Bayesian Model Selection and Model Averaging

Larry Wasserman[1]
Carnegie Mellon University

This paper reviews the Bayesian approach to model selection and model averaging. In this review, I emphasize objective Bayesian methods based on noninformative priors. I will also discuss implementation details, approximations and relationships to other methods.

KEY WORDS AND PHRASES: AIC, Bayes Factors, BIC, Consistency, Default Bayes Methods, Markov Chain Monte Carlo.

1. INTRODUCTION

Suppose we are analyzing data and we believe that the data arise from one of a set of possible models $\mathcal{M}_1, \ldots, \mathcal{M}_k$. By a "model" I mean a set of probability distributions. For example, suppose the data consist of a normally distributed outcome Y and a covariate X and that two possibilities are entertained. The first possibility is that Y is unrelated to X and the second possibility is that Y is linearly related to X. Then $\mathcal{M}_1$ consists of the distributions for which $Y \sim N(\mu, \sigma^2)$ and $\mathcal{M}_2$ consists of the distributions for which $Y \sim N(\beta_0 + \beta_1 X, \tau^2)$. This is a simple example with only two models. There could be many models under consideration and each could be very complicated.

"Model selection" refers to the problem of using the data to select one model from the list of candidate models $\mathcal{M}_1, \ldots, \mathcal{M}_k$. "Model averaging" refers to the process of estimating some quantity under each model $\mathcal{M}_j$ and then averaging the estimates according to how likely each model is. For example, we could use model $\mathcal{M}_j$ together with the data to produce a prediction $\hat{Y}_j$ of a future observation. Then our overall prediction is $\sum_{j=1}^k w_j \hat{Y}_j$ where the weight w_j is used to reflect how probable it is that the model $\mathcal{M}_j$ generated the data.

The Bayesian solution to these problems is to compute the posterior probability $Pr(\mathcal{M}_j|\text{Data})$ for each model. For model selection, we choose the model that maximizes $Pr(\mathcal{M}_j|\text{Data})$. For model averaging, we use $w_j = Pr(\mathcal{M}_j|\text{Data})$ as the weights.

In this this paper I explain how to derive $Pr(\mathcal{M}_j|\text{Data})$. I also discuss the strengths and limitations of the Bayesian approach and I compare it to other approaches. A thorough review of Bayesian model selection, aimed at statisticians, is contained in Kass and Raftery (1995).

In Section 2, I consider a simple pedagogical example that illustrates the main points. In Section 3, I review Bayesian estimation theory. Section 4 is the heart of the paper;

[1]Presented at the Mathematical Psychology Symposium on "Methods for Model Selection" held on August 3 - 4 1997 in Bloomington, Indiana.

there I provide the details about Bayesian model selection and model averaging. In Section 5, I discuss computation and simulation. In Section 6, I discuss the relationship between Bayesian methods and other methods and I consider the strengths and weaknesses of the Bayesian approach. Section 7 contains a brief discussion of what happens when all the models under consideration are wrong. Section 8 contains an example and Section 9 discusses the frequentist behavior of Bayesian model selection. Finally, Section 10 contains concluding remarks.

2. A PEDAGOGICAL EXAMPLE

Before plunging into any details, it is helpful to consider a very simple problem. We observe n independent flips of a coin. Denote the outcomes by $Y^n = (Y_1, \ldots, Y_n)$ where each Y_i is either 0 or 1 (corresponding to tails and heads on the coin). Let $\theta = Pr(Y_i = 1)$ be the unknown probability of observing $Y_i = 1$. We have two theories. Theory one says that the coin is fair, i.e. $\theta = 1/2$. Theory two say that the coin is not fair, i.e. $\theta \neq 1/2$. The probability function for a single toss is $p_\theta(y) = \theta^y (1-\theta)^{1-y}$ where $y \in \{0, 1\}$. The two theories correspond to two sets of probability distributions:

$$\begin{aligned} \mathcal{M}_1 &= \{p_\theta;\ \theta = 1/2\}, \\ \mathcal{M}_2 &= \{p_\theta;\ \theta \in [0,1], \theta \neq 1/2\}. \end{aligned}$$

One questions of interest is this: what is a reasonable numerical measure of the evidence in favor of one theory over the other? We call this the "evidence problem." In the Bayesian approach, one measure of evidence is the posterior odds of one model versus the other model, i.e. the posterior probability of the second model divided by the posterior probability of the first model. In symbols, these odds are $Pr(\mathcal{M}_2|Y^n)/Pr(\mathcal{M}_1|Y^n)$. However, a more commonly used measure of evidence is the "Bayes Factor" B_n defined by

$$B_n = \frac{Pr(\mathcal{M}_2|Y^n)}{Pr(\mathcal{M}_1|Y^n)} \div \frac{Pr(\mathcal{M}_2)}{Pr(\mathcal{M}_1)} \tag{1}$$

which is the posterior odds in favor of $\mathcal{M}_2$ divided by the prior odds in favor of $\mathcal{M}_2$. This tells us how much the data have changed our odds in favor of one model over the other. Typically, one sets the prior odds of each theory to be equal, i.e. $Pr(\mathcal{M}_1) = Pr(\mathcal{M}_2) = 1/2$. In this case, the Bayes factor and the posterior odds are the same:

$$B_n = \frac{Pr(\mathcal{M}_2|Y^n)}{Pr(\mathcal{M}_1|Y^n)}. \tag{2}$$

In Section 4 we show that this reduces to

$$B_n = \frac{\int_0^1 \mathcal{L}(\theta) p(\theta) d\theta}{\mathcal{L}(1/2)} \tag{3}$$

