Methods and Criteria for Model Selection

Summary

Model selection is an important part of any statistical analysis, and indeed is central to the pursuit of science in general. Many authors have examined this question, from both frequentist and Bayesian perspectives, and many tools for selecting the "best model" have been suggested in the literature. This paper considers the various proposals from a Bayesian decision–theoretic perspective.

Key words: AIC, Bayes Factors, BIC, Mallow's C_p, Model Averaging, Subset Selection, Variable Selection

1 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, fractional, 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 make 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 Bayesian decision–theoretic perspective was given by Key, Pericchi and Smith (1999). For an entertaining and readable look 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 (see also Miller, 2002, for a thorough discussion of variable selection in regression).

2 Why Choose a Model?

Suppose there are K models, indexed by k. Model k has parameters $\theta_k \in \Omega_k$. Then the whole parameter space is $(M, \theta_1, \theta_2, \ldots, \theta_K)$, where M denotes the model. In every statistical model, estimation may be thought of as the choice of a single value of the parameter chosen (according to some criterion) to represent the distribution. Estimation has been sharply criticized, especially by Box and Tiao (1992), because the choice of a single value may be misleading if there are several competing parameter values, distant according to some relevant metric, that are supported in some sense by the data.

Model selection can be thought of in this framework as estimation applied to the parameter M. As such, it is subject to the general criticisms of Box and Tiao. There may be occasions, just as with estimation in general, in which one model so clearly dominates the others that the choice is unobjectionable, and others in which the choice is misleading. Viewed in this light, the only special issue that comes up in model choice is that generally M is discrete, and usually has finite range.

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

3 A Conceptual Framework

Consider the following general setting. Suppose that on the parameter space $(M, \theta_1, \theta_2, \ldots, \theta_K)$ there is a prior π_k on the k^{th} model, and priors $g_k(\theta_k)$ for $k = 1, \ldots, K$. With the assumption that, given M, the priors on $\theta_1, \ldots, \theta_K$ are independent, this implies a prior on $(M, \theta_1, \theta_2, \ldots, \theta_K)$. The likelihood under model k is $f_k(x|\theta_k)$. These assumptions determine the joint distribution of $(X, M, \theta_1, \theta_2, \ldots, \theta_K)$. We are in the *M–closed* framework of Bernardo and Smith (1994), that is, we assume that one of the K 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, since it posits not only that a true model exists, but that the true model is one of those under consideration. However, it is a helpful stance for at least thinking through the ramifications of a 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 $M = k$ and θ_k is proportional to $f_k(x|\theta_k)g_k(\theta_k)\pi_k$, and the posterior probability of $M = k$ is

$$\begin{aligned} P(M_k|x) &\propto \pi_k \int_{\Omega_k} f_k(x|\theta_k)g_k(\theta_k)d\theta_k \\ &= \frac{\pi_k \int_{\Omega_k} f_k(x|\theta_k)g_k(\theta_k)d\theta_k}{\sum_{j=1}^{K} \pi_j \int_{\Omega_j} f_j(x|\theta_j)g_j(\theta_j)d\theta_j} \end{aligned} \qquad (1)$$

In a full Bayesian analysis, the priors π_k on each model and $g_k(\theta_k)$ on the parameters of model k 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 contemplated decision space is, that 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 K models, but we suggest there that it might be more faithful to most applied problems to consider choosing a subset of $\{1, \ldots, K\}$ 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 θs. Many of the methods to be considered have utilities that depend only on θ_k if model k is under consideration; some do not depend on θ 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 thinks 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). Even this, though, is a step forward, since most model selection techniques and criteria do not include utility considerations at all, and, when they do, it is usually (although not always) on the basis of a loss that most practitioners would not believe in, namely, zero–one loss.

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 examine 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. Where possible, we connect techniques back to the general framework. However, not all methods fit easily, or at all – frequentist approaches, for example, typically cannot be evaluated from this point of view, since they lack any formulation of priors or utilities. However, as we point out in later sections, bridges between frequentist and Bayesian procedures do exist, especially in the more recent literature, and in this case it may be possible to evaluate frequentist methods from a Bayesian point of view. 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, such as Box, 1980; Gelfand and Dey, 1994; Laud and Ibrahim, 1995, focus instead on predictive distributions. Finally, Bernardo and Rueda (2002) explore model choice as a problem in "Bayesian hypothesis testing."

4 Bayesian Model Selection

4.1 Bayes Factors – Variations on a Theme

Returning to the conceptual framework from Section 3, recall equation (1) for the posterior probability of model M_k; the posterior odds for model M_k is therefore

$$\begin{aligned} Odds(M_k|x) &= \frac{P(M_k|x)}{1-P(M_k|x)} \\ &= \frac{\pi_k \int_{\Omega_k} f_k(x|\theta_k) g_k(\theta_k) d\theta_k}{\sum_{j \neq k} \pi_j \int_{\Omega_j} f_j(x|\theta_j) g_j(\theta_j) d\theta_j}. \end{aligned} \tag{2}$$

In particular, when $K = 2$,

$$Odds(M_1|x) = \left(\frac{\pi_1}{\pi_2}\right) \left(\frac{\int_{\Omega_1} f_1(x|\theta_1) g_1(\theta_1) d\theta_1}{\int_{\Omega_2} f_2(x|\theta_2) g_2(\theta_2) d\theta_2}\right). \tag{3}$$

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 Kass 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 makes 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. Kadane and Dickey (1980) show that Bayes Factors are sufficient if and only if a zero–one loss obtains. Other losses are available and using them does not have to complicate Bayesian model selection – Lindley (1976) for example proposes conjugate utilities for exponential families, which work in much the same way as conjugate priors. Bernardo and Rueda (2002) consider certain continuous loss functions, which have the advantage of being more natural than step function losses. Clearly, using any such alternative loss or utility leads to criteria for model selection other than the usual Bayes factor. However, Bernardo and Rueda (2002) aim to achieve an "objective" analysis, and hence, deviate from the standard set down in Section 3.

