ISSN: 0976-3031

*Available Online at http://www.recentscientific.com*

**CODEN: IJRSFP (USA)**

*International Journal of Recent Scientific Research*
*Vol. 9, Issue, 7(B), pp. 27797-27805, July, 2018*

**International Journal of Recent Scientific Research**

**DOI: 10.24327/IJRSR**

**Research Article**

# METHODS OF BAYESIAN MODEL SELECTION

**Mariyappan P[1] and Arumugam P[2]**

[1]Research scholar Manonmaniyam Sunadaranar University, Thirnelveli, Department of statistics, Mahabararhi Engineering College, Vasudevanur
[2]Department of Statistics, Annamalai University, Annamalai Nager, Chidhambaram

**DOI: http://dx.doi.org/10.24327/ijrsr.2018.0907.2334**

**ARTICLE INFO**

***Article History:***
Received 11th April, 2018
Received in revised form 9th May, 2018
Accepted 16th June, 2018
Published online 28th July, 2018

***Key Words:***

Baysian, neighborhood, posterior probability, intrinsic, AIC and BIC

**ABSTRACT**

In this paper, we study the Bayesian model selection methods and we study the akiake information criteria (AIC) again comparison of BIC.



## INTRODUCTION

Much of modern scientific enterprise is concerned with the question of model choice. An experimenter or researcher collects data, often in the form of measurements on many different aspects of the observed units, and wants to study how these variables affect some outcome of interest. Which measures are important to the outcome? Which aren't? Are there interactions between the variables that need to be taken into account?

Statisticians are also naturally involved in the question of model selection, and so it is should come as no surprise that many approaches have been proposed over the years for dealing with this key issue. Both frequentist and Bayesian schools have weighed in on the matter, with methods such as F tests for nested models, AIC, Mallows $C_p$, exhaustive search, stepwise, backward and forward selection procedures, cross-validation, Bayes Factors of various flavors (partial, intrinsic, pseudo, posterior), BIC, Bayesian model averaging, to name some of the more popular and well-known methods. Some of these, such as stepwise selection, are algorithms for picking a "good" (or maybe useful) model; others, for example AIC, are criteria for judging the quality of a model.

Given this wealth of choices, how is a statistician to decide what to do? An approach that cannot be implemented or understood by the scientific community will not gain acceptance. This implies that at the very least we need a method that can be carried out easily and yields results that can be interpreted by scientifically and numerically literate end-users. From a statistical point of view, we want a method that is coherent and general enough to handle a wide variety of problems. Among the demands we could mane on our method would be that it obeys the likelihood principle, that it has some frequentist (asymptotic) justification, and that it corresponds to a Bayesian decision problem. Naturally, not all of these desiderata can be met at once, and this paper will do little to influence the ongoing discussion of their relative importance. An attempt to bring coherence to the field from a decision-theoretic perspective was given by Ney, Pericchi and Smith (1999). For an entertaining and readable loon at the subject of Bayesian model selection from the scientist's perspective, we recommend the article by Mackay (1992). We aim to give a more general overview.

### *Why Model Selection?*

Before getting into a review of methods of how to choose a model, it is important to address the question of "why?" At

**Corresponding author:* **Mariyappan P**
Research scholar Manonmaniyam Sunadaranar University, Thirnelveli, Department of statistics, Mahabararhi Engineering College, Vasudevanur

heart we think that the reasons are pragmatic, having to do with saving computer time and analyst attention. Viewed this way, however, there is no particular reason to choose a single best model according to some criterion. Rather it makes more sense to "deselect" models that are obviously poor, maintaining a subset for further consideration. Sometimes this subset might consist of a single model, but sometimes perhaps not. Furthermore, if it is indeed the case that model choice is driven by consideration of costs, perhaps these can be included explicitly into the process via utility functions, as suggested by Winkler (1999). Hence we think there are good reasons to challenge the traditional formulations of this problem.

***A Conceptual Frame worn***

Consider the following general setting. Suppose there are N models, indexed by n, with prior probabilities $\pi_n$, parameters $\theta_n \in \Omega_n$, likelihood $f(x|\theta_n)$ and priors $g_n(\theta_n)$ for n = 1 … N. We are in the M – closed framework of Bernardo and Smith (1994), that is, we assume that one of the N models is the "truth" (or, at least, a reasonable enough approximation thereof that we would be willing to use it in practice). This in itself is a somewhat controversial outlook, positing as it does not only that a true model exists, but that it is one of those under consideration. However, it is a helpful stance for at least thikning through the ramifications of a true Bayesian model selection procedure and the qualities we would wish to demand of it. (see also Petrone,1997; Piccinato, 1997). The posterior on the model M = n and $\theta_n$ is proportional to $f_n(x|\theta_n)g_n(\theta_n)\pi_n$, and the posterior probability of M = n is

$$P(M_n|x) \propto \pi_n \int_{\Omega_n} f_n(x|\theta_n)g_n(\theta_n)d\theta_n \quad \text{... (1)}$$

$$= \frac{\pi_n \int_{\Omega_n} f_n(x|\theta_n)g_n(\theta_n)d\theta_n}{\sum_{j=1}^{K} \pi_j \int_{\Omega_j} f_j(x|\theta_j)g_j(\theta_j)d\theta_j} \quad \text{… (2)}$$

In a full Bayesian analysis, the priors $\pi_n$ on each model and $g_n(\theta_n)$ on the parameters of model n are proper and subjective. Another important element of the full Bayesian paradigm is the utility, or loss, function. The first question to ask is what the contem-plated decision space is, among what set of decisions is the choice to be made? As discussed in Section 2, the traditional decision space for model choice is to choose one of the N models, but we suggest there that it might be more faithful to most applied problems to consider choosing a subset of $\{1, \dots, N\}$ instead.

