Some Thoughts on Future Directions in Bayesian Model Selection

Hani Doss
University of Florida

As advances in computational technology have made it possible to apply Bayesian methods to situations that are increasingly complex, using a wider variety of models that are increasingly more sophisticated, problems regarding model choice are now of central importance. As a result, the development of methods for doing model selection is now at the forefront of research in Bayesian statistics. This is reflected by many of the papers on Bayesian methods in this issue of *Statistica Sinica*, on topics ranging from variable selection in generalized linear models (Wang and George) to the need to choose between Bayesian nonparametric models and their parametric counterparts (Dunson; Bulla, Muliere, and Walker). This note describes some thoughts regarding future directions in Bayesian model selection, focusing on computational challenges. We briefly describe some unrelated approaches to Bayesian model selection that are currently used and argue that much can be gained by combining them. We proceed at a low technical level and make our points through a discussion of concrete examples; however application of the ideas is not limited to those examples.

Bayesian model selection is usually described as follows. We have data Y, and possible models $\mathcal{M}_1, \ldots, \mathcal{M}_k$, where for each j, $\mathcal{M}_j$ is defined by a family of distributions p_{θ_j}, $\theta_j \in \Theta_j$, together with a prior on Θ_j. The Θ_j's need not be of the same dimension. We may or may not have a prior distribution on the set of models. The objective is to select "the best model," or in the case where we have a prior on the set of models, the objective is to select the model with the highest posterior probability. When no single model is clearly the best, we report a set of plausible models or models with high posterior probability. It is helpful to take a slightly more general view, and not restrict the set of models to be finite. To make our points, we consider two examples of a rather different character.

Example 1 We start with the following simple three-level hierarchical model:

$$\text{conditional on } \psi_j, \qquad Y_j \overset{\text{indep}}{\sim} \mathcal{N}(\psi_j, \sigma_j^2), \quad j = 1, \ldots, m \tag{1a}$$

$$\text{conditional on } \mu, \tau, \qquad \psi_j \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \tau^2), \quad j = 1, \ldots, m \tag{1b}$$

$$(\mu, \tau) \sim \nu_c \tag{1c}$$

Here, we assume that the variances σ_j^2 are known, and ν_c is the normal/inverse gamma prior, indexed by the vector $c = (c_1, c_2, c_3, c_4)$, i.e. $1/\tau^2 \sim \text{gamma}(c_1, c_2)$, and given τ, $\mu \sim \mathcal{N}(c_3, \tau c_4)$. This model is typically used to model random effects situations: Y_j is a single summary statistic from experiment j, based on a sample of size n_j. (Usually the variance is unknown, but we assume that experiment j gives an estimate $\hat{\sigma}_j$ that is accurate enough so that assuming $\hat{\sigma}_j$ equals the true value does not cause any problem.)

Whereas the normality assumption in line (1a) is typically supported by some theoretical result, such as the asymptotic normality of maximum likelihood estimates, the normality assumption in line (1b) generally doesn't have any justification and is made solely for the sake of convenience. In certain situations, a good alternative to line (1b) is $\psi_j \overset{\text{iid}}{\sim} t_{d,\mu,\tau}$, where $t_{d,\mu,\tau}$ is the t distribution with d degrees of freedom, location μ, and scale τ. We will then want to select d, with the choice $d = \infty$ signifying the choice of the normal distribution.

In order to emphasize that d is a hyperparameter in the model we define $\theta = (\psi, \mu, \tau)$, where $\psi = (\psi_1, \dots, \psi_m)$, and recast (1) as follows:

$$\begin{aligned} \text{conditional on } \theta, \quad Y_j &\overset{\text{indep}}{\sim} \mathcal{N}(\psi_j, \sigma_j^2), \quad j = 1, \dots, m \\ \theta &\sim \nu_h, \end{aligned}$$

where the prior is now changed to

$$\nu_h(\theta) = \left(\textstyle\prod_{j=1}^m t_{d,\mu,\tau}(\psi_j)\right) \lambda_c(\mu, \tau), \tag{2}$$

where λ_c is the normal/inverse gamma prior indexed by c. Here, the set of models is the family $\{\nu_h,\, h \in \mathcal{H}\}$ and the hyperparameter is $h = (d, c)$. When looked at in this way, we see that choosing the hyperparameter of the prior ν_h involves a model selection step (choice of number of degrees of freedom d), in addition to selection of the prior on (μ, τ).

