PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Chemometrics and Spectral Frequency Selection

Philip J. Brown, Clifford H. Spiegelman and Michael C. Denham

Phil. Trans. R. Soc. Lond. A 1991 **337**, 311-322
doi: 10.1098/rsta.1991.0127

Chemometrics and spectral frequency selection

By Philip J. Brown[1], Clifford H. Spiegelman[2] and Michael C. Denham[1]

[1] *Department of Statistics and Computational Mathematics, University of Liverpool, PO Box 147, Liverpool L69 3BX, U.K.*

[2] *Department of Statistics, Texas A & M, College Station, Texas 77843, U.S.A.*

In many fields of science, the simple straight line has received more attention as a basis for calibration than any other form. This is because measuring devices have been mainly univariate and have had calibration curves which were sufficiently linear. As scientific fields become more computationally intensive they rely on more computer-driven multivariate measurement devices. The number of responses may be large. For example modern scanning infrared (IR) spectroscopes measure the absorptions or reflectances at a sequence of around one thousand frequencies. Training data may consist of the order of 10 to 100 carefully designed samples for which the true composition is either known by formulation or accurately determined by wet chemistry. In future one wishes to predict the true composition from the spectrum. In this paper we develop a variable selection approach which is both simple in concept and computationally easy to implement. Its motivation is the minimization of the width of a confidence interval. The technique for data reduction is illustrated on a mid-IR spectroscopic analysis of a liquid detergent in which the calibrating data consists of 12 observations of absorptions at 1168 frequency channels (responses) corresponding to five chemical ingredients.

1. Introduction

1.1. *Spectroscopic data*

Infrared spectroscopy involves directing a pulse of infrared (IR) light onto a substance, typically but not exclusively a liquid, noting the energizing effect over a short interval of time, and converting this to absorbances at various frequencies. Modern scanning instruments allow the simultaneous examination of a frequency range of around 1000 contiguous frequencies. The absorbances plotted at these frequencies represent the spectrum for that liquid sample. The 12 mid-IR spectra samples plotted over 1168 distinct frequencies are presented in figure 1. The samples are different mixtures of four detergent ingredients in aqueous solution. Each mixture involves a different amount of each of the five constituents. These plots appear continuous up to the resolution of the plotting device but are in reality 1168 discrete points. The 12 plots appear very similar at first sight. These data make up our calibrating data used to fit and analyse the models of this paper.

Figure 2 gives the sample mean and variance of the mid-IR spectra for the calibration data, averaging over the 12 observations. The 12 curves are broadly of a similar form to the mean curve, but there is also considerable variation, usually at those frequencies where absorbances are high. In fact we shall see that very accurate predictions of concentrations of ingredients may be obtained from these curves.

Phil. Trans. R. Soc. Lond. A (1991) **337**, 311–322
Printed in Great Britain

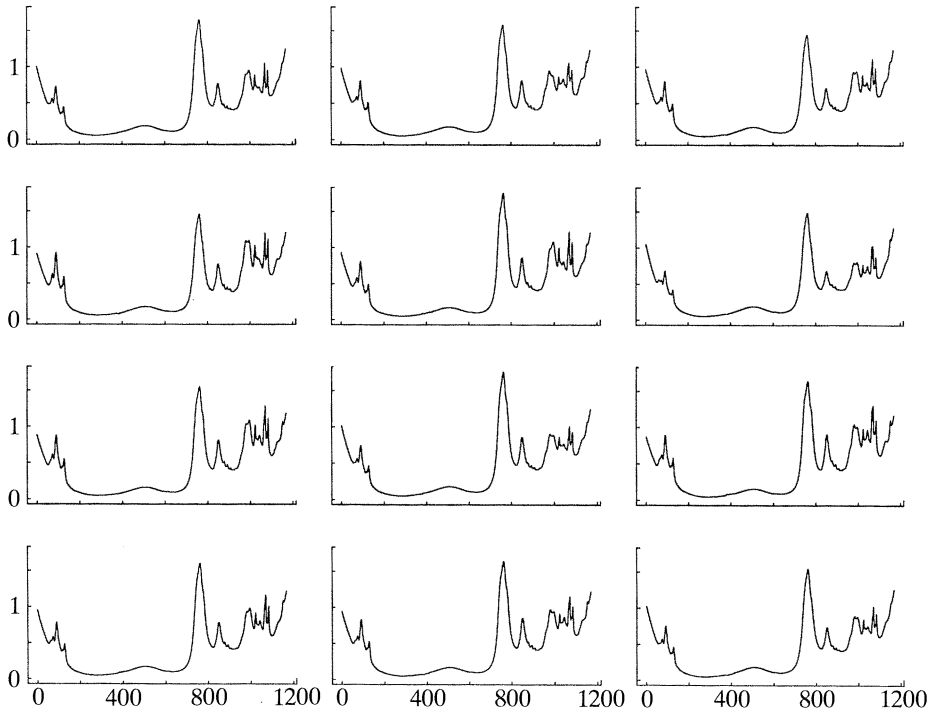


Figure 1. Absorbance spectra for the 12 detergent samples.

