J. R. Statist. Soc. B (1990)
52, *No.* 2, *pp.* 237–269

Continuum Regression: Cross-validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression

By M. STONE† and R. J. BROOKS

University College London, UK

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, October 25th, 1989*, Professor D. V. Hinkley *in the Chair*]

SUMMARY

The paper addresses the evergreen problem of construction of regressors for use in least squares multiple regression. In the context of a general sequential procedure for doing this, it is shown that, with a particular objective criterion for the construction, the procedures of ordinary least squares and principal components regression occupy the opposite ends of a continuous spectrum, with partial least squares lying in between. There are two adjustable 'parameters' controlling the procedure: 'alpha', in the continuum [0, 1], and 'omega', the number of regressors finally accepted. These control parameters are chosen by cross-validation. The method is illustrated by a range of examples of its application.

Keywords: CROSS-VALIDATION; LEAST SQUARES PREDICTION; PARTIAL LEAST SQUARES; PRINCIPAL COMPONENTS REGRESSION; SEQUENTIAL

1. INTRODUCTION

Least squares multiple regression with a single dependent variable finds application in a variety of scientific contexts. In what has been called 'hard science'—but that might, with a sense of history, be better described as *hardened science*—a given linear model is known to be an adequate representation of the truth: the number of unknown parameters in the model is usually quite small and all of them have to be estimated. A good example is provided by the pioneering work of Gauss (1826) on the triangulation of Hannover. For this, with 18 observations and seven parameters:

(a) there were no doubts about the model;
(b) the linear parameters (true angles minus good initial approximations) were large compared with their estimated standard errors and could not be arbitrarily taken to be zero;
(c) the least squares method served notoriously well.

At the other end of the hardness scale are the *soft science* applications, in which a number, sometimes a very large number, of explanatory variables is available. With little knowledge to go on and with the emphasis on prediction, the scientist would be willing to use any or all these variables in constructing *ad hoc* regressors for a predictor. Such was the problem faced by Fisher (1924) when he was asked to explain, if he could, the variation among a limited number of crop yields by means of a much larger number of meteorological variables. Fisher concluded that 'in order to arrive at unprejudiced results,

†*Address for correspondence:* Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, UK.

 0035-9246/90/52237 $2.00

(i) The meteorological variates to be employed must be chosen without reference to the actual crop record.
(ii) If multiple variates are to be used, allowance must be made for the positive bias of R^2.
(iii) Relationships of a complicated character should be sought only when long series of crop data are available.'

A contemporary illustration of a soft, but vitally useful, science is the calibration of a near infra-red reflectance spectrometer for the speedy measurement of the percentage of protein in samples of flour, which typically might have 25 observations and up to 700 explanatory variables (Fearn, 1983).

The terrain between the peaks of hardened science and the quicksands of soft science is occupied by *elastic science*. One variety of this is a mixture of the hard and the soft, which may be envisaged as a sort of bog with tussocks corresponding to *given* regressor variables that the scientist is determined to include, embedded in the soft matrix of additional *ad hoc* regressors. Examples of given, as opposed to constructed, regressors are experimental design variables known to have a major influence on the observations.

Questions of choice arise in all except the hardened science area. How do we actually construct the regressors that have to be constructed? How many should be made, if the straddling pitfalls of underfitting and overfitting are to be avoided? Some progress has been made in the last two or three decades, going well beyond Fisher's idea of an adjustment to R^2. The techniques of Mallows's C_p, Akaike's criterion and the like, as well as the more general approach of cross-validation, now provide some control of the excesses of prejudice and self-deception. However, in practice, these techniques do not often go the whole way in addressing the problem of high dimensionality of choice (Hjorth, 1989; Dijkstra, 1988).

Concentrating here entirely on prediction, we formulate a general method that brings the three separate techniques of ordinary least squares (OLS), partial least squares (PLS) and principal components regression (PCR) under the same mathematical umbrella. This formulation inspires a specific integrated procedure with a low dimensionality, 2, of adjustable control parameters chosen by cross-validation:

(a) α, a real number in the interval [0, 1], with the values 0, $\frac{1}{2}$ and 1 corresponding to OLS, PLS and PCR respectively;
(b) ω, the total number of regressors accepted ('given' plus 'constructed').

The role of α suggests the obvious title for this procedure—'continuum regression' (CR).

A feasible computational method is developed, and the paper concludes with illustrations of how the new procedure performs on some real data sets and (implicitly) how it compares with OLS, PLS and PCR.

2. SOME NOTATIONS AND SUPPOSITIONS

The generic unit of data is $(\dot{x}(1), \ldots, \dot{x}(p), \dot{y})$ or $(\dot{\mathbf{x}}, \dot{y})$ for short. The data are a sample of n such units: $(\dot{\mathbf{x}}_i, \dot{y}_i)$, $i = 1, \ldots, n$. The dots here denote basic data: their removal signifies the subtraction of sample averages, thus

$$y_i = \dot{y}_i - (\dot{y}_1 + \ldots + \dot{y}_n)/n = \dot{y}_i - \bar{\dot{y}}$$
$$x_i(j) = \dot{x}_i(j) - \{\dot{x}_1(j) + \ldots + \dot{x}_n(j)\}/n$$
$$x(j) = \dot{x}(j) - \{\dot{x}_1(j) + \ldots + \dot{x}_n(j)\}/n.$$

We write

$$\mathbf{x}_i = (x_i(1), \ldots, x_i(p))'$$
$$\mathbf{x} = (x(1), \ldots, x(p))'$$
$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}.$$

It will be supposed that the $p \times p$ matrix $\mathbf{S} = \mathbf{X}'\mathbf{X}$ has $m = \min\{n-1, p\}$ unequal non-zero eigenvalues. The 'given' regressors are denoted by

$$\dot{t}(j) = \mathbf{c}_j'\dot{\mathbf{x}}, \qquad j = 1, \ldots, g,$$

and additional constructed regressors by

$$\dot{t}(g+1) = \mathbf{c}_{g+1}'\dot{\mathbf{x}}, \quad \dot{t}(g+2) = \mathbf{c}_{g+2}'\dot{\mathbf{x}}, \ldots . \tag{1}$$

It will be supposed that $\mathbf{c}_1, \ldots, \mathbf{c}_g$ are such that no linear combination of the corresponding given regressors has zero sample variance. However, it will not be assumed that the latter are 'uncorrelated', i.e. that they have mutually zero sample correlation.

3. GENERAL METHOD

We consider a general method with three phases. Firstly, some *construction rule* uses the basic data to determine the sequence (1) of potential additional regressors. Here $\mathbf{c}_{g+1}, \mathbf{c}_{g+2}, \ldots$ are required to be vectors of unit length such that $\dot{t}(g+1)$, $\dot{t}(g+2), \ldots$ have positive sample variances and are uncorrelated both with each other and with each of the given regressors $\dot{t}(1), \ldots, \dot{t}(g)$. (These requirements mean that sequence (1) is necessarily terminating.)

Then, some *stopping rule* uses the basic data to determine the total number ω of regressors actually accepted for use in the regression predictor. Finally, with $\dot{t}(1)$, $\ldots, \dot{t}(\omega)$ now fixed, the value of $\dot{y}$ at a general value would be predicted from the (usually reduced) data

$$\begin{pmatrix} \dot{t}_1(1) & \ldots & \dot{t}_1(\omega) & \dot{y}_1 \\ \vdots & & \vdots & \vdots \\ \dot{t}_n(1) & \ldots & \dot{t}_n(\omega) & \dot{y}_n \end{pmatrix} = (\dot{\mathbf{t}}(1) \ldots \dot{\mathbf{t}}(\omega)\ \dot{\mathbf{y}})$$

by fitting the *OLS prediction formula*

$$\hat{\dot{y}} = \bar{\dot{y}} + b_1 t(1) + \ldots + b_\omega t(\omega). \tag{2}$$

By the zero sample correlation conditions imposed on $\dot{t}(g+1), \ldots, \dot{t}(\omega)$, we have, for $j = g+1, \ldots, \omega$, the simplifying formula

$$b_j = \sum_{i=1}^{n} t_i(j) y_i \Big/ \sum_{i=1}^{n} t_i(j)^2. \tag{3}$$

The possibility $\omega = g$ corresponds to no additional regressors being accepted.

In the next three sections, we formulate OLS, PCR and PLS as special cases of this general method with $g = 0$.

4. ORDINARY LEAST SQUARES

Suppose that we take

$$\mathbf{c}_1 = \frac{\mathbf{S}^- \mathbf{s}}{\|\mathbf{S}^- \mathbf{s}\|} \tag{4}$$

where $\mathbf{S} = \mathbf{X}'\mathbf{X}$, $\mathbf{s} = \mathbf{X}'\mathbf{y}$ and $\mathbf{S}^-$ is either $\mathbf{S}^{-1}$ or some generalized inverse of $\mathbf{S}$ if $\mathbf{S}$ is singular. Suppose also that we stop with $\omega = 1$.

Writing $\hat{\boldsymbol{\beta}}$ for $\mathbf{S}^- \mathbf{s} = (\mathbf{X}'\mathbf{X})^- \mathbf{X}'\mathbf{y}$, we have $t(1) = \mathbf{c}_1'\mathbf{x} = \hat{\boldsymbol{\beta}}'\mathbf{x} / \|\hat{\boldsymbol{\beta}}\|$ and

$$b_1 = \frac{\sum_{i=1}^{n} (\hat{\boldsymbol{\beta}}'\mathbf{x}_i) y_i \|\hat{\boldsymbol{\beta}}\|}{\sum_{i=1}^{n} (\hat{\boldsymbol{\beta}}'\mathbf{x}_i)^2} = \frac{\mathbf{s}'\mathbf{S}^-\mathbf{s}}{\mathbf{s}'\mathbf{S}^-\mathbf{S}\mathbf{S}^-\mathbf{s}} \|\hat{\boldsymbol{\beta}}\| = \|\hat{\boldsymbol{\beta}}\|.$$

Hence, by equations (2) and (3),

$$\hat{y} = \bar{y} + \hat{\boldsymbol{\beta}}'\mathbf{x}. \tag{5}$$

When $\mathbf{S}$ is non-singular, $\hat{\boldsymbol{\beta}}$ is the least squares estimate of $\boldsymbol{\beta}$ in the ordinary linear model $\dot{\mathbf{y}} = \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$. Whether or not $\mathbf{S}$ is singular, equation (5) is just an OLS predictor of $\dot{y}$ at $\mathbf{x}$ using all p variables as regressors. The value of (5) is independent of the choice of generalized inverse of $\mathbf{S}$ only for prediction at $\dot{\mathbf{x}}$ whose corresponding $\mathbf{x}$ lies in $\langle \mathbf{x}_1, \ldots, \mathbf{x}_n \rangle$, the subspace of R^p spanned by $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Ignoring this deficiency, we merely note that OLS is a special case of the general method of Section 3, whatever the values of n and p.

In preparation for Section 7, we also note that the choice of expression (4) for $\mathbf{c}_1$, along with $\omega = 1$, can be presented as a special case of the general method when that is driven by the criterion

$$r_c^2 = \frac{\left(\sum_{i=1}^{n} y_i \mathbf{c}'\mathbf{x}_i \right)^2}{\left(\sum_{i=1}^{n} y_i^2 \right) \sum_{i=1}^{n} (\mathbf{c}'\mathbf{x}_i)^2} = \frac{(\mathbf{c}'\mathbf{s})^2}{\|\mathbf{y}\|^2 \mathbf{c}'\mathbf{S}\mathbf{c}}, \tag{6}$$

the square of the sample correlation coefficient of $\dot{y}$ and

$$\dot{t}[\mathbf{c}] \stackrel{\text{def}}{=} \mathbf{c}'\dot{\mathbf{x}}.$$

For it may be shown, purely algebraically, that $\mathbf{c}_1$, given by equation (4), maximizes r_c^2. Alternatively, observe that the vector $(\mathbf{x}_1'\mathbf{c}, \ldots, \mathbf{x}_n'\mathbf{c})'$ can be interpreted as a scalar multiple of a generic fitted vector for y, with r_c^2 as the square of the cosine of the angle between fitted and observed y; the geometry of least squares then informs us that the OLS choice $\mathbf{c} = \mathbf{c}_1$, given by equation (4), maximizes r_c^2 among vectors $\mathbf{c}$ of unit length for which $\mathbf{t}[\mathbf{c}]'\mathbf{t}(1) = 0$, where $t_i[\mathbf{c}] = \mathbf{c}'\mathbf{x}_i$. But then

$$\mathbf{c}'\mathbf{s} = \mathbf{c}'\mathbf{X}'\mathbf{y} = \mathbf{c}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} \propto \mathbf{t}[\mathbf{c}]'\mathbf{t}(1) = 0,$$

which implies $r_c^2 = 0$, so that the sequential construction terminates with just $\dot{t}(1)$.

5. PRINCIPAL COMPONENTS REGRESSION

In marked contrast with the least squares method, $\mathbf{c}_1, \mathbf{c}_2, \ldots$ are now constructed without any reference to $\dot{\mathbf{y}}$, by use of the criterion

$$S_{\mathbf{c}} = \sum_{i=1}^{n} (\mathbf{c}'\mathbf{x}_i)^2 = \mathbf{c}'\mathbf{S}\mathbf{c}, \tag{7}$$

the 'sum of squares' of the orthogonal projections of $\dot{\mathbf{x}}_1, \ldots, \dot{\mathbf{x}}_n$ on $\langle \mathbf{c} \rangle$. Thus $\mathbf{c}_1$, maximizing $S_{\mathbf{c}}$, is the normalized eigenvector of $\mathbf{S}$ with largest eigenvalue. To follow the general method of Section 3, we have to choose $\mathbf{c}_2$ from $\mathbf{c}$ with $\mathbf{t}[\mathbf{c}]'\mathbf{t}(1) = \mathbf{c}'\mathbf{S}\mathbf{c}_1 = 0$. With $\mathbf{c}_1$ as it is, this is the same as having to choose $\mathbf{c}_2$ from $\mathbf{c}$ subject to the orthogonality condition $\mathbf{c}'\mathbf{c}_1 = 0$ conventionally imposed in principal components analysis. So $\mathbf{c}_2$ is the normalized eigenvector of $\mathbf{S}$ with the next smaller eigenvalue—and so on, until the number of $\mathbf{c}_j$s constructed reaches m, the rank of $\mathbf{S}$, beyond which point $S_{\mathbf{c}}$ would be zero. As for a stopping rule, we could use cross-validation (Wold, 1978).

6. PARTIAL LEAST SQUARES

Suppose that $\mathbf{c}_1$ is constructed to maximize the squared sample covariance, proportional to

$$\left(\sum_{i=1}^{n} y_i \mathbf{c}'\mathbf{x}_i \right)^2 = (\mathbf{c}'\mathbf{s})^2. \tag{8}$$

This immediately gives what we may call the first 'canonical covariance' variable with

$$\mathbf{c}_1 = \mathbf{s}/\|\mathbf{s}\|. \tag{9}$$

The second canonical covariance variable, $\mathbf{c}_2$, then has to maximize expression (8) for unit length $\mathbf{c}$ such that $\mathbf{c}'\mathbf{S}\mathbf{c}_1 = 0$, i.e. $\mathbf{c}$ is $\mathbf{S}$ orthogonal to $\mathbf{c}_1$. It turns out, provided that $\mathbf{s}$ is not an eigenvector of $\mathbf{S}$, that

$$\mathbf{c}_2 \propto \mathbf{s} - \left(\frac{\mathbf{s}'\mathbf{S}\mathbf{s}}{\mathbf{s}'\mathbf{S}^2\mathbf{s}} \right) \mathbf{S}\mathbf{s}. \tag{10}$$

This is because

(a) the right-hand side of equation (10) is indeed **S** orthogonal to **s**, and therefore to $\mathbf{c}_1$, and
(b) for **c** satisfying $\mathbf{c}'\mathbf{S}\mathbf{c}_1 = 0$, i.e. $\mathbf{c}'\mathbf{S}\mathbf{s} = 0$,

$$\mathbf{c}'\mathbf{s} = \mathbf{c}'\left(\mathbf{s} - \left(\frac{\mathbf{s}'\mathbf{S}\mathbf{s}}{\mathbf{s}'\mathbf{S}^2\mathbf{s}}\right)\mathbf{S}\mathbf{s}\right) .$$

Continuing thus, we could derive the remaining canonical covariance variables $\mathbf{c}_3$, $\mathbf{c}_4$, . . ., and then specify a stopping rule to give $\mathbf{c}_1, \ldots, \mathbf{c}_\omega$. Since their explicit form is not needed for the present purposes, we shall simply show that use of these $\mathbf{c}_1, \ldots, \mathbf{c}_\omega$ for predictor (2) is equivalent to PLS prediction. Using an extension of the argument in (a) and (b), Appendix A proves inductively, provided that **s**, **Ss**, . . ., $\mathbf{S}^{\omega-1}\mathbf{s}$ are linearly independent and $1 \leqslant \omega \leqslant m$, that

$$\langle \mathbf{c}_1, \ldots, \mathbf{c}_\omega \rangle = \langle \mathbf{s}, \mathbf{S}\mathbf{s}, \ldots, \mathbf{S}^{\omega-1}\mathbf{s} \rangle. \tag{11}$$

Looking for the precise connection between identity (11) and PLS, we are faced with a rich variety of formulations of the PLS method (Wold, 1984). The formulation that here serves best is the non-algorithmic version of Helland (1988), whose proposition 3.1 immediately combines with identity (11) to establish the claim of equivalence.