Example 2 In a standard formulation of the problem of Bayesian variable selection in linear regression, we have a response variable Y and a set of predictors $X_1, \dots, X_p$, each a vector of length m. For every subset γ of $\{1, \dots, p\}$ we have a potential model $\mathcal{M}_\gamma$ given by

$$Y = 1_m \beta_0 + X_\gamma \beta_\gamma + \epsilon, \tag{3}$$

where 1_m is the vector of m 1's, X_γ is the design matrix whose columns consist of the predictor vectors corresponding to the subset γ, β_γ is the vector of coefficients for that subset, and $\epsilon \sim \mathcal{N}_m(0, \sigma^2 I)$. The most commonly used prior on the unknown parameters β_γ and σ is Zellner's g-prior (Zellner 1986), indexed by a hyperparameter g. If we let p_γ denote the number of variables in the subset γ, this prior, which we will denote by π_γ, is described as follows:

$$(\sigma^2, \beta_0) \sim p(\sigma^2, \beta_0) \propto 1/\sigma^2, \qquad \text{and} \qquad \text{given } \sigma,\ \beta_\gamma \sim \mathcal{N}_{p_\gamma}\big(0, g\sigma^2 (X_\gamma' X_\gamma)^{-1}\big). \tag{4}$$

Although this prior is improper, the resulting posterior distribution is proper.

Examples 1 and 2 differ in an important aspect. In example 1, the priors ν_h are all mutually absolutely continuous, whereas in example 2 the priors π_γ are not (when a subset γ excludes a variable, in effect the regression coefficient for that variable is given a distribution that is degenerate at 0; therefore the vector β lives in a subspace of $\mathbb{R}^p$ of dimension less than p). Absolute continuity has important consequences regarding the calculation of Bayes factors.

The marginal distribution of Y when the prior is ν_h is $m_h(y) = \int \ell_y(\theta)\nu_h(\theta)\,d\theta$, where $\ell_y(\theta)$ is the likelihood function. For two hyperparameter values h_1 and h_2, the Bayes factor of the model indexed by h_2 relative to the model indexed by h_1, which we define as

$$B(h_2, h_1) = \frac{m_{h_2}(y)}{m_{h_1}(y)},$$

is often used to choose between ν_{h_1} and ν_{h_2}: when $B(h_2, h_1)$ is very small, the model indexed by h_2 is deemed less plausible.

In any problem where the priors $\nu_h,\ h \in \mathcal{H}$ are mutually absolutely continuous, in principle it is possible to conveniently estimate all possible Bayes factors. For $h \in \mathcal{H}$, let $\nu_{h,y}$ denote the posterior density of θ given $Y = y$, corresponding to the prior ν_h. If we fix an arbitrary hyperparameter value $h_1 \in \mathcal{H}$, estimation of all Bayes factors $B(h, h_1)$ can be done from a single sample (iid or Markov chain) $\theta_1, \ldots, \theta_n$ from the posterior $\nu_{h_1,y}$, and knowledge of the ratios of the *priors* ν_h/ν_{h_1} (not the posteriors). We have

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \frac{\nu_h(\theta_i)}{\nu_{h_1}(\theta_i)} &\rightarrow \int \frac{\nu_h(\theta)}{\nu_{h_1}(\theta)} \nu_{h_1,y}(\theta)\,d\theta \qquad (5) \\ &= \frac{m_h}{m_{h_1}} \int \frac{\ell_y(\theta)\nu_h(\theta)/m_h}{\ell_y(\theta)\nu_{h_1}(\theta)/m_{h_1}} \nu_{h_1,y}(\theta)\,d\theta \\ &= \frac{m_h}{m_{h_1}} \int \frac{\nu_{h,y}(\theta)}{\nu_{h_1,y}(\theta)} \nu_{h_1,y}(\theta)\,d\theta \ = \ \frac{m_h}{m_{h_1}}. \end{aligned}$$

