

BAYESIAN WAVELENGTH SELECTION IN MULTICOMPONENT ANALYSIS

P. J. BROWN,^{1*} M. VANNUCCI¹ AND T. FEARN²

¹*Institute of Mathematics and Statistics, University of Kent at Canterbury, Cornwallis Building, Canterbury CT2 7NF, U.K.*

²*Department of Statistics, University College London, Gower Street, London WC1E 6BT, U.K.*

SUMMARY

Multicomponent analysis attempts to simultaneously predict the ingredients of a mixture. If near-infrared spectroscopy provides the predictor variables, then modern scanning instruments may offer absorbances at a very large number of wavelengths. Although it is perfectly possible to use whole spectrum methods (e.g. PLS, ridge and principal component regression), for a number of reasons it is often desirable to select a small number of wavelengths from which to construct the prediction equation relating absorbances to composition. This paper considers wavelength selection with a view to using the chosen wavelengths to simultaneously predict the compositional ingredients and is therefore an example of multivariate variable selection. It adopts a binary exclusion/inclusion latent variable formulation of selection and uses a Bayesian approach. Problems of search of the vast number of possible selected models are overcome by a Markov chain Monte Carlo sampling technique. © 1998 John Wiley & Sons, Ltd.

KEY WORDS multivariate regression; Bayesian wavelength selection; Markov chain Monte Carlo (MCMC); Metropolis algorithm; NIR spectroscopy; multicomponent analysis; selection bias; model averaging

INTRODUCTION

In early work, near-infrared (NIR) instruments typically recorded absorbances at a prespecified small number of wavelengths, typically six. Selection was in effect built into the instrument by the manufacturer by the range of filters. The advent of modern scanning instruments has meant instantaneous capability of providing absorbances across the NIR range at perhaps 1000 equally spaced wavelengths. Early approaches to calibrating with these spectra concentrated on simple univariate wavelength selection strategies (e.g. stepwise selection), but ran into problems with computation time and a lack of appreciation of the bias due to extensive searching,^{1,2} resulting in over-optimistic fits. More recently there has been a shift to the use of 'full spectrum' techniques such as principal components and partial least squares (PLS) regression, and yet more recently to the use of neural network approaches.

One early attempt to address the issue of bias in extensive selection search was presented in Reference 3. It proposed a rather simple and rapid selection algorithm based on univariate correlation, but with a complexity stopping rule taking account of the search extent. Application of this technique to the calibration of sugars in research funded by Shell U.K. Ltd. demonstrated massive

* Correspondence to: P. J. Brown, Institute of Mathematics and Statistics, University of Kent at Canterbury, Cornwallis Building, Canterbury CT2 7NF, U.K. E-mail: philip.j.brown@ukc.ac.uk
Contract/grant sponsor: U.K. Engineering and Physical Sciences Research Council;
Contract/grant number: GK/K73343

generalizability improvement for 18 wavelengths selected from 700 available.⁴ The improvement stemmed from the avoidance of regions of the spectrum where interactions and non-linearity were strongly present, destroying the straightforward application of Beer's law.

There seems to be currently a reawakening of interest in selection approaches.⁵⁻⁷ At least in part this is driven by a suspicion that the more complex calibrations produced by techniques such as PLS are proving to be less robust in long-term use. There may be other, more practical reasons for wishing to select a small subset of wavelengths. If such a subset can be found, then there is a good chance that the calibration can be implemented on an instrument with a small number of filters, which would be much less costly than the scanning instruments typically used in researching the application. This is important if the aim is to use the measurement on-line in a production process, using one or more dedicated instruments. Increasingly, this is the aim of much NIR research.

Most of the approaches to selection tried so far focus on a single ingredient and look at some criterion such as prediction residual sum of squares (PRESS). Optimization is either by means of a sequential search such as stepwise regression or uses a probabilistic search so as to not get trapped in a local minimum. Kalivas *et al.*⁶ used simulated annealing, whilst Leardi *et al.*⁸ and Kubinyi⁷ used genetic algorithms. Lucasius *et al.*⁹ compared genetic algorithms with simulated annealing.

