STATISTICAL CHALLENGES IN MODERN ASTRONOMY IV
ASP Conference Series, Vol. 371, 2007
G. Jogesh Babu and Eric D. Feigelson

Discussion on "Current Challenges in Bayesian Model Choice" by Clyde et al.

William H. Jefferys

University of Texas at Austin, Department of Astronomy, 1 University Station, Austin, Texas 78712-0259, and Department of Mathematics and Statistics, University of Vermont, Burlington, VT 05401

Abstract. It is quite common to encounter problems in astronomy that require selecting amongst several models. Bayesian model selection is gaining in popularity because of its power and logical consistency. My comment here are based on our experiences in the recently-concluded SAMSI workshop on exoplanets.

1. Why Model Selection is Important in Astronomy

Model selection problems arise in many contexts in astronomy. One of these comes from the use of empirical models, where for example we want to fit a function (e.g., a light curve, the initial mass function) using basis functions of some sort (e.g., polynomials, splines, trigonometric functions, wavelets). In these empirical models we do not believe that the function being fitted is actually a polynomial (for example), but rather regard it as a useful approximation. In these problems we do not want either to underfit or overfit, but rather to choose an adequate representation. In the Bayesian context we may wish to do *model averaging*, that is, the fitted function is taken to be an average over those representations that have the highest posterior probability.

In other contexts, the models being compared may be physically believable. For example, when looking at radial velocity data on a star, the star might realistically have zero, one, two, or more planets. Or, the eccentricity of a planetary companion might be essentially zero (because of circularization mechanisms) or nonzero. Similarly, when examining the color-magnitude diagram of a cluster, a given star might be a cluster member or a field star, or a single or multiple star. All of these problems can be discussed in a uniform way using Bayesian model selection techniques.

2. Mathematical Considerations

In all such problems, we are comparing models that live on spaces of differing dimensionalities. Thus, given models $m \in M$ with parameters $\theta_i \in \Theta_m$, where $i = 1, ..., N_m$, we expect that N_m will vary from model to model. Models may be nested (that is, $\Theta_1 \subset \Theta_2 \subset \ldots \subset \Theta_m \subset \ldots$) or unnested. In the latter case,the parameters on one model need not bear any physical relationship to those in another model.

Frequentist approaches to model selection generally require nested models. A particular model is chosen as the "null model", and the tail area of the probability density under the null model that lies beyond the observed data is calculated (a "p-value"). The p-value is often misinterpreted as "the probability that the null model is true" or "the probability that the results were obtained by chance." Neither of these interpretations is correct, and indeed the interpretation of p-values is problematic.

In the Bayesian approach, no particular model is distinguished as a null model. Prior probabilities are assigned to each model under consideration, and their posterior probabilities computed. Both nested and unnested models can be handled. And, the interpretation of the posterior probabilities as "the probability that each model is true" is both natural and correct.

3. Priors

As the paper points out, selection of priors is critical, and much more difficult than in simple parameter fitting problems under a specified model. In different models the "same" parameter may have a different interpretation, and generally will require a different prior. Also, improper priors are generally disallowed in model selection problems, as they are only defined up to an arbitrary multiplicative constant, resulting in marginal likelihoods that contain arbitrary multiplicative factors.

General prescriptions for prior specification do not exist, although some useful rules are available in special cases, e.g., Zellner-Siow g-priors in linear problems, and rules using a "training sample," a subset of the data, such as Intrinsic Bayes Factors, Fractional Bayes Factors, and Expected Posterior Priors. Care must be taken to compensate for the fact that these priors are based on the data, so as to avoid using the data twice.

4. Computation

The other major difficulty is computational. We must evaluate integrals of the form

$$m(x) = \int_{\Theta_m} f(x \mid \theta_m)\pi(\theta_m)d\theta_m \tag{1}$$

where x is the (fixed) data, f is the likelihood, π the prior, and the dimension of the space Θ_m over which the integral is to be computed is in general large. This equation comes from writing Bayes' theorem in the form

$$m(x)\pi(\theta_m \mid x) = f(x \mid \theta_m)\pi(\theta_m) \tag{2}$$

and integrating over Θ_m, noting that $m(x)$ is independent of θ_m and can be removed from under the integral sign, and that the posterior density $\pi(\theta_m \mid x)$ is a normalized.

Because of high dimensionality, in many real problems this integral can be quite challenging and even unfeasable to calculate. Many appealing methods suffer from the "curse of dimensionality," that is, they only work in lower dimensions. These include cubature methods and importance sampling. When they work, they can work well, but beyond about 20 dimensions they begin to fail.

