\mu}_1$ are computed and given below for a number of values assigned for $n_2 = 10, 20, 30, 40$, $\alpha = 0.01$, and $b = 0.001$. The corresponding values of $Eff(\tilde{\mu}_1|\mu)$, $(\sqrt{n_1}/\mu)B(\tilde{\mu}_i|\mu)$, $E(n|\tilde{\mu}_i)$, $pr\{\bar{X}_1 \in R_1)$, $E(n|\tilde{\mu}_1)/n$, and $100(n_2/n)pr\{\bar{X}_1 \in R_1\}$ can be obtained from the Tables 1-6 using the corresponding constants $f = n_2/n_1$ and λ.

n_1	$\bar{X}_1$	$R_1 = [a,b]$	$n_2 = 5$	$n_2 = 10$	$n_2 = 20$	$n_2 = 30$	$n_2 = 40$
5	36.700	29.197,37.595	34.67	34.33	33.99	33.66	33.34
10	34.400	28.038,34.092	33.75	33.64	33.53	33.42	33.32

Example 2: Another data set will be used here to illustrate the calculations of the second proposed testimator $\tilde{\mu}_2$. An instructor is teaching a statistics course for many years at Nizwa University. Three groups of 120 students were registered in this course (cohort 2008) and all the students appeared for the final test. The teacher wants to estimate the average of the final score test using the prior value $\mu_0 = 82.19$ (from the last year test), and he decided the following: if $\tilde{\mu}_1 > \bar{X}_1$, he will consider $\tilde{\mu}_1$ as the sample mean of the current data and then he will modify the student's result on this basis. Based on a sample of size $n_1 = 5, 11$, $\bar{X}_1$, s, R_2 and $\tilde{\mu}_2$ are computed for a number of values assigned for $n_2 = 5, 11, 20, 35, 44$, $\alpha = 0.01$, $b = 0.001$ and given below. Some values of $Eff(\tilde{\mu}_2|\mu_0)$, $E(n|\tilde{\mu}_2)$ and $(100(n_2/n)pr\{\bar{X}_1 \in R_2\})$ are presented in Tables 7 and 8.

n_1	$\bar{X}_1$	s	$R_2 = [a,b]$	$n_2 = 5$	$n_2 = 11$	$n_2 = 20$	$n_2 = 35$	$n_2 = 44$
5	74.182	6.780	68.229,96.151	78.95	79.75	80.54	81.34	82.13
11	80.800	9.478	73.134,91.246	81.63	81.77	81.91	82.05	82.19

5 Simulation, Empirical results and Conclusions

A natural way of comparing the proposed two-stage shrinkage testimator is to study its performance with respect to the classical *MLE* $\bar{X}$ and with existing testimators given in Al-Hemyari, 2009; Arnold and Al-Bayyati, 1970; Kambo *et al.*, 1991; Katti, 1962; Waiker, Ratnaparkhi and Schuurmann, 2001; Ratnaparkhi *et al.*, 2001 and Waiker *et al.*, 1984. The comparisons were done on the basis of many properties and different

