Testing Significance in Bayesian Classifiers

Marcelo de S. Lauretto, Julio M. Stern
BIOINFO and Computer Science Dept., São Paulo University

Abstract. The Fully Bayesian Significance Test (FBST) is a coherent Bayesian significance test for sharp hypotheses. This paper explores the FBST as a model selection tool for general mixture models, and gives some computational experiments for Multinomial-Dirichlet-Normal-Wishart models.

Keywords. Mixture models, classification, significance tests

1. FBST and Model Selection

The Fully Bayesian Significance Test (FBST) is presented by Pereira and Stern, [1], 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 $\dim(\Theta_0) < \dim(\Theta)$. $f(\theta)$ 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^*, the maximum (supremum) of the posterior under the null hypothesis. The integration step consists of integrating the posterior density over the Tangential Set, T where the posterior is higher than anywhere in the hypothesis, i.e.,

$$\text{Ev}(H) = \Pr(\theta \in T \,|\, x) = \int_T f(\theta) d\theta \ , \text{ where}$$
$$T = \{\theta \in \Theta : f(\theta) > f^*\} \quad \text{and} \quad f^* = \sup_H f(\theta)$$

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

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.

2. Dirichlet-Normal-Wishart Mixtures

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. $V > 0$ is a positive definite matrix. 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. Conditioning on the missing data, we get:

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

$$f(X \mid \theta) = \prod_{j=1}^{n} f(x^j \mid \theta) \;=\; \prod_{j=1}^{n} \sum_{k=1}^{m} w_k f(x^j \mid \psi_k)$$

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:

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

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:

$$u = (1/n) \sum_{j=1}^{n} x^j = (1/n)\, X\mathbf{1}$$

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

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:

$$N(u \,|\, n, b, R) = (\tfrac{n}{2\pi})^{d/2} |R|^{1/2} \, \exp\,(\, -(n/2)(u-b)' R (u-b)\,)$$

$$W(S \,|\, e, R) = c^{-1} \, |S|^{(e-d-1)/2} \, \exp\,(\, -(1/2)\mathrm{tr}(S\,R)\,)$$

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

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

The conjugate family of priors for multivariate normal distributions is the Normal-Wishart, see [2]. For the precision matrix R, take as prior 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 [2]. 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:

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

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:

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

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

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

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

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

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

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

$$\ddot{u}^k = (1/\ddot{y}_k)(\dot{n}_k \dot{u}^k + y_k u^k)$$

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

3. 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 nothing but the MCMC generated by cyclically updating variables Z, θ, and P, by drawing θ and Z from the above distributions, see [3,4]. A multinomial variate can be drawn using a uniform generator. A Dirichlet variate w can be drawn using a gamma generator with shape and scale parameters α and β, see [5]. Johnson [6] describes a simple procedure to generate the Cholesky factor of a Wishart variate $W = U'U$ with n degrees of freedom, from the Cholesky factorization of the covariance $V = R^{-1} = C'C$, and a chi-square generator: a) $g_k = G(y_k, 1)$; b) $w_k = g_k \ / \ \sum_{k=1}^{m} g_k$; c) for $i < j$, $B_{i,j} = N(0,1)$; d) $B_{i,i} = \sqrt{\chi^2(n - i + 1)}$; and e) $U = BC$. All subsequent matrix computations proceed directly from the Cholesky factors, [7].

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 identifiable model, see [8]. 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-identifiability 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 [9,10].

The EM algorithm optimizes the log-posterior function $fl(X \mid \theta) + fl(\theta \mid \dot{\theta})$, see [11, 12,13]. 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 $p_k^j = f(z_k^j \mid x^j, \theta)$ the conditional classification probabilities. At each iteration, the log-posterior improvement is:

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

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

$$\delta(\tilde{\theta}, \theta \mid X) \;=\; fl(X \mid \tilde{\theta}) - fl(X \mid \theta) \;=\; \sum\nolimits_j \delta(\tilde{\theta}, \theta \mid x^j)$$

$$\delta(\tilde{\theta}, \theta \mid x^j) \;=\; fl(x^j \mid \tilde{\theta}) - fl(x^j \mid \theta) \;=\; \log \sum\nolimits_k \tilde{w}_k f(x^j \mid \tilde{\psi}_k) \;- fl(x^j \mid \theta) \;=$$

$$=\; \log \sum\nolimits_k \frac{p_k^j}{p_k^j} \frac{\tilde{w}_k f(x^j \mid \tilde{\psi}_k)}{f(x^j \mid \theta)} \;\;\geq\;\; \Delta(\tilde{\theta}, \theta \mid x^j) \;=\; \sum\nolimits_k p_k^j \log \frac{\tilde{w}_k f(x^j \mid \tilde{\psi}_k)}{p_k^j f(x^j \mid \theta)}$$

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

$$Q(\tilde{\theta}, \theta) = \sum\nolimits_{k,j} p_k^j \log \left(\tilde{w}_k f(x^j \mid \tilde{\psi}_k) \right) + fl(\tilde{\theta} \mid \dot{\theta})$$

