J. R. Statist. Soc. B (1998)
60, *Part* 3, *pp.* 627–641

Multivariate Bayesian variable selection and prediction

P. J. Brown† and M. Vannucci

University of Kent, Canterbury, UK

and T. Fearn

University College London, UK

[Received February 1997. Revised November 1997]

Summary. The multivariate regression model is considered with *p* regressors. A latent vector with *p* binary entries serves to identify one of two types of regression coefficients: those close to 0 and those not. Specializing our general distributional setting to the linear model with Gaussian errors and using natural conjugate prior distributions, we derive the marginal posterior distribution of the binary latent vector. Fast algorithms aid its direct computation, and in high dimensions these are supplemented by a Markov chain Monte Carlo approach to sampling from the known posterior distribution. Problems with hundreds of regressor variables become quite feasible. We give a simple method of assigning the hyperparameters of the prior distribution. The posterior predictive distribution is derived and the approach illustrated on compositional analysis of data involving three sugars with 160 near infra-red absorbances as regressors.

Keywords: Bayesian selection; Conjugate distributions; Latent variables; Markov chain Monte Carlo method; Model averaging; Multivariate regression; Prediction

1. Introduction

There is a large Bayesian literature on model choice and variable selection in the linear multiple-regression model, a skeletal pair of references being Dempster (1973) and Berger and Pericchi (1996). Some approaches focus on utility rather than on probabilistic fit; see Lindley (1968). For a detailed discussion see Bernardo and Smith (1994), and recent applications of loss approaches in Laud and Ibrahim (1995) and Marriott *et al.* (1996). Probabilistic fit in the form of latent mixture modelling has drawn considerable attention, as in George and McCulloch (1993, 1997), Geweke (1996), Wakefield and Bennett (1996), Clyde *et al.* (1996) and Chipman (1996). Our paper extends a part of this development to multivariate regression, initially in a very general distributional setting.

The practical context is situations where it is important to choose a subset (or subsets) of p regressor variables which are good for prediction of all q responses. In the application in Section 9 we wish to estimate the proportions of $q = 3$ components of a mixture of sugars from their near infra-red spectrum of $p = 160$ absorbances. For prediction we average over a set of likely (*a posteriori*) subsets in Section 8. We give algorithms for fast computing of the

†*Address for correspondence*: 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

1369–7412/98/60627

posterior distribution, although our envisaged application (Section 9) involves such a large number of regressors ($p = 160$) that Markov chain Monte Carlo (MCMC) approximations to the posterior distribution are required. Even here though fast forms for adding and deleting variables are beneficial.

To help to signpost our approach we first present the model in a fairly general setting. This setting is a generalization of George and McCulloch (1993, 1997) who concentrated on univariate Gaussian regression.

The q-variate response $Y = (Y_1, \ldots, Y_q)'$ has a distribution depending conditionally on p explanatory variables $x = (x_1, \ldots, x_p)'$. Componentwise for the response, for $l = 1, \ldots, q$, Y_l is assumed to have a mean which is $\eta(\alpha_l + \beta_l' x)$, where $\eta(\cdot)$ is some known continuous function (e.g. exp for multivariate log-linear models). Here β_l is a p-vector of unknown slope parameters and α_l is an unknown scalar parameter. With n independent observations Y_i $(q \times 1)$, conditional on x_i $(p \times 1)$, $i = 1, \ldots, n$, we have a multivariate generalized linear model. The unknown parameters are intercepts α $(q \times 1)$, slope matrix $B = (\beta_1, \ldots, \beta_q)(p \times q)$ and a further set of dispersion parameters, suggestively denoted Σ. The prior distribution of (α, B, Σ) is broadly decomposed as

$$\pi(\alpha, B, \Sigma) = \pi(\alpha, B|\Sigma)\,\pi(\Sigma)$$

and we shall lose little by assuming that α and B are independent conditionally on Σ, so that

$$\pi(\alpha, B, \Sigma) = \pi(\alpha|\Sigma)\,\pi(B|\Sigma)\,\pi(\Sigma). \qquad (1)$$

The further elaboration is of $\pi(B|\Sigma)$ through a latent binary p-vector γ. The jth element of γ, γ_j, may be either 1 or 0. If it is 1 then the covariance matrix of the corresponding row of B is 'large'; if it is 0 then it is relatively small. Since in both cases we shall assume that the prior expectation is 0, the zero value of γ_j confers a prior for the regression slope coefficients which is more concentrated about 0. In the extreme special case where this variance is 0 then $\gamma_j = 0$ effectively deletes the jth explanatory variable from the model. Thus $\pi(B|\Sigma)$ in equation (1) is elaborated to

$$\pi(B, \gamma|\Sigma) = \pi(B|\Sigma, \gamma)\,\pi(\gamma). \qquad (2)$$

The likelihood from the n observations is the product of the n densities of $Y_i|x_i$ and may be written as

$$f(Y|X, \alpha, B, \Sigma),$$

where X $(n \times p)$ is the matrix of p explanatory variables (the *model* matrix) and Y $(n \times q)$ is the matrix of q-responses. The product of this and the prior defined by equations (1) and (2) is the posterior distribution of $(\alpha, B, \Sigma, \gamma)$ up to a constant of proportionality. Thus the marginal posterior distribution of the selection latent vector γ is given by

$$\pi(\gamma|Y, X) \propto \pi(\gamma) \int f(Y|X, \alpha, B, \Sigma)\,\pi(\alpha|\Sigma)\,\pi(B|\Sigma, \gamma)\,\pi(\Sigma)\,\mathrm{d}\alpha\,\mathrm{d}B\,\mathrm{d}\Sigma. \qquad (3)$$

This posterior distribution on γ encapsulates what we need to know about the effectiveness of different explanatory variables in explaining the variation in the q-responses Y. We shall choose a particular Gaussian setting in which a natural conjugate prior distribution allows the explicit calculation of the right-hand side of expression (3) for any γ-vector. The space of γ-vectors is $\{0, 1\}^p$ so the distribution (and normalizing constant) is computable in practice for p less than about 20 since $2^{20} \approx 10^6$. However, for the number of explanatory variables in our application ($p = 160$) the number of possible γ-vectors is prohibitively large and it is necessary to adopt an MCMC approach to approximating the posterior distribution of γ.