7. CONTINUUM REGRESSION

The methods of Sections 4–6 differ in just one respect—the criterion maximized at each stage. To emphasize this point, the regressors for OLS, PCR and PLS may be referred to as 'canonical correlation', 'canonical variance' and 'canonical covariance' variables respectively. A generalized criterion that encompasses all three methods is

$$T = \|\mathbf{y}\|^2 r_{\mathbf{c}}^2 S_{\mathbf{c}}^{\alpha/(1-\alpha)} = (\mathbf{c}'\mathbf{s})^2 (\mathbf{c}'\mathbf{S}\mathbf{c})^{\alpha/(1-\alpha)-1} \tag{12}$$

where α takes some value in the continuum $0 \leqslant \alpha \leqslant 1$. The specializations are $\alpha = 0$ (OLS), $\alpha = \frac{1}{2}$ (PLS) and $\alpha = 1$ (PCR). A procedure in the general class of Section 3 with construction rule driven by T for some α, $0 \leqslant \alpha \leqslant 1$, is a *continuum regression*—only the stopping rule that determines ω remains to be specified.

In the next section, we develop a reasonably efficient algorithm for constructing the corresponding $\mathbf{c}_{g+1}$, $\mathbf{c}_{g+2}$, . . . for $0 \leqslant \alpha \leqslant 1$. Completeness of definition of the criterion requires that, for $\alpha < \frac{1}{2}$, T be taken to be zero when $\mathbf{c}'\mathbf{s} = 0$ and $\mathbf{c}'\mathbf{S}\mathbf{c} = 0$. (Since **s** is in the range of **S**, $\mathbf{c}'\mathbf{S}\mathbf{c} = 0 \Rightarrow \mathbf{c}'\mathbf{s} = 0$.) The construction sequence comes to a stop if, at any stage, T is identically zero for every **c** satisfying the **S**-orthogonality conditions corresponding to the imposed zero sample correlations. Appendix B shows that, for $\alpha \neq 0$, $\mathbf{c}_{g+1}$, $\mathbf{c}_{g+2}$, . . . must lie in the range of **S**, and that their total number is at most $m - g$ where $m = \min\{n-1, p\} = \operatorname{rank} \mathbf{S}$ (see Section 2).

Also, for $\alpha \neq 0$, the constructed regressor variables $\dot{t}(g+1)$, $\dot{t}(g+2)$, . . . are invariant only under *orthogonal* transformation of **x**. In particular, results are influenced by the choice of scales of the explanatory variables $\dot{x}(1), \ldots, \dot{x}(p)$. Our current practice is democratically to standardize these variables to unit sample standard deviation (with compensating adjustment of $\mathbf{c}_1, \ldots, \mathbf{c}_g$ to keep the given

variables unchanged) unless there are good reasons for using the given scales of measurement.

8. CONSTRUCTION ALGEBRA

In this and the next section, the algebra is simpler when $\alpha/(1-\alpha)$ in equation (12) is replaced by γ, equivalent to the substitution $\alpha=\gamma/(\gamma+1)$. The specializations are then $\gamma=0$, 1, ∞ for OLS, PLS and PCR respectively. Let $\mathbf{v}_1, \ldots, \mathbf{v}_m$ be the orthonormalized eigenvectors of $\mathbf{S}$ with increasing non-zero eigenvalues $0<e_1<\ldots<e_m$. (For $m=n-1<p$, the $\mathbf{v}_i$ and $\mathbf{e}_i$ might be computed by eigenanalysis of $\mathbf{XX}'$ rather than of $\mathbf{S}$.) Suppose that we are at the stage where $\mathbf{c}_{g+1}, \ldots, \mathbf{c}_k$, with $g\leqslant k\leqslant m-2$, have been constructed, and we want to determine $\mathbf{c}_{k+1}$ maximizing T, subject to the side-conditions $\|\mathbf{c}_{k+1}\|=1$ and

$$\mathbf{c}_j'\mathbf{S}\mathbf{c}_{k+1}=0, \qquad j=1, \ldots, k. \tag{13}$$

By the result of Appendix B, $\mathbf{c}_{k+1}$ will lie in the range of $\mathbf{S}$. So we may write

$$\mathbf{c}_{k+1}=z_1\mathbf{v}_1+\ldots+z_m\mathbf{v}_m, \tag{14}$$

whence T at $\mathbf{c}=\mathbf{c}_{k+1}$ may be expressed as

$$(d_1z_1+\ldots+d_mz_m)^2(e_1z_1^2+\ldots+e_mz_m^2)^{\gamma-1} \tag{15}$$

where $d_i=\mathbf{s}'\mathbf{v}_i$. The side-conditions are, from $\|\mathbf{c}_{k+1}\|=1$,

$$z_1^2+\ldots+z_m^2=1 \tag{16}$$

and, from equation (13),

$$a_{1j}z_1+\ldots+a_{mj}z_m=0, \qquad j=1, \ldots, k, \tag{17}$$

where $a_{ij}=e_i\mathbf{c}_j'\mathbf{v}_i$. The maximizing $z_1, \ldots, z_m$ will be a solution of the Lagrange multiplier equations

$$\frac{d_i}{\tau}+(\gamma-1)\frac{e_iz_i}{\rho}-\lambda_0z_i-\lambda_1a_{i1}-\ldots-\lambda_ka_{ik}=0 \tag{18}$$

where $\tau=d_1z_1+\ldots+d_mz_m=\mathbf{d}'\mathbf{z}$ and $\rho=e_1z_1^2+\ldots+e_mz_m^2$. Multiplying equation (18) by z_i and adding over i gives $\lambda_0=\gamma$. Writing $\sigma_i=\rho\lambda_i$,

$$\mathbf{D}=\operatorname{diag}(\gamma\rho+(1-\gamma)e_1, \ldots, \gamma\rho+(1-\gamma)e_m) \tag{19}$$

and $\mathbf{A}=(a_{ij})$, equations (17) and (18) for $\mathbf{z}$ and $\boldsymbol{\sigma}=(\sigma_1, \ldots, \sigma_k)$ then become

$$\begin{pmatrix}\mathbf{D} & \mathbf{A}\\ \mathbf{A}' & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{z}\\ \boldsymbol{\sigma}\end{pmatrix}=\rho\begin{pmatrix}\mathbf{d}\\ \mathbf{0}\end{pmatrix}\Big/\tau. \tag{20}$$

Since $\mathbf{c}_1, \ldots, \mathbf{c}_k$ are linearly independent and $e_i\neq 0$, $i=1, \ldots, m$, the $m\times k$ matrix $\mathbf{A}$ is of full rank k. By condition (16) and the standard formula for inverting a partitioned matrix, we obtain

$$\mathbf{z}=\mathbf{M}\mathbf{d}/\|\mathbf{M}\mathbf{d}\| \tag{21}$$

where

$$\mathbf{M}=\mathbf{D}^{-1}-\mathbf{D}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{D}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{D}^{-1} \tag{22}$$

for γ and ρ such that $\mathbf{D}$ is invertible. In equation (22), $\mathbf{A}$ is determined by $\mathbf{c}_1, \ldots, \mathbf{c}_k$, which are supposed already computed. For $\alpha=0$ or $\alpha=\frac{1}{2}$ ($\gamma=0$ or $\gamma=1$), $\mathbf{z}$ is then determined by equation (21). Otherwise the only unknown is the scalar ρ in $\mathbf{D}$. Writing $\mathbf{z}=\mathbf{z}(\rho)$, ρ will be a solution of

$$e_1 z_1(\rho)^2 + \ldots + e_m z_m(\rho)^2 = \rho. \tag{23}$$

Since $z_1(\rho)^2 + \ldots + z_m(\rho)^2 = 1$, we have $e_1 \leqslant \rho \leqslant e_m$ as an interval to which any search for roots of equation (23) may be confined. Since $\mathbf{z}(\rho)$ satisfies the required side-conditions for any value of ρ, it follows that the optimal ρ is the solution of equation (23) whose associated value of T is a maximum. In practice, we have found that a satisfactory approximation to the optimum is obtained by simply using the value of ρ from the finite set $\{e_i+\theta(e_{i+1}-e_i)$: $\theta=0, N^{-1}, 2N^{-1}, \ldots, 1$; $i=1, \ldots, m-1\}$ which maximizes $T(\mathbf{z}(\rho))$, for sufficiently large N. The use of a fixed grid, the same for all stages, admits a time-saving recurrence relation for the calculation of $\mathbf{M}$ for any chosen value of ρ (see Appendix C). Such a relation arises because the only change in equation (22) as we go from stage k to stage $k+1$ is that $\mathbf{A}$ is augmented by an additional column $(e_1 z_1, \ldots, e_m z_m)'$, where $\mathbf{z}=\mathbf{z}(\rho)$ is the output of the kth stage.

The special case $\alpha=\frac{1}{2}$ ($\gamma=1$) should confirm the analysis of Appendix A. With $\gamma=1$, $\mathbf{D} \propto \mathbf{I}$ and we obtain

$$\mathbf{z} \propto (\mathbf{I}-\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}')\mathbf{d}$$

which is the orthogonal projection, in eigenvector co-ordinates, of $\mathbf{s}$ on to the subspace of the range of $\mathbf{S}$ that is orthogonal to $\langle \mathbf{Sc}_1, \ldots, \mathbf{Sc}_k \rangle = \mathbf{S}\mathscr{S}_k$. This agrees with Fig. 5 of Appendix A.

We should also be able to derive PCR as $\alpha \to 1$ ($\gamma \to \infty$). However, the analysis for this is rather technical, and we shall leave this limit to be demonstrated numerically.

For $\alpha=0$ ($\gamma=0$), the construction still holds good provided that we stop after one stage. For $g=0$ and $k=0$, we have

$$\mathbf{z} \propto (d_1/e_1, \ldots, d_m/e_m)'.$$

Translating from eigenvector co-ordinates, this is just $\mathbf{S}^{-1}\mathbf{s}$ for $m=p$ (as might have been expected in the light of Section 4) and $\mathbf{S}^{+}\mathbf{s}$ for $m<p$, where $\mathbf{S}^{+}$ is the Moore–Penrose generalized inverse. In other words, in one stage we obtain the OLS predictor in which, for the case $m<p$, the choice of generalized inverse is resolved in favour of the Moore–Penrose inverse.

For $\alpha=0$ and $g>0$, the predictor obtained for $\omega=g+1$ would be the same as for $\alpha=0$, $g=0$, $\omega=1$, provided that $\mathbf{c}_1, \ldots, \mathbf{c}_g$ are in the range of $\mathbf{X}'$ (or $\mathbf{S}$) as they must be if $m=p$. Otherwise, for $m=n-1<p$, Appendix D shows that the predictor is an OLS predictor whose coefficients depend on the choice of given regressors.

As for the possible stage $k=m-1$ ($\omega=m$) with $\alpha \neq 0$, there is only one $\mathbf{z}$ (up to multiplication by ± 1) that satisfies the side-conditions applicable at that stage: the value of α is therefore not involved. This $\mathbf{z}$ would complete the spanning of the range of $\mathbf{S}$. For $m=p$, the associated predictor would just be the OLS predictor obtained at $\alpha=0$, and we may therefore exclude the case $\omega=m$ from our computations. For $m=n-1<p$, it would be an OLS predictor given by a generalized inverse of $\mathbf{S}$ dependent on the given regressors.

9. CROSS-VALIDATORY ALGEBRA

To make a cross-validatory choice (Stone, 1974) of α and ω means that, for a sufficiently large set of values of (α, ω), we carry out the following operations:

(a) for each $*=1, \ldots, n$ in turn, the datum $\dot{\mathbf{x}}_*, \dot{y}_*$ is left out of the calculation of predictor (2);
(b) this 'leave-one-out' predictor is used to calculate $\hat{y}$ at $\dot{\mathbf{x}}=\dot{\mathbf{x}}_*$ given the prediction $\hat{y}_{\setminus *}$, say, of $\dot{y}_*$;
(c) calculation of a cross-validatory assessment

$$C_{\alpha,\omega}=\frac{1}{n}\sum_{*=1}^{n} L(\dot{y}_*, \hat{y}_{\setminus *}) \tag{24}$$

of the performance of the predictor, corresponding to the choice (α, ω), of future $\dot{y}$ values at the points $\dot{\mathbf{x}}_1, \ldots, \dot{\mathbf{x}}_n$. (We shall use the quadratic loss $L(u, v)=(u-v)^2$ in our applications.)

Finally, a value of (α, ω) that gives the smallest, or nearly the smallest, value of $C_{\alpha,\omega}$ is found: this value is a *cross-validatory choice*, $(\alpha^\dagger, \omega^\dagger)$, say. For graphical exposition, we shall use the *cross-validatory index* $I_{\alpha,\omega}$, defined by

$$I_{\alpha,\omega}=1-C_{\alpha,\omega}/C_{-,g}. \tag{25}$$

The denominator here is the value of $C_{\alpha,\omega}$ for the predictor based on the g given regressors alone, which is independent of α. The index I cannot exceed unity but may take negative values.

Clearly, equation (24) could be found by applying the computational method of Section 8 to each of the n data sets in which one datum is omitted. But this would need n eigenanalyses which would be time consuming for large m. The following algebra applies only if $\alpha>0$ and shows that the single eigenanalysis of $\mathbf{S}$ (or $\mathbf{XX}'$) for the complete data is enough to do the job. When the cross-validatory modification of the recurrence relation in Appendix C associated with our grid search is used, there is a negligible increase in computation compared with a single run of the method described in Section 8. This enhances the value of our single eigenanalysis approach, particularly as m increases.

By the standard missing datum formulae, with $(\dot{\mathbf{x}}_*, \dot{y}_*)$ left out, $\mathbf{s}$ and $\mathbf{S}$ of Section 8 change to $\mathbf{s}-\nu y_*\mathbf{x}_*$ and $\mathbf{S}^*=\mathbf{S}-\nu\mathbf{x}_*\mathbf{x}_*'$ where $\nu=n/(n-1)$. The T to be maximized is then

$$(\mathbf{c}'\mathbf{s}-\nu y_*\mathbf{c}'\mathbf{x}_*)^2\{\mathbf{c}'\mathbf{S}\mathbf{c}-\nu(\mathbf{c}'\mathbf{x}_*)^2\}^{\gamma-1}. \tag{26}$$

Noting that the range of $\mathbf{S}^*$ lies in $\langle \mathbf{x}_1, \ldots, \mathbf{x}_n\rangle$ (which equals the range of $\mathbf{S}$), it follows that the new $\mathbf{c}$ vectors, $\mathbf{c}_{g+1}, \mathbf{c}_{g+2}, \ldots$, necessarily in the range of $\mathbf{S}^*$, may still be written as linear combinations of $\mathbf{v}_1, \ldots, \mathbf{v}_m$, i.e. for the construction of $\mathbf{c}_{k+1}$, $k\geqslant g$, we write $\mathbf{c}_{k+1}=z_1\mathbf{v}_1+\ldots+z_m\mathbf{v}_m$. Since $\mathbf{x}_*=f_1^*\mathbf{v}_1+\ldots+f_m^*\mathbf{v}_m$ where $f_i^*=\mathbf{x}_*'\mathbf{v}_i$, the value of expression (26) at $\mathbf{c}=\mathbf{c}_{k+1}$ equals

$$(\mathbf{d}'\mathbf{z}-\nu y_*\mathbf{f}^{*\prime}\mathbf{z})^2\{e_1z_1^2+\ldots+e_mz_m^2-\nu(\mathbf{f}^{*\prime}\mathbf{z})^2\}^{\gamma-1}. \tag{27}$$

The given $\mathbf{c}_1, \ldots, \mathbf{c}_g$ are unaffected by the omission of $\dot{\mathbf{x}}_*, \dot{y}_*$, and the zero-correlation side-conditions on $\mathbf{z}$ are $\mathbf{c}_j'\mathbf{S}^*\mathbf{c}_{k+1}=0, j=1, \ldots, k$, which in eigenvector co-ordinates become

$$a_{1j}^{*}z_1 + \ \ldots\ + a_{mj}^{*}z_m = 0, \qquad j = 1, \ \ldots, \ k, \tag{28}$$

where

$$a_{ij}^{*} = e_i \mathbf{c}_j' \mathbf{v}_i - \nu f_i^{*} \sum_{r=1}^{m} f_r^{*} \mathbf{c}_j' \mathbf{v}_r. \tag{29}$$

