FBST FOR MIXTURE MODEL SELECTION

Marcelo S. Lauretto, 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.

The computation of the evidence measure used on the FBST is performed in two steps:
1) The optimization step consists of finding f^*, the maximum (supremum) of the posterior under the null hypothesis.
2) The integration step consists of integrating the posterior density over the Tangential Set, T, where the posterior is higher than f^*. This integral is the measure of evidence against the hipothesis H.

This paper proposes the FBST as a model selection tool for multivariate normal mixture models, where the problem is to estimate the number of components in the mixture and the parameters for each component. 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. The FBST selects the m component model, rejecting H, if the evidence against the hypothesis is above a given threshold τ, and selects the $m-1$ component model, accepting H, otherwise.

We perform two simulated numerical experiments and compare the FBST performance with Mclust, a model-based clustering software. Both experimental tests show that:
- the miss rates (proportion of wrong answers) for FBST are significantly lower than Mclust;
- the miss rates convergence towards zero for increasing sample sizes is much faster for FBST than Mclust.

These results show that FBST is a robust tool for mixture models, and strongly encourage further developments.

References:
[1] J.D.Banfield, A.E.Raftery, Model-Based Gaussian and non-Gaussian Clustering. *Biometrics*, 49, 803–821 (1993).
[2] M.R.Madruga, C.A.B.Pereira, J.M.Stern, Bayesian Evidence Test for Precise Hypotheses. *Journal of Statistical Planning and Inference,* 117,185–198 (2003).
[3] J.M.Stern, Uninformative Reference Sensitivity in Possibilistic Sharp Hypotheses Tests. MaxEnt 2004, *American Institute of Physics Proceedings*, 735, 581–588 (2004).

Key Words: gaussian mixture models, significance test, model selection, cluster analysis