The natural conjugate prior distribution also allows easy closed form prediction of a future set of m independent observations at X_f ($m \times p$). This entails model averaging with respect to the marginal posterior distribution of γ given by expression (3) and is developed in Section 8.

The approach is applicable in the above general non-Gaussian setting, although MCMC or other approximation techniques then become necessary at an earlier stage, and considerations of variable dimension spaces may arise as in Carlin and Chib (1995) and Green (1995). In the following sections we specialize to Gaussian multivariate linear regression with natural conjugate prior distributions where such considerations have been by-passed by integrating out B given γ. We first review a general matrix variate notation for Gaussian and related distributions (inverse Wishart distributions) which greatly simplifies calculations, avoiding the need to string matrices as vectors and consequent Kronecker product covariance structures.

2. Matrix variate distributions

We shall follow the notation introduced by Dawid (1981) for matrix variate distributions. This has the advantage of preserving the matrix structures without the need to string by row or column as a vector. It redefines the degrees of freedom as shape parameters for both inverse Wishart and matrix variate T-distributions, to allow notational invariance under marginalization and very easy symbolic Bayesian manipulations.

With U a matrix having independent standard normal entries, $M + \mathcal{N}(\Gamma, \Sigma)$ will stand for a matrix variate normal distribution of $V = M + A'UB$ where M, A and 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 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 $B'(U'U)^{-1}B$, an inverse Wishart distribution. The shape parameter δ differs from the more conventional degrees of freedom and may be generalized, using the density function, to take on any positive real value. The matrix variate T-distribution $M + \mathcal{T}(\delta;\, \Gamma, Q)$ is the distribution of T where T follows the $M + \mathcal{N}(\Gamma, \Sigma)$ distribution conditionally on Σ, and $\Sigma \sim \mathcal{IW}(\delta;\, Q)$. Corresponding probability density functions are given in Brown (1993), appendix A.

3. The model

Conditionally on parameters α, B, γ and Σ the standard multivariate normal regression model assumed is

$$Y - \mathbf{1}\alpha' - XB \sim \mathcal{N}(I_n, \Sigma), \tag{4}$$

with $n \times q$ random matrix Y, $\mathbf{1}$ an $n \times 1$ vector of 1s, $n \times p$ model matrix X regarded as fixed and B the $p \times q$ matrix of regression coefficients. The latent vector γ is buried within the subsequent prior distribution for B. Without loss of generality we assume that columns of X have been centred by subtracting their column means, thus defining the intercept α as the expectation of Y at the data mean of the p x-variables.

The special forms of prior distributions for parameters α, B, γ and Σ in equations (1) and (2) are given as follows. Firstly, given Σ,

$$\alpha' - \alpha_0' \sim \mathcal{N}(h, \Sigma); \tag{5}$$

secondly and independently, given Σ and γ,

$$B - B_0 \sim \mathcal{N}(H_\gamma, \Sigma). \tag{6}$$

Note that from our matrix variate characterization both priors (5) and (6) have covariances that are dependent on Σ in a way that directly extends the univariate regression natural conjugate prior distributions.

Now the marginal distribution of Σ is

$$\Sigma \sim \mathcal{IW}(\delta; Q). \tag{7}$$

The prior for γ in its simplest form is multivariate Bernoulli, i.e. the γ_j are independent with $\text{Prob}(\gamma_j = 1) = w_j$ and $\text{Prob}(\gamma_j = 0) = 1 - w_j$, with hyperparameters w_j to be specified, for $j = 1, \ldots, p$. One elaboration of this prior, perhaps in the symmetric situation with $w_j = w$, would suggest that the random variable W has a beta distribution, with the parameters of this to be specified. This beta–binomial model allows greater *a priori* uncertainty than the first-stage binomial model.

Prior distributions (5)–(7) contain hyperparameters to be specified after structuring. These are discussed in detail in Section 5. One class takes the rows covariance matrix of B as

$$H_\gamma = D_\gamma R_\gamma D_\gamma, \tag{8}$$

following the univariate regression form of George and McCulloch (1993). Here D_γ is a diagonal matrix and R_γ a correlation matrix. The jth diagonal element of D_γ^2 is taken to be v_{0j} when $\gamma_j = 0$ and v_{1j} when $\gamma_j = 1$. Particular forms of v_{0j}, v_{1j} and R_γ are discussed by George and McCulloch (1997).

When R_γ is the identity matrix and the prior matrix of coefficients B_0 is the zero matrix, then the idea of a 'selection' prior may be motivated. Typically $v_{1j} \gg v_{0j}$ and in the selection prior distribution $v_{0j} = 0$. Then $\gamma_j = 0$ indicates that the jth row of B has variance 0 and the distribution is degenerate at the prior zero vector, whereas $\gamma_j = 1$ indicates that the jth row has a non-zero variance determined by v_{1j}. Although the distribution of this vector is still centred on 0 its posterior value will be data dependent.

For this selection prior, the prior distribution of B is such that each column has a singular p_γ-dimensional distribution, i.e. distribution (6) becomes

$$B_{(\gamma)} - B_{0(\gamma)} \sim \mathcal{N}(H_{(\gamma)}, \Sigma). \tag{9}$$

where $B_{(\gamma)}$ selects rows of B that have $\gamma_j = 1$. The complementary rows of B are fixed at their B_0-value with probability 1, in both the prior and the posterior distribution. Here the prior mean B_0 of B will typically be taken to be the $p \times q$ matrix of 0s.

The scale hyperparameter h of the prior distribution for α given in expression (5) will be taken to be a large value, tending to ∞, when the value ascribed to the prior mean α_0 becomes irrelevant.

