J. R. Statist. Soc. B (2002)
64, *Part* 3, *pp.* 519–536

Bayes model averaging with selection of regressors

P. J. Brown,

University of Kent at Canterbury, UK

M. Vannucci

Texas A&M University, College Station, USA

and T. Fearn

University College London, UK

[Received February 2001. Revised March 2002]

Summary. When a number of distinct models contend for use in prediction, the choice of a single model can offer rather unstable predictions. In regression, stochastic search variable selection with Bayesian model averaging offers a cure for this robustness issue but at the expense of requiring very many predictors. Here we look at Bayes model averaging incorporating variable selection for prediction. This offers similar mean-square errors of prediction but with a vastly reduced predictor space. This can greatly aid the interpretation of the model. It also reduces the cost if measured variables have costs. The development here uses decision theory in the context of the multivariate general linear model. In passing, this reduced predictor space Bayes model averaging is contrasted with single-model approximations. A fast algorithm for updating regressions in the Markov chain Monte Carlo searches for posterior inference is developed, allowing many more variables than observations to be contemplated. We discuss the merits of absolute rather than proportionate shrinkage in regression, especially when there are more variables than observations. The methodology is illustrated on a set of spectroscopic data used for measuring the amounts of different sugars in an aqueous solution.

Keywords: Bayesian model averaging; Decision theory; Multivariate general linear model; *QR*-updating; Ridge regression; Variable selection

1. Introduction

Bayesian mixture models have received considerable attention in recent years. Optimal predictions under squared error loss take the form of a Bayes model average; see Dempster (1973), Draper (1995) and Hoeting *et al.* (1999). We focus here on stochastic selection models for multivariate linear regression with univariate multiple regression as a special case. Best single-model variable selection is inherently unstable (see Breiman (1996)) and Bayesian model averaging provides a robust prediction remedy. Bayes model average predictions for the regression model have been given by Brown *et al.* (1998a), in the spirit of the univariate Bayes selection of Mitchell and Beauchamp (1988) and George and McCulloch (1997). Our formulation allowed fast computation even in the case of very many regressor variables, perhaps of the order of several hundred. It also addressed the situation when the number of observations is much less than the number of regressors.

Address for correspondence: P. J. Brown, Institute of Mathematics and Statistics, Cornwallis Building, University of Kent at Canterbury, Canterbury, Kent, CT2 7NF, UK.
E-mail: Philip.J.Brown@ukc.ac.uk

 1369–7412/02/64519

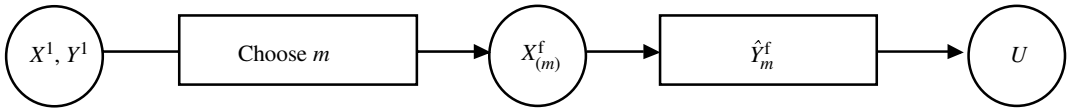

Fig. 1. Decision tree for variable selection for prediction: □, decisions; ○, random choices

One aspect of the Bayes model average prediction is that, whereas component models may involve just a few regressor variables, model averages typical involve an order of magnitude more variables. This may still be the case even when the set of models used in the averaging is restricted as in *Occam's window* (Madigan and Raftery, 1994). This may not be seen as a drawback, but it will be viewed as such if variables have costs of measurement or we are looking for interpretability of the mixture model *a posteriori*. Cost considerations were paramount in the decision theory approach of Brown *et al.* (1999). However, although Brown *et al.* (1999) broke new ground in the form of non-conjugate prior used it did not use a mixture model, rather favouring a single 'overmodel' as in Dawid (1988). This paper in contrast will address the choice of variables for prediction in the mixture model setting of Brown *et al.* (1998a). We shall later, in passing, discuss the alternative strategy of choosing a single model to approximate the Bayes model average, also resulting in reduced numbers of variables for prediction.

Selection is usually achieved either by decision theory with losses on the inclusion of variables or by a prior distribution for coefficients having spikes at zero. Our current approach is a hybrid of these two: adopting the spike prior but in addition implicitly penalizing the inclusion of variables, or rather promoting simple models.

The main work follows the paradigm of Lindley (1968), extending it in various directions, in particular to the multivariate linear regression mixture model.

We consider prediction in multivariate regression when both X of the training data and X in future are random and independent and identically distributed given parameters μ and Ω. We shall explicitly consider continuous X having a multivariate normal distribution. In the general formulation we shall have q responses Y. The decision tree in Fig. 1 consists of a series of round random and square decision nodes starting at the left with a round node generating X^l and Y^l, the $n \times p$ and $n \times q$ matrices of the learning (superscript l) data; then a square node choosing the subset identified by a p-vector of 0s and 1s denoted m, with p_m 1s. Then a future (superscript f) $X^{\mathrm{f}}_{(m)}$ $(1 \times p_m)$ is generated randomly, where suffix (m) denotes selection of those elements with $m_i = 1$, followed by a choice of the predictor of Y^{f} denoted $\hat{Y}^{\mathrm{f}}_m$. Then Y^{f} is randomly generated with the consequent pay-off $U(Y^{\mathrm{f}}, \hat{Y}^{\mathrm{f}}_m)$, where we shall look at a quadratic weighted loss. This is the direction of time. For decision analysis we fold back the tree from right to left, backwards in time, taking averages at random nodes and optimizing at decision nodes.

The paper is organized as follows: in Section 2 we set out the general multivariate linear model, the mixture prior distributions and the resulting posterior distributions. Finally this section develops the predictive distributions that are necessary for the following decision theoretic choices. Section 3.1 introduces a new fast algorithm based on data augmentation and adding and deleting variables.

Section 4.1 develops the decision theory approach to optimum prediction using a subset of regressors. We then contrast this with a method of generating a single model that approximates the Bayes model average. Section 5 presents a spectroscopic application. There follows a commentary in Section 6 on the general issue of proportional or absolute shrinkage. This critique is broader than the context of variable selection and offers a critical overview of g-priors and associated proportional (Stein) shrinkage. The paper concludes with a general discussion of the results achieved.

This paper shares with Brown *et al.* (1998a) a multivariate regression model incorporating a normal mixture prior on the regression matrix. In addition it requires the assumption of multivariate normal explanatory variables. In return it justifies a drastically reduced set of predictor variables. In analysing the training data, before contemplating prediction, our new fast algorithm for Markov chain Monte Carlo (MCMC) sampling can cope with very large numbers of explanatory variables. Our application involves 700 explanatory variables, which is far in excess of the number that is usually contemplated for this type of mixture model.

2. Bayesian analysis

2.1. The model and prior distribution

Our model for Y given X is a standard multivariate regression as in Brown *et al.* (1998a). Let the p-vector γ be the vector of 0s and 1s that defines the mixture model of Brown *et al.* (1998a), i.e. γ defines the prior structure of the regression coefficients. It represents the unknown 'true' model whereas the binary $p \times 1$ vector m denotes the variables selected. Let Y^l be the $n \times q$ response data in the learning set and X^l be the corresponding $n \times p$ set of explanatory variables. The matrix B is the $p \times q$ matrix of regression coefficients fixed by the scales adopted for the X-variables. The $q \times 1$ vector α gives intercepts for the q responses. Using the matrix variate notation of Dawid (1981), given Σ and γ, letting $\mathbf{1}_n$ denote an $n \times 1$ vector of 1s and denoting transposes by primes throughout, we have

$$Y^l - \mathbf{1}_n \alpha' - X^l B \sim \mathcal{N}(I_n, \Sigma), \tag{1}$$

$$\alpha' - \alpha_0' \sim \mathcal{N}(h, \Sigma), \tag{2}$$

$$B - B_0 \sim \mathcal{N}(H_\gamma, \Sigma), \tag{3}$$

with the marginal distribution of Σ as

$$\Sigma \sim \mathcal{IW}(\delta, Q). \tag{4}$$

In this notation both arguments of $\mathcal{N}(\cdot, \cdot)$ are covariances, the first referring to rows and the second to columns, as described for example in Brown *et al.* (1999). The notation has the advantage of preserving the matrix structure instead of reshaping the matrix as a vector. It also makes formal Bayesian manipulations much easier.

