

Current Challenges in Bayesian Model Choice

M. A. Clyde¹, J. O. Berger¹, F. Bullard¹, E. B. Ford²,
W. H. Jefferys³, R. Luo¹, R. Paulo⁴, and T. Loredo⁵

¹ *Institute of Statistics & Decision Sciences, Duke University, Durham, NC 27708-0251 USA*

² *Department of Astronomy, University of California Berkeley, Berkeley, CA 94720-3411 USA*

³ *Department of Astronomy, University of Texas, Austin, TX 78712 USA*

⁴ *Department of Mathematics, University of Bristol, Bristol, BS8 1TW UK*

⁵ *Department of Astronomy, Cornell University, Ithaca, NY 14853 USA*

Abstract. Model selection (and the related issue of model uncertainty) arises in many astronomical problems, and, in particular, has been one of the focal areas of the Exoplanet working group under the SAMSI (Statistics and Applied Mathematical Sciences Institute) Astrostatistics Exoplanet program. We provide an overview of the Bayesian approach to model selection and highlight the challenges involved in implementing Bayesian model choice in four stylized problems. We review some of the current methods used by statisticians and astronomers and present recent developments in the area. We discuss the applicability, computational challenges, and performance of suggested methods and conclude with recommendations and open questions.

1. Introduction

Model selection (and the related issue of model uncertainty) arises in many astronomical problems, and, in particular, has been one of the focal areas of the 2006 Astrostatistics Program at the Statistics and Applied Mathematical Sciences Institute. In this paper, we provide an overview of some of the recent developments in the Bayesian approach to model selection.

1.1. General Setting and Notation

One of the advantages of the Bayesian paradigm is its conceptual simplicity with regards to model choice. For the observable data \mathbf{Y} , we entertain M models, $\mathcal{M}_1, \dots, \mathcal{M}_M$, each defined by a density function $f_m(\cdot | \boldsymbol{\theta}_m)$ indexed by a model-specific parameter vector $\boldsymbol{\theta}_m \in \Theta_m$, $m = 1, \dots, M$. The parameter spaces Θ_m may be of differing dimensions, however, there is no need to restrict attention to nested models where $\Theta_k \subset \Theta_{k'}$, as in likelihood ratio tests. For the goal of making inference about models, the index \mathcal{M}_m is no different from any other parameter; hence in the Bayesian paradigm, we express prior uncertainty

regarding the collection of all unknowns $\{\boldsymbol{\theta}_m, \mathcal{M}_m, m = 1, \dots, M\}$ through a joint prior distribution:

$$\boldsymbol{\theta}_m \mid \mathcal{M}_m \sim p(\boldsymbol{\theta}_m \mid \mathcal{M}_m) \quad (1)$$

$$\mathcal{M}_m \sim p(\mathcal{M}_m). \quad (2)$$

In effect, these priors serve to embed the various separate models within one large hierarchical mixture model for the data.

Bayes theorem leads to a joint posterior distribution for $\boldsymbol{\theta}_m, \mathcal{M}_m$, which, as with the prior distributions, is conveniently expressed through the conditional distribution for model specific parameters $p(\boldsymbol{\theta}_m \mid \mathcal{M}_m, \mathbf{Y})$ and the marginal posterior distribution of models $p(\mathcal{M}_m \mid \mathbf{Y})$:

$$p(\mathcal{M}_m \mid \mathbf{Y}) = \frac{m(\mathbf{Y} \mid \mathcal{M}_m)p(\mathcal{M}_m)}{\sum_k m(\mathbf{Y} \mid \mathcal{M}_k)p(\mathcal{M}_k)} \quad (3)$$

where

$$m(\mathbf{Y} \mid \mathcal{M}_m) = \int f_m(\mathbf{Y} \mid \boldsymbol{\theta}_m)p(\boldsymbol{\theta}_m \mid \mathcal{M}_m)d\boldsymbol{\theta}_m, \quad (4)$$

is the marginal distribution of the data under model \mathcal{M}_m . This is also known as the prior predictive distribution of the data or is referred to as the integrated likelihood or marginal likelihood of the model (up to any *common* constants of proportionality).

Based on these posterior probabilities of models, pairwise comparison of any two models is given by the posterior odds

$$\frac{p(\mathcal{M}_k \mid \mathbf{Y})}{p(\mathcal{M}_j \mid \mathbf{Y})} = \frac{m(\mathbf{Y} \mid \mathcal{M}_k)}{m(\mathbf{Y} \mid \mathcal{M}_j)} \times \frac{p(\mathcal{M}_k)}{p(\mathcal{M}_j)}. \quad (5)$$

which is a function of the prior odds, $O[\mathcal{M}_k : \mathcal{M}_j] \equiv \frac{p(\mathcal{M}_k)}{p(\mathcal{M}_j)}$, and the Bayes factor

$$B[\mathcal{M}_k : \mathcal{M}_j] \equiv \frac{m(\mathbf{Y} \mid \mathcal{M}_k)}{m(\mathbf{Y} \mid \mathcal{M}_j)} \quad (6)$$

which summarizes the relative support for \mathcal{M}_k versus \mathcal{M}_j provided by the data. In fact, the posterior model probabilities (3) may be expressed entirely in terms of Bayes factors and prior odds as

$$p(\mathcal{M}_m \mid \mathbf{Y}) = \left[1 + \sum_{\mathcal{M}_k \neq \mathcal{M}_m} O[\mathcal{M}_k : \mathcal{M}_m] B[\mathcal{M}_k : \mathcal{M}_m] \right]^{-1}. \quad (7)$$

In many cases, the prior odds are taken to be equal, so that the posterior odds simplify to the Bayes factor. The posterior probabilities of models provide a complete representation of post-data model uncertainty that can be used for a variety of inferences and decisions. In this setting, model selection is conceptually no different from inference on the ‘‘parameter’’ \mathcal{M}_m . If a zero-one loss

function is assumed, the ‘best’ model is therefore the one with highest posterior probability, i.e., model \mathcal{M}_{m^*} with m^* satisfying

$$m^* = \operatorname{argmax}_{m \in \{1, \dots, M\}} p(\mathcal{M}_m | \mathbf{Y}),$$

the modal model for which $p(\mathcal{M}_m | \mathbf{Y})$ is largest. Model selection may be useful for testing a theory represented by one of a set of carefully studied models, or it may simply serve to reduce attention from many speculative models to a single useful model. However, in problems where no single model stands out, it may be preferable to report a set of high posterior models along with their probabilities to convey the model uncertainty given the available information. Bayesian model averaging is an alternative to Bayesian model selection that incorporates rather than ignores model uncertainty. For example, suppose interest focused on the distribution of \mathbf{Y}_f , future observations from the same process that generated \mathbf{Y} . The Bayesian predictive distribution of \mathbf{Y}_f is obtained as

$$p(\mathbf{Y}_f | \mathbf{Y}) = \sum_m p(\mathbf{Y}_f | \mathcal{M}_m, \mathbf{Y}) p(\mathcal{M}_m | \mathbf{Y}), \quad (8)$$

a posterior weighted mixture of the conditional predictive distributions

$$p(\mathbf{Y}_f | \mathcal{M}_k, \mathbf{Y}) = \int p(\mathbf{Y}_f | \boldsymbol{\theta}_m, \mathcal{M}_m) p(\boldsymbol{\theta}_m | \mathcal{M}_m, \mathbf{Y}) d\boldsymbol{\theta}_m. \quad (9)$$

The mixture model under model averaging may be more appropriate, for example, in design problems for selecting the next set of observations (Loredo & Chernoff 2003). For an overview of Bayesian model averaging and accounting for model uncertainty, see the review articles by Hoeting et al. (1999) and Clyde & George (2004).

Despite the apparent simplicity of the above description, model selection is actually nontrivial to implement in practice in a variety of situations. Although different contexts produce different challenges that will be discussed in more detail later, one difficulty that is always present is that of specifying prior distributions on the model-specific parameters $\boldsymbol{\theta}_m$. We discuss this aspect briefly, before turning to the challenge of obtaining posterior model probabilities.

1.2. Prior Specification

Specifying priors in a subjective fashion is most of the time infeasible because of either lack of substantive prior information regarding *joint* distributions or because of the large amount of models involved in the analysis. We provide a few words of warning about prior specifications.