criterion. The computations of $Eff(\tilde{\mu}_i|\mu)$, $(\sqrt{n_1}/\mu)B(\tilde{\mu}_i|\mu)$, $E(n|\tilde{\mu}_i)$, probability of avoiding the second stage sample $(pr\{\bar{X}_1 \in R_i\})$, the ratio $E(n|\tilde{\mu}_i)/n$, the percentage of overall sample saved $(100(n_2/n)pr\{\bar{X}_1 \in R_i\})$, were done for the two-stage shrinkage testimators $\tilde{\mu}_1$ and $\tilde{\mu}_2$. From expressions (4, 5, 6, 8, 9, 11), it is observed that $Eff(\tilde{\mu}_1|\mu)$, $MSE(\tilde{\mu}_1|\mu)$, $B(\tilde{\mu}_1|\mu)$, $E(n|\tilde{\mu}_1)$, $E(n|\tilde{\mu}_1)/n$, and $100(n_2/n)pr(\bar{X}_1 \in R_1)$ for testimator $\tilde{\mu}_1$ are functions of α, n_1, n_2, f, b, and λ, whereas R_1 and $pr(\bar{X}_1 \in R_1)$ are functions of α, n_1, b, and λ. We have computed these expressions for a number of values which were assigned for $f = 0.5, 1(1)10$, $b = 0.001, 0.01, 0.02$, $\alpha = 0.01, 0.02, 0.05, 0.01, 0.015$, and the relative variation λ takes the values $0.0(0.1)4$. This was done to provide a wide variation in the values of μ_0 around the truth. Also, from expressions (12, 17, 18, 19, 21), notice that R_2, $MSE(\tilde{\mu}_2|\mu_0)$, $B(\tilde{\mu}_2|\mu_0)$, $E(\tilde{\mu}_2|\mu_0)$, $E(\tilde{\mu}_2|\mu_0)/n$, and $pr\{\bar{X}_1 \in R_1)$ for $\tilde{\mu}_2$ are functions of α, n_1, n_2, f, and b. This was done for $\alpha = 0.01, 0.02, 0.05$, $b = 0.001, 0.01$, $n_1 = 5, 11$, and $n_2 = 1(1)55$. Some of these computations are given in Tables 1 to 7. We make the following observations from tables presented in this paper:

i) From the computations of relative efficiency given in Table 1, and as expected the double stage shrinkage estimators give higher relative efficiency in some region a round μ_0. It is observed that the estimator $\tilde{\mu}_1$ has smaller mean squared error than the classical single stage estimator $\bar{X}$ for the region $0 \leq |\lambda| \leq 3$. Thus $\tilde{\mu}_1$ may be used to improve the efficiency if the difference $\mu_0 - \mu$ is expected to belong to the effective interval (boarder range of $|\lambda|$ for which efficiency is greater than unity) $ER = [-3\sigma/\sqrt{n_1}, 3\sigma/\sqrt{n_1}]$.

ii) It is also seen that from Table 1, for fixed f, b, and α, the relative efficiency of $\tilde{\mu}_1$ is maximum when $\lambda \cong 0$ (i.e., $\mu_0 = \mu$), and much greater than the classical estimator (as much as 3500 times), whereas the relative efficiency decreases with increasing value of $|\lambda|$, and it's less than 1 for $|\lambda| > 3$ (i.e., if $(\mu_0 - \mu) \notin [-3\sigma/\sqrt{n_1}, 3\sigma/\sqrt{n_1}]$).

iii) From Tables 1 and 2, it is observed that the testimator $\tilde{\mu}_1$ is biased. The bias ratio is reasonably small if the prior point estimate μ_0 does not deviate too much from the true value μ.

iv) It is observed from our computations given in Tables 1 and 2 that the relative efficiency of $\tilde{\mu}_1$ decreases with size α of the pretest region, i.e., $\alpha = 0.01$ gives higher relative efficiency than for other values of α. As α increases, $Eff(\tilde{\mu}_1|\mu)$ remains greater than the unity, whereas for any fixed α and b, the relative efficiency is a decreasing function of n_1 when $|\lambda| \cong 0$.

v) From Table 3, the probability of avoiding the second sample is independent of n_2 and it is clearly $1 - \alpha$ at $|\lambda| = 0$ but it decreases as λ increases or n_1 increases.