We take $B_0 = 0$ a prior matrix of 0s, and $h \to \infty$ in distribution (2), a vague prior on the intercept vector so that mean centring of X^l produces a posterior mean for α of $\bar{Y}^l$ (Smith, 1973). A selection prior will have diagonal elements of H_γ of 0 corresponding to $\gamma_i = 0$, so that such coefficients are 0 with probability 1. Berger (1988) has noted that such priors act as proxies for regression coefficients β that are small in magnitude, i.e. $|\beta| < \varepsilon$ for suitably small ε. For such a selection prior B is such that each column has a singular p-variate distribution, and given γ the p_γ variables with $\gamma_i = 1$ can be selected out from distribution (3) to give

$$B_{(\gamma)} \sim \mathcal{N}(H_{(\gamma)}, \Sigma)$$

where (γ) indicates selection by $\gamma_i = 1$. The prior distribution on γ is taken to be of a Bernoulli form,

$$\pi(\gamma) = \eta^{p_\gamma} (1 - \eta)^{p - p_\gamma}. \tag{5}$$

Here typically η is prespecified, although a further less stringent beta prior distribution on η was used in Brown *et al.* (1998b). If a small value of η is specified then smaller models will be preferred *a priori*.

2.2. Posterior distributions

The posterior mean of $B_{(\gamma)}$ conditional on γ is

$$\tilde{B}_\gamma = W_\gamma \hat{B}_\gamma. \tag{6}$$

The result follows from standard normal theory calculations; see for example appendix B of Brown (1993). In equation (6), the $p_\gamma \times p_\gamma$ matrix W_γ is given as

$$W_\gamma = (H^{-1}_{(\gamma)} + X^{l'}_{(\gamma)} X^l_{(\gamma)})^{-1} X^{l'}_{(\gamma)} X^l_{(\gamma)},$$

and the $p_\gamma \times q$ matrix $\hat{B}_\gamma$ is the least squares estimate from the selected γ-model:

$$\hat{B}_\gamma = (X^{l'}_{(\gamma)} X^l_{(\gamma)})^{-1} X^{l'}_{(\gamma)} Y^l.$$

Also the posterior distribution of γ is as equation (20) of Brown *et al.* (1998a). Rewritten in a data augmentation format, this becomes

$$\pi(\gamma | Y^{l'}, X^l) \propto g(\gamma) = (|\tilde{X}^{l'}_{(\gamma)} \tilde{X}^l_{(\gamma)}|)^{-q/2} |Q_\gamma|^{-(n+\delta+q-1)/2} \pi(\gamma), \tag{7}$$

where

$$\tilde{X}^l_{(\gamma)} = \begin{pmatrix} X^l_{(\gamma)} H^{1/2}_{(\gamma)} \\ I_{p_\gamma} \end{pmatrix},$$

$$\tilde{Y} = \begin{pmatrix} Y^l \\ 0 \end{pmatrix}$$

are $(n + p_\gamma) \times p_\gamma$ and $(n + p_\gamma) \times q$ matrices respectively and

$$Q_\gamma = Q + \tilde{Y}'\tilde{Y} - \tilde{Y}'\tilde{X}^l_{(\gamma)} (\tilde{X}^{l'}_{(\gamma)} \tilde{X}^l_{(\gamma)})^{-1} \tilde{X}^{l'}_{(\gamma)} \tilde{Y}, \tag{8}$$

an updating of Q by the residual sum of products matrix from the least squares regression of $\tilde{Y}$ on $\tilde{X}_{(\gamma)}$. This form also lends itself to fast updating using the QR-decomposition where Q is orthogonal and R upper triangular and adding and subtracting columns of the augmented X-matrix, avoiding 'squaring' quantities, simply back-solving a set of triangular equations; see for example Seber (1984). A new fast form of this applicable to diagonal forms of prior covariance structure H_γ is described in Section 3.1.

2.3. Predictive distributions

In predicting a single future response vector Y^{f} $(1 \times q)$ we assume the model analogous to model (1),

$$Y^{\mathrm{f}} - \alpha' - X^{\mathrm{f}} B \sim \mathcal{N}(1, \Sigma),$$

where X^{f} is $1 \times p$. Here, since it is assumed that X^l has been mean centred, then the same centring (by $\bar{X}^l$) must be applied to X^{f}. Independently, we assume that, given a $p \times p$ covariance matrix Ω,

$$X^l - \mathbf{1}_n \mu' \sim \mathcal{N}(I_n, \Omega),$$
$$X^{\mathrm{f}} - \mu' \sim \mathcal{N}(1, \Omega),$$
$$\mu' - \mu_0' \sim \mathcal{N}(h_x, \Omega),$$

and marginally

$$\Omega \sim \mathcal{IW}(\nu; P),$$

where typically h_x is allowed to tend to ∞, equivalent to mean correction of X^l. Also perhaps $P = k_x I_p$ will be an adequate choice with ν small (e.g. $\nu = 3$) so that the prior distribution of Ω is quite vague. The necessary manipulations for the conditional distribution of X^{f} given X^l may be obtained as in the development of Brown and Mäkeläinen (1992) by stacking X^{f} $(1 \times p)$ above X^l $(n \times p)$, obtaining the matrix variate Student distribution of this, and then using the form of the conditional distribution of X^{f} given X^l, namely with $h_x \to \infty$ given X^l,

$$X^{\mathrm{f}} \sim \mathcal{T}\{\nu + n; 1 + 1/n, P + (X^l)' X^l\}. \tag{9}$$

The matrix $\mathcal{T}$ uses the notation of Dawid (1981) alluded to in Section 2.1. Hence we may obtain the marginal distribution of $X^{\mathrm{f}}_{(m)}$ and the conditional distribution of $X^{\mathrm{f}}_{(\tilde{m})}$ given $X^{\mathrm{f}}_{(m)}$ where $(\tilde{m})$ denotes the selection of variables with $m_i = 0$. In particular,

$$X^{\mathrm{f}}_{(m)} \sim \mathcal{T}(\nu + n; 1 + 1/n, V_{mm}), \tag{10}$$

$$X^{\mathrm{f}}_{(\tilde{m})} | X^{\mathrm{f}}_{(m)} \sim X^{\mathrm{f}}_{(m)} V_{mm}^{-1} V_{m\tilde{m}} + \mathcal{T}(\nu + n + p_m; a, V_{\tilde{m}\tilde{m}.m}). \tag{11}$$

Here

$$V_{mm} = X^{l'}_{(m)} X^l_{(m)} + P_{(m)(m)},$$

$$V_{\tilde{m}\tilde{m}} = X^{l'}_{(\tilde{m})} X^l_{(\tilde{m})} + P_{(\tilde{m})(\tilde{m})},$$

$$V_{\tilde{m}m} = X^{l'}_{(\tilde{m})} X^l_{(m)} + P_{(\tilde{m})(m)},$$

$$V_{\tilde{m}\tilde{m}.m} = V_{\tilde{m}\tilde{m}} - V_{\tilde{m}m} V_{mm}^{-1} V_{m\tilde{m}}, \tag{12}$$

$$a = 1 + 1/n + X^{\mathrm{f}}_{(m)} V_{mm}^{-1} X^{\mathrm{f}'}_{(m)}, \tag{13}$$

where $V_{m\tilde{m}} = V'_{\tilde{m}m}$. The linear regressions of the multivariate normal distribution are used to impute X-variables corresponding to omitted variables in the development of Section 4.1. We first consider aspects of stochastic simulation for the computation of expression (7).