The scale matrix hyperparameter Q of the prior distribution for Σ from equation (7) is given the simple form kI_q, perhaps after scaling of the response variables. Weak prior information requires a small value of δ and we usually take this to be $\delta = 3$, when $E(\Sigma) = Q/(\delta - 2) = Q$. The value $\delta = 3$ is just a convenient small value, the smallest integer value such that the expectation of Σ exists. Some sensitivity analysis is desirable in any real application.

4. Posterior distributions

The probability density function of Y from model (4) is

$$f_Y(Y|\alpha, B, \Sigma) = c(n, q)|\Sigma|^{-n/2} \exp\{-\tfrac{1}{2}\operatorname{tr}(Y - \mathbf{1}\alpha' - XB)\Sigma^{-1}(Y - \mathbf{1}\alpha' - XB)'\}. \tag{10}$$

Explicit forms for the constant $c(n, q)$, and other matrix variate densities needed below, can be found in Brown (1993), appendix A. Here *post hoc*, with equation (10) as a likelihood function, it is assumed that columns of Y in addition to those of X have been centred, and hence

$$\bar{Y}_l = 0, \qquad l = 1, \ldots, q,$$

$$\bar{x}_j = 0, \qquad j = 1, \ldots, p.$$

The prior probability density function of (α, B) given Σ, γ, is the product of

$$\pi(\alpha|\Sigma) \propto h^{-q/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2h}(\alpha - \alpha_0)'\Sigma^{-1}(\alpha - \alpha_0)\right\} \tag{11}$$

and

$$\pi(B|\Sigma, \gamma) \propto |H_\gamma|^{-q/2}|\Sigma|^{-p/2} \exp[-\tfrac{1}{2}\operatorname{tr}\{H_\gamma^{-1}(B - B_0)\Sigma^{-1}(B - B_0)'\}]. \tag{12}$$

For the selection prior, $H_\gamma \rightarrow H_{(\gamma)}$, and $B \rightarrow B_{(\gamma)}$ a $p_\gamma \times q$ matrix, and the density function columnwise of B is confined to a p_γ-dimensional hyperplane. Since $p \rightarrow p_\gamma$, the missing constant $(2\pi)^{-pq/2}$ of expression (12) reduces to $(2\pi)^{-p_\gamma q/2}$ of expression (9).

We first seek to integrate over (α, B) for given (Σ, γ). In this Gaussian setting, to do this we should 'complete the square' in α and B within the exponentiated terms of the likelihood times prior. First focusing on the likelihood given by equation (10), the exponential term is

$$-\tfrac{1}{2}\operatorname{tr}[\Sigma^{-1}\{(Y - XB)'(Y - XB) + n\alpha\alpha'\}], \tag{13}$$

using $\operatorname{tr}(AC) = \operatorname{tr}(CA)$ and the centring of both Y and X.

The α-term of expression (13) combines with the appropriate part of the exponential term in expression (11) to give

$$(h^{-1} + n)(\alpha - \bar{\alpha})(\alpha - \bar{\alpha})' + h^{-1}\alpha_0\alpha_0' - (h^{-1} + n)\bar{\alpha}\bar{\alpha}' \tag{14}$$

where $\bar{\alpha} = h^{-1}\alpha_0/(n + h^{-1})$. The exponential of the first term of expression (14) (with the factor $\frac{1}{2}\operatorname{tr}(\Sigma^{-1})$) together with the $|\Sigma|^{-1/2}$-term of expression (11) form the kernel of a Gaussian probability density and may be directly integrated out. The second and third remaining terms of expression (14) tend to 0 as h becomes large and may be ignored in our weak prior specification.

Turning to the integration of B for given Σ, γ, the first term of expression (13) plus the corresponding part of the exponential of expression (12) is

$$(Y - XB)'(Y - XB) + (B - B_0)'H_\gamma^{-1}(B - B_0).$$

Completing the square in B, this becomes

$$(B - K_\gamma^{-1}M)'K_\gamma(B - K_\gamma^{-1}M) - M'K_\gamma^{-1}M + C, \tag{15}$$

with

$$M = X'Y + H_\gamma^{-1} B_0, \tag{16}$$

$$C = Y'Y + B_0' H_\gamma^{-1} B_0, \tag{17}$$

$$K_\gamma = X'X + H_\gamma^{-1}. \tag{18}$$

The first term of expression (15) is the completed quadratic form in B. Multiplying this by $-\frac{1}{2}\mathrm{tr}(\Sigma^{-1})$, and taking its exponential, and collecting the necessary powers $-p/2$ and $q/2$ of the determinants of Σ and K_γ respectively, this forms a Gaussian probability density and may be integrated out. This leaves the likelihood marginalized over α, B, for given Σ, γ, as proportional to

$$(|H_\gamma||K_\gamma|)^{-q/2} |\Sigma|^{-n/2} \exp[-\tfrac{1}{2} \mathrm{tr}\{\Sigma^{-1}(C - M'K_\gamma^{-1} M)\}]. \tag{19}$$

In the selection prior, $H_\gamma \rightarrow H_{(\gamma)}$ is $p_\gamma \times p_\gamma$ and the changed missing constant of proportionality in expression (9) is reabsorbed in the posterior integration of $B_{(\gamma)}$, and expression (19) is unaffected except for $H_\gamma \rightarrow H_{(\gamma)}$ and $X \rightarrow X_{(\gamma)}$ in equations (16)–(18).

The probability density function of the inverse Wishart prior for Σ given by distribution (7) is of the same Σ-form as expression (19) and hence their product may be integrated out over Σ for given γ. This gives the posterior distribution of γ as

$$\pi(\gamma | Y, X) \propto g(\gamma) = (|H_\gamma||K_\gamma|)^{-q/2} |Q_\gamma|^{-(n+\delta+q-1)/2} \pi(\gamma), \tag{20}$$

where

$$\begin{aligned} Q_\gamma &= Q + C - M'K_\gamma^{-1} M \\ &= Q + Y'Y - Y'XK_\gamma^{-1} X'Y, \end{aligned} \tag{21}$$

when $B_0 = 0$. Here K_γ is given by equation (18).

