Biometrika (1999), **86,** 3, *pp.* 635–648

Printed in Great Britain

The choice of variables in multivariate regression: A non-conjugate Bayesian decision theory approach

BY PHILIP J. BROWN

Institute of Mathematics and Statistics, University of Kent at Canterbury, Canterbury, Kent, CT2 7NF, U.K.

philip.j.brown@ukc.ac.uk

TOM FEARN

Department of Statistical Science, University College London, London WC1E 6BT, U.K.

tom@stats.ucl.ac.uk

AND MARINA VANNUCCI

Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.

mvannucci@stat.tamu.edu

SUMMARY

We consider the choice of explanatory variables in multivariate linear regression. Our approach balances prediction accuracy against costs attached to variables in a multivariate version of a decision theory approach pioneered by Lindley (1968). We also employ a non-conjugate proper prior distribution for the parameters of the regression model, extending the standard normal-inverse Wishart by adding a component of error which is unexplainable by any number of predictor variables, thus avoiding the determinism identified by Dawid (1988). Simulated annealing and fast updating algorithms are used to search for good subsets when there are very many regressors. The technique is illustrated on a near infrared spectroscopy example involving 39 observations and 300 explanatory variables. This demonstrates the effectiveness of multivariate regression as opposed to separate univariate regressions. It also emphasises that within a Bayesian framework more variables than observations can be utilised.

Some key words: Bayesian decision theory; Determinism; Multivariate regression; Near infrared spectroscopy; Non-conjugate distribution; Prediction; Quadratic loss; Simulated annealing; Utility.

1. INTRODUCTION

Choice of regressor variables in linear regression has attracted considerable attention in the literature, from forward, backward and stepwise regression, model choice criteria such as Akaike's information criterion, to Bayesian techniques. We will focus on the Bayesian decision theory framework, first given by Lindley (1968) for univariate multiple regression, where costs attach to the inclusion of regressor variables. Here it is required to predict a future vector observation Y^f comprising r components. Predictions are judged by quadratic loss to which is added a cost penalty on the regressor variables,

$x_{i_1}^f, x_{i_2}^f, \ldots, x_{i_p}^f$, a subset of the full set of $q \geqslant p$ variables $\{x_i\}_1^q$. This cost typically increases as variables are added.

There is a large Bayesian literature on model choice and variable selection in the multiple regression model. Most approaches focus on probabilistic fit, see for example George & McCulloch (1997), with multivariate extensions in Brown, Vannucci & Fearn (1998a, b), and earlier work by Mitchell & Beauchamp (1988) following on Lempers (1971). Raftery, Madigan & Hoeting (1997) use model averaging over a subset of models restricted by 'Occam's Window'. Model averaging is also used by Clyde, Desimone & Parmigiani (1996) but after first orthogonalising the design space. Bernardo & Smith (1994, Ch. 6) have suggested a utility formulation that is approximated by crossvalidatory fit; see also a Nottingham Trent University technical report by J. M. Marriott, N. M. Spencer and A. N. Pettitt. Our approach contrasts to that of George & McCulloch in that we do not use a mixture prior distribution. We omit variables not because we believe their coefficients are zero, but because they cost too much relative to their predictive benefit.

The formulation we adopt assumes a joint normal distribution of the r-variate response Y $(r \times 1)$ and the full set of q regressor variables, the $(1 \times q)$ row vector $X_q = (x_1, x_2, \ldots, x_q)$. The formulation assumes random regressors, with a joint distribution the same in both training and prediction data. By focusing on the joint distribution of the response and regressor variables rather than the regression parameters in the conditional distribution of Y given X_q, we may straightforwardly assign prior distributions which cohere over different p-variate submodels.

Dawid (1988) defined determinism in the infinite regress $(q \to \infty)$ as a feature of those models in which the parameter values are such that the conditional variance of $Y|X_q$ tends to zero. Essentially it is possible to predict the response perfectly by adding enough variables, always with the proviso of knowing the regression parameters. He showed that the normal-inverse Wishart natural conjugate prior distribution embodies an implicit determinism. Mäkeläinen & Brown (1988) suggested a simple device to remedy this perceived fault. They included a component in Y which is unexplainable by X even when $q \to \infty$. More recently Fang & Dawid (2000) have used this idea to develop a non-conjugate approach to multiple regression on many variables, and we use elements of their development. Even though we will not be directly concerned with allowing an infinite set of regressors but rather with comparing different subsets of $p \leqslant q$ of the regressor variables, we prefer to avoid the conjugate prior and its limitations.

Our development has a number of features including the following: a multivariate response, a proper non-conjugate prior distribution avoiding determinism, and computational techniques for coping with very many variables. None of these items is new, but their joint use in variable selection is.

Section 2 describes the model and develops the predictive mean and covariance matrix for the non-conjugate Gaussian model. This is sufficient for subsequent utility and decision theoretic formulations in § 3. Computational aspects utilising QR decompositions and simulated annealing to maximise the expected utility are described in § 4. Section 5 describes an application to choosing wavelengths from a near infrared spectrum to predict various ingredients of biscuit dough pieces for eventual on-line production control.

2. The Bayes model

2·1. *Matrix-variate distributions*

We first review a general matrix-variate notation for Gaussian and related distributions, i.e. inverse Wishart and matrix-variate T, which greatly simplifies calculations, avoiding

the need to string matrices as vectors and consequent Kronecker product covariance structures.

We shall follow the notation introduced by Dawid (1981) for matrix-variate distributions. This has the advantage of preserving the matrix structures. It redefines the degrees of freedom as shape parameters for both inverse Wishart and matrix-variate T, to allow notational invariance under marginalisation and very easy symbolic Bayesian manipulations.

We use calligraphic lettering to distinguish the notation from more standard ones. Corresponding probability density functions for the random matrices symbolically introduced are given in Brown (1993, Appendix A).

