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.

not follow that such knowledge conveys how many predictors belong in a model. The problem is made most transparent in the following admittedly artificial setting.

A small error in the choice of the basis in an orthogonal regression can lead to a proliferation in the number of required predictors. Suppose that we wish to predict future values of a highly periodic sequence, one dominated by a single sinusoid with frequency ω. If we approach this as a problem in variable selection and use the common Fourier basis to define the collection of predictors, the number of predictors is influenced by how close the frequency of the dominant cycle comes to a Fourier frequency. Fourier frequencies are of the form $\omega_j = 2\pi j/n$, indicating sinusoids that complete precisely j cycles during our n observations. If it so happens that $\omega = \omega_k$, then our model will likely need but one sinusoid to model the response. If ω is not of this form, however, our model will require many sinusoids from the Fourier basis to fit the data well. For example, with $n = 256$ and $\omega = 2\pi 5.5/n$, it takes 8 sinusoids at Fourier frequencies to capture 90% of the variation in this signal. The optimal basis would need but one sinusoid. Adaptive thresholding — the empirical Bayes approach — is forgiving of such errors, whereas dogmatic methods that anticipate, say, a single sinusoid are not.

Information theory and the choice of priors

A difficult choice in the use of Bayesian models for variable selection is the choice of a prior, particularly a prior for the subspace identifying the predictors. We have found coding ideas drawn from information theory useful in this regard, particularly the ideas related to Rissanen's minimum description length (*MDL*). The concreteness of coding offers appealing metaphors for picking among priors that produce surprisingly different selection criteria. In the Bayesian setting, calibration also offers a framework for contrasting the range of variable selection criteria.

The problem we consider from information theory is compression. This problem is simple to state. An *encoder* observes a sequence of n random variables $Y = (Y_1, \ldots, Y_n)$, and his objective is to send a *message* conveying these observations to a *decoder* using as few bits as possible. In this context, a *model* is a completely specified probability distribution, a distribution that is shared by the encoder and decoder. Given that both encoder and decoder share a model $P(Y)$ for the data, the optimal message length (here, the so-called "idealized length" since we ignore fractional bits and the infinite precision of real numbers) is

$$\ell(Y) = \log_2 \frac{1}{P(Y)} \text{ bits.}$$

If the model is a good representation for the data, then $P(Y)$ is large and the resulting message length is small. Since the encoder and decoder share the model $P(Y)$ they can

use a technique known as arithmetic coding to realize this procedure. But what model should they use?

Common statistical models like the linear model are parametric models P_{θ_q}, indexed by a q-dimensional parameter θ_q. For example, suppose that the data Y are generated by the Gaussian linear model

$$Y = \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_q X_q + \epsilon, \quad \epsilon \sim N(0, \sigma^2) .$$

To keep the analysis straightforward, we will assume σ^2 is known (see Barron, Rissanen and Yu 1998, for the general case). Given this model, the shortest code for the data is obtained by maximizing the probability of Y, namely using maximum likelihood (i.e., least squares) to estimate θ_q and obtain a message with length

$$\ell_{\hat{\theta}_q}(Y) = \frac{\log_2 e}{2\sigma^2} RSS(\hat{\theta}_q) + \frac{n}{2} \log_2(2\pi\sigma^2) ,$$

where $RSS(\hat{\theta}_q)$ is the residual sum of squares. This code length is not realizable, however, since $P_{\hat{\theta}_q}$ is *not* a model in our sense. The normal density for Y with parameters $\hat{\theta}_q = \theta_q(Y)$ integrates to more than one, $C_{n,q} = \int_Y P_{\theta_q(Y)}(Y) dY > 1$.

Once normalized with the help of some benign constraints that make the integral finite but do not interfere with variable selection (see, e.g., Rissanen 1986), the code length associated with the model $P_{\hat{\theta}_q}/C_{n,q}$ is

$$\ell_q(Y) = \log_2 C_{n,q} + \ell_{\hat{\theta}_q}(Y) . \tag{1}$$

The need for such normalization reminds us that coding does not allow improper priors; improper priors generate codes of infinite length. We can think of the first summand in (1) as the length of a code for the parameters $\hat{\theta}_q$ (thus defining a prior for θ_q) and the second as a code for the compressed data. This perspective reveals how coding guards against over-fitting: adding parameters to the model will increase $C_{n,q}$ while reducing the length for the data.

So far, so good, but we have not addressed the problem of variable selection. Suppose that both the encoder and decoder have available a collection of p possible predictors to use in this q-variable regression. Which predictors should form the code? In this expanded context, our code at this point is incomplete since it includes $\hat{\theta}_q$, but does not identify the q predictors. It is easy to find a remedy: simply prefix the message with the p bits in γ. Since codes imply probabilities, the use of p bits to encode γ implies a prior, p_1 say, for these indicators. This prior is the iid Bernoulli model with probability $\Pr(\gamma_i = 1) = \frac{1}{2}$ for which the optimal code length for γ is indeed p,

$$\log_2 1/p_1(\gamma) = \log_2 2^p = p .$$

Since adding a predictor does not affect the length of this prefix – it's always p bits – we add the predictor X_{q+1} if the gain in data compression (represented by the reduction in RSS) compensates for the increase in the normalizing constant $C_{n,q}$. Using a so-called two-part code to approximate the code length (1), we have shown (Foster and Stine 1996) that this approach leads to a thresholding rule. For orthogonal predictors, this criterion amounts to choosing those predictors whose z-statistic $z_j = \hat{\theta}_j/\mathrm{SE}(\hat{\theta}_j)$ exceeds a threshold near 2. Such a procedure resembles the frequentist selection procedure AIC, which uses a threshold of $\sqrt{2}$ in this context.