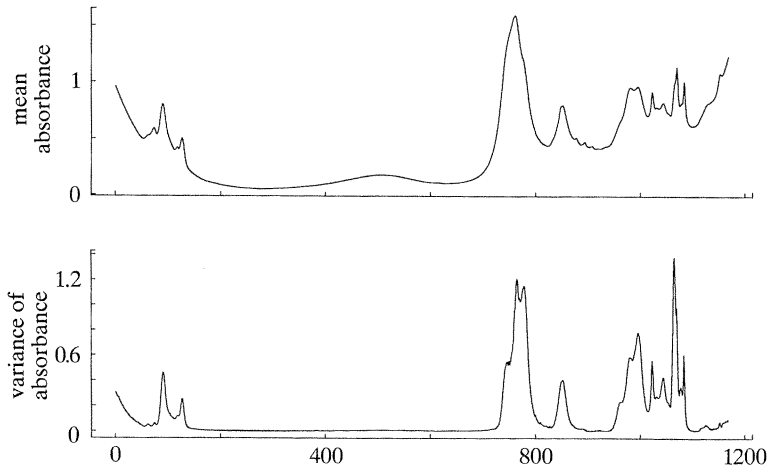


Figure 2. Mean and variance of 12 absorbance spectra.

The relationship of spectrum to ingredients may be motivated as follows. Briefly, molecules of a substance are able to absorb the incident IR radiation by moving between different vibrational and rotational energy levels of the lowest electronic energy state. In the very simple case of a diatomic A–B molecule the only vibration that can occur is a periodic stretching along the A–B bond. These stretching motions allow the vibrational frequency to be approximately predicted from Hooke's law. It is possible to consider the vibrations of individual bonds in more complex molecules in a similar manner, but other forms of vibration of individual bonds become

possible, including rocking, scissoring, twisting and wagging (see, for example, Cross & Jones 1969). Spectroscopists have verified that specific absorption bands for particular bonds or groups within a molecule occur at, or near, the expected frequencies. All these vibrational mode absorption frequencies tend to be altered in varying degrees by small changes in the remainder of the molecule, leading to highly characteristic IR spectra for organic substances.

The IR absorption of a particular mixture will be related to what chemicals are present and in what amounts. In fact in many circumstances the relationship has been found to be linear. This relationship is characterized by Beer's law (Cross & Jones 1969), which stipulates a linear relationship between the concentration of a substance dispersed in a non-absorbing medium and the amount of light absorbed by it. This is only an experimental law and departures from linearity occur for a variety of reasons, including lack of dispersion throughout the medium and scattering effects in the case of particulate solids.

Although the spectroscopist can identify bands of frequencies which give high absorbances for particular ingredients it is much harder to specify *a priori* those frequencies which best discriminate between the ingredients. Modern methods of chemometric analysis usually avoid such selection or apply some automated method. It is the purpose of this paper to explore a new method of frequency selection which is both easy to apply and demonstrated to be effective. This method is also able to cope with the high degree of indeterminacy in typical spectroscopic data where there are many more variables than observations.

1.2. *Regression models and overfitting*

There are two modes of approaching the fitting of a relationship between the spectrum ($Y, 1168 \times 1$) and the detergent and water concentrations ($x, 5 \times 1$). Suppose we regress Y on x. To predict x in future we may invert the fitted relationship, taking account of covariance structure. Or, more directly, we may regress x on Y so as to be able to predict x from Y in future. The latter is often favoured for its simplicity in the chemometrics literature. However, when Y is the response and embodies the error, statisticians have long advocated regressing in the direction of the error (see Eisenhart 1939; Williams 1969; Brown 1982). When there are fewer observations (12) than spectral variables (1168) then both approaches lead to the same indeterminate fit, with the same non-unique solution subspace (see Sundberg & Brown 1989). However, despite the conforming degeneracy, the focus of our approach is largely that of regression of spectra on concentrations, as we regard this as conceptually most satisfactory when the concentration data are designed in calibration as in the present example.