Formula (2) simplifies to

$$Odds(M_k|x) = \frac{\pi_k}{\sum_{j \neq k} \pi_j B_{k,j}}; \tag{4}$$

this is of course equivalent to (3) when $K = 2$. When $K > 2$, the odds for the k^{th} model is a function of the Bayes factor of that model with every other model. The prior probabilities $\pi_1, \pi_2, \ldots, \pi_K$ 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 $K = 2$, working with a zero–one loss, and assume that model 1 has no parameters at all. Then

$$B_{1,2} = \frac{f_1(x)}{\int f_2(x|\theta_2) g_2(\theta_2) d\theta_2}, \tag{5}$$

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 θ and variance 1. Suppose that the mean of the data is 0.3. Priors on θ are proper and normal. Depending on where the prior for θ is centered, the Bayes factor might lead us to change our opinion about which model should be favored. In other words, the decision we make 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 seek 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." (p. 187).

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. These proposals seem to us to be, for the most part, *ad hoc*, in that they are designed to solve particular problems with the ordinary Bayes factor, as opposed to arising from the coherency of the Bayesian approach. Berger and Pericchi (1996), for example, define the *intrinsic 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

$$B_{ij}(l) = \frac{m_i(x(-l)|x(l))}{m_j(x(-l)|x(l))}, \tag{6}$$

where $m_k(x(-l)|x(l))$ 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 (Key, 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 μ given σ^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

$$B_b(x) = m_1(b,x)/m_2(b,x), \tag{7}$$

where

$$m_i(b,x) = \frac{\int g_i(\theta_i) f_i(x|\theta_i) d\theta_i}{\int g_i(\theta_i) f_i(x|\theta_i)^b d\theta_i}. \tag{8}$$

$B_b(x)$ 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 factors (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 more 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, p. 2). There are two versions 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

$$B_{ij}^{M} = med(B_{ij}(l)), \tag{9}$$

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

$$B_{ij}^{RM} = \frac{med[m_i(x(-l)|x(l))]}{med[m_j(x(-l)|x(l))]}. \tag{10}$$

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

$$S = \log f_1(x|\hat{\theta}_1) - \log f_2(x|\hat{\theta}_2) - \frac{1}{2}(d_1 - d_2)\log(n), \tag{11}$$

where $\hat{\theta}_k$ is the maximum likelihood estimator under model k, d_k is the dimension of θ_k and n is the sample size (Schwarz, 1978). Minus two times this quantity is the BIC. Asymptotically, as the sample size increases,

$$\frac{S - \log B_{12}}{\log B_{12}} \to 0,$$

thus the Schwarz criterion gives a rough approximation to the logarithm of the Bayes factor, without having to specify the priors $g_k(\theta_k)$ (Kass 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. Kass 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 asymptotics, 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 overestimate the number of parameters in a model (see, for example, Geweke and Meese, 1981; Katz, 1981; Koehler 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 β_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 β_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 β_2, given the error variance σ^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 of the type

$$\Lambda(m) = \lambda - m(d_2 - d_1), \tag{12}$$

where $m = \frac{3}{2} + log\rho(n)$, λ is the likelihood ratio test statistic and $d_2 - d_1$ is the dimension of β_2. Taking $\rho(n)$ to be $e^{1/2}$ leads to AIC, and other values could just as easily be chosen. As $\rho(n)$ 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 $\rho(n) = n$ corresponds to taking a fixed prior, with variance that does not shrink with n. Under this setting, we get BIC, since $m \approx \log(n)$. AIC and BIC represent the extremes of taking $\rho(n)$ to be constant (in n) and taking $\rho(n) = n$. Looking at the criteria in this way, it is obvious that other choices for $\rho(n)$, which would impose different penalties on the larger model, are possible and perhaps desirable.

The choice of $\rho(n)$ is not a technical matter within this theory, but rather a fundamental issue of the values the statistician/scientist brings to the problem. There is a trade–off between parsimony and accuracy (in a specific sense), in which large values of $\rho(n)$ favor parsimony. Hence attempts to decree an objective, reference, or otherwise arbitrary value for $\rho(n)$ are likely to be unpersuasive, as they are for prior distributions.

4.2 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

$$P(\Delta|x) = \sum_{k=1}^{K} P(\Delta|M_k, x)P(M_k|x). \tag{13}$$

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, K) 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 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. Look 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 state M_k, then we need to draw a model $M_{k'}$ 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. As shown by

Key, 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). Because Bayesian model averaging produces a posterior in the full parameter space $(M, \theta_1, \ldots, \theta_K)$, it can be used in conjunction with any utility function reflecting the decision–maker's values.