3. Markov chain Monte Carlo sampling

The posterior for γ is directly computable through expression (7). However, the right-hand side of expression (7) must be computed for all 2^p values of the latent vector γ. This becomes prohibitive even for modern computers and fast updating when p is much greater than around 20. Sequences of models which differ successively in only one variable (Gray codes) can be used to

speed up computations substantially but will still only allow up to around $p = 25$ variables. Our applications have generally involved much larger numbers of variables; see Section 5. In such circumstances it is possible to use MCMC sampling to explore the posterior distribution. One can quite quickly identify useful variables which have high marginal probabilities of $\gamma_j = 1$. It is also possible to find promising γ-vectors even though one has explored a very small fraction of the space of 2^p possibilities. In our example we would not claim to have achieved convergence, even though marginal distributions from very disparate starting-points look similar. We do, however, find some models that predict very well. In the context of the spectroscopic application there is a considerable degree of local wavelength interchangeability, so similar but different models may have the same predictive value.

To sample γ from distribution (7) we used a Metropolis algorithm that was suggested for model selection by Madigan and York (1995) and applied to variable selection for regression by Brown *et al.* (1998b), George and McCulloch (1997) and Raftery *et al.* (1997). We generate a candidate new selection vector γ^* from the current γ by one of two possible moves: either, with probability ϕ, add or delete a variable chosen at random or, with probability $1 - \phi$, swap two variables by choosing independently at random a 0 and a 1 in γ and changing both of them. The new candidate model coded as γ^* is accepted with probability

$$\min\left\{\frac{g(\gamma^*)}{g(\gamma)}, 1\right\}, \tag{14}$$

with $g(\gamma)$ given by expression (7). There is considerable flexibility in how we design the sequence of random moves. Within the scheme above the ϕ-parameter needs to be chosen. We chose $\phi = \frac{1}{2}$, but it might be desirable to have more additions or deletions through a higher value of ϕ. Furthermore we could have chosen moves that added or subtracted or swapped two or three or more at a time, or a combination of these, or to focus also on adjacent swaps as in Sha *et al.* (2002).

As usual we seek the stationary distribution of the Markov chain. The standard procedure is that after a suitable burn-in period the realizations are monitored to see that they appear stationary. For our work we have adopted a strategy of running the chain from four different starting-points and looking at the four marginal distributions provided by the computed $g(\gamma)$ values of the visited γ. Because we know the relative probabilities we do not need to worry about creating a burn-in period: early low probability visited γ will not make any sizable contributions and we do not need to rely on equal sampling rates for visited models as, for example, in the MC^3 algorithm of Madigan and York (1995).

3.1. Fast forms of updating

The posterior distribution of γ was given in expression (7) in terms of an augmented least squares regression. The QR-decomposition of $(\tilde{X}^l_{(\gamma)}, \tilde{Y}^l)$ is given, for example, by Seber (1984), chapter 10, section 1.1b, and avoids squaring as in expressions (7) and (8). Updating `qrdelete` and `qrinsert` algorithms are then available within many computing environments, removing or adding a column. These require the number of rows to be fixed for an efficient implementation. This has meant that our original algorithm described in Brown *et al.* (1998a) required the setting up of an original $(n + p) \times p_\gamma$ matrix, formed by taking all the original regressors and augmenting by I_p the $p \times p$ identity matrix. In our earlier applications p has been as high as 350 and $p = 700$ in our later example, so the X-matrix would have 700 extra rows. In applications that we are now contemplating with deoxyribonucleic acid microarrays and expression data on genes the number of regressors may be as high as $p = 7000$; see for

example West *et al.* (2000). Clearly this approach is impractical and involves many unnecessary computations especially as generally the models being searched will involve only a small subset of the regressors, perhaps 20 or 30, and the vast majority of the rows will be zero and will not contribute to the estimation. The key insight is that all that is required of the augmenting matrix is that there is at most a single 1 in a row. This means that we can set a modest maximum dimension which can be dynamically altered if it becomes necessary. Suppose that this maximum dimension is denoted as p^*; it might for example be 20, although it would be wise initially to set it much higher for higher starting values. Suppose that our initial model involves $p_\gamma < p^*$ variables. This fast form of the algorithm works when H_γ is diagonal. We have the following steps in outline.

Step 1: initialization—augment Y with $p^* \times q$ 0s. Set up IRESERVE as the $p^* \times p^*$ identity matrix. Take the p_γ chosen X-variables in the form of an $n \times p_\gamma$ matrix. Augment with any p_γ columns taken from IRESERVE to create an $(n + p^*) \times p_\gamma$ augmented X-matrix, leaving $p^* - p_\gamma$ columns in IRESERVE. Perform a QR-decomposition.
Step 2: decide whether to delete, or insert or swap (achieved by a delete operation followed by an insertion or vice versa).
Step 3: if we *insert* then take the new $n \times 1$ X-column and augment with a column taken from IRESERVE. Use `qrinsert` as in MATLAB or its equivalent in other computing environments based on LINPACK.
Step 4: if we *delete*, note the identity of the variable chosen, use `qrdelete` to remove it and return its $p^* \times 1$ augmenting vector to IRESERVE.
Step 5: repeat the last three steps. If at any time the process wishes to include more than p^* variables, then p^* will need to be reset to a larger value and reinitialized at step 1 with a QR-decomposition. Conversely if p_γ is substantially less than p^* then the latter should be reduced.

This algorithm is available in MATLAB at `http://stat.tamu.edu/~mvannucci/webpages/codes.html`.

4. Choice of subset

4.1. Folding back the decision tree

The decision theory approach involves working backwards to the first decision node encountered. In the right-hand square node in Fig. 1 we need to choose the optimum Bayes predictor given an earlier optimum chosen subset m. We choose $\hat{Y}^{\mathrm{f}}_m$ to minimize

$$E\{(Y^{\mathrm{f}} - \hat{Y}^{\mathrm{f}}_m)L(Y^{\mathrm{f}} - \hat{Y}^{\mathrm{f}}_m)' | X^l, Y^l, X^{\mathrm{f}}_{(m)}, m\}, \tag{15}$$

where E is an expectation over Y^{f} and L is a $q \times q$ matrix of constants determining the weights of the various responses in the loss. Often we would take L to be diagonal, or equivalently $L = I$ after a suitable scaling.

Returning to expression (15), by adding and subtracting $E(Y^{\mathrm{f}}|X^l, Y^l, X^{\mathrm{f}}_{(m)}, m)$ from Y^{f} and squaring out we see that the loss is minimized by

$$\begin{aligned}
\hat{Y}^{\mathrm{f}}_m &= E(Y^{\mathrm{f}}|X^l, Y^l, X^{\mathrm{f}}_{(m)}, m) \\
&= E_{X^{\mathrm{f}}|X^{\mathrm{f}}_{(m)}, X^l, Y^l, m}\{E(Y^{\mathrm{f}}|X^l, Y^l, X^{\mathrm{f}}, m)\} \\
&= E(\bar{Y}^l + X^{\mathrm{f}}\tilde{B}_\Gamma | X^l, Y^l, X^{\mathrm{f}}_{(m)}, m),
\end{aligned}$$

with $\tilde{B}_\Gamma$ ($p \times q$) formed by filling out with 0s the $p_\gamma \times q$ posterior mean $\tilde{B}_\gamma$ of B for given γ given by equation (6). The last expectation is over both X^{f} and γ denoted as the random variable Γ.

