Bayesian Model Comparison: Review and Discussion

C. Alston[a], P. Kuhnert[b], S. Low Choy[a], R. McVinish[a], K. Mengersen[a1]

Abstract

This paper provides a brief review of the more popular methods for comparing models in a Bayesian framework. Personal experience in implementing these methods in problems requiring mixture models is also referenced.

1 Introduction

Model comparison is required for a diversity of activities, including variable selection in regression, determination of the number of components in a mixture model or the choice of parametric family. As with frequentist analogues, Bayesian model comparison will not inform about which model is 'true', but rather about the preference for the model given the data and other information. These preferences can be used to choose a single 'best' model or improve estimation via model averaging, in which expected values obtained from different models are weighted by their corresponding posterior probabilities (Congdon, 2001).

In the Bayesian arena, common methods for model comparison are based on the following: *separate estimation* including posterior predictive distributions, Bayes factors and approximations such as the Bayesian information criterion (BIC) and deviance information criterion (DIC); *comparative estimation* including distance measures such as entropy distance or Kullback-Leibler divergence; and *simultaneous estimation*, including reversible jump MCMC and birth and death processes. Each of these is briefly summarised below. Our experiences in applying these approaches to real problems involving mixture models are then described.

2 Description of Methods

2.1 Separate estimation

Consider two models M_1 and M_2, not necessarily nested. If the aim of the modelling is prediction, it is natural in a Bayesian framework to compare models in terms of their *posterior predictive distributions*. Simulations from these distributions can be compared with respect to goodness of fit or proposed inferences (Gelman et al 1995). Posterior predictive p-values and conditional p-values are also emerging as popular measures of model fit (Bayarri and Berger 2000, Perez and Berger 2002, Aitkin et al 2004).

Another natural approach is to compare models on the basis of the *posterior probability of the model* given the data. Using Bayes' rule, this is proportional to the prior probability for the model, $p(M)$ multiplied by the likelihood of the data given the model, $p(y|M)$. Thus the choice between M_1 and M_2 can be made on the basis of the ratio $p(M_2|y)/p(M_1|y) = \{p(M_2)/p(M_1)\} \times \{p(y|M_2)/p(y|M_1)\}$. A large value of this ratio gives support for M_2 over M_1. The second term in this expression, the ratio of the marginal likelihoods, is called the *Bayes factor* BF_{21} (Kass and Raftery 1995). Unlike a likelihood ratio, it is obtained by integrating over θ instead of maximising, so that $p(y|M_i) = \int p(y|M_i, \theta_i)p(\theta_i|M_i)d\theta_i, i = 1, 2$. A change to $2\log(B_{21})$ gives same scale as usual deviance and likelihood ratio statistics. Many variants of the Bayes factor have been proposed (Aitkin 1997, Berger and Pericchi 1998, Chen et al 2000) and their strengths and weaknesses have been actively debated; see for example Gelman *et al* (1995), Congdon (2001), Berger, Ghosh and Mukhopadhyay (2003) and Robert and Casella (2004), and references therein.

[1]Presenter; *a*. School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434 Brisbane; *b*. CMIS CSIRO, Cleveland Brisbane

Because the Bayes factor is often difficult or impossible to calculate, especially for models that involve many random effects, large numbers of unknowns or improper priors, a popular alternative is to adopt an approximation to the Bayes factor. For a model M, the *Bayesian Information Criterion* (BIC) is equal to $\log p(y|\hat{\theta}, M) - p/2 \log n$. The first term is the familiar probability of the data given the model, computed at the value $\hat{\theta}$ that maximises this probability. The second term promotes model parsimony by penalising models with increased model complexity (larger p) and sample size. The *Deviance Information Criterion* (DIC) also penalises against higher dimensional models (Spiegelhalter et al, 1999): with deviance denoted by D, $DIC = E[D(\theta)|y] + \{E[D(\theta|y)] - D(E[\theta|y])\} = \bar{D}(\theta) + p_D$, where $\bar{D}(\theta) = E_\theta[-2logp(y|\theta)|y] + 2\log p(y)$ and p_D denotes the effective number of parameters. It can thus be seen that the DIC comprises terms that are a function of the data alone and a measure of the complexity of the model. The use of the DIC is a topic of ongoing discussion. While its original formulation as described above is appropriate in most generalized linear modelling problems, it fails in other contexts. Celeux et al. (2003) discuss alternative representations of the DIC for latent variable models, including mixtures and missing value problems.