where $\mathcal{L}(\theta) = \prod_{i=1}^n p_\theta(y_i)$ is the likelihood function (the probability of the data given the parameter θ) and $p(\theta)$ is a prior distribution for θ under model $\mathcal{M}_2$. In Section 4, we show that for a reasonable choice of the prior $p(\theta)$, it turns out that $\log B_n \approx b$ where

$$b = \log \mathcal{L}(\hat{\theta}) - \log \mathcal{L}(1/2) - (1/2) \log n \tag{4}$$

and $\hat{\theta} = n^{-1} \sum_i y_i$ is the maximum likelihood estimator of θ under model $\mathcal{M}_2$. Thus,

$$Pr(\mathcal{M}_1|Y^n) \approx \frac{1}{1+b}, \quad \text{and} \quad Pr(\mathcal{M}_2|Y^n) \approx \frac{b}{1+b}. \tag{5}$$

A second question, which we call "the prediction problem", is to predict future outcomes as accurately as possible. Specifically, having observed $Y^n = (Y_1, \ldots, Y_n)$, what is the probability that a new observation Y_{n+1} will be heads? The answer is

$$Pr(Y_{n+1} = 1|Y^n) = Pr(Y_{n+1} = 1|Y^n, \mathcal{M}_1)Pr(\mathcal{M}_1|Y^n) + Pr(Y_{n+1} = 1|Y^n, \mathcal{M}_2)Pr(\mathcal{M}_2|Y^n) \tag{6}$$

where $Pr(Y_{n+1} = 1|Y^n, \mathcal{M}_1) = 1/2$ since model $\mathcal{M}_1$ says that the coin is fair,

$$Pr(Y_{n+1} = 1|Y^n, \mathcal{M}_2) = \int_0^1 \theta \; p(\theta|y^n) d\theta, \tag{7}$$

and

$$p(\theta|y^n) = \frac{\mathcal{L}(\theta)p(\theta)}{\int_0^1 \mathcal{L}(\theta)p(\theta)d\theta} \tag{8}$$

is the posterior density for θ under model $\mathcal{M}_2$. It is possible to show that, for large sample sizes,

$$Pr(Y_{n+1} = 1|Y^n, \mathcal{M}_2) \approx \hat{\theta}. \tag{9}$$

Putting all the pieces together we obtain

$$Pr(Y_{n+1} = 1|Y^n) \approx \left(\frac{1}{2}\right) \frac{1}{1+b} + \hat{\theta}\frac{b}{1+b}. \tag{10}$$

Model averaging differs from more traditional prediction techniques in which we would use some criterion such as AIC (Section 6), and select either model $\mathcal{M}_1$, in which case we predict $Pr(Y_{n+1} = 1) = 1/2$, or we select model $\mathcal{M}_2$, in which case we predict $Pr(Y_{n+1} = 1) = \hat{\theta}$. Model averaging avoids having to choose one model. Instead, we let the data give the competing models different weights.

3. BAYESIAN ESTIMATION THEORY

For this section, we assume we have only one model

$$\mathcal{M} = \{p_\theta(y); \theta \in \Omega\}$$

where θ is an unknown parameter in some parameter space Ω and $p_\theta(y)$ is a probability density function for Y which depends on the parameter θ. An example is the model $Y \sim N(\mu, \sigma^2)$ so that $\theta = (\mu, \sigma)$, and $p_\theta(y) = (\sigma\sqrt{2\pi})^{-1} \exp\{-(1/2)(y-\mu)^2/\sigma^2\}$.

Assume we have n independent, identically distributed observations $Y^n = (Y_1, \ldots, Y_n)$. For most of this section we will assume that θ is scalar, though everything carries over to multidimensional parameters.

The likelihood function is defined by

$$\mathcal{L}(\theta) = p_\theta(y^n) = \prod_{i=1}^{n} p_\theta(y_i) \tag{11}$$

and the log-likelihood function is $\ell(\theta) = \log \mathcal{L}(\theta)$. Note that the likelihood function is just the probability of the observed data given the parameter θ. The maximum likelihood estimator (MLE) $\hat{\theta}$ is the value of θ that maximizes $\mathcal{L}(\theta)$. Thus, $\hat{\theta}$ is the value of the parameter that makes the observed data most likely. Under weak conditions, $\hat{\theta}$ has many useful properties. First, the MLE is consistent, meaning that $\hat{\theta}$ converges to the true value of θ with probability 1. Second, it has, asymptotically in sample size, a normal distribution. To explain this last point, we need a bit more notation. The Fisher information I_θ is defined by $I_\theta = E_\theta(s_\theta^2)$ where $s_\theta = \partial \log p_\theta(y)/\partial\theta$ and E_θ refers to expectation with respect to p_θ. Then, for large n, $\hat{\theta} \approx N(\theta_0, 1/(nI_{\theta_0}))$ where θ_0 denotes the true value of θ. As a consequence, $\hat{\theta}$ is about $1/\sqrt{n}$ from the true value; formally, $\hat{\theta} - \theta_0 = O_P(n^{-1/2})$. It also follows that the interval $\hat{\theta} \pm 1.96$ SE is, asymptotically, a 95 per cent confidence interval for θ, where $\mathrm{SE}^{-1} = \sqrt{nI_{\hat{\theta}}}$. Furthermore, the MLE is known to be asymptotically optimal: of all regular estimators, $\hat{\theta}$ has smallest asymptotic variance.

