DISCUSSION

M. Clyde

Duke University

I would like to begin by thanking the authors for their many contributions to Bayesian model selection and for providing an excellent summary of the growing body of literature on Bayesian model selection and model uncertainty. The development of computational tools such as the Gibbs sampler and Markov chain Monte Carlo approaches, has led to an explosion in Bayesian approaches for model selection. On the surface, Bayesian model averaging (BMA) and model selection is straightforward to implement: one specifies the distribution of the data, and the prior probabilities of models and model specific parameters; Bayes theorem provides the rest. As the authors point out the two major challenges confronting the practical implementation of Bayesian model selection are choosing prior distributions and calculating posterior distributions. In my experience, I have found that this is especially true in high dimensional problems, such as wavelet regression or non-parametric models, where subjective elicitation of prior distributions is practically infeasible and enumeration of the number of potential models is intractable (Clyde et al. 1998; Clyde and George 2000).

Choice of Prior Distributions

The specification of prior distributions is often broken down into two parts: (1) elicitation of distributions for parameters specific to each model, such as the distribution for regression coefficients in linear models, $p(\beta|\gamma, \sigma^2)$, and (2) selection of a prior distribution over models $p(\gamma)$. For high dimensional problems one cannot specify the prior probability of each γ directly, and practical implementations of Bayesian selection have usually made prior assumptions that the presence or absence of a variable is independent of the presence or absence of other variables. As a special case of this, the uniform prior distribution over models is appealing in that posterior probabilities of models depend only on the likelihood. However, this prior distribution may not be sensible for model averaging when there are groups of similar variables and does not provide the proper "dilution" of posterior mass over similar models (see Clyde 1999; Hoeting et al. 1999), and discussion therein (George 1999; 1999a). In this regard, uniform and independent prior distributions must be used carefully with highly correlated explanatory variables and special consideration should be given to constructing the model space. Even with

Merlise Clyde is Associate Professor, Institute of Statistics and Decision Sciences, Duke University, Durham NC 27708-0251, U.S.A. email: clyde@stat.Duke.EDU.

uniform priors on models, the posterior distribution over models naturally penalizes adding redundant variables, however, this may not be enough to lead to the proper rate of dilution with nearly redundant variables. One approach to constructing dilution prior distributions is to start with a uniform prior over models and use imaginary training data to construct a posterior distribution for γ based on the training data; this posterior would become the prior distribution for γ for the observed data. Selection of the training data and hyperparameters are non-trivial issues, however, this approach would likely provide better dilution properties than starting with an independent prior. Clearly, construction of prior distributions for models that capture similarities among variables in a sensible way is a difficult task and one that needs more exploration.

For (1), by far the most common choice is a normal prior distribution for β, such as in the conjugate setup for point prior selection models (section 3.2 CGM), where $\beta_\gamma \sim N(0, \sigma^2\Sigma_\gamma)$. Again, as one cannot typically specify a separate prior distribution for β under each model, any practical implementation for Bayesian model uncertainty usually resorts to structured families of prior distributions. Another important consideration is whether prior specifications for β are "compatible" across models (Dawid and Lauritzen 2000) . For example, suppose that Model 1 contains variables 1 and 2, Model 2 contains variables 2 and 3, and Model 3 includes only variable 2. With apologies for possible abuse of notation, let β_2 denote the coefficient for variable 2 in each model. With completely arbitrary choices for Σ_γ, under Model 1 the variance for β_2 given that $\beta_1 = 0$ may not be the same as the variance for β_2 given that $\beta_3 = 0$ derived under Model 2, and both may differ from the variance for β_2 under the prior distribution given Model 3.

To avoid this incompatibility, choices for Σ_γ are often obtained from conditional specifications (i.e. conditional on $\beta_i = 0$ for $\gamma_i = 0$) derived from a prior distribution for β under the full model. For example, Zellner's g-prior (Zellner 1986) is commonly used, which leads to $\Sigma = c(X'X)^{-1}$ for the full model and $\Sigma_\gamma = c(X_\gamma' X_\gamma)^{-1}$ for model γ.

While in many cases conditioning leads to sensible choices, the result may depend on the choice of parameterization, which can lead to a Borel paradox (Kass and Raftery 1995; Dawid and Lauritzen 2000). Other structured families may be induced by marginalization or projections, leading to possibly different prior distributions. While structured prior distributions may reduce the number of hyperparameters that must be selected, i.e. to one parameter c, how to specify c is still an issue.

The choice of c requires careful consideration, as it appears in the marginal likelihoods of the data and hence the posterior model probabilities, and in my experience, can be influential. In the situation where little prior information is available, being too "non-informative" about β (taking c large) can have the un-intended consequence of favoring the null model *a posteriori* (Kass and Raftery 1995). While "default" prior distributions

(both proper and improper) can be calibrated based on information criteria such as AIC (Akaike Information Criterion - Akaike 1973), BIC (Bayes Information Criterion – Schwarz 1978), or RIC (Risk Inflation Criterion – Foster and George 1994), there remains the question of which one to use (Clyde 2000; George and Foster 2000; Fernandez et al. 1998); such decisions may relate more to utilities for model selection rather than prior beliefs (although it may be impossible to separate the two issues). Based on simulation studies, Fernandez et al.(1998) recommend RIC-like prior distributions when $n < p^2$ and BIC-like prior distributions otherwise. In wavelet regression, where $p = n$, there are cases where priors calibrated based on BIC have better predictive performance than prior distributions calibrated using RIC, and vice versa. From simulation studies and asymptotic arguments, it is clear that there is no one default choice for c that will "perform" well for all contingencies (Fernandez et al. 1998; Hansen and Yu 1999).

Empirical Bayes approaches provide an adaptive choice for c. One class of problems where BMA has had outstanding success is in non-parametric regression using wavelet bases. In nonparametric wavelet regression where subjective information is typically not available, empirical Bayes methods for estimating the hyperparameters have (empirically) led to improved predictive performance over a number of fixed hyperparameter specifications as well as default choices such as AIC, BIC, and RIC (Clyde and George 1999, 2000; George and Foster 2000; Hansen and Yu 1999) for both model selection and BMA. Because of the orthogonality in the design matrix under discrete wavelet transformations, EB estimates can be easily obtained using EM algorithms (Clyde and George 1999, 2000; Johnstone and Silverman 1998) and allow for fast analytic expressions for Bayesian model averaging and model selection despite the high dimension of the parameter space ($p = n$) and model space (2^n), bypassing MCMC sampling altogether.

George and Foster (2000) have explored EB approaches to hyperparameter selection for linear models with correlated predictors. EM algorithms for obtaining estimates of c and ω, as in Clyde and George (2000), can be adapted to the non-orthogonal case with unknown σ^2 using the conjugate point prior formulation and independent model space priors (equation 3.2 in CGM). For the EM algorithm both model indicators γ and σ^2 are treated as latent data and imputed using current estimates of c and $\omega = (\omega_1, \ldots, \omega_p)$, where ω_j is the prior probability that variable j is included under the independence prior. At iteration i, this leads to

$$\hat{\gamma}^{(i)} = \frac{p(Y|\gamma, c^{(i)})p(\gamma|\omega^{(i)})}{\sum_{\gamma'} p(Y|\gamma, c^{(i)})p(\gamma|\omega^{(i)})} \quad (1)$$

$$S_\gamma^{2(i)} = Y'Y - \frac{c^{(i)}}{1 + c^{(i)}}\mathrm{SSR}(\gamma) \quad (2)$$

where $SSR(\gamma)$ is the usual regression sum of squares and $S_\gamma^{2(i)}$ is a Bayesian version of

residual sum of squares using the current estimate $c^{(i)}$. Values of c and ω that maximize the posterior distribution for c and ω given the observed data and current values of the latent data are

$$\hat{\omega}_j^{(i+1)} = \sum_{\gamma \text{ such that } \gamma_j=1} \hat{\gamma}^{(i)} \tag{3}$$

$$\hat{c}^{(i+1)} = \max\left\{0, \left(\sum_{\gamma} \hat{\gamma}^{(i)} \frac{\mathrm{SSR}(\gamma)/(p\sum_j \hat{\omega}_j^{(i+1)})}{(\lambda\nu + S_\gamma^{2(i)})/(n+\nu)}\right) - 1\right\} \tag{4}$$

where ν and λ are hyperparameters in the inverse gamma prior distribution for σ^2 (CGM equation 3.10). These steps are iterated until estimates converge. EB estimates based on a common ω for all variables are also easy to obtain. For ridge-regression independent priors with $\Sigma_\gamma = cI$ or other prior distributions for β, estimates for ω_j have the same form, but estimates for c are slightly more complicated and require numerical optimization.

