FBST for Mixture Model Selection

Marcelo S. Lauretto and Julio M. Stern†

BIOINFO and Computer Science Dept., São Paulo University
lauretto@ime.usp.br, jstern@ime.usp.br

Abstract. The Fully Bayesian Significance Test (FBST) is a coherent Bayesian significance test for sharp hypotheses. This paper proposes the FBST as a model selection tool for general mixture models, and compares its performance with Mclust, a model-based clustering software. The FBST robust performance strongly encourages further developments and investigations.

THE FBST EVIDENCE VALUE

The Fully Bayesian Significance Test (FBST) is presented by Pereira and Stern (1999), as a coherent Bayesian significance test. The FBST is intuitive and has a geometric characterization. In this article the parameter space, Θ, is a subset of R^n, and the hypothesis is defined as a further restricted subset defined by vector valued inequality and equality constraints: $H : \theta \in \Theta_H$ where $\Theta_H = \{\theta \in \Theta \,|\, g(\theta) \leq 0 \wedge h(\theta) = 0\}$. For simplicity, we often use H for Θ_H. We are interested in precise hypotheses, with $\dim(H) < \dim(\Theta)$. $f(\theta)$ is the posterior probability density function.

The computation of the evidence measure used on the FBST is performed in two steps: The optimization step consists of finding f^* and $\widehat{f}$, the constrained (over H) and unconstrained maxima of the posterior. The integration step consists of integrating the posterior density over the Tangential Set, $\overline{T}$ where the posterior is higher than anywhere in H, i.e., $\overline{T} = \{\theta \in \Theta : f(\theta) > f^*\}$, $f^* = \max_H f(\theta) = f(\theta^*)$, $\widehat{f} = \max_\Theta f(\theta) = f(\widehat{\theta})$,

$$\overline{\mathrm{Ev}}(H) = \Pr(\theta \in \overline{T} \,|\, x) = \int_{\overline{T}} f(\theta) d\theta \quad .$$

$\overline{\mathrm{Ev}}(H)$ is the evidence against H, and $\mathrm{Ev}(H) = 1 - \overline{\mathrm{Ev}}(H)$ is the evidence supporting (or in favour of) H. Intuitively, if $\overline{\mathrm{Ev}}(H)$ is "large", $\overline{T}$ is "heavy", and the hypothesis set is in a region of "low" posterior density, meaning a "strong" evidence against H.

Let us consider the cumulative distribution of the evidence value against the hypothesis, $\overline{V}(c) = \Pr(\overline{\mathrm{Ev}} \leq c)$, given θ^0, the true value of the parameter. Under appropriate regularity conditions, for increasing sample size, $n \to \infty$, we can say the following:

- If H is false, $\theta^0 \notin H$, then $\overline{\mathrm{Ev}}$ converges (in probability) to one, that is, $\overline{V}(c) \to \delta(1)$.
- If H is true, $\theta^0 \in H$, then $\overline{V}(c)$, the confidence level, is approximated by the function $\overline{W}(t,h,c) = \mathrm{Chi2}\left(t-h, \mathrm{Chi2}^{-1}(t,c)\right)$, where $t = \dim(\Theta)$, $h = \dim(H)$ and $\mathrm{Chi2}(k,x)$ is the cumulative chi-square distribution with k degrees of freedom.

Several FBST applications and examples, efficient computational implementation, interpretations, and comparisons with other techniques for testing sharp hypotheses, can be found in the authors' papers in the reference list.

CP803, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*
edited by Kevin H. Knuth, Ali E. Abbas, Robin D. Morris, and J. Patrick Castle

DIRICHLET-NORMAL-WISHART MIXTURE MODELS

In a d-dimensional multivariate finite mixture model with m components (or classes), and sample size n, any given sample x^j is of class k with probability w_k; the weights, w_k, give the probability that a new observation is of class k. A sample j of class $k = c(j)$ is distributed with density $f(x^j \mid \psi_k)$.