In Bayesian estimation theory, we introduce a prior distribution $p(\theta)$. The prior $p(\theta)$ represents how likely different values of θ are, before seeing the data. After observing Y^n we compute the posterior distribution, using Bayes theorem:

$$p(\theta|y^n) = \frac{p(y^n|\theta)p(\theta)}{\int p(y^n|\theta)p(\theta)d\theta} = \frac{\mathcal{L}(\theta)p(\theta)}{\int \mathcal{L}(\theta)p(\theta)d\theta} = \frac{\mathcal{L}(\theta)p(\theta)}{m(y^n)} \tag{12}$$

where $m(y^n) = \int \mathcal{L}(\theta)p(\theta)d\theta$ is called the "normalizing constant." Since $m(y^n)$ is a constant not depending on θ, we often write Bayes' theorem as

$$p(\theta|y^n) \propto \mathcal{L}(\theta)p(\theta). \tag{13}$$

A point estimate of θ is obtained by finding the posterior mean, $\overline{\theta} = E(\theta|y^n) = \int \theta p(\theta|y^n)d\theta$. An interval estimate $I = (a, b)$ is found by choosing numbers a and b such that, for a given α, $Pr(\theta \in I|Y^n = y^n) = \int_a^b p(\theta|y^n)d\theta = 1 - \alpha$.

It can be shown that $\overline{\theta} - \theta_0 = O_P(n^{-1/2})$ and $\overline{\theta} - \hat{\theta} = O_P(n^{-1})$. This means that the difference between Bayes estimation and maximum likelihood estimation is an order of magnitude smaller than the estimation error.

A serious practical problem is to choose the prior $p(\theta)$. The subjective theory of Bayesian inference prescribes that the prior $p(\theta)$ is chosen to represent someone's prior opinions about θ. The more common view is objective Bayesianism in which $p(\theta)$ is chosen to be "noninformative" in some sense. It is beyond the scope of this paper to discuss this philosophically charged issue, but see Kass and Wasserman (1996) for a review of noninformative priors.

For our purposes it suffices to note that in many problems, there are agreed upon choices for noninformative priors. A commonly used noninformative prior is Jeffreys' (1961) prior which is defined by $p(\theta) \propto |I_\theta|^{1/2}$. For example, if the data are from a $N(\theta, 1)$ distribution, Jeffreys' prior turns out to be the flat prior $p(\theta) \propto 1$. In a coin flipping problem, Jeffreys' prior is $p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$. If $\theta = (\theta_1, \ldots, \theta_p)$ is a vector, Jeffreys' prior is defined in the following way. Let s_θ be a vector whose i^{th} component is $\partial \log p_\theta(y)/\partial\theta_i$. Let I_θ be a matrix defined by $I_\theta = E_\theta(s_\theta s_\theta^T)$. Then Jeffreys' prior is defined as $p(\theta) \propto |\det(I_\theta)|^{1/2}$ where $\det(A)$ is the determinant of the matrix A.

Here is one sense in which Jeffreys' prior is noninformative. Consider finding a one-sided $1-\alpha$ interval I. For example, let $I = (-\infty, a]$ where a is chosen so that $\int_{-\infty}^a p(\theta|y^n)d\theta = 1-\alpha$. Thus, $Pr(\theta \in I|Y^n) = 1-\alpha$. This is a Bayesian one-sided confidence interval. What happens if we decide to use the Bayesian interval I as if it were a confidence interval? Recall that the "coverage" of a confidence interval is the probability that the interval contains the true value of the parameter. It can be shown that the coverage of I, denoted Coverage(I), satisfies Coverage(I) $= 1 - \alpha + O(n^{-1/2})$. In other words, Bayesian confidence intervals can also be interpreted as approximate frequentist intervals. However, if Jeffreys' prior was used, then it turns out that Coverage(I) $= 1-\alpha+O(n^{-1})$. In other words, Jeffreys' prior makes Bayesian intervals behave even more like usual confidence intervals.

One feature of noninformative priors is that they are often "improper" which means that $\int p(\theta)d\theta = \infty$. Nonetheless, the posterior $p(\theta|y^n) \propto \mathcal{L}(\theta)p(\theta)$ is usually proper despite the fact that the prior is not. Note, however, that improper priors are only defined up to a constant. If $p(\theta)$ is an improper prior, we can define a new prior $q(\theta) = cp(\theta)$ where $c > 0$ is an arbitrary positive number. It is easy to check that the posterior obtained from the prior $q(\theta)$ is the same as the posterior obtained from $p(\theta)$. This undefined constant in the prior poses no problem for estimation. But we shall see that it does create trouble for model selection.

As well as estimation, we might be interested in prediction. The predictive distribution is defined by

$$\hat{p}(y) = \int p_\theta(y) p(\theta | y^n) d\theta. \tag{14}$$

We use $\hat{p}(y)$ to make predictive probability statements about a future observation Y. It can be shown that

$$\hat{p}(y) \approx p_{\hat{\theta}}(y) \tag{15}$$

which gives a simple approximation for the predictive density.

4. BAYESIAN MODEL SELECTION AND AVERAGING

Suppose we have k models $\mathcal{M}_1, \ldots, \mathcal{M}_k$ under consideration. Each model consists of a set of probability densities for a random variable Y:

$$\mathcal{M}_j = \{p_{\theta_j}(y); \theta_j \in \Omega_j\} \tag{16}$$

where θ_j is the unknown parameter in the j^{th} model. The likelihood function for model $\mathcal{M}_j$ is

$$\mathcal{L}_j(\theta_j) = \prod_i p_{\theta_j}(y_i) \tag{17}$$

the product being over the n observations $y^n = (y_1, \ldots, y_n)$. The log-likelihood is denoted by $\ell_j(\theta_j) = \log \mathcal{L}_j(\theta_j)$ and we let $\hat{\theta}_j$ denote the maximum likelihood estimate of θ_j.