In such an indeterminate, over-parametrized problem there are a variety of ways to regularize and achieve stable estimates: partial least squares regression (see Martens & Naes 1989; Brown 1990); principal components regression; ridge regression (Marquardt 1979); coherent calibration (Brown & Mäkeläinen 1992); minimum length least squares (using the Moore–Penrose generalized inverse); and continuum regression (Stone & Brooks 1990). None of these methods is unique. They depend critically on the scales of the variables used to predict. All assume implicitly, or explicitly in the case of Brown & Mäkeläinen (1992), some sort of prior assumptions for the true regression coefficients, and will fare more or less well depending on whether or not such implicit assumptions approximately hold, which can be influenced by the choice of transformation. Brown & Mäkeläinen (1992) and Denham

& Brown (1990) utilize the fact that absorbance is a continuous function of frequency, the former assuming a stationary gaussian process prior distribution, the latter using splines and autoregressive error structures.

Although it is appealing to recognize that inferentially one can in principle always do better by using more information, since one is then always free to ignore the information, in scientific research there is a tradition which emphasizes the fragility of inference with large numbers of variables and the inevitable associated degree of overfitting. Even in the closed bayesian inferential system, tractable prior distributions for large numbers of parameters have unforeseen and undesirable features (see Dawid 1988; Mäkeläinen & Brown 1988). On the other hand, in the sampling framework of statistical inference, a set of complementary problems arise. A large number of estimates of effects which are truly zero will throw up at random a proportion of large estimates of seemingly large effects; the problem of multiple comparison. Regularization or shrinkage offers one answer, but may still be open to some overfitting. Internal leave-one-out cross-validation, a valuable tool against self-deception (see Stone 1974), may flatter in its internal cross-validated mean squared errors, especially as it assumes future prediction data that is very like that sampled. In a further application of the technique of this paper, using a validation set whose range is outside that of the calibrating data to predict the concentrations of three sugars in solution, the cross-validation mean squared errors were found to be an order of magnitude too small (Brown 1991).

In the approach we adopt here we remove a large number of frequencies so as to reduce vulnerability to overfitting. This is achieved by choosing frequencies to minimize the length of a confidence interval. Before specifying our method in §4 we review approaches to variable selection in the next section. The model adopted is presented in §3 and the application described in §5. The reader wishing to skip the technical details of §4 may skip to the end of that section for a brief summary of the method before proceeding to the example.

2. Variable selection reviewed

The large number of responses and the relative paucity of observations in modern calibration make many of the standard variable selection techniques used in regression inapplicable or computationally prohibitive. Commonly used methods of variable selection are all possible regressions, best subset selection, backward elimination, forward selection and stepwise regression. For details of these approaches the reader is referred to Miller (1990). Clearly, since backward elimination relies on including all variables in the regression model and then eliminating variables, it cannot be used due to the non-uniqueness of the regression equation based on including all q variables. The use of the all possible regressions approach is also infeasible. Even if we restrict attention to those subsets which give unique estimators, the number of models which must be considered is vast. Best subset regression, which requires a fraction of the time required by all possible regressions, will still consider too many regressions for it to be practicable in many cases.

The use of forward selection methods is widespread (Fearn 1983), and some instruments have such an approach incorporated in their accompanying software (see Osborne *et al.* 1984). Fearn (1983) gives an example of calibrating protein content in ground wheat where forward selection is found wanting, partly because of high correlations between absorbances at different frequencies and low correlation

with the component to be predicted. The use of stepwise regression also suffers in a similar way when applied to this example.

Selection techniques geared to prediction of the controlled variable via the fitted regression of Y on x have the same computational drawbacks. Brown (1982) uses a test of 'additional information' to select a subset of variables, and Spezzaferri (1985) uses a bayesian information theoretic approach.

Sometimes authors use a variable selection technique after taking principal components. This does not, however, meet our aim of selecting frequencies since all of the original set of frequency channels are still retained.

So far in this section we have assumed that we are concerned with predicting only one component. With $p > 1$ components one may consider the separate predictions and take the union of the frequencies selected for each prediction. Looking for a frequency set which simultaneously discriminates among the p components may be preferable if the aim is to choose the smallest subset. It would require a weighted sum of residual mean squares as an overall measure of goodness. Such an approach would add considerably to the complexity and entail a heavy computational overhead.

In this paper we propose an extremely simple technique for response selection. It does not claim to give the smallest possible subset but it does provide a predictively good subset which substantially reduces the number of channels.

3. The model for selection

As discussed we adopt the approach of regression of the absorption spectra on composition. We also regress on only one ingredient at a time. Although it might superficially seem desirable to regress on all five ingredients simultaneously, or at least on four of them since they total 100%, we want subsequently to predict the concentration of each of the four detergent chemicals using just the spectrum without knowledge of concentrations of the other three chemicals in the sample, see Brown (1982) for further discussion.

