Statistical Science
2004, Vol. 19, No. 1, 81–94
DOI 10.1214/088342304000000035

Model Uncertainty

Merlise Clyde and Edward I. George

Abstract. The evolution of Bayesian approaches for model uncertainty over the past decade has been remarkable. Catalyzed by advances in methods and technology for posterior computation, the scope of these methods has widened substantially. Major thrusts of these developments have included new methods for semiautomatic prior specification and posterior exploration. To illustrate key aspects of this evolution, the highlights of some of these developments are described.

Key words and phrases: Bayes factors, classification and regression trees, model averaging, linear and nonparametric regression, objective prior distributions, reversible jump Markov chain Monte Carlo, variable selection.

1. INTRODUCTION

Advances in computing technology over the past few decades have allowed for the consideration of an increasingly wider variety of statistical models for data $\mathbf{Y}$. It is now often routine to consider many possible models, say $\mathcal{M}_1, \ldots, \mathcal{M}_K$, where each model $\mathcal{M}_k$ consists of a family of distributions $\{p(\mathbf{Y}|\boldsymbol{\theta}_k, \mathcal{M}_k)\}$ indexed by $\boldsymbol{\theta}_k$, a (possibly vector) parameter. For such setups, the Bayesian approach provides a natural and general probabilistic framework that simultaneously treats both model and parameter uncertainty. Coupled with the advent of Markov chain Monte Carlo (MCMC) methods for posterior computation (Gelfand and Smith, 1990; Besag and Green, 1993; Smith and Roberts, 1993; Tierney, 1994; Andrieu, Doucet and Robert, 2004, this issue, for a historical overview and discussion of recent advances), the development and application of Bayesian methods for model uncertainty (Hodges, 1987; Draper, 1995; Hoeting, Madigan, Raftery and Volinsky, 1999; Berger and Pericchi, 2001; Chipman, George and McCulloch, 2001) has seen remarkable evolution over the past decade. Before discussing some of the major innovations that have occurred, let us lay out the essential ideas of the approach.

The comprehensive Bayesian approach for multiple model setups proceeds by assigning a prior probability distribution $p(\boldsymbol{\theta}_k|\mathcal{M}_k)$ to the parameters of each model, and a prior probability $p(\mathcal{M}_k)$ to each model. This prior formulation induces a joint distribution $p(\mathbf{Y}, \boldsymbol{\theta}_k, \mathcal{M}_k) = p(\mathbf{Y}|\boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\theta}_k|\mathcal{M}_k)p(\mathcal{M}_k)$ over the data, parameters and models. In effect, these priors serve to embed the various separate models within one large hierarchical mixture model. Under this full model, the data are realized in three stages: first the model $\mathcal{M}_k$ is generated from $p(\mathcal{M}_1), \ldots, p(\mathcal{M}_K)$; second the parameter vector $\boldsymbol{\theta}_k$ is generated from $p(\boldsymbol{\theta}_k|\mathcal{M}_k)$; third the data $\mathbf{Y}$ are generated from $p(\mathbf{Y}|\boldsymbol{\theta}_k, \mathcal{M}_k)$. Through conditioning and marginalization, the joint distribution $p(\mathbf{Y}, \boldsymbol{\theta}_k, \mathcal{M}_k)$ can be used to obtain posterior summaries of interest.

Margining out the parameters $\boldsymbol{\theta}_k$ and conditioning on the data $\mathbf{Y}$ yields the posterior model probabilities

$$p(\mathcal{M}_k|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathcal{M}_k)p(\mathcal{M}_k)}{\sum_k p(\mathbf{Y}|\mathcal{M}_k)p(\mathcal{M}_k)}, \tag{1}$$

where

$$p(\mathbf{Y}|\mathcal{M}_k) = \int p(\mathbf{Y}|\boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\theta}_k|\mathcal{M}_k)\,d\boldsymbol{\theta}_k \tag{2}$$

is the marginal likelihood of $\mathcal{M}_k$. [When $p(\boldsymbol{\theta}_k|\mathcal{M}_k)$ is a discrete distribution, integration in (2) is replaced by summation.] Under the full three-stage hierarchical model interpretation for the data, $p(\mathcal{M}_k|\mathbf{Y})$ is the conditional probability that $\mathcal{M}_k$ was the actual model generated at the first stage.

Merlise Clyde is Associate Professor of Statistics, Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina, USA 27708-0251 (e-mail: clyde@isds.duke.edu). Edward I. George is Universal Furniture Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, USA (e-mail: edgeorge@wharton.upenn.edu).

Based on these posterior probabilities, pairwise comparison of models is summarized by the posterior odds

$$\frac{p(\mathcal{M}_k|\mathbf{Y})}{p(\mathcal{M}_j|\mathbf{Y})} = \frac{p(\mathbf{Y}|\mathcal{M}_k)}{p(\mathbf{Y}|\mathcal{M}_j)} \times \frac{p(\mathcal{M}_k)}{p(\mathcal{M}_j)}. \tag{3}$$

This expression reveals how the data, through the Bayes factor $B[k:j] \equiv \frac{p(y|\mathcal{M}_k)}{p(y|\mathcal{M}_j)}$, updates the prior odds $O[k:j] = \frac{p(\mathcal{M}_k)}{p(\mathcal{M}_j)}$ to yield the posterior odds. The Bayes factor $B[k:j]$ summarizes the relative support for $\mathcal{M}_k$ versus $\mathcal{M}_j$ provided by the data. Note that the Bayes posterior model probabilities (1) can be expressed entirely in terms of Bayes factors and prior odds as

$$p(\mathcal{M}_k|\mathbf{Y}) = \frac{B[k:j]O[k:j]}{\sum_k B[k:j]O[k:j]}. \tag{4}$$

Insofar as the priors $p(\boldsymbol{\theta}_k|\mathcal{M}_k)$ and $p(\mathcal{M}_k)$ provide an initial representation of model uncertainty, the model posterior $p(\mathcal{M}_1|\mathbf{Y}), \ldots, p(\mathcal{M}_K|\mathbf{Y})$ provides a complete representation of postdata model uncertainty that can be used for a variety of inferences and decisions. By treating $p(\mathcal{M}_k|\mathbf{Y})$ as a measure of the "truth" of model $\mathcal{M}_k$, a natural and simple strategy for model selection is to choose the most probable $\mathcal{M}_k$, the modal model for which $p(\mathcal{M}_k|\mathbf{Y})$ is largest. This and other strategies can be motivated by utility considerations as we will discuss in Section 6. Model selection may be useful for testing a theory represented by one of a set of carefully studied models, or it may simply serve to reduce attention from many speculative models to a single useful model. However, in problems where no single model stands out, it may be preferable to report a set of models with high posterior probability along with their probabilities to convey the model uncertainty.

Bayesian model averaging is an alternative to Bayesian model selection that incorporates rather than ignores model uncertainty. For example, suppose interest focused on the distribution of $\mathbf{Y_f}$, a future observation from the same process that generated $\mathbf{Y}$. Under the full model for the data induced by the priors, the Bayesian predictive distribution of $\mathbf{Y_f}$ is obtained as

$$p(\mathbf{Y_f}|\mathbf{Y}) = \sum_k p(\mathbf{Y_f}|\mathcal{M}_k, \mathbf{Y})\, p(\mathcal{M}_k|\mathbf{Y}), \tag{5}$$

a posterior weighted mixture of the conditional predictive distributions

$$p(\mathbf{Y_f}|\mathcal{M}_k, \mathbf{Y}) = \int p(\mathbf{Y_f}|\boldsymbol{\theta}_k, \mathcal{M}_k)\, p(\boldsymbol{\theta}_k|\mathcal{M}_k, \mathbf{Y})\, d\boldsymbol{\theta}_k. \tag{6}$$

By averaging over the unknown models, $p(\mathbf{Y_f}|\mathbf{Y})$ incorporates the model uncertainty embedded in the priors. A natural point prediction of $\mathbf{Y_f}$ is obtained as the mean of $p(\mathbf{Y_f}|\mathbf{Y})$, namely

$$E(\mathbf{Y_f}|\mathbf{Y}) = \sum_k E(\mathbf{Y_f}|\mathcal{M}_k, \mathbf{Y})\, p(\mathcal{M}_k|\mathbf{Y}). \tag{7}$$

Such model averaging or mixing procedures have been developed and advocated by Leamer (1978b), Geisser (1993), Draper (1995), Raftery, Madigan and Volinsky (1996) and Clyde, DeSimone and Parmigiani (1996).