In the case of the matrix-normal $\mathcal{N}(.,.)$ both arguments relate to covariances; the first does not relate to the mean as in the standard $N(.,.)$ multivariate normal notation. The mean in the case of the matrix-normal notation is specified separately as an additional term. With U a matrix having independent standard normal entries, $M+\mathcal{N}(\Gamma,\Sigma)$ will stand for a matrix-normal distribution of $V=M+A'UB$, where M, A, B are fixed matrices satisfying $A'A=\Gamma$ and $B'B=\Sigma$. Thus M is the matrix mean of V, and $\gamma_{ii}\Sigma$ and $\sigma_{jj}\Gamma$ are the covariance matrices of the ith row and jth column, respectively, of V. If V is a column vector then the matrix normal and multivariate normal equivalent notations are $\mathcal{N}(\Gamma,1)$ and $N(0,\Gamma)$; if V is a row vector then the matrix normal notation is $\mathcal{N}(1,\Gamma)$.

If U is of order $n\times p$ with $n\geqslant p$, the notation $\mathcal{IW}(\delta;\Sigma)$, with $\delta=n-p+1$, will stand for the distribution of $W=B'(U'U)^{-1}B$, an inverse Wishart distribution. This random matrix has expectation $\Sigma/(\delta-2)$. The shape parameter δ differs from the more conventional degrees of freedom, and may be generalised, using the density function, to take on any positive real value. Note that with $W\sim\mathcal{IW}(\delta;\Sigma)$ and W_{11} a principal submatrix of W then $W_{11}\sim\mathcal{IW}(\delta;\Sigma_{11})$, and the shape parameter δ is unchanged; this would not be the case with more standard inverse Wishart notations (Press, 1982, Ch. 5).

The matrix-variate T distribution $M+\mathcal{T}(\delta;\Sigma,Q)$ is the distribution of T where T follows the $M+\mathcal{N}(\Gamma,\Sigma)$ distribution conditional on Σ and $\Sigma\sim\mathcal{IW}(\delta;Q)$. Marginal and conditional distributions of the matrix-T are also matrix-T. Marginalisation does not affect the shape parameter. For example, if T $(p\times q)$, distributed as $\mathcal{T}(\delta;P,Q)$, is partitioned into $T'=(T_1',T_2')$ where T_i is $(p_i\times q)$, for $i=1,2$, then

$$T_2\sim\mathcal{T}(\delta;P_{22},Q) \tag{1}$$

and, conditional on $T_2=t_2$,

$$T_1-P_{12}P_{22}^{-1}t_2\sim\mathcal{T}(\delta+p_2;P_{11.2},Q+t_2'P_{22}^{-1}t_2), \tag{2}$$

with $P_{11.2}=P_{11}-P_{12}P_{22}^{-1}P_{21}$. Also note that $T'\sim\mathcal{T}(\delta;Q,P)$, so that conditioning on a set of columns follows similarly.

2·2. *The model*

We suppose the response Y $(1\times r)$ is the sum of an unobservable variable η and an unobservable error component α, independent of each other; the joint distribution of η and the explanatory variables X_q $(1\times q)$ is normal; α has an independent normal distribution. Thus the model is

$$Y=\eta+\alpha, \tag{3}$$

$$(\eta,X_q)\sim\mathcal{N}(1,\Sigma_{r+q}),\quad \alpha\sim\mathcal{N}(1,\Phi), \tag{4}$$

with (η,X_q) and α conditionally independent, given $(r+q)\times(r+q)$ and $r\times r$ covariance

matrices Σ_{r+q} and Φ, respectively. The means of all these variables are taken to be zero. In practice we centre both X and Y; see § 2·5.

If γ is a binary q-vector that identifies subsets,

$$\gamma_i = 1 \leftrightarrow x_i \quad \text{included}, \tag{5}$$

then the number of variables included in a submodel is $p = |\gamma|$, the number of ones in γ.

A particular submodel γ involving p of the explanatory variables has the distribution in (3), (4) with X_q and Σ_{r+q} replaced by X_γ, the row vector of p variables, and $\Sigma_{r+\gamma}$, the appropriate $(r+p)\times(r+p)$ submatrix of Σ_{r+q}. The joint normality of Y and X_q implies that

$$Y|X_\gamma \sim X_\gamma B_\gamma + \mathcal{N}(1, \Delta_\gamma), \tag{6}$$

$$X_\gamma \sim \mathcal{N}(1, \Sigma_{\gamma\gamma}), \tag{7}$$

where the joint covariance matrix $\Sigma_{r+\gamma}$ is partitioned as the $(r+p)\times(r+p)$ matrix

$$\Sigma_{r+\gamma} = \begin{pmatrix} \Sigma_{00} & \Sigma_{0\gamma} \\ \Sigma_{\gamma 0} & \Sigma_{\gamma\gamma} \end{pmatrix},$$

$B_\gamma = \Sigma_{\gamma\gamma}^{-1}\Sigma_{\gamma 0}$ ($p \times r$), and, with $\Sigma_{00.\gamma} = \Sigma_{00} - \Sigma_{0\gamma}\Sigma_{\gamma\gamma}^{-1}\Sigma_{\gamma 0}$, then $\Delta_\gamma = \Sigma_{00.\gamma} + \Phi$ ($r \times r$).

We will now assign an inverse Wishart prior distribution for Σ_{r+q}, and by implication for $\Sigma_{r+\gamma}$. Let

$$\Sigma_{r+q} \sim \mathcal{IW}(\delta;\, Q_{r+q}) \tag{8}$$

with shape parameter $\delta > 0$ and Q_{r+q} an $(r+q)\times(r+q)$ positive definite scale matrix.

Assigning the obvious inverse Wishart prior distribution to Φ would lead to intractable posterior distributions. In the interests of deriving an analytical solution we make the following simplifying assumption.

Assumption 1. The unexplainable error covariance matrix Φ is proportional to the residual explainable covariance matrix $\Sigma_{00.q}$. Specifically, let

$$w_q I_r = \Sigma_{00.q}(\Sigma_{00.q} + \Phi)^{-1}$$

for some scalar w_q.