CAUTIONS:

1. Although one might use the same symbol to denote a particular parameter across models, say a regression coefficient associated with a particular covariate, in actuality that parameter can in general have a different interpretation — and hence will need its own prior distribution — across distinct models. Subjective elicitation of the full joint distribution being precluded, in most circumstances one has to resort to specification of the priors by means of some formal method.

2. In general, the use of improper priors is not permitted in the context of model selection, the reason for that being that Bayes factors, and hence posterior model probabilities, are indeterminate in this circumstance. To make this point clear, note that improper priors are determined only up to an arbitrary multiplicative constant. As a consequence, recalling (4), so are marginal distributions and hence Bayes factors. In inference given a model, these arbitrary multiplicative constants cancel in the formula for the posterior distribution of the model-specific parameters; but unfortunately remain in the expressions for marginal likelihoods. Once model uncertainty is included in the analysis, those constants are important and have a high impact on the final results.

There are a few circumstances under which the use of the same arbitrary constant across models is justifiable (see Berger & Pericchi 2001; Berger et al. 1998, for discussion and examples). On the other hand, many objective Bayesian model selection techniques, say the Intrinsic Bayes Factor (IBF) (Berger & Pericchi 1996b,a, 1998), Fractional Bayes Factors (FBF) (O’Hagan 1995), and Expected Posterior (EP) Prior (Pérez & Berger 2000) are essentially ways of ‘calibrating’ these arbitrary multiplicative constants by means of some argument that guarantees that the resulting (formal) Bayes factor is indeed a sensible quantity to base inference upon. IBFs and FBFs use the idea of “training” samples to convert an improper prior (reference priors Bernardo 1979, are recommended) into a proper “posterior” distribution for θ_m that serves as a proper prior for updating using the remaining data. In the case of IBFs, a subset of the data is used as a training sample, while with FBFs a fraction b/n of the likelihood is used. Berger (1997) provides illustrations from astronomy while Berger & Pericchi (2001) review and contrast these methods with BIC and some conventional prior specifications. Because obtaining marginal likelihoods or Bayes factors with these objective priors involves the same degree of difficulty as with other conventional priors, the discussion about computations in later sections will apply.

3. Another approach to the problem is simply to derive priors on a case-by-case basis, based on arguments that are specific to the problem at hand. A word of caution against another common pitfall: since the use of improper priors is in general prohibited, one might consider the use of proper but diffuse or vague priors as a reasonable solution to the problem. It turns out that this is in general not a good idea, given the fact that often the final conclusions are essentially a function of the (arbitrary) degree of vagueness embodied in these priors. Consider a parameter θ_1 that is uniform on some finite interval, where the normalizing constant is the length of the interval. As this quantity is present in the model with θ_1 , but not others, the length of the interval acts as a penalty, hence if it is taken to be arbitrarily large, we may inadvertently over-penalize the more complex model with θ_1 . This illustrates another difficulty of model selection — its extreme sensitivity to prior specification. Inference given a model is in general relatively insensitive to small changes in the prior; for inferential problems that incorporate model uncertainty that is usually not the case, so that extra care is needed for prior specification.

We now describe four general scenarios for model selection that highlight some of the computational challenges that arise.

1.3. Scenarios for Model Selection

Whether one chooses to select a model or incorporate uncertainty regarding models, posterior model probabilities are a key component for posterior inferences and provide several computation challenges for the implementation of the Bayesian paradigm. These challenges for model selection depend on the class of problem under consideration, but may be broadly broken down based on the number of models under consideration and the availability or tractability of marginal likelihood calculations (see Clyde & George (2004) for additional references).

Enumerable Model Space and Tractable Marginals

In this scenario, the space of models is small enough that enumeration of all possible models is feasible and marginal likelihoods may be easily computed in closed form. This arises in a limited number of exponential family models with conjugate priors on the coefficients θ_m in each model, such as linear regression with Gaussian errors or decomposable graphical models (discrete multinomial or continuous multivariate normal) where the normalizing constants required to compute the marginal likelihoods are readily available. In these problems, no sampling is required as all distributions are available in closed form. In general, these are also models for which objective Bayes factors are easy to compute.

Innumerable Model Space, but Tractable Marginals

This scenario is exactly the same as the above, but the number of models is too large to enumerate. Examples include nonparametric regression models with Gaussian errors using Fourier, wavelet, or spline bases, for example. Because of the conjugate framework the posterior distributions of θ_m given \mathcal{M}_m are known, thus one may bypass sampling model specific parameters. Stochastic search algorithms base on constructing a Markov chain on the space of model, such as the SSVS algorithm of George & McCulloch (1993, 1997) or the MC³ algorithm of Madigan & Jeremy C. York (1995) or importance sampling Clyde et al. (1996), may be used to identify high probability models for model selection or to obtain posterior weighted estimates for model averaging. Because the marginal likelihoods for sampled models are readily obtainable once a model is sampled, exact comparisons between any two sample models may be made through Bayes factors (6). For calculation of posterior model probabilities, it is common to replace the summation in the denominator of (3) with a sum over the collection of sampled models. The primary challenge here is that of searching over a large discrete space (an NP hard problem).

Innumerable Model Space and Intractable Marginal Likelihoods

In this scenario because neither the marginal likelihoods or posterior distributions of model specific parameters are available, posterior inference is typically based on creating a trans-dimensional Markov chain for exploring the joint

parameter and model spaces. The recent review paper of Sisson (2005) provides an excellent summary of many current trans-dimensional methods. References of the latter include the reversible jump Markov Chain Monte Carlo (MCMC) sampler of Green (1995), the product space search of Carlin & Chib (1995), the metropolized Carlin and Chib method of Dellaportas et al. (2002) and the composite model space approach of Godsill (2001), which are variations on Metropolis-Hastings algorithms over the joint model and parameter space.

Enumerable Model Space, but Intractable Marginal Likelihoods

In this scenario, one may create a trans-dimensional MCMC algorithm to traverse both model and parameter spaces as above. Because the number of models is limited, one may alternatively choose to calculate (approximately) all marginal likelihoods or Bayes factors and use (3) or (7) to calculate the posterior model probabilities. In this group, there is perhaps the widest variety of algorithms (stochastic and deterministic) that provide approximate marginal likelihoods or ratios of marginal likelihoods. The models for detecting exoplanets using radial velocity measurements is an excellent example of this type of problem (see Ford, this volume).

There are three main approaches for estimating normalizing constants: analytic approximations (often requiring large samples), such as Laplace approximations (Tierney et al. 1989; DiCiccio et al. 1997) or the Bayes Information Criterion (Kass & Raftery 1995), numerical integration (quadrature) (see Evans & Swartz 1995), and Monte Carlo methods. Of which, the latter provide the only general approach in complex, high dimensional problems. In the remainder of this paper, we highlight some of the most popular and important methods and provide recommendations for their implementation. Many of these methods have been implemented during the SAMSI 2006 Astrostatistics Program in the ExoPlanets working group (see Ford, this volume for the application to discovery of exoplanets). We begin with a discussion of methods for evaluating the marginal likelihood for a single model. We next move to approaches that aim to estimate a ratio of normalizing constants, namely, a Bayes factor, and then follow up with a discussion of trans-dimensional methods. We conclude with a section that summarizes our findings and provides recommendations based on our experience.

2. Estimating Marginals

The methods that we will discuss in this section are designed to approximate the integral in (4) for a given model, and are best suited for model spaces that can be enumerated, although may in principle be used to refine estimates from the trans-dimensional methods discussed in Section 4. To simplify notation, we will drop the conditioning on \mathcal{M}_m .

For low dimensional parameter spaces, fixed and adaptive quadrature rules, perhaps after suitable transformations of variables, can provide highly accurate numerical estimates of the required integrals. Product quadrature rules, (repeated application of a one-dimensional integration rule), however, are well known to suffer from the curse of dimensionality, as the number of points needed to evaluate the integral for a desired degree of accuracy increases exponentially

as dimension increases. For example, in a seven dimensional space, to achieve an accuracy of $O(1/T)$, we will need on the order of T^7 function evaluations. Monte Carlo and quasi-Monte Carlo are typically viewed as being superior, as they require fewer function evaluations for the same degree of accuracy (at least theoretically).