This paragraph defines some general matrix notation. Let $r{:}s{:}t$ indicate either the vector $[r, r+s, r+2s, \ldots t]$ or the corresponding index range from r to t with step s; $r{:}t$ is a short hand for $r{:}1{:}t$. A matrix array has a superscript index, like $S^1 \ldots S^m$. So $S^k_{h,i}$ is the h-row, i-column element of matrix S^k. We may write a rectangular matrix, X, with the row (or shorter range) index subscript, and the column (or longer range) index superscript. So x_i, x^j, and x^j_i are row i, column j, and element (i,j) of matrix X. $\mathbf{0}$ and $\mathbf{1}$ are matrices of zeros and ones which dimensions are given by the context. In this paper, let h, i be indices in the range $1{:}d$, k in $1{:}m$, and j in $1{:}n$.

The classifications z^j_k are boolean variables indicating whether or not x^j is of class k, i.e. $z^j_k = 1$ iff $c(j) = k$. Z is not observed, being therefore named latent variable or missing data, see Robert (1996). Conditioning on the missing data, we get:

$$\begin{aligned} f(x^j \mid \theta) &= \textstyle\sum_{k=1}^m f(x^j \mid \theta, z^j_k) f(z^j_k \mid \theta) = \sum_{k=1}^m w_k f(x^j \mid \psi_k) \\ f(X \mid \theta) &= \textstyle\prod_{j=1}^n f(x^j \mid \theta) = \prod_{j=1}^n \sum_{k=1}^m w_k f(x^j \mid \psi_k) \end{aligned}$$

Given the mixture parameters, θ, and the observed data, X, the conditional classification probabilities, $P = f(Z \mid X, \theta)$, are:

$$p^j_k = f(z^j_k \mid x^j, \theta) = \frac{f(z^j_k, x^j \mid \theta)}{f(x^j \mid \theta)} = \frac{w_k f(x^j \mid \psi_k)}{\sum_{k=1}^m w_k f(x^j \mid \psi_k)}$$

We use y_k for the number of samples of class k, i.e. $y_k = \sum_j z^j_k$, or $y = Z\mathbf{1}$. The likelihood for the "completed" data, X, Z, is:

$$f(X, Z \mid \theta) = \textstyle\prod_{j=1}^n f(x^j \mid \psi_{c(j)}) f(z^j_k \mid \theta) = \prod_{k=1}^m \left({w_k}^{y_k} \prod_{j \mid c(j)=k} f(x^j \mid \psi_k) \right)$$

We will see in the following sections that considering the missing data Z, and the conditional classification probabilities P, is the key for successfully solving the numerical integration and optimization steps of the FBST. In this article we will focus on Gaussian finite mixture models, where $f(x^j \mid \psi_k) = N(x^j \mid b^k, R^k)$, a Normal density with mean b^k and variance matrix V^k, or precision $R^k = (V^k)^{-1}$. Next we specialize the theory of general mixture models to the Dirichlet-Normal-Wishart case.

Consider the random matrix X^j_i, i in $1{:}d$, j in $1{:}n$, $n > d$, where each column contains a sample element from a d-multivariate Normal distribution with parameters b (mean) and V (covariance), or $R = V^{-1}$ (precision). Let u and S denote the statistics:

$$u = (1/n) \textstyle\sum_{j=1}^n x^j = (1/n) X\mathbf{1} \;\;,\;\; S = \sum_{j=1}^n (x^j - b) \otimes (x^j - b)' = (X - b)(X - b)'$$

The random vector u has Normal distribution with mean b and precision nR. The random matrix S has Wishart distribution with n degrees of freedom and precision matrix

R. The Normal, Wishart and Normal-Wishart pdfs have expressions:

$$\begin{aligned} N(u\,|\,n,b,R) &= (\tfrac{n}{2\pi})^{d/2}|R|^{1/2}\exp\big(-(n/2)(u-b)'R(u-b)\big) \\ W(S\,|\,e,R) &= c^{-1}\,|S|^{(e-d-1)/2}\exp\big(-(1/2)\mathrm{tr}(SR)\big) \end{aligned}$$

with normalization constant $c = |R|^{-e/2}\,2^{ed/2}\,\pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma((e-i+1)/2)$.