Computation of the posterior distribution of γ follows directly from equation (20) once all the hyperparameters within H_γ, Q and δ have been specified.

5. Prior settings

We have already discussed giving weak prior information about the $q \times q$ covariance matrix Σ. We have generally set $\delta = 3$ and $Q = kI_q$ with k *a priori* comparable in size with the likely error variances of Y given X. Because of the small δ, having the same k for each of the q-responses is probably not critical, but some rescaling and sensitivity analysis may be advisable.

The intercept parameter vector α has been given vague prior information. This we have seen leads to our being able to ignore this parameter, provided that both X and Y have been centred. The posterior uncertainty in α does, however, enter predictions; see Section 8.

The main thrust of our modelling is through the prior for B, γ, as structured through equation (2) by distribution (6) and the multivariate Bernoulli $\pi(\gamma)$ distribution. The prior distribution for B given γ then depends on H_γ. One class that has already been discussed in Section 3 is given by equation (8). Within this we usually take $R_\gamma = I$. The specification of (v_{0j}, v_{1j}) suggested by George and McCulloch (1997) relies on a threshold of 'practical significance' for each variable marginally, $j = 1, \ldots, p$. We have been more interested in the particular case $v_{0j} = 0, j = 1, \ldots, p$, and we shall make further use of this case. It is evident that some care is needed in specifying v_{1j} relative to the lump of probability put on $\gamma_j = 0$; see for example Garthwaite and Dickey (1992) for one thoughtful approach.

An alternative automatic prior when $\upsilon_{0j} = 0$ is to take

$$H_{(\gamma)} = c(X'_{(\gamma)}X_{(\gamma)})^{-1}, \tag{22}$$

where implicitly the subset $X_{(\gamma)}$ of columns of X chosen to correspond to $\gamma_j = 1$ is of full column rank. Smith and Kohn (1996) recommended a large c in the range 10–100 and remarked that this is akin to the g-prior of Zellner (1986), and is as if the prior distribution resulted from a prior experiment with the same model matrix X. See also Raftery *et al.* (1997) for a similar form of prior. The prior is investigated in Brown *et al.* (1998).

6. Fast forms for updating

All the prior assignments of the previous section can be formulated as least squares problems, with possibly modified Y- and X-matrices. Thinking of the $\{0, 1\}^p$ space of γ as a hypercube, Gray codes may be used to visit all the vertices just once, tracing a path that involves a change in just one component of γ per step. These order the 2^p binary tuples so that adjacent tuples differ in just one of the p places; for example with $p = 3$ a possible sequence is $\{000, 001, 011, 010, 110, 100, 101, 111\}$. For algorithms to generate Gray codes see Diaconis and Holmes (1994). If γ is changed one component at a time, then fast QR deletion or addition algorithms are directly applicable. This extends the cases considered by George and McCulloch (1997) even for univariate multiple regression. We shall also show how the posterior for γ given by equation (20) can be rearranged to encompass the singular H_γ case.

Treating a prior distribution as pseudodata has a long history; see for example Marquardt (1970), where ridge regression is shown to be equivalent to a least squares problem with the design matrix augmented by $I\sqrt{k}$ and the Y-vector by p 0s. George and McCulloch (1997) used the same idea when $\upsilon_{0j} > 0$ within equation (8).

The following development allows the posterior distribution for γ, given by equation (20), to apply to both $\upsilon_{0j} > 0$ as well as to the selection prior, $\upsilon_{0j} = 0$. The relevant quantities entering equation (20) are

$$\begin{aligned} |H_\gamma||K_\gamma| &= |H_\gamma||X'X + H_\gamma^{-1}| \\ &= |H_\gamma^{1/2}X'XH_\gamma^{1/2} + I| \\ &= |\tilde{X}'\tilde{X}| \end{aligned} \tag{23}$$

where

$$\tilde{X} = \begin{pmatrix} XH_\gamma^{1/2} \\ I_p \end{pmatrix}, \tag{24}$$

an $(n+p) \times p$ matrix. Also with

$$\tilde{Y} = \begin{pmatrix} Y \\ 0 \end{pmatrix}, \tag{25}$$

an $(n+p) \times q$ matrix, then Q_γ from equation (21) becomes Q plus

$$\tilde{Y}'\tilde{Y} - \tilde{Y}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} \tag{26}$$

and this is the residual sum of products matrix from the least squares regression of $\tilde{Y}$ on $\tilde{X}$.

The computational task is further simplified by reducing $\tilde{X}$ to the $(n+p) \times p_\gamma$ matrix $\tilde{X}_{(\gamma)}$, formed by selecting the $\gamma_j = 1$ columns of $\tilde{X}$. The $p - p_\gamma$ rows of 0s could also be removed from $\tilde{X}_{(\gamma)}$ and $\tilde{Y}$ but we have found it computationally convenient to retain the same number of rows throughout and just to add or delete columns. The QR-decomposition of $(\tilde{X}_{(\gamma)}, \tilde{Y})$ is given for example by Seber (1984), chapter 10, section 1.1b, and avoids 'squaring' as in expressions (23) and (26). Updating `qrdelete` and `qrinsert` algorithms are then available within many computing environments, removing or adding a column to the reduced $(n+p) \times p_\gamma$ matrix.

The data augmentation simplification with the g-prior (22) leads to even greater computational savings. The relevant part of Q_γ is

$$Y'Y - Y'X_{(\gamma)}\{X'_{(\gamma)}X_{(\gamma)} + (1/c)X'_{(\gamma)}X_{(\gamma)}\}^{-1}X'_{(\gamma)}Y = \{c/(c+1)\}\{Y'Y - Y'X_{(\gamma)}(X'_{(\gamma)}X_{(\gamma)})^{-1}X'_{(\gamma)}Y\} + Y'Y/(c+1)$$

and all the required quantities in equation (20) are obtained from simply regressing Y on $X_{(\gamma)}$. Again QR-algorithms for fast updating may be used. Also $|H_{(\gamma)}||K_{(\gamma)}|$ in equation (20) simplifies to $(c+1)^{p_\gamma}$.