2.1. Monte Carlo Estimates

For obtaining an estimate of the marginal likelihood, the simplest stochastic integration rule is to draw an iid sample of size T , $(\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(T)})$, from the prior $p(\boldsymbol{\theta})$, and approximate the integral by the Monte Carlo (MC) sum

$$\hat{m}(\mathbf{Y})_{MC} = \frac{1}{T} \sum_t f(\mathbf{Y} | \boldsymbol{\theta}^{(t)}). \quad (10)$$

Unfortunately, when the prior is diffuse relative to the likelihood, many of the prior draws will fall in areas of low probability; in higher dimensions, finding regions of high likelihood is even more difficult due to the curse of dimensionality, which leads to a larger variance and, hence, a slower rate of convergence. For vague prior distributions, the Monte Carlo estimate will generally under-estimate the marginal likelihood, as highly peaked areas are typically under-sampled. Because MCMC is routinely used for exploring posterior distributions, estimation of marginal likelihoods from the output of the Markov chain does not require significant additional computational effort.

2.2. Harmonic Mean

Using the relation

$$\frac{1}{m(\mathbf{Y})} = \int \frac{1}{f(\mathbf{Y} | \boldsymbol{\theta})} p(\boldsymbol{\theta} | \mathbf{Y}) d\boldsymbol{\theta} \quad (11)$$

Newton & Raftery (1994) proposed the harmonic mean of the likelihood

$$\hat{m}(\mathbf{Y})_{HM} = \left[\frac{1}{T} \sum_{t=1}^T 1/f(\mathbf{Y} | \boldsymbol{\theta}^{(t)}) \right]^{-1} \quad (12)$$

evaluated at the posterior draws $(\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(T)})$ as an estimator of the marginal likelihood. While Newton & Raftery (1994) show that this converges almost surely to the required marginal, the partial sums of $1/f(\mathbf{Y} | \boldsymbol{\theta}^{(t)})$ do not always obey a Gaussian central limit theorem, hence the rate of convergence may be extremely slow. Wolpert (2002) shows that the limiting distribution is a positive Stable law with index α , ($1 < \alpha \leq 2$). While there are cases where the Gaussian CLT applies ($\alpha = 2$), in general the rate of convergence will be extremely slow ($T^{1/\alpha-1}$) when α is near one, as is the case with vague prior information and informative likelihoods. Unfortunately, the slow convergence make diagnosing problems with the estimator all the more difficult!

Weighted Harmonic Mean Since the inverse likelihood in the harmonic mean estimator may not have finite variance, multiplying by the inverse likelihood by a density with thinner tails provides a ratio with finite variance. Gelfand & Dey (1994) suggest a generalization of the Newton & Raftery (1994) estimator based on the integral,

$$\frac{1}{m(\mathbf{Y})} = \int \frac{q(\boldsymbol{\theta} | \mathbf{Y})}{f(\mathbf{Y} | \boldsymbol{\theta})p(\boldsymbol{\theta})} p(\boldsymbol{\theta} | \mathbf{Y}) d\boldsymbol{\theta} \quad (13)$$

where q plays the role of an importance sampling density (with known normalizing constant) leading to the weighted harmonic mean estimator:

$$\hat{m}(\mathbf{Y})_{WHM} = \left[\frac{1}{T} \sum \frac{q(\boldsymbol{\theta}^{(t)})}{f(\mathbf{Y} | \boldsymbol{\theta}^{(t)})p(\boldsymbol{\theta}^{(t)})} \right]^{-1}. \quad (14)$$

Note that unlike a typical importance sampling scheme where q should have heavier tails than the product of likelihood and prior, here the importance distribution q is in the numerator, so that usual roles have been reversed. This estimator also converges to $m(\mathbf{Y})$ and will have finite variance avoiding the instability problems of the Newton-Raftery harmonic mean estimator when the thinness condition is met. Unfortunately, as Chib (1995) notes, the tuning function q is quite difficult to determine, particularly in high dimensional problem; he found that choices such as multivariate normals with mean and covariance equal to their posterior counterparts do not appear to satisfy the thinness requirements.

Partial Marginalization In some problems, it is possible to integrate out a subset of the parameters and then apply the harmonic mean estimator using the partially marginalized likelihood (Satagopan et al. 2000). While there is no guarantee that this will lead to a finite variance, partial marginalization in the examples of Satagopan et al. (2000) leads to heavier tailed distributions, with finite variance for the inverse likelihood and acceleration of convergence. For problems where a subset of the parameters may be integrated analytically, this is a promising direction.

2.3. Chib's method

Chib's method (and others) are based on the following trivial but fundamental identity based on manipulating Bayes' Theorem:

$$m(\mathbf{Y})_C = \frac{p(\boldsymbol{\theta}) f(\mathbf{Y} | \boldsymbol{\theta})}{p(\boldsymbol{\theta} | \mathbf{Y})} \quad (15)$$

which is valid for all $\boldsymbol{\theta}$. The expression is often evaluated at a point of high posterior density $\boldsymbol{\theta}^*$, such as the posterior mode, posterior mean, or even the maximum likelihood estimate, depending on what is readily available. Since usually the likelihood and prior are available in closed form, as long as one can estimate the posterior at a particular point, $p(\boldsymbol{\theta}^* | \mathbf{Y})$, one immediately has an estimate of the marginal likelihood $m(\mathbf{Y})$. Chib (1995), together with Chib & Jeliazkov (2001), provide a framework under which virtually any model that can be fit using MCMC techniques can have its marginal likelihood estimated. We summarize the estimators in the simplest possible frameworks.

Chib: Two-block Gibbs Sampler Suppose $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and that both full conditionals are available in closed form, including normalizing constants. The joint distribution may be factored as

$$p(\boldsymbol{\theta}^* | \mathbf{Y}) = p(\boldsymbol{\theta}_1^* | \mathbf{Y}) p(\boldsymbol{\theta}_2^* | \boldsymbol{\theta}_1^*, \mathbf{Y}) .$$

where the second term of the right-hand side is available by assumption, whereas the first term can be estimated using a Rao-Blackwellized estimate of the density as in the conditional marginal density estimate (CMDE) of Gelfand, Smith, & Lee (1992),

$$\hat{p}(\boldsymbol{\theta}_1^* | \mathbf{Y}) = \frac{1}{T} \sum_{t=1}^T p(\boldsymbol{\theta}_1^* | \boldsymbol{\theta}_2^{(t)}, \mathbf{Y})$$

where $\{\boldsymbol{\theta}_2^{(t)}\}$ is a sample from the posterior distribution of $\boldsymbol{\theta}_2$. The method can be extended to a k -block Gibbs sampler, but requires the introduction of a “reduced run” of the Gibbs sampler (see Chib 1995, for details). The algorithm becomes more elaborate, with significant bookkeeping involved.

Chib & Jeliazkov: One Block Metropolis-Hastings Chib & Jeliazkov (2001) extend Chib’s method to situations where Metropolis-Hastings algorithms are used for sampling from the posterior. Suppose $\boldsymbol{\theta}$ is sampled in one block using a Metropolis-Hastings algorithm. Let $q(\boldsymbol{\theta}, \boldsymbol{\theta}' | \mathbf{Y})$ denote the proposal density for the transition from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$, and let

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}' | \mathbf{Y}) = \min \left\{ 1, \frac{f(\mathbf{Y} | \boldsymbol{\theta}') \pi(\boldsymbol{\theta}')}{f(\mathbf{Y} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta})} \frac{q(\boldsymbol{\theta}', \boldsymbol{\theta} | \mathbf{Y})}{q(\boldsymbol{\theta}, \boldsymbol{\theta}' | \mathbf{Y})} \right\}$$

denote the corresponding transition probability. From the detailed balance condition, for any pair $(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$,

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^* | \mathbf{Y}) q(\boldsymbol{\theta}, \boldsymbol{\theta}^* | \mathbf{Y}) p(\boldsymbol{\theta} | \mathbf{Y}) = p(\boldsymbol{\theta}^* | \mathbf{Y}) \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta} | \mathbf{Y}) q(\boldsymbol{\theta}^*, \boldsymbol{\theta} | \mathbf{Y}) .$$

which after manipulation and integration over $\boldsymbol{\theta}$ leads to

$$p(\boldsymbol{\theta}^* | \mathbf{Y}) = \frac{\mathbb{E}_p \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^* | \mathbf{Y}) q(\boldsymbol{\theta}, \boldsymbol{\theta}^* | \mathbf{Y})}{\mathbb{E}_q \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta} | \mathbf{Y})} , \quad (16)$$

where \mathbb{E}_p stands for expectation with respect to the posterior distribution $p(\boldsymbol{\theta} | \mathbf{Y})$, whereas \mathbb{E}_q denotes expectation with respect to $q(\boldsymbol{\theta}^*, \boldsymbol{\theta} | \mathbf{Y})$ (interpreted as a density on $\boldsymbol{\theta}$ for a fixed $\boldsymbol{\theta}^*$). The resulting estimate is then

$$\hat{p}(\boldsymbol{\theta}^* | \mathbf{Y}) = \frac{\frac{1}{T} \sum_{t=1}^T \alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^* | \mathbf{Y}) q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^* | \mathbf{Y})}{\frac{1}{J} \sum_{j=1}^J \alpha(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\theta}}^{(j)} | \mathbf{Y})} ,$$

where $\{\boldsymbol{\theta}^{(t)}, t = 1, \dots, T\}$ is a sample from the posterior and $\{\tilde{\boldsymbol{\theta}}^{(j)}, j = 1, \dots, J\}$ is a supplemental iid sample from $q(\boldsymbol{\theta}^*, \cdot | \mathbf{Y})$.