Thus

$$\hat{Y}^{\mathrm{f}}_m = \bar{Y}^l + E(X^{\mathrm{f}}|X^l, X^{\mathrm{f}}_{(m)}, m)\, E(\tilde{B}_\Gamma|X^l, Y^l), \tag{16}$$

since X^{f} and γ are *a priori* and *a posteriori* independent given X^l and Y^l and where the second expectation is over the posterior distribution of γ.

In what follows we shall temporarily suppress the conditioning on the training data, so that throughout conditioning is on X^l and Y^l as well as the selection m. Now $\hat{Y}^{\mathrm{f}}_m$ from equation (16) can be computed by using expression (11) for $E(X^{\mathrm{f}}|X^{\mathrm{f}}_{(m)})$ and $E(\tilde{B}_\Gamma)$ approximated by averaging over the normalized posterior probabilities of visited γ. Now returning to loss (15) with $\hat{Y}^{\mathrm{f}}_m$ given by equation (16), going back in time to the next node in the decision tree in Fig. 1, we need to average over X^{f}_m. We require

$$E(ZLZ'|X^l, Y^l, m), \tag{17}$$

where

$$\begin{aligned} Z &= Y^{\mathrm{f}} - E(X^{\mathrm{f}}|X^{\mathrm{f}}_{(m)})\, E(\tilde{B}_\Gamma) \\ &= Y^{\mathrm{f}} - X^{\mathrm{f}}B + X^{\mathrm{f}}B - E(X^{\mathrm{f}}|X^{\mathrm{f}}_{(m)})\, E(\tilde{B}_\Gamma). \end{aligned} \tag{18}$$

Squaring out expectation (17) using the pairs of terms in equation (18) we obtain the expected loss as

$$E\{(Y^{\mathrm{f}} - X^{\mathrm{f}}B)L(Y^{\mathrm{f}} - X^{\mathrm{f}}B)'\} + E(ULU') \tag{19}$$

with $U = X^{\mathrm{f}}B - E(X^{\mathrm{f}}|X^{\mathrm{f}}_{(m)})\, E(\tilde{B}_\Gamma)$, the expectation of the cross-product being 0. The first term of expression (19) is just $E\{\mathrm{tr}(L\Sigma)\}$. The second term can be expanded and squared again as three terms:

$$U = \{X^{\mathrm{f}} - E(X^{\mathrm{f}})\}\{B - E(\tilde{B}_\gamma)\} + E(X^{\mathrm{f}})\{B - E(\tilde{B}_\Gamma)\} + \{X^{\mathrm{f}} - E(X^{\mathrm{f}}|X^{\mathrm{f}}_{(m)})\}\, E(\tilde{B}_\Gamma).$$

Just the three squared terms remain since the cross-products have expectation 0. It is only the last of these squared terms that involves m, the choice of variables, so for selecting m we may ignore the other terms and evaluate this. Using the fact that the trace of a scalar is scalar this may be seen to be

$$\mathrm{tr}(E[\{X^{\mathrm{f}} - E(X^{\mathrm{f}}|X^{\mathrm{f}}_{(m)})\}'\{X^{\mathrm{f}} - E(X^{\mathrm{f}}|X^{\mathrm{f}}_{(m)})\}]\, E(\tilde{B}_\Gamma)\, L\{E(\tilde{B}_\Gamma)\}'). \tag{20}$$

The first term is the conditional covariance of X^{f} given $X^{\mathrm{f}}_{(m)}$ which from expression (11) is equal to

$$aV_{\tilde{m}\tilde{m}.m}/(\nu + n + p_m - 2), \tag{21}$$

with a a quadratic form in $X^{\mathrm{f}}_{(m)}$ given by equation (13) and $V = P + X^{l'}X^l$ from expression (9). Averaged over $X^{\mathrm{f}}_{(m)}$ using expression (10), since

$$E(X^{\mathrm{f}'}_{(m)}X^{\mathrm{f}}_{(m)}) = \left(1 + \frac{1}{n}\right)\frac{V_{mm}}{\nu + n - 2}$$

and hence

$$E(a) = \left(1 + \frac{1}{n}\right)\left\{1 + \frac{\mathrm{tr}(I_m)}{\nu + n - 2}\right\},$$

then the required average of expression (21) equals $dV_{\tilde{m}\tilde{m}.m}$, where the scalar d is given as

$$d = \frac{1 + 1/n}{\nu + n - 2}.$$

This is independent of m and may be ignored. Thus the first term of expression (20) is seen to be 0 for variables that are not chosen by m and to be otherwise proportional to $V_{\tilde{m}\tilde{m}.m}$ given by expression (12). The criterion for the choice of subset becomes choose that m which minimizes (with all conditioning reinstated)

$$\mathrm{tr}\{E(\tilde{B}_{\tilde{m}\Gamma}|X^l, Y^l)'[V_{\tilde{m}\tilde{m}} - V_{\tilde{m}m}V_{mm}^{-1}V_{m\tilde{m}}]\,E(\tilde{B}_{\tilde{m}\Gamma}|X^l, Y^l)L\}, \tag{22}$$

where the suffix $\tilde{m}$ in $\tilde{B}_{\tilde{m}\Gamma}$ extracts those coefficients corresponding to variables not selected, giving a $(p - p_m) \times q$ matrix.

The form of expression (22) is fairly natural. The optimum is to use all variables. However, if we have an implicit additional cost of including variables we may be willing to inflate expression (22) slightly if it means including substantially fewer variables. The criterion consists of expected coefficients for omitted variables combined with $V_{\tilde{m}\tilde{m}.m}$ in the square brackets, the residual covariance matrix for the prediction of x-variables not included from those included. It says that a subset m does not penalize if either

(a) the coefficients expected are near 0 for the variables omitted or
(b) the omitted variables are well predicted by the included variables although these omitted variables have sizable coefficients.

For comparison purposes it may be desirable to standardize expression (22) by dividing by the value that would result from no variables chosen, i.e. the same form but with V replacing $V_{\tilde{m}\tilde{m}.m}$ and $\tilde{B}_{\tilde{m}\Gamma} \to \tilde{B}_\Gamma$.

It may be noted that criterion (22) has the same form under many other multivariate normal model formulations, not just the mixture prior structure of expression (3) with equation (5). Indeed models without any sort of stochastic selection (e.g. ridge regression) could incorporate variable selection for prediction through a form analogous to expression (22).

4.2. Computation

One way of choosing the subset is to start with all variables included or some fairly natural maximal set of variables. This could just be all variables involved in models visited or all those variables that appear with probability at least some small amount, obtained from the marginal distribution of visited models. The algorithm would work as follows: reduce the variables, perhaps by backward elimination, successively plotting expression (22) for the minimizing subset for each size of variable subset until the plot starts to increase substantially. This 'elbow' gives the 'optimum' subset. Of course if we have explicit costs on variables then a strict optimum can be obtained, balancing an increase in expression (22) against the cost of variables, perhaps linear in the number of variables. Some judgment will need to be made about what constitutes an unacceptable increase in criterion (22). We have avoided explicit costs of variables as used by Lindley and subsequently in Brown *et al.* (1999).

Once an optimum m has been obtained then any prediction involving $X^{\mathrm{f}}_{(m)}$ is made by using

equation (16). This involves two expectations. The first expectation is obtained by computing $X^{\mathrm{f}}_{(\tilde{m})}|X^{\mathrm{f}}_{(m)}, X^l, m$ by using expression (11) and with $X^{\mathrm{f}} = (X^{\mathrm{f}}_{(\tilde{m})}, X^{\mathrm{f}}_{(m)})$ so that included variables are retained, interlaced between the imputed excluded variables.