A major appeal of the Bayesian approach to model uncertainty is its complete generality. In principle, it can be applied whenever data are treated as a realization of random variables, a cornerstone of model statistical practice. The past decade has seen the development of innovative implementations of Bayesian treatments of model uncertainty for a wide variety of potential model specifications. Each implementation has required careful attention to prior specification and posterior calculation. The evolution of these innovations is nicely illustrated in the context of the variable selection problem, on which we focus next.

2. VARIABLE SELECTION UNCERTAINTY

For a given response variable of interest $\mathbf{Y}$, and a set of potential predictors $\mathbf{X}_1, \ldots, \mathbf{X}_p$, the problem of variable selection, or subset selection as it is often called, is one of the most fundamental and widespread model selection problems in statistics (see George, 2000; Miller, 2002). Often vaguely stated as the problem of selecting the "best" predictor subset for $\mathbf{Y}$, Bayesian approaches to this problem encourage the formulation of more precise objectives. By providing an explicit description of model uncertainty, which here can be thought of as "variable selection uncertainty," the Bayesian hierarchical mixture approach transforms the problem into one of choosing the appropriate procedure to exploit posterior information. It reveals that, depending on how the solution is going to be used, model averaging might be a preferable alternative to model selection (see Section 6), a curious twist for the so-called variable selection problem.

The variable selection problem is usually posed as a special case of the model selection problem, where each model under consideration corresponds to a distinct subset of $\mathbf{X}_1, \ldots, \mathbf{X}_p$. It is most familiar in the context of multiple regression where attention is restricted to submodels of the normal linear model. Letting $\boldsymbol{\gamma}$ index the subsets of $\mathbf{X}_1, \ldots, \mathbf{X}_p$, each submodel is of the form

$$\mathcal{M}_\gamma\colon\ \mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon}, \tag{8}$$

where $\mathbf{X}_{\boldsymbol{\gamma}}$ is the design matrix whose columns correspond to the $\boldsymbol{\gamma}$th subset, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the vector of regression coefficients for the $\boldsymbol{\gamma}$th subset and $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 I)$. Many of the fundamental developments in variable selection, both Bayesian and non-Bayesian, have occurred in the context of the linear model, in large part because its analytical tractability greatly facilitates insight and computational reduction, and because it provides a simple first-order approximation to more complex relationships. Initial and fundamental Bayesian mixture model approaches to variable selection uncertainty for the general normal linear model include Leamer (1978a, b), Zellner and Siow (1980), Zellner (1984), Stewart and Davis (1986), Mitchell and Beauchamp (1988), George and McCulloch (1993), Geweke (1996), Clyde, DeSimone and Parmigiani (1996), Smith and Kohn (1996), George and McCulloch (1997) and Raftery, Madigan and Hoeting (1997). The univariate regression setup above extends naturally to multiple response models where each row of $\mathbf{Y}$ is multivariate normal. Bayesian approaches for variable selection uncertainty in multivariate regression models were developed by Brown, Vannucci and Fearn (1998, 2002).

The importance of the linear variable selection problem has been greatly enhanced by the realization that it is a canonical version for nonparametric regression, a problem of growing current interest. Letting y and $x = (x_1, \ldots, x_p)$ be elements of $\mathbf{Y}$ and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$, nonparametric regression approximates the unknown regression function $E(y|x)$ as a linear combination of a finite number of basis functions of x. For example, in the simple case where x is univariate, regression spline representations are obtained using truncated power series basis functions

$$E(y|x) = \sum_{i=0}^{q} x^i \alpha_i + \sum_{j=1}^{k} (x - t_j)_+^q \beta_j, \tag{9}$$

where q is the order of the spline, $(\cdot)_+$ is the positive-part function and $t_1, \ldots, t_k$ are the knot locations. Because removing $(x - t_j)_+^q$ is equivalent to removing the knot at t_j, uncertainty about the knot locations, which are crucial for fitting, corresponds directly to linear variable selection uncertainty. Another powerful nonparametric regression representation of $E(y|x)$ is in terms of a multiresolution wavelet basis,

$$E(y|x) = \beta_0 + \sum_{j=1}^{J-1} \sum_{i=1}^{n2^{j-J}} \phi_{ji}(x) \beta_{ji}, \tag{10}$$

where $\phi_{ji}(x) = 2^{-j/2} \psi(2^{-j} x - i)$ are scalings and translations of a mother wavelet $\psi(x)$. Variable selection uncertainty here too corresponds to uncertainty about which basis variables to include, which is crucial for determining the appropriate degree of smoothness of the regression function. Bayesian variable selection approaches for this and other nonparametric regression problems have proved to be very successful. For examples of the potential of Bayesian regression spline approaches see Smith and Kohn (1996, 1997), Denison, Mallick and Smith (1998a, c), Wood and Kohn (1998), Shively, Kohn and Wood (1999), Hansen and Yu (2001), Wood, Kohn, Shively and Jiang (2002), Liang, Truong and Wong (2001) and Hansen and Kooperberg (2002). For examples of the potential of Bayesian wavelet regression see approaches by Chipman, Kolaczyk and McCulloch (1997), Clyde, Parmigiani and Vidakovic (1998), Abramovich, Sapatinas and Silverman 1998 and Kohn, Marron and Yau (2000), for example. For further reading, see the article by Müller and Quintana (2004, this issue) and the book by Denison, Holmes, Mallik and Smith (2002). Recent developments using overcomplete representations through frames where the number of variables p is potentially greater than n show great promise for adaptive, sparse representations of functions (Wolfe, Godsill and Ng, 2004).

Finally, an important and natural generalization of the linear variable selection problem is to the class of generalized linear models (GLMs) (McCullagh and Nelder, 1989). GLMs allow for any exponential family distribution for $\mathbf{Y}$. In addition to the normal, these include the binomial, multinomial and Poisson families, which may be more appropriate when $\mathbf{Y}$ is discrete. When $\mathbf{Y}$ is discrete categorical data, such models are sometimes referred to as classification models. When there is variable selection uncertainty, each GLM subset model for the regression function relates the conditional mean of $E(\mathbf{Y}|\mathbf{X})$ to $\mathbf{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}$ via a link function g,

$$\mathcal{M}_\gamma\colon\ g\big(E(\mathbf{Y}|\mathbf{X})\big) = \mathbf{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}. \tag{11}$$

In addition to variable selection uncertainty here, g may also be treated as an unknown parameter (Ntzoufras, Dellaportas and Forster, 2003). Going further, both (8) and (11) also may be enhanced by introducing additional parameters, for example, replacing $\mathbf{Y}$ by $\mathbf{Y}^\alpha$ to allow for a Box–Cox transformation (Hoeting, Raftery and Madigan, 2002). By extending the parameter prior, the introduction of such parameters poses no essential difficulty for the

Bayesian approach. Illustrations of the success of the Bayesian approach for variable selection uncertainty in generalized linear models can be found in George, McCulloch and Tsay (1996), Raftery (1996), Clyde (1999), Ibrahim, Chen and MacEachern (1999), Chen, Ibrahim and Yiannoutsos (1999), Ibrahim, Chen and Ryan (2000), Chen, Ibrahim, Shao and Weiss (2003), Dellaportas and Forster (1999), Ntzoufras, Forster and Dellaportas (2000) and Wakefield and Bennett (1996).

3. BAYESIAN VARIABLE SELECTION EVOLVES

Implementation of the Bayesian mixture approach entails two challenges: prior specification and posterior calculation. A key consideration in meeting these challenges for the variable selection problem is that the number of subset models 2^p grows so rapidly with p. In this section, we describe a thread of developments that illustrates how this aspect influenced attempts to meet these challenges. Of course, this is only one story and there are many other interesting threads in the wide variety of papers mentioned in the previous section.

Early Bayesian mixture model formulations for the variable selection setup (8) anticipated many features of current Bayesian approaches such as particular prior specification forms and model averaging; see Leamer (1978a, b), Zellner and Siow (1980), Zellner (1984), Stewart and Davis (1986) and Mitchell and Beauchamp (1988). Recognizing the critical importance of posterior computation, especially for large p, this work also contained prescient suggestions such as importance sampling and branch-and-bound reduction strategies. Rapid advances in the speed and capacity of computing technology over the following decade would greatly enhance the potential of these methods.