In the absence of information to the contrary this offers considerable simplification. In the case of univariate Y it is non-restrictive, amounting merely to a reparameterisation. In the multivariate case contrary or confirmatory information would have to be supplied externally. Our training data cannot supply such evidence as the sufficient statistics for the Gaussian distribution can estimate $\Sigma_{00.q} + \Phi$ but not either component separately. All the arguments below are made conditional on the scalar parameter w_q. In our application we will prespecify w_q from external knowledge of the application area. Note that w_q does depend on q. In particular if we are to avoid determinism, w_q must tend to zero as $q \to \infty$.

The training or learning data consist of n independent realisations from (3) and (4) leading to Y^l ($n \times r$) and X_q^l ($n \times q$). Here and elsewhere, the superfix l is used to note explicitly that n observations of the learning data are involved, whereas f as a superfix denotes a future observation. Interest focuses on prediction of Y^f for a future case with $Y^f = \eta^f + \alpha^f$, and (η^f, X_q^f) and α^f independent realisations of model (3) and (4) conditional on the covariance matrices (Σ_{r+q}, Φ). From the characterisation of the matrix-T, we have

that

$$\begin{pmatrix} \eta^f & X_q^f \\ \eta^l & X_q^l \end{pmatrix} \tag{9}$$

is distributed as $\mathscr{T}(\delta; I_{n+1}, Q_{r+q})$.

2·3. *Bayes prediction*

Our development here follows in outline that of Fang & Dawid (2000). Suppose we wish to use a p-variate subset γ of the q regressor variables for prediction. Consider the quadratic prediction loss

$$\mathscr{L}(Y^f, \hat{Y}^f) = (Y^f - \hat{Y}^f)' L (Y^f - \hat{Y}^f) \tag{10}$$

$$= \mathrm{tr}\{L(Y^f - \hat{Y}^f)(Y^f - \hat{Y}^f)'\}, \tag{11}$$

with L any $r \times r$ positive definite matrix of weight constants. The residual sum of products matrix in (11) may be termed the matrix quadratic loss. The Bayes predictor $\hat{Y}^f$ is the predictor of Y^f assuming all variables have been measured in the learning data Y^l, X_q^l but that only the selection γ of the X_q^f is available for prediction, and is given as

$$\hat{Y}^f = E(Y^f | X_\gamma^f, X_q^l, Y^l) = E(\eta^f | X_\gamma^f, X_q^l, Y^l), \tag{12}$$

since $Y^f = \eta^f + \alpha^f$ and

$$E(\alpha^f | X_\gamma^f, X_q^l, Y^l) = 0.$$

This latter fact follows from the joint distribution of α^f and X_γ^f being independent with means zero conditional on X_γ^l, Y^l. This turns on the distribution first conditional on Σ_{r+q}, when

$$\alpha^f \sim N(0, w_q \Sigma_{00.q}), \quad X_q^f \sim N(0, \Sigma_{qq}),$$

which in turn implies

$$X_\gamma^f \sim N(0, \Sigma_{\gamma\gamma}).$$

Finally Σ_{qq} and $\Sigma_{00.q}$ are a priori and a posteriori conditionally independent, given X_q^l, Y^l, a characteristic Bartlett decomposition property of the inverse-Wishart; see (28) and for example Brown, Le & Zidek (1994).

Now we condition the right-hand side of (12) on the unobserved 'error-free' variables η^l $(n \times r)$, so that

$$E(Y^f | X_\gamma^f, X_q^l, Y^l) = E\{E(\eta^f | X_\gamma^f, \eta^l, X_q^l, Y^l)\}. \tag{13}$$

The inner conditional expectation can be simplified to

$$E(\eta^f | X_\gamma^f, \eta^l, X_q^l), \tag{14}$$

since Y^l is redundant when η^l is known in this conditional expectation. It reappears though in the outer conditional expectation in (13).

Our first task then is to evaluate (14). We can obtain the conditional distribution of η^f conditional on $X_\gamma^f, \eta^l, X_q^l$ by two applications of conditioning through result (2) applied to the array (9), first directly to the array and then to its transpose.

The first application of (2) gives

$$(\eta^f, X_q^f) | \eta^l, X_q^l \sim \mathscr{T}\{\delta + n; 1, Q_{r+q} + (\eta^l, X_q^l)'(\eta^l, X_q^l)\}. \tag{15}$$

Marginalising this distribution to the subset γ containing p of q variables, we have the predictive distribution

$$(\eta^f, X^f_\gamma)|\eta^l, X^l_q \sim \mathscr{T}(\delta + n;\, 1,\, P_{r+\gamma}), \tag{16}$$

with the $(r+p)\times(r+p)$ scale matrix $P_{r+\gamma}$ given as

$$P_{r+\gamma} = Q_{r+\gamma} + (\eta^l, X^l_\gamma)'(\eta^l, X^l_\gamma), \tag{17}$$

with $Q_{r+\gamma}$ the appropriate $(r+p)\times(r+p)$ submatrix of the $(r+q)\times(r+q)$ matrix Q_{r+q} in definition (8). If we partition the matrices $P_{r+\gamma}$ and $Q_{r+\gamma}$ into blocks corresponding to the r responses and p variables, a second application of (2) to the transpose of (16) gives

$$\eta^f|X^f_\gamma, \eta^l, X^l_q \sim X^f_\gamma P^{-1}_{\gamma\gamma} P_{0\gamma} + \mathscr{T}(\delta + n + p;\, 1 + X^f_\gamma P^{-1}_{\gamma\gamma}(X^f_\gamma)',\, P_{00.\gamma}), \tag{18}$$