`In addition to the space of decisions, utility functions also depend, in general, on the parameter space, which here consists in full generality of an indicator of a model, and all the $\theta_s$. Many of the methods to be considered have utilities that depend only on $\theta_n$ if model n is under consideration; some do not depend on $\theta$ at all. Finally, a full specification includes the functional form of the utility function. For a method to be useful, that utility function should represent how a statistician thinns about the model choice she confronts. This idea is developed to some extent by key, Pericchi and Smith(1999), for the so-called M-open perspective, in which it is desired to evaluate a set of models, none of which is believed to be true. Their approach, as mentioned previously, is decision-theoretic, taking explicit account of the utilities involved. On the other hand, they use only improper, "objective" priors, in their analyses and as such deviate from a purely Bayesian procedure (as pointed out by Bayarri, 1999).

The Bayesian proposal is then to make the decision that maximizes expected utility, where the expectation is taken with respect to the posterior distribution of M and θ. It is from this perspective that we wish to evaluate the various schemes and criteria for model selection. In particular, one question of interest is how close do the different methods come to this framework. In a similar vein, insofar as some of the techniques are approximations, how close are these approximations to a coherent Bayesian model selection?

Variations on this perspective are possible, even from the Bayesian point of view. While some practitioners, such as Raftery, Madigan and Hoeting (1997) emphasize posterior distributions, others focus instead on predictive distributions, as in Box (1980); Gelfand and Dey (1994); Laud and Ibrahim (1995).

***Bayesian Model Selections***

***Bayes Factors – Variations on a Theme***

Returning to the conceptual framework from Section 3, recall equation (2) for the posterior probability of model $M_n$; the posterior odds for model $M_n$ is therefore

$$\text{Odds}(M_n|x) = \frac{P(M_n|x)}{1-P(M_n|x)} \quad \dots \quad \text{… (3)}$$

$$= \frac{\pi_n \int_{\Omega_n} f_n(x|\theta_n)g_n(\theta_n)d\theta_n}{\sum_{j\neq n} \pi_j \int_{\Omega_j} f_j(x|\theta_j)g_j(\theta_j)d\theta_j} \quad \text{… (4)}$$

In particular, when N = 2

$$\text{Odds}(M_1|x) = \left(\frac{\pi_1}{\pi_2}\right)\left(\frac{\int_{\Omega_1} f_1(x|\theta_1)g_1(\theta_1)d\theta_1}{\int_{\Omega_2} f_2(x|\theta_2)g_2(\theta_2)d\theta_2}\right) \quad \text{… (5)}$$

The first factor is the prior odds for model 1; the second is called the Bayes Factor, written $B_{1,2}$. The Bayes Factor has been the subject of much discussion in the literature in recent years; see the review by nass and Raftery (1995) and the references therein, for a summary of the issues, although it should be noted that even within the last five years, there have been new developments in the area.

Despite its popularity, the Bayes Factor is relevant only in limited circumstances. Namely, the statistician (or scientist) is required to choose one particular model out of the two available and there must be a zero-one loss on that decision. The meaning of the second requirement is that if the statistician manes the wrong decision, it doesn't matter how far off the choice is; this is contrary to the way that statisticians think about most problems. Nadane and Dickey (1980) show that Bayes Factors are sufficient if and only if a zero-one loss obtains.

When N > 2, (4) simplifies to

$$\text{Odds}(M_n|x) = \frac{\pi_n}{\sum_{j\neq n} \pi_j B_{j,n}} \quad \text{… (6)}$$

In other words, the odds for the $n^{th}$ model is a function of the Bayes factor of that model with every other model. The prior probabilities $\pi_1, \pi_2, \cdots, \pi_N$ on the models do not come out of the sum. As contrasted with the case of inference, where often in practice the choice of prior is not crucial, for model selection, the prior continues to play a role, even asymptotically.

A similar phenomenon arises also within each model. Take the simple case where N = 2, working with a zero-one loss, and assume that model 1 has no parameters at all.

Then $B_{1,2} = \frac{f_1(x)}{\int f_2(x|\theta_2)g_2(\theta_2)d\theta_2}$, … (7)

which depends importantly on the prior over the alternative space, $g_2(\theta_2)$. An example is instructive. Consider the simple case where the first model for the data is normal, with mean 0 and variance 1, and the second model is normal, with mean $\theta$ and variance 1. Suppose that the mean of the data is 0.3. Priors on $\theta$ are proper and normal. Depending on where the prior for $\theta$ is centered, the Bayes factor might lead us to change our opinion about which model should be favored. In other words, the decision we mane will be heavily influenced by the prior, even for a large sample. The Bayes factor is not robust to the specification of prior, even when the prior is proper. If the prior $g_2(\theta_2)$, is allowed to be improper, it can be made to fit the data arbitrarily poorly, making model 2 unlikely no matter what the data turn out to be. This is the Jeffreys-Lindley paradox (Jeffreys, 1961; Good, 1950; Lindley, 1957; Shafer, 1982, among others). As a response to this paradox, Jeffreys proposed a Cauchy form for $g_2(\theta_2)$, with equal prior probability on both models, and a normal likelihood.

Phenomena such as the Jeffreys-Lindley paradox, the dependence of the Bayes factor on the specified priors and the difficulties of calculating and interpreting the Bayes factor at all when improper priors are put on the parameters of the models, have led some authors to seen automatic Bayesian methods for model selection. According to Berger and Pericchi (1996), who advocate this position, automatic methods are essential because the statistician will often, at least initially, consider a wide range of models, for which it won't usually be feasible to specify all priors subjectively (on this point, see also Laud and Ibrahim, 1995). On the other hand, as Lindley (1997) argues, impropriety (and "objective" priors, such as so-called "reference" and "noninformative" priors are often improper) rarely occurs in practice. In this perspective, with which we agree, a parameter is more than just an abstract mathematical construct; instead, it corresponds (at least we hope it does!) to something real, and, if the statistician were to think about the reality underlying the parameter, she should always be able to describe it reasonably well using a proper distribution. As Lindley (1997) phrases it, "It is unfortunately all too easy to slap on an improper prior and avoid having to think about drugs or yields….the problem [with improprieties] is not mathematical at all. It lies in the reality that is conveniently forgotten. Improper distributions in model choice have no sensible interpretation."