The ratio in the expression for $\hat{c}$ has the form of a generalized F-ratio, which is weighted by estimates of posterior model probabilities. The EM algorithm highlights an immediate difficulty with a common c for all models, as one or two highly significant coefficients may influence the EB estimate of c. For example, the intercept may be centered far from 0, and may have an absolute t-ratio much larger than the t-ratios of other coefficients. As the intercept is in all models, it contributes to all of the $SSR(\gamma)$ terms, which has the effect of increasing c as the absolute t-ratio increases. Since the same c appears in the prior variance of all other coefficients, if c becomes too large in order to account for the size of the intercept, we risk having the null model being favored (Bartlett's paradox (Kass and Raftery 1995)). While one could use a different prior distribution for the intercept (even a non-informative prior distribution, which would correspond to centering all variables), the problem may still arise among the other variables if there are many moderate to large coefficients, and a few that have extreme standardized values. Implicit in the formulation based on a common c is that the non-zero standardized coefficients follow a normal distribution with a common variance. As such, this model cannot accommodate one or a few extremely large standardized coefficients without increasing the odds that the remaining coefficients are zero. Using a heavy-tailed prior distribution for β may result in more robust EB estimates of c (Clyde and George 2000). Other possible solutions including adding additional structure into the prior that would allow for different groups of coefficients with a different c in each group. In the context of wavelet regression, coefficients are grouped based on the multi-resolution wavelet decomposition; in other problems there may not be any natural a priori groupings. Related to EB methods is the minimum description length (MDL) approach to model selection, which effectively uses a different c_γ estimated from the data

for each model (Hansen and Yu 1999). While EB methods have led to improvements in performance, part of the success depends on careful construction of the model/prior. Some of the problems discussed above highlight possible restrictions of the normal prior.

Unlike the EM estimates for orthogonal regression, the EB approach with correlated predictors involves a summation over all models, which is clearly impractical in large problems. As in the inference problem, one can base EB estimation on a sample of models. This approach has worked well in moderate sized problems, where leaps and bounds (Furnival and Wilson 1974) was used to select a subset of models; these were then used to construct the EB prior distribution, and then estimates under BMA with the estimated prior. For larger problems, leaps and bounds may not be practical, feasible, or suitable (such as CART models), and models must be sampled using MCMC or other methods. How to scale the EB/EM approach up for larger problems where models must be sampled is an interesting problem.

In situations where there is uncertainty regarding a parameter, the Bayesian approach is to represent that prior uncertainty via a distribution. In other words, why not add another level to the hierarchy and specify a prior distribution on c rather than using a fixed value? While clearly feasible using Gibbs sampling and MCMC methods, analytic calculation of marginal likelihoods is no longer an option. Empirical Bayes (EB) estimation of c often provides a practical compromise between the fully hierarchical Bayes model and Bayes procedures where c is fixed in advance. The EB approach plugs in the modal c into $g(\gamma)$ which ignores uncertainty regarding c, while a fully Bayes approach would integrate over c to obtain the marginal likelihood. As the latter does not exist in closed form, Monte Carlo frequencies of models provide consistent estimates of posterior model probabilities. However, in large dimensional problems where frequencies of most models may be only 0 or 1, it is not clear that Monte Carlo frequencies of models $p_f(\gamma|Y, S)$ from implementing MCMC for the fully Bayesian approach are superior to using renormalized marginal likelihoods evaluated at the EB estimate of c. When the EB estimate of c corresponds to the posterior mode for c, renormalized marginal likelihoods $g(\gamma)$ evaluated at the EB estimate of c are closely related to Laplace approximations (Tierney and Kadane 1986) for integrating the posterior with respect to c (the Laplace approximation would involve a term with the determinant of the negative Hessian of the log posterior). A hybrid approach where MCMC samplers are used to identify/sample models from the fully hierarchical Bayes model, but one evaluates posterior model probabilities for the unique models using Laplace approximations may provide better estimates that account for uncertainty in c.

Implementing Sampling of Models

In the variable selection problem for linear regression, marginal likelihoods are available in closed form (at least for nice conjugate prior distributions); for generalized linear models and many other models, Laplace's method of integration can provide accurate approximations to marginal distributions. The next major problem is that the model space is often too large to allow enumeration of all models, and beyond 20-25 variables, estimation of posterior model probabilities, model selection, and BMA must be based on a sample of models.

Deterministic search for models using branch and bounds or leaps and bounds algorithms (Furnival and Wilson 1974) is efficient for problems with typically fewer than 30 variables. For larger problems, such as in non-parametric models or generalized additive models, these methods are too expensive computationally or do not explore a large enough region of the model space, producing poor fits (Hanson and Kooperberg 1999). While Gibbs and MCMC sampling have worked well in high dimensional orthogonal problems, Wong et al. (1997) found that in high dimensional problems such as non-parametric regression using non-orthogonal basis functions that Gibbs samplers were unsuitable, from both a computational efficiency standpoint as well as for numerical reasons, as the sampler tended to get stuck in local modes. Their proposed focused sampler "focuses" on variables that are more "active" at each iteration, and in simulation studies provided better MSE performance than other classical non-parametric approaches or Bayesian approaches using Gibbs or reversible jump MCMC sampling.

Recently, Holmes and Mallick (1998) adapted perfect sampling (Propp and Wilson 1996) to the context of orthogonal regression. While more computationally intensive per iteration, this may prove to be more efficient for estimation than Gibbs sampling or MH algorithms in problems where the method is applicable. Whether perfect sampling can be used with non-orthogonal designs is still open.