where

$$P_{00} = Q_{00} + (\eta^l)'\eta^l, \tag{19}$$

$$P_{0\gamma} = Q_{0\gamma} + (\eta^l)'X^l_\gamma, \tag{20}$$

$$P_{\gamma\gamma} = Q_{\gamma\gamma} + (X^l_\gamma)'X^l_\gamma, \tag{21}$$

and $P_{00.\gamma} = P_{00} - P_{0\gamma}P^{-1}_{\gamma\gamma}P_{\gamma 0}$. Thus the conditional mean of η^f as given in (14) is

$$E(\eta^f|X^f_\gamma, \eta^l, X^l_q) = X^f_\gamma P^{-1}_{\gamma\gamma} P_{0\gamma},$$

and, averaging this over the distribution of η^l given X^l_q, Y^l, since it is independent of X^f_γ, according to the outer expectation of the right-hand side of (13), we have

$$\hat{Y}^f = E(Y^f|X^f_\gamma, X^l_q, Y^l) = X^f_\gamma P^{-1}_{\gamma\gamma} P_{0\gamma}(q), \tag{22}$$

where

$$P_{0\gamma}(q) = Q_{0\gamma} + E(\eta^l|X^l_q, Y^l)'X^l_\gamma. \tag{23}$$

The Bayes predictor for quadratic loss is given by (22) and is completely specified up to the calculation of $E(\eta^l|X^l_q, Y^l)$, the expectation of the latent 'true' Y^l matrix given the full training data. This will be the focus of § 2·4. This conditional expectation also turns out to be the only important missing ingredient in calculating the minimised value of the quadratic loss.

2·4. *Latent response*

We wish to calculate $E(\eta^l|X^l_q, Y^l)$. Conditional on Φ and Σ_{r+q} we have

$$Y^l|\eta^l \sim \eta^l + \mathscr{N}(I_n, \Phi), \tag{24}$$

$$\eta^l|X^l_q \sim X^l_q B_q + \mathscr{N}(I_n, \Sigma_{00.q}), \tag{25}$$

where $B_q = \Sigma^{-1}_{qq}\Sigma_{q0}$ and $\Sigma_{00.q} = \Sigma_{00} - \Sigma_{0q}\Sigma^{-1}_{qq}\Sigma_{q0}$. For given Φ and Σ_{r+q}, (25) acts as a prior distribution for η^l, with (24) providing the likelihood. Thus,

$$\eta^l|Y^l, X^l_q, \Sigma_{r+q}, \Phi \sim \eta^* + \mathscr{N}(I_n, V^*), \tag{26}$$

where

$$\eta^* = (Y^l\Phi^{-1} + X^l_q B_q \Sigma^{-1}_{00.q})(\Phi^{-1} + \Sigma^{-1}_{00.q})^{-1}, \tag{27}$$

$$V^* = (\Phi^{-1} + \Sigma^{-1}_{00.q})^{-1}.$$

This distribution depends on Σ_{r+q} only through B_q and $\Sigma_{00.q}$. The next step is to find the posterior distribution of B_q given Y^l, X_q^l, $\Sigma_{00.q}$, Φ, and then marginalise over this distribution.

The inverse Wishart prior distribution for Σ_{r+q} factorises into

$$\pi(B_q, \Sigma_{00.q})\pi(\Sigma_{qq}), \tag{28}$$

with

$$B_q|\Sigma_{00.q} \sim Q_{qq}^{-1}Q_{q0} + \mathcal{N}(Q_{qq}^{-1}, \Sigma_{00.q}) \tag{29}$$

(Dawid, 1988, Lemma 2), and

$$\Sigma_{00.q} \sim \mathcal{IW}(\delta + q;\, Q_{00.q}), \tag{30}$$

with $Q_{00.q} = Q_{00} - Q_{0q}Q_{qq}^{-1}Q_{q0}$ obtained from partitioning Q_{r+q} in (8). The likelihood for Σ_{r+q} factors in a corresponding way, with the relevant part being

$$Y^l|X_q^l \sim X_q^l B_q + \mathcal{N}(I_n, \Sigma_{00.q} + \Phi). \tag{31}$$

The aim now is to combine (29) and (31) to give a posterior distribution for B_q.

With Assumption 1 and remembering that w_q is a scalar, we can rewrite (29) as

$$B_q|\Sigma_{00.q} \sim Q_{qq}^{-1}Q_{q0} + \mathcal{N}(w_q Q_{qq}^{-1}, \Sigma_{00.q} + \Phi), \tag{32}$$

with the same row covariance matrix as in (31). Hence, using for example Appendix B of Brown (1993) we get a posteriori, given w_q, $\Sigma_{00.q}$, Y^l, X_q^l,

$$B_q - B_q^* \sim \mathcal{N}(\{Q_{qq} + w_q(X_q^l)'X_q^l\}^{-1}, \Sigma_{00.q}), \tag{33}$$

where

$$B_q^* = \{Q_{qq} + w_q(X_q^l)'X_q^l\}^{-1}\{Q_{q0} + w_q(X_q^l)'Y^l\}. \tag{34}$$

Now we can marginalise (26) over the posterior distribution of B_q given by (33). First note that, under Assumption 1, equation (27) may be written as

$$\eta^* = w_q Y^l + (1 - w_q)X_q^l B_q,$$

and also $V^* = (1 - w_q)\Sigma_{00.q}$ in (26). Thus, given w_q, $\Sigma_{00.q}$, Y^l, X_q^l the posterior distribution of η^l is given as

$$\eta^{**} + \mathcal{N}(I_n + (1 - w_q)X_q^l\{Q_{qq} + w_q(X_q^l)'X_q^l\}^{-1}(X_q^l)', (1 - w_q)\Sigma_{00.q}), \tag{35}$$

with

$$\eta^{**} = w_q Y^l + (1 - w_q)X_q^l B_q^*. \tag{36}$$

The posterior mean η^{**} still depends on w_q, but we choose not to try to average over the posterior distribution of w_q. Instead we shall specify w_q a priori in our application. It is possible to update genuinely through likelihood (31) the distribution of $\Sigma_{00.q} + \Phi$, or equivalently $\Sigma_{00.q}$ for given w_q, but not w_q itself.