Because we may already have spent a good deal of time and effort producing an MCMC sample from the posterior distribution under each model, it is tempting to try to use this sample to somehow bootstrap this information into estimates of the marginal likelihoods. Unfortunately, this is more difficult than it seems at first glance. For example, we can derive a "harmonic mean" estimate (Newton & Raftery 1994) by writing Bayes' theorem in the form

$$\frac{\pi(\theta_m)}{m(x)} = \frac{\pi(\theta_m \mid x)}{f(x \mid \theta_m} \tag{3}$$

Integrating,

$$\frac{1}{m(x)} = \int_{\Theta_m} \frac{\pi(\theta_m \mid x)}{f(x \mid \theta_m)} d\theta_m \tag{4}$$

leading to the estimate

$$\frac{1}{m(x)} \approx \left\langle \frac{1}{f(x \mid \theta_m)} \right\rangle \tag{5}$$

where the average indicated by the angle brackets is taken over the MCMC sample $\Theta_m^* = \{\theta_m^*\}$ from the posterior distribution.

Unfortunately, Eq. (5) suffers from having infinite variance. Experience shows that even though a large sample may appear to have converged, a single sample way out in the tails will cause the value to shift by a large amount, indicating lack of convergence.

Gelfand and Dey (1994) proposed multiplying Eq. (3) by a proper density $q(\theta_m)$, dividing through by $\pi(\theta_m)$ and integrating to obtain the estimate

$$\frac{1}{m(x)} \approx \left\langle \frac{q(\theta_m)}{f(x \mid \theta_m)\pi(\theta_m)} \right\rangle \tag{6}$$

This idea works, but it is difficult to implement in practice since it is not easy to choose a reasonable tuning function q, particularly in high dimensions. q needs to have thin tails relative to the posterior density.

Jim Berger's "Crazy Idea #1" proposed defining an importance function q as a mixture over a subset of the MCMC samples, with kernels that are t distributions with 4 degrees of freedom (so as to obtain a "fat-tailed" importance function). Then, by drawing a sample from q, which is relatively easy to do, the marginal likelihood can be approximated in the usual way by

$$m(x) \approx \left\langle \frac{f(x \mid \theta_m)\pi(\theta_m)}{q(\theta_m)} \right\rangle \tag{7}$$

This worked in low dimensions but is expected to fail in high dimensions.

Chib (1995) proposed solving Bayes' theorem for the marginal likelihood and evaluating at any point θ_m^* in the sample space, i.e.,

$$m(x) = \frac{f(x \mid \theta_m^*)\pi(\theta_m^*)}{\pi(\theta_m^* \mid x)} \tag{8}$$

However, one must be able to accurately estimate the posterior density at the point of evaluation. This may be very difficult to do in multimodal problems such as those that arise in the exoplanet problem, because many discrete periods for the planet may fit the data more or less well.

Berger's "Crazy Idea #2" starts with the Chib-like identity

$$m(x)\pi(\theta_m \mid x)q(\theta_m) = f(x \mid \theta_m)\pi(\theta_m)q(\theta_m) \tag{9}$$

He integrates and divides through to obtain

$$m(x) = \frac{\int f(x \mid \theta_m)\pi(\theta_m)q(\theta_m)d\theta_m}{\int \pi(\theta_m \mid x)q(\theta_m)d\theta_m} \tag{10}$$

The integral in the denominator is approximated by averaging $f\pi$ over a draw from q, while the one in the numerator is approximated by averaging q over a sample of the MCMC draws from the posterior – preferably a sample independent of that used to define q. A possible drawback of this idea is that it may oversample the modes, since each integrand is approximately the square of the posterior density.

Some methods, such as reversible-jump MCMC, can work in high dimensions, but they can be sensitive to the proposal distributions on the parameters. Poor proposals can lead to the sampler getting "stuck" on particular models with consequent poor mixing. Multimodal distributions are difficult. Parallel tempering, in which models are run in the background that mix more easily, can help alleviate this problem, as Gregory's paper (elsewhere in this volume) illustrates.

We also considered Skilling's (2004) "nested sampling" idea. This reduces a high-dimensional problem to a one-dimensional problem and is in theory quite attractive. However, we found the method to be highly sensitive to the choice of tuning parameters, and the nested sampling step to be problematic. Results were often very far off from the known values, even in simple problems.

5. The Bottom Line

Bayesian model selection is easy in concept but difficult in practice. Great care must be exercised when choosing priors. Computation is difficult in real-world problems, and no single method is likely to be useful in every situation. Calculational methods must therefore be chosen carefully, on a case-by-case basis. And finally, when several methods are available, it is useful to compare the results using these different methods.

6. Acknowledgements

I thank Jim Berger, Jogesh Babu and Eric Feigelson for organizing the 2006 SAMSI astrostatistics program and SCMA IV and for for financial assistance that made it possible for me to participate in both. I thank my many colleagues on the exoplanet working group, especially Merlise Clyde, who did the bulk of the writing on the paper that I am commenting on, and Floyd Bullard for his service as working group secretary and webmaster.