With the exception of deterministic search, most methods for sampling models rely on algorithms that sample models with replacement. In cases where $g(\gamma)$ is known, model probabilities may be estimated using renormalized model probabilities (Clyde et al. 1996). As no additional information is provided under resampling models, algorithms based on sampling models without replacement may be more efficient. Under random sampling without replacement (with equal probabilities), the estimates of model probabilities (CGM equation 3.56) are ratios of Horvitz-Thompson estimators (Horvitz and Thompson 1952) and are simulation consistent. Current work (joint with M. Littman) involves designing adaptive algorithms for sampling without replacement where sampling probabilities are sequentially updated. This appears to be a promising direction for implementation of Bayesian model selection and model averaging.

Summary

CGM have provided practical implementations for Bayesian model selection in linear and generalized linear models, non-parametric regression, and CART models, as well as spurred advances in research so that it is feasible to account for model uncertainty in a wide variety of problems. Demand for methods for larger problems seems to outpace growth in computing resources, and there is a growing need for Bayesian model selection methods that "scale up" as the dimension of the problem increases. Guidance in prior selection is also critical, as in many cases prior information is limited. For example, current experiments using gene-array technology result in high dimensional design matrices, $p > 7000$, however, the sample size may only be on the order of 10-100 (Spang et al. 2000). Identifying which genes (corresponding to columns of X) are associated with outcomes (response to treatment, disease status, etc) is a challenging problem for Bayesian model selection, from both a computational standpoint, as well as the choice of prior distributions.

ADDITIONAL REFERENCES

Clyde, M. (2000). Model uncertainty and health effect studies for particulate matter. To appear in *Environmentrics.*

Dawid, A. and Lauritzen, S. (2000). Compatible prior distributions. Technical Report.

Fernandez, C., Ley, E., and Steel, M.F. (1998). Benchmark priors for Bayesian model averaging. Technical report, Dept. of Econometrics, Tillburg Univ., Netherlands.

Furnival, G.M. and Wilson, Robert W.J. (1974). Regression by Leaps and Bounds. *Technometrics,* **16**, 499-511.

George, E.I. (1999a). Discussion of "Bayesian Model Averaging: A Tutorial" by Hoeting, J.A., Madigan, D., Raftery, A.E., and Volinsky, C.T. *Statist. Sci.* **14**, 401-404.

Hansen, M. and Yu, B. (1999). Model selection and the principle of minimum description. Technical Report *http://cm.bell-labs.com/who/cocteau/papers.*

Hanson, M. and Kooperberg, C. (1999). Spline adaptation in extended linear models. Technical Report *http://cm.bell-labs.com/who/cocteau/papers.*

Hoeting, H.A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999). Bayesian model averaging: A tutorial (with discussion). *Statist. Sci.* **14**, 382-417.

Holmes, C. and Mallick, B.K. (1998). Perfect simulation for orthogonal model mixing. Technical Report, Dept. of Math., Imperial College, London.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Asso.* **47**, 663-685.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *J. Amer. Statist. Asso.* **90**, 773-795.

Propp, J. and Wilson, D. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms,* **9**, 223-252.

Spang, R., Zuzan, H., West, M., Nevins, J, Blanchette, C., and Marks, J.R. (2000). Prediction and uncertainty in the analysis of gene expression profiles. Discussion paper, ISDS, Duke Univ.

Wong, F., Hansen,M.H., Kohn, R., and Smith, M. (1997). Focused Sampling and its Application to Nonparametric and Robust Regression. Bell Labs Technical Report. Technical Report *http://cm.bell-labs.com/who/cocteau/papers.*

Dean P. Foster and Robert A. Stine

University of Pennsylvania

We want to draw attention to three ideas in the paper of Chipman, George and McCulloch (henceforth CGM). The first is the importance of an adaptive variable selection criterion. The second is the development of priors for interaction terms. Our perspective is information theoretic rather than Bayesian, so we briefly review this alternative perspective. Finally, we want to call attention to the practical importance of having a fully automatic procedure. To convey the need for automatic procedures, we discuss the role of variable selection in developing a model for credit risk from the information in a large database.

Adaptive variable selection

A method for variable selection should be *adaptive.* By this, we mean that the prior, particularly $p(\gamma)$, should adapt to the complexity of the model that matches the data rather than impose an external presumption of the number of variables in the model. One may argue that in reasonable problems the modeler should have a good idea how many predictors are going to be useful. It can appear that a well-informed modeler does not need an adaptive prior and can use simpler, more rigid alternatives that reflect knowledge of the substantive context. While domain knowledge is truly useful, it does

Dean P. Foster and Robert P. Stine are Associate Professors, Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104-6302, U.S.A; emails: foster@diskworld.wharton.upenn.edu and stine@wharton.upenn.edu.