Thus, to recap, we have found in (36) the quantity $\eta^{**} = E(\eta^l|X_q^l, Y^l)$ needed in (23) to evaluate $\hat{Y}^f$, the Bayes predictor of Y^f given by (22).

2·5. *Prior structures*

The hyperparameters that need to be specified for any application are the matrices Q_{r+q} in (8) and the scalar w_q in Assumption 1. Typically we assume $Q_{q0} = 0$. Specification of

Q_{00} is not needed to evaluate the Bayes predictor (22) through (36) and (34), nor to compare different subsets; see § 3. Also the simplest prior structure would take $Q_{qq} = kI_q$, where k is a scalar to be specified, so that in particular $Q_{\gamma\gamma} = kI_p$. The estimator (34) then gives a ridge regression estimator for each response. In the absence of informative prior knowledge, this motivates a semi-automatic choice of k as the median of the generalised crossvalidation estimates as given by Golub, Heath & Wahba (1979).

Other structures for Q_{qq} may be appropriate, for example structures embodying correlation, perhaps simply through an autoregressive AR(1) process. Such relatively simple structures also allow simple forms of matrix square root; see for example Brown & Mäkeläinen (1992, § 5).

Throughout we have assumed that both X and Y variables have mean zero. This is an unnatural but inconsequential assumption in that in reality we have centred all the variables by their means in the learning data. This implicitly corresponds to independent vague priors, that is proportional to a constant over the real line for the mean of each variable.

Thoughts about the indeterminism parameter w_q should centre on the likely size of the unexplainable prediction variance.

3. Prediction accuracy and decision

3·1. *Prediction covariance and loss*

We noted in § 2·3 that our decision rule conditions on the full q-variable learning data X_q^l and Y^l, but looks at subsets of variables identified by γ for prediction. This same paradigm in univariate regression is formally developed by Lindley (1968) for a random experiment and later extended by Brooks (1974) to controlled experiments and mixed, controlled and random experiments. Our analysis assumes the original random paradigm in which both training X and X^f are generated by the same normal random mechanism.

The matrix $P_{r+\gamma}$ in (17) is the scale matrix of the multivariate distribution of (η^f, X_γ^f) given η^l, Y^l, X_q^l, which does not actually involve Y^l because of the conditioning on η^l. The covariance matrix is $P_{r+\gamma}/(\delta + n - 2)$. Now $Y^f = \eta^f + \alpha^f$, and, whereas the $(1 \times r)$ error vector α^f does not contribute to the conditional expectation of Y^f in (12), it does contribute to the conditional covariance of Y^f. If we include this uncertainty due to α^f, it is straightforward to derive the covariance matrix of (Y^f, X_γ^f) given Y^l, X_q^l. We assume w_q is fixed. The independent error α^f added to η^f affects only the $(r \times r)$ response covariance structure and hence only augments P_{00} given by (19). The response covariance matrix is

$$\operatorname{cov}(Y^f, X_\gamma^f \mid Y^l, X_q^l) = P_{r+\gamma}^*,$$

where $P_{r+\gamma}^*$ is partitioned as $P_{r+\gamma}$ given by equations (19), (20) and (21) with

$$P_{00}^* = [Q_{00} + E\{(\eta^l)'\eta^l \mid Y^l, X_q^l\}]/(\delta + n - 2) + E(\Phi \mid Y^l, X_q^l), \tag{37}$$

$$P_{0\gamma}^* = \{Q_{0\gamma} + E(\eta^l \mid X_q^l, Y^l)'X_\gamma^l\}/(\delta + n - 2), \tag{38}$$

$$P_{\gamma\gamma}^* = \{Q_{\gamma\gamma} + (X_\gamma^l)'X_\gamma^l\}/(\delta + n - 2), \tag{39}$$

with (38) a scaled version of $P_{0\gamma}(q)$ in (23). Also (39) is (21) scaled. Such formulae derive as follows. For example, for (37),

$$\operatorname{cov}(Y^f, Y^f \mid Y^l, X_q^l) = E\{\operatorname{cov}(Y^f, Y^f \mid Y^l, X_q^l, \Sigma_{r+q})\} = E(\Sigma_{00} + \Phi \mid Y^l, X_q^l).$$

Additionally we note that (37) does not change with the selection of regressors, and fortuitously we can avoid calculating it.

The predictor $\hat{Y}^f$ of Y^f is linear in X_γ^f and hence from standard results, see for example Mouchart & Simar (1980), the minimised quadratic loss, (10) with $L = I$, of the Bayes predictor is $\mathrm{tr}\{R(\gamma)\}$, where

$$R(\gamma) = P_{00}^* - P_{0\gamma}^* P_{\gamma\gamma}^{*-1} P_{\gamma 0}^*, \tag{40}$$

the scaled residual sum of products matrix. Only the second term on the right-hand side varies with the selection of regressors. Thus the relative merit of different subsets can be assessed without the calculation of P_{00}^*, as already intimated. This will considerably simplify and speed up calculations in § 4.

3·2. *Utility formulation*

We wish to predict an r-variate response through the Bayes predictor (22) with consequent prediction generalised quadratic loss of (40). When r is greater than 1 a Bayesian decision theoretic formulation requires us to use a scalar loss function such as the quadratic loss (10) and then add a terminal cost, a function of the cost of retaining the selected p variables.

We first assume that the r dependent variables have been put on scales that reflect their relative importance, so that the same quantitative inaccuracy on any variable is of equal loss. This may entail initial scaling of the r variables, and is allowed by our freedom to choose L in (10). One could of course take a different tack and scale one variable to dominate the rest, but a balance of losses is more natural in the examples we have considered. Given that this initial scaling of the variables has been done we wish to predict all variables accurately. Loss function (10) amounts to an amalgamated loss across the r components. Intuitively one might have wished to focus on the maximum of the prediction variances. Strict application of Bayes decision theory would then be less straightforward as the Bayes predictor would no longer be the mean of the posterior distribution.