Now suppose that p is rather large. Using the p bits to represent γ seems foolish if we believe but one or two predictors are likely to be useful. If indeed few predictors are useful, we obtain a shorter message by instead forming a prefix from the *indices* of the those $\gamma_j = 1$. Each index now costs us about $\log_2 p$ bits and implies a different prior for γ. This prior, p_2 say, is again iid Bernoulli, but with small probability $\Pr(\gamma_i = 1) = 1/p$; the optimal code length for γ under p_2 is

$$\log_2 1/p_2(\gamma) = q \log p - (p-q)\log(1-q/p) \approx q \log p \,,$$

for $q = \sum_j \gamma_j \ll p$. Notice how this code assigns a higher cost to adding a predictor to the model. Adding a predictor does not affect the length of the prefix given by $p_1(\gamma)$. With $p_2(\gamma)$ as the prior, however, adding a predictor adds both a coefficient as well as its index to the message. The prefix grows by an additional $\log_2 p$ bits. For orthogonal regression and two-part codes, this approach implies a threshold for z_j near $\sqrt{2\log p}$ which also happens to correspond roughly to another frequentist procedure. This is the well-known Bonferroni method which retains predictors whose p-value is less than α/p for some $0 \le \alpha \le 1$.

Both of these codes have some appeal and correspond to frequentist methods as well as Bayesian priors, but neither is adaptive. The prior for the first code with a fixed p-bit prefix expects half of the predictors to be useful. The second has a prior that expects only one predictor to enter the model. As codes, both are flawed. The gold standard in coding is to compress the data down to the limit implied by the entropy of the underlying process, whatever that process may be. Both $p_1(\gamma)$ and $p_2(\gamma)$ only approach that limit when they happen to be right (e.g., when in fact only one predictor is needed in the model). Alternatively, so-called universal codes exist for representing binary sequences, and these compress γ almost as well as if the underlying probability were known. Assuming that the elements γ_j are iid (one can drop this condition as well), a universal code represents γ_q using about $pH(q/p)$ bits, where H denotes the Boolean entropy function

$$H(u) = u \log_2 \frac{1}{u} + (1-u)\log_2 \frac{1}{1-u}, \quad 0 \le u \le 1 \,.$$

Universal codes are *adaptive* in that they perform well for all values of q/p, doing almost as well as either of the previous codes when they happen to be right, but much better in other cases. Returning to the setting of an orthogonal regression, a universal code also implies a threshold for adding a predictor. The threshold in this case now depends on how many predictors are in the model. One adds the predictor X_j to a model that already has q predictors if its absolute z-statistic $|\hat{\theta}_j/\mathrm{SE}(\hat{\theta}_j)| > \sqrt{2\log p/q}$. This is essentially the empirical Bayes selection rule discussed by CGM in Section 3.3. The threshold decreases as the model grows, adapting to the evidence in favor of a larger collection of predictors. Again, this procedure is analogous to a frequentist method, namely step-up testing as described, for example, in Benjamini and Hochberg (1995).

Coding also suggests novel priors for other situations when the elements of γ are not so "anonymous". For example, consider the treatment of interaction terms. In the application we discuss in the next section, we violate the principle of marginality and treat interactions in a non-hierarchical fashion. That is, we treat them just like any other coefficient. Since we start with 350 predictors, the addition of interactions raises the number of possible variables to about $p = 67,000$. Since they heavily outnumber the linear terms, interactions dominate the predictors selected for our model. Coding the model differently leads to a different prior. For example, consider a variation on the second method for encoding γ by giving the index of the predictor. One could modify this code to handle interactions by appending a single bit to all indices for interactions. This one bit would indicate whether the model included the underlying linear terms as well. In this way, the indices for X_j, X_k and $X_j * X_k$ could be coded in $1 + \log_2 p$ bits rather than $3\log_2 p$ bits, making it much easier for the selection criterion to add the linear terms.

An application of automatic, adaptive selection

Methods for automatic variable selection matter most in problems that confront the statistician with many possibly relevant predictors. If the available data set holds, say, 1000 observations but only 10 predictors, then variable selection is not going to be very important. The fitted model with all 10 of these predictors is going to do about as well as anything. As the number of predictors increases, however, there comes a point where an automatic method is necessary. What constitutes a large number of possible predictors? Probably several thousand or more.