2.2 Comparative estimation

If the 'distance' between two posterior (or posterior predictive) distributions is sufficiently small, the more parsimonious model may be preferred. Such distributional distances can be derived in a variety of ways. Mengersen and Robert (1996) use a Kullback-Liebler measure and define an 'indifference zone' within which models are accepted to be equivalent. Sahu and Cheng (2003) also discuss the use of entropy distances for model comparison.

2.3 Simultaneous model estimation

A number of model choice methods are based on enlarging the parameter space to include all of the models of interest (George and McCulloch 1993, Carlin and Chib 1995). A very popular and conceptually elegant alternative is reversible jump MCMC (RJMCMC), also termed transdimensional MCMC (Green 1995), in which the model itself is conceived as another unknown and the MCMC algorithm is enlarged to allow 'jumps' between models. A prior is required over the model space, but with judicious choice of jumps the number of models does not need to be specified in advance and each model does not require separate estimation. For example, in a mixture context in which the number of components is unknown, Richardson and Green (1997) suggest an additional Metropolis-Hastings step that involves proposals for the 'birth' of a new component or 'death' of an existing component, or a 'split' or 'combine' of two existing components. These moves then require (reversible) bridges to be built between parameters of models of different dimensions, for example the generation of a new mean from the two existing means or the collapse from three to two means. The posterior probability of a model is then estimated by the proportion of times that the particular model is accepted in the MCMC run. The RJMCMC approach has been employed and discussed in many contexts; see Robert and Casella (2004) and references therein.

An alternative *birth and death MCMC* (BDMCMC) formulation was developed by Stephens (2000). Here, the time between jumps to a model of larger dimension (for example the number of MCMC iterations to the next birth of a component) is taken to be a random variable with an underlying rate, and there is an analogous rate of time between steps to lower-dimensional models. In contrast to RJMCMC, moves between models are always accepted and the probability of a model is instead determined by the length of time that the MCMC chain remains in that model. Split and combine steps have been included by Cappé et al (2003) in their consideration of more general continuous time algorithms and comparison with RJMCMC.

3 Applications to Mixture Models

3.1 Model comparison for zero-inflated data

Kuhnert et al. (2004)describe and illustrate Bayesian modelling of the impact of grazing levels on bird density in woodland habitat, with priors based on expert opinion. The authors describe and illustrate the use of the DIC and posterior predictive checks for the comparison of two models that accommodate the excess zeros in these data: the two-component (or conditional) model and the mixture model which allows a point mass at zero. In this case, the mixture model is well defined so the effective number of parameters and the DIC were computable. Moreover, the DIC values were consistent with the graphical posterior predictive checks for both models. A second problem considered by Kuhnert and other coauthors involves quantification of components of flow or discharge entering catchments caused by heavy rainfalls and storm events. This has important consequences for the Great Barrier Reef and surrounding areas. The data may be described by mixtures of ambient or base flow conditions, wet season conditions and storm events. The standard DIC in the context of loads monitoring is unsuitable in this situation so three alternative estimators proposed specifically for mixture models by Celeux et al. (2003) are presently under investigation.

3.2 Mixture models for bioregionalisation

Bioregions are nationally accepted sets of boundaries encompassing areas considered to be homogeneous with respect to broad scale environmental elements such as climate and geology. Subregions further identify areas that share common soils as indicated by broad vegetation groups. Current methods for constructing terrestrial bioregions and subregions are either data-driven or expert-driven. Pullar, Low Choy and Rochester (2004) have investigated the feasibility of a Bayesian model that combines these two sources of information. Mixture models are used to identify regional clusters and provide information about the relative importance, average and range of values for each environmental variable within each region. Regional boundaries derived from expert-driven approaches are used as priors. Model assessment involved a comparison of Bayes factors, residual plots and the eight modified DIC measures proposed by Celeux et al (2003). The favoured measure was DIC_3 in Celeux et al.

3.3 Mixture models for CT scans

The effect of drought or different dietary regimes on sheep can potentially be assessed through the proportion of fat, muscle and bone tissue detected in CT scans. Alston et al (2004) describe and implement a Bayesian mixture representation of the greyscale frequency plots corresponding to the scans, with the number of components determined by the BIC. The application of BDMCMC and RJMCMC was also considered. In this context, the RJCMC returned a much smaller number of preferred components (2-3 compared to 5-7 under BIC) and gave a much poorer fit based on visual plots and posterior predictive checks. It is postulated that the low acceptance rates for birth/death and split/combine models is due to poor separation of components and large sample size. Simulation studies confirm this (Alston et al 2005, in preparation).