Our approach is rather different. Firstly, it is truly multivariate in that it seeks subsets that work well for all ingredients simultaneously. If the need for a subset is motivated by the desire to use a low-cost filter instrument, then we need just such a subset. If we adopt the naive approach of finding optimal subsets for each ingredient separately and then pooling them, we may miss out on much smaller combined sets that perform nearly as well on all the ingredients. Secondly, our approach does not necessarily seek a single optimum selected model. Rather, we seek to summarize the distribution of 2^p possible models, where p is the number of wavelengths and may be large. We look for selected models that fit well relative to a number of assumptions, the most significant of which is a preference for small numbers of wavelengths quantified as a probability distribution over the number of wavelengths involved. The approach is fairly automatic and uses a simulation technique, moving from one selected model to another in a Markov chain (MCMC), in the instance of this paper using the Metropolis algorithm.¹⁰

BAYESIAN LATENT VARIABLE MODEL

Suppose there are q component ingredients of interest. In the later application there are $q = 3$ sugars, sucrose, glucose and fructose, in aqueous solution. The j th of these ingredients is assumed linearly related to the full p absorbances. With n observations of composition and corresponding spectra we have the linear model

$$\mathbf{Y}_j = \mathbf{X}\boldsymbol{\beta}_j + \epsilon_j, \quad j = 1, \dots, q \quad (1)$$

where \mathbf{Y}_j and ϵ_j are random n -vectors of composition and error for the j th component, \mathbf{X} ($n \times p$) is the matrix of absorbances and $\boldsymbol{\beta}_j$ is an unknown p -vector of coefficients relating absorbance to composition. From a causal viewpoint and Beer's law the reverse relationship is more natural, relating absorbance to composition. However, wavelength selection then becomes choice of responses and this is somewhat more problematic although it is the basis of some approaches; see for example the test of additional information in Reference 11. There are errors in both absorbances \mathbf{X} and composition \mathbf{Y} . This is not of direct relevance, however, since we must needs use what measurements we have to achieve a prediction equation of the same measurable quantities. We do not seek to predict unmeasurable error-free composition. We may also comment that although the relationship assumed is linear, it may be that the dependence is non-linear at many wavelengths.

Although the model is then strictly incorrect, the approach still has the scope to predict well if either

- (i) some informative regions of the spectrum show linearity or
- (ii) a linear combination of absorbances relates linearly.

The columns of \mathbf{X} and the composition column \mathbf{Y}_j are assumed centred. The n observations are assumed independent and normally distributed. Correlation is allowed between the q compositions within each of the n observations. Our model takes (1) a stage further and assumes that there is an unknown p -vector of 1s and 0s identifying a relatively small number of wavelengths at which the coefficients are non-zero. This binary vector, denoted γ , selects the wavelengths to include by its unit entries. We condition on its unknown value to specify our prior distributions for the regression coefficients and it therefore constitutes a latent variable. The number of ones in γ is defined as $t = p_\gamma$. There are 2^p of these selection vectors. *A priori* it is assumed that the probability of a particular vector is

$$\pi(\gamma|\theta) = \theta^t(1 - \theta)^{p-t} \quad (2)$$

Thus each wavelength is included (or not) independently with probability θ (or $1 - \theta$). If θ were prespecified as say $20/p$, then *a priori* we would expect around 20 wavelengths in the reduced model. The number of ones is binomial (p, θ), so that with $p = 250$ a 95% credibility interval is approximately 20 ± 9 . To relax this assumption, whilst retaining some preference for smaller models, we assume that θ in turn has a beta distribution $\pi(\theta)$, allowing θ to be either concentrated or widely dispersed. In short, γ has *a priori* a beta mixed binomial distribution

$$\pi(\gamma) = \int \pi(\gamma|\theta)\pi(\theta)d\theta \quad (3)$$

What we have tried to do here is to quantify, in the form of a probability distribution for γ , our belief that relatively small subsets of wavelengths will give adequate predictions for composition, without specifying which wavelengths.