Table 1: *Showing $Eff(\tilde{\mu}_1|\mu)(Ef)$ and $(\sqrt{n_1}/\mu)B(\tilde{\mu}_1|\mu)/\mu(B)$ when $f = 0.5$, and different values of b, α, and λ.*

b	α	$\lvert\lambda\rvert$	0.0	0.2	0.4	0.6	0.8	1.0	2.0	3.0
0.001	0.01	Ef	15.4028	11.975	9.933	7.132	5.632	4.733	3.325	2.304
		B	0.000	−0.176	−0.208	−0.235	−1.277	−0.295	−0.397	−0.56
	0.05	Ef	11.284	9.842	7.243	5.573	4.276	3.518	2.846	2.105
		B	0.0000	−0.154	−0.189	−0.219	−0.259	−0.285	−0.364	−0.461
	0.1	Ef	9.285	7.177	6.428	5.856	4.0165	3.627	2.354	1.913
		B	0.000	−0.138	−0.171	−0.219	−0.225	−0.251	−0.339	−0.423
	0.135	Ef	6.564	5.119	4.417	3.922	3.217	2.843	2.114	1.500
		B	0.000	−0.109	−0.145	−0.173	−0.216	−0.236	−0.314	−0.399
0.01	0.01	Ef	14.829	11.284	8.345	6.823	5.426	4.064	3.156	2.163
		B	0.0000	−0.143	−0.145	−0.229	−0.264	−0.284	−0.373	−0.489
	0.05	Ef	10.372	8.043	6.824	4.414	3.890	3.099	2.184	1.778
		B	0.0000	−0.138	−0.169	−0.201	−0.233	−0.265	−0.328	−0.425
	0.1	Ef	7.393	5.784	4.627	3.864	3.432	2.835	2.159	1.471
		B	0.000	−0.121	−0.153	−0.185	−0.216	−0.243	−0.305	−0.399
	0.135	Ef	6.383	4.926	4.361	3.896	3.171	2.785	2.023	1.314
		B	0.000	−0.098	−0.137	−0.156	−0.199	−0.216	−0.297	−0.379

Table 2: *Showing $Eff(\tilde{\mu}_1|\mu)(Ef)$ and $(\sqrt{n_1}/\mu)B(\tilde{\mu}_1|\mu)/\mu(B)$ when $\alpha = 0.01$, $b = 0.001$, and different values of f and λ.*

f	$\lvert\lambda\rvert$	0.0	0.2	0.4	0.6	0.8	1.0	1.5	2.0	3.0
2	Ef	189.271	48.883	17.067	7.9723	5.4377	4.1283	2.7843	2.0900	1.0990
	B	0.000	−0.189	−0.360	−0.433	−0.471	−0.501	−0.479	−0.420	−0.399
4	Ef	280.215	57.006	19.725	8.2370	5.8850	4.3418	2.9657	1.8911	0.9940
	B	0.000	−0.198	−0.364	−0.441	−0.489	−0.501	−0.476	−0.399	−0.390
6	Ef	455.521	65.288	21.462	9.1907	5.9031	4.4310	3.0003	1.7873	0.9330
	B	0.000	−0.199	−0.364	−0.445	−0.489	−0.499	−0.447	−0.386	−0.378
8	Ef	1354.142	72.315	22.985	9.7143	6.2167	4.8733	3.1401	1.5010	0.9042
	B	0.000	−0.200	−0.365	−0.446	−0.492	−0.499	−0.428	−0.373	−0.362
10	Ef	3531.239	80.858	24.133	11.656	6.9177	5.4520	3.4213	1.0200	0.8940
	B	0.000	−0.202	−0.365	−0.446	−0.499	−0.489	−0.395	−0.358	−0.345

Table 3: *Showing $pr\{\bar{X}_1 \in R_1\}$ when $f = 0.5$, $b = 0.001$ and $\alpha = 0.01$.*

n_1 \ $\lvert\lambda\rvert$	0.0	0.2	0.4	0.6	0.8	1.0	2.0	3.0
4	0.990	0.990	0.990	0.990	0.990	0.990	0.988	0.968
8	0.990	0.990	0.990	0.990	0.988	0.983	0.822	0.714
12	0.990	0.986	0.982	0.981	0.978	0.973	0.878	0.581
16	0.990	0.984	0.981	0.979	0.975	0.971	0.816	0.500
20	0.990	0.983	0.981	0.975	0.952	0.926	0.681	0.345