However, a most influential innovation was the advent of MCMC methods such as the Gibbs sampler and the Metropolis–Hastings algorithms (Gelfand and Smith, 1990; Besag and Green, 1993; Smith and Roberts, 1993). Development of Bayesian variable selection quickly took off when it became apparent that MCMC algorithms could be used to simulate a (sequentially dependent) sample

$$\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, \boldsymbol{\gamma}^{(3)}, \ldots \tag{12}$$

that was converging in distribution to the posterior model probabilities $p(\boldsymbol{\gamma}|\mathbf{Y})$ (George and McCulloch, 1993; Smith and Kohn, 1996; Geweke, 1996; Clyde, DeSimone and Parmigiani, 1996; George and McCulloch, 1997; Raftery, Madigan and Hoeting, 1997). Such a sequence could be used to search for high probability models for model selection and to obtain posterior weighted estimates for model averaging.

The availability of such MCMC strategies for exploration of the model posterior had an interesting effect on the choice of parameter priors. A major initial appeal of MCMC methods was that they could be used with wide classes of priors, thus emancipating Bayesian analysis from the constraint of using conjugate priors that had allowed for closed form posterior computation. However, for the purpose of exploring the model posterior, it was quickly realized that the use of conjugate priors offered tremendous computational advantages both for simulating and extracting information from (12), a huge priority for the large model spaces which arose in variable selection problems. The key advantages provided by conjugate priors stemmed from the fact that they yielded rapidly computable closed form expressions for the marginal distributions $p(\mathbf{Y}|\boldsymbol{\gamma})$. The advantages were twofold.

First, closed forms allowed for Metropolis–Hastings (MH) algorithms to simulate (12) as a Markov chain directly from $p(\boldsymbol{\gamma}|\mathbf{Y})$. Given the model sequence (12) up to $\boldsymbol{\gamma}^{(k)}$, such algorithms proceed by simulating a candidate $\boldsymbol{\gamma}^*$ for $\boldsymbol{\gamma}^{(k+1)}$ from a proposal distribution $j(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^{(k)})$. Then $\boldsymbol{\gamma}^{(k+1)}$ is set equal to $\boldsymbol{\gamma}^*$ with probability

$$\min\left\{1, \frac{p(\mathbf{Y}|\boldsymbol{\gamma}^*)p(\boldsymbol{\gamma}^*)}{p(\mathbf{Y}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})} \times \frac{j(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)}{j(\boldsymbol{\gamma}^*|\boldsymbol{\gamma})}\right\}, \tag{13}$$

and otherwise $\boldsymbol{\gamma}^{(k+1)}$ remains at $\boldsymbol{\gamma}^{(k)}$. The availability of $p(\mathbf{Y}|\boldsymbol{\gamma})$ was crucial for the rapid calculation of (13). A special case is the Metropolis algorithm with a symmetric random walk on model indicators so that the acceptance ratio is just the Bayes factor for comparing model $\boldsymbol{\gamma}^*$ to model $\boldsymbol{\gamma}^{(k)}$. While the Metropolis algorithm always accepts moves to higher probability models, making it useful for finding the highest probability model, it and other MCMC algorithms occasionally accept moves to models receiving lower probability. This feature allows these algorithms to escape from local modes, unlike greedy search and stepwise methods. Attention quickly focused on the development of better and more efficient proposal distributions $j(\boldsymbol{\gamma}^*|\boldsymbol{\gamma})$, which governed the movements of the algorithm around the model space. Initial implementations of this algorithm, corresponding to different choices of j, included the conjugate version of stochastic search variable selection (SSVS) (George and McCulloch, 1997) and Markov chain Monte Carlo model composition (MC^3) (Raftery, Madigan and Hoeting, 1997).

Just as important, the availability of closed forms for $p(\mathbf{Y}|\boldsymbol{\gamma})$ made it possible to rapidly compute exact posterior odds or Bayes factors for comparison of any two of the sampled models in (12). Such exact values were far more reliable than sample frequency posterior estimates, especially for large model spaces where many of the sampled models would typically be visited only once. Within the set of sampled models, this allowed for exact selection of the modal model, and determination of the extent to which this modal model dominated the other models. Letting S stand for the set of sampled models, exact values for $p(\mathbf{Y}|\boldsymbol{\gamma})$ also allowed for the exact calculation of the renormalized estimates of posterior model probabilities

$$\hat{p}(\boldsymbol{\gamma}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}'\in S} p(\mathbf{Y}|\boldsymbol{\gamma}')p(\boldsymbol{\gamma}')}. \tag{14}$$

Such simulation consistent estimates take full advantage of the information in $p(\mathbf{Y}|\boldsymbol{\gamma})$. For the purpose of model averaging, such $\hat{p}(\boldsymbol{\gamma}|\mathbf{Y})$ can be used instead of $p(\boldsymbol{\gamma}|\mathbf{Y})$ to provide simulation consistent estimates of (5) and (7) and other quantities of interest (Clyde, DeSimone and Parmigiani, 1996; Raftery, Madigan and Hoeting, 1997). Finally, it should also be mentioned that the availability of $p(\mathbf{Y}|\boldsymbol{\gamma})$ also facilitated other viable computational alternatives such as importance sampling for model averaging estimation (Clyde, DeSimone and Parmigiani, 1996).

Of the variety of conjugate parameter prior specifications considered for the normal linear variable selection problem, Zellner's g-prior formulation (Zellner, 1986) has attracted particular attention. Letting $p_{\boldsymbol{\gamma}}$ denote the number of predictor variables in the $\boldsymbol{\gamma}$th subset, this formulation is

$$p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, g) = N_{p_{\boldsymbol{\gamma}}}\{0, g\sigma^2(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\} \tag{15}$$

for a positive hyperparameter g, and

$$p(\beta_0, \sigma^2|\boldsymbol{\gamma}) \propto 1/\sigma^2, \tag{16}$$

where all the predictors have been recentered at 0 to remove dependence on the intercept. For several reasons, this limiting version of the usual normal-inverse Gamma conjugate prior gradually emerged as a default conventional prior of choice. To begin with, it yields rapidly computable closed form expressions for $p(\mathbf{Y}|\boldsymbol{\gamma})$, in part because the prior covariance is proportional to $(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{-1}$, which avoids a ratio-of-determinants calculation. Indeed, the Bayes factor for any model $\boldsymbol{\gamma}$ with respect to the null model (intercept only) has the simple form

$$\begin{aligned} B[\boldsymbol{\gamma}:0] &= (1+g)^{(n-p_{\boldsymbol{\gamma}}-1)/2} \\ &\quad \cdot \big(1+g(1-R_{\boldsymbol{\gamma}}^2)\big)^{-(n-1)/2}, \end{aligned} \tag{17}$$

where $R_{\boldsymbol{\gamma}}^2$ is the usual coefficient of determination. To further reduce computational overhead of computing (13) for MH algorithms, such priors allow for efficient updating routines (George and McCulloch, 1997; Smith and Kohn, 1996). Such priors are also conditionally compatible in the sense that each submodel prior is obtained via a conditioning of the full model prior (Dawid and Lauritzen, 2001). Most important, such priors require only the tuning of a single hyperparameter g, which controls the expected size of the coefficients in $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, thereby facilitating their semiautomatic use. However, a drawback is that model comparisons based on g-priors have an undesirable inconsistency property, as discussed in Berger and Pericchi (2001). For any fixed g, the Bayes factor $B[\boldsymbol{\gamma}:0]$ in (17) goes to $(1+g)^{(n-p_{\boldsymbol{\gamma}}-1)/2}$ as $R_{\boldsymbol{\gamma}}^2$ goes to 1.0. Thus, for a fixed sample size, the Bayes factor is bounded no matter how overwhelmingly the data support $\boldsymbol{\gamma}$ (Berger and Pericchi, 2001). See Berger, Ghosh and Mukhopadhyay (2003) for a discussion of such inconsistency in the context of nonparametric regression.

Turning to model space prior specification, a default choice that has emerged is the independent Bernoulli prior

$$p(\boldsymbol{\gamma}|w) = w^{p_{\boldsymbol{\gamma}}}(1-w)^{p-p_{\boldsymbol{\gamma}}}, \tag{18}$$

which, like the g-prior, is controlled by a single hyperparameter $w \in (0,1)$ (George and McCulloch, 1993, 1997; Raftery, Madigan and Hoeting, 1997). Under this prior, each predictor is independently included in the model with the same probability w. This prior includes the uniform distribution over models, $w = 1/2$, which was initially considered by many as the natural "noninformative" choice. However, in the context of variable selection, the uniform distribution over models induces a Binomial distribution on the model size $p_{\boldsymbol{\gamma}}$, with prior expectation that half of the variables will be included. The more general prior $p(\boldsymbol{\gamma}|w)$ allows for the additional flexibility of controlling w, the expected proportion of predictors in the model. Another useful alternative is to assign a truncated Poisson distribution to the number of components in the model (Denison, Mallick and Smith, 1998b). This can be viewed as a limiting version of $p(\boldsymbol{\gamma}|w)$ for large p small w, and may be an appropriate way to represent prior expectations of sparsity. Elaborations of the Bernoulli prior to handle structured dependence between variables, such as occur with interactions, polynomials, lagged variables or indicator variables, were developed

by Chipman (1996). A limitation of the Bernoulli priors is that they may accumulate too much prior probability in clusters of similar models when there is severe multicollinearity (George, 1999).

