J. R. Statist. Soc. B (1997)
59, *No.* 1, *pp.* 3–54

Predicting Multivariate Responses in Multiple Linear Regression

By LEO BREIMAN and JEROME H. FRIEDMAN†

University of California, Berkeley, USA *Stanford University, USA*

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, February 14th, 1996*, Professor R. L. Smith *in the Chair*]

SUMMARY

We look at the problem of predicting several response variables from the same set of explanatory variables. The question is how to take advantage of correlations between the response variables to improve predictive accuracy compared with the usual procedure of doing individual regressions of each response variable on the common set of predictor variables. A new procedure is introduced called the curds and whey method. Its use can substantially reduce prediction errors when there are correlations between responses while maintaining accuracy even if the responses are uncorrelated. In extensive simulations, the new procedure is compared with several previously proposed methods for predicting multiple responses (including partial least squares) and exhibits superior accuracy. One version can be easily implemented in the context of standard statistical packages.

Keywords: CANONICAL REGRESSION; CROSS-VALIDATION; CURDS AND WHEY METHOD; PARTIAL LEAST SQUARES; RANK REGRESSION; RIDGE REGRESSION

1. INTRODUCTION

Increasingly, there are applications where several quantities are to be predicted using a common set of predictor variables. For instance, in a manufacturing process we may want to predict various quality aspects of a product from the parameter settings used in the manufacturing. Or, given the mass spectra of a sample, the goal may be to predict the concentrations of several chemical constituents in the sample.

Some years ago, the authors were involved in a project trying to predict changes in the valuations of the stocks in 60 industry groups by using over 100 econometric variables as predictors. In our state of knowledge at that time, prediction equations for each one of the 60 groups were derived not using the data on the other 59 responses. However, the changes in the 60 groups were strongly correlated. If we knew then what we know now, we could have taken advantage of the correlations to produce more accurate predictors.

To give a simple example of the potential improvement in estimation, suppose that the data are of the form $\{y_{n1}, y_{n2}, \mathbf{x}_n\}_1^N$ where each $\mathbf{x}_n = (x_{n1}, \ldots, x_{np})$ is a p-vector of predictor variables and there are two responses y_1 and y_2. Taking the usual path, we derive predictors for y_1 and y_2 by doing separate regressions on $(x_1, \ldots, x_p)$, i.e. the estimated regression coefficients $\hat{\mathbf{a}}_1 = (\hat{a}_{11}, \ldots, \hat{a}_{1p})$ and $\hat{\mathbf{a}}_2 = (\hat{a}_{21}, \ldots, \hat{a}_{2p})$ are solutions to

$$\hat{\mathbf{a}}_1 = \arg \min_{\mathbf{a}} \left\{ \sum_{n=1}^{N} (y_{n1} - \mathbf{a}^{\mathrm{T}} \mathbf{x}_n)^2 \right\},$$

†*Address for correspondence*: Department of Statistics and Stanford Linear Accelerator Center, Stanford University, Stanford, CA 94305, USA.
E-mail: jhf@playfair.stanford.edu

 0035–9246/97/59003

$$\hat{\mathbf{a}}_2 = \arg\min_{\mathbf{a}} \left\{ \sum_{n=1}^{N} (y_{n2} - \mathbf{a}^{\mathrm{T}}\mathbf{x}_n)^2 \right\}$$

where all variables have been centred. The prediction equations for y_1 and y_2 are $\hat{y}_1(\mathbf{x}) = \bar{y}_1 + \hat{\mathbf{a}}_1^{\mathrm{T}}(\mathbf{x} - \bar{\mathbf{x}})$ and $\hat{y}_2(\mathbf{x}) = \bar{y}_2 + \hat{\mathbf{a}}_2^{\mathrm{T}}(\mathbf{x} - \bar{\mathbf{x}})$ where $(\bar{y}_i, \bar{\mathbf{x}})$ are the corresponding sample means (before centring). Now suppose further that the (unknown) truth happens to be $y_{n1} = b_{10} + \mathbf{b}^{\mathrm{T}}\mathbf{x}_n + \epsilon_{n1}$ and $y_{n2} = b_{20} + \mathbf{b}^{\mathrm{T}}\mathbf{x}_n + \epsilon_{n2}$ where $\{\epsilon_{n1}\}_1^N$ and $\{\epsilon_{n2}\}_1^N$ are independent and identically distributed (IID) $N(0, \sigma^2)$. Here y_1 and y_2 are correlated because they have the same dependence on the predictor variables, $\mathbf{b}^{\mathrm{T}}\mathbf{x}$. It is also clear that accuracy is improved for *each* of the two responses by using the predictors

$$\tilde{y}_i = \bar{y}_i + \tfrac{1}{2}(\hat{y}_1 - \bar{y}_1) + \tfrac{1}{2}(\hat{y}_2 - \bar{y}_2) \qquad (i = 1, 2),$$

instead of $\hat{y}_1$ and $\hat{y}_2$ respectively.

1.1. *Curds and Whey Procedure*

In general, if there are q responses $\mathbf{y} = (y_1, \ldots, y_q)$ with separate least squares regressions $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_q)$, then the above example raises the possibility that if the responses are correlated we may be able to obtain a more accurate predictor $\tilde{y}_i$ of each y_i by using a linear combination

$$\tilde{y}_i = \bar{y}_i + \sum_{k=1}^{q} b_{ik}(\hat{y}_k - \bar{y}_k), \qquad i = 1, \ldots, q, \tag{1.1}$$

of the ordinary least squares (OLS) predictors

$$\hat{y}_i = \bar{y}_i + \sum_{j=1}^{p} \hat{a}_{ij}(x_j - \bar{x}_j), \tag{1.2}$$

$$\{\hat{a}_{ij}\}_{j=1}^{p} = \arg\min_{\{a_j\}_1^p} \left[\sum_{n=1}^{N} \left\{ y_{ni} - \bar{y}_i - \sum_{j=1}^{p} a_j(x_{nj} - \bar{x}_j) \right\}^2 \right], \tag{1.3}$$

rather than with the least squares predictors themselves. Note that equations (1.1) and (1.2) imply that the coefficients, but not the means, of the (OLS) estimates are modified.

To simplify the notation in all derivations that follow, we assume that the response and predictor variables are all centred by their corresponding training sample means $\{y_i \leftarrow y_i - \bar{y}_i\}_1^q$, $\{x_j \leftarrow x_j - \bar{x}_j\}_1^p$. As a result, all response estimates are centred at the corresponding response sample means $\{\hat{y}_i \leftarrow \hat{y}_i - \bar{y}_i\}_1^q$, $\{\tilde{y}_i \leftarrow \tilde{y}_i - \bar{y}_i\}_1^q$, and reference the centred predictor variables.

Assuming that equation (1.1) is an interesting possibility, the trick is to find what $\{b_{ik}\}$ to use. It turns out that there is a nearly optimal set of $\{b_{ik}\}$ that are given by what we call the 'curds and whey' (C&W) procedure. Using vector and matrix notation for the respective (centred) quantities

$$\tilde{\mathbf{y}} = \{\tilde{y}_i\}_1^q, \qquad \hat{\mathbf{y}} = \{\hat{y}_i\}_1^q, \qquad \mathbf{B} = [b_{ik}] \in R^{q \times q}, \tag{1.4}$$

equation (1.1) can be expressed as

$$\tilde{\mathbf{y}} = \mathbf{B}\hat{\mathbf{y}}. \tag{1.5}$$

We derive estimates of the matrix $\mathbf{B}$ that take the form $\mathbf{B} = \mathbf{T}^{-1}\mathbf{D}\mathbf{T}$ where $\mathbf{T}$ is the $q \times q$ matrix whose rows are the response canonical co-ordinates (see Section 2.2) and $\mathbf{D} = \text{diag}(d_1, \ldots, d_q)$ is a diagonal matrix. Two prescriptions are derived for calculating $\{d_k\}_1^q$. A generalized cross-validation (GCV) approach (Section 3.1) yields a simple formula (3.12)–(3.13). This works surprising well. Using regular (fivefold or tenfold) cross-validation (Section 3.2) to obtain the $\{d_k\}_1^q$ gives slightly better prediction.

1.2. *Statistical Background*

The C&W procedure is a form of multivariate shrinking. It transforms ($\mathbf{T}$), shrinks (multiplies by $\mathbf{D} = \{d_k\}_1^q$) and then transforms back ($\mathbf{T}^{-1}$). It derives its power by shrinking in the right co-ordinate system (canonical co-ordinates) and can be viewed as a multivariate generalization of proportional shrinkage based on cross-validation (Stone, 1974).

For a single response variable ($q = 1$) it is well known that the OLS estimate (1.2)–(1.3) can be outperformed in terms of prediction accuracy by biased (regularized) shrinkage estimates. Examples include proportional shrinkage (James and Stein, 1961; Stone, 1974; Copas, 1983, 1987), ridge regression (Hoerl and Kennard, 1970), principal components regression (Massey, 1965) and partial least squares (PLS) regression (Wold, 1975). These results suggest that there may be gains associated with treating the collection of responses as a vector-valued variable in the context of a combined shrinkage estimation procedure. Such procedures have been proposed: reduced rank regression (Izenman, 1975), two-block PLS (Wold, 1975), filtered canonical y-variate regression (FICYREG) (van der Merwe and Zidek, 1980) and multivariate forms of ridge regression (Brown and Zidek, 1980, 1982). However, they have seen little use in statistical practice. An exception is two-block PLS which is widely applied in chemometrics. The C&W method differs from the methods cited in that it has roots in both a theoretical and a cross-validation foundation. Furthermore, simulation results indicate that its performance exceeds that of the several (previous) methods with which we have compared it.

1.3. *Outline of Paper*

In Section 2 we assume (centred) predictors of the form (1.5). Taking the data to be generated from linear models plus noise, we derive (under idealized conditions) the optimal shrinkage matrix $\mathbf{B}^* = \mathbf{T}^{-1}\mathbf{D}\mathbf{T}$ where $\mathbf{T}$ is the canonical transformation and $\mathbf{D}$ is a diagonal 'shrinking' matrix. However, because of the idealized setting, the matrix $\mathbf{D}$ derived there underestimates the amount of shrinkage that is necessary. Section 3 takes a cross-validation approach to estimation of the shrinkage factors, derives a simple approximate formula and then describes the V-fold cross-validation estimates of $\mathbf{D}$.

Section 4 gives a brief description of some other methods proposed for estimating multiple responses, and these are compared with C&W in the simulation study covered in Section 5. In some fields, chemometrics for example, it is quite usual to

have fewer observations than predictor variables ($N \prec p$). The C&W method can be extended to these underdetermined systems. This procedure is described in Section 6, together with results of another simulation comparing prediction methods in this $p \succ N$ situation. Section 7 illustrates the application of the C&W method to two published data sets, one from chemometrics and the other from Scottish election results. Section 8 gives concluding remarks.

2. MULTIVARIATE PROPORTIONAL SHRINKAGE

For a single response variable y, the (centred) proportional shrinkage estimate $\tilde{y}$ can be expressed as

$$\tilde{y} = b\hat{y} = \sum_{j=1}^{p} (b\hat{a}_j)x_j \tag{2.1}$$

where $\hat{y}$ and $\{\hat{a}_j\}_1^p$ are the OLS estimates (1.2)–(1.3). Each OLS coefficient $\hat{a}_j$ is scaled by the same factor b and the overall biased estimate is a linear function of the OLS solution $\hat{y}$. Several prescriptions have been proposed for estimating the degree of shrinkage (value for b) to obtain improved expected mean-squared error

$$E(y - \tilde{y})^2 < E(y - \hat{y})^2, \tag{2.2}$$

where the expected value is over the joint distribution $F(\mathbf{x}, y)$ of the predictors $\mathbf{x}$ and the response y (see James and Stein (1961), Stone (1974) and Copas (1983, 1987)).

A natural extension of equation (2.1) to the multivariate setting is to express each biased estimate $\tilde{y}_i$ as a general linear function (1.1) of the OLS estimates $\{\hat{y}_i\}_1^q$ (1.2). In vector notation (1.4) this is expressed by equation (1.5) where $\mathbf{B}$ can be regarded as a shrinking matrix that transforms the (vector-valued) OLS estimate $\hat{\mathbf{y}}$ to the biased estimate $\tilde{\mathbf{y}}$. The goal is to obtain an estimate $\mathbf{B}$ of the optimal shrinking matrix $\mathbf{B}^*$ whose elements are defined by

$$\{b_{ik}^*\}_{k=1}^q = \arg\min_{\{\beta_k\}_1^q} \left[E\left\{ y_i - \sum_{k=1}^{q} \beta_k \hat{y}_k \right\}^2 \right], \qquad i = 1, \ldots, q. \tag{2.3}$$

Here the expected value is over the joint distribution $F(\mathbf{x}, \mathbf{y})$ of the data to be predicted. Note that the use of $\mathbf{B}^*$ (2.3) in equation (1.5) will result in reduced mean-squared prediction error for *each* response

$$E\{y_i - (\mathbf{B}^*\hat{\mathbf{y}})_i\}^2 \leqslant E\{y_i - \hat{y}_i\}^2, \qquad i = 1, \ldots, q, \tag{2.4}$$

with equality in expression (2.4) obtaining only in the (unlikely) event $\mathbf{B}^* = \mathbf{I}_q$, where $\mathbf{I}_q$ is the $q \times q$ identity matrix. Therefore, expected (squared error) loss will be reduced for every response individually, rather than only with respect to an amalgamated loss criterion involving all the responses (such as weighted average quadratic loss).

2.1. *Optimal Proportional Shrinkage*

To gain insight into the nature of the problem and its solution, we derive the optimal shrinking matrix $\mathbf{B}^*$ in an idealized setting. Here we assume that each

response is a linear function of the predictors with additive (IID) error

$$y_i = f_i(\mathbf{x}) + \epsilon_i, \tag{2.5}$$

with

$$f_i(\mathbf{x}) = \sum_{j=1}^{p} a_{ij} x_j, \qquad i = 1, \ldots, q. \tag{2.6}$$

The predictors $\mathbf{x} \in R^p$ and the errors $\epsilon \in R^q$ are random samples with respective (population) distributions $F_x(\mathbf{x})$ and $F_\epsilon(\epsilon)$ with their joint distribution given by

$$F(\mathbf{x}, \epsilon) = F_x(\mathbf{x})\, F_\epsilon(\epsilon), \tag{2.7}$$

i.e. the errors are independent of the predictor variables. Let

$$\left.\begin{aligned} E(\mathbf{x}) = E(\epsilon) &= 0, \\ E(\mathbf{x}\mathbf{x}^{\mathrm{T}}) &= \mathbf{V} \in R^{p \times p}, \\ E(\epsilon\epsilon^{\mathrm{T}}) &= \mathbf{\Sigma} \in R^{q \times q} \end{aligned}\right\} \tag{2.8}$$

where the expected values are over the joint distribution (2.7). In this setting the errors are assumed to be independent between (random) observations, but (possibly) correlated among the responses for each observation.