7. Markov chain Monte Carlo method

The posterior for γ is directly computable through equation (20). However, its right-hand side 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. The use of Gray code sequences will substantially speed up computations but will still only allow up to around $p = 25$ variables. Our applications have generally involved much larger numbers of variables; see Section 9. 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.

The simplest Gibbs sampler is obtained by generating each γ-value componentwise from the full conditional distributions,

$$\gamma_j | \gamma_{/j}, Y, X \qquad j = 1, \ldots, p,$$

where $/j = \{1, 2, \ldots, j-1, j+1, \ldots, p\}$ and we may choose any fixed or random order. The conditional probability that $\gamma_j = 1$ is $\theta_j/(\theta_j + 1)$ where

$$\theta_j = g(\gamma_j = 1, \gamma_{/j}|Y, X)/g(\gamma_j = 0, \gamma_{/j}|Y, X) \qquad (27)$$

which does not involve the proportionality constant of equation (20). The random mechanism thus requires Bernoulli random variables. As noted by George and McCulloch (1997), for each component of the iterative sequence, one of the values $g(\gamma)$ in equation (27) will be available from the previous component simulation. The other value of $g(\gamma)$ can then be found by using a fast updating QR-algorithm from Section 6. The sequence of γ-values after each completed cycle may be stored to give exact relative probabilities of the visited γ. The missing normalizing constant comes from summing over all 2^p γ-vectors. George and McCulloch (1997) show how to estimate this constant of proportionality consistently by using a pilot run to set up a target set of γ-vectors. More general Metropolis–Hastings algorithms are easy to set up and may provide faster mixing. For suggestions the reader is referred to George and

McCulloch (1997). Brown *et al.* (1998) used a Metropolis algorithm with deletion, addition and swapping moves.

In our applications we have not been concerned about strict convergence of the MCMC sampler. We have chosen five separate long runs from very different starting points and are satisfied if they show broadly similar marginal distributions, with a good indication of mixing and explorations with returns. Some illustration is given when discussing our application in Section 9. Our approach is less formal but closest perhaps to that of Gelman and Rubin (1992).

8. Prediction

There are at least two ways of proceeding when asked to predict m future Y-vectors at m given x-vectors. Let us denote the m Y-vectors as the $m \times q$ matrix Z and the given x-vectors as X_{f} $(m \times p)$. The model for Z, following the model for the training data (4), is

$$Z - \mathbf{1}\alpha' - X_{\mathrm{f}}B \sim \mathcal{N}(I_m, \Sigma)$$

where to conform with our definitions for model (4) Z and X_{f} have been adjusted by subtraction of the column means of Y and X from the training data. The first way of proceeding would be to regard the latent structure model as merely identifying good subsets. We can then use a chosen subset to provide least squares predictions. This seems implicitly to be the approach of George and McCulloch (1997). It is perhaps more satisfactory to apply the same latent structure model to prediction as well as to training. This has the practical appeal of providing averaging over a range of likely models.

Since the posterior distribution of (α, B, Σ) given γ has exactly the same normal inverse Wishart form as the prior given in Section 3, we shall firstly develop the predictive distribution for Z from the prior distribution. The predictive distribution given Y, X, will then follow by updating the hyperparameters.

Firstly arguing conditionally on Σ, and using expressions (11) and (12), we can see that

$$Z - \mathbf{1}\alpha_0' - X_{\mathrm{f}}B_0 \sim \mathcal{N}(I_m + h\mathbf{1}\mathbf{1}' + X_{\mathrm{f}}H_\gamma X_{\mathrm{f}}', \Sigma).$$

Then averaging over the prior distribution of Σ in distribution (7) gives, conditionally only on γ,

$$Z - \mathbf{1}\alpha_0' - X_{\mathrm{f}}B_0 \sim \mathcal{T}(\delta; I_m + h\mathbf{1}\mathbf{1}' + X_{\mathrm{f}}H_\gamma X_{\mathrm{f}}', Q). \tag{28}$$

The power of the Dawid (1981) notation is brought out by the symbolic simplicity of the Bayesian manipulations: for more details see Brown (1993), appendix A. The predictive distribution posterior to the training data is of the same form as expression (28) except that δ and H_γ are replaced by $\delta + n$ and K_γ^{-1} respectively, and Q is replaced by Q_γ as given in equation (21). When $h \to \infty$ and $B_0 = 0$, h is replaced by $1/n$, and α_0 and B_0 are replaced by $\hat{\alpha}$ and $\hat{B}_\gamma$ where

$$\begin{aligned} \hat{\alpha} &= 0, \\ \hat{B}_\gamma &= (I - W)\hat{B}. \end{aligned} \tag{29}$$

Here $W = K_\gamma^{-1}H_\gamma^{-1}$ and $\hat{B} = (X'X)^{-1}X'Y$, and equation (29) is a direct consequence of the centring of X and Y and the weak prior knowledge.

The formula for the mean of Z, $X_{\mathrm{f}}\hat{B}_\gamma$, can be rearranged to cover also the case where H_γ is singular, as is the case with the selection prior, $v_{0j} = 0$. We have

$$\begin{aligned}\hat{B}_\gamma &= (X'X + H_\gamma^{-1})^{-1} X'Y \\ &= H_\gamma^{1/2}(H_\gamma^{1/2\prime} X'XH_\gamma^{1/2} + I)^{-1} H_\gamma^{1/2\prime} X'Y \\ &= H_\gamma^{1/2}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}\end{aligned}$$

with $\tilde{X}$ and $\tilde{Y}$ given by equations (24) and (25). The leading $H_\gamma^{1/2}$ may be absorbed into X_f to give the prediction $\tilde{X}_\mathrm{f}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}$ with $\tilde{X}_\mathrm{f} = X_\mathrm{f} H_\gamma^{1/2}$. For the selection prior distribution, this may be dimensionally reduced further to

$$\tilde{X}_{(\gamma)\mathrm{f}}(\tilde{X}'_{(\gamma)}\tilde{X}_{(\gamma)})^{-1}\tilde{X}'_{(\gamma)}\tilde{Y} \tag{30}$$

with the $m \times p_\gamma$ and $(n+p) \times p_\gamma$ matrices $\tilde{X}_{(\gamma)\mathrm{f}}$ and $\tilde{X}_{(\gamma)}$ respectively formed by selecting the columns corresponding to $\gamma_j = 1$.