No doubt the controversy will continue. Both the objective and the subjective schools of prior specification are a part of the statistical landscape and their proponents will continue to develop methodologies for the critical activity of model selection. Many proposals have been made from the advocates of objective or noninformative priors, as a way of avoiding the difficulties associated with the dependence of Bayes factors on the priors in general, and with vague priors in particular. Berger and Pericchi (1996), for example, define the intrinstic Bayes factor. Divide the data into two parts, a training sample and a testing sample. On the training set, convert the (improper) prior distributions to proper posterior distributions. Compute the Bayes factor using the testing data, and the posterior distributions from the training set as the new priors. Letting $x(l)$ denote a minimal training set, and $x(-l)$ the rest of the sample, a Bayes factor can be defined as

$$B_{ij}(l) = \frac{m_i(x(-l)|x(l))}{m_j(x(-l)|x(l))} \quad \dots (8)$$

where $m_n(x(-l)|x(l))$ is the marginal density of the remainder of the sample, using the prior calculated from the training set. An important point is that the training set cannot increase with the sample size; rather, a minimal training sample needs to be found. For a given data set, there will be many minimal training samples (made up of different combinations of the data points); the intrinsic Bayes factor can be calculated for each one, and then an average of these, either arithmetic or geometric, is taken, yielding the arithmetic intrinsic and geometric intrinsic Bayes factor, respectively. Further modifications of these Bayes factors, such as the trimmed and median variants, are possible; see Berger and Pericchi (1996). A version of the geometric intrinsic Bayes factor is an approximate Bayesian solution to the well-posed decision problem, from within the M-open perspective, of selecting a model, on the basis of which a terminal action will be taken (predicting a single future observation), with a particular utility attached (Ney, Pericchi and Smith, 1999).

What is intrinsic about the intrinsic Bayes factor? Berger and Pericchi (1996) give the following motivation. Suppose we have data $X_i$ which are iid $N(\mu, \sigma^2)$ under the model $M_2$, whereas under $M_1$ they are $N(0, \sigma^2)$. Possible noninformative priors for the two models are $1/\sigma^2$ for $M_2$ (the Jeffreys prior) and $1/\sigma$ for $M_1$ (this is the standard noninformative prior for the normal problem). Minimal training sets are any two distinct observations. Jeffreys (1961) proposed using the standard noninformative prior for the variance, but argued for the use of a Cauchy $(0\ ,\ \sigma^2)$ conditional prior for $\mu$ given $\sigma^2$ for $M_2$. The intrinsic Bayes factor analysis gives results that are very similar to those obtained using the Cauchy prior in $M_2$. In general, the argument is that intrinsic Bayes factors reproduce Bayes factors based on "sensible" noninformative priors. However, since we question whether noninformative priors can ever really be sensible, we are still left with the question "What is intrinsic about intrinsic Bayes factors?"

If the data set is large, there will be many minimal training sets over which to average, making the Berger and Pericchi approach rather cumbersome. An alternative is suggested by O'Hagan (1995) in the form of the fractional Bayes factor. Let m denote the size of the training sample, n the size of the entire data set, and $b = m/n$. For large m and n, the likelihood based on the training set only will approximate the likelihood based on all of the data, raised to the $b^{th}$ power. Define

$$B_b(x) = m_1(b\ ,x)/m_2(b\ ,x), \quad \dots (9)$$

where

$$m_i(b,x) = \frac{\int g_i(\theta_i)f_i(x|\theta_i)d\theta_i}{\int g_i(\theta_i)f_i(x|\theta_i)^b d\theta_i}. \quad \dots (10)$$

$B_b(x)$ is the fractional Bayes factor. Note that the motivation for the fractional Bayes factor is asymptotic (in m and n ), although O'Hagan proposes it more generally for all sizes of data set.

Fractional Bayes factors have several desirable properties in common with ordinary Bayes factors that are not, however, shared by intrinsic Bayes factor (O'Hagan, 1997). The fractional bayes factor satisfies the likelihood principle, whereas intrinsic bayes factors don't. Invariance to transformations of the data is another property of fractional bayes factors which is not always enjoyed by the intrinsic version. When the two models being compared aren't nested, the arithmetic intrinsic bayes factor is not well-defined, because the researcher needs to determine which model is complex. Using an encompassing model, in which both candidates are nested, doesn't always solve the problem. O'Hagan further shows that there can be difficulties with the minimal training sample for some problems the minimal training sample requires the use of all or most of the data, in which case the intrinsic bayes factor cannot discriminate between models.

In response to the critique by O'Hagan (1997) and another, along similar lines,by Bertolino and Racugno (1997), Berger and Pericchi (1998) advocate the use of the median intrinsic bayes factor, which, they claim, may not be optimal for all situations, but is "a good IBF in virtually any situation,…."(Berger and Pericchi,1998,). There are two version of the median intrinsic bayes factor. The first is the median over training samples (instead of an arithmetic or geometric mean,take a median),that is

$$B_{ij}^{M} = med\left(B_{ij}(l)\right), \quad \dots (11)$$

with $B_{ij}(l)$ defined as above. The second is a ratio of medians,

$$B_{ij}^{RM} = \frac{med[m_i(x(-l)|x(l)]}{med[m_j(x(-l)|x(l)]}. \quad \dots (12)$$

Note that $B_{ij}^{RM}$ doesn't have to correspond to a Bayes factor arising from one of the training samples (the sample which gives the median value in the numerator might not be the same as the sample which yields the median value in the denominator). Berger and Pericchi argue that $B_{ij}^{M}$ and $B_{ij}^{RM}$ satisfy many of the desiderata outlined by O'Hagan (1997) and, in addition, are stable in a variety of situations where the arithmetic intrinsic Bayes factor fails.

Taking the general idea of splitting the data into a training set and a testing set to an extreme, Aitkin (1991) defines the posterior Bayes factor, by replacing the prior distribution $g_i(\theta_i)$ with the posterior distribution $g_i(\theta_i|x)$ in the definition of the Bayes factor. In effect, this compares the posterior means under the two models and uses the entire data set as the training sample. This method is open to a number of criticisms, not the least of which is using the data twice, once to compute the posterior (to be used as a prior) and once to calculate the Bayes factor. Furthermore, as pointed out by Lindley (1991) in his discussion, use of the posterior Bayes factor can lead to paradoxes in inference. The method does not correspond to any sensible prior, nor is it a coherent Bayesian procedure Goldstein(1991);O'Hagan(1991).