The solution to equation (2.3) is a least squares regression (through the origin) of each response y_i on the (sample-based) OLS estimates $\{\hat{y}_i\}_1^q$ over the (population) distribution (2.7),

$$\mathbf{B}^* = E(\mathbf{y}\hat{\mathbf{y}}^{\mathrm{T}})\, E(\hat{\mathbf{y}}\hat{\mathbf{y}}^{\mathrm{T}})^{-1}. \tag{2.9}$$

To simplify this derivation (only) we further assume that the sample means and covariance matrix of the predictor variables are the same as that of the population distribution. This would be the case if we condition on the design and only the errors are random. Otherwise, this can be viewed as a simplifying approximation. Denoting the 'signal' covariance matrix as

$$\mathbf{F} = E\{\mathbf{f}(\mathbf{x})\mathbf{f}^{\mathrm{T}}(\mathbf{x})\} = \mathbf{A}\mathbf{V}\mathbf{A}^{\mathrm{T}} \tag{2.10}$$

where $\mathbf{f}(\mathbf{x}) = \{f_i(\mathbf{x})\}_1^q$, and $\mathbf{A} \in R^{q \times p}$ is the matrix of (true) coefficients $\{a_{ij}\}$ (2.6), we have

$$\begin{aligned} E(\hat{\mathbf{y}}\hat{\mathbf{y}}^{\mathrm{T}}) &= \mathbf{F} + r\mathbf{\Sigma}, \\ E(\mathbf{y}\hat{\mathbf{y}}^{\mathrm{T}}) &= F \end{aligned} \tag{2.11}$$

where

$$r = p/N \tag{2.12}$$

is the ratio of the number of predictor variables to the training sample size. Therefore, from equation (2.9),

$$\mathbf{B}^* = \mathbf{F}(\mathbf{F} + r\mathbf{\Sigma})^{-1} = (\mathbf{I}_q + r\mathbf{R})^{-1} \tag{2.13}$$

where

$$\mathbf{R} = \mathbf{\Sigma}\mathbf{F}^{-1} \tag{2.14}$$

is the 'noise-to-signal' matrix. This result shows that the optimal shrinking matrix $\mathbf{B}^*$ is determined by the noise-to-signal structure in the response space as reflected by the matrix $\mathbf{R} \in R^{q\times q}$. Since both $\mathbf{\Sigma}$ and $\mathbf{F}$ are unknown this result is of no direct use except to illustrate that they need not be separately determined; only an estimate of the product (2.14) is required. In the next section we show that $\mathbf{R}$ is related to the canonical co-ordinates of the joint distribution of the predictors and responses.

2.2. *Canonical Analysis*

In terms of a population distribution, canonical analysis can be formulated as follows. Let $F(\mathbf{x}, \mathbf{y})$ be the (population) joint distribution of the (population centred) predictors $\mathbf{x}$ and the responses $\mathbf{y}$. The goal is to find vectors $\mathbf{t} \in R^q$ and $\mathbf{v} \in R^p$ such that the correlation between the linear combinations $\mathbf{t}^{\mathrm{T}}\mathbf{y}$ and $\mathbf{v}^{\mathrm{T}}\mathbf{x}$ is maximized. More generally, canonical analysis seeks $K = \min(p, q)$ such pairs of linear combinations such that each successive pair maximizes correlation under the constraint of being uncorrelated with the previous pairs

$$(\mathbf{t}_k, \mathbf{v}_k) = \arg \max_{\substack{\{\mathrm{corr}(\mathbf{t}^{\mathrm{T}}\mathbf{y},\, \mathbf{t}_l^{\mathrm{T}}\mathbf{y})=0\}_1^{k-1} \\ \{\mathrm{corr}(\mathbf{v}^{\mathrm{T}}\mathbf{x},\, \mathbf{v}_l^{\mathrm{T}}\mathbf{x})=0\}_1^{k-1}}} \{\mathrm{corr}(\mathbf{t}^{\mathrm{T}}\mathbf{y}, \mathbf{v}^{\mathrm{T}}\mathbf{x})\}. \tag{2.15}$$

The vectors $\{\mathbf{t}_k\}_1^K$ and $\{\mathbf{v}_k\}_1^K$ are (respectively) called the $\mathbf{y}$ and $\mathbf{x}$ canonical co-ordinates, and their respective correlations

$$\{c_k = \mathrm{corr}(\mathbf{t}_k^{\mathrm{T}}\mathbf{y}, \mathbf{v}_k^{\mathrm{T}}\mathbf{x})\}_1^K \tag{2.16}$$

are known as the canonical correlations of $F(\mathbf{x}, \mathbf{y})$. The criterion (2.15) is invariant to, and thus does not restrict, the scales of the linear combinations; this ambiguity is usually resolved by standardizing them all to have unit variances

$$E(\mathbf{t}_k^{\mathrm{T}}\mathbf{y})^2 = E(\mathbf{v}_k^{\mathrm{T}}\mathbf{x})^2 = 1, \qquad k = 1, \ldots, K. \tag{2.17}$$

It is well known (see for example Anderson (1957)) that the solutions to equations (2.15) and (2.17) for $\{\mathbf{t}_k\}_1^K$ are obtained from an eigenanalysis of the $q \times q$ matrix

$$\mathbf{Q} = E(\mathbf{y}\mathbf{y}^{\mathrm{T}})^{-1}E(\mathbf{y}\mathbf{x}^{\mathrm{T}})E(\mathbf{x}\mathbf{x}^{\mathrm{T}})^{-1}E(\mathbf{x}\mathbf{y}^{\mathrm{T}}) = \mathbf{T}^{\mathrm{T}}\mathbf{C}^2\mathbf{T}^{-\mathrm{T}} \in R^{q\times q}. \tag{2.18}$$

(Although $\mathbf{Q}$ is not symmetric, it is the product of two symmetric matrices, so that the eigendecomposition (2.18) exists and is straightforward to obtain (see Golub and van Loan (1989)).) The rows of the $q \times q$ matrix $\mathbf{T}$ (eigenvectors) are the $\mathbf{y}$ canonical co-ordinates $\{\mathbf{t}_k\}_1^q$ and the diagonal matrix

$$\mathbf{C}^2 = \mathrm{diag}\{c_1^2, \ldots, c_K^2\} \tag{2.19}$$

contains the respective squared canonical correlations (2.16). The $\mathbf{x}$ canonical co-ordinates are obtained by an eigenanalysis of a matrix analogous to $\mathbf{Q}$ (2.18) where $\mathbf{x}$ and $\mathbf{y}$ are interchanged.

Generally, canonical analysis is used to obtain a set of descriptive statistics for the joint distribution $F(\mathbf{x}, \mathbf{y})$. However, in our regression model (2.5)–(2.7) it provides a

means for obtaining the optimal shrinking matrix $\mathbf{B}^*$ (2.13). Under that model $\mathbf{Q}$ (2.18) becomes

$$\mathbf{Q} = (\mathbf{F} + \boldsymbol{\Sigma})^{-1}\mathbf{F} = (\mathbf{I}_q + \mathbf{R}^{\mathrm{T}})^{-1} \tag{2.20}$$

so that

$$\mathbf{B}^* = \{(1 - r)\mathbf{I}_q + r\mathbf{Q}^{-\mathrm{T}}\}^{-1} \tag{2.21}$$

where is r given by equation (2.12). This result (2.21) shows that $\mathbf{B}^*$ is diagonal in the $\mathbf{y}$ canonical co-ordinate system (2.18)

$$\mathbf{B}^* = \mathbf{T}^{-1}\mathbf{D}\mathbf{T}, \qquad \mathbf{D} = \mathrm{diag}\{d_1, \ldots, d_q\} \tag{2.22}$$

with

$$d_i = \frac{c_i^2}{c_i^2 + r(1 - c_i^2)}, \qquad i = 1, \ldots, q, \tag{2.23}$$

where by definition $\{c_i = 0\}_{K+1}^q$. Substituting equation (2.22) into equation (1.5) we have

$$\mathbf{T}\tilde{\mathbf{y}} = \mathbf{D}(\mathbf{T}\hat{\mathbf{y}}) \tag{2.24}$$

so that equation (1.5) reduces to separate proportional shrinking of each OLS solution in the $\mathbf{y}$ canonical co-ordinate system. This leads to the following prescription for optimal multivariate proportional shrinking.

(a) Transform $\mathbf{y}$ to the canonical co-ordinate system, $\mathbf{y}' = \mathbf{T}\mathbf{y}$.
(b) Perform a separate OLS regression of each y'_i on $\mathbf{x}$ ($i = 1, \ldots, q$), obtaining $\{\hat{y}'_i\}_1^q$.
(c) Separately scale (shrink) each $\hat{y}'_i$ by the factor d_i (2.23), obtaining $\tilde{\mathbf{y}}' = \{d_i\hat{y}'_i\}_1^q$.
(d) Transform back to the original $\mathbf{y}$ co-ordinate system, $\tilde{\mathbf{y}} = \mathbf{T}^{-1}\tilde{\mathbf{y}}'$.

Fig. 1 shows graphs of the canonical co-ordinate shrinkage factors d_i (2.23) as a function of the corresponding squared canonical correlations c_i^2, for various values of r (2.12). For small values of r there is very little shrinking of the OLS solutions in the canonical co-ordinate system, except for very small values of c_i^2, whereas for large values the shrinkage factor decreases roughly linearly with decreasing c_i^2. In all cases, $0 \leqslant d_i \leqslant 1$.

To estimate $\mathbf{B}^*$ (2.21) we need a sample-based estimate of $\mathbf{Q}$ (2.18). A natural choice would be the 'plug-in' estimate

$$\hat{\mathbf{Q}} = (\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{2.25}$$

where

$$\begin{aligned} \mathbf{Y} &= (y_{ni}) \in R^{N\times q}, \\ \mathbf{X} &= (x_{nj}) \in R^{N\times p}, \end{aligned} \tag{2.26}$$

are the respective (centred) data matrices. Although this choice improves the OLS estimates, it does not provide enough shrinkage, and more improvement is possible.

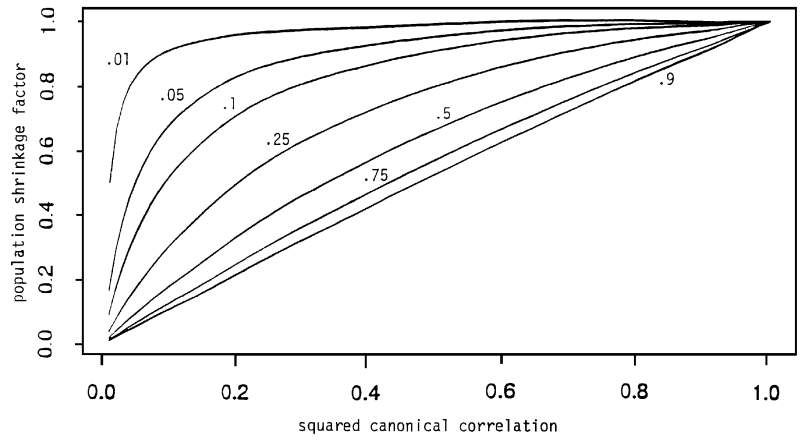

Fig. 1. Population canonical co-ordinate shrinkage factors (2.23) as a function of squared (population) canonical correlation, for various ratios of parameter-to-observation count

The reason is that the sample canonical correlations $\{\hat{c}_i\}_1^q$ overestimate their corresponding population values $\{c_i\}_1^q$ (2.19) so that using these sample-based estimates in equation (2.23) reduces the amount of shrinkage from that which would be obtained by using the correct (unbiased) population values. The problem is that the same sample is used to estimate both the OLS solution and its goodness of fit as reflected by the inflated (resubstitution) $\{\hat{c}_i\}_1^q$-values. This is a common problem in model selection. To estimate the proper amount of shrinkage a better (less biased) estimate of goodness of fit is needed. One commonly used method for this is cross-validation (Stone, 1974).

3. CROSS-VALIDATORY MULTIVARIATE SHRINKAGE

The optimal shrinking matrix **B*** (2.3) is obtained by a regression of the responses $\{y_i\}_1^q$ on the (sample-based) OLS estimates $\{\hat{y}_i\}_1^q$ over (all future) data that are not part of the training sample. This procedure can be approximated through cross-validation. Each observation $(\mathbf{y}_n, \mathbf{x}_n)$ is (in turn) removed from the training sample and treated as a 'future' observation. The corresponding (cross-validation) analogue to equation (2.3) then becomes

$$\{b_{ik}\}_{k=1}^q = \arg\min_{\{\beta_k\}_1^q} \left\{ \sum_{n=1}^N \left(y_{ni} - \sum_{k=1}^q \beta_k \hat{y}_{\backslash nk} \right)^2 \right\}, \qquad i = 1, \ldots, q, \tag{3.1}$$

where $\hat{y}_{\backslash nk}$ is the OLS prediction of the kth response for the nth observation, obtained with it removed from the training sample. For a single response ($q = 1$) this approach was proposed by Stone (1974) and called 'flattening'. From standard matrix updating formulae we obtain

$$\hat{\mathbf{y}}_{\backslash n} = (1 - g_n)\mathbf{y}_n + g_n \hat{\mathbf{y}}_n \tag{3.2}$$

where $\hat{\mathbf{y}}_n$ is the OLS estimate on the full sample and

$$g_n = 1/(1 - h_{nn}) \tag{3.3}$$

with $\{h_{nn}\}_1^N$ being the diagonal elements of the 'hat' matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} \in R^{N\times N}, \tag{3.4}$$

where $\mathbf{X}$ is the predictor data matrix (2.26). Substituting equation (3.2) into equation (3.1) we obtain the cross-validated estimate of the shrinking matrix $\mathbf{B}$.

3.1. *Generalized Cross-validation-based Multivariate Shrinking*

To simplify estimate (3.1) we first consider an approximation to the cross-validation procedure (3.2)–(3.4). We approximate each h_{nn} (3.3) by its average over the N observations

$$h_{nn} \approx \bar{h} = \frac{1}{N}\sum_{m=1}^{N} h_{mm} = \frac{1}{N}\,\mathrm{trace}(\mathbf{H}) = r \tag{3.5}$$

with r given by equation (2.12). This approximation is equivalent to GCV proposed by Craven and Wahba (1979). Using this approximation the solution for the elements of the shrinking matrix $\mathbf{B}$ (3.1) becomes

$$\{b_{ik}\}_{k=1}^{q} = \arg\min_{\{\beta_k\}_1^q}\left(\sum_{n=1}^{N}\left[y_{ni} - \sum_{k=1}^{q}\beta_k\{(1-g)y_{nk} + g\hat{y}_{nk}\}\right]^2\right), \qquad i = 1, \ldots, q, \tag{3.6}$$

where

$$g = 1/(1 - r). \tag{3.7}$$

The normal equations for the solution (in matrix notation) are

$$\mathbf{B}\{(1-g)\mathbf{Y}^{\mathrm{T}} + g\hat{\mathbf{Y}}^{\mathrm{T}}\}\{(1-g)\mathbf{Y} + g\hat{\mathbf{Y}}\} = (1-g)\mathbf{Y}^{\mathrm{T}}\mathbf{Y} + g\mathbf{Y}^{\mathrm{T}}\hat{\mathbf{Y}} \tag{3.8}$$

where $\mathbf{Y}$ is the response data matrix (2.26) and $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} \in R^{N\times q}$ (3.4) is the corresponding matrix of OLS predictions. After a little matrix algebra equation (3.8) reduces to

$$\mathbf{B}\{(1-g)^2\mathbf{I}_q + (2g - g^2)\hat{\mathbf{Q}}^{\mathrm{T}}\} = (1-g)\mathbf{I}_q + g\hat{\mathbf{Q}}^{\mathrm{T}} \tag{3.9}$$

where $\hat{\mathbf{Q}}$ is the sample canonical correlation matrix (2.25). Equation (3.9) shows that the solution $\mathbf{B}$ is a diagonal matrix in the same co-ordinate system that diagonalizes $\hat{\mathbf{Q}}^{\mathrm{T}}$,

$$\hat{\mathbf{Q}}^{\mathrm{T}} = \hat{\mathbf{T}}^{-1}\hat{\mathbf{C}}^2\hat{\mathbf{T}}, \qquad \hat{\mathbf{C}}^2 = \mathrm{diag}\{\hat{c}_1^2, \ldots, \hat{c}_q^2\}. \tag{3.10}$$

Here $\hat{\mathbf{T}}$ is the matrix of *sample* canonical co-ordinates and $\{\hat{c}_i\}_1^q$ are the *sample* canonical correlations. Using equation (3.10) in equation (3.9) the solution for the GCV shrinkage matrix becomes

$$\mathbf{B} = \hat{\mathbf{T}}^{-1}\hat{\mathbf{D}}\hat{\mathbf{T}}, \qquad \hat{\mathbf{D}} = \mathrm{diag}\{\hat{d}_1, \ldots, \hat{d}_q\} \tag{3.11}$$

with

$$\hat{d}_i = \frac{(1-r)(\hat{c}_i^2 - r)}{(1-r)^2\hat{c}_i^2 + r^2(1-\hat{c}_i^2)}, \qquad i = 1, \ldots, q. \tag{3.12}$$

Examination of equation (3.12) shows that $\hat{d}_i$ is negative whenever $\hat{c}_i^2 < r$. As is usually done, we perform 'positive part' shrinkage in this case by setting $\hat{d}_i = 0$, so that

$$\hat{d}_i \leftarrow \max(\hat{d}_i, 0) \tag{3.13}$$

in equation (3.12).