The Lagrangian minimization of expression (27) then gives, in place of equation (21),

$$\mathbf{z} = \mathbf{M}^{*}\mathbf{d}^{*} / \| \mathbf{M}^{*}\mathbf{d}^{*} \|, \tag{30}$$

where

$$\mathbf{d}^{*} = \mathbf{d} - \nu y_{*} \mathbf{f}^{*} \tag{31}$$

$$\mathbf{M}^{*} = \mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{A}^{*}(\mathbf{A}^{*\prime}\mathbf{Q}^{-1}\mathbf{A}^{*})^{-1}\mathbf{A}^{*\prime}\mathbf{Q}^{-1} \tag{32}$$

$$\mathbf{Q} = \mathbf{D}^{*} - (1-\gamma)\nu \mathbf{f}^{*}\mathbf{f}^{*\prime} \tag{33}$$

$$\mathbf{D}^{*} = \operatorname{diag}\{\gamma\rho^{*} + (1-\gamma)e_1, \ \ldots, \ \gamma\rho^{*} + (1-\gamma)e_m\} \tag{34}$$

$$\rho^{*} = e_1 z_1^2 + \ \ldots\ + e_m z_m^2 - \nu(\mathbf{f}^{*\prime}\mathbf{z})^2 \tag{35}$$

$$\mathbf{A}^{*} = (a_{ij}^{*}), \qquad m \times k. \tag{36}$$

For the inverse of $\mathbf{Q}$, we have, by a standard formula,

$$\mathbf{Q}^{-1} = \mathbf{D}^{*-1} + \frac{(1-\gamma)\nu \mathbf{D}^{*-1}\mathbf{f}^{*}\mathbf{f}^{*\prime}\mathbf{D}^{*-1}}{1 - (1-\gamma)\nu \mathbf{f}^{*\prime}\mathbf{D}^{*-1}\mathbf{f}^{*}}. \tag{37}$$

The rest of the computation of $\mathbf{c}_{k+1}$ follows the lines of Section 8. The only change relates to the interval in which ρ^{*} is known to lie. Appendix E shows that, for cases with $n > p+1$ and rank $\mathbf{S}^{*} = \text{rank } \mathbf{S}$ $(= m = p)$, we have to widen the interval to

$$\frac{e_1}{1 + u_2 e_1/(1-u_1)} \leqslant \rho^{*} \leqslant e_m, \tag{38}$$

where

$$u_2 = \nu \sum_{i=1}^{n} f_i^{*2}/e_i^2$$

and

$$u_1 = \nu \sum_{i=1}^{n} f_i^{*2}/e_i = 1 - |\mathbf{S}^{*}|/|\mathbf{S}| < 1$$

for $\mathbf{x}_{*} \neq \mathbf{0}$. Appendix E shows that, in all other cases, we may continue to use $e_1 \leqslant \rho^{*} \leqslant e_m$. Appendix C gives a recurrence relation for the calculation of $\mathbf{M}^{*}$, based on that used for $\mathbf{M}$ in Section 8.

For the computation of $\hat{y}_{\backslash *}$ in equation (24), first observe that this is the value for $k = \omega$ of $\hat{y}_{\backslash *}^{(k)}$, defined as the prediction of $\dot{y}_{*}$, using predictor (2) with $\omega = k$ at $\dot{\mathbf{x}} = \dot{\mathbf{x}}_{*}$ when $(\dot{\mathbf{x}}_{*}, \ \dot{y}_{*})$ is omitted from the basic data. Then

$$\hat{y}_{\backslash *} = \hat{y}_{\backslash *}^{(g)} + \sum_{k=g}^{\omega-1} (\hat{y}_{\backslash *}^{(k+1)} - \hat{y}_{\backslash *}^{(k)}). \tag{39}$$

By a well-known result for an omitted datum, we have

$$\hat{\dot{y}}^{(g)}_{\setminus *} = (\hat{\dot{y}}^{(g)}_{*} - h_* \dot{y}_*)/(1 - h_*) \tag{40}$$

where $\hat{\dot{y}}^{(g)}_{*}$ is the value of formula (2) at $\dot{\mathbf{x}} = \dot{\mathbf{x}}_*$ with $\omega = g$, i.e. the fitted value of $\dot{y}_*$ in least squares regression of $\dot{y}$ on just the given regressors $\dot{\imath}(1), \ldots, \dot{\imath}(g)$, while h_* is the corresponding diagonal element of the associated projection, or 'hat', matrix. The increment $\hat{\dot{y}}^{(k+1)}_{\setminus *} - \hat{\dot{y}}^{(k)}_{\setminus *}$ is, by the leave-$*$-out modification of b_{k+1} in formula (2),

$$\frac{\sum\limits_{i \neq *} \mathbf{c}'_{k+1}(\dot{\mathbf{x}}_i - \bar{\dot{\mathbf{x}}}_{\setminus *})\dot{y}_i}{\sum\limits_{i \neq *} \{\mathbf{c}'_{k+1}(\dot{\mathbf{x}}_i - \bar{\dot{\mathbf{x}}}_{\setminus *})\}^2} \mathbf{c}'_{k+1}(\dot{\mathbf{x}}_* - \bar{\dot{\mathbf{x}}}_{\setminus *}) \tag{41}$$

where $\bar{\dot{\mathbf{x}}}_{\setminus *} = \sum_{r \neq *} \dot{\mathbf{x}}_r/(n-1)$. In 'undotted' terms, expression (41) is equal to

$$\frac{\nu \mathbf{c}'_{k+1}\mathbf{x}_* \sum\limits_{i \neq *} \{\mathbf{c}'_{k+1}\mathbf{x}_i + \mathbf{c}'_{k+1}\mathbf{x}_*/(n-1)\}y_i}{\sum\limits_{i \neq *} \{\mathbf{c}'_{k+1}\mathbf{x}_i + \mathbf{c}'_{k+1}\mathbf{x}_*/(n-1)\}^2} = \frac{\nu \mathbf{f}^{*\prime}\mathbf{z} \sum\limits_{i \neq *} \{\mathbf{f}^{i\prime}\mathbf{z} + \mathbf{f}^{*\prime}\mathbf{z}/(n-1)\}y_i}{\sum\limits_{i \neq *} \{\mathbf{f}^{i\prime}\mathbf{z} + \mathbf{f}^{*\prime}\mathbf{z}/(n-1)\}^2}. \tag{42}$$

Formulae (40) and (42) are suitable for computation, using the $\mathbf{z}$ in hand after stage k.

For $\alpha = 0$ ($\gamma = 0$), this cross-validatory algebra breaks down if $\mathbf{Q}$ in equation (33) is singular for any omitted datum, equivalent to singularity of $\mathbf{S}^*$. For $n \leqslant p+1$, this singularity is omnipresent and therefore $I_{0,g+1}$ is not calculable using this approach. For this case, we might transfer attention from the OLS predictor at $\alpha = 0$ to the OLS predictor given by $\alpha = \frac{1}{2}$, $\omega = m - 1$ for which the cross-validatory index *is* calculable. (The value $\frac{1}{2}$ for α is chosen because then $\gamma = 1$, $\mathbf{z}$ in equation (30) is free of ρ^* and the calculations are non-iterative.)

10. PERFORMANCE AND COMPARISONS

Throughout, we use a grid (Section 8) with $N = 5$ or $N = 10$ subintervals between adjacent eigenvalues, and $I_{\alpha,\,\omega}$ is defined with quadratic loss.

10.1. *Example 1: Cement Heat Evolution Data*

The experimental data were produced and analysed by least squares by Wood *et al.* (1932) and have been extensively reanalysed. For ease of comparison, we shall use the data selected by Hald (1952) and reused by Draper and Smith (1981). We shall follow these researchers in taking a purely formal approach to this data set and ignore the interesting scientific reanalysis of the original data by Daniel and Wood (1971). In our notation, $n = 13$, $p = 4$, $\dot{y}$ is the heat evolved in calories per gram from cement samples in the first 180 days after addition of water and $\dot{x}(1), \ldots, \dot{x}(4)$ are rounded estimated percentages by weight of the four compounds that make up most

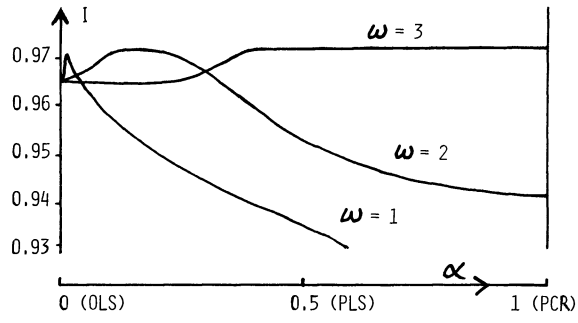

Fig. 1. I plot for the standardized cement heat data

of the cement ($\dot{x}(1) + \ldots + \dot{x}(4)$ varies from 95 to 99). The four $\dot{x}$ variables are highly collinear; the condition number of the matrix $\mathbf{S}$, e_m/e_1, is 1379.

Using 10 subintervals and standardized explanatory variables, Fig. 1 shows the complete I plot, i.e. the value of $I_{\alpha,\omega}$ for any value of (α, ω). With $p=4$, $(\alpha, 4)$ is equivalent to (0, 1). The absolute maximum of I is 0.9717 given by $\omega=3$ with any $\alpha>0.41$, which includes the optimal PLS and PCR choices. The relative maxima were $I=0.9713$ at $\alpha=0.12$, $\omega=2$, and $I=0.9708$ at $\alpha=0.006$, $\omega=1$. A notable feature of the output (recalling the ridge trace method with an ill-conditioned design matrix) is the steep rise in I between $\alpha=0$ and $\alpha=0.006$, for $\omega=1$. For any value of (α, ω), we may calculate the associated predictor (2) in terms of the unstandardized $\dot{x}(1), \ldots, \dot{x}(p)$. The comparisons in Table 1 are of interest.

The predictors for the values of (α, ω) with I at or close to the maximum are seen to be very similar, and different from the somewhat less optimal OLS predictor. The predictor for the drastically non-optimal $(\frac{1}{2}, 1)$, equivalent to stopping PLS after the first stage, is also reassuringly quite different. Some further comparisons are of interest, taking us outside the family of CRs. Draper and Smith (1981) used the data to illustrate a variety of variable-selection techniques and modifications of OLS. They showed that just the two variables $\dot{x}(1)$ and $\dot{x}(2)$ are selected by application of any one of the three techniques 'Mallows's C_p', 'backward elimination with 10% critical region' and 'stepwise regression with 10% critical region'. The associated least squares predictor was

$$52.6 + 1.47\dot{x}(1) + 0.66\dot{x}(2) \tag{43}$$

TABLE 1

α	ω	I	*Predictor coefficient* $\dot{x}(1)$	$\dot{x}(2)$	$\dot{x}(3)$	$\dot{x}(4)$
0	1	0.965	1.55	0.51	0.10	−0.14
0.006	1	0.9708	1.31	0.30	−0.14	−0.35
0.12	2	0.9713	1.33	0.29	−0.13	−0.36
>0.41	3	0.9717	1.31	0.27	−0.14	−0.38
0.5	1	0.94	0.84	0.35	−0.56	−0.33

which is closer to OLS with all four variables than to the optimal CR predictor (for $\alpha > 0.41$, $\omega = 3$) given by

$$85.4 + 1.31\dot{x}(1) + 0.27\dot{x}(2) - 0.14\dot{x}(3) - 0.38\dot{x}(4). \tag{44}$$

By contrast with the output of the variable-selection methods, ridge regression with a trace parameter 0.013 delivered the predictor

$$83.4 + 1.30\dot{x}(1) + 0.30\dot{x}(2) - 0.14\dot{x}(3) - 0.35\dot{x}(4), \tag{45}$$

while PCR gave

$$85.9 + 1.32\dot{x}(1) + 0.27\dot{x}(2) - 0.15\dot{x}(3) - 0.38\dot{x}(4). \tag{46}$$

Two principal components, the first and the third, were selected for predictor (46) by the default method in BMDP4R (Dixon, 1983). Predictor (46) happens to be close to that given by use of all the first three components. This accounts for the closeness of expressions (46) and (44). Fig. 1 suggests that the conventional two-component PCR predictor would be quite different: given by our method with $\alpha = 1$ and $\omega = 2$, it is

$$89.0 + 0.79\dot{x}(1) + 0.36\dot{x}(2) - 0.60\dot{x}(3) - 0.33\dot{x}(4).$$

10.2. *Example 2: Road Accident Data*

For a thoroughly soft application of CR, we turn to Table 8.1 of Weisberg (1980), which gives the accident rates $\dot{y}$ for a variety of stretches of road in Minnesota during 1973. Selecting for analysis just the data for 'minor arterial highways', and using only the first nine potential explanatory variables, gives $n = 13$ and $p = 9$.

With standardized variables, the condition number is 433, somewhat less than that for example 1, but the predictive value uncovered is much less. Fig. 2 gives a partial I plot for 10 subintervals for $1 \leqslant \omega \leqslant 3$; higher values of ω give suboptimal values of I.

OLS performs so badly, with erratic one-out predictions giving an I value of -12, that it does not even appear on the figure. Moving away from $\alpha = 0$, as far as $\alpha = \frac{1}{2}$ (PLS) with $\omega = 1$, is sufficient to ensure reasonable performance. However, PCR with one component is the clear winner in this example with $I = 0.65$.

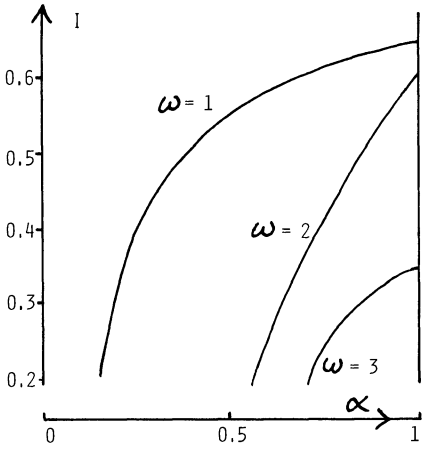

Fig. 2. I plot for the standardized accident data

The associated predictor has coefficients that are almost uniformly nondescript; with the $\dot{y}$ variable also standardized, the coefficients are, in correlation form,

$$-0.11,\ 0.13,\ -0.12,\ 0.12,\ 0.05,\ -0.06,\ 0.14,\ 0.13,\ 0.13.$$

It is doubtful whether much meaning may be extracted from such an outcome.

10.3. *Example 3: Near Infra-red Calibration for Protein*

We shall put the first 12 items in Table 1 of Fearn (1983) through our new mill, without pursuing all the further questions raised for these data by Farebrother (1984), Hoerl *et al.* (1985) and Næs *et al.* (1986). The $\dot{y}$ variable is protein percentage and, with $p=6$, the tabulated explanatory variables $L_1, \ldots, L_6$ are log(1/reflectance) values at six wavelengths. If we were to take $\dot{x}(j)=L_j$, $j=1, \ldots, 6$, unstandardized, we would find $\alpha^\dagger=0.6$, $\omega^\dagger=5$, with $I=0.947$, which compares with $I=0.939$ for OLS. If we were to take $\dot{x}(6)=\bar{L}=(L_1+\ \ldots\ +L_6)/6$ and $\dot{x}(j)=L_j-\bar{L}$, $j=1, \ldots, 5$, we would find $\alpha^\dagger=0.3$, $\omega^\dagger=4$, with $I=0.958$. However, both $\dot{x}(5)$ and $\dot{x}(6)$ then have roughly twice the sample standard deviation of $\dot{x}(1), \ldots, \dot{x}(4)$: with standardization of all these variables, CR gives $\alpha^\dagger=0.35$, $\omega^\dagger=2$, with $I=0.960$. Fig. 3 shows the broader picture of associated I values for 10 subintervals. The coefficients of $L_1, \ldots, L_6$ in the predictors generated by OLS and the three choices of variables for CR are shown in Table 2. The differences illustrate the effect on the CR technique of linear transformation of explanatory variables and of standardization.