Consideration of Bayes factors also leads to two of the more common criteria used for model selection-the Bayes Information Criterion (or BIC) and the Akaike Information Criterion (or AIC). The Schwarz Criterion is defined as

$$S = \log f_1(x|\theta_1) - \log f_2(x|\theta_2) - \frac{1}{2}(d_1 - d_2)\log(n), \quad \dots (13)$$

Where $\hat{\theta}_n$ is the maximum likelihood estimates under model n, $d_n$ is the dimension of $\theta_n$ and n is the sample size (Schwarz,1978). Minus two times this quantity is the BIC. Asymptotically, as the sample size increases,

$$\frac{S - \log B_{12}}{\log B_{12}} \to 0, \quad \dots (14)$$

thus the Schwarz criterion gives a rough approximation to the logarithm of the Bayes factor, without having to specify the priors $g_n(\theta_n)$ (Nass and Raftery,1995). however, even for very large samples exp(S) is not equal to $B_{12}$, as the relative error tends to be of order O(1). that is, the approximation does not achieve the correct value of the Bayes factor. Nass and Raftery (1995) note, though, that the Schwarz criterion should, for large samples, give an indication of the evidence for or against a model.

The AIC is given by AIC= -2(log maximized likelihood) +2(number of parameters); as a model selection criterion, the researcher should choose the model that minimizes AIC (Akaike, 1973). One justification for the AIC is Bayesian (Akaike, 1983), namely, that asymptotically, comparisons based on Bayes factors and on AIC are equivalent, if the precision of the prior is comparable to the precision of the likelihood. This requirement that the prior change with the sample size is unusual asymptotic, and furthermore is usually not the case. Rather, the data tend to provide more information than the prior. In this situation, the model which minimizes

BIC= -2(log maximized likelihood) + (log n) (number of parameters) has the highest posterior probability. As can be seen by comparing the expressions for AIC and BIC, these two criteria differ only by the coefficient multiplying the number of parameters, in other words, by how strongly they penalize large models. In general, models chosen by BIC will be more parsimonious than those chosen by AIC. The latter has been shown to overestimates the number of parameters in a model (see, for example, Gewene and Meese, 1981; Natz, 1981; Noehler and Murphree, 1988). It's also worth pointing out that, even though AIC has a Bayesian justification, nowhere does a prior appear in the expression for the criterion itself.

Smith and Spiegelhalter (1980) study the relation between the ordinary Bayes factor and selection criteria such as AIC and BIC in the setting of nested regression models.

Denote by $\beta_2$ the vector of regression coefficients unique to the encompassing model, that is, the parameters which are in the larger model, but not in the smaller model. The choice of prior on $\beta_2$ is crucial in the form of the Bayes factor. Letting the matrix of additional (assumed orthogonal) columns in the encompassing model be $X_2$, Smith and Spiegelhalter consider priors on $\beta_2$, given the error variance $\sigma^2$, that have covariance matrix of the form $\sigma^2\rho(n)(X_2^tX_2)^{-1}$. Minus twice the logarithm of the approximate Bayes factor obtained from priors of this sort is the type

$$\Lambda(m) = \lambda - m(d_2 - d_1) \quad \dots (15)$$

where $m = \frac{3}{2} + \log\rho(n)$, $\lambda$ is the likelihood ratio test statistic and $d_2 - d_1$ is the dimension of $\beta_2$. Taking ρ(n) to be $e^{1/2}$ leads to AIC, and other value could just as easily be chosen. As ρ(n)

increases, support for the simpler model also rises. When the elements of $X_2^t X_2$ are of order n for large n, the choice ρ(n) = n corresponds to taking a fixed prior, with variance that does not shrink with n. under this settings, we get BIC ,since $m \approx \log(n)$. AIC and BIC represent the extremes of taking ρ(n) to be constant (in n) and taking ρ(n) = n. looking at the criteria in this way, it is obvious that other choices for ρ(n), which would impose different penalties on the larger model, are possible and perhaps desirable.

***Bayesian Model Averaging***

When working with Bayes factors, the decision space involves the choice of a model, or possibly several models, which are then used for inference or prediction. If the chosen model is only one of many possibilities, the statistician runs the risk that model uncertainty will be ignored (Draper, 1995). In this light, it makes sense to look at the panoply of models and the inferences or predictions they would give. A formal Bayesian solution to this problem, as outlined in the conceptual framework posed in the opening sections, was proposed by Leamer (1978). Suppose there is a quantity of interest, denoted Δ; the posterior distribution of this quantity, given the data is

$$P(\Delta|x) = \sum_{n=1}^{K} P(\Delta|M_{n,x}) \, P(M_n|x). \quad \ldots (16)$$

This is a weighted average of the posterior probabilities of Δ under each model, where the weights are given by the posterior probabilities of the models in question. Raftery, Madigan and Hoeting (1997) call this approach Bayesian model averaging (Draper, 1995, does not use this specific terminology, but advocates the same idea). As pointed out by those authors, averaging over all models increases predictive ability, compared to basing conclusions about Δ on any of the single models under consideration; however, the process itself can be very difficult, since it often involves integrals that are hard to evaluate, and the number of terms in the sum (that is, the number of models, N) may be too large to be easily handled.

The latter problem can be tackled by using the Occam,s window algorithm for Bayesian model averaging (Madigan and Raftery, 1994). Based on two common-sense principles of model selection, namely (1) that if a model predicts the data much worse than the best model, it should be dropped from further consideration and (2) that models that predict the data less well than any of their nested submodels should be discarded, this algorithm often drastically reduces the number of models that need to be considered in the average. Now, the problem is one of finding the class of models to be included in the average. Occam's window compares at each step two models, where one model, call it $M_0$, is a submodel of the other, $M_1$. Loon at the logarithm of the posterior odds for $M_0$; if this is positive (or, in general, greater than some set constant), that is, the data give evidence in favor of the smaller model, reject $M_1$; if it is negative but small, consider both models, since there isn't enough evidence one way or another; if it is negative and large, then reject $M_0$ from further consideration. If $M_0$ is rejected, so are all of its submodels. Using either an "up" or a "down" procedure to move around the space of all possible models, models are eliminated, until the set of potentially acceptable models to go into the averaging is found.