While the method of Chib & Jeliazkov may be extended to multiple parameter-blocks, as with Chib’s method the bookkeeping in the algorithm becomes much more involved. Neal (1999) has noted that in multi-modal problems, such as mixture models, Chib’s method may give inaccurate estimates, if the Gibbs sampler has not adequately visited all modes.

2.4. Importance Sampling

As a variation on Monte Carlo methods, importance sampling (IS) replaces drawing from the prior distribution by drawing instead from an importance function $q_m(\boldsymbol{\theta})$ whose support contains that of the posterior and places more mass in the “important” regions of the parameter space Θ_m . The integral in (4) may be rewritten as

$$\begin{aligned} m(\mathbf{Y}) &= \int \frac{f(\mathbf{Y} | \boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ \hat{m}_{IS}(\mathbf{Y}) &= \frac{1}{T} \sum_{t=1}^T \frac{p(\boldsymbol{\theta}^{(t)})f(\mathbf{Y} | \boldsymbol{\theta}^{(t)})}{q(\boldsymbol{\theta}^{(t)})} = \frac{1}{T} \sum_{t=1}^T w_t \end{aligned} \quad (17)$$

where $\{\boldsymbol{\theta}^{(t)}, t = 1, \dots, T\}$ is a sample (often iid, but not necessarily) from $q(\cdot)$ and $w_1 \dots, w_T$ are the importance weights $w_t \equiv f(\mathbf{Y} | \boldsymbol{\theta}^{(t)})p(\boldsymbol{\theta}^{(t)})/q(\boldsymbol{\theta}^{(t)})$.

Importance sampling has a long and rich history of use in estimating normalizing constants or ratios of normalizing constants, however, the efficiency depends critically on the choice of proposal distributions. Simple importance sampling with multivariate t densities (four degrees of freedom is one default choice) with location equal to MLE estimates and scale matrix based on the inverse Fisher information are easy to construct. Alternatively, location and scale parameters may be estimated from MCMC output, if it is available. In practice, these choices tend to perform reasonably whenever the posterior is unimodal and reasonably well approximated by a normal density (perhaps after suitable transformations of parameters).

It is well-known that the IS approach will be as good as the importance function chosen; ideally, $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta} | \mathbf{Y}) \propto f(\mathbf{Y} | \boldsymbol{\theta})p(\boldsymbol{\theta})$, so that the weights are constant. Since the target $p(\boldsymbol{\theta} | \mathbf{Y})$ is known only up to the normalizing constant $m(\mathbf{Y})$, the effective sample size (ESS) based on the normalized weights $\tilde{w}_t = w_t/\bar{w}$,

$$\text{ESS} \equiv \frac{T}{1 + \text{var}_q(\tilde{w}_t)} \quad (18)$$

provides a rule of thumb for judging how different the proposal distribution q is from the target $p(\boldsymbol{\theta} | \mathbf{Y})$ (Liu 2001, page 34). Other diagnostics such as histograms and boxplots, provide effective tools to judge the importance proposal q .

For multi-modal or posteriors exhibiting strong nonlinear structure, a single multivariate t distribution is not adequate. It becomes more difficult to find good importance functions as the dimensionality grows, and a common problem is that a few extreme weights (due to large values of likelihood and prior relative to the importance function) will dominate the estimate. Instead better approximations to $p(\boldsymbol{\theta} | \mathbf{Y})$ may be constructed using mixtures of parametric densities (such as a t with 4 degrees of freedom)

$$q(\cdot) = \sum_k \omega_j t_4(\cdot; \mu_j, \alpha \Sigma_j), \quad (19)$$

where ω_j are mixture weights, μ_j is a location parameter and $\alpha \Sigma_j$ is the scale matrix for the j th mixture component. A “defensive mixture” may be obtained

by including the prior density as one of the mixture components (Hesterberg 1995). Owen & Zhou (2000) suggest the use of mixtures and control variates to reduce the infinite variance that may arise when q is a close match to the target, but decreases to zero faster than the target. Another issue with IS is that there may be an overabundance of very small weights (q may be too over-dispersed relative to the posterior or just a poor match in some regions). Liu (2001) suggests rejection control as a way of avoiding small weights. As samples from q are obtained, additional adaptive refinement of the IS proposal q to match the target $p(\boldsymbol{\theta} \mid \mathbf{Y})$ is possible, leading to adaptive importance sampling (AIS); such adaptive importance sampling schemes using mixtures have been proposed by West (1993); Oh & Berger (1992); Givens & Raftery (1996); Raghavan & Cox (1998) and more recently by Douc et al. (2006), among others.

As MCMC algorithms may be preferable for exploring the posterior distribution, it is possible to utilize information from the MCMC output in constructing an importance sampler with the purpose of calculating the marginal likelihood. Using a subset of the MCMC draws M_T , we take $\mu_j = \boldsymbol{\theta}^{(j)}$, $j \in M_T$ and estimate a local covariance Σ_j using a proportion of the neighboring points of $\boldsymbol{\theta}^{(j)}$ (distance may be measured using a scaled Mahalanobis distance, for example). We have explored using weights $\omega_t \propto [f(\boldsymbol{\theta})p(\boldsymbol{\theta})/q(\boldsymbol{\theta}; \alpha)]^\rho$, where q is based on the mixture model in (19). The values of both α and ρ may be chosen to minimize the Kullback-Leibler divergence between $p(\boldsymbol{\theta} \mid \mathbf{Y})$ and $q(\boldsymbol{\theta}; \alpha, \rho)$ evaluated at the remaining MCMC draws. Using on the order of 400 components, we have been able to accurately estimate marginal likelihoods in complicated mixtures in up to 10 dimensions, before seeing degradation in the IS weights. Using more components combined with adaptive refinement of the weights and IS proposal distribution, should lead to a more robust procedure.

2.5. Thermodynamic Integration

Thermodynamic integration (TI), used in physics and chemistry for computing the free energy difference (see the review article Frenkel (1986)), appears to have been independently discovered by several communities. In the numerical analysis literature, it is better known as Ogata’s method Ogata (1989). Another version in the astronomy literature appears in Gregory (2005), who uses TI in conjunction with parallel tempering to obtain normalizing constants. In the statistics community, Gelman & Meng (1998) show that thermodynamic integration is a special case of bridge sampling, and generalize the method using the more flexible and efficient path sampling.