What the analysis will do is update this prior probability distribution using the observed data to give a posterior probability distribution for γ . Before we can carry out this Bayesian analysis, we need to specify prior probability distributions for all the other unknowns involved. The most important of these are the coefficients β_j corresponding to the included wavelengths for any particular γ .

In this paper we take the included coefficients to be drawn from a normal distribution with mean vector zero and covariance matrix proportional to

$$c\{\mathbf{X}_{(\gamma)}^T \mathbf{X}_{(\gamma)}\}^{-1} \quad (4)$$

with the error variance for the j th regression being the proportionality factor. Here the suffix (γ) indicates the included coefficients. A moderate value of the constant c is recommended. This so-called g -prior of Reference 12 assumes that the prior distribution of the selected coefficients is as if it were formed from a prior experiment with a similar design matrix, most importantly reflecting similar correlation structures. The Bayesian assumptions and derivation are summarized in the Appendix. The full details of prior to posterior analysis are given in Reference 13, which offers a range of other prior structures, but neither the implementation of the g -prior nor the Metropolis algorithm of the next section is explored there. The approach is a multivariate generalization of References 14 and 15. In the case of the pseudo-data g -prior structure the posterior distribution of the p -vector γ is

$$\pi(\gamma|\mathbf{Y}, \mathbf{X}) \propto g(\gamma) = (c + 1)^{-p\gamma q/2} |\mathbf{Q}_\gamma|^{-(n+\delta+q-1)/2} \pi(\gamma) \quad (5)$$

with \propto indicating proportionality, prior $\pi(\gamma)$ given by (3) and where the $q \times q$ matrix \mathbf{Q}_γ is a prior

estimate plus a weighted average of the residual sum of products matrix and the total sum of products matrix,

$$\mathbf{Q}_\gamma = \mathbf{Q} + [c/(c+1)][\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}_\gamma(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^T\mathbf{Y}] + \mathbf{Y}^T\mathbf{Y}/(c+1)$$

Here \mathbf{Y} ($n \times q$) comprises the q -component \mathbf{Y} -vectors and $\mathbf{Q}/(\delta-2)$ ($q \times q$) is the prior expectation of the error covariance matrix, usually taken to be the form $k\mathbf{I}_q$, where \mathbf{I}_q is the $q \times q$ identity matrix and k is a suitably chosen constant. The uncertainty in this prior error covariance structure is summarized in the parameter δ , which we will generally take to be a small value, corresponding to high uncertainty; in particular, $\delta = 3$ in our applications.

MARKOV CHAIN MONTE CARLO

Equation (5) gives the posterior probability of each of the 2^p different γ vectors and thus of each choice of subset. In principle, all that remains is to compute these probabilities and select a few subsets with the highest probabilities for further investigation. When p is much greater than about 25, there are too many subsets for this to be feasible. Fortunately, we have available simulation methods that will find the γ vectors that have relatively high posterior probabilities. The method we use, Markov chain Monte Carlo, is designed to produce a sample from the posterior distribution (5). It does this by starting from a randomly chosen γ and then moves through a sequence of further values of γ , with each step in the sequence having an element of randomness. At each point in the sequence a new candidate γ is generated by randomly modifying the current one. If this has a higher probability, as given by (5), than the current one, then we move to it. If not, then we may still accept it, but now only with a certain probability. If we reject the new γ , then we just try again. The sequence of γ s thus generated is a realization of a Markov chain, and with an appropriate choice of acceptance probabilities we can ensure that the equilibrium distribution of this chain is the distribution given by (5). In particular, γ vectors with high posterior probability have more chance of appearing in the sequence, and we might expect that a long run of such a chain will therefore visit many of the best subsets. It may of course miss some, but short of an exhaustive search this is a risk we have to run.