The second expectation is the model average estimated coefficients given the learning data. The term $E(\tilde{B}_\Gamma|X^l, Y^l)$ can be computed for once and all. It can be approximated by the average over the visited models from the MCMC search. We need to be cautious about such averages when prior distributions are vague, for then Bayes factors become arbitrary. This does not strictly arise with our proper priors, but see Fernández *et al.* (2001). For a given selection vector γ the Bayes estimate of the coefficient matrix as given by equation (6) can be written in alternative forms that are convenient for computation, either

$$(I + H_{(\gamma)} X^{l'}_{(\gamma)} X^l_{(\gamma)})^{-1} H_{(\gamma)} X^{l'}_{(\gamma)} Y^l_{(\gamma)}$$

or the symmetric forms of augmentation suggested in section 8 of Brown *et al.* (1998a) and facilitating the use of a QR-decomposition with insertions and deletions, but with the overhead of greatly increased dimensions. For diagonal H this overhead is reduced by the new fast algorithm in Section 3.1.

4.3. Single-model approximations

Our approach as developed above has been to retain model averaging but to consider a reduced set of regressors for prediction. An alternative would be to approximate the Bayes model average $E(\tilde{B}_\Gamma|X^l, Y^l)$ by means of a single model. We explore this approach briefly in this section for comparison with our main approach. The single model will in itself tend to use fewer variables than the model averaging. The natural metric stemming from expression (22) is to seek a model m which minimizes

$$\mathrm{tr}\{(E(\tilde{B}_\Gamma) - B_m)' V (E(\tilde{B}_\Gamma) - B_m) L\}. \tag{23}$$

Here V replaces the conditional variance since we are not intending to choose fewer variables for prediction except incidentally through single-model selection. For an alternative formulation where the loss is on the logarithmic scale see San Martini and Spezzaferri (1984). We make the following observations if $X^{l'} X^l$ and H_γ are diagonal.

(a) The least squares and Bayes estimates for any row of B do not change with γ, given only that the row corresponds to a $\gamma = 1$ row.
(b) The Bayes model average for the ith row is then a weighted average of the zero vector and the Bayes estimator for the variable included with weights the marginal posterior probabilities of $\gamma_i = 0$ and $\gamma_i = 1$ respectively. To minimize expression (23) then choose to include the variable if and only if the marginal posterior probability of $\gamma_i = 1$ is at least $\frac{1}{2}$.

This has been called the *median model* rule by Barbieri and Berger (2002), who derived it by geometric arguments. The orthogonal X-setting is rather artificial for our purposes, although it would be relevant for the selection of derived variables such as principal components; see Clyde *et al.* (1996) and Clyde and George (2000), who exploited orthogonality further and also considered scale-mixed long-tailed distributions.

When orthogonality of both the prior and x-variables cannot be assumed then both least squares and Bayes coefficients for the ith row of B depend on which other variables are included and no explicit rule is possible. However, we can search through models to minimize expression (23). In Section 5.2 we see how suboptimal the median model is in our general non-orthogonal

setting by searching through a sequence of marginal models constructed from thresholding at levels less than the median cut-off of $\frac{1}{2}$.

5. The application

The sugars data consisted of three sugars each at five levels in aqueous solution making 125 observations in a full 5^3-factorial. Each of these was presented to a near infra-red spectrometer, and the second difference absorbance spectra were recorded from 1100 to 2498 nm at 2 nm intervals. The data set is described in Brown (1993). Predictive ability was judged on a separate validation data set of 21 observations. The data can be attained from

`http://www.blackwellpublishers.co.uk/rss/`

5.1. Hyperparameter settings

Our philosophy here is to rely on variable selection to provide the regularization, and consequently to 'make do' with very weak prior assumptions in terms of hyperparameter settings. The sugars data have previously been analysed in Brown *et al.* (1998a). Here for the first time we utilize the full 700-variable spectrum, which is now feasible with our new algorithm of Section 3.1.

The regularization of the $X'X$ matrix for imputation did not enter previous work. We take $\nu = 3$ the smallest integer value such that the prior mean exists. This means that the specification of the scale matrix P is not too critical. We assumed $P = k_x I_p$ to regularize $X'X$ and chose a small value for k_x. The size of k_x is judged relative to the non-zero eigenvalues of $X'X$. There are $g = \min(n-1, p)$ of these non-zero eigenvalues with probability 1. We ordered the eigenvalues from largest, $\lambda_{(1)}$, to smallest, $\lambda_{(g)}$, and took the lower decile of these, i.e. $k_x = \lambda_{([0.9g])}$, where $[\cdot]$ indicates the integer part. The other hyperparameters are H_γ from the prior assumption on the coefficients conditional on γ in equation (3) and δ and Q from the prior for the error covariance specified by distribution (4). We also take $\delta = 3$, with the same argument as for ν above, and $Q = k_e I_q$. The coefficient prior scale matrix is such that $H_{(\gamma)} = c[I_{(\gamma)}]$ rather than the diagonalized g-prior in Brown *et al.* (1998a). For a further general discussion of the form of H see Section 6. We take $k_e = 0.01$ after response standardization, corresponding to a prior expectation of 99% variation explained. Because of the low prior degrees of freedom the value of k_e is unlikely to be critical. Also c was chosen by marginal maximum likelihood inflated by 700/20 where 20 is the prior expected number of included variables out of the full 700. This empirical Bayes estimate of c is derived in Brown *et al.* (2001).

5.2. Results

We ran four Metropolis chains of 100 000 iterations each. Initial numbers of variables were

(a) 350,
(b) 100,
(c) 50 and
(d) 10

variables selected in wavelength order (from 1100 nm). These were regarded as sufficiently disparate and different from the prior expectation of 20 variables to explore very different regions of space initially, in line with the philosophy of Gelman and Rubin (1992). The probability of moves by either adding, or deleting or swapping was $\phi = \frac{1}{2}$. Graphs of marginal probabilities

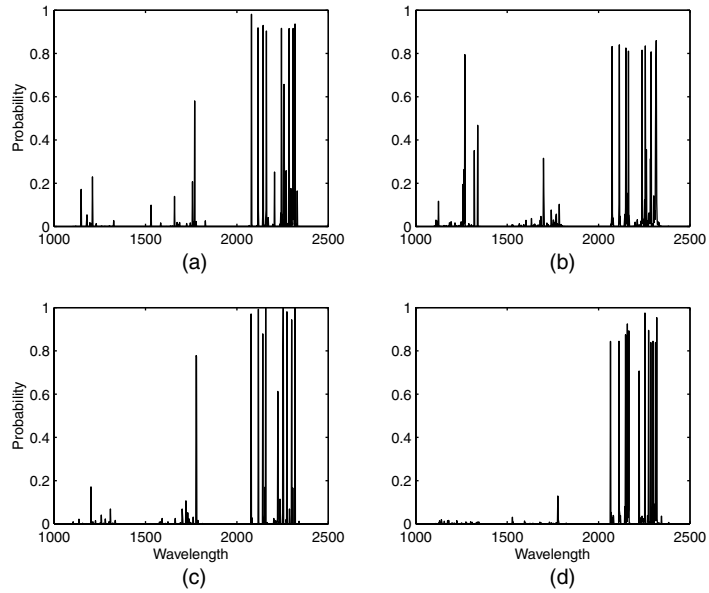

Fig. 2. Normalized marginal probabilities of the components of γ for four starting values of (a) 350, (b) 100, (c) 50 and (d) 10 variables randomly selected