Given two un-normalized densities $g_0(\boldsymbol{\theta})$ and $g_1(\boldsymbol{\theta})$, both indexed by a common parameter $\boldsymbol{\theta}$, Gelman & Meng (1998) use a scalar parameter $\phi \in [0, 1]$ to construct a continuous “path” that links the two densities; for example a geometric path,

$$g(\boldsymbol{\theta} \mid \phi) = g_0(\boldsymbol{\theta})^{1-\phi} g_1(\boldsymbol{\theta})^\phi, \quad (20)$$

where

$$m(\phi) \equiv \int_{\Theta} g(\boldsymbol{\theta} \mid \phi) d\boldsymbol{\theta}. \quad (21)$$

Our interest, of course, is in the ratio of the normalizing constants $m(0)$ and $m(1)$. When g_0 and g_1 correspond to two posterior distributions for $\boldsymbol{\theta}$ under

different models with the same support, this will be the Bayes factor. In the case of models with different supports, we may take $g_0(\boldsymbol{\theta})$ to be the prior of $\boldsymbol{\theta}$ and $g_1(\boldsymbol{\theta}) = f(\boldsymbol{\theta} | \mathbf{Y})p(\boldsymbol{\theta})$, so that $m(1)$ is the desired marginal likelihood $m(\mathbf{Y})$. The key identity to TI and path sampling is

$$\frac{d}{d\phi} \log(m(\phi)) = \int \frac{1}{m(\phi)} \frac{d}{d\phi} g(\boldsymbol{\theta} | \phi) d\boldsymbol{\theta} = \mathbb{E}_\phi \left[\frac{d}{d\phi} \log[g(\boldsymbol{\theta} | \phi)] \right] = \mathbb{E}_\phi [U(\boldsymbol{\theta}, \phi)] \quad (22)$$

where the expectation is taken with respect to the distribution $p(\boldsymbol{\theta} | \phi)$. Integrating (22) from 0 to 1, provides an estimate of the log ratio of the normalizing constants

$$\log[m(1)] - \log[m(0)] = \int_0^1 \mathbb{E}_\phi [U(\boldsymbol{\theta}, \phi)] d\phi \quad (23)$$

If one introduces a prior density $p(\phi)$ on $[0, 1]$, then (23) may be re-expressed as

$$\log[m(1)] - \log[m(0)] = \mathbb{E} \left[\frac{U(\boldsymbol{\theta}, \phi)}{p(\phi)} \right] \quad (24)$$

where the expectation is taken to the joint distribution of $p(\boldsymbol{\theta} | \phi)p(\phi)$. This leads to the estimator

$$\log[\hat{m}(1)] - \log[\hat{m}(0)] = \frac{1}{T} \sum_{t=1}^T \frac{U(\boldsymbol{\theta}^{(t)}, \phi^{(t)})}{p(\phi^{(t)})} \quad (25)$$

where $(\boldsymbol{\theta}^{(t)}, \phi^{(t)})$ are draws from $p(\boldsymbol{\theta} | \phi)p(\phi)$. Numerical integration may be used in place of Monte Carlo integration of ϕ . Gelman & Meng (1998) discuss the optimal choice of priors for ϕ and optimal path functions; in particular, they suggest that the geometric path in (20) is suboptimal and that a reduction in variance may be obtained thru alternative linking paths, although construction of optimal paths is an area that needs more research. See Gelman & Meng (1998) for more details and discussion of the connections between bridge, path and importance sampling and thermodynamic integration.

2.6. Nested Sampling

Skilling (2006) develops a novel Monte Carlo method called nested sampling that provides foremost an estimate of the marginal likelihood (evidence), but also generates samples of $\boldsymbol{\theta}$, if so desired. Nested sampling employs a change of variables from $\boldsymbol{\theta}$ to the scalar $\mathcal{L}(\boldsymbol{\theta})$ (the likelihood);

$$m(\mathbf{Y}) = \int_{\Theta} \mathcal{L}(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_0^\infty \mathcal{L} F(d\mathcal{L}) \quad (26)$$

reducing the problem to a one dimensional integral with respect to the distribution of \mathcal{L} . Using a second change of variables, $\chi = F(\mathcal{L})$, the original multidimensional integral is now represented as a one-dimensional integral

$$m(\mathbf{Y}) = \int_0^1 \mathcal{L}(\chi) F(d\chi) \quad (27)$$

on $[0, 1]$, which in principle, may be evaluated by standard quadrature rules. The integration to obtain the distribution of χ after the change of variables is of course intractable, but is handled via Monte Carlo sampling. Nested sampling starts by drawing N independent values of $\boldsymbol{\theta}$ from the prior. At each iteration, the minimum likelihood value, $\mathcal{L}_{(1)}(\boldsymbol{\theta}^{(t)})$, is recorded, and a subsequent N values of $\boldsymbol{\theta}^{(t+1)}$ are drawn from the prior, but now constrained to be inside the likelihood contour of $\mathcal{L}_{(1)}(\boldsymbol{\theta}^{(t)})$, hence the name “nested” sampling. At each iteration of the algorithm, the nested sampling scheme forces the likelihood uphill. While $\mathcal{L}_{(1)}(\boldsymbol{\theta}^{(t)}) \equiv \mathcal{L}(\chi_{(1)}^{(t)})$ is the minimum of N order statistics, we do not know the corresponding value of $\chi_{(1)}^{(t)}$ (except in special cases). However, we can sample $\chi_{(1)}^{(t)}$ as the minimum of N uniform random variables. Given an estimate of the density, an estimate of the marginal likelihood using numerical integration is

$$m_N(\mathbf{Y}) = \sum_t \mathcal{L}_{(1)}(\boldsymbol{\theta}^{(t)}) w_t \quad (28)$$

where w_t are the random weights, $w_t = .5(\chi_{(1)}^{(t-1)} - \chi_{(1)}^{(t)})$ associated with the numerical integration rule, (the trapezoid rule in this case). For more details see Skilling (2006).

The algorithm has two tuning parameters; T the number of iterations, and N the number of samples from the prior. The choice of N determines how quickly the likelihood contours are traversed. The number of iterations T is also critical. For $T = 1$ nested sampling reduces to naive Monte Carlo sampling from the prior. In order for the initial samples to be replaced with higher likelihood values, T must be much larger than N , but then like annealing, higher likelihood values are automatically obtained. The difficulty with implementing nested sampling, particularly in multi-modal problems, is the replacement step. At each iteration, we need N draws from the prior constrained to the current likelihood contour. As the previous $N - 1$ (after removing the minimum) are iid draws from the prior that satisfy the constraint, one needs to generate one additional point in the nested region. Rejection sampling from the prior is too inefficient to use in practice. Because MCMC transitions maintain the invariant distribution, it is possible to create a new point by “splitting” an existing point (possibly the discarded point), and allowing the new point (or entire population of N points) to evolve according to the MCMC transition kernel. In practice, this requires carefully tuning of the Markov chain and a large enough N to cover the possible modes. Based on our experiences with model selection in exoplanet models, MCMC followed by importance sampling is perhaps more efficient and easier to implement than nested sampling because of the multi-modal nature of the posterior distribution. Nested sampling has been used successfully by Mukherjee et al. (2006), who apply the method to a cosmological model selection problem.

2.7. Variational Methods

Variational methods provide a simple and efficient way of providing approximations and bounds on integrals that arise in calculating normalizing constants, and are popular in the machine/statistical learning community (see

<http://www.variational-bayes.org> for papers and software). Graphical models and mixture models, with massive data, are particular model selection problems where variational methods have been successful.

3. Estimating Bayes Factors

Bridge sampling (Meng & Wong 1996), path sampling (Gelman & Meng 1998), ratio importance sampling (RIS) (Chen & Shao 1997b) build on standard importance sampling. While RIS, with the optimal choice of proposal distribution is theoretically more efficient than bridge or path sampling, the optimal proposal distribution depends on the unknown Bayes factor. The book by Chen et al. (2000) discuss relationships among these methods, and extensions to models with differing dimensions.

3.1. Estimating Ratios via Importance Sampling

The Ratio Importance Sampling identity of Chen & Shao (1997b) provides a simple method of evaluating the Bayes factor

$$B[\mathcal{M}_0 : \mathcal{M}_1] = \frac{\mathbb{E}_q f(\mathbf{Y} | \boldsymbol{\theta}_0, \mathcal{M}_0) p(\boldsymbol{\theta}_0 | \mathcal{M}_0) / q(\boldsymbol{\theta}_0)}{\mathbb{E}_q f(\mathbf{Y} | \boldsymbol{\theta}_1, \mathcal{M}_1) p(\boldsymbol{\theta}_1 | \mathcal{M}_1) / q(\boldsymbol{\theta}_1)}, \quad (29)$$

where q is an IS density (possibly un-normalized) defined over $\Theta_0 \cup \Theta_1$, the union of the supports of each of the densities. The paper provides several practical and theoretical reasons why using the same importance function q for both denominator and numerator is actually a very good idea.