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

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

$$Q(\tilde{\theta}, \theta) = \sum\nolimits_{k=1}^{m} \sum\nolimits_{j=1}^{n} p_k^j \big(\log \tilde{w}_k + \log N(x^j \mid \tilde{b}^k, \tilde{R}^k) \big) + fl(\tilde{\theta} \mid \dot{\theta})$$

$$fl(\tilde{\theta} \mid \dot{\theta}) = \log D(\tilde{w} \mid \dot{y}) \; + \sum\nolimits_{k=1}^{m} \log NW(\tilde{b}^k, \tilde{R}^k \mid \dot{n}_k, \dot{e}_k, \dot{u}^k, \dot{S}^k)$$

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

$$y = P\mathbf{1} \;\;,\;\; \tilde{w}_k = (y_k + \dot{y}_k - 1) \,/\, \Big(n - m + \sum\nolimits_{k=1}^{m} \dot{y}_k \Big)$$

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

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

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, for example [14].

The EM is a local optimizer, but the MCMC provides plenty of 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

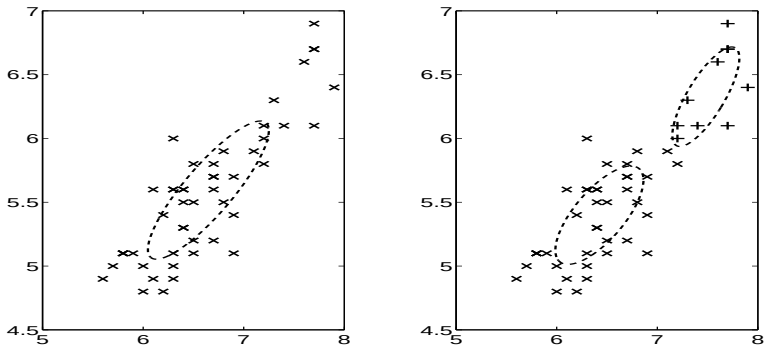

Figure 1. Iris virginica data and models with one (left) and two (right) components

MCMC output using a clustering algorithm, and select a starting point from each cluster. For better efficiency, or more complex problems, the Stochastic EM algorithm can be used to provide starting points near each important local maximum, see [15,16,17].

4. Experimental Tests and Final Remarks

Our test case is the *Iris virginica* data set, with sepal and petal length of 50 specimens (1 discarded outlier), where the botanical problem consists of determining whether or not there are two distinct subspecies in the population, [18,19]. Here, the data X are assumed to follow a mixture of bivariate normal distributions with unknown parameters, including the number of components. Figure 1 presents the dataset and posterior density level curves for the parameters, θ^* and $\widehat{\theta}$, optimized for the 1 and 2 component models.

In the FBST formulation of the problem, the 2 components is the base model, and the hypothesis to be tested is the constraint of having only 1 component. The FBST selects the 2 component model, rejecting H, if the evidence against the hypothesis is above a given threshold, $\text{Ev}(H) > \tau$, and selects the 1 component model, accepting H, otherwise. The threshold τ is chosen by empirical power analysis, see [21,22,23]. Let θ^* and $\widehat{\theta}$ represent the constrained and unconstrained (1 and 2 components) maximum a posteriori (MAP) parameters optimized to the Iris dataset. Generate two collections of t simulated datasets of size n, the first collection at θ^*, and the second at $\widehat{\theta}$. $\alpha(\tau)$ and $\beta(\tau)$ are the empirical type 1 and type 2 statistical errors, i.e., the rejection rate in the first collection and the acceptance rate in the second collection. A small, $t = 500$, calibration run sets the threshold τ so to minimize the total error, $(\alpha(\tau) + \beta(\tau))/2$. Other methods like sensitivity analysis, see [24,25,26], and loss functions, see [27], could also be used.

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. This problem is particularly serious with the Iris dataset because of the small precision, only 2 significant digits, of the measurements. A standard way to avoid this inconvenience is to use flat or minimally informative priors, instead of non-informative priors, see [20]. 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 [20] uses, with similar effects, $\dot{e} = 6$, $\dot{S} = (1.5/n)S$.

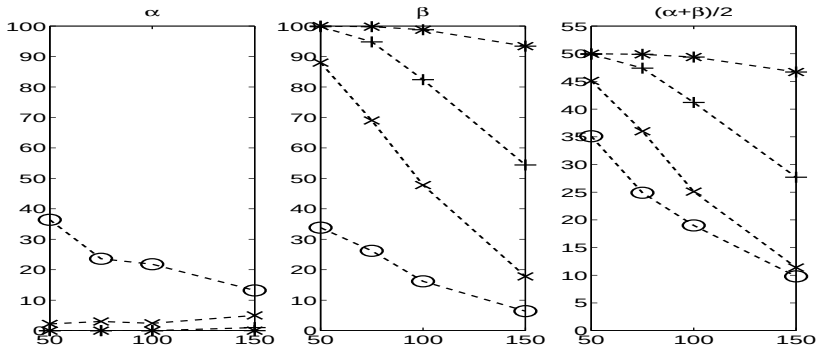

Figure 2. FBST(O), AIC(X), AIC3(+) and BIC(*): Type 1, 2 and total error rates for different sample sizes.

Biernacki and Govaert [28] studied similar mixture problems and compared several selection criteria, pointing as the best overall performers: AIC - Akaike Information Criterion, AIC3 - Bozdogan's modified AIC, and BIC - Schwartz' Bayesian Information Criterion. These are regularization criteria, weighting the model fit against the number of parameters, see [29]. If λ is the model log-likelihood, κ its number of parameters, and n the sample size, then,

$$AIC = -2\lambda + 2\kappa \ , \ \ AIC3 = -2\lambda + 3\kappa \text{ and } \ BIC = -2\lambda + \kappa \log(n) \ .$$

Figure 2 shows α, β, and the total error $(\alpha + \beta)/2$. The FBST outperforms all the regularization criteria. For small samples, BIC is very biased, always selecting the 1 component model. AIC is the second best criterion, caching up with the FBST for sample sizes larger than $n = 150$.

Finally, let us point out a related topic for 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 functional forms. In this case it is usual to call them "separate" models (or hypotheses). Atkinson [30], although in a very different theoretical framework, was the first to analyse this problem using a mixture formulation,

$$f(x \,|\, \theta) = \sum\nolimits_{k=1}^{m} w_k f_k(x, \psi_k) \ .$$

The theory for mixture models presented here can be adapted to analyse the problem of discriminating between separate hypotheses. This is the subject of the authors' forthcoming articles with Carlos Alberto de Bragança Pereira and Basílio de Bragança Pereira.

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

References

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

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

[3] W.R.Gilks, S.Richardson, D.J.Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice.* NY: CRC Press.
[4] O.Häggström (2002). *Finite Markov Chains and Algorithmic Applications.* Cambridge Univ.
[5] J.E.Gentle (1998). *Random Number Generator and Monte Carlo Methods.* NY: Springer.
[6] M.E.Johnson (1987). Multivariate Statistical Simulation. NY: Wiley.
[7] M.C.Jones (1985). Generating Inverse Wishart Matrices. *Comm. Statist. Simula. Computa.* 14, 511–514.
[8] M.Stephens (1997). *Bayesian Methods for Mixtures of Normal Distributions.* Oxford Univ.
[9] C.H.Bennett (1976). Efficient Estimation of Free Energy Differences from Monte Carlo Data. *Journal of Computational Physics* 22, 245-268.
[10] X.L.Meng, W.H.Wong (1996). Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration. *Statistica Sinica*, 6, 831-860.
[11] A.P.Dempster, N.M.Laird, D.B.Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Royal Stat. Soc. B*, 39, 1-38.
[12] 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.
[13] S.Russel (1988). Machine Learning: The EM Algorithm. Unpublished note.
[14] J.M.Martinez (2000). BOX-QUACAN and the Implementation of Augmented Lagrangian Algorithms for Minimization with Inequality Constraints. *Comp. Appl. Math.* 19, 31-56.
[15] 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.
[16] G.C.Pflug (1996). *Optimization of Stochastic Models: The Interface Between Simulation and Optimization*. Boston: Kluwer.
[17] J.C.Spall (2003). *Introduction to Stochastic Search and Optimization.* Hoboken: Wiley.
[18] E.Anderson (1935). The Irises of the Gaspé Peninsula. *Bulletin of the American Iris Society*, 59, 2-5.
[19] G.McLachlan, D.Peel (2000). *Finite Mixture Models.* NY: Wiley.
[20] C.P.Robert (1996). Mixture of Distributions: Inference and Estimation. In [3].
[21] 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.
[22] 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.
[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.
[24] 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.
[25] J.M.Stern (2004a). Paraconsistent Sensitivity Analysis for Bayesian Significance Tests. SBIA'04, *Lecture Notes Artificial Intelligence*, 3171, 134–143.
[26] J.M.Stern (2004b). Uninformative Reference Sensitivity in Possibilistic Sharp Hypotheses Tests. MaxEnt 2004, *American Institute of Physics Proceedings*, 735, 581–588.
[27] M.Madruga, L.G.Esteves, S.Wechsler (2001). On the Bayesianity of Pereira-Stern Tests. *Test*,10,291–299.
[28] C.Biernacki G.Govaert (1998). Choosing Models in Model-based Clustering and Discriminant Analysis. Technical Report INRIA-3509-1998.
[29] C.A.B.Pereira, J.M.Stern, (2001). Model Selection: Full Bayesian Approach. *Environmetrics*, 12, 559–568.
[30] A Method for Discriminating Between Models. *J. Royal Stat. Soc. B*, 32, 323-354.