For the moment assume the simpler linear regression models,

$$Y_{kj} = \alpha_j + \beta_j x_k + \epsilon_{kj}, \tag{1}$$

where $j = 1, \ldots, q$ indexes the $q = 1168$ frequencies, and $k = 1, \ldots, n$ the $n = 12$ observations. Here x_k is the concentration of one of the ingredients, for the kth sample. In this model, errors $\{\epsilon_{kj}\}$ have zero mean and variance σ_j^2 and are assumed to be uncorrelated. The error variance at frequency j, σ_j^2, incorporates both measurement error and residual effects due to omitted constituents of the sample. The uncorrelatedness of the errors is perhaps an over-simplification since correlation is induced by these omitted constituents. In this case one could entertain the possibility of correlation of Y_{kj} across channels within an observation, but with the not insubstantial additional problem of a very large unknown covariance matrix; see Denham & Brown (1990) for some possible ways of structuring such problems. If such correlations are substantial it will be important to use them explicitly for prediction or to use them implicitly by a regularized method, such as partial least squares, that regresses x on Y.

As already discussed, the formulation is in terms of controlled calibration, where the explanatory variables are of future interest in prediction. Controlled calibration is most appropriate when these explanatory variables are initially fixed in some designed calibration experiment, as in the present example.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

The n observations $\{Y_{kj}, x_k; j = 1, \ldots, q; k = 1, \ldots, n\}$ serve to estimate the unknown parameters, $\alpha_j, \beta_j, \sigma_j^2$ in model (1). With these q estimated calibrated relationships we may predict any further unknown sample compositions $\xi_t, t = 1, \ldots, T$, where we have designated unknown x by ξ. For this prediction problem the analogous model to (1) is

$$Z_{tj} = \alpha_j + \beta_j \xi_t + \epsilon_{tj}, \tag{2}$$

where ϵ_{tj} are uncorrelated errors with variance $\sigma_j^2 (j = 1, \ldots, q; t = 1, \ldots, T)$.

We use the single-component method to tackle the multicomponent situation as in our illustration in §5 by successively applying the single composition results marginally to the p components and amalgamating the p sets of selected wavelengths. That is, we look to discriminate in turn between each ingredient and the other ingredients taken together.

One approach to choice of model dimension is to choose the size of model by calculating the mean squared error of prediction. This mean squared error is the sum of a variance and a bias squared term. Larger models have increased variance but decreased bias and an optimum compromise between the two is possible. A related but different approach which we adopt focuses on prediction intervals. The fitted model from (1) is not the true model. It is estimated once but may be used repeatedly to predict future values of ξ, each prediction incorporating the same bias of fitting. In addition to this source of error from the calibrating experiment there is the error generated by the postcalibration or prediction experiment. These two sources of error are affected by the size of model or number, q, of frequency channels adopted and determine the width of the confidence interval. We do not here look at the simultaneous satisfaction of all future use prediction intervals with an ascribed probability, rather we focus marginally on each single use interval. Again we are able to achieve an optimal compromise which minimizes the width of interval and gives a unique choice of selected channels.

When q is large, the numerous systematic errors will give rise to a proportion which is excessively large and spawns a related large literature of methods which seek to shrink estimates towards zero.

Quite distinct but equally problematic considerations arise within a bayesian framework. Suppose $\xi_1, \ldots, \xi_T$ are viewed as exchangeable *a priori*, that is, having a joint distribution which is invariant to permutations of the indices and consequently not in general independent; then $Z_1, \ldots, Z_t$ provide information on the form of this exchangeability, and posterior inference about ξ_t will depend on $Z_1, \ldots, Z_{t-1}$ as well as Z_t and the training data. Additionally, the multiple use of the calibration with its prior necessitates a careful assessment of the stability of inference to alterations of this prior distribution (see Berry 1988).

For the selection of frequency channels in this paper we work within the sampling theory inferential framework. Our basic idea is to choose that subset of q' of the q channels such that the approximated length of each single use prediction interval is minimized. Both q' and the corresponding subset of frequencies are chosen by the method.

4. The selection method

Since x and ξ are scalars, we seek a linear combination of the response to each instrument. We can then use the well-developed univariate methodology. Henceforth in the prediction model (2) we refer to the tth future Z and drop the subscript t to it. Let $Z = (Z_1, \ldots, Z_q)^{\mathrm{T}}$ and $Z_\theta = \Sigma \theta_j Z_j$. Similarly let $\alpha_\theta = \Sigma \theta_j \alpha_j$ and $\beta_\theta = \Sigma \theta_j \beta_j$. We

look for values of $\theta = (\theta_1, \ldots, \theta_q)^{\mathrm{T}}$, with typically $q' < q$ non-zero components, which minimize the approximate length of certain confidence intervals. In models (1), (2) in addition to the second-order error assumptions the errors are taken to be normally distributed. Readers wishing to skip the following technical derivation may proceed to the last paragraph of this section.