However, as noted in Chen & Shao (1997a), if the parameters under the two models have different dimensions (say $\boldsymbol{\theta}_0$ under \mathcal{M}_0 and $(\boldsymbol{\theta}_1, \boldsymbol{\psi})$ under \mathcal{M}_1 with $\dim(\boldsymbol{\theta}_0) = \dim(\boldsymbol{\theta}_1)$), then RIS is not directly applicable. Instead, they suggest introducing an auxiliary distribution $\omega(\boldsymbol{\psi} | \boldsymbol{\theta})$, a completely known conditional density. Then,

$$\mathcal{M}_0(\mathbf{Y}) = \iint f(\mathbf{Y} | \boldsymbol{\theta}_0, \mathcal{M}_0) p(\boldsymbol{\theta}_0 | \mathcal{M}_0) \omega(\boldsymbol{\psi} | \boldsymbol{\theta}) d\boldsymbol{\theta} d\boldsymbol{\psi}$$

and as a consequence of using RIS,

$$B[\mathcal{M}_0 : \mathcal{M}_1] = \frac{\mathbb{E}_q f(\mathbf{Y} | \boldsymbol{\theta}_0, \mathcal{M}_0) p(\boldsymbol{\theta}_0 | \mathcal{M}_0) \omega(\boldsymbol{\psi} | \boldsymbol{\theta}_0) / q(\boldsymbol{\theta}_0, \boldsymbol{\psi})}{\mathbb{E}_q f(\mathbf{Y} | \boldsymbol{\theta}_1, \boldsymbol{\psi}) p(\boldsymbol{\theta}_1, \boldsymbol{\psi} | \mathcal{M}_1) / q(\boldsymbol{\theta}_1, \boldsymbol{\psi})}.$$

The choice $q(\boldsymbol{\theta}, \boldsymbol{\psi}) = p(\boldsymbol{\theta}, \boldsymbol{\psi} | \mathbf{Y}, \mathcal{M}_1) \equiv p_1$ is particularly interesting in that the Bayes factor simplifies to

$$B[\mathcal{M}_0 : \mathcal{M}_1] = \mathbb{E}_{p_1} \frac{f(\mathbf{Y} | \boldsymbol{\theta}, \mathcal{M}_0) p(\boldsymbol{\theta} | \mathcal{M}_0) \omega(\boldsymbol{\psi} | \boldsymbol{\theta})}{f(\mathbf{Y} | \boldsymbol{\theta}, \boldsymbol{\psi}) p(\boldsymbol{\theta}, \boldsymbol{\psi} | \mathcal{M}_1)}. \quad (30)$$

Note that, according to (30), we only need a sample from the posterior under \mathcal{M}_1 in order to estimate the ratio of the two normalizing constants. Of course, $\Theta_0 \subset \Theta_1$ for the method to make any sense, but in principle \mathcal{M}_0 does not need to be nested inside \mathcal{M}_1 . However, in the context of model selection in nested models, RIS may be simplified even more.

3.2. RIS and Nested models

Consider a series of models, where there is a single model, $\mathcal{M}_{\mathbf{E}}$ with parameters $\boldsymbol{\theta}$, that encompasses all other models (the other models are nested in $\mathcal{M}_{\mathbf{E}}$). For any model $\mathcal{M}_{\mathbf{S}}$ with parameter $\boldsymbol{\theta}_{\mathbf{S}}$, we may partition the parameter vector in the encompassing model as $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathbf{S}}, \boldsymbol{\theta}_{(-\mathbf{S})})$, where $\boldsymbol{\theta}_{(-\mathbf{S})}$ denotes the parameters that are not included in $\mathcal{M}_{\mathbf{S}}$, i.e. $\boldsymbol{\theta}_{(-\mathbf{S})} \equiv \mathbf{0}$ under $\mathcal{M}_{\mathbf{S}}$. In principle, using only a sample from $\mathcal{M}_{\mathbf{E}}$, one can estimate the Bayes factor for every model to $\mathcal{M}_{\mathbf{E}}$ using (30), and as a consequence one can compute every posterior model probability (Ibrahim, Chen, & MacEachern 1999; Chen, Ibrahim, & Yiannoutsos 1999).

If the priors on the model-specific parameters are compatible by conditioning, i.e., if

$$p(\boldsymbol{\theta}_{\mathbf{S}} | \mathcal{M}_{\mathbf{S}}) = p(\boldsymbol{\theta}_{\mathbf{S}} | \boldsymbol{\theta}_{(-\mathbf{S})} = \mathbf{0}, \mathcal{M}_{\mathbf{E}})$$

then Bayes factor simplifies to the Savage-Dickey density ratio,

$$B[\mathcal{M}_{\mathbf{S}} : \mathcal{M}_{\mathbf{E}}] = \frac{p(\boldsymbol{\theta}_{(-\mathbf{S})} = \mathbf{0} | \mathbf{Y}, \mathcal{M}_{\mathbf{E}})}{p(\boldsymbol{\theta}_{(-\mathbf{S})} = \mathbf{0} | \mathcal{M}_{\mathbf{E}})}, \quad (31)$$

where $p(\boldsymbol{\theta}_{(-\mathbf{S})} = \mathbf{0} | \mathbf{Y}, \mathcal{M}_{\mathbf{E}})$ and $p(\boldsymbol{\theta}_{(-\mathbf{S})} = \mathbf{0} | \mathcal{M}_{\mathbf{E}})$ are the marginal posterior and prior densities, respectively, of $\boldsymbol{\theta}_{(-\mathbf{S})}$ under $\mathcal{M}_{\mathbf{E}}$ evaluated at zero, and as such, the RIS estimator in (30) may be viewed as an IS estimate of the generalized Savage-Dickey density ratio (Verdinelli & Wasserman 1995) for Bayes factors.

Alternatively, the marginal distribution evaluated at $p(\boldsymbol{\theta}_{(-\mathbf{S})} = \mathbf{0} | \mathbf{Y}, \mathcal{M}_{\mathbf{E}})$ in the numerator of (31) may be estimated using the importance weighted marginal density estimate of Chen (1994),

$$\hat{p}(\boldsymbol{\theta}_{(-\mathbf{S})} = \mathbf{0} | \mathbf{Y}, \mathcal{M}_{\mathbf{E}}) = \frac{1}{T} \sum_{t=1}^T \frac{p(\boldsymbol{\theta}_{(-\mathbf{S})} = \mathbf{0}, \boldsymbol{\theta}_{\mathbf{S}}^{(t)} | \mathbf{Y}, \mathcal{M}_{\mathbf{E}})}{p(\boldsymbol{\theta}_{(-\mathbf{S})}^{(t)}, \boldsymbol{\theta}_{\mathbf{S}}^{(t)} | \mathbf{Y})} \omega(\boldsymbol{\theta}_{(-\mathbf{S})}^{(t)} | \boldsymbol{\theta}_{\mathbf{S}}^{(t)}),$$

where $\{\boldsymbol{\theta}^{(t)} \ t = 1, \dots, T\}$ is a (possibly dependent) sample from $p(\boldsymbol{\theta} | \mathbf{Y}, \mathcal{M}_{\mathbf{E}})$.

The choice $p(\boldsymbol{\theta}_{(-\mathbf{S})} | \boldsymbol{\theta}_{\mathbf{S}}, \mathbf{Y}, \mathcal{M}_{\mathbf{E}})$ for ω which results in the conditional marginal density estimator (CMDE) of Gelfand, Smith, & Lee (1992), which is optimal among all IWMDE (Chen 1994). As this is typically not available, the empirical method of Chen (1994) approximates the posterior for $\boldsymbol{\theta}$ by a multivariate Gaussian with mean and variance equal to the estimated posterior mean and variance; the density ω is then given by the conditional Gaussian for $\boldsymbol{\theta}_{\mathbf{S}}$ obtained by conditioning on $\boldsymbol{\theta}_{(-\mathbf{S})} = \mathbf{0}$ (Chen 1994).

The key feature to using RIS in nested models is that the method only requires MCMC output from the posterior distribution for the encompassing model to estimates all Bayes Factors, but does not require knowledge of the MCMC sampler as with Chib (1995) or Chib & Jeliazkov (2001) or generation of additional samples as in other IS approaches. In practice, RIS using the posterior of the larger model may be unstable, however, if the values of parameters in the simpler model have very little support under the posterior of the more complex model, *i.e.* the point $\boldsymbol{\theta}_{(-\mathbf{S})} = \mathbf{0}$ is in the tails of the posterior distribution of the encompassing model.