For this paper we generate a candidate new selection vector γ^* from the current γ by one of two possible moves, the first with probability ϕ or the second with probability $1 - \phi$.

1. (Adding or deleting) Choose one of the p wavelengths at random. If the wavelength is currently in the model, delete the variable; if it is not currently in the model, add it to the model. Thus the new candidate γ^* differs from γ in one of its entries.
2. (Swapping) Choose at random one of the currently included wavelengths and at random one of the currently excluded wavelengths. For the new candidate model exclude the previously included wavelength and include the previously excluded wavelength.

The new candidate model coded as γ^* is accepted with probability

$$\min \left\{ \frac{g(\gamma^*)}{g(\gamma)}, 1 \right\} \quad (6)$$

There is considerable flexibility in how one designs 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/deletions through a higher value of ϕ . Furthermore, one could have chosen moves that added or subtracted or swapped two or three or more at a time, or a combination of these. In fact, if the new candidate γ^* is generated from the current γ according to a probability distribution $r(\gamma^*|\gamma)$ and is

accepted with probability

$$\min \left\{ \frac{r(\gamma|\gamma^*)g(\gamma^*)}{r(\gamma^*|\gamma)g(\gamma)}, 1 \right\} \quad (7)$$

then this is the Hastings¹⁶ generalization of the Metropolis algorithm and $r(\gamma^*|\gamma)$ is called the proposal distribution. If $r(\cdot|\cdot)$ is symmetric in its two arguments, then the Metropolis–Hastings acceptance probability (7) reverts to the Metropolis acceptance probability (6). Gibbs sampling provides another particular case of the Metropolis–Hastings algorithm and was used in Reference 13. One very useful feature of these algorithms is that they do not require the normalizing constant in (5): it is only the relative probabilities provided by $g(\cdot)$ that are needed.

Under fairly general conditions, provided that the Markov chain defined by the proposal distribution $r(\cdot|\cdot)$ is irreducible (every state can be reached from every other state), then the chain will converge to a stationary distribution given by (5). In general practice, 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 five different starting points and looking at the five marginal distributions provided by the computed $g(\cdot)$ values of the visited γ . Because we know the relative probabilities, we do not need to worry about creating a burn-in period.

APPLICATION

We illustrate the methodology on near-infrared data of three sugars, sucrose, glucose and fructose, present in varying concentrations in aqueous solution, originally analysed in References 4 and 17. 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 to 2500 nm in steps of 2 nm. There were 125 training samples and 21 further samples reserved for later prediction. For illustration, and to reduce computation, we chose 250 from the 700 wavelengths, equally spaced from 1100 to 2500 nm, by linear interpolation. Thus the number of explanatory variables is $p = 250$.

In keeping with the methodology suggested, we selected the g -prior distribution with $c = 4$. The value of four ($\approx 1.96^2$) is suggested by the sampling standard error of estimation: the prior suggests that important coefficients are around two standard errors (or larger).

In line with the number of explanatory variables needed in similar applications, and to induce a ‘small’ model, we chose the Bernoulli prior (2) with a beta mixing prior having an expectation of 20. This was specified relatively weakly by having a ‘flat’ beta distribution with the sum of its two parameters being two. 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 MCMC sampling runs; all but the last randomly permuted the wavelengths first and then chose $\gamma_j =$ (i) all 1s, (ii) half 1s, (iii) 20 1s, (iv) one 1 and (v) first 20 1s (unpermuted). There were 25 000 iterations in each run, where an iteration consisted of an attempt to move by either adding/deleting or swapping as given by steps 1 and 2 taken with equal probability $\phi = \frac{1}{2}$. Each run took about 2 h on a SUN Sparcstation. Acceptance of the possible move by (6) was by generation of a Bernoulli random variable. Computation of $g(\gamma)$ and $g(\gamma^*)$ was done using the QR decomposition and *qrdelete* and *qrinsert* of MATLAB (The Math works, Inc., Natick, Mass., U.S.A.). Every tenth 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.