4.3 Bayesian Linear Models

Another direction of research tackles the standard variable selection problem from a Bayesian perspective. These methods, like the frequentist ones we will discuss below, aim to find one, or a few, "best" models. They differ from the frequentist techniques in that they incorporate prior information into the analysis, and only approximate the fully Bayesian solution described in our general conceptual framework. For the regression problem, Mitchell and Beauchamp (1988) propose placing "spike and slab" 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, 1997) 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 β_j is a two–component normal mixture, with each component centered about zero, and having different variance. A latent variable determines to which component β_j belongs.

In contrast to Mitchell and Beauchamp's prior, no point mass is placed on zero. Denoting the latent parameter by γ_i, the prior is

$$\beta_j|\gamma_j \sim (1-\gamma_j)N(0,\tau_j^2) + \gamma_j N(0, c_j^2\tau_j^2). \tag{14}$$

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 τ_j and c_j. The former should be small, so that if $\gamma_j = 0$, β_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 β_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 work of George and McCulloch, 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 γ_j to be the indicator for the j^{th} variable being in the model, the regression for subject i is written as

$$y_i = \sum_{j=1}^{p} \beta_j \gamma_j x_{ij} + \epsilon_i. \tag{15}$$

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 γ_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 "spike and slab" prior, we have a "spike and bell." Therefore, as in Mitchell and Beauchamp (1988), 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.

Brown, Vannucci and Fearn (1998) extend some of these ideas to multivariate generalized linear models. Here, the response for an individual is a vector, that is, there is more than one outcome of interest. Let the number of explanatory variables be p, and the length of the response vector be q. The model specification is

$$Y_m = \eta(\alpha_m + \beta_m^T x), \tag{16}$$

where $\eta(\cdot)$ is a known, continuous function, α_m is a scalar intercept term, and β_m is a vector of slopes. Interest centers on the unknown parameters, α (a $q \times 1$ vector of intercepts), B (a $p \times q$ matrix of slopes) and Σ, a matrix of dispersion parameters. The prior on the unknown parameters, $\pi(\alpha, B, \Sigma)$ is taken to be of the form $\pi(\alpha, B, \Sigma) = \pi(\alpha|\Sigma)\pi(B|\Sigma)\pi(\Sigma)$; in addition, the authors elaborate $\pi(B|\Sigma)$ as $\pi(B, \gamma|\Sigma) = \pi(B|\Sigma, \gamma)\pi(\gamma)$, where γ is a latent binary vector of length p. Roughly speaking, $\gamma_j = 1$ when the covariance of the appropriate row of B is spread out, and $\gamma_j = 0$ when it is concentrated. Since priors are centered at 0 in their formulation, the two possible values of γ_j correspond to explanatory variables that should be included in, or excluded from, the model, respectively. In addition to generalizing the class of problems that can be handled by this latent parameter approach, Brown, Vannucci and Fearn (1998) introduce fast and efficient MCMC algorithms for the case when the number of explanatory variables is large (for instance, on the order of 100).

Within this same general model specification, Brown, Fearn and Vannucci (1999) describe a Bayesian decision–theoretic approach to the problem of variable selection. The setting is the multivariate linear regression, where costs are associated with the inclusion of explanatory variables. Typically, although not necessarily, the cost increases with the number of variables; the simplest cost function is additive, with common cost for each explanatory variable, although other scenarios are possible. This is a generalization of Lindley (1968), who considered the univariate multiple regression case. The goal is to predict a future response, Y^f; the criterion for judging predictors is quadratic loss, to which the cost function is added. Brown *et al.* (1999) point out that when this method omits variables, it is not because the researcher believes that the coefficients are truly zero, but rather because the omitted variables simply cost too much, relative to the benefits derived from them in terms of prediction.

4.4 Predictive Methods

The framework proposed in Section 3 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 jackknife 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 $M_1, \ldots, M_K$, p_k is the probability that model M_k is true and $p_k(y)$ is the predictive density of a future observation y based on the model M_k. 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

$$E_1[u(p_1(*), y) - u(p_2(*), y)]p_1 > E_2[u(p_2(*), y) - u(p_1(*), y)]p_2, \tag{17}$$

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 framework of Smith and Spiegelhalter (1980), with a penalty term that increases as the distance between the two models (as measured by the likelihood ratio statistic) increases. Recall from Section 4.1 that Smith and Spiegelhalter (1980) discussed Bayes factors of the form $\Lambda(m) = \lambda - m(d_2 - d_1)$, equation (12). Taking different utilities leads to different values of m; the method developed by San Martini and Spezzaferri has $m = \log(nC^{2/(d_2-d_1)})$, where C is a transformation of the likelihood ratio statistic.

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_k have likelihood $f(x|\theta_k)$. 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 $\{1, 2, \ldots, n\}$ and let S be a subset of S_n. Define

$$L(\theta_k|x_S) = \prod_{i=1}^{n} f(x_i|\theta_k)^{d_k}, \tag{18}$$

where d_k is the indicator for $k \in S$. As before, we denote the prior for θ_k under model M_k by $g_k(\theta_k)$. For prediction purposes, Gelfand and Dey propose consideration of the conditional density

$$f(x_{S_1}|x_{S_2}, M_k) \quad = \int L(\theta_k|x_{S_1}) g_k(\theta_k|x_{S_2}) d\theta_k$$

$$= \frac{\int L(\theta_k|x_{S_1})L(\theta_k|x_{S_2})g_k(\theta_k)d\theta_k}{\int L(\theta_k|x_{S_2})g_k(\theta_k)d\theta_k} \tag{19}$$