We assign prior probabilities $Pr(\mathcal{M}_j)$ to the models. In this paper we assume, as is commonly done, that $Pr(\mathcal{M}_j) = 1/k$, $j = 1, \ldots, k$. For each model, we also specify a prior $p_j(\theta_j)$ for the parameter θ_j. The posterior for model $\mathcal{M}_j$ is easily found from Bayes' theorem to be

$$Pr(\mathcal{M}_j | Y^n = y^n) = \frac{p(y^n | \mathcal{M}_j) Pr(\mathcal{M}_j)}{\sum_r p(y^n | \mathcal{M}_r) Pr(\mathcal{M}_r)}. \tag{18}$$

¿From elementary probability we find that $p(y^n|\mathcal{M}_j) = \int p_{\theta_j}(y^n) p_j(\theta_j) d\theta_j$ and, recalling that $Pr(\mathcal{M}_j) = 1/k$ for each model, and that $p_{\theta_j}(y^n) = \mathcal{L}_j(\theta_j)$, we see that

$$Pr(\mathcal{M}_j | Y^n = y^n) = \frac{m_j}{\sum_{r=1}^k m_r} \tag{19}$$

where

$$m_r = \int \mathcal{L}_r(\theta_r) p_r(\theta_r) d\theta_r. \tag{20}$$

The "Bayes factor" for $\mathcal{M}_i$ versus $\mathcal{M}_j$ is defined to be

$$B_{ij} = \frac{Pr(\mathcal{M}_i | Y^n = y^n)}{Pr(\mathcal{M}_j | Y^n = y^n)} = \frac{m_i}{m_j} = \frac{\int \mathcal{L}_i(\theta_i) p_i(\theta_i) d\theta_i}{\int \mathcal{L}_j(\theta_j) p_j(\theta_j) d\theta_j}. \tag{21}$$

Bayes Factor	Interpretation
$B_{ij} < 1/10$	Strong Evidence For $\mathcal{M}_j$
$1/10 < B_{ij} < 1/3$	Moderate Evidence For $\mathcal{M}_j$
$1/3 < B_{ij} < 1$	Weak Evidence For $\mathcal{M}_j$
$1 < B_{ij} < 3$	Weak Evidence For $\mathcal{M}_i$
$3 < B_{ij} < 10$	Moderate Evidence For $\mathcal{M}_i$
$B_{ij} > 10$	Strong Evidence For $\mathcal{M}_i$

Table 1. Jeffreys' Scale of Evidence For Bayes Factors.

The Bayes factor gives a measure of the evidence for model $\mathcal{M}_j$ versus model $\mathcal{M}_i$. If $B_{ij} = 10$, for example, then model $\mathcal{M}_i$ is ten times more likely than model $\mathcal{M}_j$, given the data. Jeffreys recommended a scale of evidence for interpreting Bayes factors. Table 1 gives Jeffreys' scale. (Our version is a slight modification of his scale.)

To predict a new observation Y, we use the predictive distribution

$$\hat{p}(y) = \sum_{j=1}^{k} \hat{p}_j(y) Pr(\mathcal{M}_j | Y^n) \tag{22}$$

where

$$\hat{p}_j(y) = \int p_{\theta_j}(y) \pi(\theta_j | y^n) d\theta_j \tag{23}$$

is the predictive distribution from model $\mathcal{M}_j$.

There are two practical problems to be solved if we are to make use of these ideas. First, we have to choose the priors $p_j(\theta_j)$ and second, we have to compute the integrals in (21).

If we try to use a noninformative prior for $p_j(\theta_j)$ we run into a problem. Recall that noninformative priors are often improper and that improper priors are only defined up to a constant. So if $p_j(\theta_j)$ is an improper prior for θ_j and c_j is an arbitrary positive constant, then $q_j(\theta_j) = c_j p_j(\theta_j)$ could also be used as a prior. But now the Bayes factor using these priors q_j is

$$B_{ij} = \frac{c_i}{c_j} \frac{\int \mathcal{L}_i(\theta_i) p_i(\theta_i) d\theta_i}{\int \mathcal{L}_j(\theta_j) p_j(\theta_j) d\theta_j} \tag{24}$$

so the Bayes factors and the posterior probabilities are ill-defined since there are arbitrary constants floating around in the equations.

In some cases, we can solve the prior problem and the integration problem in the following simple way. Let $\hat{\ell}_j = \ell_j(\hat{\theta}_j)$ and let d_j be the dimension of Ω_j. It can be shown that (Kass and Wasserman 1995) $m_j = \hat{m}_j + O_P(1)$ where

$$\hat{m}_j = \hat{\ell}_j - \frac{d_j}{2} \log n. \tag{25}$$

This result means that m_j can be approximated by $\hat{m}_j$ which requires no integration and does not depend on the prior. The catch is that the error $O_P(1)$ does not go to 0 as n gets large. This is not as bad as it sounds for two reasons. First, quantities like m_j typically tend to ∞ or $-\infty$ as sample size increases. Hence, the error of the approximation relative to the quantity we are estimating does tend to 0. In other words, $|\hat{m}_j - m_j|/|m_j| \to 0$ in probability. Second, there are certain priors for which the approximation (25) has an error of size $O_P(n^{-1/2})$. One example of such a prior is a "unit information prior" which is discussed in Kass and Wasserman (1995). (For example, if $\mathcal{M}_2 = \{p_\theta; \theta \in \mathcal{R}\}$ and $\mathcal{M}_1 = \{p_\theta; \theta = \theta_0\}$, then the unit information prior is $N(\theta_0, I_{\theta_0}^{-1})$, where I_{θ_0} is the Fisher information based on a single observation.) A second prior that justifies the smaller error term is Jeffreys' prior. As noted in the previous section, Jeffreys' prior is usually improper and thus is plagued by the arbitrary constant. But if we decree that, as a matter of convention, we shall define the arbitrary constant in front of Jeffreys' prior to be $c_j = (2\pi)^{-d_j/2}$, then it turns out that again, the error in (25) is $O_P(n^{-1/2})$. In short, if we adopt the noninformative prior $p_j(\theta_j) = \{2\pi\}^{-d_j/2}|\det(I_\theta)|^{1/2}$ then $\hat{m}_j$ is a fairly accurate approximation of m_j.