For each run of the successful moves of 25 000 iterations we recorded the γ s and their corresponding $g(\gamma)$ relative probability. There were around 2000–3000 successful moves for each of

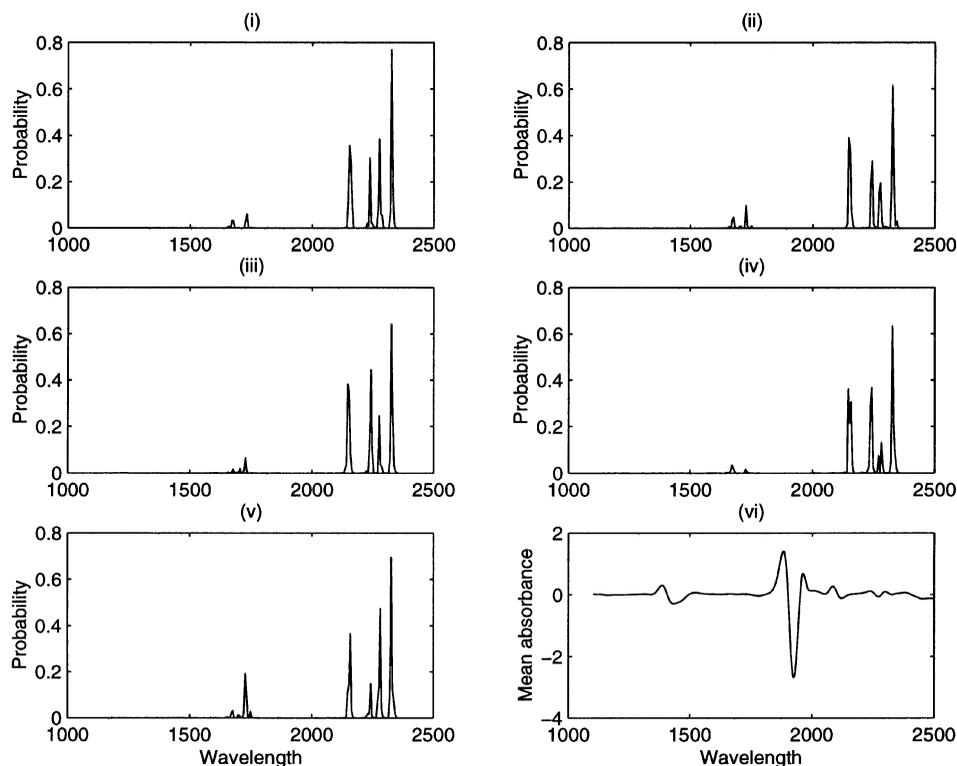


Figure 1. (i)–(v) Marginal probabilities of components of γ for five runs against wavelength (nm). (vi) Mean spectrum

the five runs. Of these successful moves, most (around 95%) were swaps. The successful move vectors were reduced to the set of distinct γ s, and most of the successful moves in each run were distinct. The relative probabilities of the set of distinct γ s were then normalized to unity over the visited γ vectors. The marginal probabilities for the 250 components of γ , $P(\gamma_j = 1)$, are plotted in (i)–(v) of Figure 1 where j runs from 1100 to 2500 nm in steps of 5.6 nm. The spikes are where wavelengths have been included with high probability. The locations of these spikes may be judged relative to the mean spectrum of graph (vi) of this figure. Despite some differences, the five 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 one of the runs, (iii), two further plots are given. Firstly, the number of 1s in γ is plotted over the 25000 iterations in Figure 2(a) and looks to have settled down to around five after starting at 20. Secondly, the $\log(\text{relative probabilities}) \log(g(\gamma))$ of visited γ s are plotted in their occurrence order in the iterative sequence in Figure 2(b). The $g(\gamma)$ s quickly increase to the level at which they settle down.