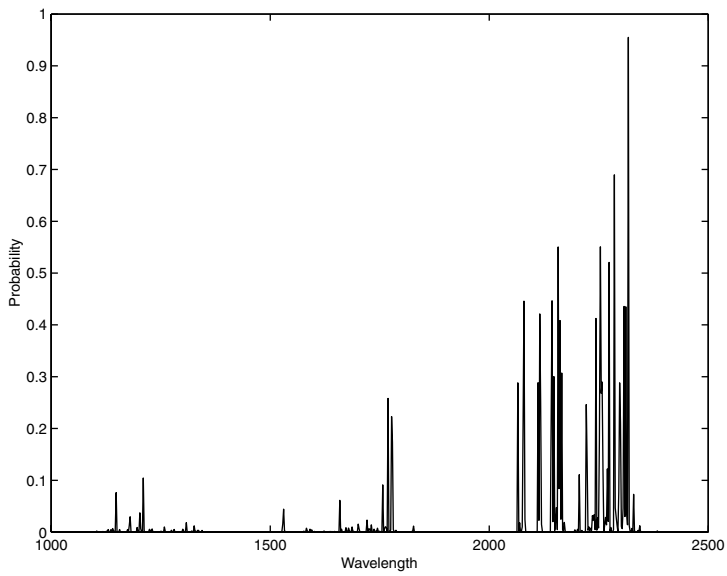

Fig. 3. Renormalized marginal probabilities of the components of γ: results for four runs pooled

for the wavelengths of visited models for the four runs are given in Fig. 2 and show reasonable consistency. The pooled plot is shown in Fig. 3.

The top 500 visited models in the pooled four runs had 164 distinct variables and accounted for 99.8% of the total visited probability. Starting with these 164 variables we applied criterion (22) with $L = I_3$ (and no scaling) to the learning data to select a subset of variables, eliminating variables one at a time, each time removing the variable which did least to in-

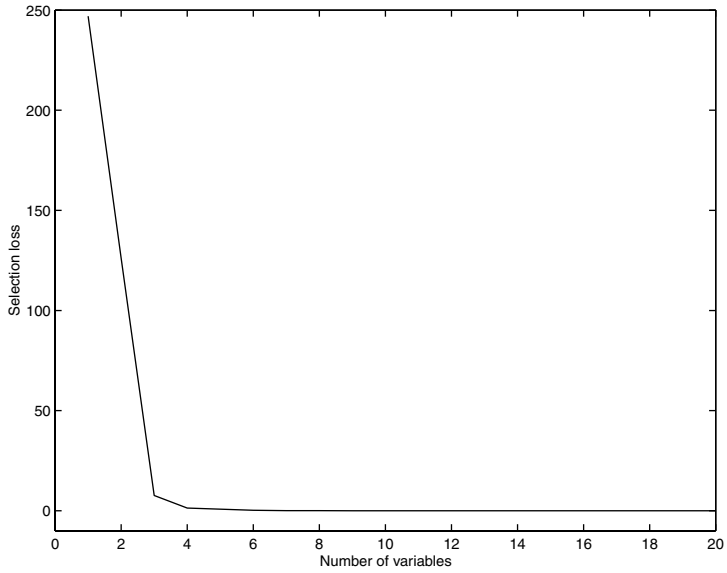

Fig. 4. Elbow plot of selection loss (22) against number of variables for the first 20 retained

crease the criterion. Plotted in Fig. 4 an elbow suggests around six variables; this choice results in mean-square errors (MSEs) using predictor (16) of (0.43, 0.11, 0.32) for the three sugars (sucrose, glucose and fructose), compared with the unrestricted approximate Bayes model average MSE of (0.23, 0.39, 0.27), so there was very little penalty for restricting to six variables; in fact there is an improvement for glucose. The six variables from the 700 in total were (2078 nm, 2114 nm, 2252 nm, 2272 nm, 2298 nm, 2316 nm) from the range 1100–2498 nm in steps of 2 nm. The highest probability model among those visited by the MCMC algorithm had a comparable prediction MSE of (0.30, 0.52, 0.19) but used 12 variables.

Finally, in Section 4.3 we promised to illustrate the approximation to the Bayes model average by using a single model. The 'median' model chose just five variables and had a much worse MSE on validation of (0.53, 1.25, 0.19). The plot of the loss given by expression (23) plotted for 143 models ordered by the threshold from 0.5 to 0.001 in steps of 0.00005 is given in Fig. 5. The best model in that plot had a comparable MSE on validation of (0.35, 0.10, 0.35) but used 12 variables. The best model according to the criterion found by searching through models ordered by posterior probability had a comparable MSE on validation of (0.31, 0.20, 0.10) but used 21 variables. Plots are given of the logarithm of criterion (23) in Fig. 6. Although there may be questions about whether the original chain had converged, the relative probabilities of models are exact irrespective of convergence.

It might be wondered whether a separate prediction for each of the three sugars would offer greater savings. The univariate regression approach was applied to these data in Brown (1992) both for the full 700 wavelengths and for subsets chosen by the method of Brown *et al.* (1991). These do not show an improvement on the multivariate Bayes approach here.

5.3. Interpretation

One advantage of reducing the number of variables in the prediction equation is that it increases the scope for interpretation. The wavelength range, roughly 2050–2350 nm, in which our six selected variables lie is one where the three sugars show spectral differences. Some of

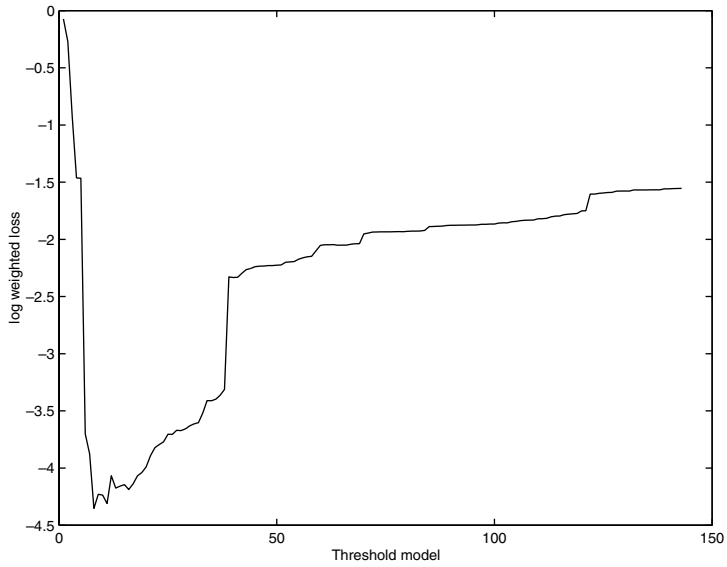

Fig. 5. Plot of log-weighted-loss from expression (23) against marginal models ordered by the level of the threshold

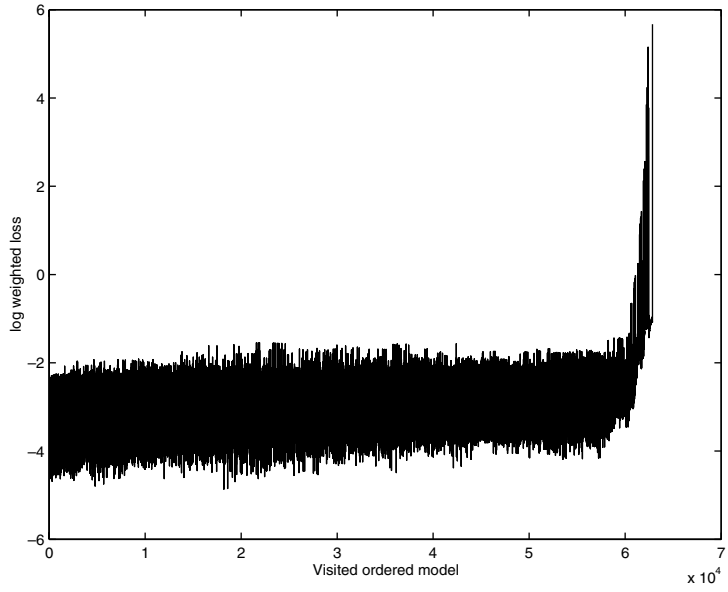

Fig. 6. Plot of log-weighted-loss from expression (23) against visited models ordered by probability for the sugars data

these wavelengths, or ones very close to them, have arisen before in similar investigations. For example, Lanza and Li (1984) identified 2256, 2270 and 2314 nm (among others) as useful for the measurement of these sugars in model mixtures and in fruit juices, and Osborne *et al.* (1993) assigned a peak at 2080 nm to a bond vibration associated with sucrose, as well as reporting that 2073 nm has been used for sucrose measurements.