Comparing results (3.11)–(3.13) with those of equations (2.22)–(2.24), we see that multivariate proportional shrinking based on GCV leads to the same prescription as that for (population) optimal proportional shrinking derived in Section 2.2, but with all population quantities replaced by their sample-based estimates, and using equations (3.12) and (3.13) in place of equation (2.23) for the shrinking factors in the (sample) canonical co-ordinate system. Fig. 2 shows graphs of $\hat{d}_i$ (3.12)–(3.13) as a function of the corresponding (sample) squared canonical correlations $\hat{c}_i^2$, for the same values of r as in Fig. 1. The GCV canonical shrinkage factors $\hat{d}_i$ are universally smaller valued (more shrinkage) than the corresponding population-based values d_i (2.23) (assuming $c_i = \hat{c}_i$) for all values of $\hat{c}_i^2$ and r. This compensates for the upward bias in the estimates $\{\hat{c}_i\}_1^q$ of the population values $\{c_i\}_1^q$. This effect becomes more pronounced as r increases because the GCV estimate of the upward bias becomes larger with increasing r (2.12).

Although GCV optimal shrinking (3.11)–(3.13) results in a similar prescription to that of Section 2.2, it was derived without recourse to the specific model and assumptions of Section 2.1, except for the IID assumption required for cross-validation. The validity of the GCV result rests on the suitability of equation (3.1) as an estimate of equation (2.3), and the GCV approximation (3.5). This latter approximation can be removed by the use of full cross-validation to estimate the shrinkage matrix **B**.

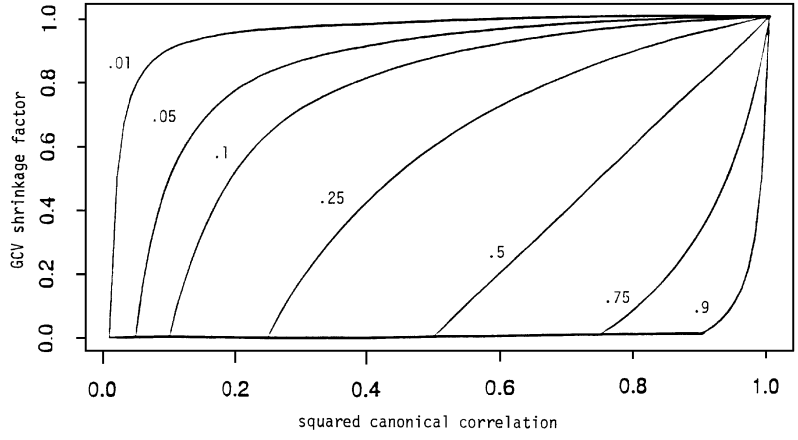

Fig. 2. Sample-based canonical co-ordinate shrinkage factors (3.12) and (3.13) as a function of squared (sample) canonical correlation, for the same ratios of parameter-to-observation count as in Fig. 1

3.2. *Fully Cross-validated Multivariate Shrinking*

As shown in Section 3.1, the GCV approximation (3.5) leads to a very simple and interpretable solution for the shrinking matrix **B** in terms of the sample canonical co-ordinates, and shrinking based on a simple formula. Resulting prediction accuracy (2.4) may be impaired, however, by the lack of validity of approximation (3.5). To overcome this, we define **B** by

$$\mathbf{B} = \hat{\mathbf{T}}^{-1}\mathbf{D}\hat{\mathbf{T}}, \qquad \mathbf{D} = \mathrm{diag}\{d_1, \ldots, d_q\}, \tag{3.14}$$

with $\hat{\mathbf{T}}$ being the sample **y** canonical co-ordinate transformation matrix (3.10) and **D** the solution to

$$\mathbf{D} = \arg \min_{\mathbf{\Delta}=\mathrm{diag}} \left[\sum_{i=1}^{q} \sum_{n=1}^{N} \{y_{ni} - (\hat{\mathbf{T}}_{\setminus n}^{-1} \mathbf{\Delta} \hat{\mathbf{T}}_{\setminus n} \hat{\mathbf{y}}_{\setminus n})_i\}^2 \right]. \tag{3.15}$$

Here the subscript $\setminus n$ on a quantity refers to that quantity calculated with the nth observation removed. Note that equation (3.15) is a purely quadratic criterion in $\mathbf{\Delta} = \mathrm{diag}\{\delta_1, \ldots, \delta_q\}$ so that the solution for **D** is unique and can be obtained by straightforward linear algebra given the other quantities appearing in equation (3.15).

For the cases studied previously (Sections 2.2 and 3.1) the solution values (2.23) and (3.12) for the canonical co-ordinate shrinkage factors were monotone functions of the respective canonical correlations. We impose a similar constraint on equations (3.14) and (3.15) by replacing the elements of **D**, $\{d_i\}_1^q$, by the closest fit to those values that are monotone in the sample canonical correlations $\{\hat{c}_i\}_1^q$ (3.10). Positivity is then imposed by replacing all negative elements of **D** by 0,

$$d_i \leftarrow \max(d_i, 0), \tag{3.16}$$

in equation (3.14).

Equations (3.14)–(3.16) generalize the GCV approach by removing approximation (3.5), and accounting for the variability in the estimate of the sample canonical co-ordinate transformation $\hat{\mathbf{T}}$ (3.10), in the estimation of the canonical co-ordinate shrinkage factors $\{d_i\}_1^q$ (3.14). This usually results in increased shrinkage. This is accomplished at the expense of considerably increased computational complexity. The quantities $\hat{\mathbf{T}}_{\setminus n}$ and $\hat{\mathbf{y}}_{\setminus n}$ must be calculated for each observation ($n = 1, \ldots, N$) removed. In practice this 'N-fold' cross-validation procedure is approximated by V-fold cross-validation in which successive subsets of N/V observations are removed and the values of $\hat{\mathbf{T}}$ and $\hat{\mathbf{y}}$, computed on the remaining (training) observations, are used for all the observations in the left-out subset. This reduces the computation by a factor of V/N. Common choices are $V = 5$ or $V = 10$.

4. COMPETITORS

In terms of common statistical practice the primary competitor to the procedures that we propose (C&W–GCV and C&W–CV) is OLS, i.e. a separate least squares regression (1.2)–(1.3) of each response y_i on the predictor variables **x**. However, it is well known that OLS is inadmissible (James and Stein, 1961) and in fact can be

(sometimes) substantially dominated, in terms of (single-response) prediction accuracy, by a variety of biased (regularized) alternatives (Frank and Friedman, 1993). Thus, when comparing our multivariate approaches with a strategy of separate marginal univariate regressions, the best among these biased methods should provide worthier competition.

As noted in Section 1.2 several multivariate multiple-regression procedures have been proposed with the same goal as ours; they attempt to exploit the correlational structure among the responses to improve prediction accuracy. In Sections 4.2–4.4 we include a brief description of some of these and examine their relationship to our procedures. In Section 5 we compare performances through an extensive simulation study.

4.1. *Separate Ridge Regressions*

Ridge regression (RR) (Hoerl and Kennard, 1970) is one of the more popular and best performing (Frank and Friedman, 1993) alternatives to (single-response) OLS. A reasonable multiple-response strategy would be to perform a separate RR on each individual response y_i (1.2). The regression coefficient estimates are the solution to a penalized least squares criterion

$$\{\hat{a}_{ij}\}_{j=1}^{p} = \arg\min_{\{a_j\}_1^p} \left\{ \sum_{n=1}^{N} \left(y_{ni} - \sum_{j=1}^{p} a_j x_{nj} \right)^2 \right\} + \lambda_i \sum_{j=1}^{p} a_j^2, \qquad i = 1, \ldots, q. \quad (4.1)$$

This biases the coefficient estimates towards smaller absolute values and discourages dispersion among their values. The 'ridge' parameters $\{\lambda_i\}_1^q$ (4.1) regulate the strength of this effect and their values are estimated through model selection. We employed cross-validation to estimate each (separate) ridge parameter

$$\hat{\lambda}_i = \arg\min_{\lambda} \left\{ \sum_{n=1}^{N} (y_{ni} - \hat{y}_{\setminus ni})^2 \right\}, \qquad i = 1, \ldots, q, \quad (4.2)$$

with $\hat{y}_{\setminus ni}$ being the RR estimate (1.2)–(4.1) obtained with the nth observation removed from the training sample. Although this separate RR approach ignores the correlational structure of the response variables $\{y_i\}_1^q$, it can provide considerably more accurate estimates than OLS can (see Section 5).

4.2. *Reduced Rank Regression*

Reduced rank regression (Izenman, 1975) places a rank constraint on the matrix of estimated regression coefficients (1.2). Consider regression model (2.5)–(2.6) and suppose that we wish to find the coefficient matrix $\tilde{\mathbf{A}}_r \in R^{q \times p}$ of rank $r \leqslant \min(p, q)$ that minimizes

$$\tilde{\mathbf{A}}_r = \arg\min_{\operatorname{rank}(A)=r} \{E(\mathbf{y} - \mathbf{A}\mathbf{x})^{\mathrm{T}} \Sigma^{-1} (\mathbf{y} - \mathbf{A}\mathbf{x})\} \quad (4.3)$$

with $\boldsymbol{\Sigma}$ given by equation (2.8). The solution to equation (4.3) is

$$\tilde{\mathbf{A}}_r = \mathbf{B}_r \hat{\mathbf{A}} \quad (4.4)$$

where $\hat{\mathbf{A}} \in R^{q \times p}$ is the matrix of OLS estimates and the reduced rank 'shrinking' matrix $\mathbf{B}_r \in R^{q \times q}$ is given by

$$\mathbf{B}_r = \mathbf{T}^{-1}\mathbf{I}_r\mathbf{T} \tag{4.5}$$

with $\mathbf{T}$ being the (population) canonical co-ordinate matrix (2.18) and

$$\mathbf{I}_r = \text{diag}\{\mathbf{1}(i \leqslant r)\}_1^q. \tag{4.6}$$

In applications of reduced rank regression the sample canonical co-ordinates $\hat{\mathbf{T}}$ (2.25) and (3.10) are taken as estimates of the corresponding population quantities in equation (4.5) and the rank value r (4.6) is regarded as a regularization parameter of the procedure whose value is estimated through model selection. We employed cross-validation (the analogue of equation (4.2)). This estimate (2.25), (3.10), (4.5) and (4.6) has the same form as C&W–GCV but with a different diagonal matrix ($\mathbf{I}_r$ (4.6) *versus* $\mathbf{D}$ (3.11)–(3.13)).

4.3. *Filtered Canonical y-variate Regression*

FICYREG was proposed by van der Merwe and Zidek (1980). The estimated coefficient matrix $\tilde{\mathbf{A}} \in R^{q \times p}$ takes the form

$$\tilde{\mathbf{A}} = \mathbf{B}_f\hat{\mathbf{A}} \tag{4.7}$$

where again $\hat{\mathbf{A}} \in R^{q \times p}$ is the matrix of OLS estimates and the shrinking matrix $\mathbf{B}_f \in R^{q \times q}$ is given by

$$\mathbf{B}_f = \hat{\mathbf{T}}^{-1}\mathbf{F}\hat{\mathbf{T}}. \tag{4.8}$$

Here $\hat{\mathbf{T}}$ is the sample canonical co-ordinate matrix (2.25) and (3.10) and

$$\mathbf{F} = \text{diag}\{f_1, \ldots, f_q\} \tag{4.9}$$

with

$$f_i = \left(\hat{c}_i^2 - \frac{p-q-1}{N}\right) \Big/ \hat{c}_i^2\left(1 - \frac{p-q-1}{N}\right) \tag{4.10}$$

and

$$f_i \leftarrow \max(0, f_i). \tag{4.11}$$

The $\{\hat{c}_i^2\}_1^q$ in equation (4.10) are the sample (squared) canonical correlations (3.10).

Like reduced rank regression FICYREG shrinkage (4.7)–(4.11) also has the same form as C&W–GCV, here with the matrix $\mathbf{F}$ (4.9)–(4.11) replacing $\mathbf{D}$ (3.11)–(3.13). One difference between equations (4.10) and (3.12) is that the canonical co-ordinate shrinkage factors $\{f_i\}_1^q$ (4.10) depend on the number of responses q as well as on the number of predictor variables p and corresponding squared sample canonical correlations $\{\hat{c}_i^2\}_1^q$. For the same values of $\hat{c}_i^2$ and p, equations (4.10) and (4.11) shrink less for a larger number of responses. The corresponding C&W–GCV factors $\{d_i\}_1^q$ (3.12) and (3.13) depend only on $\{\hat{c}_i^2\}_1^q$ and p irrespective of the number of responses. For all values $q \geqslant 1$ we have

$$\{d_i < f_i\}_1^q, \tag{4.12}$$

i.e. FICYREG always shrinks less than C&W–GCV. As the number of responses increases effect (4.12) becomes more pronounced. In fact, if we set $q = -1$ in expressions (4.10) and (4.11) shrinkage values almost identical with those of expressions (3.12) and (3.13) are produced for the same value of $\hat{c}_i^2$ and p.

4.4. *Two-block Partial Least Squares*

PLS regression (Wold, 1975) is very popular in chemometrics. The multiple ($q > 1$) response version ('two-block' PLS) begins with a 'canonical covariance' analysis. This is similar to canonical correlation analysis (Section 2.2) with the covariance between the linear combination pairs $\mathrm{cov}(\mathbf{t}^{\mathrm{T}}\mathbf{y}, \mathbf{v}^{\mathrm{T}}\mathbf{x})$ replacing $\mathrm{corr}(\mathbf{t}^{\mathrm{T}}\mathbf{y}, \mathbf{v}^{\mathrm{T}}\mathbf{x})$ in equation (2.15), and the constraints in equations (2.17) replaced by $\{\mathbf{t}_k^{\mathrm{T}}\mathbf{t}_k = \mathbf{v}_k^{\mathrm{T}}\mathbf{v}_k = 1\}_1^q$. The (ordered) set of canonical covariance $\mathbf{x}$ linear combinations

$$\{z_k = \mathbf{v}_k^{\mathrm{T}}\mathbf{x}\}_1^p \tag{4.13}$$

are then used to form an ordered sequence of coefficient estimates for each response

$$\{\tilde{a}_{ik}^{(K)}\}_{k=1}^K = \arg\min_{\{a_k\}_1^K} \left\{ \sum_{n=1}^N \left(y_{ni} - \sum_{k=1}^K a_k z_{nk} \right)^2 \right\}, \tag{4.14}$$

$$\tilde{y}_1^{(K)} = \sum_{k=1}^K \tilde{a}_{ik}^{(K)} z_k, \qquad i = 1, q. \tag{4.15}$$

This is a separate OLS regression of each response y_i on the first K $\mathbf{x}$ canonical covariance linear combinations (4.13). Coefficients (4.14) reference the linear combinations (4.13) as predictor variables. They can be easily transformed to reference the original predictors $\{x_j\}_1^p$. The number of 'components' K (4.14) and (4.15) is a regularization parameter of the procedure; its value is determined through cross-validation (the analogue of equation (4.2)).

The relationship between two-block PLS and other multiple-response regression procedures is not obvious. It was introduced by Wold (1975) as an iterative computational algorithm and much effort has been expended since then to try to understand it statistically. Frank and Friedman (1993) provided some insight by comparing its results with that of a particular formulation of multivariate RR derived from a particular joint prior on the true regression coefficients and assumptions on the error covariance matrix $\mathbf{\Sigma}$ (2.8).