Finally to predict Z under quadratic loss we need the expectation of the predictive distribution given by expression (28) unconditionally, i.e. averaging over the posterior distribution of γ, given by equation (20). This gives

$$\hat{Z} = \sum_\gamma X_\mathrm{f}\hat{B}_\gamma\, \pi(\gamma|X, Y), \tag{31}$$

where expression (30) may replace $X_\mathrm{f}\hat{B}_\gamma$ for the selection prior distribution. We might choose to approximate this by some restricted set of γ-values, perhaps the r most likely values from the MCMC simulation, or perhaps in the spirit of Occam's window (Madigan and Raftery, 1994).

9. Application

We illustrate the methodology on near infra-red data of three sugars, sucrose, glucose and fructose, present in varying concentrations in aqueous solution, originally analysed in Brown (1992, 1993). For our purposes the concentrations of the three sugars represent the $q = 3$ responses. For each sample the absorbances were recorded at 700 wavelengths, from 1100 nm to 2498 nm in steps of 2 nm. There were 125 training samples and 21 further samples reserved for later prediction. These 21 samples were especially designed to be difficult to predict, with compositions outside the range of the other 125: see Brown (1993), chapter 1, for further details. It would not therefore be natural to amalgamate all 146 samples and to leave out other subsets for validation.

For illustration, and to reduce the computation, we chose 160 from the 700 wavelengths, equally spaced from 1100 nm to 2500 nm by linear interpolation. Any larger number could have been chosen with correspondingly increased computational time, but intervals of around 9 nm retain adequate resolution in the near infra-red. Thus the number of explanatory variables is $p = 160$.

In keeping with the methodology developed we selected the prior distribution with $v_{0j} = 0$. We chose $H_{(\gamma)} = c[\mathrm{diag}(X'X)^-]_{(\gamma)}$, where the minus superscript denotes a generalized inverse, allowing fast updating because of its diagonal structure. The value of c was 0.8, intentionally around 1, and comes from its full model relationship to a g-prior, the full model prior standard deviation being about the same order of magnitude as the full data least squares standard error. In other applications it may be necessary to experiment with other values of c.

Because of its diagonal structure, $\surd v_{1j}$ may also be considered as a scaling multiplier of the jth explanatory variable, a variant on the scalar divisor of $\surd\{\mathrm{diag}(X'X)/n\}$ that is more conventionally used for *autoscaling*.

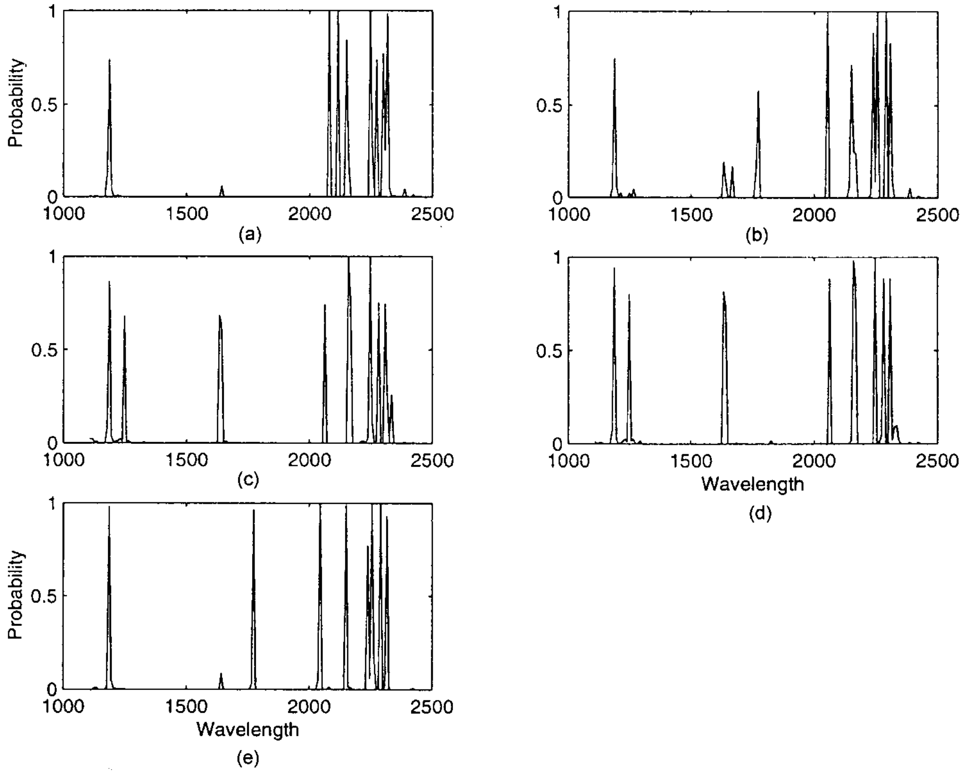

Fig. 1. Marginal probabilities of components of γ for the five runs (a)–(e)