Table 4: *Showing* $E(n|\tilde{\mu}_1)$ *when* $\alpha = 0.01$, $b = 0.001$ *and* $n_1 = 12$.

f \ $\lvert\lambda\rvert$	0.0	0.2	0.4	0.6	0.8	1.0	2.0	3.0
0.5	12.048	12.081	12.101	12.112	12.123	12.157	12.725	14.507
1	12.096	12.162	12.212	12.229	12.249	12.318	13.455	17.020
2	13.192	12.325	12.430	12.450	12.508	12.640	14.918	22.050
3	12.289	12.488	12.643	12.673	12.770	12.691	16.380	27.074
4	12.385	12.651	12.861	12.902	13.026	13.280	17.841	32.105
5	12.481	12.814	13.077	13.131	13.284	13.602	19.302	37.131
10	12.962	13.629	14.156	14.260	14.571	15.211	26.610	62.266

Table 5: *Showing* $E(n|\tilde{\mu}_1)/n$ *when* $\alpha = 0.01$, $b = 0.001$ *and* $n_1 = 12$.

f \ $\lvert\lambda\rvert$	0.0	0.2	0.4	0.6	0.8	1.0	2.0	3.0
0.5	0.699	0.671	0.673	0.673	0.674	0.675	0.707	0.806
1	0.502	0.506	0.509	0.509	0.511	0.513	0.561	0.709
2	0.335	0.342	0.345	0.346	0.347	0.351	0.414	0.613
3	0.251	0.260	0.263	0.264	0.266	0.270	0.341	0.564
4	0.206	0.211	0.214	0.215	0.217	0.221	0.297	0.535
5	0.167	0.178	0.182	0.182	0.184	0.189	0.268	0.516
10	0.098	0.103	0.107	0.108	0.110	0.115	0.202	0.472

Table 6: *Showing* $(100x(n_2|n)\ (pr\{\bar{X}_1 \in R_i\})$ *when* $\alpha = 0.01$, $b = 0.001$ *and* $n_1 = 12$.

f \ $\lvert\lambda\rvert$	0.0	0.2	0.4	0.6	0.8	1.0	2.0	3.0
0.5	33.066	32.872	32.730	32.7.1	32.615	32.431	29.268	19.360
1	49.599	49.315	49.089	49.052	48.824	48.649	43.905	29.048
2	66.132	65.752	65.461	65.364	65.097	64.719	58.541	38.731
3	74.398	73.975	73.641	73.583	73.372	72.973	65.861	43.569
4	79.358	78.910	79.552	78.481	78.279	77.851	70.253	46.477
5	82.665	82.193	81.827	81.752	81.540	81.089	73.179	48.412
10	90.180	89.661	89.260	89.177	88.941	88.460	79.823	52.813

vi) It is seen from Tables 4 and 5, that the expected sample size is close to n_1 when $\lambda = 0$ and increases very slowly with increases of $|\lambda|$ and f, whereas for any fixed α, b and n_1, the ratio $E(n|\tilde{\mu}_1)/n$ (which reflects the profligacy ratio in experimental units) is minimum when $|\lambda| = 0$, and decreases with increasing value of f.

vii) From Table 6, it is observed that the percentage of saving in sample is maximum when μ is close to μ_0 but it decreases as $|\lambda|$ increases. However, decreases in

Table 7: *Showing $Eff(\tilde{\mu}_2|\mu_0)$, when $\alpha = 0.01, 0.5, 0.1$, $b = 0.001$, $n_1 = 5, 11$ and n.*