4.5. *Discussion*

Our proposals, C&W–GCV and C&W–CV, were introduced in Sections 3.1 and 3.2 respectively. Four additional approaches (separate RRs, reduced rank regression, FICYREG and two-block PLS) were described in Sections 4.1–4.4. These are not the only ones that have been proposed. Brown and Zidek (1980, 1982) suggested a variety of multivariate generalizations of RR along the lines of FICYREG. The four competitors described above have seen use on data and two (separate RRs and two-block PLS) are very popular.

Of the six procedures described above, four (C&W–CV, reduced rank regression, separate RRs and two-block PLS) require sample reuse (cross-validation) to estimate regularization parameters. Therefore they can be expected to be much more computationally intense than the other two (C&W–GCV and FICYREG) which do not require sample reuse to estimate such parameters. All the procedures except two (separate RRs and two-block PLS) are equivariant under all non-singular affine (translation, rotation and/or scaling) transformations of either the responses $\mathbf{y}$ or the predictors $\mathbf{x}$. Separate RRs are clearly equivariant under response scale changes but not under rotations in the response space. They are equivariant under (rigid) rotations of the $\mathbf{x}$ co-ordinates but not equivariant under scale changes of the predictors or their linear combinations. Two-block PLS is rotationally equivariant in both the $\mathbf{y}$- and the $\mathbf{x}$-spaces but not equivariant under scale changes in either space. Both RR and PLS are equivariant under translation in both spaces.

Although motivated from very different perspectives, four of the six procedures discussed above (the affine equivariant ones) all have the same (generic) form

$$\tilde{\mathbf{y}} = (\hat{\mathbf{T}}^{-1}\mathbf{G}\hat{\mathbf{T}})\hat{\mathbf{A}}\mathbf{x} \tag{4.16}$$

where $\hat{\mathbf{T}}$ is the matrix of sample canonical co-ordinates (2.25) and (3.10), and the diagonal $q \times q$ matrix $\mathbf{G}$ contains the shrinkage factors for scaling the OLS solutions $\hat{\mathbf{A}}$ in the canonical co-ordinate system. C&W–GCV (3.11)–(3.13) and C&W–CV (3.14)–(3.16) were motivated by the cross-validation approximation (3.1) to optimal proportional shrinking (2.3). Reduced rank regression (4.5) and (4.6) derives its motivation from the 'naturalness' of regularizing OLS through a rank restriction on the matrix of estimated coefficients (4.3). FICYREG is based on Zidek (1978) which contains the only previous theoretical justification for transforms of the form (4.16). Zidek assumed that the data $\{\mathbf{y}_n, \mathbf{x}_n\}_1^N$ are an IID sample from a joint normal distribution. A set of transformations of the data is defined together with a particular (amalgamated) invariant loss function. The equivariant coefficient estimates are then given by equation (4.16) where the elements of $\mathbf{G}$ depend only on the sample canonical correlations. Zidek (1978) then showed that, for the particular loss function defined, the form of $\mathbf{G}$ used in FICYREG (4.10) and (4.11) gives estimates dominating OLS. It is perhaps no surprise that many multivariate multiple-regression procedures involve canonical co-ordinates at a basic level, since, as shown in Sections 2.1 and 2.2, the canonical co-ordinate system emerges as the natural system for optimal proportional shrinkage (2.22) and (2.24).

5. SIMULATION STUDY

An important issue is whether any of the multivariate multiple-regression procedures offer sufficient improvement over separate (uniresponse) multiple regressions (OLS or separate RR) to justify their consideration as viable alternatives. And, among those that do, which provide the best trade-off between accuracy improvement and increased complexity, both in terms of implementation and computation? The answers to these questions may depend on the detailed nature of the problem at hand in terms of the number of observations N, the number of response variables q, their correlational structure, signal-to-noise ratio, collinearity of the predictor

variables, etc. In this section we attempt to provide some answers to these questions by means of an extensive simulation study.

5.1. *Design*

In all situations covered by this study the number of predictor variables was taken to be $p = 50$. There were two training sample sizes ($N = 100$ and $N = 400$) and three values for the number of responses ($q = 5$, $q = 10$ and $q = 20$). For each (random) replication of each situation the predictor variables were generated according to a normal distribution with zero mean and covariance matrix $\mathbf{V}$,

$$\mathbf{x} \sim N(\mathbf{0}, \mathbf{V}). \tag{5.1}$$

The covariance matrix $\mathbf{V}$ (5.1) was itself random with a different realization for each replication

$$V_{ij} = r^{|i-j|} \tag{5.2}$$

with r a random number generated from a uniform distribution

$$r \sim U[-1, 1]. \tag{5.3}$$

Thus for some replications ($|r| \simeq 1$) there was a high degree of collinearity among the predictors, whereas for others ($|r| \simeq 0$) they are nearly independent. A range of possibilities (5.3) in between these extremes was also produced.

Each response y_i was computed from equations (2.5) and (2.6). The errors $\{\epsilon_i\}_1^q$ were generated from a normal distribution with zero mean and covariance matrix $\mathbf{\Sigma}$ (2.8)

$$\{\epsilon_i\}_1^q \sim N(\mathbf{0}, \mathbf{\Sigma}). \tag{5.4}$$

Two covariance structures among the errors were studied:

$$\begin{aligned} \mathbf{\Sigma} &= \sigma^2 \mathbf{I}_q, \\ \mathbf{\Sigma} &= \sigma^2 \operatorname{diag}(\{i^2\}_1^q). \end{aligned} \tag{5.5}$$

In the first, the error variance associated with each response is the same, whereas in the second they are very different. More complicated (non-diagonal) error covariance structures were not considered since they are included for the signal covariance matrix $\mathbf{F}$ (2.10) (see below), and the relevant quantity is the relationship between the signal and noise covariances as captured by the noise-to-signal matrix $\mathbf{R}$ (2.14). Two values of σ^2 (5.5) were studied. They were chosen so that (on average) signal-to-noise ratios of 1.0 and 3.0 were produced.

The ('true') coefficients a_{ij} (2.6) were generated through

$$a_{ij} = \sum_{k=1}^{10} c_{ik}\, g(j, k) \tag{5.6}$$

with

$$g(j, k) = h_k(l_k - |j - j_k|)_+^2 \tag{5.7}$$

where the value of h_k is adjusted so that

$$\sum_{j=1}^{50} g(j, k) = 1. \tag{5.8}$$

The quantities j_k and l_k in equation (5.7) are integers with random values sampled from uniform distributions in the ranges [1, 50] and [1, 6] respectively. The coefficients $\{c_{ik}\}_{i=1}^{q}$ (5.6) are each randomly sampled (separately) from a (q-dimensional) Gaussian distribution

$$\{c_{ik}\}_{i=1}^{q} \sim N(0, \mathbf{\Gamma}) \tag{5.9}$$

with the covariance matrix being

$$\Gamma_{mn} = \rho^{|m-n|}. \tag{5.10}$$

Thus, the coefficients c_{ik} (5.6) are independent for different k but correlated among the responses i, with the degree of that correlation controlled by the value of the parameter ρ (5.10). Finally, all coefficient values were normalized by the same scale factor so that the average ('signal') variance for each response was equal to 1.0.

Each $g(j, k)$ in equation (5.7), when viewed as a function of the predictor variable index j, represents a (normalized) 'bump' centred at j_k with support (non-zero values) in the interval $[j_k - l_k, j_k + l_k]$. Thus the coefficient vector (5.6) for each response is a (different) random superposition of the (same) 10 such bumps, each bump centred at a random location j_k, with (random) width l_k. Since the coefficients multiplying each of the individual bumps are independent of each other, the (average) correlation among the response variables is completely determined by the covariance matrix $\mathbf{\Gamma}$ (5.9) controlled by the parameter ρ (5.10). Therefore, the (true) response functions (2.6) are (randomly) different for each replication (of each situation). Some have coefficients $\{a_{ij}\}_{j=1}^{50}$ that have roughly the same (absolute) values, whereas others have coefficients with very different (absolute) values (e.g. a few large values and the others very small). A variety of sets of coefficient values in between these extremes are also realized.

The design of this simulation comprises two samples sizes ($N = 100$ and $N = 400$), three values for the number of responses ($q = 5$, 10, 20), five values for the average correlation among the response functions (2.6)

$$\text{ave}_{i \neq j}|\text{corr}(f_i, f_j)| = \pm 0.7, \ \pm 0.35, \ 0.0 \tag{5.11}$$

(controlled by ρ (5.10)), two error covariance structures (5.5) and two signal-to-noise ratios (1.0 and 3.0). A complete factorial design over all these levels gives rise to $2 \times 3 \times 5 \times 2 \times 2 = 120$ situations. Each situation was replicated 250 times giving rise to 30000 runs. Each of the competitors (OLS, separate RR, reduced rank, FICYREG, two-block PLS, C&W–GCV and C&W–CV) were applied to the data for each run. Thus, the entire simulation study consists of 210000 (multiple-response) regressions.

5.2. *Performance Measures*

For each replication, the mean-squared estimation error of the ith response for a particular method m is given by

$$e_i^2(m) = \int \{(\mathbf{a}_i - \tilde{\mathbf{a}}_i(m))^{\mathrm{T}}\mathbf{x}\}^2 \, p(\mathbf{x}) \, d\mathbf{x}$$
$$= (\mathbf{a}_i - \tilde{\mathbf{a}}_i(m))^{\mathrm{T}}\mathbf{V}(\mathbf{a}_i - \tilde{\mathbf{a}}_i(m)) \tag{5.12}$$

where $\mathbf{a}_i = \{a_{i1}, \ldots, a_{ip}\}$ is the true coefficient vector (2.6) and (5.6) for the ith response and $\tilde{\mathbf{a}}_i(m)$ is the corresponding estimate for each method. Here $p(\mathbf{x})$ is the probability density (5.1) from which the predictors $\mathbf{x}$ are sampled and $\mathbf{V}$ is the corresponding (population) covariance matrix (2.8) and (5.2). Several summary measures of relative performance are derived based on different combinations of $\{e_i^2(m)\}_1^q$ (5.12). The first is the overall average mean-squared error

$$A(m) = \sum_{i=1}^{q} e_i^2(m) \Big/ \sum_{i=1}^{q} e_i^2(\mathrm{OLS}) \tag{5.13}$$

relative to the overall average of the OLS mean-squared estimation errors $\{e_i^2(\mathrm{OLS})\}_1^q$. The second performance measure is the average of the individual ratios of each response mean-squared error to that of its OLS estimate

$$I(m) = \frac{1}{q} \sum_{i=1}^{q} \frac{e_i^2(m)}{e_i^2(\mathrm{OLS})}. \tag{5.14}$$

The third measure is the worst individual mean-squared error relative to OLS,

$$W(m) = \max_{m=1,q} \left\{ \frac{e_i^2(m)}{e_i^2(\mathrm{OLS})} \right\}. \tag{5.15}$$

The fourth and fifth measures are derived from the first two; they are the ratio of each to the corresponding minimum value over all six methods being compared,

$$\mathrm{RA}(m) = A(m) / \min_{k=1,6} \{A(k)\}, \tag{5.16}$$

$$\mathrm{RI}(m) = I(m) / \min_{k=1,6} \{I(k)\}. \tag{5.17}$$

Criteria (5.13) and (5.14) provide a means of comparing each of the six methods with OLS in terms of how much average (squared) error reduction each gives relative to OLS. Criterion (5.15) measures the degree of caution associated with each method. Values of $W(m) > 1$ indicate that the method produced at least one response estimate that is less accurate than its corresponding OLS estimate. The last two measures (5.16) and (5.17) allow comparisons between the six biased methods themselves. For each individual replication, the value of expression (5.16) or (5.17) is 1.0 for the corresponding best (minimum error) method, and greater than that for the other methods. If a particular method happened to be best for every replication then the corresponding distribution of its values (5.16) and (5.17) over all replications would be a point mass at the minimum value (1.0).

5.3. *Results*

The results of the simulation study are summarized by the respective means of the performance measure values (5.13)–(5.17) for each method over the 250 replications for each situation. Figs 3–6 display box plots of the mean values of equations (5.13), (5.14), (5.16) and (5.17) respectively over all the 120 situations covered by the simulation study, i.e. each box plot summarizes the distribution of 120 (mean) values. Fig. 3 summarizes the distribution of the average overall mean-squared error ratio $A(m)$ (5.13) for each of the six methods. All are seen to provide substantial improvement over OLS ($A(\text{OLS}) = 1$). All the multivariate methods, except two-block PLS, also show substantial improvement over separate (uniresponse) RRs. The average overall mean-squared error associated with reduced rank regression and FICYREG are comparable, with the latter exhibiting considerably less variability. C&W–CV and C&W–GCV show comparable performance with each other, and somewhat better than the rest. The best of these methods, C&W–CV, provides over a factor of 2 improvement over OLS, as averaged over all 120 situations, and about a 61% improvement over separate RRs.

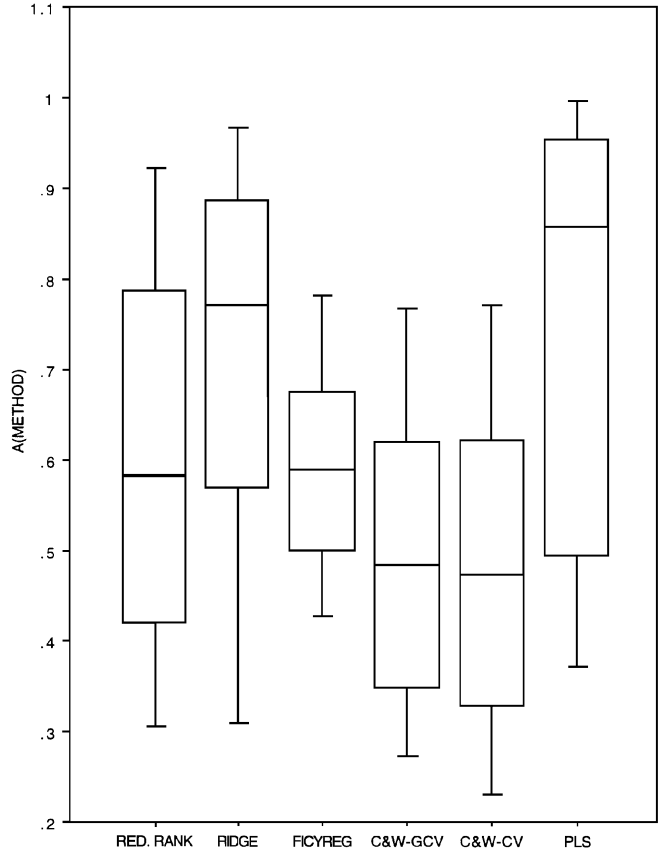

Fig. 3. Distribution over all 120 situations ($p \prec N$) of the overall average response mean-squared error relative to OLS (5.13) for each biased method

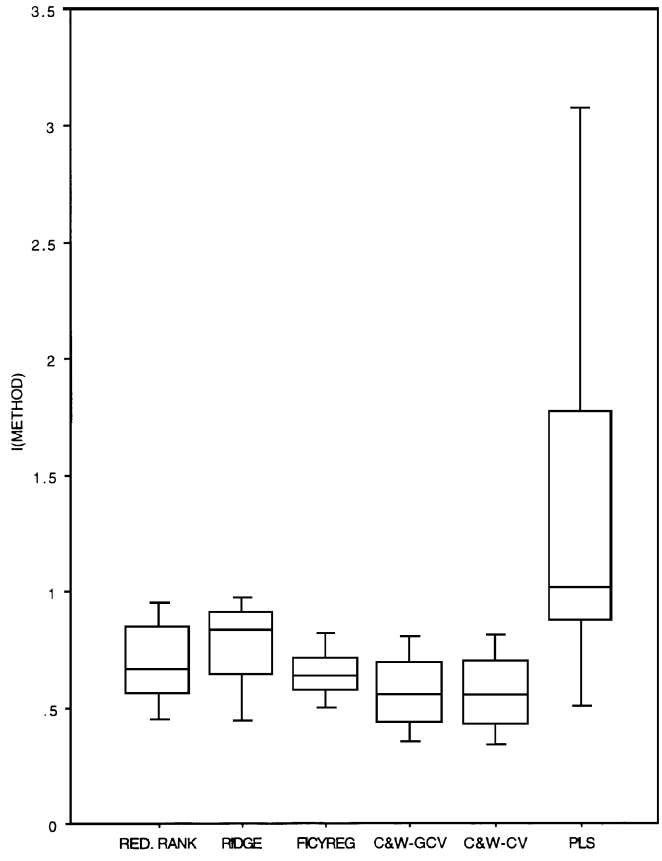

Fig. 4. Distribution over all 120 situations ($p \prec N$) of the average individual response mean-squared error relative to OLS (5.14) for each biased method

Fig. 4 shows the distribution of average individual mean-squared error ratio $I(m)$ (5.14). These distributions are fairly similar to the corresponding distributions for the $A(m)$- (5.13) values, except for two-block PLS. The $I(m)$-values for two-block PLS tend to be substantially larger than its $A(m)$-values. This indicates that two-block PLS suffers a 'Robin Hood' effect where responses that are well estimated by OLS (low error) are made substantially worse (relatively) by PLS to achieve modest (relative) improvements in those that are poorly estimated by OLS (and PLS). Comparing Figs 3 and 4 we see that the other methods do not exhibit the Robin Hood effect; they produce a roughly equal relative improvement across all responses.