If we use the approximation (25), then

$$Pr(\mathcal{M}_j|Y^n) \approx \frac{e^{\hat{m}_j}}{\sum_{r=1}^k e^{\hat{m}_r}}. \tag{26}$$

Equation (26) is the most important equation in this paper. It gives a very easy, albeit approximate, way to compute the posterior probability of each model.

¿From (25) and the definition of B_{ij} we observe that

$$\log B_{ij} \approx \hat{\ell}_i - \hat{\ell}_j + \frac{d_j - d_i}{2} \log n \tag{27}$$

which is also known as BIC or the Schwarz criterion (Schwarz 1978, Kass and Raftery 1995). Thus, BIC can be regarded as an approximation to the log Bayes factor.

If we are interested in prediction, then can substitute (26) into (22). Further, we can approximate (23) by $p_{\hat{\theta}_j}(y)$ leading to

$$\hat{p}(y) \approx \sum_j p_{\hat{\theta}_j}(y) \frac{e^{\hat{m}_j}}{\sum_{r=1}^k e^{\hat{m}_r}} \tag{28}$$

The above approximate method seems to work well in well behaved problems with moderate to large sample sizes. But in some irregular cases, these approximations can break down.

An alternative theory, still being developed, that might be effective in these more delicate problems is the theory of "intrinsic Bayes factors" developed by Berger and Pericchi (1996).

Briefly, the idea is this. Suppose we are comparing two models $\mathcal{M}_1$ and $\mathcal{M}_2$. (The extension to several models is straightforward.) We start with improper noninformative priors $p_j(\theta_j)$ for each model. A small subset S of the data Y^n, called the training set, is used to update the prior by Bayes' theorem. Denote this posterior by p_j^S. The training set is chosen to be minimal, which means that p_j^S is proper but no subset of S will yield a proper posterior. We can compute the Bayes factor using the priors p_1^S and p_2^S; denote the resulting Bayes factor by B_{ij}^S. Note that B_{ij}^S is well defined since proper priors have no undefined constants. But B_{ij}^S will depend on the choice of training set. Berger and Pericchi suggest computing B_{ij}^S for each possible minimal training set. They define the intrinsic Bayes factor as the average over all the resulting Bayes factors.

It is beyond the scope of this paper to explain this methodology in detail. The reader is referred to Berger and Pericchi (1996). A related technique, called fractional Bayes factors, is discussed in O'Hagan (1995) and De Santis and Spezzaferri (1997).

5. COMPUTING POSTERIOR PROBABILITIES

Suppose that for each model $\mathcal{M}_j$ we do specify a prior $p_j(\theta_j)$ for the parameter θ_j of that model. We may want to compute $Pr(\mathcal{M}_j|Y^n)$ exactly instead of using the approximation (26). ¿From (19), it follows that to compute $Pr(\mathcal{M}_j|Y^n)$, we need to be able to compute $m_j = \int \mathcal{L}_j(\theta_j)p_j(\theta_j)d\theta_j$.

Let us thus set aside the model selection problem for the moment and ask, more generally, how do we compute an integral of the form $m_j = \int \mathcal{L}_j(\theta_j)p_j(\theta_j)d\theta_j$? One approach is to use an analytic approximation. ¿From Tierney and Kadane (1986) it follows that $m_j = \hat{m}_j(1 + O_P(n^{-1}))$ where

$$\hat{m}_j = \mathcal{L}_j(\hat{\theta}_j)p(\hat{\theta}_j)\{2\pi\}^{d_j/2}|\det(H)|^{1/2}. \tag{29}$$

In (29), $\hat{\theta}_j$ is the mode of the posterior $p_j(\theta_j|y^n)$ and H is the matrix of second derivatives of the log-posterior, evaluated at $\hat{\theta}_j$. This approximation arises from a method called "Laplace's method."

A more exact method has become popular lately, namely simulation. The idea is this: we draw a random sample $\theta_j^1, \ldots, \theta_j^N$ from the posterior $p_j(\theta_j|y^n)$. We then try to find a way to use the sample to estimate m_j.

These days, the most common way to simulate from a posterior is to use Markov chain Monte Carlo (MCMC). Briefly, MCMC works like this. Pick a starting point θ_j^0. Draw a "candidate value" ψ from a Normal centered at θ_0 with some variance b^2. Let

$$r = \min\left\{\frac{\mathcal{L}_j(\psi)p_j(\psi)}{\mathcal{L}_j(\theta_j^0)p_j(\theta_j^0)}, 1\right\}. \tag{30}$$

Now draw $U \sim (0,1)$, i.e. U is a random number between 0 and 1. If $U < r$ set $\theta_j^1 = \psi$. Otherwise, set $\theta_j^1 = \theta_j^0$. Now draw a candidate from a Normal centered at θ_j^1 with variance b^2 and so on. Continue the process until we have N draws. It can be shown that the sample approximates a random draw from the posterior. This is, of course, a brief sketch of the idea. See Carlin and Louis (1996), Gelfand and Smith (1990), Gelman, Carlin, Stern and Rubin (1995), Gilks, Richardson, and Spiegelhalter (1995), Tanner and Wong (1987) and Tierney (1994) for more details.