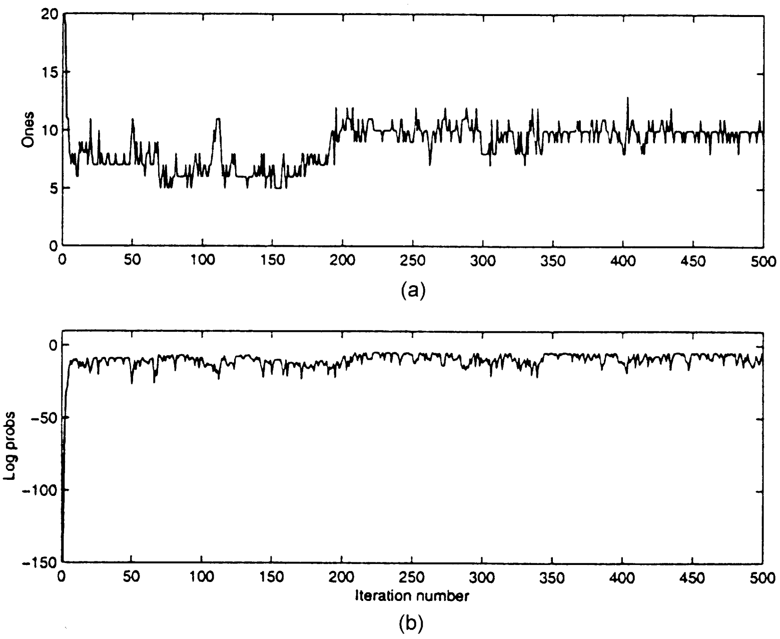

Fig. 2. Plots in sequence order for run (c): (a) the number of 1s; (b) $\log\{g(\gamma)\}$

In line with the number of explanatory variables that are needed in similar applications, and to induce a 'small' model, we chose the binomial prior for the exchangeable γs to have an expectation of 20. Other hyperparameters were specified to give weak prior knowledge; specifically $\delta = 3$ and $k = 0.2$, where k was also commensurate with the sort of accuracy expected and hoped for. We chose five widely different starting points for the five Gibbs sampling runs, all except the last randomly permuted the wavelengths first, and then chose $\gamma_j = 1$s

(a) all 1s,
(b) half 1s,
(c) 20 1s,
(d) 20 1s and
(e) the first 20 1s.

It was thought to be important to break up the positive correlation structure of the spectrum to aid Gibbs convergence, and randomization of the wavelengths would help. There were 500 iterations in each run, where an iteration consisted of a Bernoulli random draw for each of the 160 components γ_j in the predefined order (random in all except run (e)). This was done using the QR-decomposition and `qrdelete` and `qrinsert` of MATLAB (MathWorks, 1996). Every 10th run we recalculated using the QR-decomposition to check on the possible build-up of rounding error in using the updating algorithms. Rounding error turned out not to be a problem. To users of MCMC algorithms 500 iterations may seem very modest. However, we were not concerned with strict convergence and we were in the unusual position of knowing the exact posterior probabilities up to a constant of proportionality for the sampled γs. Thus no 'burn-in' was necessary, since relatively unlikely γs would automatically be downweighted in our analysis. The ultimate justification was in the excellent predictions that were obtained from the most probable γ generated.

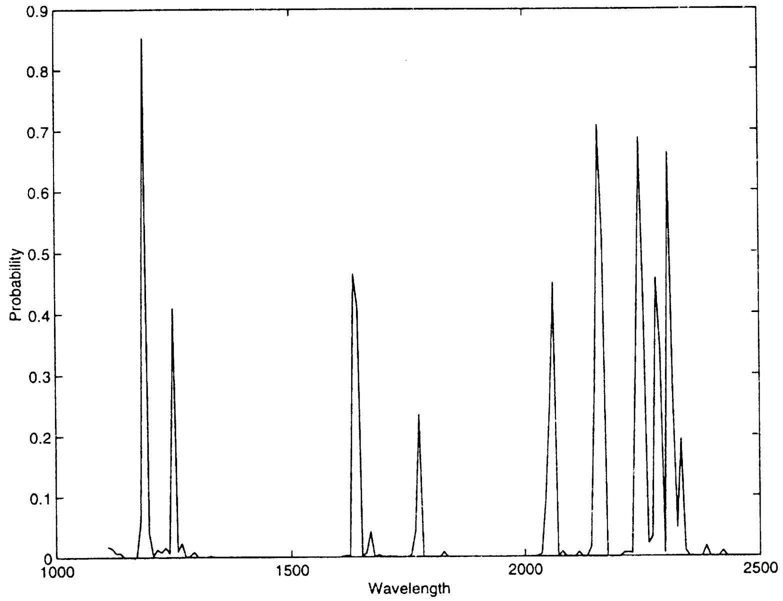

Fig. 3. Renormalized marginal probabilities of components of γ

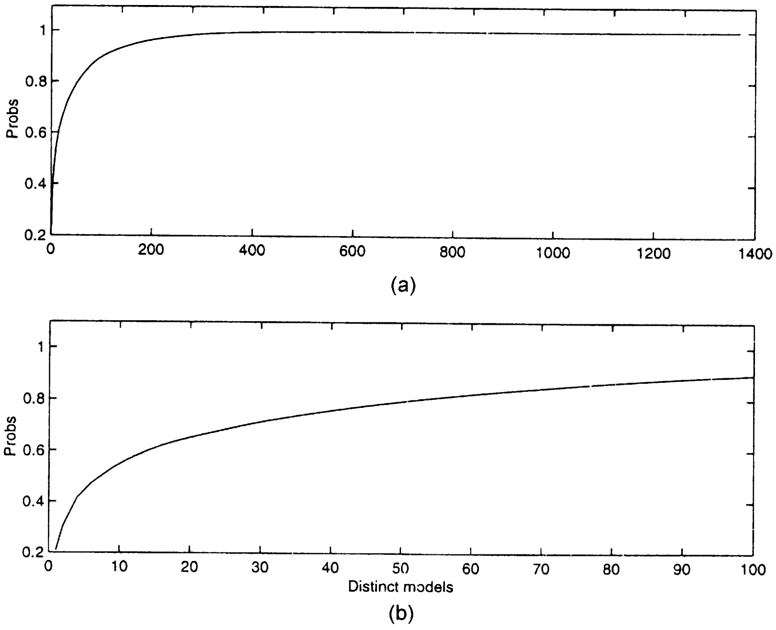

Fig. 4. Cumulative ordered relative probabilities of distinct γs: (a) all visited γs; (b) first 100 γs