Now consider the matrix X as above, with unknown mean b and unknown precision matrix R, and the statistic

$$S = \sum_{j=1}^{n}(x^j-u)\otimes(x^j-u)' = (X-u)(X-u)'$$

The conjugate family of priors for multivariate Normal distributions is the Normal-Wishart. Take as prior distribution for the precision matrix R the Wishart distribution with $\dot{e} > d-1$ degrees of freedom and precision matrix $\dot{S}$ and, given R, take as prior for b a multivariate Normal with mean $\dot{u}$ and precision $\dot{n}R$, i.e. let us take the Normal-Wishart prior $NW(b,R\,|\,\dot{n},\dot{e},\dot{u},\dot{S})$. Then, the posterior distribution for R is a Wishart distribution with $\ddot{e}$ degrees of freedom and precision $\ddot{S}$, and the posterior for b, given R, is k-Normal with mean $\ddot{u}$ and precision $\ddot{n}R$, i.e., we have the Normal-Wishart posterior:

$$\begin{aligned} NW(b,R\,|\,\ddot{n},\ddot{e},\ddot{u},\ddot{S}) &= W(R\,|\,\ddot{e},\ddot{S})\;N(b\,|\,\ddot{n},\ddot{u},R) \\ \ddot{n} &= \dot{n}+n\;\;,\;\;\ddot{e}=\dot{e}+n\;\;,\;\;\ddot{u}=(nu+\dot{n}\dot{u})/\ddot{n} \\ \ddot{S} &= S+\dot{S}+(n\dot{n}/\ddot{n})(u-\dot{u})\otimes(u-\dot{u})' \end{aligned}$$

All covariance and precision matrices are supposed to be positive definite, and proper priors have $\dot{e}\geq d$, and $\dot{n}\geq 1$. Non-informative Normal-Wishart improper priors are given by $\dot{n}=0$, $\dot{u}=0$, $\dot{e}=0$, $\dot{S}=0$, i.e. we take a Wishart with 0 degrees of freedom as prior for R, and a constant prior for b, see DeGroot (1970). Then, the posterior for R is a Wishart with n degrees of freedom and precision S, and the posterior for b, given R, is d-Normal with mean u and precision nR.

The conjugate prior for a multinomial distribution is a Dirichlet distribution:

$$\begin{aligned} M(y\,|\,n,w) &= \big(n!/y_1!\ldots y_m!\big)\;{w_1}^{y_1}\ldots{w_m}^{y_m} \\ D(w\,|\,y) &= \big(\Gamma(y_1+\ldots+y_k)/\Gamma(y_1)\ldots\Gamma(y_k)\big)\prod_{k=1}^{m}{w_k}^{y_k-1} \end{aligned}$$

with $w>\mathbf{0}$ and $w\mathbf{1}=1$. Prior information given by $\dot{y}$, and observation y, result in the posterior parameter $\ddot{y}=\dot{y}+y$. A non-informative prior is given by $\dot{y}=\mathbf{1}$.

Finally, we can write the posterior and completed posterior for the model as:

$$f(\theta\,|\,X,\dot{\theta}) = f(X\,|\,\theta)\,f(\theta\,|\,\dot{\theta})$$

$$f(X\,|\,\theta) = \prod_{j=1}^{n}\sum_{k=1}^{m}p_k^j w_k N(x^j\,|\,b^k,R^k)$$

$$f(\theta\,|\,\dot{\theta}) = D(w\,|\,\dot{y})\prod_{k=1}^{m}NW(b^k,R^k\,|\,\dot{n}_k,\dot{e}_k,\dot{u}^k,\dot{S}^k)$$

$$p_k^j = w_k N(x^j\,|\,b^k,R^k)\;/\;\sum_{k=1}^{m}w_k N(x^j\,|\,b^k,R^k)$$

$$f(\theta \mid X, Z, \dot{\theta}) = f(\theta \mid X, Z) f(\theta \mid \dot{\theta}) = D(w \mid \ddot{y}) \prod_{k=1}^{m} NW(b^k, R^k \mid \ddot{n}_k, \ddot{e}_k, \ddot{u}^k, \ddot{S}^k)$$