Having obtained a sample from the posterior, we still need to somehow use the sample to estimate m_j. There are many ways to do this. Most of them are reviewed in DiCiccio, Kass, Raftery and Wasserman (1997). Here, I will describe only one which is perhaps best called the "density estimation approach." Recall that from Bayes' theorem we have

$$p_j(\theta_j|y^n) = \frac{\mathcal{L}_j(\theta_j)p_j(\theta_j)}{m_j}. \tag{31}$$

Hence,

$$m_j = \frac{\mathcal{L}_j(\theta_j)p_j(\theta_j)}{p_j(\theta_j|y^n)}. \tag{32}$$

This identity holds for all values of θ_j. So we can pick any value $\hat{\theta}_j$ of θ_j and calculate $m_j = \mathcal{L}_j(\hat{\theta}_j)p_j(\hat{\theta}_j)/p_j(\hat{\theta}_j|y^n)$. Evaluating $\mathcal{L}_j(\hat{\theta}_j)$ and $p_j(\hat{\theta}_j)$ is usually easy since these are given functions. The difficulty is to evaluate $p_j(\hat{\theta}_j|y^n)$. How do we evaluate the height of the posterior density at the point $\hat{\theta}_j$? Since we have a sample from this density, this is just a standard density estimation problem.

In general, if we have a sample $X_1, \ldots, X_n$ from a distribution with density $f(x)$ and we want to estimate $f(x)$, there are many techniques for doing so. The most common are histograms and kernel density estimation. Silverman (1986) is an excellent, readable introduction to density estimation. Thus, we apply any density estimation technique to use the sample $\theta_j^1, \ldots, \theta_j^N$ to get an estimate $\hat{p}_j(\hat{\theta}_j|y^n)$ of $p_j(\hat{\theta}_j|y^n)$. Then our estimate of m_j is $\hat{m}_j = \mathcal{L}_j(\hat{\theta}_j)p_j(\hat{\theta}_j)/\hat{p}_j(\hat{\theta}_j|y^n)$. This process is repeated for each model to get estimates $\hat{m}_1, \ldots, \hat{m}_k$. Finally, we get the posterior probabilities of the models by

$$Pr(\mathcal{M}_j|Y^n = y^n) = \frac{m_j}{\sum_{r=1}^k m_r} \approx \frac{\hat{m}_j}{\sum_{r=1}^k \hat{m}_r}. \tag{33}$$

6. AIC, BIC AND BAYES

There are many methods for choosing between competing models. We have discussed the Bayesian approach. The two most common other approaches are AIC and BIC. In the AIC approach (Akaike 1973), one chooses the model that maximizes

$$\log \mathcal{L}_j(\hat{\theta}_j) - d_j. \tag{34}$$

The BIC (Schwarz 1978) method is to choose the model that maximizes

$$\log \mathcal{L}_j(\hat{\theta}_j) - (d_j/2)\log n. \tag{35}$$

As we saw in Section four, BIC and Bayes methods are asymptotically equivalent under weak conditions. For the purposes of this section, I will therefore lump BIC and Bayes together and talk of them as if they were the same procedure.

Other papers from this conference discuss AIC so I will not dwell on it in great detail. However, a brief comparison is worthwhile. AIC is aimed at solving the following prediction problem: find the model $\mathcal{M}_j$ that produces estimates of the density $\hat{f}_j \equiv p_{\hat{\theta}_j}$ which is close, on average to the true density. Here, close is measured by the Kullback-Leibler distance $K(f,g) = \int f(y)\log[f(y)/g(y)]dy$. Let's call this the "Akaike prediction problem." There appears to be some debate on the relative merits of AIC versus BIC. It should be noted, however, that AIC and BIC were designed for different purposes. AIC was designed to solve the Akaike prediction problem while – at least from the Bayesian perspective – BIC was designed to find the most probably model given the data. Here are a few more comments on this debate:

(1) If one of the models is correct, then, asymptotically in sample size, BIC chooses the correct model but AIC does not. However, this does not preclude the fact that AIC might choose the correct model more often in small samples. Zhang (1993) have some results on this issue.

(2) As discussed in Section 4, we can exponentiate BIC and treat it as an approximate posterior. Then the models can be averaged. I am unaware of any way to do model averaging using AIC. Thus, one is confined to selecting one model.

(3) Here is a technical comment. Suppose the models are nested, which means that $\mathcal{M}_1 \subset \mathcal{M}_2 \subset \cdots \subset \mathcal{M}_k$ and that the true density is in $\mathcal{M}_k$ but is not in any of the other models. If the sample size is large relative to the dimension of the largest model, then the optimal solution to the Akaike prediction problem is to choose $\mathcal{M}_k$. This follows since each sub-model has non-zero asymptotic bias while $\mathcal{M}_k$ has bias and variance going to 0. So why does AIC not choose $\mathcal{M}_k$ asymptotically? The reason is that Akaike assumed that the true density is in $\mathcal{M}_k$ but is close, relative to sampling error, to a sub-model. Technically, the true density is changing as sample size changes. In this sense, AIC is adding the prior information that the true model might be close to a sub-model. This suggests that it would be interesting to compare the performance of Bayesian model averaging to AIC since Bayesian methods more directly include the prior information that sub-models are probably. To my knowledge, a careful comparison has not been done.

(4) There are some arguments that AIC is better when none of the models is true. We discuss that general issue in the next section.

There are yet more methods for choosing among models. For example: minimum description length, cross-validation, modified versions of AIC etc. We shall not consider these techniques further in this paper.