This conditional density is a predictive density; it averages the joint density of x_{S_1} with respect to the prior $g_k(\theta_k)$, 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

$$\prod_{r=1}^{n} f(x_r|x_{(r)}, M_1) / \prod_{r=1}^{n} f(x_r|x_{(r)}, M_2), \tag{20}$$

where $x_{(r)} = \{x_1, x_2, \ldots, x_{r-1}, x_{r+1}, \ldots, x_n\}$. $S_1 = S_2 = S$ 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 some of 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 Kadane and Dickey (1980) that Bayes factors 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 of replicates of the data, where the replicates are assumed to have the same distribution 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 no 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

$$L(y_{rep}, a; y_{obs}) = L(y_{rep}, a) + kL(y_{obs}, a) \tag{21}$$

for $k \geq 0$ 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, k, 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

$$D_k(m) \equiv \sum_{l=1}^{n} \min_{a_l} E_{y_{l,rep}|y_{obs},m} L(y_{l,rep}, a_l; y_{obs}); \tag{22}$$

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

$$D_k(m) = \sum_{l=1}^{n} \min_{a_l} \{E_{y_{l,rep}|y_{obs},m} L(y_{l,rep}, a_l) + kL(y_{l,obs}, a_l)\}. \tag{23}$$

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

$$\sigma_l^2 + (a_l - \mu_l)^2 + k(a_l - y_{l,obs})^2, \tag{24}$$

where σ_l^2 is the variance of $y_{l,rep}$ given y_{obs} and m, and μ_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 $(k+1)^{-1}(\mu_l + ky_{l,obs})$. If this is inserted back into the expression for $D_k(m)$, the result is

$$D_k(m) = \frac{k}{k+1}\sum_{l=1}^{n}(\mu_l - y_{l,obs})^2 + \sum_{l=1}^{n}\sigma_l^2. \tag{25}$$

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 μ_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 β a $N(\mu_b, \Sigma)$ distribution. If the prior is very imprecise, that is, Σ is large, then $y_{rep}|y_{obs}$ has an approximate $N(X\hat{\beta}, \sigma^2[I + X(X^TX)^{-1}X^T])$ distribution. The two summands in $D_k(m)$ 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 $D_k(m)$ 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.

4.5 Practical Issues: Elicitation and Computation

While our focus is on methods for model selection and the criteria that relate to them, we would be remiss if we did not mention the practical problems of elicitation and computation.

Of the quantities introduced in Section 3, only the data x have a claim of being agreed to as part of the statement of the problem. Each of the other quantities are “states of mind, not states of nature” in L.J. Savage’s elegant phrase. In particular, the models included in the model choice parameter M, the parameter spaces θ_k, the likelihoods $f_k(x|\theta_k)$, the priors π_k and $g_k(\theta_k)$ and the losses or utilities, are all matters of opinion on which conscientious statisticians and users of statistics can legitimately disagree without making a provable or logical error.

Elicitation of expert opinion is a feasible way of obtaining proper, subjective priors to incorporate into the model averaging procedure (as well as other Bayesian model selection techniques) and is the subject of a growing literature, much of it in the last ten years or so (see, for example, Kadane, Dickey, Winkler, Smith and Peters, 1980; Dickey, Dawid and Kadane, 1986; Garthwaite and Dickey, 1992; Kadane and Wolfson, 1998; O’Hagan, 1998; Garthwaite and Al–Awadhi, 2001). Garthwaite and Al–Awadhi, for example, propose a method for quantifying expert opinion about multivariate normal distributions. The basic idea is to simplify the elicitation by concentrating on one type of parameter at a time, asking the expert a series of questions, for example relating to the quantiles of the predictive distributions, as recommended by Kadane and Wolfson (1998).

A referee raises the question “if one has enough prior information to use a proper informative prior on each parameter of a particular model, why don’t they have enough information to know what the underlying model is without resorting to a model selection procedure.” While it is possible to have a prior that is opinionated with respect to what model obtains (*i.e.* know the model which model obtains with certainty regardless of the data), it is also possible to be less certain about which model obtains. Both states of belief are consistent with the subjective Bayesian position taken as the viewpoint of this review.

Regarding computation, it is worth noting that several 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). A recent review (Han and Carlin, 2001) compares reversible jump, marginal likelihood, and other approaches that use proper priors, in terms of computational ease, need for preprocessing, speed and accuracy. According to Han and Carlin (2001), "... *all* methods ... require significant human and computer effort, and this suggests that less formal Bayesian model choice methods may offer a more realistic alternative in many cases." (pg. 1122) Combining computation and asymptotic approximations, as shown by DiCiccio, Kass, Raftery and Wasserman (1997) is also an effective way of computing Bayes Factors for model comparison. Here, too, it was found that no one method is optimal in all situations, although a simple bridge sampler (Meng and Wong, 1996; Gelman and Meng, 1998) in conjunction with the Laplace approximation worked well in most cases. See DiCiccio *et al.* (1997) for details on the methods and their comparison, on both simulated and real data sets.

5 Frequentist Approaches to Model Selection

5.1 Techniques

Classical statistics has also dealt extensively with the problem of model selection. Every introductory book on regression analysis, for example, contains chapters on ways of choosing among competing models. In contrast to most of the Bayesian methods, classical approaches generally have had to focus on the comparison of nested models, as non–nested models are usually difficult to treat. Much of model choice in the classical setting is based on the principle of *extra sums of squares*, that is, comparing the residual sums of squares from models with and without particular sets of variables. Valid comparisons can be made for models that differ in that, in the smaller model, some of the parameters (coefficients on the variables) in the larger model are set to zero. In contrast, when using various criteria for model selection (as in the next section), models can be compared without being nested. For details on many of the methods to be considered in the rest of this section we refer readers to Miller (2002). Taken as a whole, the frequentist techniques and criteria do not fit in to the general Bayesian framework described in Section 3, since they do not specify priors or utilities. Where connections exist we point them out.