MCMC model composition (Madigan and York, 1995) is another approach for evaluating $P(\Delta|x)$. A Markov chain is built on the model space, with stationary distribution $P(M_i|x)$, and steps through it are taken by moving in a small neighborhood of the current model. More specifically, the neighborhood of a model consists of all those models with one variable more or one variable less than the one under consideration at a given stage of the chain. Transition probabilities are defined such that the probability of moving to a model outside of the neighborhood is zero, and the probability of moving to a model within the neighborhood is the same for all models in the neighborhood. If the chain is currently at stage $M_n$, then we need to draw a model $M_n$ from the neighborhood.

The model averaging method described by Raftery, Madigan and Hoeting (1997) uses flat priors over the range of "plausible" values of the parameters. Further, for some of the parameters the priors are data dependent, involving both the dependent and the independent variables from a linear regression model. In that sense, their approach is only an approximation to the fully Bayesian analysis that would be achieved by the use of subjective priors. Elicitation of expert opinion (see, for example, Nadane, Dickey, Winkler, Smith and Peters, 1980; Garthwaite and Dicney, 1992; Nadane and Wolfson, 1998) is a feasible way of obtaining proper, subjective priors to incorporate into the model averaging procedure. As shown by Ney, Pericchi and Smith (1999), model averaging is also a solution to a well-posed Bayesian decision problem from the M-closed perspective, specifically, that in which a terminal decision is made directly (for instance, predicting a new observation).

Although our focus is not on computation, it is worth noting that several other schemes have been developed for the calculation of posterior probabilities over model spaces of varying dimension. In particular, the reversible jump approach (Green, 1995; Richardson and Green, 1997) has been gaining popularity in Bayesian circles in recent years. Chib (1995) proposes an alternative method, which is based on the computation of marginal likelihoods, and hence allows the computation of Bayes factors as well. See also Carlin and Chib (1995) and Carlin and Polson (1991).

For the regression problem, Mitchell and Beauchamp (1988) propose a Bayesian approach to variable selection. They place "spine and slap" priors on each of the coefficients in the regression equation, i.e. a point mass on $\beta_j = 0$ for each j, with the rest of the prior probability spread uniformly over some defined (and large) range. In a similar vein, George and McCulloch (1993) describe a Gibbs sampling technique for "stochastic search variable selection" in regression, which selects promising subsets of variables. George and McCulloch suggest embedding the problem in a hierarchical Bayes normal mixture model, with latent variables to identify subsets. Models with high posterior probabilities are picked out for additional study by the procedure. The prior on $\beta_j$ is a two-component normal mixture, with each component centered about zero, and having different variance. A latent variable determines to which component $\beta_j$ belongs. In contrast to Mitchell and Beauchamp's prior, no point mass is placed on zero. Denoting the latent parameter by $\gamma_i$, the prior is

$$\beta_j|\gamma_j \sim (1-\gamma_j)N(0,\tau_j^2) + \gamma_j N(0, c_j^2\tau_j^2). \quad \ldots (17)$$

The latent variable is equal to 1 with probability $p_j$. In this formulation, the statistician needs to devote some thought to the values of $\tau_j$ and $c_j$. The former should be small, so that if $\gamma_j = 0$, $\beta_j$ is small and might be closely estimated by zero. On the other hand, $c_j$ should be large. Thus if $\gamma_j = 1$, a non-zero estimate of $\beta_j$ would lead to including this variable in a model. Under this interpretation, $p_j$ can be thought of as the prior probability that variable j should be in the model.

Building on the worn of George and McCulloch (1993), Kuo and Mallick (1998) also explore the use of Markov Chain Monte Carlo to identify models with high posterior probability. Where the former build a hierarchical model, Kuo and Mallick start from a regression equation that embeds all models within it. Taking $\gamma_j$ to be the indicator for the $j^{th}$ variable being in the model, the regression for subject $i$ is written as

$$y_i = \sum_{j=1}^{p} \beta_j \gamma_j x_{ij} + \epsilon_i. \quad \ldots (18)$$

When $\gamma_j = 1$, predictor j is included in the model and when $\gamma_j = 0$, we omit predictor j. Standard priors are assumed on the parameters – normal for the vector of coefficients, inverse gamma for the variance of the errors, and the $\gamma_j$ are independent Bernoullis. Note that in this formulation, the prior on $\beta_j \gamma_j$ is a mixture – it has a point mass at 0 with a certain probability, and the rest of the mass is normally distributed. Instead of a "spine and slap" prior, we have a "spine and bell". Therefore, as in Mitchell and Beauchamp (1998), a privileged position is given to the particular hypothesis that $\beta_j = 0$. The posterior distribution of the vector of indicators is supported on each of the $2^p$ submodels, and gives a measure of the probability of each. In this way, it is possible to evaluate the models and consider the ones with highest posterior probability. The model with the highest posterior probability corresponds to a Bayes decision rule with zero-one loss (see also discussion of Bayes factors). Calculation of the posterior distributions is via Gibbs sampling.

***Predictive Methods***

The framework proposed in Section 4.2 looks at the posterior probability assigned to each model. Alternatively, it should be possible to look at the predictions from the various models. Now the question of interest shifts slightly, from "Which models best explain the observed data?" to "Which models give the best predictions of future observations generated from the same process as the original data?" Ideally, we would like to compare predictions and choose the model which gives the best overall predictions of future values. However, we don't know these "future values"- if we did, we could just use them directly. Most predictive methods, then, use some sort of jacknkife approach, under the assumption that future observations from the process that generated the data would be similar to those actually in the sample. That is, the data are assumed to be exchangeable. This is the idea behind the "quasi-Bayes" approach of Geisser and Eddy (1979), a blend of Bayesian and sample-reuse ideas. For each model, compute the likelihood as the product of "predicting densities", that is, the density of the $j^{th}$ observation, calculated on the rest of the data with the $j^{th}$ observation deleted, under a specific model (this gives a predicted value for observation j based on the rest of the data). The model for which this likelihood is maximized is chosen as the most suitable of those models being considered.