For prescribed θ, the compound response Z_θ is normally distributed with, as mean, the calibration line, $\theta^{\mathrm{T}}E(Z) = \alpha_\theta + \beta_\theta \xi$, and variance $\sum \theta_j^2 \sigma_j^2$. We proceed by means of the Cauchy–Schwarz inequality. First, with prescribed confidence level $1-\gamma$ and fixed θ with q' specific non-zero components,

$$|Z_\theta - \sum \theta_j(\alpha_j + \beta_j \xi)| = |\sum \theta_j \sigma_j \epsilon_j^*| \leqslant \sqrt{[\sum \theta_j^2 \sigma_j^2]}\sqrt{[\chi^2_{1-\gamma}(q')]} \tag{3}$$

since ϵ_j^* are now independent standard normal. Secondly and similarly, with probability $1-\delta$,

$$|\sum \theta_j(\hat{\alpha}_j + \hat{\beta}_j \xi) - \sum \theta_j(\alpha_j + \beta_j \xi)| \leqslant \sqrt{[\sum \theta_j^2 \sigma_j^2 s^2(\xi)]}\sqrt{[\chi^2_{1-\delta}(q')]}, \tag{4}$$

where $s^2(\xi) = [1/n + \{(\xi - \bar{x})^2 / \sum (x_k - \bar{x})^2\}]$ and q' is the number of prescribed non-zero θ_j. These statements will also be true for θ chosen by the calibrating data in model (1) provided the randomness induced does not effect the actual channels selected. We give sufficient conditions for this later.

Let B be the event represented by (3) and A the event represented by (4). Then

$$\begin{aligned} P(B|A) &= P(A,B)/P(A) \\ &= [P(A) + P(B) - P(A \cup B)]/P(A) \geqslant 1 + [(1-\gamma) - 1]/(1-\gamma) \\ &= 1 - \gamma/(1-\delta) \\ &\doteq 1 - \gamma. \end{aligned}$$

Thus, for 'good' calibrating events prescribed by (4) which occur with probability $(1-\delta)$, we have future predictions prescribed by (3) with probability approximately $(1-\gamma)$, the approximation being better the smaller the value of γ and δ. Inequality (4) may be adapted to give a statement for all future ξ, and the χ^2 degrees of freedom increase to $2q'$. With inequality (3) this would form a basis for simultaneous repeated-use confidence intervals (see also Scheffé 1973; Carroll, *et al.* 1988). We, however, prefer the single-use statements.

We let $c_1 = \sqrt{[\chi^2_{1-\gamma}(q')]}$ and $c_2 = \sqrt{[\chi^2_{1-\delta}(q')]}$. Now, the triangle inequality gives $|Z_\theta - \hat{E}(Z_\theta)| \leqslant |Z_\theta - E(Z_\theta)| + |E(Z_\theta) - \hat{E}(Z_\theta)|$, which enables us to combine (3) and (4) into a single inequality for the divergence of Z_θ from its fitted value for given ξ. Solving this for ξ would give the single-use confidence region. More simply, thinking of a graph of Z against ξ for a limited region of ξ values, we may approximate the width of the confidence interval by 'height' divided by 'slope', essentially a local linear Taylor series expansion (see, for example, Carroll & Spiegelman 1986).

The approximate half-width is thus

$$\left\{c_1 \sqrt{\left[\sum_1^{q'} \theta_j^2 \sigma_j^2\right]} + c_2 \sqrt{\left[\sum_1^{q'} \theta_j^2 \sigma_j^2 s^2(\xi)\right]}\right\} \Big/ \left|\sum_1^{q'} \theta_j \hat{\beta}_j\right|. \tag{5}$$

This is the quantity we will minimize with respect to the q' channels with non-zero θ_j.

When σ_j^2 have to be estimated from the data they are replaced by the usual unbiased estimators $\hat{\sigma}_j^2$, but for this paper we have made no attempt to adjust the probability statements appropriately.

The non-zero θ_j which minimize (5) may be easily seen to be proportional to $\hat{\theta}_j$, where

$$\hat{\theta}_j = \left(\frac{\hat{\beta}_j}{\sigma_j^2}\right) \Big/ \left(\sum_1^{q'} \frac{\hat{\beta}_j^2}{\sigma_j^2}\right), \tag{6}$$

$j = 1, \ldots, q'$, and for notational convenience we have assumed that the selected frequencies are the first q' out of q. Our estimator of ξ obtained from the linear compound of the $Z_j, j = 1, \ldots, q$, obtained from (6) is

$$\hat{\xi} = Z_{\hat{\theta}} - \hat{\alpha}_{\hat{\theta}}, \tag{7}$$

since $\hat{\beta}_{\hat{\theta}} = \Sigma \hat{\beta}_j \hat{\theta}_j, = 1$.