The various stepwise procedures, in which we include also forward selection and backward elimination, are among the most popular and widespread techniques. They all provide systematic ways of searching through models, where at each stage new models are obtained by adding or deleting one variable from the models at the previous stages. While these techniques originated for regression models to aid in the variable selection problem, they can also be applied in settings that extend the basic linear model, such as generalized linear models (Lawless and Singhal, 1978; Hastie and Pregibon, 1992), contingency tables (Agresti, 1990) and graphical models (Whittaker, 1990); for these other types, residual sum of squares would be replaced by deviance or other relevant measures. We frame our discussion in the regression context, with the understanding that the search philosophy can be used in other settings as well.

With forward selection, start with the null model and, one at a time, consider variables for inclusion in the model. At the first step, include the variable that makes the biggest individual contribution, assuming that the F–test for a model with that variable versus the null model is greater than a predetermined threshold. At each step the procedure continues in this way, adding in the variable that has the largest effect given the variables already in the model, if its F statistic is above the cutoff. When there is no candidate variable that meets the criterion, the algorithm stops. Another option is to set in advance the size of the largest model to be considered, and stop the procedure when that point is reached (Draper and Smith, 1981).

Backward elimination is similar, but moves in the opposite direction. That is, starting with the full model, at each step consider eliminating the variable with the least effect on the model, given that the other variables are included. Again, a predetermined threshold for dropping variables from the model decides whether or not the candidate will indeed be taken out. When no candidates for removal meet the criterion, stop.

In both forward selection and backward elimination, once a variable has been acted upon, that decision cannot be reversed. Hence, a variable that was eliminated at some point during a backward procedure, for example, will never be allowed back in to the model. This lack of flexibility is remedied in the stepwise approach to variable selection. Here, at each step each variable is considered for inclusion or elimination. Thus, a variable might be included in an early stage, but taken out later; or, a variable that was taken out of the model might be allowed back in.

While these procedures are widely used and readily available in most statistics packages, they should be used with care. Since none of the stepwise regression methods correspond to a specific criterion for choosing a model (Weisberg, 1985, p. 211), the selected model need not be optimal in any other sense than that it is the result of the algorithm applied to the data set. Indeed, working on the same data set, the forward selection and backward elimination might not result in the same final model (Graybill, 1976). Due to the way that the algorithms work, furthermore, not all models will even be looked at. The lack of a clear criterion for model choice makes it difficult to see how these procedures fit at all into our general Bayesian framework, or, indeed, into a frequentist approach, since they each involve a complex sequential testing strategy with a dynamically changing null hypothesis.

An alternative to stepwise regression is to do an exhaustive search across all models and in such a fashion to find subsets of the variables that yield a good model, according to some criterion (see below for a discussion of possible choices). These are usually used as a starting point for further study. This approach, even with advances in computing power and memory, as well as the development of algorithms that allow the user to avoid calculating most of the models (for instance, Furnival and Wilson, 1974), is feasible mostly when the number of variables is moderate. In any case, exhaustive search over all possible models is usually naive – the statistician or the scientist often has ideas about which candidate models make substantive sense.

5.2 Criteria for Subset Selection

As described above, the exhaustive search, or all possible regressions, compares models according to a specific criterion. Those models that perform well according to the chosen criterion may be considered for a more in–depth investigation. Over the years, many criteria have been suggested. Some of them, such as AIC and BIC, have already been discussed. They have a role in classical model choice no less than in the Bayesian counterpart.

Most of the popular criteria for model selection are readily computed as byproducts of the ordinary regression calculations, but don't necessarily have counterparts in other common model settings; hence this section discusses only the problem of variable selection in regression. R^2, for instance, is defined as the ratio of the sum of squares for regression to the total sum of squares, $\sum(y_i - \bar{y})^2$. The problem with using this measure as a criterion, specifically for comparing models of different sizes, is that the sum of squares for regression, and hence R^2 itself, increases the more variables there are in the model. For this reason, an adjusted version of R^2, which takes into consideration the number of parameters in the model, is usually used instead. It is defined to be

$$R^2_{adj} = 1 - \frac{(n-1)}{(n-p)}(1 - R^2), \tag{26}$$

where n is the sample size and p is the number of variables in the model (including the intercept term).

A related criterion is the C_p statistic (Mallows, 1973),

$$C_p = RSS_p/\hat{\sigma}^2 + (2p - n), \tag{27}$$

with RSS_p the residual sum of squares for a model with p terms, and $\hat{\sigma}^2$ the estimate of the error variance based on the full model. C_p is closely related to R^2_{adj} (Kennard, 1971). A number of features of this statistic make it useful for model comparison. For a model that fits the data adequately, $E(C_p)$ is approximately p, and therefore C_p itself should be approximately equal to p for an adequate model (of which there may be several in a given problem). For the full model, with, say k parameters, this holds exactly, that is, $C_k = k$. The criterion can clearly be used for comparing subsets of the same size, but it can also be used more generally, by looking for those models for which $C_p \approx p$. The purpose of C_p is to guide the researcher in the process of subset selection (Mallows, 1995; George, 2000); choosing the model that minimizes the criterion and then estimating the parameters of the model via least squares, although a widespread practice, is prone to selection bias and should be avoided (Mallows, 1995; the problem is that the common procedure does not account for the fact that the selected subset depends on the observed data). See the discussion of the Risk Inflation Factor, below, for more on this question.