San Martini and Spezzaferri (1984) give a different twist on the predictive approach to model selection, defining their criterion in terms of utility. Here, priors on the models and the parameters are incorporated. They define an average criterion, which, like those of Akaike and Schwarz, corrects the likelihood ratio statistic by taking account of the differences in model dimension. It differs from other similar criteria in that it also accounts for the distance between two models. Assume that the models under consideration are $M_1, \ldots, M_n$, $p_n$ is the probability that model $M_n$ is true and $p_n(y)$ is the predictive density of a future observation y based on the model $M_n$. Now let $u(p(*), y)$ be a utility function for choosing the density $p(*)$ as the predictive distribution of y (the unknown future observation). The procedure picks the model whose expected utility is the largest. If there are two models, for example, the first will be chosen if

$$E_1[u(p_1(*), y) - u(p_2(*), y)]p_1 > E_2[u(p_2(*), y) - u(p_1(*), y)]p_2,$$

expectations $E_i$ being taken with respect to the predictive distribution $p_i(*)$. In addition, San Martini and Spezzaferri (1984) show that their criterion fits into the frameworn of Smith and Spiegelhalter (1980), with a penalty term that increases as the distance between the two models (as measured by the linelihood ratio statistic) increases.

A predictive version of a general Bayesian model selection framework is given in Gelfand and Dey (1994). Observed (independent) data are $x_1, \ldots, x_n$, which under model $M_n$ have likelihood $f(x|\theta_i)$. For simplicity, Gelfand and Dey restrict attention to the case where only two models are being considered; as they point out, comparisons are generally done pairwise, so nothing is lost by this. Denote by $S_n$ the index set {1,2,…,n} and let S be a subset of $S_n$. Define

$$L(\theta_n|x_s) = \prod_{i=1}^{n} f(x_i|\theta_n)_i^d, \quad \ldots (19)$$

where $d_i$ is the indicator for $i \in S$. As before, we denote the prior for $\theta_n$ under model $M_n$ by $g_n(\theta_n)$. For prediction purposes, Gelfand and Dey propose consideration of the conditional density

$$f(x_{S_1}|x_{S_2}, M_n) = \int L(\theta_n|x_{S_1}) g_n(\theta_n|x_{S_2}) d\theta_n$$
$$= \frac{\int L(\theta_n|x_{S_1}) L(\theta_n|x_{S_2}) g_n(\theta_n) d\theta_n}{\int L(\theta_n|x_{S_2}) g_n(\theta_n) d\theta_n} \quad \ldots (20)$$

This conditional density is a predictive density; it averages the joint density of $x_{S_1}$ with respect to the prior $g_n(\theta_n)$, updated by $x_{S_2}$. Both $S_1$ and $S_2$ are taken to be subsets of S, and different choices correspond to predictive techniques in the Bayesian literature. For instance, $S_1$ ={r} and $S_2$ = S - {r} gives the Geisser and Eddy (1979) cross-validation density and hence the pseudo-Bayes factor

$$\prod_r f(x_r|x_r, M_1) / \prod_r f(x_r|x_r, M_2). \quad \ldots (21)$$

$S_1 = S_2 = S$ results in Aitkin's (1991) posterior predictive density and the posterior Bayes factor. When $S_2$ is a minimal subset and $S_1 = S - S_2$, we can obtain the different versions of the intrinsic Bayes factor.

Gelfand and Ghosh (1998) also adopt a predictive outlook to model selection, building on the observation by Nadane and Dickey (1980) that Bayes factor correspond to a 0 – 1 loss. Other loss functions are possible, and they base their method on the idea of evaluating models by comparing observed data to predictions. For each model, minimize the expected posterior loss over all possible predictions as the observed data; then, choose the model for which this minimum is minimized. Note that in this framework, as opposed to our general outline of the model selection process, there is no notion of one of the models being "true"; furthermore, there are priors assigned to the models themselves.

The goal of this approach is to obtain good predictions for replicates of the observed data, but at the same time to be faithful to the observed values. In order to attain this objective, a loss of the general form

$$L(y_{rep}, a; y_{obs}) = L(y_{rep}, a) + nL(y_{obs}, a) \quad \dots (22)$$

for $n \geq 0$ is proposed, where $y_{obs}$ are the observed data, $y_{rep}$ are the replicates to be predicted (assumed to come from the same distribution as the observed data) and $a$ is the "action" or estimate. The action is a compromise between the observation and the prediction, with the weight, n, expressing how important it is to be close to $y_{obs}$, relative to $y_{rep}$. Gelfand and Ghosh show that for a range of models and appropriate choices of the loss $L(y, a)$, the form above results (asymptotically or approximately) in a goodness of fit term plus a penalty term, similar to criteria such as AIC and BIC.

Let's consider a simple example in more detail; this example is given in Gelfand and Ghosh (1998) and we repeat it here to highlight the essentials of the method, which is somewhat different in spirit than others we have considered so far. Take

$$D_n(m) = \sum_{i=1}^{n} \min_{a_l} E_{y_{l,res} \setminus y_{obs,m}} L(y_{l,rep}, a_l; y_{obs}); \quad \dots (23)$$

m represents the model relative to which calculations are carried out. For the general form of the loss described above, this becomes

$$D_n(m) = \sum_{i=1}^{n} \min_{a_l} \{E_{yl,res \setminus y_{obs,m}} L(y_{l,rep}, a_l) + nL(y_{l,obs}, a_l)\}. \quad \dots (24)$$

For a fixed $a_l$, and $L(y, a) = (y - a)^2$, the $l^{th}$ term in this sum is

$$\sigma_l^2 + (a_l - \mu_l)^2 + n(a_l - y_{l,obs})^2, \quad \dots (25)$$

where $\sigma_l^2$ is the variance of $y_{l,rep}$ given $y_{obs}$ and $m$, and $\mu_l$ is the expected value of $y_{l,rep}$ given $y_{obs}$ and $m$; in both of these we have suppressed the dependence on the model in the notation for simplicity.