	$n_1 = 5$			$n_1 = 11$		
n_2	0.01	0.05	0.1	0.01	0.05	0.1
5	231.551	54.289	30.654	197.237	49.272	29.384
8	392.692	89.978	49.476	278.389	82.727	42.833
11	598.497	134.088	71.912	373.825	109.894	62.077
14	851.458	186.612	97.728	483.797	140.599	78.452
17	1155.00	247.725	126.812	608.622	174.924	96.406
20	1513.40	317.792	159.145	748.681	212.811	115.874
23	1932.51	379.387	194.803	904.420	254.254	136.803
26	2419.42	487.322	233.950	1076.40	299.268	159.153
29	2983.30	588.691	276.841	1265.10	347.884	182.890
32	3636.01	702.929	323.835	1471.40	400.147	207.992
35	4392.22	831.902	375.403	1695.90	456.126	234.444
38	5271.40	978.029	432.157	1939.60	515.903	262.241
41	6298.92	1144.51	494.873	2203.40	579.583	291.383
44	7508.31	13335.32	564.540	2488.60	647.291	321.881

Table 8: *Showing $E(n|\tilde{\mu}_2)(E_2)$, and $(100x(n_2|n)\ (pr\{\bar{X}_1 \in R_2\})\,(E_3)$ when $\alpha = 0.01$, $b = 0.001$, $n_1 = 5, 11$ and n.*

	$n_1 = 5$		$n_1 = 11$	
n_2	E2	E3	E2	E3
5	5.050	49.500	11.050	68.063
8	8.080	38.077	11.080	57.316
11	5.110	30.938	11.110	49.500
14	5.140	26.053	11.140	43.560
17	5.170	22.500	11.170	38.893
20	5.200	19.800	11.200	35.129
23	5.230	17.679	11.230	32.029
26	5.260	15.968	11.260	29.432
29	5.290	14.559	11.290	27.225
32	5.320	13.378	11.320	25.326
35	5.350	12.376	11.350	23.674
38	5.380	11.512	11.380	22.224
41	5.410	10.761	11.410	20.942
44	5.440	10.102	11.440	19.800

percentage overall sample saved with increase in $|\lambda|$ is very slow irrespective f, e.g. for $\alpha = 0.01$, percentage sample saved is almost constant up to $|\lambda|$ as high as 0.8 even for f as high as 10.

viii) As the main purpose of a two-stage shrinkage testimator is to cut down the sample size without reducing efficiency, we shall like to study empirically the relation between efficiency, λ and $f = (n_2/n_1)$. Indeed the value of n_1 is dictated by the availability of the experimental data and the second sample n_2 can be produced

whenever necessary by performing a new experiment. It is observed from our computation given in Table 2, that (for $0 \leq |\lambda| \leq 0.5$) the increment of the maximum increase in relative efficiency decreases with f and is between 19 % to 5.5 % approximately. The corresponding increment of increase in f (or in n) is fixed and is 100 %. Thus the choice $f \cong 4(n_2 \cong 4n_1)$, is recommended (which corresponds to maximum increment in relative efficiency).

ix) The behavioural pattern of testimator $\tilde{\mu}_2$ is similar to that of $\tilde{\mu}_1$ as for expected sample size, relative efficiency, probability of avoiding the second stage sample and the percentage of overall samples saved are concerned.

x) Testimator $\tilde{\mu}_1$ is better than that of Katti (Katti, 1962), Arnold and Al-Bayyati (Arnold and Al-Bayyati, 1970), Waiker, Schuurmann and Raghunathan (Waiker, Schuurmann and Raghunathan, 1984), Kambo, Handa and Al-Hemyari (Kambo *et al.*, 1991), and Waiker, Ratnaparkhi, and Schuurmann (Waiker *et al.*, 2001) and Ratnaparkhi *et al.*, 2001) both in terms of higher relative efficiency and boarder range of the effective interval. Also comparing these results with the Tables 1 and 5 of Al-Hemyari (Al-Hemyari, 2009) it is observed that the testimator $\tilde{\mu}_1$ performs better in the sense of higher relative efficiency for $0 \leq |\lambda| \leq 2$. Comparing Table 7 with the results of Al-Hemyari, 2009; Arnold and Al-Bayyati, 1970; Kambo *et al.*, 1991; Waiker *et al.*, 2001; Ratnaparkhi *et al.*, 2001 and Waiker *et al.*, 1984, it is seen that $\tilde{\mu}_2$ is also much better in terms of higher relative efficiency than the existing testimators with unknown σ^2.