$$y = Z\mathbf{1} \ , \ \ddot{y} = \dot{y} + y \ , \ \ddot{n} = \dot{n} + y \ , \ \ddot{e} = \dot{e} + y$$

$$u^k = (1/y_k) \sum_{j=1}^{n} z_k^j x^j \ , \ S^k = \sum_{j=1}^{n} z_k^j (x^j - u^k) \otimes (x^j - u^k)'$$

$$\ddot{u}^k = (1/\ddot{y}_k)(\dot{n}_k \dot{u}^k + y_k u^k) \ , \ \ddot{S}^k = S^k + \dot{S}^k + (\dot{n}_k y_k / \ddot{n}_k)(u^k - \dot{u}^k) \otimes (u^k - \dot{u}^k)'$$

GIBBS SAMPLING, INTEGRATION AND OPTIMIZATION

In order to integrate a function over the posterior measure, we use an ergodic Markov Chain. The form of the Chain below is known as Gibbs sampling, and its use for numerical integration is known as Markov Chain Monte Carlo, or MCMC.

Given θ, we can compute P. Given P, $f(z^j \mid p^j)$ is a simple multinomial distribution. Given the latent variables, Z, we have simple conditional posterior density expressions for the mixture parameters:

$$f(w \mid Z, \dot{y}) = D(w \mid \ddot{y}) \ , \ f(R^k \mid X, Z, \dot{e}_k, \dot{S}^k) = W(R \mid \ddot{e}_k, \ddot{S}^k)$$

$$f(b^k \mid X, Z, R^k, \dot{n}_k, \dot{u}^k) = N(b \mid \ddot{n}_k, \ddot{u}^k, R^k)$$

Gibbs sampling is the MCMC generated by cyclically updating variables Z, θ, and P, by drawing θ and Z from the above distributions, Häggström (2002), Johnson (1987).

Given a mixture model, we obtain an equivalent model renumbering the components $1:m$ by a permutation $\sigma([1:m])$. This symmetry must be broken in order to have an identificable model, Stephens (1997). Let us assume there is an order criterion that can be used when numbering the components. If the components are not in the correct order, Label Switching is the operation of finding permutation $\sigma([1:m])$ and renumbering the components, so that the order criterion is satisfied.

If we want to look consistently at the classifications produced during a MCMC run, we must enforce a label switching to break all non-identificability symmetries. For example, in the Dirichlet-Normal-Mixture model, we could choose to order the components (switch labels) according to the the rank given by: 1- A given linear combination of the vector means, $c' * b^k$; 2- The variance determinant $|V^k|$. The choice of a good label switching criterion should consider not only the model structure and the data, but also the semantics and interpretation of the model.

The semantics and interpretation of the model may also dictate that some states, like certain configurations of the latent variables Z, are either meaningless or invalid, and shall not be considered as possible solutions. The MCMC can be adapted to deal with forbidden states by implementing rejection rules, that prevent the chain from entering the forbidden regions of the complete and/or incomplete state space, see Bennett (1976), Meng and Wong (1996).

The EM algorithm optimizes the log-posterior function $fl(X \mid \theta) + fl(\theta \mid \dot{\theta})$, see Dempster et al. (1977), Ormoneit and Tresp (1995), Russel (1988). The EM is derived from the conditional log-likelihood, and the Jensen inequality: If $w, y > \mathbf{0}, w'\mathbf{1} = 1$ then $\log w'y \geq w' \log y$. Let θ and $\tilde{\theta}$ be our current and next estimate of the MAP (Maximum

a Posteriori), and $p_k^j = f(z_k^j \,|\, x^j, \theta)$ the conditional classification probabilities. At each iteration, the log-posterior improvement is:

$$\delta(\tilde{\theta}, \theta \,|\, X, \dot{\theta}) \;=\; fl(\tilde{\theta} \,|\, X, \dot{\theta}) - fl(\theta \,|\, X, \dot{\theta}) \;=\; \delta(\tilde{\theta}, \theta \,|\, X) + \delta(\tilde{\theta}, \theta \,|\, \dot{\theta})$$