Fig. 5 shows the distributions of RA(m) (5.16), and Fig. 6 the logarithm of RI(m) (5.17). The C&W–CV method is seen to have the best average performance, or within a few per cent of the best, in every one of the 120 situations. The C&W–GCV method is seen to be next closest to the best, with median performance only 2% worse than C&W–CV and seldom more than 10% worse. The other methods substantially lag behind these two, relative to the best performer.

Figs 3 and 4 show that, averaged over all responses, all the six biased methods considered here provide improved performance over OLS. That improvement was

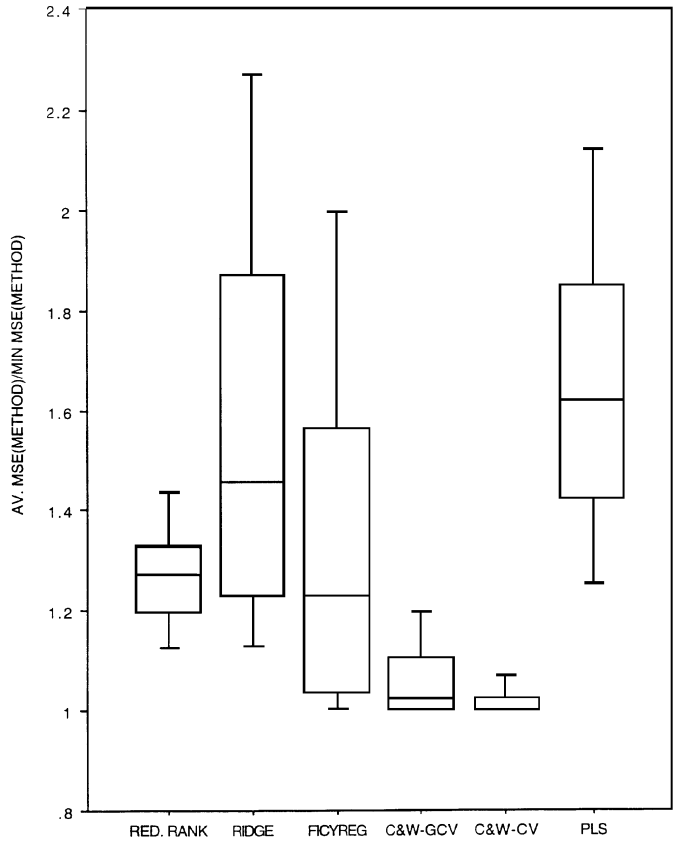

Fig. 5. Distribution over all 120 situations ($p \prec N$) of the ratio of overall average response mean-squared error for each method, to that of the best method (5.16)

fairly dramatic for some of the methods. From a perspective of caution one might ask how probable is it that an individual response estimate by one of these methods will be less accurate that its OLS estimate, i.e. how often do they make things worse? We already have an indication that two-block PLS has a tendency to degrade the most accurate OLS estimates. Fig. 7 addresses this issue for all the methods by showing the distribution (over all 30000 replications associated with the 120 situations) of the fraction of responses (in each replication) for which the accuracy of the biased estimate was worse than that for OLS. We see that the most cautious method by this measure is FICYREG. On average less than 3% of its response estimates are worse than OLS. The C&W–GCV method is seen to be only slightly less cautious, its estimates being worse than the corresponding OLS estimates an average of 5% of the time. The C&W–CV method also exhibits fairly cautious behaviour by this measure, degrading the OLS estimate on average 7% of the time. At the other extreme is two-block PLS which degrades the OLS estimate an average of 35% of the time, providing further evidence of its susceptibility to the Robin Hood effect.

Another measure of caution is the worst individual mean-squared error ratio $W(m)$ (5.15). Fig. 8 shows the distribution of the logarithm of this quantity for each

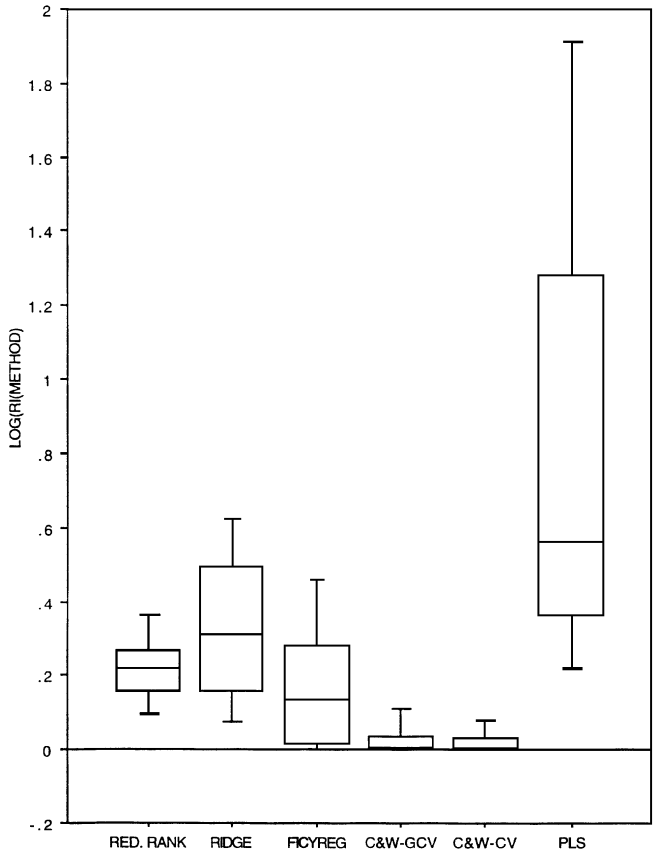

Fig. 6. Distribution over all 120 situations ($p \prec N$) of the logarithm of the ratio of average individual response mean-squared error (relative to OLS) for each method, to that of the best method (5.17)

method, separately for each of the two error variance structures (5.5). The left-hand box plot for each method m summarizes the distribution of the averages of $W(m)$ for the 60 situations in which the (population) error variances are all equal, $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_q$, and the right-hand box plot is the corresponding distribution over the other 60 for which they are very unequal, $\mathbf{\Sigma} = \sigma^2 \operatorname{diag}(\{i^2\}_1^q)$. We see that for the most cautious methods (FICYREG, C&W–GCV and C&W–CV) $W(m)$ seldom becomes much larger than 1.0, indicating that these methods seldom produce a substantial degradation of the OLS estimate for any response for either error variance structure. These methods are seen to be slightly less cautious for highly dissimilar error variances than for equal variances. In contrast, the caution associated with two-block PLS is seen to depend dramatically on the structure of the error variances of the respective responses. Although, even for equal error variances, it is the least cautious of the methods considered here, PLS at least does not produce disastrous results in this case. When the errors of the individual responses have highly unequal variances, however, two-block PLS typically degrades the OLS error (squared) of at least one of the responses (usually the best one(s)) by a factor of 10, and factors of 20 are quite

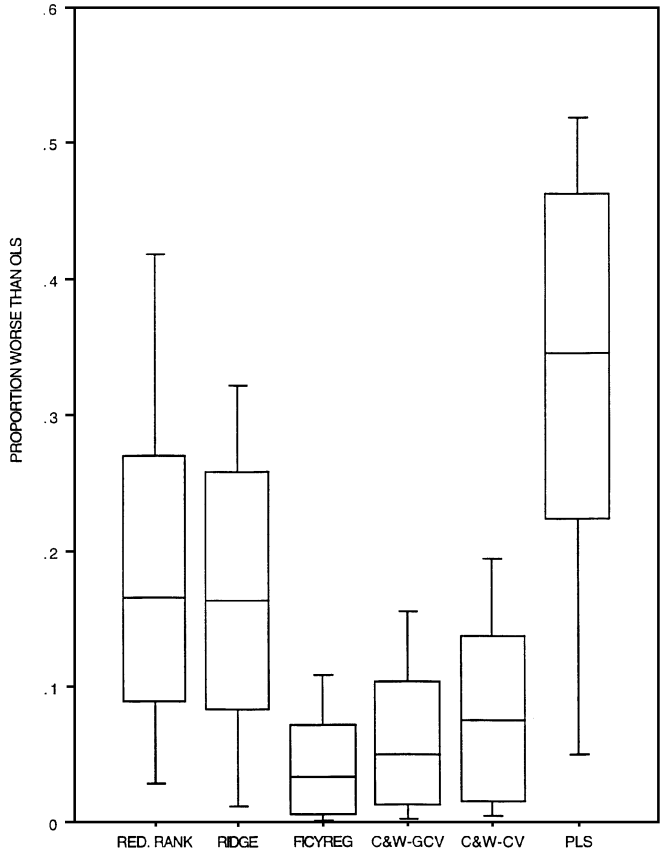

Fig. 7. Distributions over all the 30000 replications ($p \prec N$) of the fraction of responses in each, for which the respective biased methods were less accurate than the corresponding OLS estimate

common. Frank and Friedman (1993) argued that an intrinsic (implicit) assumption associated with two-block PLS is the simple error covariance structure $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_q$. The results shown in Fig. 8 tend to confirm this.

As noted in Section 4.3, FICYREG always shrinks less than C&W–GCV (4.12), which in turn shrinks less (on average) than C&W–CV. Shrinking less aggressively causes less modification of the OLS estimates resulting in less chance of making things worse. However, this more cautious approach limits the gains that are possible as a result of the shrinking strategy. If caution is an important issue, C&W–GCV would appear to be the best compromise since it results in nearly as much caution as the most cautious method FICYREG (Fig. 7), while providing nearly as much accuracy as the most accurate method, C&W–CV (Figs 5 and 6).

Fig. 9 shows the (first-order) interaction effects between the method (m) and the factors of the simulation design. Plotted on the vertical scale is the average of $A(m)$ (5.13) over all situations for which the particular factor was at the given level indicated on the horizontal axis. We see from Fig. 9(a) that separate RRs are unaffected by the degree of correlation among the responses (5.11) whereas the

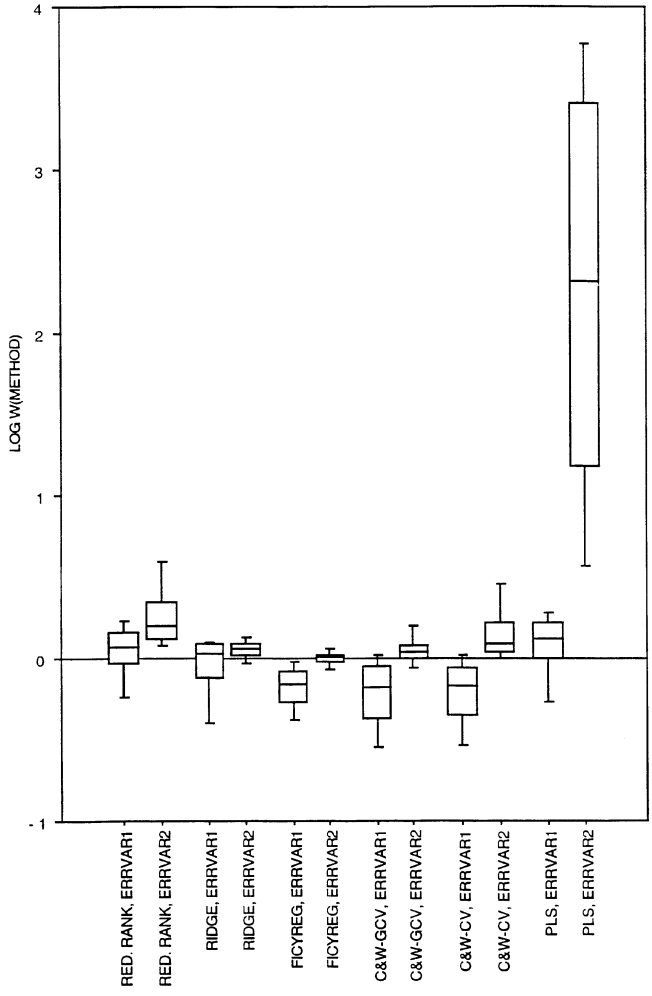

Fig. 8. Distribution ($p \prec N$) of the logarithm of the worst individual response mean-squared error relative to OLS (5.15) of each of the six biased methods, for each of the two error covariance matrix structures (5.5) (ERRVAR1 and ERRVAR2)

multivariate methods all perform better with higher (positive or negative) correlation, as would be expected. Fig. 9(b) shows that the performance (relative to OLS) of all methods, except two-block PLS, is better with highly unequal error variances (5.5). As we would expect, all methods improve (relative to OLS) with decreasing sample size (Fig. 9(c)) and decreasing signal-to-noise ratio (Fig. 9(d)), but FICYREG seems to enjoy less improvement than the others. Fig. 9(e) shows the dependence of $A(m)$ on the number of responses q. The performance of separate RRs is independent of q (as would be expected), whereas that of all the multivariate methods, except FICYREG, improves (monotonically) with more response variables. FICYREG's relative inability to take advantage of an increasing number of responses q is probably due to the dependence of its shrinkage factors (4.10) on q, as discussed in Section 4.3. Two-block PLS shows only modest performance gain with increasing q,

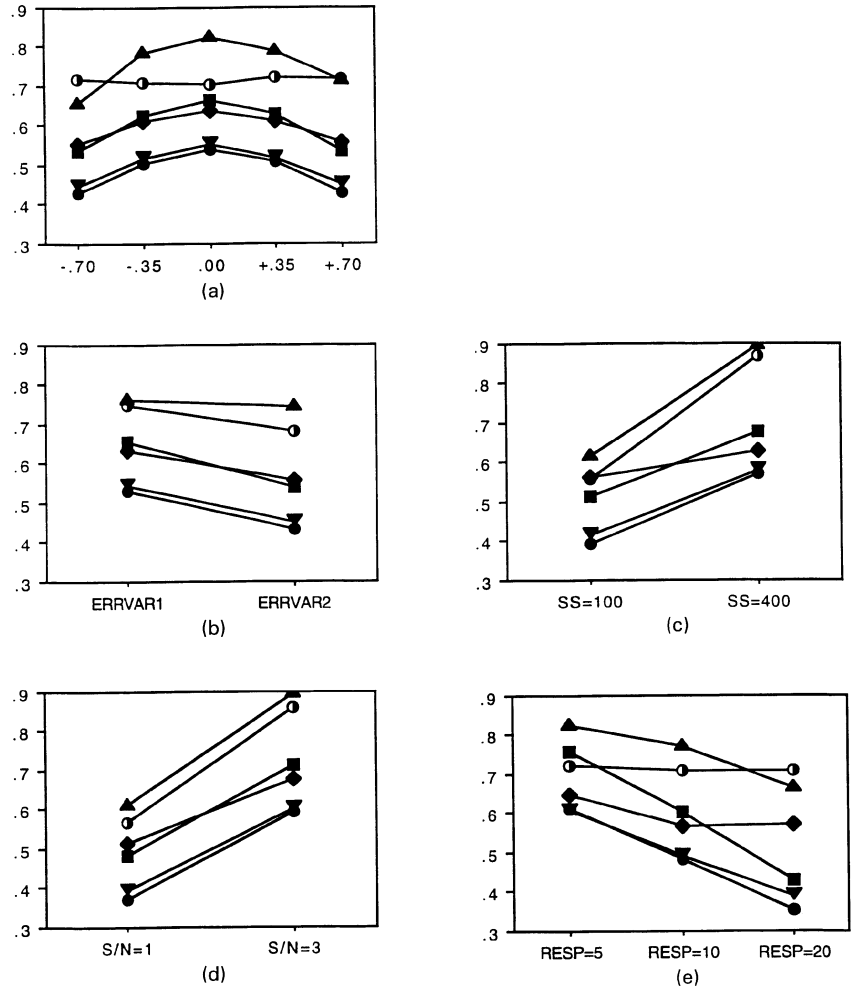

Fig. 9. Interaction of method with the other factors of the ($p \prec N$) simulation design (the ordinate is average response mean-squared error relative to OLS (5.13); number is the average response correlation, RESP is the number of responses, SS is the sample size and S/N is the signal-to-noise ratio): ■, reduced rank regression; ◑, RR; ◆, FICYREG; ▼, C&W–GCV; ●, C&W–CV; ▲, PLS

whereas reduced rank regression shows the most rapid (relative) gain. Note that the two C&W procedures dominate the others at all levels of all the design factors, with C&W–CV always being (slightly) the better.