We now add the terminal cost of using a particular γ subset of the q regressors, with a general terminal cost $g(\gamma)$. Our overall loss is then

$$\mathrm{tr}\{R(\gamma)\} + g(\gamma), \tag{41}$$

assuming that after appropriate scaling L is the $r \times r$ identity matrix. The simplest form of $g(\gamma)$ is additive with a cost c_i of including variable X_i^f in prediction. If these costs have a common value, c, then the criterion reduces to

$$C(\gamma) = \mathrm{tr}\{R(\gamma)\} + cp. \tag{42}$$

It is however easy to envisage applications where the more general form (41) would be required. There may for example be a premium on restricting the number p of regressors to be at most some small number. It would not add significantly to the computations to use more general forms for the costs.

4. Computational aspects

4·1. *Synthesis*

Suppose the posterior mean η^{**} of η has been computed. This can be done once and for all. We will see that we can then address an equivalent minimisation problem by modifying the first term, not involving γ, in (40) used in the overall loss (42). Both these steps may profitably be implemented by QR decompositions of appropriately defined

matrices. We assume throughout that w_q in Assumption 1 is prespecified. For simplicity of exposition we assume that the prior matrix $Q_{q0} = 0$, a $(q \times r)$ matrix of zeros, so that in particular $Q_{\gamma 0} = 0$.

4·2. *Computation of latent response*

From (36) the $(n \times r)$ estimated latent response matrix η^{**} is a simple weighted average. The only ingredient in this weighted average requiring computation is B_q^* given by (34). Both these calculations involve all q variables and not the chosen subset γ. A little rearrangement of (34) enables one to see that this regression matrix is the least squares solution of the new augmented regression problem of Y^* on X^*, where

$$Y^* = \begin{pmatrix} Y^l \\ 0 \end{pmatrix}, \quad X^* = \begin{pmatrix} X_q^l \\ (Q_{qq}/w_q)^{\frac{1}{2}} \end{pmatrix}$$

are $(n+q) \times r$ and $(n+q) \times q$ matrices, respectively. Thus the B_q^* matrix can be calculated using the QR decomposition, avoiding the need to 'square' variables.

4·3. *Computation of loss*

Let us now define an augmented regression of $\tilde{Y}$ on $\tilde{X}$, where

$$\tilde{Y} = \begin{pmatrix} \eta^{**} \\ 0 \end{pmatrix}, \quad \tilde{X} = \begin{pmatrix} X_\gamma^l \\ Q_{\gamma\gamma}^{\frac{1}{2}} \end{pmatrix}, \tag{43}$$

$(n+p) \times r$ and $(n+p) \times p$ matrices, respectively. This regression matrix gives $P_{0\gamma}(q)P_{\gamma\gamma}^{-1}$ and hence the Bayes predictor (22). What is more, the residual sum of products matrix differs only by an additive constant, not depending on γ, from $R(\gamma)$ given in (40). Hence one can use the QR decomposition; QR-delete and QR-insert algorithms can be used to remove or add a variable (Seber, 1984, Ch. 10.1.1b), avoiding the need to 'square' variables.

4·4. *A stochastic search*

With q variables there are 2^q possible subsets, typically too many for a complete evaluation of expected losses. What we need is a search method that has a good chance of finding at least some of the best subsets. We employ simulated annealing.

We use the binary q-vector γ that identifies subsets. The search algorithm moves sequentially through the space of all possible binary vectors trying to find good ones, i.e. low cost ones. Our cost function is given by (42), with p being the number of nonzero components of γ. At each step the algorithm constructs γ^{new} from γ^{old} by choosing at random between 3 types of move, as follows.

Move 1: *A*. Add a variable by choosing at random a 0 in γ^{old} and changing it to a 1. Move chosen with probability P_A.

Move 2: *D*. Delete a variable by choosing at random a 1 in γ^{old} and changing it to a 0. Move chosen with probability P_D.

Move 3: *S*. Swap two variables by choosing independently at random a 0 and a 1 in γ^{old} and changing both of them. Move chosen with probability $1 - P_A - P_D$.

At the boundaries, with all variables included or no variable present, only deletion or

addition is possible respectively, and we choose this move with probability 1. At each step $d = C(\gamma^{\text{new}}) - C(\gamma^{\text{old}})$ is calculated. If $d < 0$, γ^{new} is accepted. Otherwise it is accepted with probability $\exp(-d/T)$, where T is a control parameter called temperature. We chose a cooling schedule of the form $T_i = \rho T_{i-1}$ $(0 < \rho < 1)$, reducing temperature at each iteration i. Allowing moves to 'worse' subsets may help to avoid local minima. The starting configuration described below involves specifying parameter θ and a random set of chosen wavelengths with expected size $q\theta$. We stop when the temperature becomes so low that the system essentially stops moving. Every m steps we calculate an acceptance ratio AR, the proportion of m steps that have been accepted, and stop if $\text{AR} \leqslant \tau$.

Let $\tilde{\gamma}$ be the vector with minimum cost given by the stochastic search. Good practice is to 're-heat' by starting a new annealing with $\gamma^0 = \tilde{\gamma}$. This will allow a 'jump' from $\tilde{\gamma}$ in an attempt to avoid being trapped in a local minimum.

5. Application

5·1. *The data*

We apply the methodology to data from an experiment designed to investigate the feasibility of measuring biscuit dough composition on-line using near infrared (NIR) spectroscopy. The experiment involved 39 biscuit doughs in the calibration set and a further 39 available for prediction or validation. The aim is to predict the four major constituents, fat, sucrose, dry flour and water, of the dough using 300 equally spaced NIR reflectance measurements from 1202 to 2400 nm. Spectral data were centred. Compositional response data were both centred and standardised, both training and validation data using the training set mean and standard deviation. The data were first analysed in Osborne et al. (1984).