$$\delta(\tilde{\theta}, \theta \,|\, \dot{\theta}) \;=\; fl(\tilde{\theta} \,|\, \dot{\theta}) - fl(\theta \,|\, \dot{\theta})$$

$$\delta(\tilde{\theta}, \theta \,|\, X) \;=\; fl(X \,|\, \tilde{\theta}) - fl(X \,|\, \theta) \;=\; \textstyle\sum_j \delta(\tilde{\theta}, \theta \,|\, x^j)$$

$$\delta(\tilde{\theta}, \theta \,|\, x^j) \;=\; fl(x^j \,|\, \tilde{\theta}) - fl(x^j \,|\, \theta) \;=\; \log \textstyle\sum_k \tilde{w}_k f(x^j \,|\, \tilde{\psi}_k) \;-\; fl(x^j \,|\, \theta) \;=$$

$$=\; \log \textstyle\sum_k \dfrac{p_k^j}{p_k^j} \dfrac{\tilde{w}_k f(x^j \,|\, \tilde{\psi}_k)}{f(x^j \,|\, \theta)} \;\;\geq\;\; \Delta(\tilde{\theta}, \theta \,|\, x^j) \;=\; \textstyle\sum_k p_k^j \log \dfrac{\tilde{w}_k f(x^j \,|\, \tilde{\psi}_k)}{p_k^j f(x^j \,|\, \theta)}$$

Hence, $\Delta(\tilde{\theta}, \theta \,|\, X, \dot{\theta}) = \Delta(\tilde{\theta}, \theta \,|\, X) + \delta(\tilde{\theta}, \theta \,|\, \dot{\theta})$, is a lower bound to $\delta(\tilde{\theta}, \theta \,|\, X, \dot{\theta})$. Also $\Delta(\theta, \theta \,|\, X, \dot{\theta}) = \delta(\theta, \theta \,|\, X, \dot{\theta}) = 0$. So, under mild differentiability conditions, both surfaces are tangent, assuring convergence of EM to the nearest local maximum. But maximizing $\Delta(\tilde{\theta}, \theta \,|\, X, \dot{\theta})$ over $\tilde{\theta}$ is the same as maximizing

$$Q(\tilde{\theta}, \theta) \;\;=\;\; \textstyle\sum_{k,j} p_k^j \log\left(\tilde{w}_k f(x^j \,|\, \tilde{\psi}_k)\right) + fl(\tilde{\theta} \,|\, \dot{\theta})$$

and each iteration of the EM algorithm breaks down in two steps:
E-step: Compute $P = E(Z \,|\, X, \theta)$. M-step: Optimize $Q(\tilde{\theta}, \theta)$, given P.

For the Gaussian mixture model, with a Dirichlet-Normal-Wishart prior,

$$\begin{aligned} Q(\tilde{\theta}, \theta) \;\;&=\;\; \textstyle\sum_{k=1}^m \sum_{j=1}^n p_k^j \left(\log \tilde{w}_k + \log N(x^j \,|\, \tilde{b}^k, \tilde{R}^k)\right) + fl(\tilde{\theta} \,|\, \dot{\theta}) \\ fl(\tilde{\theta} \,|\, \dot{\theta}) \;\;&=\;\; \log D(\tilde{w} \,|\, \dot{y}) + \textstyle\sum_{k=1}^m \log NW(\tilde{b}^k, \tilde{R}^k \,|\, \dot{n}_k, \dot{e}_k, \dot{u}^k, \dot{S}^k) \end{aligned}$$