The minimizing $a_l$ is $(n + 1)^{-1}(\mu_1 + ny_{l,obs})$. If this is inserted bacn into the expression for $D_n(m)$, the result is

$$D_n(m) = \frac{n}{n+1} \sum_{l=1}^{n} (\mu_1 - y_{l,obs})^2 + \sum_{l=1}^{n} \sigma_l^2. \quad \dots (26)$$

The first summand can be thought of as a goodness-of-fit measure (how close are the predictions to the observed data) and the second is a type of penalty term. If $y_l$ comes from a normal distribution, the first term is equivalent to the likelihood ratio statistic with $\mu_l$ replacing the MLE of the mean of $y_l$. Extending the example, suppose that y comes from a normal linear model. Put as a prior on the parameters $\beta$ a $N(\mu_b, \Sigma)$ distribution. If the prior is very imprecise, that is Σ is large, then $y_{rep} \setminus y_{obs}$ has an approximate $N(X\hat{\beta}, \sigma^2[I + X(X^TX)^{-1}X^T])$ distribution. The two summands in $D_n(m)$ become (again, approximately) $(y - X\hat{\beta})^T(y - X\hat{\beta})$ and $\sigma^2(n + p)$.

As pointed out in Gelfand and Ghosh (1998), this is one example where the calculation of $D_n(m)$ can be explicitly made. In general, however, a combination of asymptotic expansions and Monte Carlo simulation for the evaluation of integrals will need to be employed.

***Akaike Information Criterion (AIC)***

The Akaike information criterion (AIC) is a measure of the relative quality of a statistical model, for a given set o data. As such, AIC provides a means for model selection. AIC deals with the trade-off between the goodness of fit of the model And the complexity of the model .It is founded on information entropy: AIC does not provide a test of a model in the sense of testing a null hypothesis; i.e. AIC can tell nothing about the quality of the model in an absolute sense .If all the candidate models fit poorly, AIC will not give any warning of that.

In the general case, the AIC is

$$AIC = 2n - 2\ln(L) \quad \dots (27)$$

Where n is the number of parameters in the statistical model and L is the maximized value of the likelihood function for the estimated model. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Hence AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. This penalty discourages over fitting (increasing the number of free parameters in the model improves the goodness of the fit, regardless of the number of free parameters in the data-generating process).

**AICc**

AICc is AIC with a correction for finite sample sizes:

$$AICc = AIC + \frac{2n(n+1)}{n-n-1} \quad \dots (28)$$

Where n denotes the sample size. Thus, AICc is AIC with a greater penalty for extra parameters. Burnham & Anderson (2002) strongly recommend using AICc, rather than AIC, if n is small or n is large. Since AICc converges to AIC as n gets large, AICc generally should be employed regardless. Using AIC, instead of AICc, When n is not many times larger than $n^2$, increases the probability of selecting models that have too many parameters, i.e. of over fitting. The probability of AIC over fitting can be substantial, in some cases. Brockwell & Davis (1991) advise using AICc as the primary criterion in selecting the orders of an ARMA model for time series. McQuarrie & Tsai (1998) ground their high opinion of AICc on extensive simulation worn with regression and time series.

AICc was first proposed by Hurvich & Tsai (1989). Different derivations of it are given by Brockwell & Davis (1991). Burnham & Anderson, and Canvanaugh (1997). All the derivations assume a univariate linear model with normally distributed errors (conditional upon regressors); if that assumption does not hold, then the formula for AICc will usually change. Further discussion of this, with examples of

other assumptions, is given by Burnham & Anderson (2002). In particular, bootstrap estimation is usually feasible. Note that when all the models in the candidate set have the same $n$, then AICc and AIC will give identical (relative) valuations. In that situation, then, AIC can always be used.

***History of Akaike Information Criterion***

The Akaike information criterion was developed by Hirotugu Akaike , under the name of " an information criterion". It was first published by Akaike in 1974. The original derivation of AIC relied upon some strong assumptions. Takeuchi (1976) showed that the assumptions could be made much weaker. This worn, however, was in Japanese, and was not widely known outside Japan for many years.

AICc was originally proposed for linear regression (only) by Sugiura (1978).That instgated the worn of Hurvich &Tsai (1989), and several further papers by the same authors, which extended the situations in which AICc could be applied. The worn of Hurvich &Tsai contributed to the decision to publish a second edition of the volume by Brockwell & Davis (1991) which is the standard reference for linear time series; the new edition states, "Our prime criterion for model selection [among ARMA (p, q) models] will be the AICc. The volume by Burnham & Anderson (2002) was the first attempt to set out the information –theoretic approach in a general context. It includes an English exposition of the results of Takeuchi. The volume led to far greater use of the information-theoretic approach, and now has over 20000 citations on Google Scholar.

Akaike originally called his approach an "entropy maximizations principal". Burnham & Anderson (2002) discuss and expand on this, and trace the approach back to the worn of Ludwig Boltzmann on thermodynamics. Briefly, minimizing AIC in a statistical model is essentially equivalent to maximizing entropy in a thermodynamic system. In other words, the information –theoretic approach in statistics is essentially applying the second Law of thermodynamics.