4. Trans-Dimensional Methods

Reversible jump MCMC (RJ-MCMC) (Green 1995) and variations that sample the model space and parameter space jointly do not require exhaustive enumeration of the model space, and can be used in moderate to infinite dimensional problems, such as sparse nonparametric regression models for mass spectroscopy (Clyde & Wolpert 2006). Steps in the RJ-MCMC algorithm may be divided into within model steps (the usual MCMC transitions) or across model steps, in which case one proposes a jump to a new model and new parameter values for the model. With the model viewed as just another parameter, the basic model jumping step in a RJ-MCMC algorithm is no different than the usual Metropolis-Hastings transitions. The model jumping step can be described as follows and applies to extremely general model selection problems.

Given the current state $(\boldsymbol{\theta}_k, \mathcal{M}_k)$, Propose a jump to a new model \mathcal{M}_j $q(\mathcal{M}_j | \mathcal{M}_k, \boldsymbol{\theta}_k, \mathbf{Y})$ given the current model \mathcal{M}_k . Given the new model, generate $\boldsymbol{\theta}_j$ from distribution $q(\boldsymbol{\theta}_j | \boldsymbol{\theta}_k, \mathcal{M}_k, \mathcal{M}_j, \mathbf{Y})$. Accept the proposed move to $(\boldsymbol{\theta}_j, \mathcal{M}_j)$ with probability

$$\alpha((\boldsymbol{\theta}_k, \mathcal{M}_k), (\boldsymbol{\theta}_j, \mathcal{M}_j)) = \min \{1, H((\boldsymbol{\theta}_k, \mathcal{M}_k), (\boldsymbol{\theta}_j, \mathcal{M}_j))\}$$

where the Hastings ratio $H((\boldsymbol{\theta}_k, \mathcal{M}_k), (\boldsymbol{\theta}_j, \mathcal{M}_j))$ is

$$\frac{p(\mathbf{Y} | \boldsymbol{\theta}_j, \mathcal{M}_j)p(\boldsymbol{\theta}_j | \mathcal{M}_j)p(\mathcal{M}_j) q(\boldsymbol{\theta}_k, \mathcal{M}_k | \boldsymbol{\theta}_j, \mathcal{M}_j, \mathbf{Y})}{p(\mathbf{Y} | \boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\theta}_k | \mathcal{M}_k)p(\mathcal{M}_k) q(\boldsymbol{\theta}_j, \mathcal{M}_j | \boldsymbol{\theta}_k, \mathcal{M}_k, \mathbf{Y})}.$$

The key to implementing efficient RJ-MCMC algorithms involves constructing model jumping proposals and efficient proposals for $\boldsymbol{\theta}$ under the new model. Ideally the proposal for $\boldsymbol{\theta}_k | \mathcal{M}_k$ is the posterior distribution for $\boldsymbol{\theta}_k$ under model \mathcal{M}_k . Often, the proposals have to be tailored to each specific class of problems and may require significant tuning. Relationships of RJ-MCMC and MH/Gibbs sampling in the linear model setting are discussed in Clyde (1999) and Godsill (2001). Recent papers by Dellaportas et al. (2002), Brooks et al. (2003), Godsill (2001) and Green (2003) discuss variations of RJ-MCMC algorithms and construction of efficient/automatic proposal distributions.

The proposal for a new $\boldsymbol{\theta}_j^*$ is often constructed by “reusing” part of the parameter vector of the current model \mathcal{M}_k . When moving up in dimension, one may generate a vector u , and then construct a one-to-one mapping $g : (\boldsymbol{\theta}_k, u) \rightarrow (u^*, \boldsymbol{\theta}_j^*)$. Because u is generated, rather than $\boldsymbol{\theta}_j^*$, this accounts for the Jacobian term in many descriptions of RJ-MCMC. Such a mapping may be difficult to construct, particularly for non-nested models, although Green (2003) describes an automatic procedure. Ideally, the proposal distribution for $\boldsymbol{\theta}^*$ should be as close to the posterior distribution $p(\boldsymbol{\theta}_j | \mathcal{M}_j, \mathbf{Y})$ as possible. If MLEs are available for each model, we may construct an independent proposal distribution based on a t density centered at the MLE and with a scale matrix proportional to the inverse of the Fisher information at the MLE. For nested models, another possible strategy is to fit the full model by maximum likelihood and use that information to specify the parameters of all conditional distributions proposals, so that one does not have to fit a model every time a new model is proposed.

The relative frequencies of each model provide a Monte Carlo estimate of the posterior model probabilities,

$$\hat{p}_{RJ}(\mathcal{M}_m | \mathbf{Y}) = \frac{f_m}{T}$$

where f_m is the number of times the Markov chain visited model \mathcal{M}_m out of the T iterations. In the case of the models with very small posterior probabilities, one will need to run the Markov chain for a very large number of iterations in order to be able to accurately estimate the posterior probabilities using the Monte Carlo frequencies. This does not appear to be a problem if one is interested in model averaging rather than model selection. An alternative, is to “bias” the prior model probabilities so that low probability models are more likely to be visited, and then correct the results by an importance re-weighting. (This is practical only in low dimensional model spaces).

Recently, Bartolucci, Scaccia, & Mira (2006) proposed a class of estimators of the Bayes factor using the output of the reversible jump algorithm, based on an extension of the bridge sampling identity of Meng & Wong (1996). The simplest estimator is of the form

$$\hat{B}_{BSM}[Mk : Mj] = \frac{f_k \sum_{i=1}^{f_j} \alpha((\boldsymbol{\theta}_j^{(i)}, \mathcal{M}_j^{(i)}), (\boldsymbol{\theta}_k^{(i)}, \mathcal{M}_k^{(i)}))}{f_j \sum_{i=1}^{f_k} \alpha((\boldsymbol{\theta}_k^{(i)}, \mathcal{M}_k^{(i)}), (\boldsymbol{\theta}_j^{(i)}, \mathcal{M}_j^{(i)}))} \quad (32)$$

and may be easily computed using the RJ-MCMC output. Note that the estimator is restricted to RJ-MCMC algorithms where the jumps from \mathcal{M}_k are limited to “neighboring” models \mathcal{M}_{k-1} and \mathcal{M}_{k+1} (with wrapping or reflection at $k = 1$ or $k = M$), however, Bayes factors $B[\mathcal{M}_k : \mathcal{M}_l]$ for non-sequential models \mathcal{M}_l may be obtained by the recursive property of Bayes factors, $B[Mk : Ml] = B[\mathcal{M}_k : \mathcal{M}_{k+1}]B[\mathcal{M}_{k+1} : \mathcal{M}_{k+2}] \dots B[\mathcal{M}_{l-1} : \mathcal{M}_l]$. Bartolucci et al. (2006) discuss other estimators and show that these may lead to a substantial gain of efficiency in estimating the Bayes factor over the naive Monte Carlo frequencies.

5. Conclusions/Recommendations

The recent review paper by Han & Carlin (2001) uses several examples to compare MCMC approaches for computing Bayes Factors, such as Chib’s marginal likelihood approach, the product space search of Carlin & Chib (1995), the Metropolized product space method from Dellaportas et al. (2002) (a RJ variation of Carlin and Chib), and the Composite Model search of Godsill (2001) (a RJ algorithm that takes advantage of common parameters in the context of variables selection). Han and Carlin found that joint model/parameter space methods worked adequately, but could be difficult to tune, particularly the RJ formulations. The marginal likelihood methods were easiest to program and tune, although they note the blocking structure required may limit applications.

Chib’s method has performed reasonably well in the cases where we have implemented it, but as we mentioned before if the sampling mechanisms become more involved (multiple parameter blocks), the associated algorithms become

harder to implement and possibly less stable. We have found the method of Chib & Jeliazkov (2001) to be much more unstable. Contrary to Han & Carlin (2001), we have been able to implement RJ-MCMC in a several applications (in particular infinite dimensional problems), but do agree that Monte Carlo frequencies can result in poor estimates of the posterior model probabilities. It will be interesting to see if the Rao-Blackwellized estimates of Bartolucci *et al.* (2006) provide significant improvements for estimating Bayes Factors.

In a variety of low dimensional problems (up to 15 dimensions), running MCMC within a model followed by importance sampling using the MCMC output to construct a mixture proposal distribution has turned out to be the easiest to implement and has performed better than other methods in comparisons. While we have found cases where the integrated harmonic mean is reliable, IS is still more accurate.