5.4. *Discussion*

Overall, the simulation studies demonstrate that some multivariate multiple-regression methods can produce increased (expected) prediction accuracy (for each response) over separate multiple regressions (OLS or RR). Of the methods compared here, only two-block PLS provided results inferior to separate RRs. If prediction accuracy were the only criterion for choosing a method then Figs 5 and 6 suggest

C&W–CV as the method of choice. It attained the highest average accuracy, or very close to it, in every one of the 120 situations comprising our simulation study. However, C&W–GCV is a worthy contender, typically performing almost as well as C&W–CV in relation to the best method in every situation.

If (minimax) caution were a primary concern then FICYREG might be a good choice. However, C&W–GCV is only slightly less cautious (Figs 7 and 8) while producing substantially greater gains in accuracy (Figs 5 and 6). C&W–CV is also seen to be fairly cautious, being only slightly less so than C&W–GCV. In terms of implementational simplicity and computational speed FICYREG and C&W–GCV stand out. Neither requires sample reuse (cross-validation) to estimate the values of model selection parameters, and both are easily implemented in any statistical package that provides canonical correlation analysis. Again C&W–GCV appears to be the logical choice between these two owing to its higher performance in terms of accuracy in our simulation study.

Two-block PLS emerges from this simulation study as consistently the poorest performer from every perspective. It is the least cautious and produces the least accuracy among all the biased methods considered here. In fact, it is dominated in accuracy by separate RRs. This, coupled with the fact that it is (by far) the computationally slowest method, and that it is affine equivariant in neither the predictor nor the response space, would tend to exclude it from consideration. This is somewhat surprising since it is one of the most popular and highly promoted methods for multivariate multiple regression, especially in chemometrics. By contrast, single-response PLS is competitive with other (single-response) biased regression methods, performing almost as well as RR (Frank and Friedman, 1993). This together with the fact that separate RRs substantially outperform two-block PLS suggest that, in environments where PLS for some reason must be used, performing separate (uni-response) PLS regressions on each individual response would be a better strategy than employing (multivariate) two-block PLS. This is especially the case if the error variances among the responses are not equal (Fig. 8). The superiority of separate PLS regressions over two-block PLS has been noted by Frank and Friedman (1993) and Garthwaite (1994). The simulation results of Section 5.3 suggest, however, that using one of the better multivariate multiple-regression procedures should provide considerably enhanced performance over a strategy of separate uniresponse PLS regressions since they consistently outperformed separate RRs.

It is important to note that all these conclusions are based on the results of the simulation study described in Section 5.1. Although considerable effort was involved in attempting to make it as comprehensive as possible, every conceivable situation cannot be covered by any such study. Just as one can seldom verify whether a particular data set conforms to the assumptions associated with any theoretical result, we cannot be sure that it is represented within the scope of our simulation study. It is possible that for factor values that are very different from those represented in our design the results would be different, in the same way that violation of the assumptions of a theorem may alter its conclusions.

6. UNDERDETERMINED SYSTEMS

Separate RRs and two-block PLS do not require the response and/or predictor sample covariance matrices, $\mathbf{Y}^T\mathbf{Y}$ and $\mathbf{X}^T\mathbf{X}$ (2.26) respectively, to be non-singular.

Therefore no special problems arise with these procedures when $q > N$ and/or $p > N$. However, the other multivariate multiple-regression procedures considered here (reduced rank regression, FICYREG, C&W–GCV and C&W–CV) are not strictly defined when either $\mathbf{Y}^T\mathbf{Y}$ or $\mathbf{X}^T\mathbf{X}$ is singular. Therefore these methods must be suitably generalized to be applicable to such settings. Situations for which $p > N$, especially, represent an important class of applications.

Singular $\mathbf{Y}^T\mathbf{Y}$ causes no special problem. The response linear combinations (eigenvectors of $\mathbf{Y}^T\mathbf{Y}$) corresponding to zero variance (eigenvalues) are simply defined to have zero (canonical) correlation with the predictors, and the usual canonical correlation analysis (2.25) and (3.10) is then confined to the non-zero variance subspace of the responses by using the generalized inverse of $\mathbf{Y}^T\mathbf{Y}$ in equation (2.25). Dealing with singular $\mathbf{X}^T\mathbf{X}$, however, must be done with care.

One possibility for treating singular $\mathbf{X}^T\mathbf{X}$ is by analogy with that for singular $\mathbf{Y}^T\mathbf{Y}$. Perform an eigenanalysis of the predictor covariance matrix

$$\left.\begin{aligned} \mathbf{X}^T\mathbf{X} &= \mathbf{U}\mathbf{E}^2\mathbf{U}^T, \\ \mathbf{U}^T\mathbf{U} &= \mathbf{U}\mathbf{U}^T = \mathbf{I}_p, \\ \mathbf{E}^2 &= \operatorname{diag}\{e_1^2, \ldots, e_r^2, 0, \ldots\} \end{aligned}\right\} \tag{6.1}$$

where $r < p$ is the rank of $\mathbf{X}^T\mathbf{X}$, and the eigenvalues $\{e_1^2, \ldots, e_r^2\}$ are in descending order. The matrix $\mathbf{Z}_r \in R^{N\times r}$ formed by the first r columns of the rotated predictor data matrix

$$\mathbf{Z} = \mathbf{X}\mathbf{U} \in R^{N\times p} \tag{6.2}$$

is then used in equation (2.25) in place of $\mathbf{X}$. The regression coefficient estimates associated with the last $p - r$ columns are then all defined to have 0 value. This is equivalent to using the generalized inverse of $\mathbf{X}^T\mathbf{X}$ in equation (2.25).

A problem with this approach is that the resulting (non-zero) coefficient estimates are likely to be highly variable owing to the fact that $\mathbf{Z}_r^T\mathbf{Z}_r$ is still likely to be poorly conditioned. This can be remedied by making the rank value r a model selection parameter to be estimated through cross-validation by analogy with (single-response) principal components regression (Massey, 1965). This approach would tend to rule out reduced rank regression and C&W–CV since several model selection parameters would then have to be estimated through sample reuse with limited data. Since it consistently outperformed FICYREG for $p < N$, we chose C&W–GCV for this combined implementation.

6.1. *Curds and Whey–Ridge Regression*

Although the technique described above for combining C&W–GCV with principal components regression provided satisfactory performance, we found that using a similar strategy based on RR worked consistently better. With this approach the coefficient matrix $\hat{\mathbf{A}}_\lambda \in R^{q\times p}$ is obtained from separate RRs of each response on the predictors

$$\hat{\mathbf{A}}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y} \tag{6.3}$$

using a common value of the ridge parameter λ for all responses. This leads to the

corresponding RR response estimates $\hat{\mathbf{y}}(\lambda) \in R^q$ through

$$\hat{\mathbf{y}}(\lambda) = \hat{\mathbf{A}}_\lambda \mathbf{x}. \tag{6.4}$$

The value $\hat{\lambda}$ of the (common) ridge parameter λ is chosen by (fivefold) cross-validation

$$\hat{\lambda} = \arg\min_\lambda \left[\sum_{i=1}^{q} \sum_{n=1}^{N} \{y_{ni} - \hat{y}_{\setminus ni}(\lambda)\}^2 \right]. \tag{6.5}$$

The C&W–RR estimates are then given by

$$\tilde{\mathbf{y}} = (\hat{\mathbf{T}}^{-1}\mathbf{D}\hat{\mathbf{T}})\hat{\mathbf{A}}_{\hat{\lambda}}\mathbf{x}, \tag{6.6}$$

$$\mathbf{D} = \text{diag}\{d_1, \ldots, d_q\}, \tag{6.7}$$

where $\hat{\mathbf{T}} \in R^{q\times q}$ is obtained by a canonical correlation analysis between the sample responses $\mathbf{Y}$ and their corresponding ridge estimates $\hat{\mathbf{Y}}_{\hat{\lambda}} \in R^{N\times q}$

$$\mathbf{Y}^{\mathrm{T}}\hat{\mathbf{Y}}_{\hat{\lambda}}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1}\hat{\mathbf{Y}}_{\hat{\lambda}}^{\mathrm{T}}\mathbf{Y}(\hat{\mathbf{Y}}_{\hat{\lambda}}^{\mathrm{T}}\hat{\mathbf{Y}}_{\hat{\lambda}})^{-1} = \hat{\mathbf{T}}^{-1}\hat{\mathbf{C}}^2\hat{\mathbf{T}}, \tag{6.8}$$

$$\hat{\mathbf{C}}^2 = \text{diag}\{\hat{c}_1^2, \ldots, \hat{c}_q^2\}. \tag{6.9}$$

The diagonal matrix $\mathbf{D}$ (6.7) is given by the C&W–GCV formula (3.12)–(3.13) with $\{\hat{c}_i^2\}_1^q$ given by equation (6.9) and with

$$\hat{r} = \frac{1}{N}\,\text{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \hat{\lambda}\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\} \tag{6.10}$$

replacing r (2.12) in equation (3.12). Note that this C&W–RR procedure generalizes C&W–GCV in that it reduces to C&W–GCV when $\hat{\lambda} = 0$.

Unlike C&W–GCV, C&W–RR is not affine equivariant in either the response or predictor spaces. Although it is equivariant under (rigid) rotations in both spaces, changing the relative scales of the responses and/or the predictors (or their linear combinations) changes the predictive model. As in ordinary RR, principal components regression and PLS, this ambiguity is usually resolved by standardizing ('autoscaling') all variables before the analysis is performed.

For poorly determined systems ($p/N \cong 1$) the least squares estimates (though defined) can be highly variable, potentially causing difficulty for procedures based on proportional shrinking like C&W–GCV and C&W–CV. The ridge estimates (6.3)–(6.5) have less variance at the expense of (additional) bias. It is therefore possible that C&W–RR may outperform C&W–GCV and C&W–CV in such poorly (but not ill-) conditioned situations. In the simulation study described in Section 5 ($p/N = \frac{1}{2}$ and $p/N = \frac{1}{4}$) C&W–RR exhibited substantially inferior performance to that of both C&W–GCV and C&W–CV. However, for substantially larger values of p/N ($\cong 1$) C&W–RR may have the best performance. This will probably depend on other aspects of the problem such as sample size and (unknown) signal-to-noise ratio. A reasonable strategy would be to compare the methods by using cross-validated error estimates as a guide.

6.2. *Simulation Study*

For $p > N$ the competitors to C&W–RR (Section 6.1) are separate RRs (Section 4.1) and two-block PLS (Section 4.4). To study their respective performance in a variety of situations we performed another (less ambitious) simulation study. For all replications the training sample size was $N = 25$. There were two values for the number of responses ($q = 5$ and $q = 10$) and two values for the number of predictor variables ($p = 50$ and $p = 100$). Two error covariance structures were studied, expressions (5.4) and (5.5), each with two values of σ^2 chosen to give (average) signal-to-noise ratios of 1.0 and 3.0 respectively. Three different signal covariance structures $\mathbf{F}$ (2.10) were studied corresponding to average correlations among the signals (2.6) and (5.11) of 0.0, 0.35 and 0.70. For each replication the predictors were generated from expressions (5.1) and (5.2) with r assigned three values: $r = 0.0$, $r = 0.90$ and $r = 0.99$. The response values were computed from equations (2.5) and (2.6) with the true coefficient values $\{a_{ij}\}$ generated in the same manner as described in Section 5.1. A full factorial design over all the above levels gives rise to 144 situations; 100 replications were performed for each. Thus, the entire simulation study comprised 14400 replications.

The performance measure used to compare the three methods is

$$\mathrm{RA}(m) = \sum_{i=1}^{q} e_i^2(m) \Big/ \min_{k=1,3} \left\{ \sum_{i=1}^{q} e_i^2(k) \right\}, \qquad m = 1, 2, 3, \tag{6.11}$$

with $\{e_i^2(m)\}_1^q$ given by equation (5.12). This measures the error squared (averaged over the responses) of each method relative to the corresponding minimum over all the methods. For each replication, equation (6.11) will have the value 1.0 for the best (minimum average error squared) method and larger values for the other two methods. The results of this simulation study are summarized by the average of equation (6.11) over the 100 replications for each of the 144 situations.

Fig. 10 shows box plots for each method of the distribution of the 144 averages of equation (6.11) over all situations. C&W–RR is seen to produce the best average error (squared), or within a few per cent of the best, in every situation. The corresponding quantity for separate RRs is typically 22% larger than the best, and that for two-block PLS is 30% larger. However, the dispersion of values for two-block PLS about its median is somewhat less than that for separate RRs.

Fig. 11 shows the (first-order) interaction effects between method (m) and the design factors of this simulation study, based on RA(m) (6.11), in the same manner as that of Fig. 9. We can see from Fig. 11(a) that for low (population) collinearity all three methods perform comparably, C&W holding a slight edge. This is due to the fact that for $p \prec N$ and low collinearity none of the three methods can produce predictions that are much more accurate than simply the response means. In higher (population) collinearity settings more accurate prediction is possible and the C&W procedure is seen to be much more dominant over the other two. This is especially the case for the highest collinearity ($r = 0.99$) where it is typically 42% better than two-block PLS and 75% better than separate RRs.

The relative advantage of C&W–RR over the other two methods is seen to increase with decreasing signal-to-noise ratio (Fig. 11(b)) and increasing dispersion among the response error variances (Fig. 11(c)). Its competitive advantage is slightly less for more responses (Fig. 11(d)) and more predictor variables (Fig. 11(f)). The

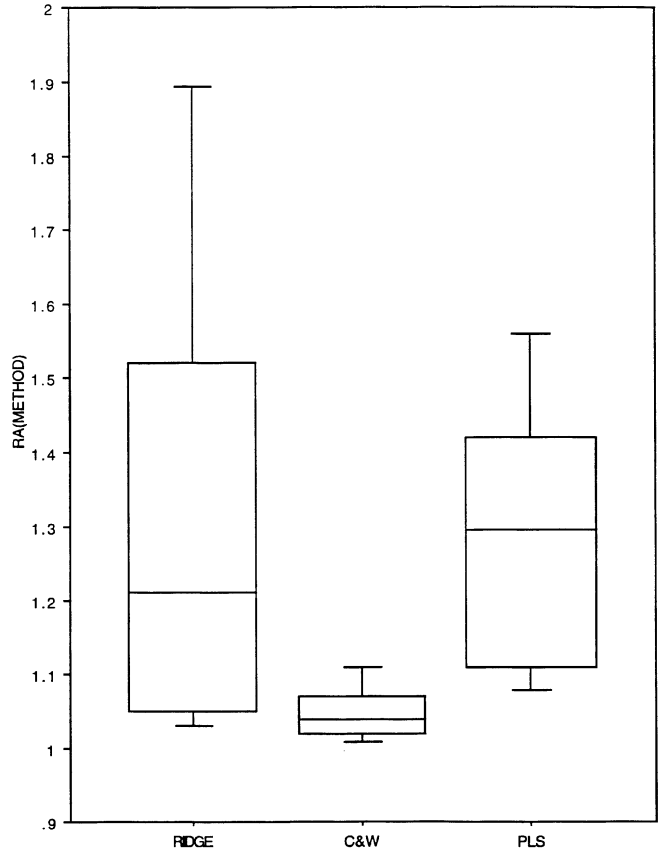

Fig. 10. Distribution over all 144 ($p \succ N$) situations of the ratio of the overall average response mean-squared error for each method, to that of the best method (6.11)

degree of correlation among the responses does not seem to affect its advantage strongly (Fig. 11(e)). The performance of C&W–RR is seen to dominate that of separate RRs and two-block PLS for every level of every factor.