Implementation of the Bernoulli g-prior combination requires values for the two hyperparameters g and w. For this purpose, it was quickly realized that setting g arbitrarily large, a typical "noninformative" strategy for estimation problems, could lead to misleading results in the model uncertainty context. Why? Because very large g values can induce the well-known Lindley–Bartlett paradox (Bartlett, 1957), where Bayes factors tend to overwhelmingly favor the null model for all but very extreme parameter estimates. Thus, a variety of default choices with $w = 1/2$ but g no larger than 10,000 were initially recommended on the basis of reasonable performance in simulations and applications (Clyde, DeSimone and Parmigiani, 1996; Smith and Kohn, 1996; George and McCulloch, 1997; Raftery, Madigan and Hoeting, 1997). To shed further light on the effect of different values of g and w, George and Foster (2000) showed that, for fixed values of σ^2, different choices of g and w corresponded exactly to popular model selection criteria, such as the Akaike information criterion (AIC; Akaike, 1973), the Bayesian information criterion (BIC; Schwarz, 1978) and the risk inflation criterion (RIC; Foster and George, 1994), in the sense that the highest posterior model would be the same as that model selected by the criteria. Through simulation studies, Fernández, Ley and Steel (2001) recommended RIC calibrated priors when $n < p^2$ and BIC calibrated priors otherwise. In nonparametric models, such as wavelet regression where $p = n$, there are cases where priors calibrated to BIC have better predictive performance than prior distributions calibrated to RIC, and vice versa (Clyde, Parmigiani and Vidakovic, 1998). It gradually became clear, through simulation studies and asymptotic arguments, that no one default choice for g and w would "perform" well for all contingencies (Fernández, Ley and Steel, 2001; Hansen and Yu, 2001; Berger and Pericchi, 2001).

The essential difficulty of using fixed values for g and w was that different values put different prior weights on model features. For example, small w and large g concentrate the prior on parsimonious models with large coefficients, whereas large w and small g concentrate the prior on saturated models with small coefficients. To avoid the difficulty of preselecting g and w, George and Foster (2000) and Clyde and George (2000) proposed and developed empirical Bayes (EB) methods that used estimates $\hat{g}$ and $\hat{w}$ based on the data. Such methods provided automatic prior specifications and had the computational convenience of the g-prior formulation. Motivated by information theory, Hansen and Yu (2001) developed related approaches that use model specific (local EB) estimates of g. The global EB procedure (one common g in all models) borrows strength from all models in estimating g (Clyde, 2001), but can be difficult to implement in conjunction with stochastic search in high-dimensional problems; the one exception where global EB is easier to implement is orthogonal regression, which arises naturally in the wavelet setting (Clyde and George, 2000).

A natural alternative to these EB methods are fully Bayes (FB) treatments that put priors on w and/or g. Putting a uniform or Beta prior on w induces a Beta-binomial prior on $\boldsymbol{\gamma}$, and putting an inverse Gamma$(1/2, n/2)$ prior on g, as recommended by Zellner and Siow (1980), leads to a multivariate Cauchy prior on $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. Such priors have heavier tails than the Bernoulli g-prior combination and are often recommended from a Bayesian robustness perspective. Such FB approaches, including the use of Strawderman priors $p(g) \propto (1+g)^{-a/2}$ (Strawderman, 1971) that yield closed form marginals with Cauchy-like tails, have been recently investigated (Liang et al., 2003; Cui, 2002; Johnstone and Silverman, 2004; Wang, 2002). For the wavelet regression problem, Johnstone and Silverman (2004) show that empirical Bayes estimation of w coupled with heavy tailed priors for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, such as the Cauchy or double exponential, yields adaptive thresholding rules that yield optimal rates of convergence for various smoothness classes of functions.

4. BEYOND VARIABLE SELECTION UNCERTAINTY

Rapid advances in computational power and MCMC allowed Bayesian treatment of model uncertainty in other classes of problems, in particular, tree models and graphical models. The appeal of these models, as with other hierarchical models, is that they exploit local dependencies (and hence can take advantage of local calculations) to model complex global structures.

4.1 Tree Models

Motivated by the CART formulation of Breiman, Friedman, Olshen and Stone (1984), tree models offer a flexible alternative to additive regression models such as (8) and (11). The basic idea is to partition

the $\mathbf{X}$ values so that the distribution of $\mathbf{Y}$ within each subset of the partition is captured by a (hopefully simple) parametric model. The partition is accomplished by a binary tree T that assigns each observation (y, x) in $(\mathbf{Y}, \mathbf{X})$ to a subset of the partition with simple splitting rules of the form $\{x \in A\}$ or $\{x \notin A\}$. Beginning with a splitting rule at the root node, each x is assigned to one of the terminal nodes of T by a sequence of splitting rules at each of the intermediate nodes. The terminal node of T then associates the observation with a probability distribution for $y|x$.

Letting $T_1, \ldots, T_b$ denote the b terminal nodes of a particular tree T, and letting $p_j(y|x, \theta_j)$ denote the distribution corresponding to T_j, the tree model for each observation can be expressed as

$$M_T\colon\ p(y|x) = \sum_{j=1}^{b} p(y|x, \theta_j)\, I\{x \in T_j\}. \tag{19}$$

For a fixed parametric family of terminal node distributions, the model uncertainty here stems from the choice of a partition tree T. Initial Bayesian treatments of this problem (Buntine, 1992; Chipman, George and McCulloch, 1998; Denison, Mallick and Smith, 1998b) considered simple parametric distributions for $p(y|x, \theta_j)$ such as the Bernoulli or Normal that did not depend on x. More recently, extensions using linear and generalized linear models for $p(y|x, \theta_j)$ have been developed by Chipman, George and McCulloch (2001, 2003). For further references on these and closely related partition models, see the book by Denison et al. (2002).

4.2 Graphical Models

Graphical models (see Jordan, 2004, this issue) provide graph theoretic representations of probability models that greatly facilitate the formulation of multivariate models for complex phenomena. Recent developments concerning model uncertainty have focused on identifying latent graphical structure that encodes conditional independence relationships with the presence or absence of edges connecting variables in the graph. Bayesian treatments of model selection and accounting for model uncertainty for discrete graphical models, such as directed acyclic graphs, were considered by Madigan and Raftery (1994), Madigan and York (1995) and Dellaportas and Forster (1999). For multivariate Gaussian data, the model selection problem can be viewed as a problem in covariance selection (Dempster, 1972), where zeros in the precision matrix (the inverse covariance matrix) encode various conditional independence specifications (Giudici and Green, 1999; Smith and Kohn, 2002; Wong, Carter and Kohn, 2003). With decomposable graphical models and conjugate priors, explicit marginal likelihoods are available, allowing the use of MH to stochastically explore the model space. However, even with a moderate number of variables, the model space is astronomical in size so that efficient proposal distributions are needed. Extensions to nondecomposable models add additional complexities as marginal likelihoods are not available and potentially high-dimensional integrals must be approximated (Dellaportas, Giudici and Roberts, 2003; Roverato, 2002; Atay-Kayis and Massam, 2002). This is typical of many other model selection and variable selection problems where closed form marginal likelihoods and Bayes factors are unavailable.

5. ESTIMATING BAYES FACTORS AND MARGINALS

While the class of models that permits analytically tractable marginal likelihoods covers a wide range of applications, generalized linear models, hierarchical mixed or random effects models, nondecomposable Gaussian graphical models, for example, do not allow closed form expressions for marginal likelihoods. Methods based on computing the marginal likelihoods for each model using Monte Carlo methods of integration, such as importance sampling, are often difficult to implement in moderate to high-dimensional models. Such models, however, are highly amenable to MCMC methods for sampling from model specific posteriors for parameters, leading to a range of approaches to estimate either marginals or Bayes factors using the output from MCMC. These methods can be broken down into two groups; those that involve running a single chain for each model and indirectly estimating marginal likelihoods or Bayes factors from the output, or methods based on constructing one Markov chain that samples from the joint parameter–model space. Han and Carlin (2001) provide a recent comparison of several approaches that have broad applicability for model selection.