6 Summary

It has been seen that the suggested general two-stage shrunken testimators have considerable gain in relative efficiency for many choices of constants involved in it. It is recommended that one should not consider the substantial gain in efficiency in isolation, but also the wider range of $|\lambda|$. It is really interesting that the proposed testimator gives high relative efficiency for small first sample (or large f), which reduces the cost of the experimentation, and also for large first sample (or small f) and for a broad range of $|\lambda|$. Accordingly, even if the experimenter has less confidence in the guessed value μ_0 (if $\bar{X}_1 \notin R$), the relative efficiency is also greater than the classical and all the existing testimators. Moreover, the efficiency of the suggested testimators can be increased considerably by choosing the scalars α, n_1, n_2 and b appropriately. Thus it is recommended to use the proposed testimators in practice.

Acknowledgment

The author is thankful to the referees for constructive suggestions and valuable comments which resulted in the improvement of this article.

References

Al-Hemyari, Z. A., Kurshid, A. and A. Al-Gebori, A. (2009). On Thompson testimator for the mean of normal distribution. *Investigación Operacional*, 30, 109-116.

Al-Hemyari, Z. A. (2009). On Stein type two-stage shrinkage testimator. *International Journal of Modelling and Simulation*, 29, 1-7.

Al-Hemyari, Z. A. and Al-Bayyati, H. A. (1981). On double stage estimation for the multiple linear regression model. *Zanco*, the Scientific J. of Salahuddin University, 4, 134-155.

Arnold, J. C. and Al-Bayyati, H. A. (1970). On double stage estimation of the mean using prior knowledge. *Biometrics*, 26, 787-800.

Kambo, N. S., Handa, B. R. and Al-Hemyari, Z. A. (1992). On Huntsberger type shrinkage estimator. *Communications in Statistics-Theory and Methods*, 2, 823-841.

Kambo, N. S., Handa, B. R. and Al-Hemyari, Z. A. (1991). Double stage shrunken estimator of the mean of a normal distribution. *Journal of Information and Optimization Sciences*, 12, 1-11.

Katti, S. K. (1962). Use of some a prior knowledge in the estimation of means from double samples. *Biometrics*, 18, 139-147.

Saxena, S. and Singh, H. P. (2006). From ordinary to shrinkage square root estimators. *Communications in Statistics-Theory and Methods*, 35, 1037-1058.

Thompson, J. R. (1968). Some shrinkage techniques for estimating the mean. *Journal of the American Statistical Association*, 63, 953-963.

Waiker, V. B., Ratnaparkhi, M. V. and Schuurmann, F. J. (2001). Improving the efficiency of the two-stage shrinking estimators using bootstrap methods. *Proceedings of the International Conference on Monto Carlo and Quasi Methods*. Springer, Verlage, Hong Kong,

Ratnaparkhi, M. V., Waiker, V. B. and Schuurmann, F. J. (2001). Selection of the shrinkage factor for the two-stage testimator of the normal mean using bootstrap likelihood. www.galary.gmu.edu/interface/101/2001.

Waiker, V. B., Schuurmann, F. J. and Raghunathan, T. E. (1984). On a two stage shrinkage testimator of the mean of a normal distribution. *Communications in Statistics-Theory and Methods*, A 13, 1901-1913.

Information for authors and subscribers

Information for authors and subscribers

Submitting articles to SORT

Guidelines for submitting articles

SORT accepts for publication only original articles that have not been submitted to any other journal in the areas of Statistics, Operations Research, Official Statistics and Biometrics. Articles may be either theoretical or applied, and may include computational or educational elements. Publication will be exclusively in English, with an abstract in Catalan (abstract translation into Catalan can be done by the journal's staff).

All articles submitted to thematic sections will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board, except for those articles specifically invited by the journal or reprinted with permission. Reviewer comments will be sent to the primary author if changes are requested in form or content.