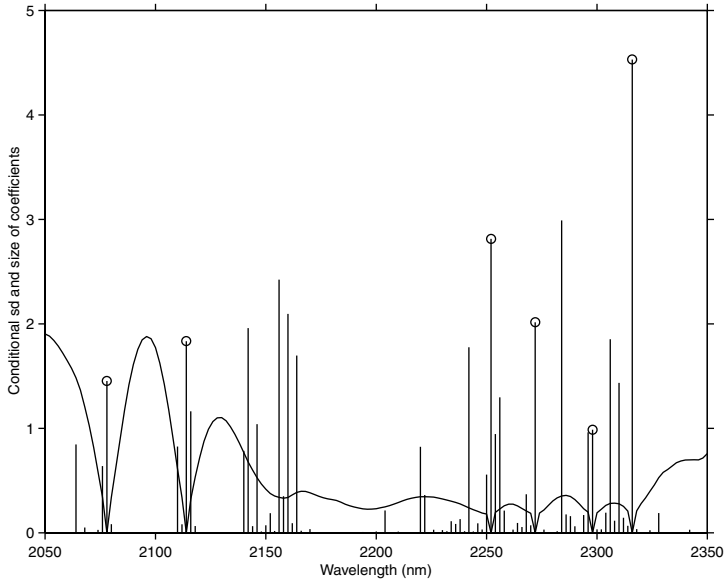

Fig. 7. Decomposition of criterion (22) showing for each wavelength the square root of the conditional variance given the six chosen variables (——) and the square root of the sum of squares of the three Bayes model average coefficients (|): ◯, selected wavelength

It is possible to obtain some insight into why these six variables are chosen from the many in the visited models by a further exploration of criterion (22). In Fig. 7 the full curve is a plot of the square root of the diagonal elements of the conditional variance (in the square brackets in expression (22)) given the selected six variables against the corresponding wavelengths. The vertical lines show the size of the (full) Bayes model average coefficients at each wavelength, the plotted quantity being the square root of the sum of squares of the three coefficients at that wavelength. The six selected wavelengths are marked with a circle at the top of the line. To improve the visual clarity, only the range 2050–2350 nm, which includes all the selected wavelengths and all the visited ones of any importance, has been shown.

The relevance of the plot is that it is the square of the product of the two quantities shown at each wavelength (the coefficient and conditional standard deviation) that contributes to criterion (22). Of course the conditional variance matrix is far from diagonal, so this is only part of the picture, but we can still obtain some insight from the plot.

It is apparent that what the variable selection is not doing is picking the largest coefficients. Rather, it is choosing a spread of wavelengths in important regions. This 'pins down' the conditional variance, which is 0 at an included wavelength and small in some neighbourhood of it, over a whole region and enables the dropping of some other variables even though their coefficients in the Bayes model average may have been sizable.

6. Scaling, invariance and proportional *versus* absolute shrinkage

We have applied a form of regularization, shrinking estimates towards 0. This shrinkage has been achieved by the prior assumption of a subset model γ and by more continuous shrinkage through the prior on B given γ, embodied in H_γ. Because our primary objective was to use the former subset shrinkage we have generally chosen to avoid strong assumptions and have

therefore opted in general in our examples for rather minimal continuous shrinkage. However, several important issues are connected with H_γ, and we shall discuss them here.

Broadly two forms of continuous shrinkage are popularly adopted, which we choose to call *absolute* and *proportional*. Proportional shrinkage has its roots in Stein minimax estimation; see also Copas (1983) and Breiman and Friedman (1997), with Bayesian versions through g-priors (Zellner, 1986). Absolute shrinkage has roots in ridge regression (Hoerl and Kennard 1970a, b). An earlier insightful exposition is given by Dempster (1973). Ridge forms of regression cannot be minimax for estimation of the regression coefficients when the matrix X is ill conditioned; see for example Brown and Zidek (1980). But this should not be viewed as a drawback especially in the context of fewer observations than variables with its inevitable ill conditioning. Even in the context of prediction rather than parameter estimation absolute ridge shrinkage became beneficial in the development of Breiman and Friedman (1997) when $p > n$.

For simplicity consider univariate multiple regression ($q = 1$), with $B \to \beta$. When $p > n$ the x-variables lie in an $(n-1)$-dimensional subspace of Euclidean p-space. To gain more insight we use the singular value decomposition of X^l: with T and V orthonormal,

$$T' X^l V = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_s}, 0, \ldots, 0),$$

with $s = n - 1$. Thus we may write $U = T'Y^l$ and $\theta = V'\beta$ so our univariate model for the part that depends on θ reduces to

$$U_i = \sqrt{\lambda_i}\theta_i + \varepsilon_i, \qquad i = 1, \ldots, s.$$

The prior for θ is such that θ_i are independent $N(0, c\sigma^2/\lambda_i)$ for the g-prior but $N(0, c\sigma^2)$ for the ridge prior. For the g-prior the posterior distribution of θ_i has expectation $\{c/(c+1)\}\hat{\theta}_i^{\mathrm{LS}}$ whereas the ridge expectation is $\{\lambda_i/(\lambda_i + c)\}\hat{\theta}_i^{\mathrm{LS}}$. Thus for any fixed c the ridge estimator shrinks greatly in directions of small eigenvalues, whereas the proportional estimator retains much more of the least squares estimator in ill-conditioned directions. The g-prior is suspect from a fundamentalist Bayesian viewpoint also, since it depends on the data. This might be mitigated by the ancillarity of X^l, but to specify a weak prior on ill-conditioned directions where prior information is important and a strong prior on those well-estimated directions where a strong prior is not needed is perhaps perverse.

Even for prediction, especially when $p > n$, the proportional shrinkage does too little to reduce the instability of small eigenvalues, the point being that

$$E\{(Y^{\mathrm{f}} - \hat{Y}^{\mathrm{f}})'(Y^{\mathrm{f}} - \hat{Y}^{\mathrm{f}})\} = E\{(\theta - \hat{\theta})' V X^{\mathrm{f}'} X^{\mathrm{f}} V' (\theta - \hat{\theta})\}$$

and that $X^{\mathrm{f}'} X^{\mathrm{f}}$ will with probability 1 be different from $X^{l'} X^l$ and training regressors and have features outside the space spanned by X^l. It is the basis of minimax and invariance arguments that the matrices X^l and X^{f} are the same, which they are not.

A further facet of this is the issue of scaling and autoscaling. The use of autoscaling has an effect that is similar to the Bayesian use of a g-prior. Many of the modern technologies that generate almost unlimited explanatory variates do so on a common scale of measurement. In the case of infra-red spectroscopy the common scale is that of absorbance or reflectance at a range of wavelengths. Autoscaling then blows up variates that vary little over the data and scales down variates that vary greatly. The corresponding parameters are reduced or increased respectively, since the product of the variable and coefficient is unchanged. The consequence is vulnerability to blowing up noise; see for example the degradation in predictions of nitrate levels in waste water that is shown in Karlsson *et al.* (1995) as also discussed by Sundberg (1999). This also argues against our historic use of H_γ proportional to $\mathrm{diag}([X^{l'} X^l]^-)$. It argues more for our

use of a prior of the form $H_\gamma(i, j) = \sigma^2\rho^{|i-j|}$ in the wavelet regression application of Brown *et al.* (2001). In this same application the empirical Bayes approach in the form of marginal maximum likelihood was used to estimate the hyperparameters σ^2 and ρ.