3.4 Mixture models for nonparametric density estimation

Nonparametric density estimation on $[0, 1]$ has been investigated using Bernstein polynomials (Petrone 1999) and triangular distributions (Perron and Mengersen 2001) McVinish et al (2004, in preparation) establish conditions for strong and weak consistency of the posterior distribution under two forms of triangular mixtures (fixed partitions and variable weights; fixed weights and variable partitions) that compare favourably with convergence of Bernstein polynomial representations. The behaviour of the Bayes factor in testing a uniform or

parametric family against a nonparametric alternative using these triangular mixtures is also considered. Consistency of the Bayes factor can be obtained provided the nonparametric prior does not place too much probability near the parametric family.

References

Aitkin, M. (1997) The calibration of P-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood (with discussion). *Statist. And Computing* **7**, 253-272.

Aitkin, M., Boys, R.J. and Chadwick, T. (2004) Bayesian point null hypothesis testing via the posterior likelihood ratio. Technical report.

Alston, C.L., Mengersen, K., Thompson, J.M., Littlefield, P.J., Perry, D., Ball, A.J. (2004) Statistical analysis of sheep CAT scan images using a Bayesian mixture model. *Aust. J. Agricultural Research* **55**, 57-68.

Bayarri, M.J. and Berger, J. (2000) P-values for composite null models (with discussion). *J. American Statist. Assoc.* **95**, 1127-1142.

Berger, J., Ghosh, J.K. and Mukhopadhyay, N. (2003) Approximations to the Bayes factor in model selection problems and consistency issues. *J. Statist. Planning and Inference* **112**, 241-258.

Berger, J. and Pericchi, L. (1998) Accurate and stable Bayesian model selection: the median intrinsic Bayes factor. *Sankhya B***60**, 1-18.

Cappé, O., Robert, C. and Rydén, T. (2002) Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers. *J. Royal Statist. Society Series B* **65**(3), 679-700.

Carlin, B. and Chib, S. (1995) Bayesian model choice via Markov chain Monte Carlo methods. *J. Royal Statist. Society Series B* **57**(3), 473-484.

Congdon, P. (2001) *Bayesian Statistical Modelling*. Wiley, England.

Cappé, O., Robert, C. and Rydén, T. (2002) Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers. *J. Royal Statist. Soc. Series B*, **65**(3), 679-700.

Celeux, G., Forbes, F., Robert, C.P. and Titterington, D.M. (2003) Deviance information criteria for missing data models. *Cahiers du Ceremade* 0325.

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

George, E. and McCulloch, R. (1993) Variable selection via Gibbs sampling. *J. American Statist. Association* **88**(423), 881-889.

Green, P. (1995) Reversible jump MCMC computation and Bayesian model determination. *Biometrika* **82**(4), 711-732.

Kass, R. and Raftery, A. (1995) Bayes factors. *J. American Statist. Assoc.* **90**, 773-795.

Kuhnert, P.M., Martin, T.G., Mengersen, K. and Possingham, H.P. (2004) Assessing the impacts of grazing levels on bird density in woodland habitat: A Bayesian approach using expert opinion, *Environmetrics*. In press.

Mengersen, K. and Robert, C.P. (1996) Testing for mixtures: a Bayesian entropic approach. *In* Bayesian statistics, 5 (Alicante, 1994), 255-276, Oxford Sci. Publ., Oxford Univ. Press, New York.

Perez, J.M. and Berger, J. (2002) Expected posterior prior distributions for model selection. *Biometrika* **89**, 491-512.

Perron, F. and Mengersen, K. (2001) Bayesian nonparametric modelling using mixtures of triangular distributions. *Biometrics* **57** 518-528.

Petrone, S. (1999) Bayesian density estimation using Bernstein polynomials. *Canadian J. Statist.* **27** 105-126.

Richardson, S. and Green, P. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Statist. Soc. Series B* **59** 731-792.

Robert, C. and Casella, G. (2004) *Monte Carlo Statistical Methods*. Springer-Verlag, New York, second edition.

Sahu, S. and Cheng, R. (2003) A fast distance based approach for determining the number of components in mixtures. *Canadian J. Statistics* **31**, 3-22.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A. (2002) Bayesian measures of model complexity and fit. *J. Royal Statist. Society Series B* **64**(3), 583-639.

Stephens, M. (2000) Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *Ann. Statist.* **28**, 40-74.