7. EXAMPLES

In this section we illustrate the application of the C&W method to two published data sets and compare its performance with OLS. In a simulation study one can consider a wide range of situations and accurately estimate expected performance by averaging accuracy over many replicated samples drawn from each. A real data set by contrast represents only a single sample from one (unknown) situation. Also, the mean-squared prediction error from that single sample is unknown and must be estimated with uncertainty. This limits the substantive conclusions that can be drawn. None-the-less, empirical success on real data, though not definitive, lends some support to the merit of the approach.

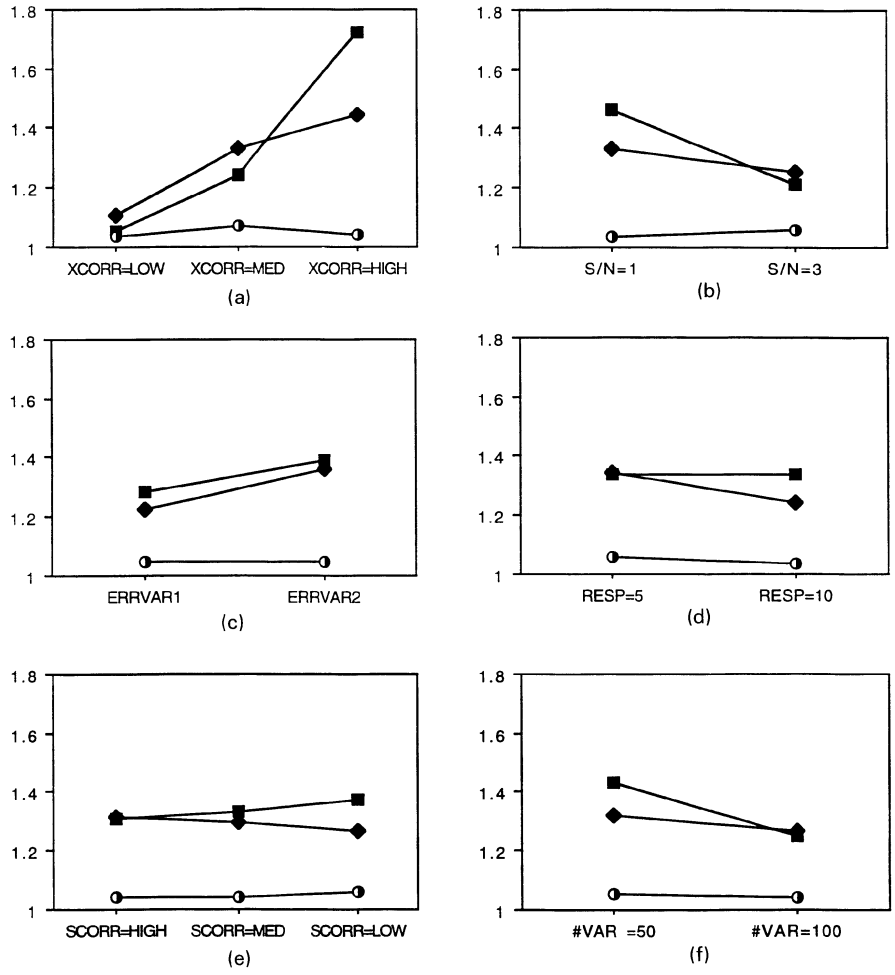

Fig. 11. Interactions of method with the other factors of the $p \succ N$ simulation design (the ordinate is the ratio of overall average response mean-squared error for each method, to that of the best method (6.11); XCORR is the predictor variable collinearity (5.1)–(5.2) (LOW, $r = 0.0$; MED, $r = 0.9$; HIGH, $r = 0.99$); SCORR is the average signal correlation (5.11) (LOW= 0.0; MED= 0.35; HIGH= 0.70)): ■, RR; ◑, C&W; ◆, PLS

7.1. *Chemometrics Example*

The chemometrics data are taken from Skagerberg *et al.* (1992). There are $N = 56$ observations, each with $p = 22$ predictor variables and $q = 6$ responses. The data are taken from a simulation of a low density tubular polyethylene reactor. The predictor variables consist of 20 temperatures measured at equal distances along the reactor together with the wall temperature of the reactor and the feed rate. The responses are the output characteristics of the polymers produced: y_1, the number-average molecular weight; y_2, the weight-average molecular weight; y_3, the frequency of long chain branching; y_4, the frequency of short chain branching; y_5, the content of vinyl groups; y_6, the content of vinylidene groups.

TABLE 1
Response correlation matrix for the chemometrics example

1.000	0.957	0.064	0.254	0.255	0.259
0.957	1.000	−0.130	0.283	0.266	0.276
0.064	−0.130	1.000	−0.502	−0.486	−0.481
0.254	0.283	−0.502	1.000	0.974	0.978
0.255	0.266	−0.486	0.974	1.000	0.976
0.259	0.276	−0.481	0.978	0.976	1.000

TABLE 2
Predictive squared error for the chemometrics example

Response	*Results from the following methods:*	
	OLS	*C&W*
1	0.566	0.550
2	1.171	0.808
3	0.259	0.250
4	0.123	0.111
5	0.277	0.193
6	0.190	0.148
Average	0.431	0.343

Because the distributions of the values of all the response variables are highly skewed to the right, the analysis was performed using the logarithms of their corresponding values. For interpretational convenience all were then standardized to unit variance. The average (absolute) correlation between the (transformed) responses is 0.48 and the correlations between the individual pairs are given in Table 1. Responses y_1 and y_2 are seen to be strongly correlated, and y_4, y_5 and y_6 form another strongly correlated group. The third response y_3 is more weakly correlated with the others.

The predictive accuracy of each method was estimated through leave-one-out cross-validation, i.e. the predictive equations were estimated using 55 of the 56 observations and squared error measured on the left-out case. This was repeated 56 times, each time leaving out a different case, and the 56 errors (squared) averaged. Note that the predictive accuracy being estimated here is larger than the corresponding mean-squared estimation error (5.12) since it includes the contribution of the irreducible error ϵ (2.5).

Table 2 shows the estimated squared prediction error for OLS (second column) and C&W–GCV (third column) for each of the (transformed) responses (rows). The C&W method is seen to improve the predictive accuracy of all the responses, with that improvement being substantial for three of them (y_2, y_5 and y_6). On the whole the C&W method decreased the squared error by about 20%. The GCV shrinkage factors (3.11) and (3.12) are $\hat{\mathbf{D}} = \text{diag}\{0.994, 0.973, 0.864, 0.172, 0.142, 0.000\}$. This indicates that the effective response dimension is around 3.

7.2. *Scottish Elections*

Brown (1980) lists electoral results for all 71 Scottish constituencies in the British general elections of February and October 1974. The raw data given in Brown (1980)

consist of the total votes for each of the four parties (Conservative, Labour, Liberal and Nationalist) in each election, together with a categorical variable listing the location of the constituency by six regions, and the size of the electorate in each constituency. The constituencies are listed in the order that they were declared in the February election. The objective is to use the February and October results from part of the constituencies to predict the remaining October results from the corresponding February data.

Following Brown (1980), we use as response variables $\mathbf{y} = (y_1, y_2, y_3, y_4)$ the difference between the vote in October and February for each party divided by the electorate size. There are $p = 7$ predictor variables. The first four are the votes in February for each party divided by the electorate size. The next three are binary variables:

(a) $x_5 = 0.5$ if Liberal intervenes (Liberal vote in October $\succ$ 0; Liberal vote in February $=$ 0); otherwise $x_5 = 0$;
(b) $x_6 = 0.5$ if the constituency is in a rural area; otherwise $x_6 = 0$;
(c) $x_7 = 0.5$ if Labour or Nationalist won in February and $|x_2 - x_4| \prec 0.2$; otherwise $x_7 = 0$.

The average (absolute) correlation between the responses is 0.435. The response correlation matrix is given in Table 3. We use the data from the first 30 constituencies to form October prediction equations and then test these equations on the data from the remaining 41 constituencies. Table 4 gives the mean-squared prediction error for OLS (third column) and C&W (fourth column) multiplied by 1000. As a base-line, we include the predictor consisting of the average of each October response over the 30 constituencies (second column). The GCV shrinkage factors (3.11) and (3.12) are $\hat{\mathbf{D}} = \text{diag}\{0.96, 0.52, 0.20, 0.00\}$ indicating an effective response dimensionality of less than 2.

TABLE 3
Response correlation matrix for the Scottish election example

1.000	0.537	−0.415	−0.495
0.537	1.000	−0.376	−0.393
−0.415	−0.376	1.000	−0.395
−0.495	−0.393	−0.395	1.000

TABLE 4
Predictive squared-error (times 1000) for the Scottish election example

Response	*Mean*	*Results from the following methods:*	
		OLS	*C&W*
1	1.10	1.83	0.98
2	0.43	0.87	0.58
3	2.04	0.30	0.38
4	2.15	2.67	1.92
Average	1.43	1.42	0.97

8. CONCLUSION

The results presented in this paper strongly suggest that the conventional (statistical) wisdom, that we should avoid combining multiple responses and treating them in a multivariate manner, may not be the best advice. Our simulation studies indicate that the best of the multiple-response procedures considered here can provide large gains in expected prediction accuracy (for each individual response), over separate single-response regressions, with surprisingly little risk of making things worse. In the fields of neural networks and chemometrics, by contrast, the conventional wisdom has always been in favour of combining multiple responses. The results of this paper generally validate that intuition, but it is not clear that the respective recommended approaches in each of those fields best serve that purpose. For example, the two-block PLS approach commonly used in chemometrics was seen in our simulation studies to provide generally lower accuracy than separate RRs.

The C&W procedure tends to improve the expected prediction accuracy for every response. This suggests the intriguing prospect that, even when there is only a single response of interest, if variables are available that are correlated with it, then prediction for the response of interest may be improved by introducing the other variables as additional responses. Of course, if the values of these variables will also be available for (future) prediction, they should be regarded as predictors (rather than responses) and included in the regression equation. In some circumstances, however, the (training) data may include measurements of variables whose values will not be available in the prediction setting.

In the neural network literature such variables are known as 'coaches'. These are variables whose values are available for use during training but not available for future prediction. Examples might be expensive or difficult-to-obtain medical measurements that were available at the hospital where the training data were collected, but not available in the field or at smaller hospitals where the predictions are made. In financial forecasting, 'future' values of other quantities, thought to be correlated with the response, might be included as coaches. The results presented in this paper suggest that the inclusion of such coaching variables as extra responses during training using C&W may indeed improve prediction accuracy.

ACKNOWLEDGEMENTS

The work of Leo Breiman was partially supported by National Science Foundation grant DMS-9212419; that of Jerome Friedman by Department of Energy contract DEAC03-76SF00515.

REFERENCES

Anderson, T. W. (1957) *An Introduction to Multivariate Analysis*. New York: Wiley.

Brown, P. J. (1980) Aspects of multivariate regression (with discussion). In *Bayesian Statistics* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 247–292. Valencia: Valencia University Press.

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

——— (1982) Multivariate regression shrinkage estimators with unknown covariance matrix. *Scand. J. Statist.*, **9**, 209–215.

Copas, J. B. (1983) Regression, prediction and shrinkage (with discussion). *J. R. Statist. Soc.* B, **45**, 311–354.

———(1987) Cross-validation shrinkage of regression predictors. *J. R. Statist. Soc.* B, **49**, 175–183.
Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, 317–403.
Frank, I. E. and Friedman, J. H. (1993) A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, **35**, 109–148.
Garthwaite, P. H. (1994) An interpretation of partial least squares. *J. Am. Statist. Ass.*, **89**, 122–127.
Golub, G. H. and van Loan, C. F. (1989) *Matrix Computations*. Baltimore: Johns Hopkins University Press.
Hoerl, A. E. and Kennard, R. W. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **8**, 27–51.
Izenman, A. J. (1975) Reduced-rank regression for the multivariate linear model. *J. Multiv. Anal.*, **5**, 248–264.
James, W. and Stein, C. (1961) Estimation with quadratic loss. In *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability* (ed. J. Neyman), vol. I, pp. 361–379. Berkeley: University of California Press.
Massey, W. F. (1965) Principal components regression in exploratory statistical research. *J. Am. Statist. Ass.*, **60**, 234–246.
van der Merwe, A. and Zidek, J. V. (1980) Multivariate regression analysis and canonical variates. *Can. J. Statist.*, **8**, 27–39.
Skagerberg, B., MacGregor, J. and Kiparissides, C. (1992) Multivariate data analysis applied to low-density polyethylene reactors. *Chemometr. Intell. Lab. Syst.*, **14**, 341–356.
Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc.* B, **36**, 111–147.
Wold, H. (1975) Soft modeling by latent variables; the nonlinear iterative partial least squares approach. In *Perspectives in Probability and Statistics: Papers in Honour of M. S. Bartlett* (ed. J. Gani). New York: Academic Press.
Zidek, J. (1978) Deriving unbiased risk estimators of multinormal mean and regression coefficient estimators using zonal polynomials. *Ann. Statist.*, **6**, 769–782.

DISCUSSION OF THE PAPER BY BREIMAN AND FRIEDMAN

Paul H. Garthwaite (University of Aberdeen): We have been presented with a constructive and useful paper and the authors are to be congratulated. Multivariate regression methods are seldom used in practice — even when there are data for which they would be appropriate, separate univariate regressions are generally used instead. This is unfortunate, as multivariate regression should enable individual regressions to 'borrow strength' from each other and hence it should improve the accuracy of their predictions. This paper has introduced three new methods for multivariate regression and, largely through simulation, it has demonstrated that they can bring very clear benefits. This should encourage practitioners to apply the methods rather than to use univariate regression methods.

Four of the methods described in the paper (C&W–GCV, C&W–CV, reduced rank regression and FICYREG) give prediction equations that have the form

$$\tilde{\mathbf{y}} = \hat{\mathbf{T}}^{-1}\hat{\mathbf{D}}\hat{\mathbf{T}}\hat{\mathbf{y}}.$$

The methods differ only in their choice of $\hat{\mathbf{D}}$, the diagonal matrix of shrinkage factors. The first method, C&W–GCV, assumes that the prediction equation has the form $\tilde{\mathbf{y}} = \mathbf{B}\hat{\mathbf{y}}$, that observations are independent and identically distributed, so that cross-validation may be used, and that the diagonal elements of $\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ can reasonably be approximated by their average. From these fairly minimal assumptions, Section 3.1 uses elegant mathematics to show that $\hat{\mathbf{B}}$ may be expressed as $\hat{\mathbf{T}}^{-1}\hat{\mathbf{D}}\hat{\mathbf{T}}$ and to determine a formula for the non-zero elements of $\hat{\mathbf{D}}$ (equation (3.12)).

The approach adopted by C&W–CV is more direct. It assumes that $\tilde{\mathbf{y}} = \hat{\mathbf{T}}^{-1}\hat{\mathbf{D}}\hat{\mathbf{T}}\hat{\mathbf{y}}$ and estimates $\hat{\mathbf{D}}$ by cross-validation. This method is more flexible than reduced rank regression, FICYREG and C&W–GCV since, in principle, it could yield the same estimate of $\mathbf{D}$ as any one of these other methods if cross-validation suggested that it would be optimal. Hence, C&W–CV may be expected to perform at least as well as these other methods when there are sufficient data for cross-validation to give a good estimator of $\mathbf{D}$. The simulations show that moderately large sample sizes ($N = 100$ and $N = 400$) can be sufficient

in some circumstances and Fig. 9(c) suggests that smaller sizes may sometimes also be adequate. In the simulations, the degree of improvement given by the C&W methods over other methods is impressive.