A final comment links model averaging to the ridge form of shrinkage. Leamer and Chamberlain (1976) showed that the ridge form of shrinkage is equivalent to a weighted average of all 2^p subsets of models fitted by least squares. They also gave the explicit form of the weights.

7. Commentary

This subset selection procedure provides an attractive justifiable way of choosing those variables that influence prediction while retaining the robustness of model average predictions. It has a different motivation from that of Occam's window that was advocated by Madigan and Raftery (1994). It extends the notion of model averaging while restricting to a subset of variables, optimally chosen. In many substantive applications the reduction in the number of variables and the identity of these chosen variables can offer highly interpretable insight into the informative predictors.

We have also shown how this model average may be approximated by a single model and illustrated this and other choices. The data set involved 700 explanatory variables on 125 observations. The fast form of updating algorithm that was developed should be able to cope with many more variables than this, even the several thousand variables in many current gene expression data sets. A further application to a spectroscopic analysis of biscuits data using wavelets is presented in Vannucci *et al.* (2001).

Finally we have given a critique of standard proportionate shrinkage in favour of more ridge-like regularization.

Many standard classic text-books in statistics would baulk at the excess of variables over observations, but clearly good predictions can be achieved in such circumstances. Although it is a truism that you are always better off with more information, how it is used is crucial.

Acknowledgements

Marina Vannucci was supported by the Texas Higher Advanced Research Program, grant 010366-0075 and National Science Foundation CAREER award DMS-0093208.

References

Barbieri, M. M. and Berger, J. O. (2002) Optimal predictive model selection. *Technical Report 02-02*. Duke University, Durham.

Berger, J. O. (1988) Discussion of the paper by Mitchell and Beauchamp. *J. Am. Statist. Ass.*, **83**, 1033–1034.

Breiman, L. (1996) Heuristics of instability and stabilisation in model selection. *Ann. Statist.*, **24**, 2350–2383.

Breiman, L. and Friedman, J. H. (1997) Predicting multivariate responses in multiple linear regression (with discussion). *J. R. Statist. Soc.* B, **59**, 3–54.

Brown, P. J. (1992) Wavelength selection in multicomponent near-infrared calibration. *J. Chemometr.*, **6**, 151–161.

———(1993) *Measurement, Regression, and Calibration*. Oxford: Clarendon.

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.

———(2001) Bayesian wavelet regressions on curves with application to a spectroscopic calibration problem. *J. Am. Statist. Ass.*, **96**, 398–408.

Brown, P. J. and Mäkeläinen, T. (1992) Regression, sequenced measurements and coherent calibration. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith), pp. 97–108. Oxford: Clarendon.

Brown, P. J., Spiegelman, C. H. and Denham, M. C. (1991) Chemometrics and spectral frequency selection. *Phil. Trans. R. Soc. Lond.* A, **337**, 311–322.

Brown, P. J., Vannucci, M. and Fearn, T. (1998a) Multivariate Bayesian variable selection and prediction. *J. R. Statist. Soc.* B, **60**, 627–641.
———(1998b) Bayesian wavelength selection in multicomponent analysis. *J. Chemometr.*, **12**, 173–182.
Brown, P. J. and Zidek, J. V. (1980) Adaptive multivariate ridge regression. *Ann. Statist.*, **8**, 64–74.
Clyde, M., Desimone, H. and Parmigiani, G. (1996) Prediction via orthogonalised model mixing. *J. Am. Statist. Ass.*, **91**, 1197–1208.
Clyde, M. and George, E. I. (2000) Flexible empirical Bayes estimation for wavelets. *J. R. Statist. Soc.* B, **62**, 681–698.
Copas, J. B. (1983) Regression, prediction and shrinkage (with discussion). *J. R. Statist. Soc.* B, **45**, 311–354.
Dawid, A. P. (1981) Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, **68**, 265–274.
———(1988) The infinite regress and its conjugate analysis. In *Bayesian Statistics 3* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 95–110. Oxford: Clarendon.
Dempster, A. P. (1973) Alternatives to least squares in multiple regression. In *Multivariate Statistical Inference* (eds D. G. Kabe and R. P. Gupta), pp. 25–40. New York: Elsevier.
Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion). *J. R. Statist. Soc.* B, **57**, 45–97.
Fernández, C., Ley, E. and Steel, M. F. J. (2001) Benchmark priors for Bayesian model averaging. *J. Econometr.*, **100**, 381–427.
Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.*, **7**, 457–472.
George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection. *Statist. Sin.*, **7**, 339–373.
Hoerl, A. E. and Kennard, R. W. (1970a) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.
———(1970b) Ridge regression: applications to nonorthogonal problems. *Technometrics*, **12**, 69–82.
Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999) Bayesian model averaging: a tutorial. *Statist. Sci.*, **14**, 382–417.
Karlsson, M., Karlberg, B. and Olsson, R. J. O. (1995) Determination of nitrate in municipal waste water by UV spectroscopy. *Anal. Chem. Acta*, **312**, 107–113.
Lanza, E. and Li, B. W. (1984) Application of near infrared spectroscopy for predicting the sugar content of fruit juices. *J. Food Sci.*, **49**, 995–998.
Leamer, E. E. and Chamberlain, G. (1976) A Bayesian interpretation of pretesting. *J. R. Statist. Soc.* B, **38**, 85–94.
Lindley, D. V. (1968) The choice of variables in multiple regression (with discussion). *J. R. Statist. Soc.* 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. Am. Statist. Ass.*, **89**, 1535–1546.
Madigan, D. and York, J. (1995) Bayesian graphical models for discrete data. *Int. Statist. Rev.*, **63**, 215–232.
Mitchell, T. J. and Beauchamp, J. J. (1988) Bayesian variable selection in linear regression. *J. Am. Statist. Ass.*, **83**, 1023–1036.
Osborne, B. G., Fearn, T. and Hindle, P. H. (1993) *Practical NIR Spectroscopy*. Harlow: Longman.
Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997) Bayesian model averaging for linear regression models. *J. Am. Statist. Ass.*, **92**, 179–191.
San Martini, A. and Spezzaferri, F. (1984) A predictive model selection criterion. *J. R. Statist. Soc.* B, **46**, 296–303.
Seber, G. A. F. (1984) *Multivariate Observations*. New York: Wiley.
Sha, N., Vannucci, M. and Brown, P. J. (2002) Bayesian variable selection in multinomial models with application to spectral data and DNA microarrays. *Technical Report UKC/IMS/02/05*. University of Kent at Canterbury, Canterbury.
Smith, A. F. M. (1973) A general Bayesian linear model. *J. R. Statist. Soc.* B, **35**, 67–75.
Sundberg, R. (1999) Multivariate calibration—direct and indirect regression methodology. *Scand. J. Statist.*, **26**, 161–207.
Vannucci, M., Brown, P. J. and Fearn, T. (2001) Predictor selection for model averaging. In *Bayesian Methods with Applications to Science, Policy and Official Statistics* (eds E. I. George and P. Nanopoulos), pp. 553–562. Luxemburg: Eurostat.
West, M., Nevins, J. R., Marks, J. R., Spang, R. and Zuzan, H. (2000) Bayesian regression analysis in the "large p, small n" paradigm with application in DNA microarray studies. *Technical Report*. Duke University, Durham.
Zellner, A. (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques—Essays in Honour of Bruno de Finetti* (eds P. K. Goel and A. Zellner), pp. 233–243. Amsterdam: North-Holland.