Therefore, the estimate in the left side of (5) is a consistent estimate of the Bayes factor $B(h, h_1)$. (The equality $\int \nu_h(\theta)/\nu_{h_1}(\theta)\, \nu_{h_1,y}(\theta)\,d\theta = m_h/m_{h_1}$ appears in many different guises in the literature, including in incomplete data problems in frequentist inference, in which θ plays the role of missing data and h is the unknown parameter; see section 6.5 of Nicolae et al. (2007)).

We illustrate this on a toy example considered by Bayarri and Berger (2004). Data was generated as follows: Given $\psi_1, \ldots, \psi_5$, $Y_j \overset{\text{ind}}{\sim} \mathcal{N}(\psi_j, 1/2)$, $j = 1, \ldots, 5$; also, $\psi_j \overset{\text{iid}}{\sim} \mathcal{N}(1, 1)$, $j = 1, \ldots, 4$, and independently $\psi_5 \sim \mathcal{N}(5, 1)$. The resulting data vector turned out to be $Y = (1.560, 0.641, 1.982, 0.014, 6.964)$ with Y_5 being over six standard deviations away from the mean of the other four Y_j's. Bayarri and Berger (2004) were interested in testing the null hypothesis that the distribution of the random effects in model (1) is normal, and here we will replace line (1b) with $\psi_j \overset{\text{iid}}{\sim} t_{d,\mu,\tau}$, and we will be interested in selecting the degrees of freedom parameter. We ran a Markov chain that gave samples from the posterior distribution for the model where the t distribution in (2) has 3 degrees of freedom and (μ, τ) has the normal/inverse gamma distribution with parameter $c = (.1, .1, 0, 1000)$. Keeping the prior on (μ, τ) fixed and varying the degrees of freedom parameter, we estimated the Bayes factor using (5), producing Figure 1. The plot suggests that a model with a t distribution with 1 or 2 degrees of freedom is reasonable, but that the normal model (1) is not appropriate (the Bayes factor for the t_1 distribution relative to the normal distribution is 7.8).

There is a substantial literature devoted to devising estimates that improve on (5), most of which focuses on estimating a single Bayes factor. Important references are Meng and Wong

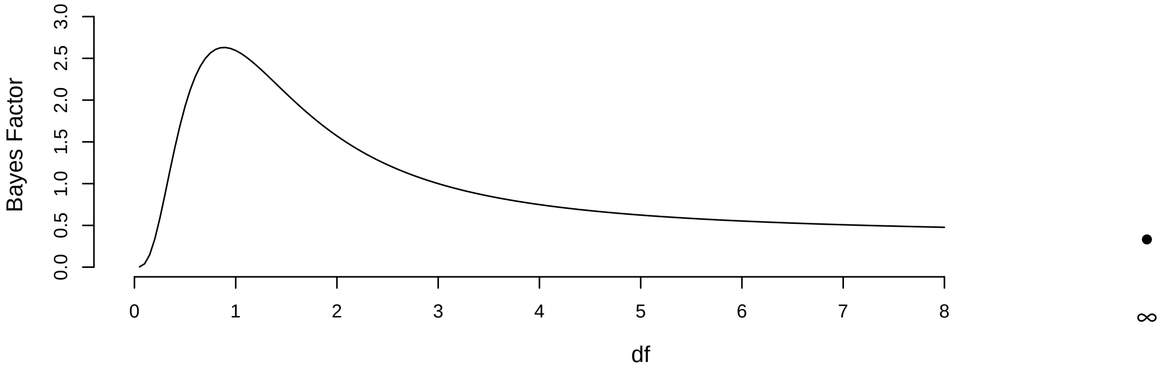

Figure 1: Bayes factors for the Bayarri-Berger example. The degrees of freedom parameter varies but the hyperparameters of the normal/inverse gamma prior are fixed.