One of the motivations for the C_p statistic is as an estimate of the mean square error for prediction. It is possible instead to use cross–validation to get such an measure. Delete observation i for each of $i = 1, \ldots, n$ and fit the regression model with the i^{th} observation deleted. Using the fitted values, it is possible to obtain a "prediction" for the deleted point, which can be compared to its actual value. The difference in the two is sometimes called the *deleted resid-*

ual. The sum of the squared deleted residuals is the predicted residual sum of squares, PRESS (Allen, 1974). Good models will have small values of this criterion. Similar thinking drives the pseudo–Bayes method of Geisser and Eddy (1979) discussed previously. It is important to note that, at least in theory, one needs to go through the procedure on each of the models being considered, which could be a computational burden if the number of models is large.

5.3 Modern Frequentist Developments

As in the Bayesian world, refinements and innovations on frequentist procedures continue to appear (George, 2000). New criteria, such as the risk inflation criterion (Foster and George, 1994; Donoho and Johnstone, 1994) and the covariance inflation criterion (Tibshirani and Knight, 1999) have been proposed within the last decade. Advances in computation have created new opportunities, with the now–standard cross–validation and bootstrap (Efron, 1979, 1982; Stone, 1974) as well as more exotic procedures such as the "little bootstrap" (Breiman, 1992), the nonnegative garrote (Breiman, 1995) and the lasso (Tibshirani, 1996) coming into play.

Foster and George (1994) note that the variable selection problem in regression is actually a two stage process – first, a "best" subset of predictors is selected, and then the coefficients of the chosen subset are calculated by least squares. The second stage proceeds as if the predictors are known to be the correct ones, rather than having been chosen. The Risk Inflation Criterion, or RIC, is defined to be the maximum possible increase in risk due to selecting the variables in the model, as opposed to knowing which the "correct" ones are. The inflation comes from comparing the risk of the fitted model to the risk of the ideal model which uses only the "right" variables. RIC turns out to be related to other criteria we have already encountered, such as AIC, C_p and BIC, the difference being in the penalty it imposes on the dimensionality of the model – $2 \log k$, where k is the dimension of the full model, using all predictors. This same penalty was arrived at by Donoho and Johnstone (1994) for a wavelet model choice problem. More recent work by George and Foster (2000) shows that the criteria in this family correspond to a Bayesian model selection procedure under a particular class of priors. Their work provides a bridge between frequentist and Bayesian criteria. An empirical Bayes analysis results here in an adaptive dimension penalty, as opposed to the fixed penalties of AIC, BIC, C_p, RIC and the like. Additional advantages of the empirical Bayes approach of George and Foster (2000) are that it automatically allows for shrinkage of the least squares estimates of the selected variables, and that it fits quite naturally into a model averaging framework.

The covariance inflation criterion (Tibshirani and Knight, 1999) has a similar motivation to the RIC. It is a criterion for model selection in prediction problems, whereby a model is chosen based on a training set of data to find the best predictor of future data. The method adjusts the training error by the average covariance of the response and the predictors, when the model is applied to permutations of the original data set.

Some of the other more recent developments in the area – the little bootstrap, the nonnegative garrote and the lasso, mentioned above, also take advantage of advances in computing power. Breiman's (1995) nonnegative garrote grows out of an attempt to keep the strengths of both subset selection and ridge regression. The advantage of the former is that it does select out variables; however, it is highly unstable, in that small changes in the data set can lead to very different models. Ridge regression, on the other hand, is very stable, but does not eliminate any variables, leading to possibly cumbersome models that are (or can be) hard to interpret. Again, in the linear model setting, let $\hat{\beta}_i$ be the original least squares estimates of the coefficients, and take c_i to minimize

$$\sum_j (y_j - \sum_i c_i \hat{\beta}_i x_{ij})^2 \tag{28}$$

subject to the constraints that $c_i \geq 0$ for all i and that $\sum_i c_i \leq s$. By decreasing s, more of the c_i become zero, and the ones that don't are shrunk, thereby also shrinking the remaining parameter estimates, $\hat{\beta}_i(s) = c_i \hat{\beta}_i$. This "garrote" is relatively stable, while eliminating some variables from consideration. It tends to lead to larger models than ordinary subset regression, but on the other hand it is, in many instances, more accurate (in terms of prediction). The "little bootstrap" (Breiman, 1992) or cross–validation (Stone, 1974; Efron, 1982) can be used to estimate the value of the garroting parameter, s.

A similar idea is captured by Tibshirani's lasso (1996), which chooses β_is to minimize

$$\sum_j (y_j - \sum_i \beta_i x_{ij})^2, \tag{29}$$

under the constraint that $\sum_i |\beta_i| \leq s$. Here, s controls the amount of shrinkage. As noted by Tibshirani, a main difference between the lasso and the garrote is that the latter modifies the ordinary least squares estimates, and hence its behavior is, at least in part, dependent on theirs. In contrast, with the lasso there is no explicit use of the least squares estimates. Tibshirani also offers a Bayesian interpretation of the lasso estimates, as the posterior mode under independent double–exponential priors on the βs.

6 Conclusions

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 and interesting of problems. With such a wealth of methods, it can be 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 Key, 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. Objective Bayesians try to avoid the discomfort of selecting a subjective (proper) prior, that is, they hope to "have their Bayesian cake and eat it too.”

Acknowledgments: The authors thank the two anonymous referees and the Reviews editor for their helpful comments on an earlier draft of this manuscript.

REFERENCES

Agresti, A. (1990) *Categorical Data Analysis*. New York: John Wiley & Sons.

Aitkin, M. (1991) Posterior Bayes factors (with discussion). *Journal of the Royal Statistical Society, Series B*, **53**, 111–142.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, Petrox, B.N. and Caski, F. (eds). Budapest: Akademiai Kiado, 267–281.

Akaike, H. (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.

Bayarri, M.J. (1999) Discussion of Bayesian model choice: What and why? In *Bayesian Statistics 6*, Bernardo, J.M., Berger, J.O., Dawid, A.P. and Smith, A.F.M. (eds). Oxford: Oxford University Press, 357–359.

Berger, J.O. and Pericchi, L.R. (1996) The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109–122.

Berger, J.O. and Pericchi, L.R. (1997) On criticisms and comparisons of default Bayes factors for model selection and hypothesis testing (with discussion). In *Proceedings of the Workshop on Model Selection*, Racugno, W. (ed). Bologna: Pitagora Editrice, 1–50.

Berger, J.O. and Pericchi, L.R. (1998) Accurate and stable Bayesian model selection: The median intrinsic Bayes factor. *Sankhyā*, *B*, **60**, 1–18.

Bernardo, J.M. and Rueda, R. (2002) Bayesian hypothesis testing: A reference approach. *International Statistical Review*, **70**, 351–372.

Bernardo, J.M. and Smith, A.F.M. (1994) *Bayesian Theory*. Chichester: John Wiley & Sons.

Bertolino, F. and Racugno, W. (1997) Is the intrinsic Bayes factor intrinsic? *Metron*, **LIV**, 5–15.

Box, G.E.P. (1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society, Series A*, **143**, 383–430.

Box, G.E.P. and Tiao, G.C. (1992) *Bayesian Inference in Statistical Analysis*. New York: John Wiley & Sons.

Breiman, L. (1992) The little bootstrap and other methods for dimensionality selection in regression: X–fixed prediction error. *Journal of the American Statistical Association*, **87**, 738–754.

Breiman, L. (1995) Better subset selection using the nonnegative garrote. *Technometrics*, **37**, 373–384.

Brown, P.J., Vannucci, M. and Fearn, T. (1998) Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B*, **60**, 627–641.

Brown, P.J., Fearn, T. and Vannucci, M. (1999) The choice of variables in multivariate regression: A non–conjugate Bayesian decision theory approach. *Biometrika*, **86**, 635–648.

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

Carlin, B.P. and Polson, N.G. (1991) Inference for nonconjugate Bayesian models using the Gibbs sampler. *Canadian Journal of Statistics*, **19**, 399–405.

Chib, S. (1995) Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, **90**, 1313–1321.

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

Dickey, J.M., Dawid, A.P. and Kadane, J.B. (1986) Subjective–probability assessment methods for multivariate–t and matrix–t models. In *Bayesian Inference and Decision Techniques*, Goel, P.K. and Zellner, A. (eds). Amsterdam: Elsevier Science Publishers B.V., 177–195.

Donoho, D.L. and Johnstone, I.M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B*, **57**, 45–97.

Draper, N. and Smith, H. (1981) *Applied Regression Analysis*, Second Edition. New York: John Wiley & Sons.

Efron, B. (1979) Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**, 1–26.

Efron, B. (1982) *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Foster, D.P. and George, E.I. (1994) The risk inflation criterion for multiple regression. *The Annals of Statistics*, **22**, 1947–1975.

Furnival, G. and Wilson, R. (1974) Regression by leaps and bounds. *Technometrics*, **16**, 499–511.

Garthwaite, P.H. and Al–Awadhi, S.A. (2001) Non–conjugate prior distribution assessment for multivariate normal sampling. *Journal of the Royal Statistical Society, Series B*, **63**, 95–110.

Garthwaite, P.H. and Dickey, J.M. (1992) Elicitation of prior distributions for variable–selection problems in regression. *The Annals of Statistics*, **20**, 1697–1719.

Geisser, S. and Eddy, W.F. (1979) A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153–160.

Gelfand, A.E. and Dey, D.K. (1994) Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B*, **56**, 501–514.

Gelfand, A.E. and Ghosh, S.K. (1998) Model choice: A minimum posterior predictive loss approach. *Biometrika*, **85**, 1–11.

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

George, E.I. (2000) The variable selection procedure. *Journal of the American Statistical Association*, **95**, 1304–1308.

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

George, E.I. and McCullogh, R.E. (1993) Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

George, E.I. and McCullogh, R.E. (1997) Approaches for Bayesian variable selection. *Statistica Sinica*, **7**, 339–373.

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

Goldstein, M. (1991) Discussion of Posterior Bayes factors. *Journal of the Royal Statistical Society, Series B*, **53**, 134.

Good, I.J. (1950) *Probability and the Weighting of Evidence*, London: Charles Griffin.

Graybill, F.A. (1976) *Theory and Application of the Linear Model*, Pacific Grove: Wadsworth & Brooks/Cole.

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

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

Hastie, T.J. and Pregibon, D. Generalized linear models. In *Statistical Models in S*, Chambers, J.M. and Hastie, T.J. (eds). Pacific Grove: Wadsworth & Brooks/Cole.

Jeffreys, H. (1961) *Theory of Probability* (3rd ed.), London: Oxford University Press.

Kadane, J.B. and Dickey, J.M. (1980) Bayesian decision theory and the simplification of models. In *Evaluation of Econometric Models*, Kmenta, J. and Ramsey, J. (eds). New York: Academic Press, 245–268.

Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S. and Peters, S.C. (1980) Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, **75**, 845–854.

Kadane, J.B. and Wolfson, L.J. (1998) Experiences in elicitation. *The Statistician*, **47**, 3–19.

Kass, R.E. and Raftery, A.E. (1995) Bayes Factors. *Journal of the American Statistical Association*, **90**, 773–795.

Katz, R.W. (1981) On some criteria for estimating the order of a Markov chain. *Technometrics*, **23**, 243–249.

Kennard, R.W. (1971) A note on the C_p statistic. *Technometrics*, **13**, 899–900.

Key, J.T., Pericchi, L.R. and Smith, A.F.M. (1999) Bayesian model choice: What and why? (with discussion) In *Bayesian Statistics 6*, Bernardo, J.M., Berger, J.O., Dawid, A.P. and Smith, A.F.M. (eds). Oxford: Oxford University Press, 343–370.

Koehler, A.B. and Murphree, E.S. (1988) A comparison of the Akaike and Schwarz criteria for selecting model order. *Applied Statistics*, **37**, 187–195.

Kuo, L. and Mallick, B. (1998) Variable selection for regression models. *Sankhyā*, **60**, 65–81.

Laud, P.W. and Ibrahim, J.G. (1995) Predictive model selection. *Journal of the Royal Statistical Society, Series B*, **57**, 247–262.

Lawless, J.F. and Singhal, K. (1978) Efficient screening of non–normal regression models. *Biometrics*, **43**, 318–327.

Leamer, E.E. (1978) *Specification Searches*. New York: John Wiley & Sons.

Lindley, D.V. (1957) A statistical paradox. *Biometrika*, **44**, 187–192.

Lindley, D.V. (1968) The choice of variables in multiple regression (with Discussion). *Journal of the Royal Statistical Association, Series B*, **30**, 31–66.

Lindley, D.V. (1976) A class of utility functions. *The Annals of Statistics*, **4**, 1–10.

Lindley, D.V. (1991) Discussion of Posterior Bayes factors. *Journal of the Royal Statistical Society, Series B*, **53**, 130–131.

Lindley, D.V. (1997) Some comments on Bayes factors. *Journal of Statistical Planning and Inference*, **61**, 181–189.

MacKay, D.J.C. (1992) Bayesian interpolation. *Neural Computation*, **4**, 415–447.

Madigan, D. and Raftery, A.E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**, 1535–1546.

Madigan, D. and York, J. (1995) Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215–232.

Mallows, C.L. (1973) Some comments on C_p. *Technometrics*, **15**, 661–675.

Mallows, C.L. (1995) More comments on C_p. *Technometrics,* **37**, 362–372.

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

Miller, A.J. (2002) *Subset Selection in Regression,* Second Edition. London: Chapman & Hall.

Mitchell, T.J. and Beauchamp, J.J. (1988) Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association*, **83**, 1023–1036.

O'Hagan, A. (1991) Discussion of Posterior Bayes factors. *Journal of the Royal Statistical Society, Series B*, **53**, 136.

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

O'Hagan, A. (1997) Properties of intrinsic and fractional Bayes factors. *Test*, **6**, 101–118.

O'Hagan, A. (1998) Eliciting expert beliefs in substantial practical applications. *Statistician*,

47, 21–35.

Petrone, S. (1997) Discussion of Choosing among models when none of them are true. In *Proceedings of the Workshop on Model Selection*, Racugno, W. (ed). Bologna: Pitagora Editrice, 355–358.

Piccinato, L. (1997) Discussion of Choosing among models when none of them are true. In *Proceedings of the Workshop on Model Selection*, Racugno, W. (ed). Bologna: Pitagora Editrice, 350–354.

Raftery, A.E., Madigan, D. and Hoeting, J.A. (1997) Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, **92**, 179–191.

Richardson, S. and Green, P.J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B*, **59**, 731–792.

San Martini, A. and Spezzaferri, F. (1984) A predictive model selection criterion. *Journal of the Royal Statistical Society, Series B, 46*, 296–303.

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

Shafer, G. (1982) Lindley's paradox. *Journal of the American Statistical Association*, **77**, 325–351.

Smith, A.F.M. and Spiegelhalter, D.J. (1980) Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society, Series B*, **42**, 213–220.

Stone, M. (1974) Cross–validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B*, **36**, 111–147.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Tibshirani, R. and Knight, K. (1999) The covariance inflation criterion for model selection. *Journal of the Royal Statistical Society, Series B*, **61**, 529–546.

Weisberg, S. (1985) *Applied Linear Regression*, Second Edition. New York: John Wiley & Sons.

Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics*. New York: John Wiley & Sons.

Winkler, R.L. (1999) Discussion of Bayesian model choice: What and why? In *Bayesian Statistics 6*, Bernardo, J.M., Berger, J.O., Dawid, A.P. and Smith, A.F.M. (eds). Oxford: Oxford University Press, 367–368.