Figure 3 corresponds to the union of γ s from the five runs, normalizing the relative probabilities of the distinct γ s and displaying the marginal probabilities $P(\gamma_j = 1)$, where j indexes the 250 wavelengths from 1100 to 2500 nm in equally spaced steps of 5.6 nm.

We decided to summarize the posterior probability on wavelengths by including all wavelengths whose visited posterior probability was at least 0.05. This gives 14 wavelengths: {1727, 2147, 2153, 2158, 2237, 2242, 2248, 2270, 2276, 2282, 2321, 2326, 2332, 2338} nm. Using these 14 wavelengths for a simple least squares prediction for the 21 observation validation data (with centring correction

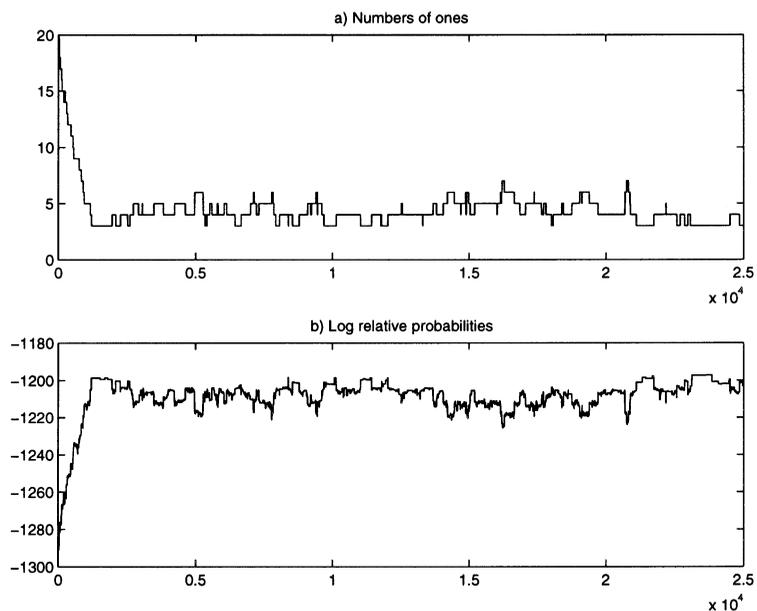


Figure 2. Plots in time sequence order for run (iii) for 25 000 iterations

using training sample means of composition and spectrum) gave mean squared errors of 0.44, 0.54 and 0.14 for prediction of sucrose, glucose and fructose respectively. The validation sample has ingredients at three levels, 0, 12 and 25 mass%, with a variance of 156. This implies that more than

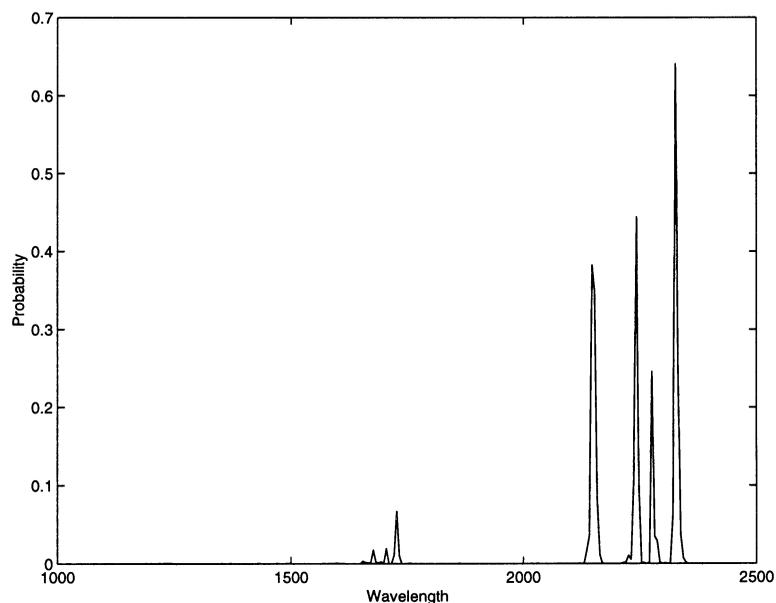


Figure 3. Marginal probabilities of components of γ against wavelength (nm)