(1996), Chen and Shao (1997), Gelman and Meng (1998), and Kong et al. (2003), this last also dealing with the problem of estimating multiple Bayes factors. When we need to estimate $B(h, h_1)$ for a wide range of h's, we face the problem that the estimate (5) is unstable when h is far from h_1. So it is better to select k hyperparameter points $h_1, \dots, h_k$, and get Markov chain samples from $\nu_{h_l,y}$ for each $l = 1, \dots, k$. The prior ν_{h_1} in the denominator of the left side of (5) is replaced by a mixture $w_1\nu_{h_1} + \cdots + w_k\nu_{h_k}$, with appropriately chosen weights, and the average is taken over the combined output of the k Markov chains. This results in accurate estimation of the Bayes factor for a wider range of hyperparameter values. The selection of the points $h_1, \dots, h_k$ is an interesting design issue.

We now return to example 2. A common approach for dealing with this variable selection problem is to introduce a hierarchical model that involves a prior distribution on the variables to include, and a particularly appealing choice is the independence Bernoulli prior

$$\rho_w(\gamma) = w^{p_\gamma}(1 - w)^{p - p_\gamma}, \tag{6}$$

indexed by a hyperparameter $w \in (0, 1)$ (recall that $p_\gamma \le p$ is the number of variables in subset γ). Under this prior, each variable has probability w of being included, independently of all the other variables. Here, the parameter is $\theta = (\gamma, \beta_0, \beta_\gamma, \sigma)$, and the two-level hierarchy (6) and (4) determines its prior distribution, which we will denote $\nu_{g,w}$.

There exist Markov chain Monte Carlo (MCMC) methods for dealing with this situation, where the state space includes the subset indicator γ. These produce Markov chains $(\gamma^{(1)}, \beta_0^{(1)}, \beta_\gamma^{(1)}, \sigma^{(1)}), (\gamma^{(2)}, \beta_0^{(2)}, \beta_\gamma^{(2)}, \sigma^{(2)}), \dots$ whose stationary distribution is the posterior distribution of $(\gamma, \beta_0, \beta_\gamma, \sigma)$ given $Y = y$, and are considerably more involved than MCMC methods for situations where the dimension of the parameter is not changing. We mention in particular Green's (1995) reversible jump MCMC, Carlin and Chib (1995), Godsill (2001), Dellaportas et al. (2002), the very recent paper Bartolucci et al. (2006) and the review paper by Han and Carlin (2001). From the subsequence $\gamma^{(1)}, \gamma^{(2)}, \dots$ the posterior distribution of γ given Y can be estimated, which enables variable selection.

This is by no means the end of the story, since such a method presupposes we have made a choice of the hyperparameters g and w to specify the prior. Loosely speaking, when w is large and g is small, the prior encourages models with many variables and small coefficients, whereas when w is small and g is large, the prior concentrates its mass on parsimonious models with large coefficients. Therefore, the hyperparameter $h = (g, w)$ plays a very important role, and in fact the choice of h in effect determines the model that will be used to carry out variable selection. For this reason there has been considerable work in finding good ways to choose h; see the references cited in section 3 of Clyde and George (2004). In particular, George and Foster (2000) show that the marginal distribution of Y can be written as $m_{g,w}(y) = \sum_\gamma p(y \mid \gamma)\, \rho_w(\gamma)$, where $p(y \mid \gamma)$ is available in closed form, so that $m_{g,w}(y)$ is available in closed form. Therefore in principle maximization of the function $m_{g,w}(y)$ with respect to g and w can be carried out, and the maximizing values can then be used. However, this is really feasible only if p is relatively small because of the large number of terms that go into the sum. (An exception arises if the design matrix X is orthogonal, in which case substantial simplifications arise and the numerical maximization of $m_{g,w}(y)$ becomes feasible even for moderately large p.)