Lagrange optimality conditions give a simple analytical solutions for the M-step:

$$y = P\mathbf{1} \;\;,\;\; \tilde{w}_k = \left(y_k + \dot{y}_k - 1\right) \,/\, \left(n - m + \textstyle\sum_{k=1}^m \dot{y}_k\right)$$

$$u^k = \tfrac{1}{y_k} \textstyle\sum_{j=1}^n p_k^j x^j \;\;,\;\; S^k = \textstyle\sum_{j=1}^n p_k^j (x^j - \tilde{b}^k) \otimes (x^j - \tilde{b}^k)'$$

$$\tilde{b}^k = \dfrac{\dot{n}_k \dot{u}^k + y_k u^k}{\dot{n}_k + y_k} \;\;,\;\; \tilde{V}^k \;\;=\;\; \dfrac{S^k + \dot{n}_k (\tilde{b}^k - \dot{u}^k) \otimes (\tilde{b}^k - \dot{u}^k)' + \dot{S}^k}{y_k + \dot{e}_k - d}$$

In more general (non-Gaussian) mixture models, if an analytical solution for the M-step is not available, a robust local optimization algorithm can be used, see for example Birgin et al. (2004). The EM is only a local optimizer, but the MCMC provides plenty of good starting points, so we have the basic elements for a global optimizer. To avoid using many starting points going to a same local maximum, we can filter the (ranked by the posteriori) top portion of the MCMC output using a clustering algorithm, and select a starting point from each cluster. For better efficiency, or more complex problems, the Stochastic EM or SEM algorithm can be used to provide starting points near each important local maximum, see Celeux et al. (1996), Pflug (1996) and Spall (2003).

MODEL SELECTION AND COMPARATIVE RESULTS

The problem under study is to determine the number of components (or classes) in a population, given a sample X drawn from that population. Each component k is assumed to follow a multivariate Normal distribution, whose mean vector b^k and variance matrix V^k must also be estimated.

In the FBST formulation of the problem, the base model has m components, and the hypothesis to be tested is the constraint of having $m-1$ components, i.e., components m and $m-1$ are identical. The FBST selects the m component model, rejecting H, if the evidence against the hypothesis is above a critical level, $\overline{\mathrm{Ev}}(H) > c$, and selects the $m-1$ component model, accepting H, otherwise. In order to determine the number of components, we apply the FBST in the base model with $m = 2,3,...$ components, and stop the process at the lowest m such that the hypotheses is accepted, m_f. The elected model has $m_f - 1$ components.

Several methods can be used to choose the critical level c. Empirical power analysis, see Stern and Zacks (2002), Lauretto et al. (2003), and sensitivity analysis, Stern (2004), require calibration procedures. Loss functions, Madruga et al. (2001), require decision theoretical interpretations. Application of these methods will be discussed in forthcoming papers. Following an anonymous referee suggestion, we proceed with a traditional power analysis. This is a form of the Rule of Parsimony, or Occam's Razor: Accept H, the smaller model, unless there is strong evidence not to do so.

We use approximate (asymptotic) critical levels corresponding to the standard Fisher confidence level of $1-\alpha$ for $\alpha = 0.01$. For example (see section 1), at the $m = 3$ base model, $t = 17$ and $h = 11$, giving $c = 0.53$.

When implementing the FBST one has to be careful with trapping states on the MCMC. These typically are states where one component has a small number of sample points, that become (nearly) collinear, resulting in a singular posterior. A standard way to avoid this inconvenience is to use flat or minimally informative priors, instead of non-informative priors, see Robert (1996). We used as flat prior parameters: $\dot{y} = \mathbf{1}$, $\dot{n} = 1$, $\dot{u} = u$, $\dot{e} = 3$, $\dot{S} = (1/n)S$. Robert (1996) uses, with similar effects, $\dot{e} = 6$, $\dot{S} = (1.5/n)S$.

In this work we compare the FBST performance with Mclust, a software for model-based cluster analysis, see Banfield and Raftery (1993) and Fraley and Raftery (1999). Mclust is available at the authors' internet site as an easy to use and ready to run software package, that has been extensively and successfully used in many applications. Also, Mclust has no extra parameters that need to be adjusted or calibrated to the specific application. These characteristics motivated our choice of Mclust for a first comparison with the FBST. Forthcoming articles will include other well published methods, based on Dirichlet processes, jump-diffusion and birth-death MCMC.

In Mclust, the variance structure and the number of components are selected via Bayesian Information Criterion (BIC), see Schwarz (1978): $BIC = 2\Lambda - \kappa\log(n)$, where Λ is the maximum model log-likelihood, κ its number of parameters, and n the sample size. BIC is a regularization criteria, weighting the model fit against the number of parameters. A larger BIC score indicates stronger evidence for the corresponding model.

Our numerical experiments are based on the *Old Faithful* dataset, see Stephens (1997), which consists of 272 eruptions observations of the Old Faithful geyser in the Yellow-

stone National Park. Each observation has the eruption duration and waiting time before the next eruption. The problem is to decide how many classes of eruptions there exist. Old Faithful is a standard dataset for experiments in the area, allowing our results to be easily reproduced, but our general conclusions have been confirmed in several randomly generated datasets.

Two numerical experiments on simulated data were performed, using parameters θ^* and $\widehat{\theta}$, the maximum likelihood estimators for 2 and 3 component models in the original dataset. In the first experiment, our interest was to analyze the overestimate and underestimate rates on the number of components, for FBST and Mclust. We used Mclust library to generate a random collection of 500 datasets with 272 points each using parameter θ^* and a second collection of 500 datasets with 272 points each using parameter $\widehat{\theta}$. Table 1 shows the number of datasets according to the estimated number of components by FBST and Mclust. Each column corresponds to one of the collections, at θ^* and $\widehat{\theta}$, and each row represents the estimated number of components.

In the second numerical experiment we examine the FBST and Mclust choice between the 2 and 3 component models, as the sample size n increases. For each $n \in \{200, 300, 400, 500, 600\}$, we simulated two collections of 500 datasets with n points each, one using the parameter θ^*, and the other using parameter $\widehat{\theta}$. Table 2 shows the number of missclassifications for FBST and Mclust, at each of the 10 collections.

These (preliminary) results corroborate the authors' previous findings, indicating that the FBST is a robust Bayesian sharp hypothesis test, and a promissing tool for model selection, deserving further development and investigation.

Finally, let us point out a related topic for further research: The problem of discriminating between models consists of determining which of m alternative models, $f_k(x, \psi_k)$, more adequately fits or describes a given dataset. In general the parameters ψ_k have distinct dimensions, and the models f_k have distinct (unrelated) functional forms. In this case it is usual to call them "separate" models (or hypotheses). Atkinson (1970), although in a very different theoretical framework, was the first to analyse this problem using a mixture formulation, $f(x \mid \theta) = \sum_{k=1}^{m} w_k f_k(x, \psi_k)$. The general theory for mixture models presented in this article can be adapted to analyse the problem of discriminating between separate hypotheses. This is the subject of the authors' ongoing research with C.A.B.Pereira and B.B.Pereira, to be presented in forthcoming articles.

The authors are grateful for the support of CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico, and FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo.

Estimated components	FBST θ^*	FBST $\widehat{\theta}$	Mclust θ^*	Mclust $\widehat{\theta}$
1	0	0	0	0
2	498	187	500	280
3	2	288	0	218
4	0	25	0	2
–	–	–	–	–

Table 1: Datasets (in 500), according to estimated number of components.

Dataset Size	FBST θ^*	FBST $\widehat{\theta}$	Mclust θ^*	Mclust $\widehat{\theta}$
200	4	356	0	390
300	2	82	0	235
400	1	47	0	156
500	5	3	0	69
600	6	0	0	31

Table 2: Missclassifications (in 500), according to dataset size.

REFERENCES

A.C.Atkinson (1970). A Method for Discriminating Between Models. *J. Royal Stat. Soc. B*, 32, 323-354.

J.D.Banfield,A.E.Raftery (1993). Model Based Gaussian and nonGaussian Clustering. *Biometrics*,803-21.

C.H.Bennett (1976). Efficient Estimation of Free Energy Differences from Monte Carlo Data. *Journal of Computational Physics* 22, 245-268.

E.G.Birgin, R.Castillo, J.M.Martinez (2004). Numerical comparison of Augmented Lagrangian algorithms for nonconvex problems. to appear in *Computational Optimization and Applications.*

W.Borges,J.M.Stern (2005). *On the Truth Value of Complex Hypotheses.* Tech.Rep.MAC-IME-USP-05-5.

G.Celeux, G.Govaert (1995). Gaussian Parsimonious Clustering Models. *Pattern Recog.* 28, 781-793.

G.Celeux, D.Chauveau, J.Diebolt (1996). On Stochastic Versions of the EM Algorithm. An Experimental Study in the mixture Case. *Journal of Statistical Computation and Simulation,* 55, 287–314.

M.H.DeGroot (1970). *Optimal Statistical Decisions.* NY: McGraw-Hill.

A.P.Dempster, N.M.Laird, D.B.Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. of the Royal Statistical Society B.* 39, 1-38.

C.Fraley, A.E.Raftery (1999). Mclust: Software for Model-Based Cluster Analysis. *J. Classif.*,16,297-306.

W.R.Gilks, S.Richardson, D.J.Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice.* NY: CRC.

O.Häggström (2002). *Finite Markov Chains and Algorithmic Applications.* Cambridge University Press.

M.E.Johnson (1987). Multivariate Statistical Simulation. NY: Wiley.

M.Lauretto, C.A.B.Pereira, J.M.Stern, S.Zacks (2003). Comparing Parameters of Two Bivariate Normal Distributions Using the Invariant FBST. *Brazilian Journal of Probability and Statistics,* 17, 147-168.

M.Madruga, L.G.Esteves, S.Wechsler (2001). On the Bayesianity of Pereira-Stern Tests. *Test*,10,291–299.

M.R.Madruga, C.A.B.Pereira, J.M.Stern (2003). Bayesian Evidence Test for Precise Hypotheses. *Journal of Statistical Planning and Inference,* 117,185–198.

X.L.Meng, W.H.Wong (1996). Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration. *Statistica Sinica*, 6, 831-860.

D.Ormoneit, V.Tresp (1995). Improved Gaussian Mixtures Density Estimates Using Bayesian Penalty Terms and Network Averaging. *Advances in Neural Information Processing Systems 8*, 542–548. MIT.

C.A.B.Pereira, J.M.Stern, (1999). Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses. *Entropy Journal*, 1, 69–80.

C.A.B.Pereira, J.M.Stern, (2001). Model Selection: Full Bayesian Approach. *Environmetrics*,12,559–68.

G.C.Pflug (1996). *Optimization of Stochastic Models.* Boston: Kluwer.

C.P.Robert (1996). Mixture of Distributions: Inference and Estimation. in Gilks et al. (1996).

S.Russel (1988). Machine Learning: The EM Algorithm. Unpublished note.

G.Schwarz (1978). Estimating the Dimension of a Model. *Ann. Stat.*, 6, 461-464.

J.C.Spall (2003). *Introduction to Stochastic Search and Optimization.* Hoboken: Wiley.

M.Stephens (1997). *Bayesian Methods for Mixtures of Normal Distributions.* Oxford University.

J.M.Stern (1992). Simulated Annealing with a Temperature Dependent Penalty Function. *ORSA Journal on Computing,* 4, 311-319.

J.M.Stern (2003). Significance Tests, Belief Calculi, and Burden of Proof in Legal and Scientific Discourse. Laptec'03, *Frontiers in Artificial Intelligence and its Applications*, 101, 139–147.

J.M.Stern (2004a). Paraconsistent Sensitivity Analysis for Bayesian Significance Tests. SBIA'04, *Lecture Notes Artificial Intelligence*, 3171, 134–143.

J.M.Stern (2004b). Uninformative Reference Sensitivity in Possibilistic Sharp Hypotheses Tests. MaxEnt 2004, *American Institute of Physics Proceedings*, 735, 581–588.

J.M.Stern (2005). Cognitive Constructivism, Eigen-Solutions, and Sharp Statistical Hypotheses. *Proc. 3rd Conference on the Foundations of Information Science, FIS-2005*, 1–23.

J.M.Stern, S.Zacks (2002). Testing the Independence of Poisson Variates under the Holgate Bivariate Distribution. The Power of a New Evidence Test. *Statistical and Probability Letters*, 60, 313–320.