Coincidentally, (6) are the generalized least squares estimators conventionally used in multivariate controlled calibration, and represent a special unicomponent case of equation (2.16) of Brown (1982) with diagonal covariance structure. The minimized half-length of interval (5), substituted by (6), is

$$[c_1 + c_2 s(\xi)] \Big/ \sqrt{\left[\sum_1^{q'} \frac{\hat{\beta}_j^2}{\sigma_j^2}\right]}. \tag{8}$$

With c_1 and c_2 functions of q' and linked by $s(\xi)$ in the numerator of (8), our choice of q', and that subset of q' of the q channels, depends on the unknown ξ. However, if we choose equal probability levels $\gamma = \delta$, so that $c_1 = c_2 = c$, then minimization of (8) over q' becomes independent of ξ. Moreover, whether or not the levels are chosen equal, (8) is easily minimized by ordering the absolute values of the standardized slope coefficients $\hat{\beta}_j/\sigma_j$ and choosing the largest q' of these, and then taking that q' for which (8) is a minimum.

If we do not want the conditional interpretation for single-use curves characterized by two probability levels γ and δ, then inequalities (3) and (4) are replaced by

$$|Z_\theta - \Sigma \theta_j(\hat{\alpha}_j + \hat{\beta}_j \xi)| = |\Sigma \theta_j \sigma_j \epsilon_j^*| \leqslant \sqrt{[\Sigma \theta_j^2 \sigma_j^2 (1 + s^2(\xi))]} \sqrt{[\chi^2_{1-\eta}(q')]},$$

where $1 - \eta$ is the single confidence level. Taking the same linearizing approximation to this gives a narrower interval but the same form of half-width and the same selection, provided $\eta = \gamma = \delta$.

Sufficient conditions for their to be little variation in the channels selected are as follows.

1. The variances of all the estimated regression slope coefficients $\{\hat{\beta}_j\}$ are small.
2. The slopes $\{\hat{\beta}_j\}$ belong to one of two non-empty sets. The first has absolute value far from zero, $|\beta_j| \gg 0$. The second group has slopes approximately equal to zero. For all slopes β_j and β_k in the first group it is assumed that the distances $\|\beta_j| - |\beta_k\|$ are large.

It is evident that the number of components selected is a monotonic decreasing function of the both γ and δ, so that if they are both chosen large enough only one component is selected. Typical values $\gamma = \delta = 0.1$ allow enough information to be retained in our detergent example.

In summary, our method orders the absolute values of $\hat{\beta}_j/\sigma_j$, with $\hat{\sigma}_j$ estimating σ_j when as usual the error standard deviation is unknown. It chooses the frequencies corresponding to the largest q' of these, where q' minimizes (8), and here c_1^2 and c_2^2 are tabulated χ^2 percentage points on q' degrees of freedom. Dependence on $s(\xi)$ disappears if the two confidence levels are equal.

Figure 3

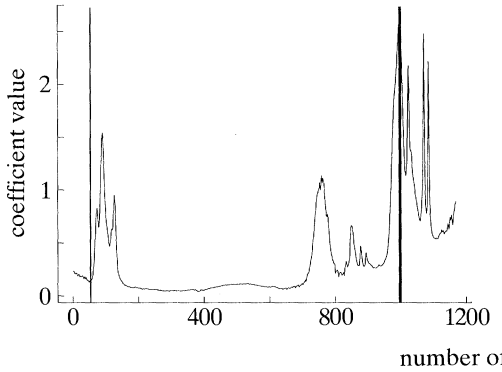
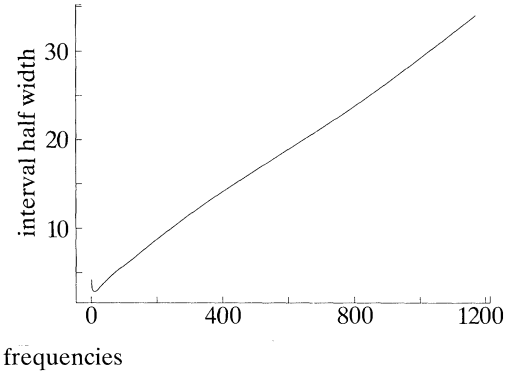


Figure 4



number of frequencies

Figure 3. Component 1 multiple regression coefficients with selected frequency bands.

Figure 4. Component 1 confidence interval half-width by number of chosen frequencies. Minimum occurs at 11.