Such problems are not simply imaginary scenarios and are the common fodder for "data mining". Here is one example of such a problem, one that we discuss in detail in Foster and Stine (2000). The objective is to anticipate the onset of bankruptcy

for credit card holders. The available data set holds records of credit card usage for a collection of some 250,000 accounts. For each account, we know a variety of demographic characteristics, such as place and type of residence of the card holder. When combined with several months of past credit history and indicators of missing data, we have more than 350 possible predictors. The presence of missing data adds further features, and indeed we have found the absence of certain data to be predictive of credit risk. Though impressive at first, the challenge of choosing from among 350 features is nonetheless small by data mining standards. For predicting bankruptcy, we have found interactions between pairs or even triples to be very useful. Considering pairwise interactions swells the number of predictors to over 67,000. It would be interesting to learn how to apply a Gibbs sampler to such problems with so many possible features.

Though challenging for any methodology, problems of this size make it clear that we must have automated methods for setting prior distributions. To handle 67,000 predictors, we use adaptive thresholding and stepwise regression. Beginning from the model with no predictors, we identify the first predictor X_{j_1} that by itself explains the most variation in the response. We add this predictor to the model if its t-statistic $t_{11} = \hat{\beta}_{j_1,1}/se(\hat{\beta}_{j_1,1})$ (in absolute value) exceeds the threshold $\sqrt{2\log p}$. If $\hat{\beta}_{j_1,1}$ meets this criterion, we continue and find the second predictor X_{j_2} that when combined with X_{j_1} explains the most variation in Y. Rather than compare the associated t-statistic $t_{j_2,2}$ to the initial threshold, we reduce the threshold to $\sqrt{2\log p/2}$, making it now easier for X_{j_2} to enter the model. This process continues, greedily adding predictors so long as the t-statistic for each exceeds the declining threshold,

$$\text{Step } q\text{: Add predictor } X_{j_q} \quad \Longleftrightarrow \quad |t_{j_q,q}| > \sqrt{2\log p/q}$$

Benjamini and Hochberg (1995) use essentially the same procedure in multiple testing where it is known as step-up testing. This methodology works in this credit modelling in that it finds structure without over-fitting. Figure 1 shows a plot of the residual sum of squares as a function of model size; as usual, RSS decreases with p. The plot also shows the cross-validation sum of squares (CVSS) computed by predicting an independent sample. The validation sample for these calculations has about 2,400,000 observations; we scaled the CVSS to roughly match the scale of the RSS. Our search identified 39 significant predictors, and each of these — with the exception of the small "bump" — improves the out-of-sample performance of the model. Although the CVSS curve is flat near $p = 39$, it does not show the rapid increase typically associated with over-fitting.

Gibbs sampling and the practical Bayesian methods discussed by CGM offer an interesting alternative to our search and selection procedure. They have established the foundations, and the next challenge would seem to be the investigation of how their search procedure performs in the setting of real applications such as this. Greedy

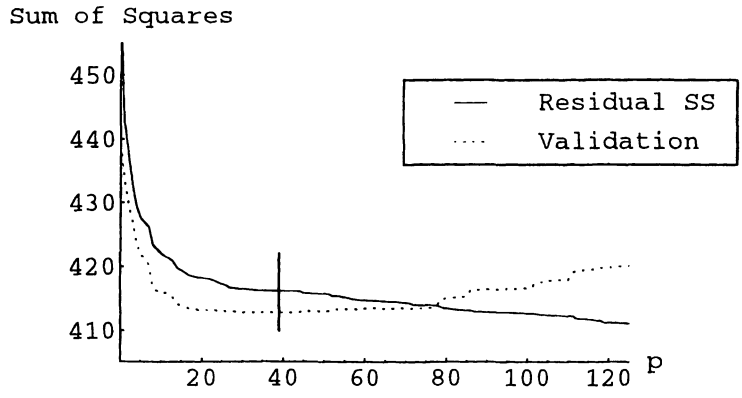

Figure 2: *Residual and cross-validation sums of squares for predicting bankruptcy.*

selection methods such as stepwise regression have been well-studied and probably do not find the best set of predictors. Once X_{j_1} becomes the first predictor, it must be in the model. Such a strategy is clearly optimal for orthogonal predictors, but can be 'tricked' by collinearity. Nonetheless, stepwise regression is fast and comparisons have shown it to be competitive with all possible subsets regression (e.g., see the discussion in Miller 1990). Is greedy good enough, or should one pursue other ways of exploring the space of models via Gibbs sampling?

ADDITIONAL REFERENCES

Barron, A., Rissanen, J. and Yu, B. (1998). The minimum description length principle in coding and modelling. *IEEE Trans. Info. Theory* **44**, 2743-2760.

Benjamini, Y. and Hoohberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289-300.

Foster, D.P. and Stine, R.A. (1996). Variable selection via information theory. Technical Report, Northwestern Univ., Chicago.

Foster, D.P. and Stine, R.A. (2000). Variable selection in data mining: Building a predictive model for bankruptcy. Unpublished Manuscript.

Miller, A.J. (1990). *Subset Selection in Regression.* Chapman and Hall, London.

Rissanen, J. (1986). Stochastic complexity and modelling. *Ann. Statist.* **14**, 1080-1100.