The adopted CR predictor, namely (d) with $\alpha^\dagger=0.35$, $\omega^\dagger=2$, has a calibration $(\text{PRESS})^{1/2}$ of 0.29, compared with 0.36 for OLS. When these two predictors are compared on the remaining 12 items of Table 1 of Fearn (1983) and on the 26 items in Fearn's validation sample, we find the results in Table 3. No attempt will be made

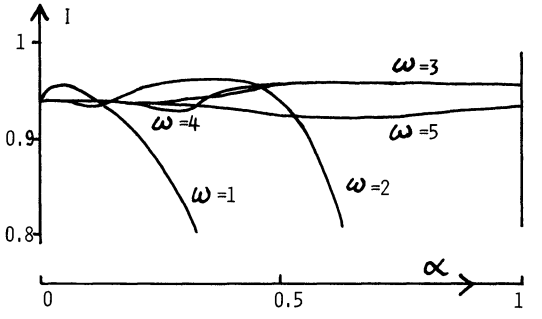

Fig. 3. I plot for standardized $\bar{L}$, $L_1-\bar{L}, \ldots, L_5-\bar{L}$

TABLE 2

	Predictor coefficients					
	L_1	L_2	L_3	L_4	L_5	L_6
(a) OLS	0.31	0.02	0.01	−0.44	0.02	0.14
(b) CR on $L_1, \ldots, L_6$	0.23	−0.06	0.15	−0.39	0.01	0.11
(c) CR on $L_1-\bar{L}, \ldots, L_5-\bar{L}, \bar{L}$	0.09	0.08	0.13	−0.27	0.01	−0.01
(d) CR on (c), standardized	−0.02	0.09	0.16	−0.21	0.01	−0.01

TABLE 3

Validation set	*Root-mean-square prediction error*	
	CR(0.35, 2)	*OLS*
$n=12$	0.44	0.61
$n=26$	0.45	0.83

here to adjust our analysis of the $n=12$ calibration data in the retrospective light of these larger-than-expected values. If we interpret the results obtained by Fearn (1983) to mean that, even if there is a true linear calibration formula, random error has a standard deviation of about 0.2, then biases with a root-mean-square error of 0.4 would be needed to explain the 0.45 root-mean-square prediction error for our CR(0.35, 2) predictor. By various stratagems, other analysts of both the $n=12$ and the $n=24$ calibration sets have evaded bias of this order (Hoerl *et al.*, 1985; Næs *et al.*, 1986). We simply note that, apart from the reversal of correlations of L_i and $\dot{y}$ ($i=1, \ldots, 6$) between the $n=24$ calibration and the $n=26$ validation set noted by Hoerl *et al.*, there are appreciable differences in $\bar{L}$. Even between the two halves of the $n=24$ set, the mean of $\bar{L}$ jumps from 264 to 299 (two-tailed Mann–Whitney, $P<0.002$) so that we are really asking our optimal CR predictor to be a good extrapolator, from a data set that perhaps does not cover a sufficiently wide range of experimental items.

10.4. *Example 4: Near Infra-red Calibration for Ethanol*

Our first example with $p>n$ uses data associated with Table 3 of Bjørsvik and Martens (1989). For this subset, $n=11$, $p=101$ and $\dot{y}$ is the percentage of ethanol in a mixture of ethanol, methanol and *n*-propanol, while $\dot{x}(1), \ldots, \dot{x}(101)$ constitute the 'spectrum' of log(1/reflectance) values at intervals of 5 nm between the wavelengths 1100 nm and 1600 nm. Some I values, calculated for 10 subintervals and *without* standardization of $\dot{x}(1), \ldots, \dot{x}(101)$, are exhibited in Table 4.

In this example, there is a case for not standardizing the $\{\dot{x}(j)\}$ before use of CR: all the variables are on the same physical scale and the variables with large sample standard deviations may be expected to include those expressing informative differences

TABLE 4

ω	*I values for the following values of* α†*:*							
	0.1	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.75*	*1*
1	0.989	0.95	0.7	0.5	0.0	−0.1	−0.2	−0.3
2	0.997	0.996	0.986	0.97	0.96	0.94	0.90	0.8
3	*0.9980*	0.997	0.996	0.996	0.996	0.996	0.995	0.995
4	0.9980	*0.9981*	*0.9981*	0.9979	0.9977	0.997	0.996	0.996
5	0.9980	0.9980	0.9981	*0.9982*	*0.9983*	*0.9982*	0.9981	0.997
6	0.9980	0.9980	0.9980	0.9980	0.9981	0.9981	*0.9982*	0.9978
7	0.9980	0.9980	0.9980	0.9980	0.9980	0.9980	0.9981	*0.9980*

†The maximum value in each column is in italics.

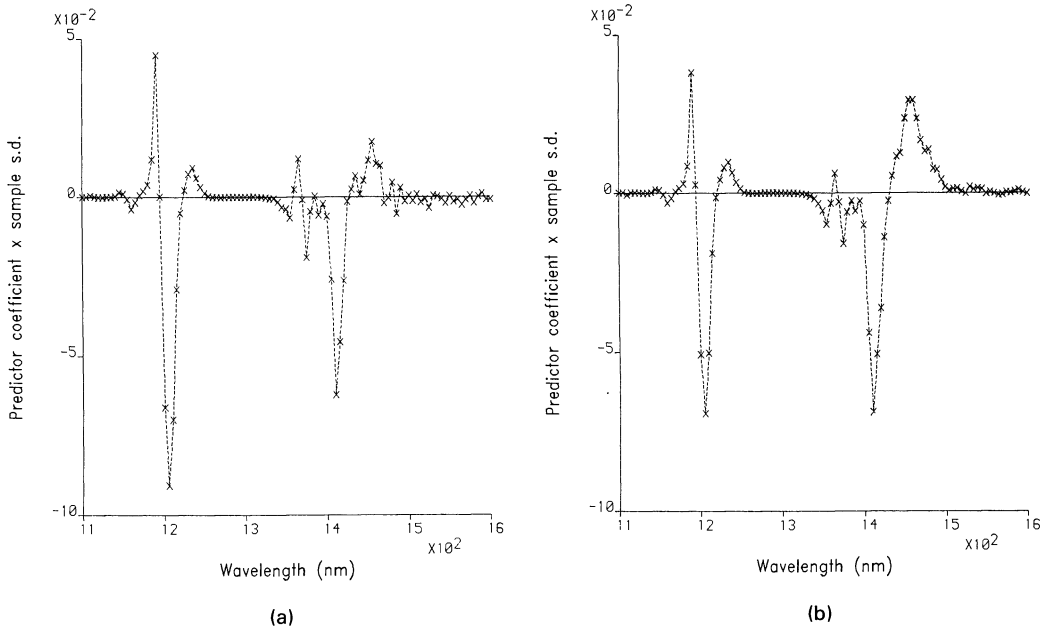

Fig. 4. Predictor coefficients for protein calibration: (a) $\alpha = \frac{1}{2}$, $\omega = 5$; (b) $\alpha = 1$, $\omega = 2$

between the mixtures. (This expectation is borne out by the finding that, with standardization, the maximum value of I drops dramatically to 0.951.)

Selecting $\alpha = 0.5$, $\omega = 5$ as sufficiently near optimal, the corresponding predictor coefficients are shown in Fig. 4(a) (multiplied by the sample standard deviations of the variables for more meaningful comparison). For comparison, Fig. 4(b) shows the predictor coefficients for $\alpha = 1$, $\omega = 2$.

10.5. *Example 5: Near Infra-red Calibration for Fat in Biscuit Dough*

Our final example involves a very large set of data with $n = 39$ pieces of dough and $p = 601$ wavelengths, a part of those analysed by Osborne *et al.* (1984). The variables $\{\dot{x}(j)\}$ are log(1/reflectance) values at intervals of 2 nm between 1200 nm and 2400 nm, jointly scaled to reduce the influence of variations in background reflectance from dough piece to dough piece. The $\dot{y}$ variable is the percentage of fat, which varies around 30%.

The sample standard deviations of $\{\dot{x}(j)\}$ are not very unequal but we have, following example 4, used CR without standardization and with five subintervals per grid interval. Our findings were as follows. For $\alpha = 0.5$ and $\alpha = 1$, the optimal values of ω were both 18, with I values of 0.965 and 0.956 respectively. Small values of α with $\omega = 1$ (approximating the Moore–Penrose limit) give $I = 0.962$. For ω values of 2, 3 and 4, the optimal values of α were 0.05, 0.11 and 0.17 respectively with I values 0.963, 0.964 and 0.963 respectively. The root-mean-square, one-out prediction error for the PLS choice $\alpha = 0.5$, $\omega = 18$ is 0.37% which may be compared with the value 0.43% of the residual standard deviation for the best fitting of the $0.5 \times 601 \times 600 = 180\,300$ pairs of wavelengths analysed by Osborne *et al.* (1984).

11. DISCUSSION

In the absence of any appreciable and presentable theory for CR prediction, it may be unwise to try to extract general conclusions from just five examples. None-theless, some of the following observations are supported by other trials that we have made. They suggest theoretical problems, answers to which may have to rest on Monte Carlo simulation.

(a) Predictors associated with quite different values of (α, ω), but having nearly optimal values of I, are themselves effectively identical. This was remarked on in example 1 but is also to be found in example 3, where the coefficients of $L_1, \ldots, L_6$ for $\alpha = 0.99$, $\omega = 3$ and $I = 0.957$ are -0.02, 0.10, 0.17, -0.21, 0.00 and -0.01 respectively, which differ by at most 0.01 from the coefficients for the optimal $\alpha^\dagger = 0.35$, $\omega^\dagger = 2$ and $I = 0.960$. The same phenomenon persists even more markedly for the data set of example 4 with large p. This robustness seems reasonable when expressed as 'near uniqueness of the near best' and is particularly useful in that it is not necessary to be numerically neurotic in the determination of the optimum. Our I plots can be, and have been, drawn with a thick pen.

(b) For any given value of α, the index $I_{\alpha, \omega}$ tends to be unimodal in ω and often shows the by now well-known cross-validatory hump—'not too little, not too much, but just right'. That we easily achieve only approximate unimodality is shown by the case of $\alpha = 0.35$ in example 3.

(c) In example 1, the value of α maximizing $I_{\alpha, \omega}$ is an increasing function of ω. In example 3, it increases up to $\omega = 4$ but then effectively drops to zero at $\omega = 5$. This further manifestation of a possible unimodality may be interpretable as the complex outcome of a battle between biases and variance in the estimation of the final predictor. As we consider larger values of ω for a given value of α, we lose on variance but gain on biases. To reduce the variance, we need to move away from the correlational adaptability of least squares, i.e. to increase α. But, for the largest value of ω, it is possible that, with the bulk of the variance penalty paid for all but the smaller values of α, it pays to return to least squares to pick up the reduced bias benefits.

(d) The effects of standardization of explanatory variables in example 3 were not as great as those of a preceding linear transformation, while non-standardization was necessary for example 4. More studies are clearly needed, especially if we want to devise rules for non-standardization to be used as a device for prior weighting of the variables.

(e) The calculations are faster the smaller the value of N (Section 8), at the expense of a poorer approximation to the procedure as defined. However, we have found that even $N = 1$ usually gives a reasonable approximation. Moreover the $N = 1$ procedure may be regarded as defined in its own right, so that concern about the closeness of the approximation to a procedure that is itself somewhat arbitrary would not be justifiable.

(f) In examples 4 and 5, the Moore–Penrose OLS predictors were hardly less optimal than the predictors that we selected. To see why this particular choice of OLS predictor should do so well, we simply note that it corresponds to $\mathbf{c}_1 = \mathbf{S}^+\mathbf{s}/\|\mathbf{S}^+\mathbf{s}\|$, and that this is the standardization to unit length of the particular $\boldsymbol{\beta}$ in $\boldsymbol{\beta}'\mathbf{x}$ that satisfies the OLS 'normal equations' which, in the

singular, perfect fit case, are simply $\boldsymbol{\beta}'\mathbf{x}_i = y_i$, $i = 1, \ldots, n$, and minimizes $\|\boldsymbol{\beta}\| = (\beta_1^2 + \ldots + \beta_p^2)^{1/2}$. Now in examples 4 and 5, it seems that the informative $\dot{x}(j)$ are associated with large sample standard deviation. The required minimization then ensures that the choice of $\boldsymbol{\beta}$ is resolved in favour of just those informative $\dot{x}(j)$: hence the near optimal performance. Indeed, in example 5, it is likely that it is the residual adaptability of $\boldsymbol{\beta}$ within the set of informative $\dot{x}(j)$ that has given the Moore–Penrose predictor an edge over PCR.

The CR prediction method is not at all directed towards *selection* of variables, as opposed to their construction. For that purpose, CR cannot therefore be expected to be uniformly superior to sensibly deployed selection methods such as stepwise regression, especially in problems where there really are only a small number of informative explanatory variables.

The procedure proposed in this paper is similar in some respects to the independent work of Frank (1987), who called her method 'intermediate least squares' (ILS). This may be restated as another special case of the general method of Section 3 with $g = 0$. At stage k, calculate the vector, **cov**, of sample covariances of $\dot{x}(1), \ldots, \dot{x}(p)$ with the $\dot{y}$ residuals in the regression of $\dot{y}$ on $\dot{t}(1), \ldots, \dot{t}(k)$. Then $\mathbf{c}_{k+1}$ is taken proportional to the vector in which the α smallest components of **cov**, in magnitude, are set equal to zero. Here α is a control parameter with an integer value from $\{0, 1, \ldots, p-1\}$. As with CR, the control parameters α and ω are chosen by cross-validation. The ILS method generalizes PLS, which corresponds to $\alpha = 0$. However,

(a) we do not see it as a generalization of conventional stepwise regression,
(b) it is necessary to take $\omega = p$, if that is feasible, to obtain OLS and
(c) the 'spectrum' of α values, $\{0, 1, \ldots, p-1\}$, does not include PCR.

An alternative generalization of PLS has been proposed by Lorber *et al.* (1987) giving, like CR, a continuum of procedures around PLS with control parameter given by the 'power' in the eigenanalysis power method. Along with CR, these alternative approaches contrast sharply with the work of Brown (1982), Næs (1985) and Sundberg and Brown (1988), which would model our $\{\dot{x}(j)\}$ as multivariate normal conditional on the value of $\dot{y}$.

We have not yet found any striking examples of the value of the option $g > 0$.

Potential users of CR should not be too bothered about the deficiencies of cross-validatory assessment, or estimation, documented by Stone (1977), Efron (1983) and Bunke and Droge (1984). Success with CR depends on effective *choice* rather than assessment. With quadratic loss for the cross-validatory comparison of two prediction rules, we have

$$\sum_*(y_* - \hat{y}^{(1)}_{\setminus *})^2 - \sum_*(y_* - \hat{y}^{(2)}_{\setminus *})^2 = \left\{\sum_*(\eta_* - \hat{y}^{(1)}_{\setminus *})^2 - \sum_*(\eta_* - \hat{y}^{(2)}_{\setminus *})^2\right\} - 2\sum_*(y_* - \eta_*)(\hat{y}^{(1)}_{\setminus *} - \hat{y}^{(2)}_{\setminus *}).$$

where $E(y_*) = \eta_*$, $* = 1, \ldots, n$. The difference in braces is an improved comparison measure using $\{\eta_*\}$ rather than $\{y_*\}$ as the 'one-out predictees'. The final sum has expectation zero and should have a standard deviation that is small compared with either of the sums in the braces and even, in some comparisons, with their difference.

A Fortran subroutine for CR will be submitted for publication elsewhere.

ACKNOWLEDGEMENTS

We are grateful to H. Martens and T. Fearn for the data of examples 4 and 5 respectively, and to R. A. Stone for otherwise missing references.

APPENDIX A

Proof of identity of subspaces

We know that equation (11) holds for $\omega = 1$. Proceeding inductively, suppose that it is true for $\omega = k$. Let $\mathcal{S}_k$ denote $\langle \mathbf{s}, \mathbf{Ss}, \ldots, \mathbf{S}^{k-1}\mathbf{s}\rangle$ and $\mathbf{h}$ the unit length vector in $\mathcal{S}_{k+1}$ that is orthogonal to $\mathbf{S}\,\mathcal{S}_k = \langle \mathbf{Ss}, \ldots, \mathbf{S}^k\mathbf{s}\rangle$, and therefore $\mathbf{S}$ orthogonal to $\mathcal{S}_k$. Write $\mathbf{h} = \lambda(\mathbf{s} + \mathbf{s}_k)$ where $\mathbf{s}_k$ is in $\mathbf{S}\,\mathcal{S}_k$ (which determines λ and $\mathbf{s}_k$ uniquely). Then, for any $\mathbf{c}$ that is $\mathbf{S}$ orthogonal to $\langle \mathbf{c}_1, \ldots, \mathbf{c}_k\rangle$ ($= \mathcal{S}_k$ by the inductive assumption), we will have $\mathbf{c}'\mathbf{s} = \mathbf{c}'(\mathbf{s} + \mathbf{s}_k) = \mathbf{c}'\mathbf{h}/\lambda$. Hence, for $\mathbf{c}_{k+1}$ to maximize $(\mathbf{c}'\mathbf{s})^2$, we must have $\mathbf{c}_{k+1} = \mathbf{h}$ and $\langle \mathbf{c}_1, \ldots, \mathbf{c}_{k+1}\rangle = \mathcal{S}_k \oplus \langle \mathbf{h}\rangle = \mathcal{S}_{k+1}$, whence equation (11) would hold for $\omega = k + 1$.