5. The detergent example

This data-set consists of absorptions at $q = 1168$ mid-IR, equally spaced frequencies (channels) in the range 3100 to 750 cm^{-1}. This detergent is a solution in water of four chemicals, with water making up the fifth component of the mixture. Only $n = 12$ carefully designed samples were available with the five percentages of ingredients recorded for each of these. The data are further described and analysed in Denham & Brown (1990), using the continuity of absorbance as a function of frequency. Since the calibration design was quite strictly controlled, we adopt a controlled calibration mode of analysis throughout, regressing absorptions on ingredients and then 'inverting' this relationship to predict a future composition from the absorptions for the sample. However, we have here concentrated on a single x-variable, whereas the data have five. We adopt the following approach to this. First, treating the data as five completely separate data-sets, we select sets of frequencies for each of these data-sets. We then are able either to continue to treat the five datasets as quite separate and predict each component from the inverted simple linear regressions, using (7); or to adopt a hybrid multiparameter approach involving (*a*) pooling the five sets of chosen frequency channels and (*b*) 'inverting' the $\Sigma q' = Q$ multiple regressions of absorption on components to predict each component in future from a set of absorptions. This inversion would itself be a weighted multiple regression of the Q-vector $Z_{(Q)}$ on vectors of regression coefficients with, as design matrix, the $Q \times 5$ matrix of regression coefficients and Q different variances from the earlier multiple regression of $Z_{(Q)}$ on components. The estimator is thus provided by (2.16) of Brown (1982), namely,

$$\hat{\xi} = (\hat{B}S^{-1}\hat{B}^{\mathrm{T}})^{-1}\hat{B}S^{-1}(Z_{(Q)} - \hat{\alpha}),$$

with the S the full error covariance matrix of that paper replaced by a diagonal $Q \times Q$ covariance matrix.

For component 1, figure 3, gives the bands of frequency channels selected for chosen probability levels $\gamma = \delta = 0.1$. The positions of the chosen frequencies are indicated by vertical lines, and are superimposed on the graph of slope coefficients for this component from multiple regression of absorbances on components. For this component there are 11 selected channels, in order {998, 999, 997, 996, 1000, 995,

Table 1. *Root mean squared prediction error*

	component				
method	1	2	3	4	5
simple LS	3.44	0.95	1.77	0.36	1.38
multiple LS	0.39	0.47	0.40	0.11	0.30
select simple LS	1.53	0.43	0.81	0.39	0.94
select multiple LS	0.14	0.21	0.15	0.11	0.23

54, 1001, 994, 53, 55}, making up two quite separate groups of frequencies, {53, 54, 55} and {994–1000}. Figure 4 is proportional to the half-width (8) as a function of q' for component one, depicting the minimum at $q' = 11$ and a sharp increase as q' increases, with the value at the minimum being around one tenth of that with all 1168 frequencies included. Figure 4 is typical of plots of (8) for the other components, and these have consequently been omitted, as have the other component plots paralleling figure 3. The number of frequencies selected for components two to five are 9, 3, 7, 17, and all the selected channels are quite distinct, so that the pooled set of frequencies for the hybrid multiparameter prediction involves $Q = 47$ selected channels.

Table 1 gives leave-one-out root mean squared prediction errors for the five components corresponding to four different methods. The tabulated results are thus cross-validated so that the selection of wavelengths is based on 11 observations and the 12th is predicted for every subset of size 11. Two of the methods incorporate all 1168 wavelengths and two involve selection of wavelengths as indicated above. The two methods are further dichotomized by whether they are uni- or multicomponent. As a consequence of the diagonal covariance structure, there is a substantial reduction of prediction error in using multicomponent methods, and within either uni- or multicomponent the selection procedures of this paper are beneficial for most components. The variance of the four detergent ingredient percentages were 16.4, 5.5, 4.9, 2.9, respectively, so that predictions are generally very accurate with very high percentages of variation explained. As a further reference point, partial least squares on all frequencies with four latent factors gave root mean square prediction errors of 0.20, 0.29, 0.15, 0.11, 0.18 for the four components and water. Although these are only marginally less good than our preferred last row of table 1, further work on the calibration of sugars in Brown (1991) shows partial least squares on all frequencies to lack robustness with respect to different prediction sets. After selection of frequencies by developments of the method of this paper, such non-robustness disappears.

Note that we have used mean squared error as a prediction criterion. For simple linear regression the mean squared error is infinite. However, the use of mean squared error is justified for $q \geqslant 3$, since it is finite if and only if the number q of frequencies is at least three (see Brown & Spiegelman 1991). This paper provides for the selection process and sharpens the results of Lieftinck-Koeijers (1988).

6. Commentary

We have provided a new method of channel frequency selection. The method is strictly applicable to unicomponent calibration, but we have demonstrated by example that a hybrid multicomponent method is also effective. This hybrid version

coped with the multicomponents through $\hat{\sigma}_j$. If at a particular frequency other components also influenced the response, this would inflate the corresponding σ, reducing the standardized coefficient. It relies on discriminating between each component and the rest in turn. One can, however, envisage situations where the selection of frequencies should be based on a completely multicomponent method. This would be the case if for example particular frequencies were very good at discriminating between a pair of components on the one hand against the rest, but offered little discrimination between the pair, whose resolution could be achieved from other channels.