5.1 Single Chain Methods

Chib (1995) proposed a method of estimating marginal likelihood based on inverting the identity behind Bayes's theorem,

$$p(\mathbf{Y}|\mathcal{M}_k) = \frac{p(\mathbf{Y}|\boldsymbol{\theta}_k, \mathcal{M}_k)\, p(\boldsymbol{\theta}_k|\mathcal{M}_k)}{p(\boldsymbol{\theta}_k|\mathbf{Y}, \mathcal{M}_k)}, \tag{20}$$

which holds for any $\boldsymbol{\theta}_k$, in particular for $\boldsymbol{\theta}_k^*$, a *fixed* point of high probability or the MLE. When

$p(\boldsymbol{\theta}_k|\mathbf{Y}, \mathcal{M}_k)$ is unavailable, but closed form full conditionals are available, Chib (1995) uses the Gibbs sampler to construct an estimator $\hat{p}(\boldsymbol{\theta}_k|\mathbf{Y}, \mathcal{M}_k)$ to use in estimating the marginal likelihood (20). Chib's method for constructing $\hat{p}(\boldsymbol{\theta}_k|\mathbf{Y}, \mathcal{M}_k)$ involves partitioning $\boldsymbol{\theta}_k$ into blocks of parameters each having closed form full conditional distributions (given the other blocks of parameters). In the case of two blocks, $\boldsymbol{\theta}_k = (\boldsymbol{\theta}_{k1}, \boldsymbol{\theta}_{k2})$, the method is straightforward to implement; however, extensions to B blocks require an additional $(B-1)$ Gibbs samplers (per model) and extensive bookkeeping. More recently Chib and Jeliazkov (2001) extended the approach to Metropolis–Hastings algorithms by exploiting the detailed balance of MH algorithms. When $\boldsymbol{\theta}_k$ is generated in more than one block, multiple chains per model must be executed to estimate $p(\boldsymbol{\theta}_k^*|\mathbf{Y})$. While theoretically the methods of Chib (1995) and Chib and Jeliazkov (2001) can be applied with any MCMC scheme, the extra sampling and bookkeeping may limit practical application to models where efficient MCMC algorithms exist for low-dimensional blocked samplers.

Importance sampling (IS) has a long history of use in estimating normalizing constants or ratios of normalizing constants as in Bayes factors. However, its efficiency depends critically on the choice of proposal distributions and related IS weights. For low dimensional variable selection problems, simple importance sampling using t-densities with location and scale parameters based on the output of the Gibbs sampler or even based on MLEs can often be very efficient and should not be overlooked. Bridge sampling (Meng and Wong, 1996), path sampling (Gelman and Meng, 1998), ratio importance sampling (RIS; Chen and Shao, 1997) build on standard importance sampling (see also Andrieu, Doucet and Robert, 2004, this issue). While RIS, with the optimal choice of proposal distribution, is theoretically more efficient than bridge or path sampling, the optimal proposal distribution depends on the unknown Bayes factor. Chen, Shao and Ibrahim (2000) discuss relationships among these methods, and extensions to models with differing dimensions. For variable selection, Ibrahim, Chen and MacEachern (1999) and Chen, Ibrahim and Yiannoutsos (1999) combine RIS with the importance weighted marginal density estimator (IWMDE; Chen, 1994) to estimate Bayes factors of submodels $\mathcal{M}_k$ of the full model. This can be viewed as an estimate of the generalized Savage–Dickey density ratio (Verdinelli and Wasserman, 1995) for Bayes factors. The key feature is that the method only requires MCMC output from the posterior distribution for the full model to estimate all Bayes factors.

The above methods require an exhaustive list of models, but can be combined with some additional search strategy to calculate Bayes factors for a subset of models. The "leaps and bounds" algorithm of Furnival and Wilson (1974) has been adapted to a wide variety of settings by Volinsky, Madigan, Raftery and Kronmal (1997), and can be used to rapidly identify a subset of models for further evaluation. Alternatively, single chain methods, such as reversible jump, can be used for both search and estimation of model probabilities.

5.2 MCMC over Combined Model–Parameter Spaces

Single chain methods required creating a Markov chain over a fixed dimensional space as in the product space search of Carlin and Chib (1995), or using dimension matching at each iteration as in reversible jump MCMC (RJ-MCMC; Green, 1995). Unlike the product-space and single-chain-per-model approaches, RJ-MCMC and variations that sample over the model space and parameter space jointly do not require exhaustive enumeration of the model space and theoretically can be used in moderate and large dimensional problems. The basic iteration step in RJ-MCMC algorithms can be described as follows and applies to extremely general model selection problems.

Given the current state $(\boldsymbol{\theta}_k, \mathcal{M}_k)$:

1. Propose a jump to a new model $\mathcal{M}_j$, $j(\mathcal{M}_j|\mathcal{M}_k, \mathbf{Y})$ given the current model $\mathcal{M}_k$.
2. Generate a vector $\mathbf{u}$ from a continuous distribution $q(\mathbf{u}|\boldsymbol{\theta}_k, \mathcal{M}_k, \mathcal{M}_j, \mathbf{Y})$.
3. Set $(\boldsymbol{\theta}_j, \mathbf{u}^*) = g(\boldsymbol{\theta}_k, \mathbf{u}; \mathcal{M}_k, \mathcal{M}_j)$, where g is a bijection between $(\boldsymbol{\theta}_k, \mathbf{u})$ and $(\boldsymbol{\theta}_j, \mathbf{u}^*)$ and the lengths of $\mathbf{u}$ and $\mathbf{u}^*$ satisfy $p_{\mathcal{M}_k} + \dim(\mathbf{u}) = p_{\mathcal{M}_j} + \dim(\mathbf{u}^*)$, where $p_{\mathcal{M}_k}$ and $p_{\mathcal{M}_j}$ are the dimensions of $\mathcal{M}_k$ and $\mathcal{M}_j$, respectively.
4. Accept the proposed move to $(\boldsymbol{\theta}_j, \mathcal{M}_j)$ with probability

$$\alpha = \min\left\{1, \frac{p(\mathbf{Y}|\boldsymbol{\theta}_j, \mathcal{M}_j)p(\boldsymbol{\theta}_j|\mathcal{M}_j)p(\mathcal{M}_j)}{p(\mathbf{Y}|\boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\theta}_k|\mathcal{M}_k)p(\mathcal{M}_k)} \cdot \frac{j(\mathcal{M}_k|\mathcal{M}_j, \mathbf{Y})q(\mathbf{u}^*|\boldsymbol{\theta}_j, \mathcal{M}_j, \mathcal{M}_k, \mathbf{Y})}{j(\mathcal{M}_j|\mathcal{M}_k, \mathbf{Y})q(\mathbf{u}|\boldsymbol{\theta}_k, \mathcal{M}_k, \mathcal{M}_j, \mathbf{Y})} \cdot \left|\frac{\partial g(\boldsymbol{\theta}_k, \mathbf{u}; \mathcal{M}_k, \mathcal{M}_j)}{\partial(\boldsymbol{\theta}_k, \mathbf{u})}\right|\right\}. \tag{21}$$

The introduction of the variables $\mathbf{u}$ and $\mathbf{u}^*$ ensures that the numerator and denominator in the acceptance ratio

are all defined with respect to a common measure, so that at each iteration (locally) the dimensions of the two augmented spaces are equal. The key to implementing efficient RJ-MCMC algorithms involves constructing model jumping proposals j, efficient proposals for $\mathbf{u}$ and an appropriate function g mapping between the two models. These often have to be tailored to each specific class of problems and may require significant tuning. Relationships of RJ-MCMC and MH–Gibbs sampling in the linear model setting are discussed in Clyde (1999) and Godsill (2001). Recent papers by Dellaportas, Forster and Ntzoufras (2002), Brooks, Giudici and Roberts (2003), Godsill (2001) and Green (2003) discuss variations of RJ-MCMC algorithms and construction of efficient or automatic proposal distributions.