For each run of 500 iterations we recorded the γs, and their corresponding $g(\gamma)$ relative probability. The 500 vectors were reduced to the set of distinct γs. Each run showed around half the 500 vectors as distinct, and a few vectors were repeated many times. The relative probabilities of the set of distinct γs were then normalized to 1 over the visited γ-vectors. The marginal probabilities for components of γ, $P(\gamma_j = 1)$, $j = 1, \ldots, 160$, are plotted in Fig. 1, where j ranges over the range 1100–2498 nm in 160 equally spaced steps. The spikes are where regressor variables have been included with high probability. Despite some differences the plots are broadly similar and, although we would not claim convergence, the localities explored are not too disparate, even with the widely different starting values.

For run (c), two further revealing plots are given. Firstly, the number of 1s is plotted over the 500 iterations in Fig. 2(a), and looks to have settled down to around 10 after starting at 20. Secondly, the log-relative-probabilities $\log\{g(\gamma)\}$ of visited γs are plotted in their order of occurrence in the iterative sequence in Fig. 2(b). The $g(\gamma)$ quickly increase to the level at which they settle down. We also looked at the number of component switches (out of 160) from iteration to iteration. This was initially large but soon settled down to around two or three per iteration.

It is also interesting to look at more conventional measures of fit. Suppose that we take as $\gamma_j = 1$ all those that have a marginal (normalized) probability of 1 of at least 0.1, say. This gives a selection of variables. Least squares prediction using this subset, and internal to the training data, gave percentages of variation explained of 99.7%, 99.5% and 99.5% for the three sugar concentrations for starting set (c), and 10 variables selected: 1183, 1244, 1625, 1634, 2054, 2150, 2159, 2238, 2273 and 2299 nm. For the other four starting values the subsets differed a little but similar explanations of variance were achieved.

Fig. 3 corresponds to the union of γs from the five runs, normalizing the relative probabilities of the distinct γs, and displayed as marginal probabilities, $P(\gamma_j = 1), j = 1, \ldots, 160$, equally spaced from 1100 to 2498 nm.

Fig. 4 gives the cumulative probability distribution of visited γs, against number of γs, where the γs are ordered according to probability. This enables us to choose a cut-off for the γs to be averaged for prediction of the concentrations for the validation data. This uses equation (31) but with an approximation to the posterior of γ obtained from the most likely γ-vectors of the MCMC simulation, renormalized to sum to 1. The 10 most likely γs account for 55% of the visited probability and use 23 of the wavelengths. It gives mean-squared errors of 0.116, 0.361 and 0.351 for the prediction of sucrose, glucose and fructose respectively. This implies that more than 99.8% of the variation for each of the three sugars in the 21 prediction samples is explained. This very commendable and uniform accuracy is for a particularly difficult set of prediction samples designed to be largely outside the range of the training data. The most likely γ, with 10 wavelengths selected, accounts for 20% of the visited probability and gives slightly worse prediction mean squares of 0.210, 0.446 and 0.510, so that model averaging is seen to be beneficial.

MATLAB software used in this paper may be obtained on the World Wide Web at

`http://stork.ukc.ac.uk/IMS/statistics/people/M.Vannucci.html`

Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council with a grant under the Stochastic Modelling in Science and Technology Initiative, grant GK/K73343. We are grateful to Shell UK for providing the data.

References

Berger, J. O. and Pericchi, L.-R. (1996) The intrinsic Bayes factor for linear models. In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 25–44. Oxford: Clarendon.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. Chichester: Wiley.

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., Vannucci, M. and Fearn, T. (1998) Bayesian wavelength selection in multicomponent analysis. *J. Chemometr.*, **12**, in the press.

Carlin, B. P. and Chib, S. (1995) Bayesian model choice via Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **57**, 473–484.

Chipman, H. (1996) Bayesian variable selection with related predictors. *Can. J. Statist.*, **24**, 17–36.

Clyde, M., Desimone, H. and Parmigiani, G. (1996) Prediction via orthogonalised model mixing. *J. Am. Statist. Ass.*, **91**, 1197–1208.

Dawid, A. P. (1981) Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, **68**, 265–274.

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.

Diaconis, P. and Holmes, S. (1994) Gray codes for randomisation procedures. *Statist. Comput.*, **4**, 287–302.

Garthwaite, P. H. and Dickey, J. M. (1992) Elicitation of prior distributions for variable-selection problems in regression. *Ann. Statist.*, **20**, 1697–1719.

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. (1993) Variable selection via Gibbs sampling. *J. Am. Statist. Ass.*, **88**, 881–889.

——(1997) Approaches for Bayesian variable selection. *Statist. Sin.*, **7**, 339–373.

Geweke, J. (1996) Variable selection and model comparison in regression. In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 609–620. Oxford: Clarendon.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Laud, P. W. and Ibrahim, J. G. (1995) Predictive model selection. *J. R. Statist. Soc.* B, **57**, 247–262.

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.

Marquardt, D. W. (1970) Generalised inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics*, **12**, 591–612.

Marriott, J. M., Spencer, N. M. and Pettitt, A. N. (1996) A Bayesian approach to selecting covariates for prediction. *Technical Report*. Nottingham Trent University, Nottingham.

MathWorks (1996) *MATLAB Version 5.0.0 .4064 on SOLZ*. Matick: MathWorks.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997) Bayesian model averaging for linear regression models. *J. Am. Statist. Ass.*, **92**, 179–191.

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

Smith, M. and Kohn, R. (1996) Nonparametric regression using Bayesian variable selection. *J. Econometr.*, **75**, 317–343.

Wakefield, J. and Bennett, J. (1996) The Bayesian modelling of covariates for population pharmacokinetic models. *J. Am. Statist. Ass.*, **91**, 917–927.

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.