It is easy to see qualitatively how one might consider components simultaneously, and at the same time incorporate a correlation structure across frequencies. Some straightforward calculations following from Brown & Sundberg (1987) give the asymptotic, observed, second-derivatives, $p \times p$ matrix of the profile log-likelihood of ξ at the maximum likelihood estimate, that is, the information matrix for the vector of component ξs, as $\hat{B}\Gamma^{-1}\hat{B}^{\mathrm{T}}$. Here B is a $p \times q$ matrix of coefficients from the full version of model (1) formed by regressing the q absorptions on the $p = 4$ components. It is of interest to note that, when $p = 1$ and the $q \times q$ covariance matrix Γ is diagonal, then our criterion just amounts to accumulating the q' highest information components. Thus the ordering based on $\hat{\beta}_j^2/\sigma_j^2$ is quite natural. The use of information in itself does not give one a stopping rule which adequately takes account of the dimensionality of estimation.

It probably requires more experience with other data-sets to see whether or not the confidence levels γ and δ, can be prespecified and yet retain sufficient information. In the application of the methodology to the calibration of three sugars (Brown 1991), $\gamma = \delta = 0.001$ was chosen by cross-validation. That report also contains a simple modification to deal with an autoregressive error structure.

We are grateful for comments by Dr Rolf Sundberg on an earlier version, which led to an improved presentation. P.J.B. and M.C.D. are grateful to the SERC for providing a grant under the Complex Stochastic Systems Initiative. Shell UK also provided funding for work on selection in calibration with many variables. C.H.S. was funded by the US National Science Foundation and the Shell Development company.

References

Berry, D. A. 1988 Multiple comparisons, multiple tests and data dredging: a Bayesian perspective. In *Bayesian statistics* 3 (ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley & A. F. M. Smith), pp. 79–94. Oxford: Clarendon Press.

Brown, P. J. 1982 Multivariate calibration (with discussion). *Jl R. statist. Soc.* B, **44**, 287–321.

Brown, P. J. 1990 Partial least squares in perspective. *Analyt. Proc. R. Soc. Chem.* 303–306.

Brown, P. J. 1991 Variable selection with more variables than observations. Report to Shell Research, Sittingbourne.

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 & A. F. M. Smith). Oxford University Press. (In the press.)

Brown, P. J. & Spiegelman, C. H. 1991 Mean squared error and selection in multivariate calibration. *Statist. Prob. Lett.* **12**, 157–159.

Brown, P. J. & Sundberg, R. 1987 Confidence and conflict in multivariate calibration. *Jl R. statist. Soc.* B **49**, 46–57.

Carroll, R. J. & Spiegelman, C. H. 1986 The effect of ignoring small measuring errors in precision instrument calibration. *J. Quality Control* **18**, 170–173.

Carroll, R. J., Spiegelman, C. H. & Sacks, J. 1988. A quick and easy multiple-use calibration-curve procedure. *Technometrics* **30**, 137–141.

Cross, A. D. & Jones, R. A. 1969 *An introduction to practical infra-red spectroscopy*, 3rd edn. London: Butterworth.

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 & A. F. M. Smith), pp. 95–110. Oxford University Press.

Denham, M. C. & Brown, P. J. 1990 Calibration with many variables. Technical Report, Liverpool University.

Eisenhart, C. 1939 The interpretation of certain regression methods and their use in biological and industrial research. *Ann. math. Statist.* **10**, 162–186.

Fearn, T. 1983 A misuse of ridge regression in the calibration of a near infrared reflectance instrument. *Appl. Statist.* **32**, 73–79.

Lieftinck-Koeijers, C. A. J. 1988 Multivariate calibration: a generalisation of the classical estimator. *J. Multivariate Analysis* **25**, 31–44.

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 & A. F. M. Smith), pp. 677–696. Oxford University Press.

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

Martens, H. & Naes, T. 1989 *Multivariate calibration*. Chichester: Wiley.

Miller, A. 1990 *Subset selection in regression*. London: Chapman and Hall.

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.

Scheffé, H. 1973 A statistical theory of calibration. *Ann. Statist.* **1**, 1–37.

Spezzaferri, F. 1985 A note on multivariate calibration experiments. *Biometrics* **41**, 267–272.

Stone, M. 1974 Cross-validatory choice and assessment of statistical predictions. *Jl R. statist. Soc.* B **36**, 111–147.

Stone, M. & Brooks, R. J. 1990 Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Jl R. statist. Soc.* B **52**, 237–269.

Sundberg, R. & Brown, P. J. 1989 Multivariate calibration with more variables than observations. *Technometrics* **31**, 365–371.

Williams, E. J. 1969 Regression methods in calibration problems. In *Bull. Int. Statist. Inst.*, vol. 43, pp. 17–28.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES