REJOINDER

Hugh Chipman, Edward I. George and Robert E. McCulloch

First of all, we would like to thank the discussants, Merlise Clyde, Dean Foster and Robert Stine, for their generous discussions. They have each made profound contributions to model selection and this comes through in their insightful remarks. Although there is some overlap in the underlying issues they raise, they have done so from different vantage points. For this reason, we have organized our responses around each of their discussions separately.

Clyde

Clyde raises key issues surrounding prior selection. For choosing model space priors for the linear model with redundant variables, she confirms the need to move away from uniform and independence priors towards dilution priors. This is especially true in high dimensional problems where independence priors will allocate most of their probability to neighborhoods of redundant models. Clyde's suggestion to use an imaginary training data to construct a dilution prior is a very interesting idea. Along similar lines, we have considered dilution priors for the linear model where $p(\gamma)$ is defined as the probability that $Y^* \sim N_n(0, I)$ is "closer" to the span of X_γ than the span of any other $X_{\gamma'}$. Here Y^* can be thought of as imaginary training data reflecting ignorance about the direction of Y. Further investigation of the construction of effective dilution priors for linear models is certainly needed.

Clyde comments on the typical choices of Σ_γ for the coefficient prior $p(\beta_\gamma \mid \sigma^2, \gamma) = N_{q_\gamma}(\bar{\beta}_\gamma, \sigma^2\Sigma_\gamma)$ in (3.9), namely $\Sigma_\gamma = c\,(X_\gamma' X_\gamma)^{-1}$ and $\Sigma_\gamma = c\,I_{q_\gamma}$. In a sense, these choices are the two extremes, $c\,(X_\gamma' X_\gamma)^{-1}$ serves to reinforce the likelihood covariance while $c\,I_{q_\gamma}$ serves to break it apart. As we point out, the coefficient priors under such Σ_γ, are the natural conditional distributions of the nonzero components of β given γ when $\beta \sim N_p(0, c\,\sigma^2(X'X)^{-1})$ and $\beta \sim N_p(0, c\,\sigma^2 I)$, respectively. The joint prior $p(\beta_\gamma, \gamma \mid \sigma^2)$ then corresponds to a reweighting of the conditional distributions according to the chosen model space prior $p(\gamma)$. With respect to such joint priors, the conditional distributions are indeed compatible in the sense of Dawid and Lauritzen (2000). Although not strictly necessary, we find such compatible specifications to provide an appealingly coherent description of prior information.

We agree with Clyde that the choice of c can be absolutely crucial. As the calibration result in (3.17) shows, different values of c, and hence different selection criteria such as AIC, BIC and RIC, are appropriate for different states of nature. For larger models

with many small nonzero coefficients, smaller values of c are more appropriate, whereas for parsimonious models with a few large coefficients, larger values of c are better. Of course, when such information about the actual model is unavailable, as is typically the case, the adaptive empirical Bayes methods serve to insure against poor choices. It is especially appealing that by avoiding the need to specify hyperparameter values, empirical Bayes methods are automatic, a valuable feature for large complicated problems. Similar features are also offered by fully Bayes methods that margin out the hyperparameters with respect to hyperpriors. The challenge for the implementation of effective fully Bayes methods is the selection of hyperpriors that offer strong performance across the model space while avoiding the computational difficulties described by Clyde.

Clyde points out an important limitation of using empirical Bayes methods with conditional priors of the form $p(\beta_\gamma \mid \sigma^2, \gamma) = N_{q_\gamma}(\bar{\beta}_\gamma, \sigma^2 c\, V_\gamma)$. When the actual model has many moderate sized coefficients and but a few very large coefficients, the few large coefficients will tend to inflate the implicit estimate of c, causing the moderate sized coefficients to be ignored as noise. In addition to the heavy-tailed and the grouped prior formulations she describes for mitigating such situations, one might also consider elaborating the priors by adding a shape hyperparameter.

Finally, Clyde discusses the growing need for fast Bayesian computational methods that "scale up" for very large high dimensional problems. In this regard, it may be useful to combine heuristic strategies with Bayesian methods. For example, George and McCulloch (1997) combined globally greedy strategies with local MCMC search in applying Bayesian variable selection to build tracking portfolios. In our response to Foster and Stine below, we further elaborate on the potential of greedy algorithms for such purposes.