***Comparison with BIC***

The AIC penalizes the number of parameters less strongly than does the Bayesian information criterion (BIC). A comparison of AIC/AICc and BIC is given by Burnham & Anderson (2002). The authors show that AIC and AICc can be derived in the same Bayesian framework as BIC, just by using a different prior. The authors also argue that AIC/AICc has theoretical advantages over BIC. First, because AIC/AICc is derived from principles of information; BIC is not, despite its name. Second, because the (Bayesian framework) derivation of BIC has a prior of $1/R$(where R is the number of candidate model), which is "not sensible", since the prior should be a decreasing function of n. Additionally , they present a few simulation studies that suggest AICc tends to have practical/performance advantages over BIC. See to Burnham & Anderson (2004).Further comparison of AIC and BIC, in the context of regression, is given by Yang (2005). In particular, AIC is asymptotically optimal in selecting the model with the least mean squared error, under the assumption that the exact "true" model is not in the candidate set (as is virtually always the case in practice);BIC is not asymptotically optimal under the assumption. Yang further shows that the rate at which AIC converges to the optimum is, in certain sense, the best Possible.

***Bayesian information criterion (BIC)***

The Bayesian information criterion was introduced by Schwarz (1978) as a competitor to the Akaike (1973, 1974) information criterion.Schwarz derived BIC to serve as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model.In large-sample settings, the fitted model favored by BIC ideally corresponds to the candidate model which is a posterior most probable; i.e., model which is rendered most plausible by the data at hand.The computation of BIC is based on the empirical log-likelihood and does not require the specification of priors.In Bayesian applications, pair wise comparisons between models are often based on Bayes factors. Assuming two candidate models are regarded as equally probable a prior, a Bayes factor represents the ratio of the posterior probabilities of the models. The model which is a posterior most probable is determined by whether the Bayes factor is less than or greater than one. In certain settings, model selection based on BIC is roughly equivalent to model selection based on Bayes factors (Nass and Raftery ,1995; Nass and Wasserman, 1995)
The Bayesian information criterion is

$$BIC = -2lnf\left(y/\widehat{\theta_n}\right) + n\ln n. \quad \ldots (29)$$

AIC and BIC feature the same goodness-fit term. The penalty term of BIC is more stringent than the penalty term of AIC (For $n \geq 8, n\ln n \ exceeds\ 2n$ ).Consequently; BIC tends to favor smaller models than AIC. BIC provides a large-sample estimator of transformation Of the Bayesian posterior probability associated with the approximating model. BIC provides a large-sample estimator of a transformation of the Bayesian posterior probability associated with the approximating model. By choosing the fitted candidate model corresponding to the minimum value of BIC, one is attempting to select the candidate model corresponding to the highest Bayesian posterior probability.BIC was justified by Schwarz (1978) "for the case of independent, identically distributed observations, and linear models," under the assumption that the likelihood is from the regular exponential family.

An endeavor as basic to the pursuit of science as model choice and selection is bound to generate a plethora of approaches. Bayesian and classical statisticians have both put forth proposals for solving this most difficult, as we have argued, for a researcher to know what is the "proper" way to proceed.

The unifying conceptual framework we proposed is an attempt to bring order to this often chaotic field. From this perspective, a "model" is just a discrete parameter in a larger super-model. Model averaging, with proper priors, provides a principled and coherent Bayesian approach to the problem at hand. Regarding other Bayesian techniques, such as the various flavors of Bayes factors, while they may be solutions to specific decision theoretic problems, as described in Ney, Perrichi and Smith (1999), they are more narrow in focus and in applicability. Indeed, applicability of the "default prior" methods, embodied in intrinsic and fractional Bayes factors, needs to be checked on a case by case basis (Berger and Perrichi, 1997) and in that sense they don't necessarily offer an advantage even over frequentist methods.

Frequentist approaches to model selection of course do not fit neatly into the proposed Bayesian framework, and suffer from

the lack of a guiding principle. New methods are developed apparently on ad hoc grounds. To be fair, many of the so-called objective Bayesian techniques also seem to us to be derived more as a response to something else not working, than from proper Bayesian considerations, and this is perhaps not coincidental.

## Reference


Akaike, Hirotugu (1983). "Information Measures and Model Selection", *Bulletin Of the International Statistical Institute 50: 277- 290.*

*Allen, D.M.(*1974) The relationship between variable selection and prediction. "Technometrics," 16, 125-127.

Berger,J.O and Perichi, L.R. (1998) Accurate and stable Bayesian Model selection: The median Intrinsic Bayes Factor . Sankhya, B, 60, 1-18.

Billio M,Monfort A,Robert CP (1999) ."Bayesian estimation of switching ARMA models. *Journal of Econometrics*, Vo l93, pp 229-255

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *J. Am. Statist. Ass*. 74, 153-160; correction, 75 (1980), 765.

Geweke, J.F. and Meese, R.A. (1981). Estimating regression models of finite but unknown order. *International Economics Review*, 22, 55 – 70.

Han Hong and Bruce Preston (2008). "Bayesian Averaging, prediction and Non-nested Model selection,"

Harvey, A.C. And S.Peters (1990). "Estimation procedures for structural time series Models". *Journal of Forcasting*, vol.9, 89-108.

Kadane, J. B.and Lazar, N.A.(2004), "Methods and criteria for model selection, " *Journal Of the American Statistical Association,* 99, 279-290.

Land, P. W, and Ibrahim, J.G. (1995) Predictive model selection. *Journal of the Royal statistical Socity, Series* B, 57, 247-262.

Martin, R.D.(1980). "Robust Estimation of Autoregressive Models"(with discussion), in Directions in time series, eds.D.R.Brillinger and G.C. Tiao, Haywood, CA: Institute of mathematical Statistical, pp 228-254.

McQuarrie, A. D. R.; Tsai, C.-L. (1998), Regression and Time Series Model Selection, World Scientific, ISBN 981-02-3242-X.

Phillips, P.C.B. (1995a). "Bayesian model selection and prediction with empirical applications*." Journal of Econometrics*, 69: 289 – 331.

Ricardo S.Ehlers and Stephen P.Brooks(2003). Bayesian analysis of order Uncertainty in ARIMA models.

Sally Wood, Ori Rosen and Robert Kohn.(1993). "Bayesian Mixture of Autoregressive Models".



**How to cite this article:**

Mariyappan P and Arumugam P.2018, Methods of Bayesian Model Selection. *Int J Recent Sci Res.* 9(7), pp. 27797-27805.
DOI: http://dx.doi.org/10.24327/ijrsr.2018.0907.2334


*******