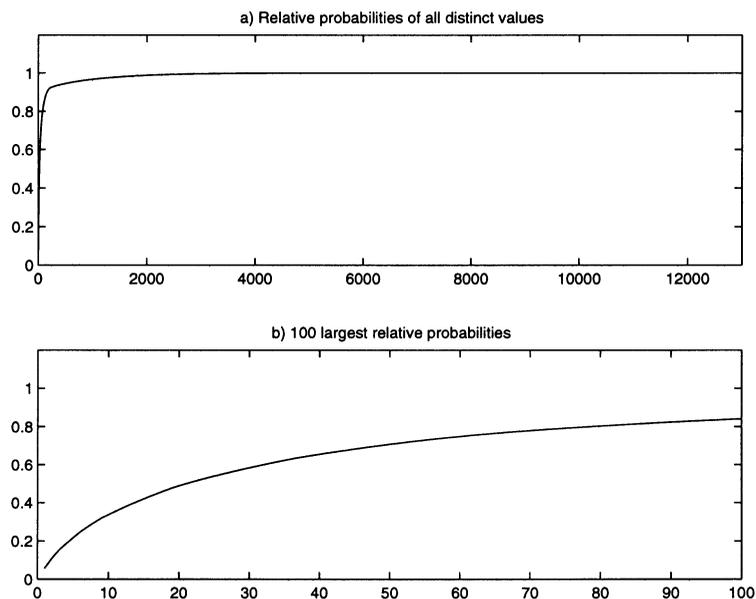


Figure 4. Cumulative relative probabilities versus number of distinct γ s ordered by probability

99.6% of 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.

Figure 4 gives the cumulative probability distribution of visited γ s against number of γ s, where γ s are ordered according to probability. This shows that around 100 of the 13 000 distinct visited models account for about 80% of the visited posterior probability. The single best model of the 13 000 visited includes just three wavelengths, {2153, 2242, 2326} nm, but with somewhat worse prediction mean squared errors of 0.69, 0.80 and 0.63, for prediction of sucrose, glucose and fructose respectively. It seems natural to avoid banding of wavelengths in a model. Choice by the marginal distribution on wavelengths above has this undesirable banding feature. This banding can be sidestepped by choosing say the top ten most likely γ s or by some probability cut-off applied to the cumulative probabilities in Figure 4. One can then average the least squares predictions, perhaps by their relative probability amongst the chosen models. In this instance the model averaging of the 100 best γ s gives mean squared errors of 0.33, 0.67 and 0.31, improving on the single best model. Model averaging also provides one insurance against any bias due to extensive searching in that some of those models in the averaging process may be good models and not spuriously thrown up. The Bayesian prior structure itself with its shrinkage-type predictions used after selection also tends to lessen selection effects; see for example Reference 18.

The mean squared errors found here are comparable in size with the best selections of Reference 17 and give orders-of-magnitude improvement over whole spectrum PLS approaches described there.

MATLAB software used in this paper may be obtained on the Web at: <http://stork.ukc.ac.uk/IMS/statistics/people/M.Vannucci.html>.

CONCLUSIONS

A new method of multivariate linear regression variable selection has been presented. The application

is to multicomponent NIR calibration, seeking to predict the composition of a mixture of sugars from 250 wavelengths, equally spaced from 1100 to 2500 nm. The methodology is Bayesian, but with prior distributions specified in a fairly automated fashion. In order to compute posterior probabilities over the vast number of selection models, the Metropolis algorithm is used. Our implementation of this specific Markov chain Monte Carlo method uses single-variable addition/deletion and swap moves to probabilistically explore different subset models. Model averaging for prediction may be used to give a more robust prediction which does not just focus on one over-selected model.