Foster and Stine

Foster and Stine begin by emphasizing the need for adaptive procedures. We completely agree. The adaptive empirical Bayes methods described in Section 3.3 offer improved performance across the model space while automatically avoiding the need for hyperparameter specification. For more complicated settings, adaptivity can be obtained by informal empirical Bayes approaches that use the data to gauge hyperparameter values, such as those we described for the inverse gamma distribution in Sections 3.2 and 4.1.2. In the sinusoid modelling example of Foster and Stine, a simple adaptive resolution is obtained by a Bayesian treatment with a prior on ω_k. This nicely illustrates the fundamental adaptive nature of Bayesian analysis. By using priors rather than fixed arbitrary values to describe the uncertainty surrounding the unknown characteristics in a statistical problem, Bayesian methods are automatically adaptive. We attribute the adaptivity

of empirical Bayes methods to their implicit approximation of a fully Bayes approach.

Foster and Stine go on to discuss some revealing analogies between strategies for minimum length coding and formulations for Bayesian model selection. The key idea is that the probability model for the data, namely the complete Bayesian formulation, also serves to generate the coding strategy. Choosing the probability model that best predicts the data is tantamount to choosing the optimal coding strategy. Foster and Stine note that improper priors are unacceptable because they generate infinite codes. This is consistent with our strong preference for proper priors for model selection. They point out the potential inefficiencies of Bernoulli model prior codes for variable selection, and use them to motivate a universal code that adapts to the appropriate model size. This is directly analogous to our observation in Section 3.3 that different hyperparameter choices for the Bernoulli model prior (3.15) correspond to different model sizes, and that an empirical Bayes hyperparameter estimate adapts to the appropriate model size. It should be the case that the universal prior corresponds to a fully Bayes prior that is approximated by the empirical Bayes procedure. Finally, their coding scheme for interactions is interesting and clearly effective for parsimonious models. Such a coding scheme would seem to correspond to a hierarchical prior that puts a Bernoulli $1/p$ prior on each potential triple - two linear terms and their interaction - and a conditionally uniform prior on the elements of the triple.

The credit risk example given by Foster and Stine raises several interesting issues. It illustrates that with this large dataset, an automatic stepwise search algorithm can achieve promising results. Figure 1 shows how their adaptive threshold criterion guards against overfitting, although the cross validation results seem also to suggest that a smaller number of terms, around 20, is adequate for prediction. Another automatic adaptive alternative to consider here would be a stepwise search based on the empirical Bayes criterion C_{CML} in (3.22). It would also be interesting to investigate the potential of one of the Bayesian variable selection approaches using the hierarchical priors described in Section 3.1 to account for potential relationships between linear and interaction terms. As opposed to treating all potential predictors independently, such priors tend to concentrate prior mass in a smaller, more manageable region of the model space.

For example, Chipman, Hamada and Wu (1997) considered an 18 run designed experiment with 8 predictors used in a blood-glucose experiment. The non-orthogonal design made it possible to consider a total of 113 terms, including quadratic terms and interactions. They found that independence priors of the form (3.2) led to such a diffuse posterior that, in 10,000 steps of a Gibbs sampling run, the most frequently visited model was visited only 3 times. On the other hand, hierarchical priors like (3.7) raised posterior mass on the most probable model to around 15%. In the same problem stepwise meth-

ods were unable to find all the different models identified by stochastic search. In effect, priors that account for interactions (or other structure, such as correlations between predictors which can lead to the dilution problem discussed in Section 3.1) can narrow the posterior to models which are considered more "plausible". We note, however, that the credit risk example is much larger than this example, and because the number of observations there is much larger than the number of predictors, such a hierarchical prior may have only a minor effect.

The credit risk example is also a clear illustration of the everlasting potential of greedy search algorithms on very large problems. At the very least, greedy algorithms can provide a "baseline" against which MCMC stochastic search results can be compared and then thrown out if an improvement is not found. Furthermore, greedy algorithms can provide a fast way to get rough estimates of hyperparameter values, and can be used directly for posterior search. Greedy algorithms also offer interesting possibilities for enhancement of stochastic search. At the most basic level, the models identified by greedy algorithms can be used as starting points for stochastic searches. Stochastic search algorithms can also be made more greedy, for example, by exponentiating the probabilities in the accept/reject step of the MH algorithms. The use of a wide variety of search algorithms, including MCMC stochastic search, can only increase the chances of finding better models.