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

As with the MH methods in linear models, estimates of model probabilities using Monte Carlo frequencies of models from RJ-MCMC may be very slow to converge to $p(\boldsymbol{\gamma}|\mathbf{Y})$. While perhaps less important for model averaging than say model selection, construction of efficient proposal distributions and more efficient estimates of Bayes factors and marginal likelihoods given the output are still critical areas for future developments. Using RJ-MCMC for search only and alternative approaches for estimating marginal likelihoods, such as the Laplace approximation (Tierney and Kadane, 1986) or the Metropolized–Laplace estimators (DiCiccio, Kass, Raftery and Wasserman, 1997; Lewis and Raftery, 1997), can provide more accurate results for model selection. Sampling without replacement from the model space (Clyde, 1999) using adaptive proposals is another alternative for model search and appears to perform well for variable selection.

5.3 Default Bayes Factors

Despite tremendous progress in Monte Carlo methods, significant effort is required to implement Monte Carlo methods for estimating Bayes factors. As a result, the simplicity of BIC,

$$B[\mathcal{M}_k : \mathcal{M}_j]_{\mathrm{BIC}} = \frac{p(\mathbf{Y}|\hat{\boldsymbol{\theta}}_k)}{p(\mathbf{Y}|\hat{\boldsymbol{\theta}}_j)} n^{(p_{\mathcal{M}_j} - p_{\mathcal{M}_k})/2}, \tag{22}$$

has made it popular as an approximation to Bayes factors (Kass and Raftery, 1995), as it requires just the MLE of $\boldsymbol{\theta}$ under each model. In combination with deterministic or stochastic search, BIC provides a default method for approximating model probabilities and is appealing in practical applications with many models and/or where conventional prior specification is difficult (Hoeting et al., 1999). The software designed for use in Hoeting et al. (1999) [as well as other programs and articles on Bayesian model averaging (BMA)] can be found at the BMA web page (http://www.research.att.com/∼volinsky/bma.html). One of the difficulties with using BIC, however, is determining the effective sample size n in nonindependent settings, such as hierarchical models (Pauler, 1998; Pauler, Wakefield and Kass, 1999). BIC is also not appropriate in problems where the number of parameters increases with the sample size or other irregular asymptotics prevails (Berger, Ghosh and Mukhopadhyay, 2003).

In addition to the concerns over the general applicability and accuracy of BIC, the overwhelming need for objective Bayesian approaches for model selection has led to a wealth of new procedures for obtaining "default" Bayes factors, such as intrinsic Bayes factors (IBF; Berger and Pericchi, 1996a, b, 1998), fractional Bayes factors (FBF; O'Hagan, 1995) and expected posterior (EP) prior (Pérez and Berger, 2000). Berger and Pericchi (2001) review and contrast these methods with BIC and conventional prior specifications in the context of linear models. It is well known that marginal likelihoods constructed using improper priors lead to indeterminacies of Bayes factors and posterior model probabilities. IBFs and FBFs use the idea of "training" samples to convert an improper prior (reference priors are recommended) into a proper posterior for $\boldsymbol{\theta}_k$. In the case of IBFs, a subset of the data is used as a training sample, while with FBFs a fraction b/n of the likelihood is used. This proper distribution is then used as a prior to define the Bayes factors based on the remaining part of the data. While the Bayes factors do not depend on any arbitrary scaling in the

improper priors, they do depend on the choice of training samples. In the case of IBFs, this dependency on the training sample is eliminated by "averaging" over all possible training samples. Two popular choices include the arithmetic IBF (AIBF), defined by arithmetic average of IBFs over training samples, and the median IBF (MIBF), which is the median of the IBFs over all minimal training samples. With more than two models under consideration, IBFs are not coherent in the sense that $B[i:j] \neq B[i:k]/B[k:j]$; nevertheless they can be used to define *formal* posterior model probabilities (Casella and Moreno, 2002).

The EP prior also uses the idea of taking training samples $\mathbf{Y}^*$ from a marginal distribution $m(\mathbf{Y}^*)$. As with the IBF approach, the training sample is used to convert an improper prior distribution into a proper posterior distribution given $\mathbf{Y}^*$; the expectation of the resulting distribution with respect to $m(\mathbf{Y}^*)$ leads to the expected posterior prior, which can then be used to construct objective Bayes factors. While subjective distributions for $m(\mathbf{Y}^*)$ are of course allowable, a default choice can be obtained by sampling from the empirical distribution of the data. Like the IBF and FBF, there is no problem of indeterminacies in the definition of Bayes factors. The EP priors are also automatically compatible, a feature that may be difficult to achieve with nonnested models.

Modulo computation of the Bayes factors themselves, these default approaches have wide applicability, particularly in nonnested models, or where conventional prior distributions are unavailable. Many of the approaches lead to an "intrinsic" prior which can be contrasted with conventional priors. In linear models, the intrinsic priors associated with AIBFs behave like a mixture of multivariate Cauchy distributions. Recently, Casella and Moreno (2002) have explored intrinsic priors for Bayesian model selection in linear models and have developed algorithms for computation and search in moderate to high dimensional problems. EP priors also show promise for more complicated problems in that they are amenable to MCMC sampling and hence potentially can be combined with other methods for computing Bayes factors and model probabilities.

6. DECISION THEORETIC CONSIDERATIONS

The key object provided by the Bayesian approach is the posterior quantification of postdata uncertainty. Whether to proceed by model selection or model averaging is determined by additional considerations that can be formally motivated by decision theoretic considerations (Gelfand, Dey and Chang, 1992; Bernardo and Smith, 1994). Letting $u(a, \Delta)$ be the utility or negative loss of action a given the unknown of interest Δ, the optimal a maximizes the posterior expected utility

$$\int u(a, \Delta) p(\Delta|\mathbf{Y})\, d\Delta, \tag{23}$$

where $p(\Delta|\mathbf{Y})$ is the predictive distribution of Δ given $\mathbf{Y}$ under the full three stage model specification. For example, highest posterior model selection corresponds to maximizing 0–1 utility for a correct selection. The model averaged point prediction $E(\mathbf{Y_f}|\mathbf{Y})$ corresponds to minimizing quadratic loss with respect to the actual future value $\mathbf{Y_f}$. The predictive distribution $p(\mathbf{Y_f}|\mathbf{Y})$ minimizes Kullback–Leibler loss with respect to the actual predictive distribution $p(\mathbf{Y_f}|\boldsymbol{\theta}_k, \boldsymbol{\mathcal{M}}_k)$. Model selection can also be motivated with these latter utility functions by restricting the action space to selection. For example, for such a restriction, Barbieri and Berger (2004) show that, for sequences of nested models, the median posterior model minimizes quadratic predictive loss. San Martini and Spezzaferri (1984) investigate selection rules that maximize posterior weighted logarithmic divergence.

Several authors have proposed Bayesian model selection approaches that use parameter priors $p(\boldsymbol{\theta}_k|\boldsymbol{\mathcal{M}}_k)$ but entirely avoid model space priors $p(\boldsymbol{\mathcal{M}}_1), \ldots, p(\boldsymbol{\mathcal{M}}_K)$. Such an approach using maximum utility (23) can be used where $p(\Delta|\mathbf{Y})$ is the posterior distribution of Δ under an all-encompassing model, that is, a model under which every other model is nested. In one of the earliest papers on Bayesian variable selection, Lindley (1968) developed such an approach where costs for including variables were included in the utility function and the encompassing model was the model with all variables included. This approach was extended to multivariate regression by Brown, Fearn and Vannucci (1999). For some other Bayesian selection approaches that avoid the model space prior see Gelfand and Ghosh (1998), Draper and Fouskakis (2000) and Dupuis and Robert (2003).

Another interesting modification of the decision theory setup is the so-called M-open framework under which the "true" model is not any one of the $\boldsymbol{\mathcal{M}}_k$ under consideration, a commonly held perspective in many applications. One way of incorporating this aspect into a utility analysis is by using a cross validation training sample estimate of the actual predictive density in place of $p(\Delta|\mathbf{Y})$; see Bernardo and Smith (1994), Berger and Pericchi (1996b), Key, Pericchi and Smith (1999) and Marriott, Spencer and Pettitt (2001).

7. FUTURE DIRECTIONS

The Bayesian treatment of model uncertainty, coupled with advances in posterior search and computation, has led to an explosion of research in model selection and model averaging. To illustrate the rapid evolution of these methods, we have described the highlights of some of these developments, but due to space limitations have left out much more. What is clear, however, is that the evolution and impact of these Bayesian methods is far from over. New model uncertainty challenges continue to arise in a wide variety of areas, including bioinformatics, data-mining, "inverse" problem analysis, nonparametric function estimation, overcomplete representation and spatial–temporal modeling. New computational advances such as automatic RJ-MCMC (Green, 2003) and adaptive MCMC samplers (see Andrieu, Doucet and Robert, 2004, this issue) portend powerful new approaches for exploration of model space posteriors. Continuing developments in objective Bayesian methodology hold the promise of improved automatic prior specifications and a greater understanding of the operating characteristics of these methods. The potential of Bayesian methods for model uncertainty has only begun to be realized.