From our experience in various SAMSI programs, it is important to try to implement more than one method and test code on examples with known marginals, if nothing else because it is very easy to make mistakes in coding! To this regard, we would also like to mention that, in the case of methods based on the MCMC output, it is often necessary to run the Markov chains for a much larger number of iterations than what would be needed for parameter estimation only. Failure to do so, especially paired with the use of only one computing method, may result in severe errors in the final answer. And finally, if the Bayes factors or model probabilities seem to be at odds with the data, do not forget the prior distributions on model specific parameters may have a huge influence on the resulting inferences about models.

This material is based upon work supported by the National Science Foundation under Grants 0112069, 0422400, and 0507481. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

- Bartolucci, F., Scaccia, L., & Mira, A. 2006, *Biometrika*, 93, 41
- Berger, J. O. 1997, in *Statistical Challenges in Modern Astronomy II*, ed. G. H. Babu & E. D. Feigelson (Springer), 15–39
- Berger, J. O., & Pericchi, L. R. 1996a, in *Bayesian Statistics 5*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. Smith (Oxford, UK: Oxford Univ. Press), 25–44
- Berger, J. O., & Pericchi, L. R. 1996b, *J. Am. Stat. Assoc.*, 91, 109
- . 1998, *Sankhyā*, Ser. B, 60, 1
- . 2001, in *Lecture Notes in Statistics*, Vol. 38, *Model Selection*, ed. P. Lahiri (Hayward, CA: Inst. Math. Statist.), 135–193
- Berger, J. O., Pericchi, L. R., & Varshavsky, J. A. 1998, *Sankhyā*, Ser. A, 60, 307
- Bernardo, J. M. 1979, *J. Roy. Stat. Soc. B*, 41, 113
- Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F., & West, M., eds. 2007, *Bayesian Statistics 8* (Oxford, UK: Oxford Univ. Press)
- Brooks, S. P., Giudici, P., & Roberts, G. O. 2003, *J. Roy. Stat. Soc. B*, 65, 3
- Carlin, B. P., & Chib, S. 1995, *J. Roy. Stat. Soc. B*, 57, 473
- Chen, M.-H. 1994, *J. Am. Stat. Assoc.*, 89, 818
- Chen, M.-H., Ibrahim, J. G., & Yiannoutsos, C. 1999, *J. Roy. Stat. Soc. B*, 61, 223

- Chen, M.-H., & Shao, Q.-M. 1997a, *Stat. Sinica*, 7, 607
 —. 1997b, *Ann. Stat.*, 25, 1563
 Chen, M.-H., Shao, Q.-M., & Ibrahim, J. G. 2000, *Monte Carlo methods in Bayesian computation* (Springer-Verlag), 386
 Chib, S. 1995, *J. Am. Stat. Assoc.*, 90, 1313
 Chib, S., & Jeliazkov, I. 2001, *J. Am. Stat. Assoc.*, 96, 270
 Clyde, M. 1999, in *Bayesian Statistics 6*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. Smith (Oxford, UK: Oxford Univ. Press), 157–185
 Clyde, M., DeSimone, H., & Parmigiani, G. 1996, *J. Am. Stat. Assoc.*, 91, 1197
 Clyde, M., & George, E. I. 2004, *Stat. Sci.*, 19, 81
 Clyde, M. A., & Wolpert, R. L. 2006, in *Bayesian Statistics 8*, ed. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. Smith, & M. West (Oxford, UK: Oxford Univ. Press), to appear
 Dellaportas, P., Forster, J. J., & Ntzoufras, I. 2002, *Stat. Comput.*, 12, 27
 DiCiccio, T. J., Kass, R. E., Raftery, A. E., & Wasserman, L. 1997, *J. Am. Stat. Assoc.*, 92, 903
 Douc, R., Guillin, A., Marin, J.-M., & Robert, C. 2006, *Convergence of adaptive mixtures of importance sampling schemes*, Tech. rep., Ceremade - Universit Paris-Dauphine
 Evans, M., & Swartz, T. 1995, *Stat. Sci.*, 10, 254
 Frenkel, D. 1986, in *Molecular Dynamics Simulation of Statistical Mechanical Systems*, ed. G. Ciccotti & W. Hoover (North-Holland, Amsterdam), 151–188
 Gelfand, A. E., & Dey, D. K. 1994, *J. Roy. Stat. Soc. B*, 56, 501
 Gelfand, A. E., Smith, A. F. M., & Lee, T.-M. 1992, *J. Am. Stat. Assoc.*, 87, 523
 Gelman, A., & Meng, X.-L. 1998, *Stat. Sci.*, 13, 163
 George, E. I., & McCulloch, R. E. 1993, *J. Am. Stat. Assoc.*, 88, 881
 —. 1997, *Stat. Sinica*, 7, 339
 Givens, G. H., & Raftery, A. E. 1996, *J. Am. Stat. Assoc.*, 91, 132
 Godsill, S. J. 2001, *J. Comput. Graph. Stat.*, 10, 230
 Green, P. J. 1995, *Biometrika*, 82, 711
 Green, P. J. 2003, in *Highly Structured Stochastic Systems*, 179–206
 Gregory, P. 2005, *Bayesian Logical Data Analysis for the Physical Sciences* (Cambridge)
 Han, C., & Carlin, B. P. 2001, *J. Am. Stat. Assoc.*, 96, 1122
 Hesterberg, T. 1995, *Technometrics*
 Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. 1999, *Stat. Sci.*, 14, 382, corrected version at <http://www.stat.washington.edu/www/research/online/hoeting1999.pdf>
 Ibrahim, J. G., Chen, M.-H., & MacEachern, S. N. 1999, *Can. J. Stat.*, 27, 701
 Kass, R. E., & Raftery, A. E. 1995, *J. Am. Stat. Assoc.*, 90, 773
 Liu, J. 2001, *Monte Carlo Strategies in Scientific Computing* (Springer-Verlag)
 Loredo, T. J., & Chernoff, D. F. 2003, *Bayesian adaptive exploration (Statistical Challenges in Astronomy)*, 57–70
 Madigan, D., & Jeremy C. York, N. 1995, *Int. Stat. Rev.*, 63, 215
 Meng, X.-L., & Wong, W. H. 1996, *Stat. Sinica*, 6, 831
 Mukherjee, P., Parkinson, D., & Liddle, A. R. 2006, *ApJ*, 638, L51
 Neal, R. M. 1999, *Erroneous Results in “Marginal Likelihood from the Gibbs Output”*, Tech. rep., University of Toronto
 Newton, M. A., & Raftery, A. E. 1994, *J. Roy. Stat. Soc. B*, 56, 3
 Ogata, Y. 1989, *Numer. Math.*, 55, 137
 Oh, M.-S., & Berger, J. O. 1992, *Journal of Statistical Computation and Simulation*, 41, 143
 O’Hagan, A. 1995, *J. Roy. Stat. Soc. B*, 57, 99
 Owen, A., & Zhou, Y. 2000, *J. Am. Stat. Assoc.*, 95, 135
 Pérez, J. M., & Berger, J. O. 2000, *Expected posterior prior distributions for model selection*, Tech. Rep. 00-08, Duke University ISDS, USA

- Raghavan, N., & Cox, D. D. 1998, *Journal of Statistical Computation and Simulation*, 60, 237
- Satagopan, J., Newton, M., & Raftery, A. 2000, Easy estimation of normalizing constant and Bayes factors from posterior simulation: Stabilizing the harmonic mean estimator, Technical report, University of Washington
- Sisson, S. A. 2005, *J. Am. Stat. Assoc.*, 100, 1077
- Skilling, J. 2006, in *Bayesian Statistics 8*, ed. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. Smith, & M. West (Oxford, UK: Oxford Univ. Press)
- Tierney, L., Kass, R. E., & Kadane, J. B. 1989, *J. Am. Stat. Assoc.*, 84, 710
- Verdinelli, I., & Wasserman, L. 1995, *J. Am. Stat. Assoc.*, 90, 614
- West, M. 1993, *J. Roy. Stat. Soc. B*, 55, 409
- Wolpert, R. 2002, Stable Limit Laws for Marginal Probabilities from MCMC Streams: Acceleration of Convergence, <http://www.stat.duke.edu/courses/Spring02/sta205/mcbf.pdf>