For our purposes the reflectances represent the $q = 300$ regressor variables and the percentages of the four constituents represent the $r = 4$ responses. In future one wants to predict the true composition from the spectrum or from selected parts of it.

5·2. *Hyperparameters*

We chose the hyperparameter $\delta = 3$, for minimally informative prior knowledge; $k = 0{\cdot}0085^2$ was chosen by taking the generalised crossvalidation ridge estimates for each of the four ingredients and taking the median of these four.

The unpredictable part of Y is made up principally of sampling and measurement errors in the reference values that are to be predicted. With perfect spectral data we might hope to get the prediction errors down to this level. Experience of a wide range of applications suggests that for 'good' calibrations with this type of instrumentation the predictions have a variance about twice this value. Thus, $\Sigma_{00.q} + \Phi$ is roughly twice as large as Φ, which gives a value $w_q = \frac{1}{2}$ for the indeterminism parameter.

One of the motivations for finding small subsets of wavelengths in this application was a desire to implement the measurements on-line using a low-cost instrument that measures the spectrum only at selected wavelengths using a filter for each wavelength. The cost of the instrument will increase with the number of filters required. In reality this cost function will be quite complex, with some standard filters being cheaper than specially made ones and perhaps with large jumps in cost as certain critical numbers of filters are exceeded. This cost needs to be balanced against the value of improved measurement accuracy, as measured by the predictive variance. We have no easy access to the true detailed costs, but we do have some feel for their rough values. We have tried to use this knowledge to

specify a value for c in our linear cost function that gives a realistic exchange rate between variables and variance at the point where this trade-off matters. The four Y's have been scaled to have variance 1. We expect to reduce this to a predictive variance of around 0·1 with a small number of variables. We estimate that it would be worth an additional 2 variables per Y, that is 8 extra variables, to reduce this by 25% for all Y, thus reducing the sum of the 4 variances by $\frac{1}{4} \times 4 \times 0{\cdot}1$. This implies a c of (0·1)/8 or 1/80.

5·3. *Optimisation*

We ran the simulated annealing sampling first with γ^0 all ones. The initial temperature was $T_0 = 300$. This value was found by reversing the annealing: starting from $T_0^{(r)}$ small, we increase the temperature by $T_j = sT_{j-1}$ with $s > 1$ and stop when $\text{AR} \geqslant \beta$ with $0{\cdot}9 < \beta < 1$, where AR is the acceptance ratio. We chose $\beta = 1$ here. The final temperature reached by the reversed annealing will ensure that worse cases are likely to be accepted and can thus be used as starting temperature T_0. The temperature was updated after each step as $T_i = \rho T_{i-1}$ with $\rho = 0{\cdot}999$. Adding and deleting steps were chosen with probabilities $P_A = P_D = \frac{1}{3}$, and swapping steps with probability $1 - P_D - P_A = \frac{1}{3}$. The acceptance ratio, AR, was calculated every 500 iterations, $m = 500$, and the search stopped when $\text{AR} = 0$, that is $\tau = 0$. Our chosen $\text{AR} = 0$ and large m, in combination, enable us to say confidently that the annealing has frozen. Also QR decomposition matrices were recomputed every 500 iterations to avoid the build-up of rounding errors.

The simulated annealing stopped after 13 500 steps, of which 9233 were accepted, involving 2867 additions, 3162 deletions and 3204 swaps, giving a vector $\tilde{\gamma}$ with a minimum cost of 0·1858, and five variables, wavelengths (1626, 1718, 1994, 2066, 2194), selected. Re-heating was performed, starting a new annealing with $\gamma^0 = \tilde{\gamma}$ and temperature $T_0/3 = 100$. The re-heating stopped after a further 12 500 steps, including 2659 additions, 2659 deletions and 2740 swaps, giving the same best model. This gave Bayes prediction root mean squared errors of 0·14 for fat, 0·19 for sucrose, 0·18 for flour and 0·26 for water.

On this response-standardised scale the prediction root mean squares of Osborne et al. (1984) were 0·11 for fat, 0·28 for sucrose, 0·30 for flour and 0·29 for water. We do just about as well on fat and water but improve considerably for sugar and flour. They also used around a dozen regressors with a strategy of amalgamating the regressors selected for each of the four responses in turn. The combined-response loss has evidently induced a more selective and accurate predictor.

The five-wavelength predictor explains (98, 96, 96, 90)% of variation in the four constituents. Figure 1(a) plots the cost function against iteration number, first the original sequence followed by re-heating. The corresponding number of wavelengths selected is given in Fig. 1(b). Note that when the temperature is high, both initially and on first re-heating, there is a good probability of choosing moves that worsen the cost and in fact give more than 39 wavelengths selected, despite there being only 39 data points. Within the Bayesian framework with proper prior distributions no numerical problem arises with such overfitting. One does fairly quickly move to more reasonable and parsimonious models.

We also tried some very different starting values: first 20 wavelengths selected; last 20 selected; random selection with each of the 300 wavelengths selected independently with probability $\frac{20}{300}$. Very similar selections, all of just five wavelengths, resulted with similar mean squared errors. We do not claim to have found the optimum subset; only an exhaustive search will justify such a claim. We have, however, found some very good ones.