The geometrical relationships are shown in Fig. 5, in which □ and [S] denote Euclidean and S-orthogonality respectively, following the conventions set out in Stone (1987).

APPENDIX B

The constructed regressors, at most $m - g$, lie in the range of $\mathbf{S}$

Let $\mathcal{G}$ denote the subspace of $\mathcal{R}_\mathbf{S}$, the range of $\mathbf{S}$, that is spanned by the orthogonal projections of $\mathbf{c}_1, \ldots, \mathbf{c}_g$ on to $\mathcal{R}_\mathbf{S}$. By the suppositions made for $\mathbf{c}_1, \ldots, \mathbf{c}_g$, $\dim \mathcal{G} = g$. For unit length $\mathbf{c}$ in R^p, let $\mathbf{c}_0$ denote its orthogonal projection on to $\mathcal{R}_\mathbf{S}$ and $\tilde{\mathbf{c}} = \mathbf{c}_0/\|\mathbf{c}_0\|$. (We here suppose that $\mathbf{c}$ is such that $\mathbf{c}_0 \neq \mathbf{0}$: otherwise $\mathbf{c}'\mathbf{Sc} = 0$ and $T = 0$.) Then $\mathbf{c}'\mathbf{Sc} = \mathbf{c}_0'\mathbf{Sc}_0$, and also $\mathbf{c}'\mathbf{s} = \mathbf{c}_0'\mathbf{s}$ since $\mathbf{s} \in \mathcal{R}_\mathbf{S}$. Hence T at $\mathbf{c}$ equals

$$(\mathbf{c}_0'\mathbf{s})^2(\mathbf{c}_0'\mathbf{Sc}_0)^{\alpha/(\alpha-1)-1} \leqslant (\tilde{\mathbf{c}}'\mathbf{s})^2(\tilde{\mathbf{c}}'\mathbf{S}\tilde{\mathbf{c}})^{\alpha/(\alpha-1)-1}$$

for $\alpha > 0$, since $\|\mathbf{c}_0\| \leqslant 1$. Moreover $\tilde{\mathbf{c}}$ is of unit length and has the same S-orthogonality properties as $\mathbf{c}$. So, to maximize T, for $\alpha > 0$ we would always replace $\mathbf{c}$ in R^p by $\tilde{\mathbf{c}}$ in $\mathcal{R}_\mathbf{S}$. It may therefore be seen that the construction process, starting with that of $\mathbf{c}_{g+1}$ in $\mathcal{R}_\mathbf{S}$, progressively builds up S-orthogonal $\mathbf{c}$s from the S-orthogonal complement of $\mathcal{G}$ in $\mathcal{R}_\mathbf{S}$, until the dimensionality of $\mathcal{R}_\mathbf{S}$ is exhausted with $\mathcal{G} \oplus \langle \mathbf{c}_{g+1}, \ldots, \mathbf{c}_m\rangle = \mathcal{R}_\mathbf{S}$.

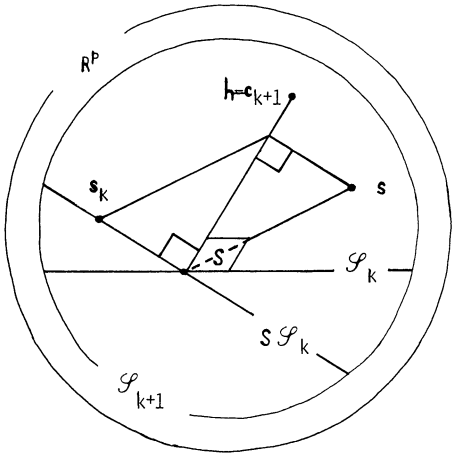

Fig. 5. Inductive identification of $\mathbf{c}_{k+1}$

APPENDIX C

Recurrence relations for $\{\mathbf{M}_k\}$ *and* $\{\mathbf{M}_k^*\}$

(a) Write $\mathbf{M}$ and $\mathbf{A}$ in equation (22) as $\mathbf{M}_k$ and $\mathbf{A}_k$ to exhibit their dependence on k. We have $\mathbf{A}_{k+1}=(\mathbf{A}_k\colon \mathbf{a})$, where $\mathbf{a}'=(e_1z_1, \ldots, e_mz_m)$. Writing $\mathbf{B}_k=\mathbf{A}_k'\mathbf{D}^{-1}\mathbf{A}_k$, we have

$$\mathbf{B}_{k+1}=\begin{pmatrix}\mathbf{B}_k & \mathbf{b}\\ \mathbf{b}' & c\end{pmatrix}$$

where $\mathbf{b}'=\mathbf{A}_k'\mathbf{D}^{-1}\mathbf{a}$ and $c=\mathbf{a}'\mathbf{D}^{-1}\mathbf{a}$ ($\mathbf{z}$, $\mathbf{a}$, $\mathbf{b}$ and c refer to the kth-stage output only). Then

$$\mathbf{B}_{k+1}^{-1}=\begin{pmatrix}\mathbf{B}_k^{-1}+\mathbf{B}_k^{-1}\mathbf{b}\mathbf{b}'\mathbf{B}_k/d & -\mathbf{B}_k^{-1}\mathbf{b}/d\\ -\mathbf{b}'\mathbf{B}_k^{-1}/d & 1/d\end{pmatrix}$$

where $d=(c-\mathbf{b}'\mathbf{B}_k^{-1}\mathbf{b})$. Writing $\mathbf{C}_k=\mathbf{A}_k\mathbf{B}_k^{-1}\mathbf{A}_k'$, we have $\mathbf{M}_k=\mathbf{D}^{-1}-\mathbf{D}^{-1}\mathbf{C}_k\mathbf{D}^{-1}$. Also we find $\mathbf{d}=\mathbf{a}'\mathbf{M}_k\mathbf{a}$ and $\mathbf{C}_{k+1}=\mathbf{C}_k+(\mathbf{a}-\mathbf{C}_k\mathbf{D}^{-1}\mathbf{a})(\mathbf{a}-\mathbf{C}_k\mathbf{D}^{-1}\mathbf{a})'/d$, whence

$$\mathbf{M}_{k+1}=\mathbf{M}_k-(\mathbf{M}_k\mathbf{a})(\mathbf{M}_k\mathbf{a})'/\mathbf{a}'\mathbf{M}_k\mathbf{a}.$$

This recurrence on k starts with the calculation of $\mathbf{M}_0=\mathbf{D}^{-1}$ if $g=0$ and of $\mathbf{M}_g=\mathbf{D}^{-1}-\mathbf{D}^{-1}\mathbf{C}_g\mathbf{D}^{-1}$ if $g>0$.

(b) The recurrence relation for $\mathbf{M}_k^*$, equal to the $\mathbf{M}^*$ of equation (32), has the same formulation. All we need to do is to replace $\mathbf{D}$ and $\mathbf{A}_k$ by $\mathbf{Q}$ and $\mathbf{A}_k^*$ respectively. The recurrence starts with the calculation of $\mathbf{M}_0^*=\mathbf{Q}^{-1}$ if $g=0$ and of

$$\mathbf{M}_g^*=\mathbf{Q}^{-1}-\mathbf{Q}^{-1}\mathbf{A}_g^*(\mathbf{A}_g^{*\prime}\mathbf{Q}^{-1}\mathbf{A}_g^*)^{-1}\mathbf{A}_g^{*\prime}\mathbf{Q}^{-1}$$

if $g>0$. For $\mathbf{a}$ we have, from equation (29),

$$a_i=e_iz_i-\nu f_i^*\sum_{r=1}^{m}f_r^*z_r$$

where $\mathbf{z}^*=\mathbf{M}_k^*\mathbf{d}^*/\|\mathbf{M}_k^*\mathbf{d}^*\|$.

APPENDIX D

Construction at $\alpha=0$ *with* $g>0$

For $\alpha>0$, $g>0$ and $k=g$, the translation of $\mathbf{Md}$ in equation (21) from eigenvector coordinates gives $\mathbf{c}_{g+1}$ proportional to $\mathbf{S}^{-1}\mathbf{s}$ (or $\mathbf{S}^{+}\mathbf{s}$) minus a vector $a_1\mathbf{c}_{10}+\ \ldots\ +a_g\mathbf{c}_{g0}$ in $\mathscr{S}$ (see Appendix B) where $\mathbf{c}_{10}, \ldots, \mathbf{c}_{g0}$ are the orthogonal projections of $\mathbf{c}_1, \ldots, \mathbf{c}_g$ respectively on to $\langle\mathbf{v}_1, \ldots, \mathbf{v}_m\rangle=\mathscr{R}_{\mathbf{S}}$. (The columns of $\mathbf{D}^{-1}\mathbf{A}$ translate into $\mathbf{c}_{10}, \ldots, \mathbf{c}_{g0}$.) Moreover, by the argument used at the end of Section 4, any $\mathbf{c}_{g+2}$ satisfying the required S-orthogonality conditions with $\mathbf{c}_1, \ldots, \mathbf{c}_g, \mathbf{c}_{g+1}$ would have r_c^2, and hence T, zero. For $\omega=g+1$, the predictor is then given by 'with-a-constant' least squares regression of $\dot{y}$ on the $g+1$ ($\leqslant m-1$) regressors $\dot{t}(1), \ldots, \dot{t}(g), \hat{\boldsymbol{\beta}}'\dot{\mathbf{x}}-a_1\mathbf{c}_{10}'\dot{\mathbf{x}}-\ \ldots\ -a_g\mathbf{c}_{g0}'\dot{\mathbf{x}}$, whose data vectors are linearly independent. Since $t_i(j)=\mathbf{c}_j'\mathbf{x}_i=\mathbf{c}_{j0}'\mathbf{x}_i$, $i=1, \ldots, n$, $j=1, \ldots, g$, it follows that the resulting predictor is

$$\bar{y}+\hat{\boldsymbol{\beta}}'\mathbf{x}+a_1(\mathbf{c}_1-\mathbf{c}_{10})'\mathbf{x}+\ \ldots\ +a_g(\mathbf{c}_g-\mathbf{c}_{g0})'\mathbf{x}.$$

APPENDIX E

Interval for ρ^*

For $n > p+1$ and rank $\mathbf{S}^* = \text{rank}\,\mathbf{S} = m = p$ (i.e. both full) the largest eigenvalues of $\mathbf{S}^{*-1}$ and $\mathbf{S}^{-1}$ are e_1^{*-1}, say, and e_1^{-1} respectively. Since $\mathbf{S} - \mathbf{S}^*$ is non-negative definite, $e_1^* \leqslant$ (minimum eigenvalue of $\mathbf{S}$) $= e_1$. Also

$$\begin{aligned}\mathbf{S}^{*-1} &= (\mathbf{S} - \nu \mathbf{x}_* \mathbf{x}_*')^{-1} \\ &= \mathbf{S}^{-1} + \nu \mathbf{S}^{-1}\mathbf{x}_*\mathbf{x}_*'\mathbf{S}^{-1}/(1 - \nu \mathbf{x}_*'\mathbf{S}^{-1}\mathbf{x}_*).\end{aligned}$$

Hence, by the perturbation result (41.8) of Wilkinson (1965),

$$e_1^{*-1} - e_1^{-1} \leqslant \nu \mathbf{x}_*'\mathbf{S}^{-2}\mathbf{x}_*/(1 - \nu \mathbf{x}_*'\mathbf{S}^{-1}\mathbf{x}_*),$$

which, with the fact that $\rho^* \geqslant e_1^*$, gives the lower bound in equation (38) after conversion to eigenvector co-ordinates. The upper bound follows from

$$\begin{aligned}\rho^* \leqslant e_m^* &\overset{\text{def}}{=} \text{maximum eigenvalue of } \mathbf{S}^* \\ &\leqslant \text{maximum eigenvalue of } \mathbf{S} = e_m.\end{aligned}$$

For the remaining cases, rank $\mathbf{S}^* = m - 1$ and we shall prove geometrically that $e_1^* \geqslant e_1$.

In Fig. 6, $\bar{\dot{\mathbf{x}}} + \mathscr{R}_{\mathbf{S}}$ is the m-dimensional flat in R^p that contains the points $\dot{\mathbf{x}}_1, \ldots, \dot{\mathbf{x}}_n$; $\mathscr{A}^*$ is an arbitrary $(m-1)$-dimensional subspace containing $\mathbf{x}_* = \dot{\mathbf{x}}_* - \bar{\dot{\mathbf{x}}}$; $\bar{\dot{\mathbf{x}}} + \mathscr{A}^*$ and $\dot{\mathbf{x}}_i + \mathscr{A}^*$ are the translates of $\mathscr{A}$ through $\bar{\dot{\mathbf{x}}}$ and $\dot{\mathbf{x}}_i$ respectively; $\bar{\dot{\mathbf{x}}}_{\setminus *}$ is the average of $\dot{\mathbf{x}}_i$: $i \neq *$; $\bar{\dot{\mathbf{x}}}_{\setminus *} + \mathscr{R}_{\mathbf{S}^*}$ is the translate through $\bar{\dot{\mathbf{x}}}_{\setminus *}$ of the $(m-1)$-dimensional subspace $\mathscr{R}_{\mathbf{S}^*}$; the symbol $\square$ denotes orthogonal projection. Then, by principal components theory applied first to $\mathbf{S}^*$ then to $\mathbf{S}$,

$$e_1^* = \inf_{\mathscr{A}^*} \sum_{i \neq *} (P_* P_i^*)^2 \geqslant \inf_{\mathscr{A}^*} \sum_{i \neq *} (P_* P_i)^2 = \inf_{\mathscr{A}^*} \sum_{i=1}^{n} (P_* P_i)^2 \geqslant e_1.$$

(For the first equality, we use the fact that the intersection of $\bar{\dot{\mathbf{x}}}_i + \mathscr{A}^*$ and $\bar{\dot{\mathbf{x}}}_{\setminus *} + \mathscr{R}_{\mathbf{S}^*}$ is an arbitrary subflat of $\bar{\dot{\mathbf{x}}}_{\setminus *} + \mathscr{R}_{\mathbf{S}^*}$ through the average of $\dot{\mathbf{x}}_i$: $i \neq *$. The first inequality is by Pythagoras. The second equality holds because $P_* P_* = 0$. The second inequality is a consequence of the restriction on $\mathscr{A}^*$.)

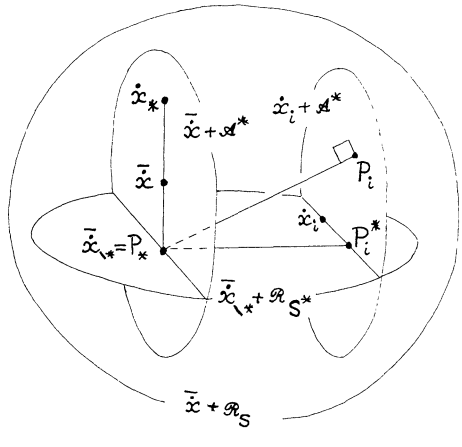

Fig. 6. Construction for the case when rank $\mathbf{S}^* = m - 1$

REFERENCES

Bjørsvik, H. R. and Martens, H. (1989) Data analysis: PLS—calibration of NIR instruments by PLS regression. In *Near-infrared-analysis* (ed. D. A. Burns). New York: Dekker.

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

Bunke, O. and Droge, B. (1984) Bootstrap and cross-validation estimates of the prediction error for linear regression models. *Ann. Statist.*, **12**, 1400–1424.

Daniel, C. and Wood, F. S. (1971) *Fitting Equations to Data*. New York: Wiley-Interscience.

Dijkstra, T. K. (1988) On model uncertainty and its statistical implication. *Lect. Notes Econ. Math. Syst.*, **307**.

Dixon, W. J. (ed.) (1983) *BMDP Statistical Software*. Berkeley: University of California Press.

Draper, N. R. and Smith, H. (1981) *Applied Regression Analysis*. New York: Wiley.

Efron, B. (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Statist. Ass.*, **78**, 316–331.

Farebrother, R. W. (1984) A note on Fearn's "Misuse of ridge regression". *Appl. Statist.*, **33**, 74–75.

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

Fisher, R. A. (1924) The influence of rainfall on the yield of wheat at Rothamsted. *Phil. Trans. R. Soc. Lond.* B, **213**, 89–142.

Frank, I. E. (1987) Intermediate least squares regression method. *Chemomet. Intell. Lab. Syst.*, **1**, 233–242.

Gauss, C. F. (1826) Theoria combinationis observationum erroribus minimis obnoxiae. *Werke*, **4**, 1–93.

Hald, A. (1952) *Statistical Theory with Engineering Applications*. New York: Wiley.

Helland, I. S. (1988) On the structure of partial least squares regression. *Communs Statist. Simuln*, **17**, 581–607.

Hjorth, U. (1989) On model selection in the computer age. *J. Statist. Planng Inf.*, **23**, 101–115.