The approach that involves the importance sampling estimate (5) can be helpful here. It is easy to see that for this Bayesian formulation, for $h_1 \neq h_2$, the prior distributions ν_{h_1} and ν_{h_2} are mutually absolutely continuous. Indeed, if $h_1 = (g_1, w_1)$ and $h_2 = (g_2, w_2)$, the Radon-Nikodym derivative (the likelihood ratio) is given very simply by

$$\left[\frac{d\nu_{h_1}}{d\nu_{h_2}}\right](\gamma, \beta_0, \beta_\gamma, \sigma) = \left(\frac{w_1}{w_2}\right)^{p_\gamma} \left(\frac{1-w_1}{1-w_2}\right)^{p-p_\gamma} \times \frac{\phi_{p_\gamma}\big(\beta_\gamma; 0, g_1\sigma^2(X_\gamma' X_\gamma)^{-1}\big)}{\phi_{p_\gamma}\big(\beta_\gamma; 0, g_2\sigma^2(X_\gamma' X_\gamma)^{-1}\big)}, \tag{7}$$

where $\phi_{p_\gamma}(u; a, V)$ is the density of the p_γ-dimensional normal distribution with mean a and covariance V, evaluated at u. (We note that the priors ν_h are distributions on $\{0,1\}^p \times \mathbb{R}^{p+1} \times (0, \infty)$ which are not absolutely continuous with respect to the product of counting measure on $\{0,1\}^p$ and Lebesgue measure on $\mathbb{R}^{p+1} \times (0, \infty)$, and this is the reason why we refer to (7) as a Radon-Nikodym derivative.) Therefore, we can apply (5) directly: we fix a particular hyperparameter $h_1 = (g_1, w_1)$, run a chain corresponding to the prior ν_{h_1}, and use the output to estimate the Bayes factors $B(h, h_1)$ simultaneously for "all" h. An important feature of this approach is that even though the ν_h's are probability measures that give mass to various sets of different dimensions, this does not cause any problem, because the calculation of $[d\nu_{h_1}/d\nu_{h_2}]$ in (7) takes place at a *point*.

To conclude, posterior distributions that consist of mixtures of distributions that live in spaces of different dimensions arise in a variety of settings, of which the variable selection problem discussed here is but one. Other situations include hierarchical models based on mixtures of Dirichlet process priors (Doss 2007), and a long list of examples is given in Green (1995). There is a developing methodology for running Markov chains in spaces that are changing dimension. There is also a developing methodology for doing model selection based on importance sampling and variants thereof. I hope that the discussion above will show that these methods can be usefully combined.

References

Bartolucci, F., Scaccia, L. and Mira, A. (2006). Efficient Bayes factor estimation from the reversible jump output. *Biometrika* **93** 41–52.

Bayarri, M. J. and Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science* **19** 58–80.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B: Methodological* **57** 473–484.

Chen, M.-H. and Shao, Q.-M. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics* **25** 1563–1594.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science* **19** 81–94.

Dellaportas, P., Forster, J. J. and Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing* **12** 27–36.

Doss, H. (2007). Estimation of Bayes factors for nonparametric Bayes problems via Radon-Nikodym derivatives. Preprint.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science* **13** 163–185.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747.

Godsill, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics* **10** 230–248.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732.

Han, C. and Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association* **96** 1122–1132.

Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D. and Tan, Z. (2003). A theory of statistical models for Monte Carlo integration (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* **65** 585–618.

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica* **6** 831–860.

Nicolae, D., Meng, X.-L. and Kong, A. (2007). Quantifying the fraction of missing information for hypothesis testing in statistical and genetic studies (with discussion). *Statistical Science* (to appear).

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. K. Goel and A. Zellner, eds.). Elsevier/North-Holland [Elsevier Science Publishing Co., New York; North-Holland Publishing Co., Amsterdam].