7. WHEN ALL THE MODELS ARE WRONG

Whenever people discuss model selection, it is inevitable that someone says "but surely all the models are wrong." Except in special cases, this is probably true. The methods in Section 4 unabashedly assume that one of the models under consideration contains the true distribution. So does it make sense to compare any finite list of models when we don't literally believe any of them? The answer is yes for several reasons.

First, we would hope that, while none of the models is exactly correct, at least one is approximately correct. It behooves the data analyst to do common sense exploratory analysis – checking residuals for example – to make sure that not all of the models are heinously wrong. Surely, no competent data analyst compares models which are all horribly wrong. So the methods in Section 4 can be regarded as being useful under the reasonable assumption that at least one model is approximately correct.

Second, even when all models are wrong, it is useful to consider the relative merits of two models. Newtonian physics and general relativity are both wrong. Yet it makes sense to compare the relative evidence in favor of one or the other. Our conclusion would be: "under the tentative working hypothesis that one of these two theories is correct, we find that the evidence strongly favors general relativity." It is understood that the working hypothesis that "one of the models is correct" is wrong. But it is a useful, tentative hypothesis and, proceeding under that hypothesis, it makes sense to evaluate the relative posterior probabilities of those hypotheses.

If we are seriously concerned about the correctness of all the models, and we have a large sample, then one might consider doing a nonparametric analysis. For example, one can consider an infinite list of increasingly complex models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ with the property that $\cup_{j=1}^{\infty} \mathcal{M}_j$ is infinite dimensional. Such a set of models is called a "sieve" (Grenander 1981, Wong and Shen 1985). Choosing a model from a sieve is a hard problem and is an important current research topic in statistics.

8. A REGRESSION EXAMPLE

Consider the problem of choosing which covariates to include in a regression model. For the i^{th} subject, we have a vector of data $(Y_i, X_{i1}, \ldots, X_{ip})$. We have data on n subjects and we are contemplating a model of the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i \tag{36}$$

where the $\epsilon_i' s$ are independent $N(0, \sigma^2)$.

Suppose we have reason to believe that some of the covariates should be dropped from the model. Let S represent a subset of $\{0, 1, \ldots, p\}$ and let $\mathcal{M}_S$ represent the model in which all the $\beta_i' s$ in (36) are zero except for $\{\beta_j; j \in S\}$. For example, if $S = \{0, 1, 3\}$ then $\mathcal{M}_S$ refers to the model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \epsilon_i$. Note that there are 2^{p+1} models under consideration each corresponding to including or leaving out various covariates. We will now compute the posterior probability of each model using the approximation outlined in Section 4.

We may write model $\mathcal{M}_S$ as $Y \sim N(X_S \beta_S, \sigma^2)$ where X_S is a matrix with one column for each covariate in the model. For example, if $S = \{0, 1\}$, then the first column of X_S is all one's and the second column is $(X_{11}, \ldots, X_{n1})$. Also, β_S refers to the regression parameters of the model. Let $\theta_S = (\beta_S, \sigma)$ refer to all the parameters in model $\mathcal{M}_S$. Since we are assuming normality, the log-likelihood for model $\mathcal{M}_S$ is, after some algebra,

$$\begin{aligned} \ell_S(\theta_S) &= --\frac{n}{2}\log(2\pi) - \log\sigma - -\frac{1}{2\sigma^2}(Y - X_S\beta_S)^T(Y - X_S\beta_S) &(37)\\ &= --\frac{n}{2}\log(2\pi) - \log\sigma - -\frac{1}{2\sigma^2}[v_S A_S^2 + (\beta_S - \hat{\beta}_S)^T X_S^T X_S(\beta_S - \hat{\beta}_S)] &(38)\end{aligned}$$

where $\hat{\beta}_S = (X_S^T X_S)^{-1} X_S^T Y$, $v_S = n - c_S$, c_S is the number of columns of X_S, $A_S^2 = (Y - \hat{Y}_S)^T (Y - \hat{Y}_S)/v_S$ and $\hat{Y}_S = X_S \hat{\beta}_S$. Maximizing this log-likelihood, we see that the MLE of β_S is just $\hat{\beta}_S$. The MLE of σ is $\hat{\sigma}_S = v_S A_S^2/n$. Inserting these MLE's into the above equation, we find that (25) becomes

$$\hat{m}_S = -n\log\hat{\sigma}_S - \frac{n}{2}\log(2\pi) - \frac{n}{2}. \tag{39}$$

Then

$$Pr(\mathcal{M}_S|Data) \approx \frac{e^{\hat{m}_S}}{\sum_S e^{\hat{m}_S}} \tag{40}$$

the sum being over all possible models.

Suppose we have a new subject with covariate values $(X_1^*, \ldots, X_p^*)$ and we want to predict his outcome Y. Model $\mathcal{M}_S$ gives the predicted value $\hat{Y}_S = X_S^* \hat{\beta}_S$ where X_S^* is a matrix with one row only, corresponding the new subject's covariate values. We predict Y using model averaging by using the following equation:

$$\hat{Y} = \sum_S \hat{Y}_S Pr(\mathcal{M}_S|Data) \approx \frac{\sum_S X_S^* \hat{\beta}_S e^{\hat{m}_S}}{\sum_S e^{\hat{m}_S}} \tag{41}$$

where again the sums are over all sub-models.

This approach provides an alternative to the usual stepwise procedures. If the number of predictors p is large, then summing over all sub-models is not feasible. In this case, deterministic or random searches over some subset of the set of all models must be used.