Hoerl, A. E., Kennard, R. W. and Hoerl, R. W. (1985) Practical use of ridge regression: a challenge met. *Appl. Statist.*, **34**, 114–120.

Lorber, A., Wangen, L. E. and Kowalski, B. R. (1987) A theoretical foundation for the PLS algorithm. *J. Chemomet.*, **1**, 19–31.

Næs, T. (1985) Multivariate calibration when the error covariance matrix is structured. *Technometrics*, **27**, 301–311.

Næs, T., Irgens, C. and Martens, H. (1986) Comparison of linear statistical methods for calibration of NIR instruments. *Appl. Statist.*, **35**, 195–206.

Osborne, B. G., Fearn, T., Miller, A. R. and Douglas, S. (1984) Application of Near Infrared Reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs. *J. Sci. Food Agric.*, **35**, 99–105.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc.* B, **36**, 111–147; corrigendum, **38** (1976), 102.

——— (1977) Asymptotics for and against cross-validation. *Biometrika*, **64**, 29–35.

——— (1987) *Coordinate-free Multivariable Statistics: an Illustrated Geometric Progression from Halmos to Gauss and Bayes*. Oxford: Clarendon.

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

Weisberg, S. (1980) *Applied Linear Regression*. New York: Wiley.

Wilkinson, J. H. (1965) *The Algebraic Eigenvalue Problem*. Oxford: Clarendon.

Wold, H. (1984) PLS regression. In *Encyclopaedia of Statistical Sciences* (eds N. L. Johnson and S. Kotz), vol. 6, pp. 581–591. New York: Wiley.

Wold, S. (1978) Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, **20**, 397–405.

Wood, H., Steinour, H. H. and Starke, H. R. (1932) Effect of composition of Portland cement on heat evolved during hardening. *Indust. Engng Chem.*, **24**, 1207–1214.

DISCUSSION OF THE PAPER BY STONE AND BROOKS

Professor P. J. Brown (Liverpool University): It is always a pleasure to receive an elegant and stimulating paper by Professor Stone and this paper, joint with Dr Brooks, is no disappointment. The idea behind the paper is simple; it seeks to combine three methods for prediction of a response y from a vector of explanatory variables x, when the predictor is linear in the explanatory variables. Indeed it elegantly ties together two established methods, ordinary least squares (OLS) and principal component regression (PCR), and brings in the newly emerging method of partial least squares (PLS), shedding some new light on this in the process. A continuum of intervening possibilities is also instated.

I particularly liked the insight provided by the different optimality criteria of OLS, PCR and PLS and especially the canonical covariance depiction of PLS. My own struggle to understand PLS was enlightened earlier by the work of Helland (1988). He stripped away its algorithmic cover and came nearest to specifying a model. For me an understanding of PLS lies at the root of the appreciation of the paper. Let me first pause to restate the salient features of PLS. The model matrix X of n observations on p explanatory variables can be described in bilinear factor form:

$$X = t_1 p_1' + t_2 p_2' + \ldots + t_k p_k' + E_k$$

where the scores t_i are n-vectors. They are the latent variables and the p-vectors p_j are the loadings. The residual matrix E_k is small in some sense. The crucial idea of PLS is that the relationship between X and y is conveyed through the latent variables. Thus we also have the decomposition

$$y = t_1 q_1 + t_2 q_2 + \ldots + t_k q_k + f_k$$

for scalar q_j and the *same* scores or latent factors. Incidentally PLS can quite naturally include vector y leading to multivariate regression. Various conditions need to be imposed for uniqueness. We could force the scores to be mutually orthogonal in R^n or the loadings to be mutually orthogonal in R^p. We have to shy away from imposing both simultaneously since then the t_i would be eigenvectors of XX' and p_j eigenvectors of $X'X$ and the latent factors would be entirely determined by the X data without reference to y, as in the PCR method. There are thus two main algorithms for PLS depending on whether the scores or the loadings are determined as orthogonal. The algorithms are sequential, starting with no factors adding a factor at each stage. For one algorithm, latent variables are formed as weighted averages of the X-residuals from the previous step, with weights proportional to covariances with the y-residuals from the previous step. Only the simplest least squares algorithms are required. Although the construction algebra of the paper removes the arbitrariness in the specification of weights in PLS, for general continuum regression it does have the quite substantial overhead of the iterative numerical solution of equation (23).

What emerges in PLS is that the regression coefficients are formed as

$$\hat{B}_k = H_{(k)} X' y$$

where $H_{(k)}$ is a rank k approximation to the inverse of $X'X$ and the usual algorithm is indeed just the conjugate gradient method of forming an inverse; see Westlake (1968). In PLS the conjugate directions are formed with respect to y, whereas in PCR the approximating inverses are formed on the basis of $X'X$ alone. We see also that PLS and PCR can be viewed as shrinkage methods, although the shrinkage of PLS is decidedly more obscure than that of say ridge regression, with its implicit Bayesian assumption of exchangeability of regression coefficients.

PLS has this motivational notion of latent factors and I wonder whether much is gained by adding to PLS a continuum of other possibilities. After all PLS with $\min(p, n-1)$ factors gives OLS, or minimum length OLS when $n-1 < p$. Also PCR with $\min(p, n-1)$ factors gives the same Moore–Penrose minimum length solution. Augmenting PLS with PCR having fewer than the saturated number of factors would cover all three techniques. With continuum regression the (α, ω) two-dimensional parameter space is peculiar. Points in this two-dimensional space far apart can be coincident in a model difference sense. It is unclear how in this space points relate to the regression coefficients. The authors refer to this in Section 11, point (a). However, I wonder whether it is satisfactory to ascribe virtue to the similarity of coefficients for seemingly very different 'models'. How much is due to the particular data and how much to the overlaying nature of the methodology? I note that the only attempt at cross-validatory assessment as opposed to cross-validatory choice was made in example 3, and the validation set there

did not behave quite as anticipated. When we add in the choice of scaling and metric we are left with a bewildering range of possibilities. Perhaps it is more than enough to try to contend with PLS.

If I am permitted some criticism it is that overriding all these methods I feel unease at the 'black box' approach and lack of attention to prior knowledge encompassed in the substantive application. They are biased non-linear shrinkage methods and alternatives like ridge regression are more explicit in their implicit prior assumptions, guiding users on when and how to use it. The method is prescriptive and is not embedded in an inferential framework to judge the relative merits of the prescriptions. Modelling as such is eschewed. I wonder whether we are being treated to *synthetic* rather than *soft* science? The chemometrician or statistician user is left with the comfortable notion that he can collect a batch of data; he does not have to worry too much about how he collects it, or what past knowledge there is. He can apply continuum regression with the assurance that after a little fine tuning he will have a good predictor for all future unspecified purposes!

I am grateful to the authors for mentioning, in Section 11, papers in which I had a hand. I am sorry that our intended message did not get through. *Modelling* is paramount. Whether to regress x on y or y on x would depend on the way that the training data have been collected, whether x or y had been controlled. But when $n-1 < p$ then regressing either way leads to the same least squares estimates with a $(p-n+1)$-dimensional degree of undeterminedness. I believe that prior information is then crucial in forming a unique estimator. There is much about the infra-red data of examples 3–5 which needs modelling. For finely ground solids, particle size has a substantial influence on relectance, shifting bodily the reflectance curve. Also looking at one of these reflectance curves as a function of wavelength one is immediately struck by a beautiful continuity. Try jumbling up the p wavelengths, destroying this continuity, and apply continuum regression. Out comes the same answer as before. Continuity is not utilized. A start in accounting for continuity, through spline fitting and autoregressive error, is given in Brown and Denham (1989) and is the subject of a Science and Engineering Research Council Complex Stochastic Systems project.

Finally I wonder whether criterion (12) of the paper can be justified as a utility in a wider decision theory framework. This is pertinent because of the emphasis on estimation and prediction rather than modelling and inference.

This paper offers many insights and impressively ties together three different prescriptions in what the authors have referred to as soft science. Although I may have reservations concerning the downgraded role of modelling, I have no hesitation in proposing the vote of thanks.

Dr T. Fearn (University College London): It is a pity, given the elegance of the appendixes, to resort to the use of co-ordinates in discussing this paper. However, it is interesting to look at continuum regression (CR) using the canonical form of the linear model in which ridge regression is often explored (see for example Goldstein and Smith (1974)):

$$y_i \sim N(\beta_i \sqrt{e_i},\ \sigma^2) \qquad i=1, \ldots, p$$
$$y_i \sim N(0,\ \sigma^2) \qquad i=p+1, \ldots, n.$$

In this form $\mathbf{S}$ is just $\mathrm{diag}(e_1, \ldots, e_p)$, where the non-standard ordering $e_1 < \ldots < e_p$ corresponds to that in Section 8, and β_i is the regression coefficient associated with the ith eigenvector of $\mathbf{S}$. The least squares estimates are just $\hat{\beta}_i = y_i/\sqrt{e_i}$, and the simple form of the ridge regression estimator

$$\beta_i^{\mathrm{RR}} = \frac{e_i}{\varkappa + e_i}\hat{\beta}_i,$$

where $\varkappa > 0$ is the ridge constant, shows how ridge regression shrinks preferentially in directions associated with small eigenvalues of $\mathbf{S}$. The corresponding estimator for CR with $g=0$, $w=1$, is

$$\beta_i^{\mathrm{CR}} = \frac{e_i}{\gamma\rho + (1-\gamma)e_i}\hat{\beta}_i. \tag{47}$$

Here γ varies from zero (OLS) via unity (PLS) to infinity (PCR) and ρ, which varies with γ and depends on both y and e, lies in the interval $e_1 \leqslant \rho \leqslant e_p$. The behaviour of equation (47) as $\gamma \to \infty$ is not obvious at first glance; what appears to happen is that $\rho \to e_p$ so that $\beta_p^{\mathrm{CR}} \to \hat{\beta}_p$ and $\beta_i^{\mathrm{CR}} \to 0$ for $i \neq p$.

Comparison of equation (47) with β_i^{RR} shows that CR may behave like ridge regression for small γ, with $\gamma\rho$ playing the role of $\varkappa$. In example 1 the e_i are 0.002, 0.187, 1.576 and 2.236; when $\alpha \approx \frac{1}{3} = 0.006$, ρ is about 1.9, $\gamma\rho$ about 0.011 and the shrinkage factors for CR are very close to those for ridge regression with $\varkappa = 0.011$. For these choices of γ and $\varkappa$ both CR and ridge regression effectively delete the direction corresponding to the smallest eigenvalue, giving results almost identical with $\alpha = 1$, $\omega = 3$, as noted in Section 10.1. The similarity to ridge regression here depends on the spread of the e_i and the fact that ρ is large for small γ. This need not always hold, and the general question of when the two methods are similar would bear further investigation.

Although CR is a 'shrinker' in the sense that $\|\beta^{CR}\| \leqslant \|\hat{\beta}\|$ it does not, unlike ridge regression, shrink each individual $\hat{\beta}_i$, as can be seen from equation (47) and the fact that $e_1 \leqslant \rho \leqslant e_p$. In fact the regression coefficients associated with large eigenvalues are inflated by CR. At first sight this seems a little disturbing after so many years of being told that shrinkage is a Good Thing, although the authors might well retort that the componentwise behaviour is irrelevant.

As may be deduced from my attempts to introduce one, I find the lack of any model, in the usual sense, something of a problem with CR. When a procedure is defined by an algorithm (as PLS usually is) or as the result of maximizing some arbitrary criterion it is easier to implement it than to understand when its use might be appropriate, or more importantly not appropriate. Clearly CR is more appropriate if you believe in some sort of latent structure model for the x than if you do not. Even then, however, it is not at all obvious why maximizing the particular criterion chosen is a good way to proceed. The authors offer little help in this direction, making no attempt to justify their procedure.

As the authors point out in their discussion, CR makes no attempt to select variables and cannot be expected to be uniformly superior to 'sensibly deployed . . . stepwise regression'. This point is particularly important in interpreting, for example, Fig. 1, where OLS on the left-hand side is not allowed to select variables even when p approaches or even exceeds n, and PCR at the other end is not allowed to select components according to their correlation with y. This type of contest—take the standard methods, restrict them severely and then compare them against a novel method—is much favoured in the 'soft modelling' literature. Given the framework in which the comparison is made we would expect to see values of α in the middle of the interval doing better than the extremes. This does not necessarily mean that PLS is better than sensibly used OLS or PCR.

These criticisms apply to the whole soft modelling area, and it is perhaps unfair to level them at a paper which is atypical in that it illuminates rather than obscures the issues involved. As one who has struggled with the literature of PLS for some time I am particularly grateful for that illumination.

It gives me great pleasure to second the vote of thanks.

The vote of thanks was passed by acclamation.

Mike Denham (University of Liverpool): In this paper the authors have illustrated their continuum regression approach with five examples. In the three near infra-red calibration examples there is a clear continuity among the x variables. I should like to discuss some work on accounting for this continuity alluded to by Professor Brown.

Brown and Denham (1989) consider modelling the variables $\{\dot{x}(j)\}$ conditional on the value of variable $\dot{y}$ which is now generalized to be multivariate, i.e. $\{\dot{y}(k)\}$ where $k = 1, \ldots, q$. We do this through a standard linear model of the form

$$X = YB + E \tag{48}$$

where $B = (B_{jk})$ is a $q \times p$ matrix of unknown regression coefficients and

$$E = (\epsilon_1, \ldots, \epsilon_n)^T \qquad E(\epsilon_i) = 0;\ E(\epsilon_i \epsilon_i^T) = \Gamma;\ E(\epsilon_i \epsilon_l^T) = 0, \quad i \neq l.$$

Prediction is then performed by using generalized least squares methods conditional on the estimates obtained from model (48).

To introduce continuity between the x variables, we impose a cubic spline structure on the matrix B so that model (48) becomes

$$X = Y\Theta D_\tau + E \tag{49}$$

where D_τ depends only on the knot sequence τ of the underlying spline function. The choice of knot

sequence could be made to reflect the prior beliefs about the smoothness of B or alternatively could be estimated from the data in some way (see for example Smith (1979)).

In addition we also consider structuring the variance–covariance matrix Γ by assuming that the error vectors ϵ_i are independent realizations of an autoregressive process of order m and derive an approximate generalized least squares predictor for autoregressive processes of reasonably low order.

In random calibration where we model the y variables conditional on the x variables by

$$Y = XB_* + E_* \tag{50}$$

we consider estimation of B_* by Moore–Penrose ordinary least squares as a limiting case of the multivariate ridge approach of Brown and Zidek (1980) where a ridge estimator of B_* is given by

$$\hat{B}_*(K) = (X^{\mathrm{T}}X \otimes I_q + I_p \otimes K)^{-1}(X^{\mathrm{T}}X \otimes I_q)\hat{B}_* \tag{51}$$

with K a $q \times q$ positive definite ridge matrix and $\hat{B}_*$, $\hat{B}_*(K)$ vectorized versions of $\hat{B}_*$ and $\hat{B}_*(K)$ obtained by stringing out the matrices row by row as column vectors. A generalization of this which would take account of the continuity of B_* would be to replace I_p by a second ridge matrix L to give

$$\hat{B}_*(K, L) = (X^{\mathrm{T}}X \otimes I_q + L \otimes K)^{-1}(X^{\mathrm{T}}X \otimes I_q)\hat{B}_*. \tag{52}$$

Professor A. C. Atkinson (London School of Economics and Political Science): The methods that we have heard so elegantly described are based on aggregate statistics, i.e. on quantities summed over all data or cases. However, one or a few cases may unduly influence the conclusions of an analysis of data. For least squares regression, methods based on the deletion of single cases are widely used for diagnosing unsuspected influence and other model failures (Cook and Weisberg, 1982; Atkinson, 1985). The algebra for these calculations is similar to some of that given by the authors in Section 9, especially, as they imply, around equation (40). It seems that diagnostic information may therefore be available as a by-product of the authors' algorithm. Is this so? Have the authors any experience in the use of diagnostic methods for detecting the effect of individual data or cases on their inferences?

Docent Urban Hjorth (Linköping University): Cross-validation was first a technique for evaluation of a single estimated model. Stone (1974) extended this into a method for the estimation of parameters. The potential of the method lies in its generality. The computational approach allows almost any kind of parameter to be estimated with optimal predictions as a goal. I have found this approach very useful for model selection as a stopping rule in stepwise regression, for selecting time series models and for some other estimation and selection problems (Hjorth, 1989; Hjorth and Holmqvist, 1981). Stone and Brooks use it here for the complex parameters α and ω, which is also a kind of model selection.