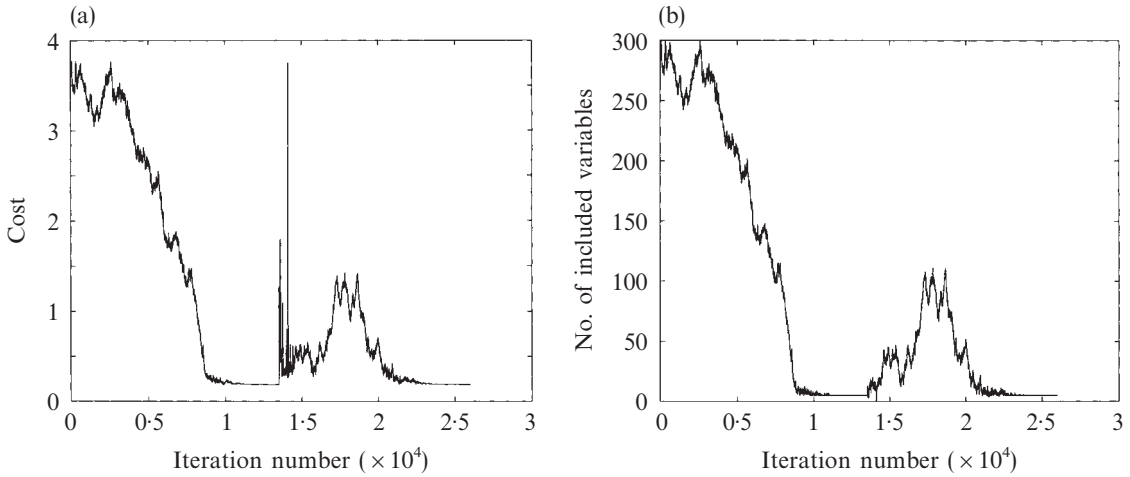

Fig. 1. Simulated annealing with final re-heating. (a) Cost function by iteration, and (b) number of 1's of γ vectors by iteration.

From knowledge of the NIR spectra of the separate ingredients one can say that wavelength 1718 nm is in a region of the spectrum where fat has a characteristic absorbance, and that wavelength 2066 nm is in a characteristic region of sucrose. The other three wavelengths are less easy to assign a priori but evidently provide the discriminatory power to separate water and flour from the other two ingredients.

The parameter w_q was prespecified at $\frac{1}{2}$. We examined the sensitivity of results to this chosen value by trying both $w_q = 1$ and $w_q = \frac{1}{4}$. Both searches stopped at a five variable model, and the best models had total mean squared errors of 0·8460 and 0·7966, respectively, whereas that of our chosen $w_q = \frac{1}{2}$ is 0·7724. Thus there is no great sensitivity to the chosen value of w_q, although it is satisfying that $w_q = 0{\cdot}5$ improves on the natural conjugate case, $w_q = 1$.

5·4. *Discussion*

In applying our methodology to this example we have made a number of rather arbitrary choices of parameters and costs. Some of these could doubtless be challenged as unrealistic. Realistic or not, the result was to identify a much smaller set of variables than did the original investigators, without any loss of predictive performance. In the context of on-line implementation this result has considerable value. Some of the improvement may have come from our more extensive search of possibilities, stepwise methods being used originally, but it seems clear that treating this problem as a multivariate one is highly beneficial.

The simulated annealing program for optimisation was written in MATLAB and we plan to make it available on the Web.

Acknowledgement

This work is supported by the U.K. Engineering and Physical Sciences Research Council with a grant under the Stochastic Modelling in Science and Technology Initiative. We are grateful to the Flour Milling and Baking Research Association for providing the data.

REFERENCES

BERNARDO, J. M. & SMITH, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

BROOKS, R. J. (1974). On the choice of an experiment for prediction in linear regression. *Biometrika* **61**, 303–11.

BROWN, P. J. (1993). *Measurement, Regression, and Calibration*. Oxford: Clarendon.

BROWN, P. J., LE, N. D. & ZIDEK, J. V. (1994). Inference for a covariance matrix. In *Aspects of Uncertainty: A Tribute to D. V. Lindley*, Ed. P. R. Freeman and A. F. M. Smith, pp. 77–92. Chichester: Wiley.

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

BROWN, P. J., VANNUCCI, M. & FEARN, T. (1998a). Bayesian wavelength selection in multi-component analysis. *J. Chemomet.* **12**, 173–82.

BROWN, P. J., VANNUCCI, M. & FEARN, T. (1998b). Multivariate Bayesian variable selection and prediction. *J. R. Statist. Soc.* B **60**, 627–41.

CLYDE, M., DESIMONE, H. & PARMIGIANI, G. (1996). Prediction via orthogonalised model mixing. *J. Am. Statist. Assoc.* **91**, 1197–208.

DAWID, A. P. (1981). Some matrix-variable distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68**, 265–74.

DAWID, A. P. (1988). The infinite regress and its conjugate analysis. In *Bayesian Statistics 3*, Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, pp. 95–110. Oxford: Clarendon.

FANG, B. Q. & DAWID, A. P. (2000). Nonconjugate Bayesian regression on many variables. *J. Statist. Plan. Infer.* To appear.

GEORGE, E. I. & MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7**, 339–73.

GOLUB, G. H., HEATH, M. & WAHBA, G. (1979). Generalised cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215–23.

LEMPERS, F. B. (1971). *Posterior Probabilities of Alternative Models*. Rotterdam: Rotterdam University Press.

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

MÄKELÄINEN, T. & BROWN, P. J. (1988). Coherent priors for ordered regressions. In *Bayesian Statistics* 3, Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, pp. 677–96. Oxford: Clarendon.

MITCHELL, T. J. & BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Am. Statist. Assoc.* **83**, 1023–36.

MOUCHART, M. & SIMAR, L. (1980). Least squares approximation in Bayesian analysis. In *Bayesian Statistics 1*, Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, pp. 207–22. Valencia University Press.

OSBORNE, B. G., FEARN, T., MILLER, A. R. & DOUGLAS, S. (1984). Application of near infrared reflectance spectroscopy to compositional analysis of biscuits and biscuit doughs. *J. Sci. Food Agric.* **35**, 99–105.

PRESS, S. J. (1982). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*, 2nd ed. Malabar, FL: Krieger.

RAFTERY, A. E., MADIGAN, D. & HOETING, J. A. (1997). Bayesian model averaging for linear regression models. *J. Am. Statist. Assoc.* **92**, 179–91.

SEBER, G. A. F. (1984). *Multivariate Observations*. New York: Wiley.

[*Received February* 1998. *Revised November* 1998]