9. THE FREQUENTIST BEHAVIOR OF BAYESIAN MODEL SELECTION

Let us assume that the true distribution which generates the data is in exactly one of the candidate models. Let $\mathcal{M}_j$ denote the model containing the true density. Then, under weak conditions, it can be shown that, for $i \neq j$,

$$\frac{Pr(\mathcal{M}_i|Y^n)}{Pr(\mathcal{M}_j|Y^n)} \to 0 \tag{42}$$

in probability. This means that the posterior probability of the true model goes to one and that the posterior probability of the other models go to zero. BIC has the same asymptotic behavior: it selects the true model asymptotically.

When the models are nested, the true density can be in more than one model. For example, suppose model 1 says that $Y \sim N(\beta_0, \sigma^2)$ and model 2 says that $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$. Suppose that $\beta_1 = 0$. Then both model 1 and model 2 are true. In such a case, the posterior of the smallest (lowest dimensional) model tends to one and the others tend to zero. Thus, Bayesian model selection automatically incorporates Occam's razor.

For separate models, things work even better. Suppose for example that $\mathcal{M}_1$ and $\mathcal{M}_2$ have no overlap and, moreover, that no distribution in one model is the limit of distributions from the other. Then, again, the posterior of the true model converges to one. And in this case, the convergence can be shown to happen very quickly. Technically, the convergence is at an exponential rate. These facts follow from standard asymptotic theorems. See Haughton (1988) and Kass and Wasserman (1995) for example.

An interesting question is how the small sample behavior of Bayesian model selection and model averaging compare to competitors like AIC. As far as I know, there are not any systematic comparisons which is unfortunate. Another interesting question is what happens to the posterior probabilities when the true distribution is in none of the models. Under regularity conditions, what happens is this: the posterior probability of the model that contains the closest distribution to the true distribution, tends to one.

10. CONCLUSION

When faced with several candidate models, the analyst can either choose one model or average over the models. Bayesian methods provide a set of tools for these problems.

Bayesian methods also give us a numerical measure of the relative evidence in favor of competing theories. The main points of this paper are:

(1) Bayesian model selection and model averaging is a conceptually simple, unified approach.

(2) For well-behaved models, and moderate to large sample sizes, BIC provides a useful approximation to the log Bayes factor.

(3) For non-standard problems, intrinsic Bayes factor might be a useful approach.

(4) There is no need to choose one model. It is possible to average the predictions from several models.

(5) Simulation methods make it feasible to compute posterior probabilities in many problems.

(6) I have emphasized objective Bayesian methods that do not require subjective prior distributions. This reflects my bias that subjective Bayes is unreliable, impractical, and not scientific. But there is vocal, and perhaps legitimate opposition to this viewpoint in the statistical community. Also, there is an approach called Robust Bayesian inference (Berger 1984, Berger and Delampady 1987) in which Bayesian inference is performed with a set of priors rather than a single prior. This provides a bridge between subjective and objective Bayesian methods.

ACKNOWLEDGMENTS

I am grateful to the organizers of the workshop for inviting me to speak and for financial support. I would also like to thank Fulvio De Santis, Rob Kass, Richard Golden and Isabella Verdinelli for providing feedback on an earlier version of this paper. Research was supported by NIH grant RO1-CA54852 and NSF grants DMS-9303557 and DMS-9357646.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory,* eds. B.N. Petrov and F. Csaki. Akad. Kiado, Budapest, 267-281.

Berger, J. (1984). The robust Bayesian viewpoint (with discussion), In *Robustness in Bayesian Statistics*, ed. J. Kadane, North Holland, Amsterdam.

Berger, J. and Delampady, M. (1987). Testing precise hypotheses (with discussion), *Statistical Science*, **2**, 317-352.

Berger, J.O. and Pericchi, L. (1994). The intrinsic Bayes factor for model selection and prediction, *The Journal of the American Statistical Association,* **91**, 109-122.

Carlin, B. and Louis, T. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall, New York.

De Santis, F. and Spezzaferri, F. (1997). Alternative Bayes factors for model selection, *The Canadian Journal of Statistics,* to appear.

DiCiccio, T., Kass, R., Raftery, A. and Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations, *The Journal of the American Statistical Association,* **92**, 903-915.

Gelfand, A. and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities, *Journal of the American Statistical Association,* **85**, 398-409.

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

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1995). *Markov Chain Monte Carlo in Practice.* London: Chapman and Hall.

Grenander, U. (1981). *Abstract Inference.* Wiley: New York.

Haughton, D. (1988). On the choice of a model to fit data from an exponential family, *The Annals of Statistics*, **16**, 342-355.

Jeffreys, H. (1961). *Theory of Probability. Third Edition.* Clarendon Press: Oxford.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors, *Journal of the American Statistical Association*, **90**, 377-395.

Kass, R.E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion, *Journal of the American Statistical Association*, **90**, 928-934.

Kass, R.E. and Wasserman, L. (1996). The selection of prior distributions by formal rules, *Journal of the American Statistical Association*, **91**, 1343-1370.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion), *Journal of the Royal Statistical Society Series B,* **57**, 99-138.

Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics,* **6**, 461-464.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall: New York.

Tanner, M. and Wong, W. (1987). The Calculation of Posterior Distributions by Data Augmentation, *The Journal of the American Statistical Association,* **82**, 528-540.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions (with discussion), *The Annals of Statistics*, 22, 1701-1762.

Tierney, L. and Kadane, J. (1986). Accurate Approximations for Posterior Moments and Marginal Densities, *Journal of the American Statistical Association*, 81, 82-86.

Wong, W. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLE's. *The Annals of Statistics*, 23, 339-362.

Zhang, P. (1993). On the convergence rate of model selection criteria, *Communications in Statistics*, **22**, 2765-2775.