In many estimation problems, cross-validation has tough competition with bootstrap analysis, and the latter should be preferred in some problems, but I cannot see how a bootstrap analysis can handle, for example, model choice in regression without introducing a large amount of irrelevant noise due to variability of the design matrix. Can Professor Stone comment on whether the bootstrap can in principle be applied to his kind of analysis? Stone and Brooks work with linear models in such a way that all available predictors are involved. Their method for noise reduction is to compromise between regression on $X_1, \ldots, X_k$ and regression on principal components $P_1, \ldots, P_k$. The parameter α, $0 \leqslant \alpha \leqslant 1$, is natural as a description of where to take this compromise. However, their second parameter ω works for principal components regression but not at all for ordinary least squares. This difference between the two ends will favour large α-values. A corresponding parameter at $\alpha = 0$ would require a natural order of the predictors or perhaps use of Y to determine such an order by stepwise regression.

The situation is illustrated in Fig. 7 where the empty upper left-hand side indicates lack of alternatives. Do the authors see any useful extension of ω for α close to zero?

Partial least squares is usually regarded as suboptimal and perhaps slightly heuristic. One of the most interesting conclusions here is that with a proper balance between noise and signal (quite a lot of noise), partial least squares can be the optimal compromise between ordinary least squares and principal components.

Statisticians working on multivariate applications with linear models, principal components etc. will have a valuable new tool here. With its built-in cross-validation some decisions will be very expedient even if, at the regression end of the spectrum, some alternatives are worth studying as indicated by the author's discussion.

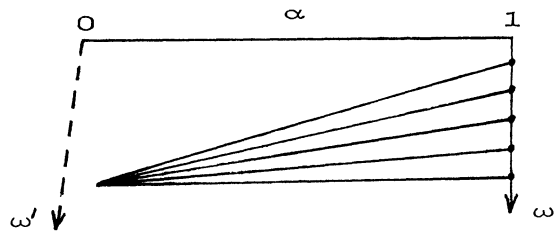

Fig. 7

Dr R. W. Farebrother (University of Manchester): The authors are to be congratulated for establishing that ordinary least squares, partial least squares and principal components regression predictions may be obtained from special cases ($\mathbf{A}=\mathbf{ss}'$, $\mathbf{B}=\mathbf{S}$; $\mathbf{A}=\mathbf{ss}'$, $\mathbf{B}=\mathbf{I}$; and $\mathbf{A}=\mathbf{S}$, $\mathbf{B}=\mathbf{I}$) of the constrained maximization problem

$$\text{maximize } \mathbf{c}'\mathbf{Ac}/\mathbf{c}'\mathbf{Bc} \text{ subject to } \mathbf{c}'\mathbf{SD}=0,\ \mathbf{c}'\mathbf{c}=1$$

where $\mathbf{D}=[c_1, c_2, \ldots, c_{k-1}]$. However, it should be pointed out

(a) that the authors' choice of a one-parameter family based on $T=(\mathbf{c}'\mathbf{s})^2(\mathbf{c}'\mathbf{Sc})^{\gamma-1}$ is not the only possible choice and
(b) that it may not be appropriate to include principal components regression in such a family. My experience of 'soft science' suggests that low order principal components are often more important explanatory variables than high order components. Perhaps a variant based on latent root regression as a limiting case would be more appropriate; see Mason (1986).

Tormod Næs (MATFORSK, Ås): Although I am in general quite pessimistic with respect to the potential for significant improvements over principal components regression (PCR) when prediction is a sensible thing to do, the introduction of partial least squares and later *continuum regression* is very important for the understanding of linear prediction. I very much like the philosophy of extracting components and using these in regression. First it is good for the understanding of prediction as a balancing of X and Y information. Secondly, components, e.g. principal components, are useful for interpretation and understanding of the data. Thirdly, results are more easily communicated to practitioners in terms of such methods.

The reason for my scepticism with respect to the potential for improvement over PCR (which is also supported by the examples in the paper) is described in Martens and Næs (1989). That discussion shows that eigenvector directions in X space with moderate to large eigenvalue and moderate to large correlation with y should always be used for prediction. In contrast, directions with small eigenvalue and small correlation should always be deleted. The directions with moderate to large eigenvalue and small correlation could also in general be deleted, but this will usually have little impact since the variability is properly spanned along these axes. This is also supported by example 1 in the paper. The really complicated directions in X space are those with small eigenvalue and moderate to large correlation with y. If the correlation is real, these directions should be used. However, we should be aware that the prediction ability will usually be poor in such cases, owing to the lack of variability in X space along the actual axes. In other words, prediction will give poor results and should usually be avoided in such situations. If the correlation is accidental (as it will be in most reasonably well-designed experiments), the comparison of the mean-squared errors shows that it is very important to delete the eigenvectors from prediction. For me, all this shows that a PCR with inclusion of components starting with the component with largest eigenvalue and which stops according to a criterion related to increase in cross-validated prediction error is a strategy that in most situations handles all these cases reasonably well.

Dr Lars Ståhle (Karolinska Institute, Stockholm): The paper is a most interesting approach to the evergreen linear predictor problem. In view of recent experiences with partial least squares (PLS) I would like to comment on some aspects of continuum regression (CR). Firstly, there are many apparently competitive ways of regarding regression methods as special cases of one another. The method used in the paper is different from that of, for example, Höskuldsson (1988) who noted that PLS and principal

components regression (PCR) are just two special choices of subspaces in the measurement space. Any subspace with a smaller dimensionality than the recorded data might be chosen by means of some criterion. What is important is a clear statement of the criterion. The authors have chosen the cross-validation quadratic loss function L of the outcome variable (their equation (24)). We have recently developed a variant of PLS in which the subspace is determined by sequential calculation of weights ($\mathbf{w}$) such that L is minimized (prediction-optimized PLS). This is done by introducing a $p \times p$ diagonal matrix $\mathbf{A}$ the elements of which are chosen by cross-validation. In the PLS algorithm an extra step is added: $\mathbf{w} \rightarrow \mathbf{Aw}$. For many data sets with a small number of observations ($n < 20$) the solution becomes seriously pathological with only a minute proportion of the variance in $\mathbf{X}$ explaining almost 100% of the variance in $\mathbf{Y}$. A reformulation of the criterion was therefore necessary using cross-validation of both $\mathbf{X}$ and $\mathbf{Y}$. This suggests that much of the success of PLS as a predictive method lies in that the covariance between $\mathbf{X}$ and $\mathbf{Y}$ is maximized (Höskuldsson, 1988). We shall come back to this in the near future.

The purpose of this comment is thus to point out that restriction of cross-validation to $\mathbf{Y}$ with models that are not clear in their intention (obviously the intention of PLS is different from that of ordinary least squares or PCR) is associated with a risk of systematically finding spurious correlations. I do not know whether this will occur in CR but, if so, it would be unfortunate considering that the intention of the authors is precisely the opposite. Nevertheless, it will be most interesting to see to what extent CR PLS, intermediate least squares and other methods (prediction-optimized PLS?) can be shown to reduce the prediction error on real data in truly prospective studies.

The following contributions were received in writing after the meeting.

F. Y. Chan (University of Winnipeg) **and T. K. Mak** (Concordia University, Montreal): The idea of 'continuum regression', which unifies the classical ordinary least squares, partial least squares and principal components regression, is elegant. Its implementation requires first the computation of the $\mathbf{c}_j$ and Professor Stone and Dr Brooks have outlined a viable approach in Section 8. We would like to discuss here an alternative approach for computing the $\mathbf{c}_j$. The method is conceptually simple and involves at each stage the solution of an explicit equation in a scalar variable. Furthermore, the computations of eigenvalues and eigenvectors are avoided, which may save some computational effort when a large number of explanatory variables is involved.

Following the notation of Stone and Brooks, suppose that $\mathbf{c}_{g+1}, \ldots, \mathbf{c}_k$ have been constructed and we want to find the $\mathbf{c}_{k+1}$ which maximizes $T = (\mathbf{c}'\mathbf{s})^2 (\mathbf{c}'\mathbf{S}\mathbf{c})^{\gamma-1}$ subject to the constraints $\|\mathbf{c}\|^2 = 1$ and $\mathbf{c}'\mathbf{S}\mathbf{c}_j = 0$, $j = 1, \ldots, k$. Using Lagrange multipliers, we have to differentiate with respect to $\mathbf{c}$.

$$\mathbf{f} = (\mathbf{c}'\mathbf{s})^2 (\mathbf{c}'\mathbf{S}\mathbf{c})^{\gamma-1} - \lambda(\mathbf{c}'\mathbf{c} - 1) - \mathbf{c}'\mathbf{S}\boldsymbol{\psi}\mathbf{d},$$

where $\boldsymbol{\psi} = [c_1, c_2, \ldots, c_k]$, λ and $\mathbf{d} = (\lambda_1, \ldots, \lambda_k)'$ are Lagrange multipliers. We have

$$\frac{\partial \mathbf{f}}{\partial \mathbf{c}} = 2(\mathbf{c}'\mathbf{s})\,(\mathbf{c}'\mathbf{S}\mathbf{c})^{\gamma-1}\mathbf{s} + 2(\gamma-1)\,(\mathbf{c}'\mathbf{s})^2\,(\mathbf{c}'\mathbf{S}\mathbf{c})^{\gamma-2}\mathbf{S}\mathbf{c} - 2\lambda\mathbf{c} - \mathbf{S}\boldsymbol{\psi}\mathbf{d} = 0.$$

It can be shown, using $\mathbf{c}'\mathbf{c} = 1$ and $\mathbf{c}'\mathbf{S}\boldsymbol{\psi} = 0$, that

$$\lambda = \gamma(\mathbf{c}'\mathbf{s})^2\,(\mathbf{c}'\mathbf{S}\mathbf{c})^{\gamma-1}$$

and

$$\mathbf{d} = (\boldsymbol{\psi}'\mathbf{S}\boldsymbol{\psi})^{-1}\{2(\mathbf{c}'\mathbf{s})\,(\mathbf{c}'\mathbf{S}\mathbf{c})^{\gamma-1}\boldsymbol{\psi}'\mathbf{s} - 2\gamma(\mathbf{c}'\mathbf{s})^2\,(\mathbf{c}'\mathbf{S}\mathbf{c})^{\gamma-1}\boldsymbol{\psi}'\mathbf{c}\}.$$

Substituting these expressions in $\partial\mathbf{f}/\partial\mathbf{c}$ and simplifying, we obtain the matrix equation

$$(\mathbf{c}'\mathbf{S}\mathbf{c})\mathbf{s} + (\gamma-1)\,(\mathbf{c}'\mathbf{s})\mathbf{S}\mathbf{c} - \gamma(\mathbf{c}'\mathbf{s})\,(\mathbf{c}'\mathbf{S}\mathbf{c})\mathbf{c} - (\mathbf{c}'\mathbf{S}\mathbf{c})\mathbf{S}\boldsymbol{\psi}(\boldsymbol{\psi}'\mathbf{S}\boldsymbol{\psi})^{-1}\boldsymbol{\psi}'\mathbf{s} + \gamma(\mathbf{c}'\mathbf{s})\,(\mathbf{c}'\mathbf{S}\mathbf{c})\mathbf{S}\boldsymbol{\psi}(\boldsymbol{\psi}'\mathbf{S}\boldsymbol{\psi})^{-1}\boldsymbol{\psi}'\mathbf{c} = 0.$$

Writing $\mathbf{A}$ for $\mathbf{I}-\mathbf{S}\psi(\psi'\mathbf{S}\psi)^{-1}\psi'$ and ρ for $\mathbf{c}'\mathbf{S}\mathbf{c}$, the equation can be rewritten as

$$(\mathbf{c}'\mathbf{s})^{-1}\mathbf{A}\mathbf{s}-\{(1-\gamma)\rho^{-1}\mathbf{S}+\gamma\mathbf{A}\}\mathbf{c}=0.$$

It then becomes

$$(\mathbf{c}'\mathbf{s})\mathbf{c}=\{\gamma\mathbf{I}+(1-\gamma)\rho^{-1}\mathbf{R}^{-1}\}^{-1}\mathbf{s}$$

where $\mathbf{R}=\mathbf{S}^{-1}-\psi(\psi'\mathbf{S}\psi)^{-1}\psi'$ does not involve γ, $\mathbf{c}$ and ρ. It can be seen that for a given ρ

$$\mathbf{c}=\frac{\mathbf{H}(\rho)\mathbf{s}}{\{\mathbf{s}'\mathbf{H}(\rho)\mathbf{s}\}^{1/2}},$$

where $\mathbf{H}(\rho)=\{\gamma\mathbf{I}+(1-\gamma)\rho^{-1}\mathbf{R}^{-1}\}^{-1}$. Thus, to find $\mathbf{c}_{k+1}$, we only have to find a root of

$$\rho-\frac{\mathbf{s}'\mathbf{H}(\rho)\,\mathbf{S}\mathbf{H}(\rho)\mathbf{s}}{\mathbf{s}'\mathbf{H}(\rho)\mathbf{s}}=0$$

and substitute it into the expression of $\mathbf{c}$.

Since $\mathbf{H}(\rho)$ is an explicit function of ρ and γ, $\mathbf{c}_{k+1}$ may be expressed as an explicit function of γ when the solution of the last equation can be expressed as an explicit function $g(\gamma)$. One may explore the existence of the $g(\gamma)$ using a symbolic computing package such as MATHEMATICA.

We plan to compare our method with that of Stone and Brooks. We are also interested in investigating whether our method may simplify the cross-validatory algebra in Section 9.

Ildiko E. Frank (JerIl, Inc., Stanford) **and Jerome H. Friedman** (Stanford University): Combining ordinary least squares (OLS), partial least squares (PLS) and principal components regression (PCR) into a single framework is helpful in understanding their relationships. The statistical motivation is less apparent. The assumption that if a method contains others as special cases it is necessarily superior to those others, while often plausible, is not always true and must be demonstrated. This premise is especially suspect here because the two parameters (α, ω) essentially regulate the same thing in only slightly different ways. Both control the strength of the penalty imposed on solutions $\mathbf{c}$ for smaller $\mathbf{c}^{\mathrm{T}}\mathbf{S}\mathbf{c}$. Increasing α does this directly. Decreasing ω does this indirectly by restricting $\mathbf{c}$ to lower dimensional subspaces chosen so that any $\mathbf{c}$ within them cannot achieve a value of $\mathbf{c}^{\mathrm{T}}\mathbf{S}\mathbf{c}$ that is too small. The subspaces are nested and (for a given value of α) ordered on their $\mathbf{c}^{\mathrm{T}}\mathbf{S}\mathbf{c}$ bound. Varying α would seem to be more natural since it is a continuous variable allowing finer tuning. This α–ω equivalence is clearly reflected in the examples where low α, low ω solutions are seen to be equivalent to high α, high ω solutions, certainly within the accuracy of the cross-validated estimate of prediction error. The only exception is example 2 where the signal-to-noise ratio is sufficiently small that maximal penalty is required. The discreteness of ω requires its next higher value to be $\omega=2$, which is already too large.

If we assume the correctness of an underlying linear model $\mathbf{c}^*$ with no prior information on $\mathbf{c}^*/\|\mathbf{c}^*\|$, then the optimal method (in terms of expected squared error loss) chooses $\mathbf{c}$ to maximize

$$(\mathbf{c}^{\mathrm{T}}\mathbf{s})^2/(\mathbf{c}^{\mathrm{T}}\mathbf{S}\mathbf{c}+\lambda), \qquad \mathbf{c}^{\mathrm{T}}\mathbf{c}=1,$$

and then takes as the estimate for the coefficient vector

$$\hat{\mathbf{c}}=\mathbf{c}\{\mathbf{c}^{\mathrm{T}}\mathbf{s}/(\mathbf{c}^{\mathrm{T}}\mathbf{S}\mathbf{c}+\lambda)\}$$

(ridge regression). The parameter λ plays a dual role of regulating the $\mathbf{c}^{\mathrm{T}}\mathbf{S}\mathbf{c}$ penalty and the shrinkage of the final solution vector. In its former capacity it is quite similar to the role played by α. We could generalize ridge regression by the paradigm outlined in the paper to construct additional variables. This generalized procedure includes ridge regression as a special case ($\omega=1$) but the theory tells us that it would not perform as well.

PCR and PLS basically do well to the extent that they emulate ridge regression. This will tend to be the case in highly collinear settings where the effect of the $\mathbf{c}^{\mathrm{T}}\mathbf{S}\mathbf{c}$ penalty dominates the shrinkage of the solution norm. A large simulation study (Frank, 1989) shows that ridge regression, PLS and PCR behave quite similarly, all with vastly superior performance to OLS, and with ridge regression dominating PLS and PCR (sometimes only slightly) in all situations considered.

Professor Inge S. Helland (Agricultural University of Norway, Aas): As Professor Stone and Dr Brooks mention in their discussion, predictors that are close to being optimal are themselves effectively equal. For principal components regression (PCR) and partial least squares (PLS), this can be related to the following results from Helland (1989): for each $m = 1, 2, \ldots, p$ a hypothesis H_m can be formulated in terms of $\mathrm{var}(\mathbf{x}) = \Sigma$ and $\mathrm{cov}(\mathbf{x}, y) = \sigma$, saying that m components in $\mathbf{x}$ are relevant for the prediction of y. (One formulation is that m eigenvectors of Σ have components along σ; another, equivalent, is that $\dim(\sigma, \Sigma\sigma, \Sigma^2\sigma \ldots) = m$.) Under this hypothesis the population versions of PCR (with the correct ordering) and PLS will both stop after m steps, and they will both give the best linear prediction of y. Since sample versions of the two predictors are continuously dependent on the (co)variances, they must be close after m steps when H_m is true, and then they are also close to the optimal solution. For other values of m, PCR and PLS can be very different.