ACKNOWLEDGMENTS

This material was based upon work supported by NSF Grants DMS-97-33013, DMS-01-30819 and DMS-01-12069. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

ABRAMOVICH, F., SAPATINAS, T. and SILVERMAN, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60** 725–749.

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (B. Petrov and F. Csáki, eds.) 267–281. Akadémiai Kiadó, Budapest.

ANDRIEU, C., DOUCET, A. and ROBERT, C. (2004). Computational advances for and from Bayesian analysis. *Statist. Sci.* **19** 118–127.

ATAY-KAYIS, A. and MASSAM, H. (2002). The marginal likelihood for decomposable and nondecomposable graphical Gaussian models. Technical report, Dept. Mathematics, York Univ.

BARBIERI, M. M. and BERGER, J. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897.

BARTLETT, M. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika* **44** 533–534.

BERGER, J. O., GHOSH, J. K. and MUKHOPADHYAY, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *J. Statist. Plann. Inference* **112** 241–258.

BERGER, J. O. and PERICCHI, L. R. (1996a). The intrinsic Bayes factor for linear models. In *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 25–44. Oxford Univ. Press.

BERGER, J. O. and PERICCHI, L. R. (1996b). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91** 109–122.

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

BERGER, J. O. and PERICCHI, L. (2001). Objective Bayesian methods for model selection: Introduction and comparison (with discussion). In *Model Selection* (P. Lahiri, ed.) 135–207. IMS, Beachwood, OH.

BERNARDO, J. M. and SMITH, A. F. M. (1994). *Bayesian Theory*. Wiley, New York.

BESAG, J. and GREEN, P. J. (1993). Spatial statistics and Bayesian computation (with discussion). *J. Roy. Statist. Soc. Ser. B* **55** 25–37.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

BROOKS, S. P., GIUDICI, P. and ROBERTS, G. O. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 3–55.

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.

BROWN, P. J., VANNUCCI, M. and FEARN, T. (1998). Multivariate Bayesian variable selection and prediction. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60** 627–641.

BROWN, P. J., VANNUCCI, M. and FEARN, T. (2002). Bayes model averaging with selection of regressors. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 519–536.

BUNTINE, W. (1992). Learning classification trees. *Statist. Comput.* **2** 63–73.

CARLIN, B. P. and CHIB, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B* **57** 473–484.

CASELLA, G. and MORENO, E. (2002). Objective Bayes variable selection. Technical Report 2002-023, Dept. Statistics, Univ. Florida.

CHEN, M.-H. (1994). Importance-weighted marginal Bayesian posterior density estimation. *J. Amer. Statist. Assoc.* **89** 818–824.

CHEN, M.-H., IBRAHIM, J. G., SHAO, Q.-M. and WEISS, R. E. (2003). Prior elicitation for model selection and estimation in generalized linear mixed models. *J. Statist. Plann. Inference* **111** 57–76.

CHEN, M.-H., IBRAHIM, J. G. and YIANNOUTSOS, C. (1999). Prior elicitation, variable selection and Bayesian computation for logistic regression models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 223–242.

CHEN, M.-H. and SHAO, Q.-M. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.* **25** 1563–1594.

CHEN, M.-H., SHAO, Q.-M. and IBRAHIM, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, New York.

CHIB, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90** 1313–1321.

CHIB, S. and JELIAZKOV, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *J. Amer. Statist. Assoc.* **96** 270–281.

CHIPMAN, H. A. (1996). Bayesian variable selection with related predictors. *Canad. J. Statist.* **24** 17–36.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (1998). Bayesian CART model search (with discussion). *J. Amer. Statist. Assoc.* **93** 935–960.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2001). The practical implementation of Bayesian model selection (with discussion). In *Model Selection* (P. Lahiri, ed.) 65–134. IMS, Beachwood, OH.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2003). Bayesian treed generalized linear models (with discussion). In *Bayesian Statistics 7* (J. M. Bernardo, M. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 85–103. Oxford Univ. Press.

CHIPMAN, H. A., KOLACZYK, E. D. and MCCULLOCH, R. E. (1997). Adaptive Bayesian wavelet shrinkage. *J. Amer. Statist. Assoc.* **92** 1413–1421.

CLYDE, M. (1999). Bayesian model averaging and model search strategies (with discussion). In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 157–185. Oxford Univ. Press.

CLYDE, M. (2001). Discussion of "The practical implementation of Bayesian model selection," by H. A. Chipman, E. I. George and R. E. McCulloch. In *Model Selection* (P. Lahiri, ed.) 117–124. IMS, Beachwood, OH.

CLYDE, M., DESIMONE, H. and PARMIGIANI, G. (1996). Prediction via orthogonalized model mixing. *J. Amer. Statist. Assoc.* **91** 1197–1208.

CLYDE, M. and GEORGE, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62** 681–698.

CLYDE, M., PARMIGIANI, G. and VIDAKOVIC, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85** 391–401.

CUI, W. (2002). Variable selection: Empirical Bayes vs. fully Bayes. Ph.D. dissertation, Dept. Management Science and Information Systems, Univ. Texas, Austin.

DAWID, A. and LAURITZEN, S. (2001). Compatible prior distributions. In *Bayesian Methods with Applications to Science, Policy, and Official Statistics, Selected Papers from ISBA 2000*: *The Sixth World Meeting of the International Society for Bayesian Analysis* (E. I. George, ed.) 109–118. Eurostat, Luxembourg.

DELLAPORTAS, P. and FORSTER, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86** 615–633.

DELLAPORTAS, P., FORSTER, J. J. and NTZOUFRAS, I. (2002). On Bayesian model and variable selection using MCMC. *Statist. Comput.* **12** 27–36.

DELLAPORTAS, P., GIUDICI, P. and ROBERTS, G. (2003). Bayesian inference for nondecomposable graphical Gaussian models. *Sankhyā Ser. A.* **65** 43–55.

DEMPSTER, A. M. (1972). Covariance selection. *Biometrics* **28** 157–175.

DENISON, D. G. T., HOLMES, C., MALLICK, B. K. and SMITH, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley, New York.

DENISON, D. G. T., MALLICK, B. K. and SMITH, A. F. M. (1998a). Automatic Bayesian curve fitting. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60** 333–350.

DENISON, D. G. T., MALLICK, B. K. and SMITH, A. F. M. (1998b). A Bayesian CART algorithm. *Biometrika* **85** 363–377.

DENISON, D. G. T., MALLICK, B. K. and SMITH, A. F. M. (1998c). Bayesian MARS. *Statist. Comput.* **8** 337–346.

DICICCIO, T. J., KASS, R. E., RAFTERY, A. and WASSERMAN, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *J. Amer. Statist. Assoc.* **92** 903–915.

DRAPER, D. (1995). Assessment and propagation of model uncertainty (with discussion). *J. Roy. Statist. Soc. Ser. B* **57** 45–97.

DRAPER, D. and FOUSKAKIS, D. (2000). A case study of stochastic optimization in health policy: Problem formulation and preliminary results. *J. Global Optimization* **18** 399–416.

DUPUIS, J. A. and ROBERT, C. P. (2003). Variable selection in qualitative models via an entropic explanatory power. *J. Statist. Plann. Inference* **111** 77–94.

FERNÁNDEZ, C., LEY, E. and STEEL, M. F. (2001). Benchmark priors for Bayesian model averaging. *J. Econometrics* **100** 381–427.

FOSTER, D. P. and GEORGE, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975.

FURNIVAL, G. M. and WILSON, R. W., JR. (1974). Regression by leaps and bounds. *Technometrics* **16** 499–511.

GEISSER, S. (1993). *Predictive Inference. An Introduction*. Chapman and Hall, London.

GELFAND, A. E., DEY, D. K. and CHANG, H. (1992). Model determination using predictive distributions, with implementation via sampling-based methods (with discussion). In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 147–167. Oxford Univ. Press.

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

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

GELMAN, A. and MENG, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185.

GEORGE, E. I. (1999). Discussion of "Bayesian model averaging and model search strategies," by M. Clyde. In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 175–177. Oxford Univ. Press.

GEORGE, E. I. (2000). The variable selection problem. *J. Amer. Statist. Assoc.* **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 MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–374.

GEORGE, E. I., MCCULLOCH, R. and TSAY, R. (1996). Two approaches to Bayesian model selection with applications. In *Bayesian Analysis in Statistics and Econometrics*: *Essays in Honor of Arnold Zellner* (D. Berry, K. Chaloner and J. Geweke, eds.) 339–348. Wiley, New York.

GEWEKE, J. (1996). Variable selection and model comparison in regression. In *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 609–620. Oxford Univ. Press.

GIUDICI, P. and GREEN, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika* **86** 785–801.

GODSILL, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *J. Comput. Graph. Statist.* **10** 230–248.

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

GREEN, P. J. (2003). Trans-dimensional Markov chain Monte Carlo. In *Highly Structured Stochastic Systems* (P. J. Green, N. L. Hjort and S. Richardson, eds.) 179–206. Oxford Univ. Press.

HAN, C. and CARLIN, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *J. Amer. Statist. Assoc.* **96** 1122–1132.

HANSEN, M. H. and KOOPERBERG, C. (2002). Spline adaptation in extended linear models (with discussion). *Statist. Sci.* **17** 2–51.

HANSEN, M. H. and YU, B. (2001). Model selection and the principle of minimum description length. *J. Amer. Statist. Assoc.* **96** 746–774.

HODGES, J. S. (1987). Uncertainty, policy analysis and statistics (with discussion). *Statist. Sci.* **2** 259–275.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial (with discussion). *Statist. Sci.* **14** 382–417. (Corrected version available at http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.)

HOETING, J. A., RAFTERY, A. E. and MADIGAN, D. (2002). Bayesian variable and transformation selection in linear regression. *J. Comput. Graph. Statist.* **11** 485–507.

IBRAHIM, J. G., CHEN, M.-H. and MACEACHERN, S. N. (1999). Bayesian variable selection for proportional hazards models. *Canad. J. Statist.* **27** 701–717.

IBRAHIM, J. G., CHEN, M.-H. and RYAN, L. M. (2000). Bayesian variable selection for time series count data. *Statist. Sinica* **10** 971–987.

JOHNSTONE, I. and SILVERMAN, B. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649.

JORDAN, M. I. (2004). Graphical models. *Statist. Sci.* **19** 140–155.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

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

KOHN, R., MARRON, J. S. and YAU, P. (2000). Wavelet estimation using Bayesian basis selection and basis averaging. *Statist. Sinica* **10** 109–128.

LEAMER, E. E. (1978a). Regression selection strategies and revealed priors. *J. Amer. Statist. Assoc.* **73** 580–587.

LEAMER, E. E. (1978b). *Specification Searches*: *Ad Hoc Inference with Nonexperimental Data*. Wiley, New York.

LEWIS, S. M. and RAFTERY, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace–Metropolis estimator. *J. Amer. Statist. Assoc.* **92** 648–655.

LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. and BERGER, J. (2003). Gaussian hyper-geometric and other mixtures of g-priors for Bayesian variable selection. Technical report, Statistical and Applied Mathematical Sciences Inst., Research Triangle Park, NC.

LIANG, F., TRUONG, Y. and WONG, W. H. (2001). Automatic Bayesian model averaging for linear regression and applications in Bayesian curve fitting. *Statist. Sinica* **11** 1005–1029.

LINDLEY, D. V. (1968). The choice of variables in multiple regression (with discussion). *J. Roy. Statist. Soc. Ser. B* **30** 31–66.

MADIGAN, D. and RAFTERY, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89** 1535–1546.

MADIGAN, D. and YORK, J. (1995). Bayesian graphical models for discrete data. *Internat. Statist. Rev.* **63** 215–232.

MARRIOTT, J. M., SPENCER, N. M. and PETTITT, A. N. (2001). A Bayesian approach to selecting covariates for prediction. *Scand. J. Statist.* **28** 87–97.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

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

MILLER, A. J. (2002). *Subset Selection in Regression*, 2nd ed. Chapman and Hall, London.

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression (with discussion). *J. Amer. Statist. Assoc.* **83** 1023–1032.

MÜLLER, P. and QUINTANA, F. A. (2004). Nonparametric Bayesian data analysis. *Statist. Sci.* **19** 95–110.

NTZOUFRAS, I., DELLAPORTAS, P. and FORSTER, J. J. (2003). Bayesian variable and link determination for generalised linear models. *J. Statist. Plann. Inference* **111** 165–180.

NTZOUFRAS, I., FORSTER, J. J. and DELLAPORTAS, P. (2000). Stochastic search variable selection for log-linear models. *J. Statist. Comput. Simulation* **68** 23–37.

O'HAGAN, A. (1995). Fractional Bayes factors for model comparison (with discussion). *J. Roy. Statist. Soc. Ser. B* **57** 99–138.

PAULER, D. K. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika* **85** 13–27.

PAULER, D. K., WAKEFIELD, J. C. and KASS, R. E. (1999). Bayes factors and approximations for variance component models. *J. Amer. Statist. Assoc.* **94** 1242–1253.

PÉREZ, J. and BERGER, J. O. (2000). Expected posterior prior distributions for model selection. Technical Report 00-08, Institute of Statistics and Decision Sciences, Duke Univ.

RAFTERY, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* **83** 251–266.

RAFTERY, A. E., MADIGAN, D. and HOETING, J. A. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.* **92** 179–191.

RAFTERY, A. E., MADIGAN, D. and VOLINSKY, C. T. (1996). Accounting for model uncertainty in survival analysis improves predictive performance. In *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 323–349. Oxford Univ. Press.

ROVERATO, A. (2002). Hyper inverse Wishart distribution for nondecomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Statist.* **29** 391–411.

SAN MARTINI, A. and SPEZZAFERRI, F. (1984). A predictive model selection criterion. *J. Roy. Statist. Soc. Ser. B* **46** 296–303.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SHIVELY, T. S., KOHN, R. and WOOD, S. (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior (with discussion). *J. Amer. Statist. Assoc.* **94** 777–806.

SMITH, A. F. M. and ROBERTS, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. Roy. Statist. Soc. Ser. B* **55** 3–23.

SMITH, M. and KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75** 317–343.

SMITH, M. and KOHN, R. (1997). A Bayesian approach to nonparametric bivariate regression. *J. Amer. Statist. Assoc.* **92** 1522–1535.

SMITH, M. and KOHN, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *J. Amer. Statist. Assoc.* **97** 1141–1153.

STEWART, L. and DAVIS, W. W. (1986). Bayesian posterior distributions over sets of possible models with inferences computed by Monte Carlo integration. *The Statistician* **35** 175–182.

STRAWDERMAN, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.* **42** 385–388.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22** 1701–1728.

TIERNEY, L. and KADANE, J. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86.

VERDINELLI, I. and WASSERMAN, L. (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *J. Amer. Statist. Assoc.* **90** 614–618.

VOLINSKY, C. T., MADIGAN, D., RAFTERY, A. E. and KRONMAL, R. A. (1997). Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke. *Appl. Statist.* **46** 433–448.

WAKEFIELD, J. and BENNETT, J. (1996). The Bayesian modeling of covariates for population pharmacokinetic models. *J. Amer. Statist. Assoc.* **91** 917–927.

WANG, X. (2002). Bayesian variable selection for generalized linear models. Ph.D. dissertation, Dept. Management Science and Information Systems, Univ. Texas, Austin.

WOLFE, P. J., GODSILL, S. J. and NG, W.-J. (2004). Bayesian variable selection and regularisation for time–frequency surface estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* To appear.

WONG, F., CARTER, C. and KOHN, R. (2003). Efficient estimation of covariance selection models. *Biometrika* **90** 809–830.

WOOD, S. and KOHN, R. (1998). A Bayesian approach to robust binary nonparametric regression. *J. Amer. Statist. Assoc.* **93** 203–213.

WOOD, S., KOHN, R., SHIVELY, T. and JIANG, W. (2002). Model selection in spline nonparametric regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 119–139.

ZELLNER, A. (1984). Posterior odds ratios for regression hypotheses: General considerations and some specific results. In *Basic Issues in Econometrics* (A. Zellner, ed.) 275–305. Univ. Chicago Press.

ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques*: *Essays in Honor of Bruno de Finetti* (P. K. Goel and A. Zellner, eds.) 233–243. North-Holland, Amsterdam.

ZELLNER, A. and SIOW, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics. Proceedings of the First Valencia International Meeting Held in Valencia (Spain)* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 585–603. Valencia Univ. Press.