When an original article is received, the journal sends the author a dated receipt; this date will appear in the publication as "reception date". The date in which its final version is received will be the "acceptance date".

To submit an article, the author must send it in PDF format to the electronic address of the journal (sort@idescat.cat) especifying his e-mail address and/or a phone number. Once the article has been approved, the author must send a final electronic version, following the instructions of the editor responsible for that article. It is recommended that this final version be submitted using a LaTeX 2_ε .

In any case, upon request the journal secretary will provide authors with LaTeX 2_ε templates and appropriate references to the MSC2000 classification of the American Mathematical Society. For more information about the criteria for using this classification system, authors may consult directly the Mathematics Subject Classification 2000 (MSC2000) at the website of the Institut d'Estadística de Catalunya (http://www.idescat.es/sort/Normes.stm).

Publishing rights and authors' opinions

Authored articles represent the opinions of the authors; the journal does not necessarily agree with the opinions expressed in authored articles.

Formal guidelines for articles

The first page of the article must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (100-200 words) followed by MSC2000 classification of the American Mathematical Society and the keywords. Before submitting an article, the author(s) would be well advised to ensure that the text uses correct English.

Bibliographic references have to be listed alphabetically at the end of the article, according to the following examples:

Citations
Mahalanobis (1936), Rao (1982b)

Journal articles
Pfeffermann, D. and Barnard, C. H. (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business & Economic Statistics*, 9, 73-84.

Books
Platek, R., Rao, J. N. K., Särndal, C. E. and Singh, M. P. (eds.) (1987). *Small Area Statistics: an International Symposium*. New York: John Wiley and Sons.

Parts of books
Lee, E.W., Wei, L.J. and Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, J.P. Klein and P.K. Goel (eds.), Kluwer Academic Publishers, 237-247.

Web files or "pages"
Nielsen, S. F. (2001). *Proper and improper multiple imputation*
http://www.stat.ku.dk/~feodor/publications/ (10th May 2003).

Multiple publications by a single author are to be listed chronologically. Explanatory notes should be numbered sequentially and placed at the bottom of the corresponding page. Tables and figures should also be numbered sequentially and will be reproduced directly from the submitted originals if it is impossible to include them in the electronic text.

SORT *(Statistics and Operations Research Transactions)*
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel.: +34-93 412 09 24 or +34-93 412 00 88
Fax: +34-93 412 31 45
E-mail: sort@idescat.cat

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.

Subscription form
SORT *(Statistics and Operations Research Transactions)*

Name _____

_____

Organisation _____

_____

Street Address _____

_____

Zip/Postal code _____ City _____

State/Country _____ Tel. _____

Fax _____ NIF/VAT Registration Number _____

E-mail _____

Date _____

Signature

I wish to subscribe to **SORT *(Statistics and Operations Research Transactions)*** for the year 2009 (volume 33)

Annual subscription rates:

– Spain: €22 (4% VAT included)
– Other countries: €25 (4% VAT included)

Price for individual issues (current and back issues):

– Spain: €15/issue (4% VAT included)
– Other countries: €17/issue (4% VAT included)

Method of payment:

☐ Bank transfer to account number 2013-0100-53-0200698577

☐ Automatic bank withdrawal from the following account number

☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

☐ Check made payable to the Institut d´Estadística de Catalunya

Please send this subscription form (or a photocopy) to:

SORT *(Statistics and Operations Research Transactions)*
Institut d´Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45

Bank copy

Authorisation for automatic bank withdrawal in payment for

SORT *(Statistics and Operations Research Transactions)*

The undersigned _____

authorises Bank/Financial institution _____

located at (Street Address) _____

Zip/postal code _____ City _____

Country _____

to draft the subscription to **SORT *(Statistics and Operations Research Transactions)*** from my account

number ☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

Date _____

Signature

SORT *(Statistics and Operations Research Transactions)*
Institut d´Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45