The prediction resulting from maximum likelihood estimation of the covariance structure under H_m and multinormality is now under investigation. Numerically it is easiest to handle a stepwise version of this, which in the notation of the present paper corresponds to minimizing

$$g(\mathbf{c}) = \{\|\mathbf{y}\|^2(\mathbf{c}'\mathbf{S}\mathbf{c}) - (\mathbf{c}'\mathbf{s})^2\}\mathbf{c}'\mathbf{S}^{-}\mathbf{c}.$$

Using this procedure on the Hald data set (example 1) gives very promising results: a likelihood ratio test rejects $m = 1$ ($P = 0.06$), and a two-component predictor leads directly to the optimal solution of this paper ($I = 0.97177$ and the same predictor coefficients) without use of cross-validation.

For Fearn's data (example 3), the result depends on the pretreatment of the data. Suppose that the variables used are $L_1 - \bar{L}, \ldots, L_6 - \bar{L}$ in the notation of this paper. (This is permissible: the mean is non-significant in the full regression model, and the methods work also when Σ is singular.) Then maximum likelihood gives its best solution for $m = 1$. The I-value for this solution is low, but the root-mean-square prediction errors are lower than those of the continuum regression solution for both validation sets.

Unfortunately, the maximum likelihood approach does not give the best solution in all cases. In some simulated examples it performs definitely worse with respect to prediction than PCR and PLS. In general the maximum likelihood solution is close to the alternative PCR version which includes the components with large values of Student's t. As in the area as a whole, further studies are definitely needed.

Dr I. T. Jolliffe (University of Kent, Canterbury): I have two comments, each leading to a question. First, the way in which principal component regression is presented is a somewhat restricted version, with the strategy for selecting components based only on the size of their eigenvalues. It is well known (Jolliffe, 1982) that small variance components are quite frequently included among those which have the largest correlations with the dependent variable. For this reason, various more sophisticated strategies have been developed for deciding which principal components to include in the regression equation—see, for example, Hill *et al.* (1977). Is there any sense in which moving away from $\alpha = 1$ has a similar effect to these more sophisticated strategies?

My second point concerns data, such as those in the later examples, where the large number of variables represents a discrete set of points along a continuous curve. The proposed family of techniques does not appear to take any account of the underlying continuity. Would it not be preferable to incorporate the knowledge of continuity in some way, such as that proposed by Ramsay (1982)?

Professor Bruce R. Kowalski and Mary Beth Seasholtz (University of Washington, Seattle): The authors propose a new multivariate regression method based on the cross-validated selection of two parameters, the number of latent variables or regressors and α, a real number in the interval $[0, 1]$. The selection of the former parameter is in common with most biased regression methods. The parameter α, however, leads to the name of the proposed method, continuum regression, and the theory behind the method presented by the authors places three popular methods, ordinary least squares (OLS), principal components regression (PCR) and partial least squares (PLS) on the same continuum making them special cases of the authors' method.

For convenience, the authors switch from α to γ $[0, \infty]$ which puts OLS at $\gamma = 0$, PLS at $\gamma = 1$ and PCR at $\gamma = \infty$. It is not surprising that their algorithm has problems at the two extreme values. Nevertheless, they do manage to find optimal values on some examples that are quite convincing.

We wish to amplify the similarity of the authors' continuum regression to our unnamed method described in Lorber *et al.* (1987). Our method is also based on a continuum, where the power multiplying

the matrix in the power method varies in $[0, \infty]$. Like the author's γ OLS is at zero, PLS at unity and PCR at infinity. The authors' method maximizes T in equation (12) while ours minimizes the standard error of prediction.

The authors' formulation is based on the original data for all regressors as is ours for the first latent variable. Thereafter the similarity ends as we orthogonalize after each regressor is calculated and then select the next using the residual matrix. It will be interesting to see whether the two methods are the same for all values of γ.

Harald Martens (Unscrambler AS, Aas): I congratulate Professor Stone and Dr Brooks on their continuum regression (CR) paper. The CR method is a more controllable extension between ordinary least squares (OLS) and principal components regression (PCR) than partial least squares regression (PLSR), where the balance is purely data driven (Martens and Næs, 1989).

PLSR has the property that, when $\mathbf{X}'\mathbf{X}$ lacks dominant eigenvalues, then PLSR will come close to OLS, and each PLS factor uses up some degrees of freedom in y. The same is probably true for CR with $\alpha < 1$. Hopefully the number of degrees of freedom is more easy to estimate in CR than in PLSR.

I should like to point out, as a chemometrician, a way in which CR and the other bilinear regression methods may be made more useful in practice. When experts in a domain analyse their empirical data, it is important for them to find an adequate balance between data-driven and knowledge-driven modelling. Today's bilinear regression methods do not accommodate *a priori* knowledge sufficiently well.

The most important aspect of the bilinear regression methods is *how the directions c_g in X space are determined.* We can use several different criteria of 'interestingness' (Hastie and Tibshirani, 1986) such as **X**–**Y** covariances, **X**–**X** eigenvectors and combinations such as that of CR, as well as various robust, non-linear or entropy-oriented criteria. For each factor g, these various candidate vectors for **X** direction c_g can be joined with various candidate vectors expected *a priori* to be reliable and relevant to y (e.g. a previous bilinear model) into a matrix of interestingness M_g spanning the 'signal covariance'. Likewise, various **X** directions that the final c_g should avoid can form a matrix of 'uninterestingness' N_g spanning the 'noise covariance' for each factor. These directions could, for example, be estimated as eigenvectors of between-replicates covariances, or from *a priori* knowledge about undesired **X** interferences. The vectors in M_g and N_g can further be orthogonalized to previous factors, smoothed etc., and then scaled according to their *a priori* expected importance. When analysed together, e.g. as the first component in a generalized least squares principal component analysis, matrices M_g and N_g can yield a more flexible and informative determination of **X** direction c_g than any of the above criteria alone.

In practical data analysis, the mental model developing in the scientist's mind is more important than any mathematical model. Multivariate data analytic methods require a balance between predictive ability and interpretability. Cross-validated CR provides such a desired balance.

The **authors** replied later, in writing, as follows.

The discussants have done far more than fill in some of the gaps in our account of continuum regression (CR) and its relationships. They have reinforced and deepened our own tentative understanding of the technique, for which we are grateful.

It might have disarmed some of Professor Brown's and Dr Fearn's cogent criticisms, if the paper had truthfully labelled CR as a superficial, non-scientific, empirical prediction method that should always give way to any hardening of the scientific context in which it may be used. That is what we meant by 'soft'. Naturally our censorious inner statisticians do not warm to such an uninspiring technique, craving rather the excitements of assumptive modelling. But is there not a need for interim procedures that do not grossly mislead the ordinary user? Especially when based on a 'representative' sample, CR should help to meet this need, based as it is on cross-validatory choice of only two control parameters.

Professor Brown *could* be right to suggest that partial least squares (PLS), principal components regression (PCR) and their variants together do all that is necessary—for the world as we find it, it might be added. If so, then CR may have served some purpose in drawing the attention of more statisticians to the widely neglected technique of PLS, as well as in reiterating the value of the Moore–Penrose generalized inverse in some contexts.

We are sorry that Professor Brown could not accept our gross summary of the shape of his original, elegant contribution (Brown, 1982) to the problem, namely that, for multivariate calibration, linear modelling starts with $\dot{\mathbf{x}}$ on y (our notation) but may with advantage under extra assumptions lead to

y on $\dot{\mathbf{x}}$. Provoked to a second reading of Sundberg and Brown (1988), we have to question the second author's present interpretation of what was there proved for the $n-1<p$ case: the equivalence referred to is, we think, to be found only after canonical transformation which, in the singular case, is not the innocuous mathematical device that it is for $n \geqslant p+1$.

Our only quibble with Dr Fearn's canonical transformation, which greatly simplifies and illuminates the comparison of CR with ridge regression, is that it disrupts the CR format to conjoin a transformation of $(y_1, \ldots, y_n)$: only the transformation to the principal component variables $\mathbf{v}_j'\dot{\mathbf{x}}$, $j=1, \ldots, p$, is needed for the comparison.

Let us hope that the interesting work described by Mr Denham is published. Will the world as we find it conform to the prior beliefs that have to be built into the spline method and its conceptually awkward knots? Only detailed case studies will answer that question.

The 'estimation' posited by Dr Hjorth is presumably of the values of α and ω that minimize $\mathrm{MSEP}(\alpha, \omega)=E\{Y-\hat{y}(\mathbf{X}; \alpha, \omega)\}^2$ where $(\mathbf{X}, Y)$ is a random $(\mathbf{x}, y)$ and $\hat{y}(\mathbf{x}; \alpha, \omega)$ is the CR prediction of y at $\mathbf{x}$ for choice (α, ω). Boundless bootstrappers would presumably use simulation to estimate $\mathrm{MSEP}(\alpha, \omega)$ after giving $(\mathbf{X}, Y)$ the empirical distribution of $(\mathbf{x}_i, y_i)$, $i=1, \ldots, n$. No problem!—but the message of our penultimate paragraph in Section 11 is some defence against the bootstrappers' bolder claims. (We are genuinely puzzled by Dr Hjorth's figure, since there is a free choice of ω at every value of α.)

In reply to Professor Atkinson, a necessarily modest form of 'undue influence' analysis would be to exclude the item with the largest one-out residual (or some multiple of it that allows for 'due influence'), and then to see whether there is any serious change in the CR predictor for the previously optimal values of α and ω. If so, the optimum would be recalculated on the remaining $n-1$ items, and the outlier procedure repeated.

Dr Farebrother's legitimate concern for 'low order' principal components should be alleviated by the fact that higher order components pay a relatively small cross-validatory price to be included. Hence valuable low order components do have a good chance of being brought into play.

We guess that Dr Næs might agree on that, since he goes so far as to suggest that cross-validated PCR should be enough—outbidding Professor Brown. We should mention here that the user of the CR program has the option of a print-out of the constructed 'component' vectors $\mathbf{c}_1, \ldots, \mathbf{c}_w$ for any w, $0 \leqslant w \leqslant \omega$, to await interpretation. But would anyone not addicted to difficult or impossible interpretations want to use PCR just because, in the CR framework, it usually generates more components than the OLS end of the spectrum? All those components may be just trying to simulate a single meaningful component that would have emerged intact near OLS. This is just one reason why we would like CR to have its day in court.

It would be nice if Dr Ståhle's 'prediction-optimized PLS' became popular, but we fear that his cross-validatory choice of as many as p control parameters may be misguided. As Professor Brown noted, choice is not the same as assessment. For $n<20$, the control parameters should be few in number, unless statistical noise is very low.

The 'ρ' of Professor Chan and Professor Mak is the same as ours, so their equation for it is probably some rearrangement of our equation (23). But their version involves a $p \times p$ matrix $\mathbf{S}$ which, for large p as in example 5, generates much more calculation and storage than our method, in which $\mathbf{c}$ is calculated from the m $(=n-1 \ll p)$ n-dimensional eigenvectors of the $n \times n$ matrix $\mathbf{XX}'$ and the $m \times m$ matrix $\mathbf{M}$ (or $\mathbf{M}^*$). Furthermore our use of expression (27) for T means that we do not have to calculate $\mathbf{c}$ itself during cross-validatory choice. Our own search for optimal ρ was initially based on solution of equation (23), but existence of multiple roots meant that subdivision of the interval for ρ was needed in the use of the library numerical routine to ensure that no roots were missed. Moreover, at each stage, the matrix $\mathbf{M}$ (or $\mathbf{M}^*$) had to be computed from its definition (22) (or (32)) for each value of ρ called by the routine. The fixed grid method avoids this by use of the recurrence relation for $\mathbf{M}$ (or $\mathbf{M}^*$) between successive stages. Because the relations for $\mathbf{M}$ and $\mathbf{M}^*$ have the same form, the computational work in finding an optimal ρ on the grid is similar in main and cross-validatory runs. For all these reasons, we think that the simplification and saving envisaged by Professor Chan and Professor Mak will not be attainable.

We hope that we will be able to see the details of the alternative ideas of Dr Frank and Professor Friedman. Were the simulations referred to chosen on empirical grounds? It is empirical evidence, possibly translated into coarse models, that must underpin final judgments of competing techniques.

Professor Helland drops us a novel morsel from what is probably a well-stocked table. Does his remark that 'further studies are definitely needed' mean that we may look forward to a plateful?

Broadly speaking, the answer to Dr Jolliffe's first question is yes. All the sophisticated variants of

PCR, including the latent root regression mentioned by Dr Farebrother, aim to involve the y-values in the choice of components. This is what taking $\alpha < 1$ does. We do not yet know of any more specific similarity, however. As for Dr Jolliffe's second question, the 'continuity' that Ramsay (1982) deals with is the denseness of the indexing set for the explanatory variables, e.g. $x(t)$, $0 \leqslant t \leqslant 1$. Ramsay is not primarily concerned with the continuity of $x(t)$ as a function of t, ideas about which do have implications (see Brown and Denham (1989)) for the way that we might handle the discretization of t that is necessary unless we use analogue computers. Ramsay would like statisticians to learn functional analysis as an aid to thinking about the continuous limit.

We hope that there is no monotone transformation from γ to the 'power' of the PLS variant of Professor Kowalski and his colleagues that would make the two methods the same. For then we would have to try to prove mathematically what would be a most remarkable equivalence. Similarity rather than sameness is a possibility. Professor Kowalski and Dr Martens are leading experts in the new multivariate analysis, and we are particularly intrigued by Dr Martens's proposals for incorporating prior knowledge into the CR mechanism.

REFERENCES IN THE DISCUSSION

Atkinson, A. C. (1985) *Plots, Transformations and Regression*. Oxford: Clarendon.

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

Brown, P. J. and Denham, M. C. (1989) Calibration with many variables. To be published.

Brown, P. J. and Zidek, J. V. (1980) Adaptive multivariate ridge regression. *Ann. Statist.*, **8**, 64–74.

Cook, R. D. and Weisberg, S. (1982) *Residuals and Influence in Regression*. London: Chapman and Hall.

Frank, I. E. (1989) A comparative Monte Carlo study of biased regression techniques. *Technical Report LCS 105*. Department of Statistics, Stanford University.

Goldstein, M. and Smith, A. F. M. (1974) Ridge-type estimators for regression analysis. *J. R. Statist. Soc.* B, **36**, 284–291.

Hastie, T. and Tibshirani, R. (1986) Generalized additive models. *Statist. Sci.*, **1**, 297–318.

Helland, I. S. (1988) On the structure of partial least squares regression. *Communs Statist. Simuln*, **17**, 581–607.

——— (1989) Partial least squares regression and statistical models. *Scand J. Statist.*, to be published.

Hill, R. C., Fomby, T. B. and Johnson, S. R. (1977) Component selection norms for principal components regression. *Communs Statist.* A, **6**, 309–334.

Hjorth, U. (1989) On model selection in the computer age. *J. Statist. Planng Inf.*, **23**, 101–115.

Hjorth, U. and Holmqvist, L. (1981) On model selection based on validation with applications to pressure and temperature prognosis. *Appl. Statist.*, **30**, 264–274.

Höskuldsson, A. (1988) PLS regression methods. *J. Chemomet.*, **2**, 211–220.

Jolliffe, I. T. (1982) A note on the use of principal components in regression. *Appl. Statist.*, **31**, 300–303.

Lorber, A., Wangen, L. E. and Kowalski, B. R. (1987) A theoretical foundation for the PLS algorithm. *J. Chemomet.*, **1**, 19–31.

Martens, H. and Næs, T. (1989) *Multivariate Calibration*. Chichester: Wiley.

Mason, R. L. (1986) Latent root regression: a biased regression methodology for use with collinear predictor variables. *Communs Statist.* A, **15**, 2651–2678.

Næs, T. (1987) PLS versus some other statistical calibration methods. In *Proc. Sem. PLS Data Approximation*, May 19th (ed. M. Martens).

Ramsay, J. O. (1982) When the data are functions. *Psychometrika*, **47**, 379–396.

Smith, P. L. (1979) Splines as a useful and convenient statistical tool. *Am. Statistn*, **33**, 57–62.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc.* B, **36**, 111–147; corrigendum, **38** (1976), 102.

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

Westlake, J. R. (1968) *A Handbook of Numerical Matrix Inversion and Solution of Linear Equations*, p. 47. New York: Wiley.