APPENDIX: BAYESIAN THEORY

This appendix gives a brief summary of the model and prior to posterior Bayesian analysis derived in Reference 13.

The data consist of $n = 125$ independent observations of $q = 3$ composition variables. The data for the j th composition variable are assumed to arise from normal regression model (1), regressing on the vectors of absorbances in the model determined by the latent selection vector γ . The regression model for all the compositions also includes covariances and is multivariate so that the likelihood or probability of the data given the parameters is obtained from multivariate normal probability densities. The full set of unknown parameters is \mathbf{B} , $\mathbf{\Sigma}$, γ , where columns of $\mathbf{B}(p \times q)$ are β_j of (1), $\mathbf{\Sigma}$ is the $q \times q$ error covariance matrix and γ is the $p \times 1$ binary selection vector and defines rows of \mathbf{B} to be included in the model. Conditional on γ and $\mathbf{\Sigma}$ the prior for the selected rows of \mathbf{B} is matrix-variate normal, with column covariance structure $\mathbf{\Sigma}$ and row covariance structure defined by the g -prior (4). In turn, $\mathbf{\Sigma}$ is assumed to have an inverse Wishart distribution with expectation $\mathbf{Q}/(\delta-2)$, with δ the shape parameter. The selection vector γ is assumed to have a Bernoulli structure (2) conditional on θ with a beta distribution on θ . All that is required analytically is to multiply the likelihood by the prior density of \mathbf{B} , $\mathbf{\Sigma}$, γ as structured above and then integrate out \mathbf{B} and $\mathbf{\Sigma}$ given γ to give the posterior relative probabilities (5). Because of the nature of the assumed conjugate distributions, this is straightforward and the reader is referred to Reference 13 for the details.

ACKNOWLEDGEMENTS

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

REFERENCES

1. H. Mark, *Principles and Practice of Spectroscopic Calibration*, Wiley, New York (1991).
2. W. R. Hrushka, in *Near Infrared Technology in the Agriculture and Food Industries*, ed. by P. Williams and K. Norris, AACC, St Paul, MN, 35–55 (1987).
3. P. J. Brown, C. H. Spiegelman and M. C. Denham, *Philos. Trans. R. Soc. A*, **337**, 311–322 (1991).
4. P. J. Brown, *Measurement, Regression, and Calibration*, Clarendon, Oxford (1993).
5. K. G. Kowalski, *Chemometrics Intell. Lab. Syst.* **9**, 177–184 (1990).
6. J. H. Kalivas, N. Roberts and J. M. Sutter, *Anal. Chem.* **61**, 2024–2030 (1989).
7. H. Kubinyi, *J. Chemometrics*, **10**, 119–133 (1996).
8. R. Leardi, R. Boggia and M. Terrile, *J. Chemometrics*, **6**, 267–281 (1992).
9. C. B. Lucasius, M. L. M. Beckers and G. Kateman, *Anal. Chim. Acta*, **286**, 135–153 (1994).
10. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, *J. Chem. Phys.* **21**, 1087–1091 (1953).
11. P. J. Brown, *J. R. Statist. Soc. B*, **44**, 287–321 (1982).
12. A. Zellner, in *Bayesian Inference and Decision Techniques—Essays in Honour of Bruno de Finetti*, ed. by P. K. Goel and A. Zellner, pp 233–243, North-Holland, Amsterdam (1986).
13. P. J. Brown, M. Vannucci and T. Fearn, *J. R. Statist. Soc. B.* **60**, in press.

14. E. I. George and R. E. McCulloch, *J. Am. Statist. Soc.* **88**, 881–889 (1993).
15. E. I. George and R. E. McCulloch, *Statist. Sinica*, **7**, 339–373 (1997).
16. W. K. Hastings, *Biometrika*, **57**, 97–109 (1970).
17. P. J. Brown, *J. Chemometrics*, **6**, 151–161 (1992).
18. J. B. Copas, *J. R. Statist. Soc. B*, **45**, 311–354 (1983).