Ordinary least squares (OLS) is an inadmissible procedure for predicting a single response variable and, as shown in another nice discussion paper co-authored by Professor Friedman (Frank and Friedman, 1993), several univariate biased regression methods dominate it substantially. Moreover, OLS cannot be used if the sample size N is less than the number of predictor variables (p) and OLS estimates are very poor if N is little bigger than p. Hence, it is natural to consider multivariate predictors of the form

$$\tilde{\mathbf{y}} = \hat{\mathbf{T}}^{-1}\hat{\mathbf{D}}\hat{\mathbf{T}}\hat{\mathbf{y}}_{\mathrm{b}},$$

where $\hat{\mathbf{y}}_{\mathrm{b}}$ is the vector of response estimates given by a *biased* univariate regression method, $\hat{\mathbf{D}}$ is diagonal and $\hat{\mathbf{T}}$ is given by an equation similar to equation (6.8). Intuitively, one feels that there should be a method of determining $\hat{\mathbf{y}}_{\mathrm{b}}$ and $\hat{\mathbf{D}}$ that compares well both with C&W–CV and with univariate regression methods for all sample sizes.

One form of this method, C&W–RR, is suggested in Section 6. In simulations, for underdetermined systems it compared well with two-block partial least squares and separate univariate ridge regressions, but its performance was much poorer than that of C&W–CV and C&W–GCV when N was substantially bigger than p. I was surprised by the poor performance of C&W–RR. One possible explanation is that $\hat{\mathbf{D}}$ is meant to be a shrinkage matrix, so each component of $\hat{\mathbf{y}}_{\mathrm{b}}$ should require shrinkage. C&W–RR forces the ridge parameter λ to have the same value for each of the univariate ridge regressions that determine $\hat{\mathbf{y}}_{\mathrm{b}}$. The value that is optimal overall (assuming that no further shrinkage is to be applied) is likely to give too much shrinkage to some components of $\hat{\mathbf{y}}_{\mathrm{b}}$ and these may need to be unshrunk (stretched?) by $\hat{\mathbf{D}}$. Overshrinking followed by stretching sounds a recipe for reduced performance. (The situation is somewhat more complicated than this, as combining univariate estimates by using multivariate regression may increase the degree of shrinkage that is optimal.)

While discussing C&W–RR, could I ask the authors to mention in their response the route that leads to equations (6.6)–(6.10), which detail the estimator given by C&W–RR. Do they start with an equation similar to equation (3.1), but with $\hat{\mathbf{y}}_{\backslash nk}$ replaced by a ridge estimate, and use only the approximation that the diagonal elements of $\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}$ equal their average value? Stronger assumptions or further approximation might contribute to the poorer performance of C&W–RR.

The direct approach adopted by C&W–CV makes it straightforward to extend the method to use with any biased estimator. $\hat{\mathbf{T}}$ may be determined from equation (6.8) with $\hat{\mathbf{Y}}_{\hat{\lambda}}$ replaced by the estimates of $\mathbf{Y}$ that the biased estimator gives. $\mathbf{D}$ could then be estimated by cross-validation. My earlier comments suggest that each component of $\hat{\mathbf{y}}_{\mathrm{b}}$ should be estimated separately (for example, if ridge estimators are used, then λ should be allowed to vary for the different components) and it may be desirable to use estimators which are cautious in their degree of shrinkage.

My experience with multivariate regression problems stems largely from the task of predicting the concentration of several constituents in a sample from its near infra-red reflectance (NIR) spectrum. This task is alluded to briefly in Section 1 and is a common potential application of multivariate regression methods. I shall restrict myself to two comments about it. The first is that NIR data are usually analysed as a standard regression model even when they should be analysed as a calibration problem. The robustness of multivariate regression methods to misuse in this way merits study through simulation, as such misuse seems likely if multivariate regression gains popularity. The second comment is to note that NIR data often consist of measurements on 700 or more X-variables and fewer than 15 Y-variables. Thus there is a clear need for methods that can be used with underdetermined systems. Also, there will often be adequate data to estimate parameters by cross-validation provided that the number of parameters is determined by the dimension of $\mathbf{y}$, and not $\mathbf{x}$. (This would be the case for methods that combine C&W–CV with appropriate biased regression procedures.)

As you can see, I have found the paper very stimulating and it gives me great pleasure to propose the vote of thanks.

Philip J. Brown (University of Kent, Canterbury): On first reading this paper, it felt as if I had entered a time warp. I savoured all the old debates of the 1960s and 1970s. I congratulate the authors for breathing new life into Stein shrinkage estimation for the general linear model. They use a canonical variates reduction and the class of equivariant estimators, first introduced by Jim Zidek, and consequently they implicitly assume normality. They have, however, elegantly added cross-validated

shrinkage choice and enriched the class. The results of their simulations seem to show significant gains. Whether they have dispelled some of the doubts that surfaced in the earlier debate though remains to be seen.

The focus of the paper is that of prediction as opposed to estimation of the regression coefficients in the linear model. Although this is appropriate for some problems, it may need to be overlaid with knowledge of the science. Invariance may not be persuasive where explanatory variables have meaning. Ordinary least squares is the yardstick for comparison but in reality is seldom used in its pure form, except perhaps in simulation studies to justify competing estimators. 'Shrinkage' in the form of variable selection might more often be applied. I am sympathetic to more continuous forms of shrinkage, as adopted in this paper, but with a motivation and applicability that is perhaps more flexible, inferential and generalizable to other error structures.

The general linear model adopted may be written as

$$Y = XA + E$$

where Y $(n \times q)$ is the random matrix of n observations on q responses, X $(n \times p)$ is the regressor matrix of conditionally fixed p explanatory variables *common* to each response. The matrix of errors E $(n \times q)$ would typically be assumed uncorrelated across observations and correlated among responses, although not the latter in this paper. The matrix of regression coefficients A $(p \times q)$ is unknown and the problem can be viewed as estimating pq unknown means of pq responses, seen perhaps more transparently through a singular value decomposition $X = Q\Lambda^{1/2}P$ where

$$\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_p\} \qquad \lambda_1 \geqslant \ldots \geqslant \lambda_p \geqslant 0,$$

Q $(n \times p)$ and P $(p \times p)$ being orthonormal matrices. This leads to

$$Z = Q^{\mathrm{T}} Y = \Lambda^{1/2} A^* + E^*$$

with $A^* = PA$ $(p \times q)$ and Z $(p \times q)$, where we have ignored that part of the orthogonal transformation of Y which produces $n - p \times q$ responses with zero means utilizable to estimate the covariance structure. Shrinkage to zero (or another value) can be across the q responses, across the p regressors or most naturally through a two-way exchangeability of both rows and columns, after suitable scaling. What emerges clearly from the earlier decision theoretic debate is that the loss function is central to the choice and performance of an estimator, and it should provide an amalgamation and add up to the component losses in a fairly even-handed way for strict dominance over the whole parameter space; see Brown (1966).

In the context of the precision asymmetry imposed by the unequal singular values $\Lambda^{1/2}$ then prediction loss in the rather restricted sense of predicting at past X compensates conveniently for the precision differences of different directions of X-space. To illustrate this, suppose that we wish to predict m future observations at X_{f} $(m \times p)$; then, apart from the new error component of Y_{f} $(m \times q)$, the prediction mean-square error

$$\text{trace}[E\{(Y_{\mathrm{f}} - \hat{Y}_{\mathrm{f}})^{\mathrm{T}}(Y_{\mathrm{f}} - \hat{Y}_{\mathrm{f}})\}]$$

is

$$\text{trace}[E\{(A - \hat{A})^{\mathrm{T}} X_{\mathrm{f}}^{\mathrm{T}} X_{\mathrm{f}} (A - \hat{A})\}]. \tag{1}$$

The authors take expectation over the ancillary X_{f} and in the averaging process regard them as typical of past X. Their claims of gains in all q responses would seem only possible through the weighted averaging effect of prediction loss across the p regressors and certainly not possible for all parameter values when $p \leqslant 2$ and such amalgamation is insufficient. The amalgamation role of the loss function leads to the wonderland of unrelated responses collected just to improve prediction loss.

If we are really interested in regression parameters and not prediction at the past then I would advocate a more open weighting of the parameter space rather than indirectly through the choice of loss function. The assumption of multivariate normality and for example the prior assumption of q-variate random effects about 0 for each of the rows of the regression matrix A leads to a multivariate version of

ridge regression as in Brown and Zidek (1980, 1982),

$$\hat{\alpha}_i(K) = \hat{\alpha}_i \lambda_i (\lambda_i I_q + K)^{-1},$$

where α_i is the ith row, $i = 1, \ldots, p$, of A^*, the orthogonally transformed matrix of $p \times q$ regression coefficients. Here $K = \Gamma_\alpha^{-1} \Sigma$, the matrix ratio of error to random effects covariance matrices. Richer choices might focus on a two-way exchangeability of rows and columns of regression matrix A. The nice thing about ridge estimators, touched on in Section 6, is that they shrink strongly in ill-estimated (small λ_i) directions and relatively less in well-estimated directions, and indeed they were first introduced for numerical stability in linear regression. In this sense they form a continuous and less extreme form of selection shrinkage. Simple estimates of the $q \times q$ matrix K are also possible not requiring iterations or expensive cross-validation. They are very different from the proportional shrinkers of C&W–CV and C&W–GCV. A version, ridge with K diagonal, has been used for election night forecasting in every British general and European election since 1974 (see for example Brown and Payne (1975, 1984) and Payne and Brown (1981)) and looks like doing so in 1996–97 also. Here early in the declaration the number of available observations is small and builds up through the election night. The regressor variables are ordered for inclusion by their importance. Speculative variables are subjected to stronger shrinkage than a few *a priori* established variables.

I would have liked to have seen more of this wider selection of estimators compared in the simulation study. Both predictive and estimative ($X_f^T X_f$ or V in expression (1) replaced by I_p) loss functions would be of interest with a wider range of implied regression coefficients, including 'lumpy' arrangements, not just those that comply with the symmetries of cross-validatory choice and prediction loss. And how would C&W–CV fare with the odd outlier or non-normality?

It is traditional that the seconder of the vote of thanks is allowed some critical licence, but let it not cover up my genuine enjoyment of this paper. It gives me great pleasure to second the vote of thanks.

The vote of thanks was passed by acclamation.

David J. Hand (The Open University, Milton Keynes): This paper makes two valuable contributions. Firstly, it provides a unification of several multivariate multiple-regression shrinkage methods via the transformation to canonical variate space. And, secondly, it finds the optimal extent of that shrinkage. The simulation comparisons and the real data examples support the contention that the method proposed is optimal.

However, I have a couple of questions relating to the simulations.

Firstly, to me the striking thing about Fig. 9 is the general lack of interaction between the method and the other factors you examine. The curves in these figures stack up, one above the other. However, you draw attention to some apparent relatively small interactions. In no cases are these interactions such that a method other than your new proposals is superior *in the ranges of the factors that you consider*. Nevertheless, the mere existence of such interactions leads to the suggestion that some other method might be the best outside the ranges of the factors that you have examined. I wondered whether you had any comment on this.

My second question arises from some joint work that I am carrying out with Wojtek Krzanowski on assessing the performance of classification rules. Almost all of your simulation conclusions are based on the 120 mean values, each an average of 250 replications. However, each of those 250 replications is based on a *different*, randomly generated, true set of regression functions. Now surely it is inappropriate to average over different true situations in this way. What one really wants to know is the performance for a fixed (if unknown) regression function. Does not this mean that the proper thing to do is to average over random data sets, each of which arises from the *same* true regression situation, summarizing your results in terms of factors describing those true situations? That is, should you not condition your simulations on the true regressions?

Svante Wold (Umeå University): Breiman and Friedman develop shrunken canonical correlation models (curds and whey methods, C&W), which predict better than separate ordinary least squares (OLS) regression models for each response, both in simulations and in two published data sets.

In the simulations they also compare the predictive performance of C&W with that of other multivariate regression methods, including OLS and two-block partial least squares (PLS) (Nomikos and MacGregor, 1995; Wold, 1995; Wold *et al.*, 1984, 1987). The simulation results indicate that PLS

does not predict as well as the other methods. For the two examples no comparison is made with PLS or other methods.

Breiman and Friedman explicitly refer to PLS as it is used in chemometrics, e.g. with chemical manufacturing processes, environmental chemistry, drug design and medicinal chemistry, with typically 5–10 *y*s, sometimes up to several hundreds. Hence, I expected them to use PLS as it is used today. Instead they use cross-validation (CV) based *only* on the total prediction error (over all *y*s) to estimate the number of PLS components in the simulations. This 'total' CV rule often gives too few PLS components and is why PLS looks less good in the simulations, and accounts for the so-called 'Robin Hood' effect. The total CV rule should be supplemented by a criterion based on the CV prediction error of the separate responses. This has been done in chemometrics since around 1987. The improvement by this second rule is seen in example 2 (Table 5).

The simulations are not representative for any real data I have seen. The predictors (X) are generated to be full rank, $\min(N, p)$, and without noise, and so that the main canonical axes of X (with highest explained variance) usually not parallel to those of Y. In contrast, PLS is designed for the common situation where the main canonical axes of X and Y are nearly parallel.

In the two examples, PLS predicts better than C&W and OLS (Table 5). The C&W solution, like OLS, suffers from 'inverse Robin Hood' effects in both examples, where noisy responses are 'sacrificed' for better modelled responses, making the average prediction error deteriorate. In example 2 both OLS and C&W predict response 2 even worse than the mean! PLS shows no apparent 'Robin Hood' effect.

There are several other criteria for evaluating the success of a model, e.g. parsimony, bias, interpretability and diagnostics for outliers and other data inhomogeneities, but they are not mentioned by Breiman and Friedman.

How to use the information in correlated responses is important, and the approach proposed by the authors is interesting. However, they also make strong and general statements about their C&W method in comparison with other methods. These statements are based on unrealistic simulations, using an inadequate implementation of CV with PLS, and moreover contradicted by their own examples, and hence are unwarranted and misleading.

TABLE 5
Breiman and Friedman's Tables 2 and 4, extended for multiresponse PLS†

Predictive variance	*OLS*	*C&W*	*Mean*	*PLS2, A = 7*	*PLS2, A = 1*	*PLS2, A = 2*	*CV, Q2 (15 groups)*	
							A = 1	*A = 2, change*
Example 1: N = 56, p = 23								
1 inv MW	0.57	0.55		0.20				
2 inv MN	1.17	0.81		0.28				
3 LCB	0.26	0.25		0.28				
4 SCB	0.12	0.11		0.28				
5 VNL	0.28	0.19		0.27				
6 VND	0.19	0.15		0.28				
Average	0.43	0.34		0.26				
Example 2: $N_1 = 30$, $N_2 = 41$, p = 7								
1 YCons	1.83	0.98	1.10		0.89	0.91	0.176	−0.076
2 YSoc	0.87	0.58	0.43		0.43	0.40	0.393	−0.047
3 YLib	0.30	0.38	2.04		0.41	0.39	0.798	0.028
4 YNat	2.67	1.92	2.15		2.03	1.95	0.169	0.007
Average	1.42	0.97	1.43		0.94	0.91	0.38	−0.02

†PLS2, multiresponse PLS; CV indicates seven significant components in example 1 and one significant component in example 2, using the 'total' criterion; the 'single-response' CV criterion indicates seven significant components in example 1 and two significant components in example 2 (last two columns); Q2 denotes the cross-validated multiple-correlation coefficient (R^2). The first column shows the numbers and names of the responses y_m.

D. R. Cox (Nuffield College, Oxford): This is a very interesting paper. The idea that, if we are regressing Y_1 on a set of explanatory variables x_1, then there may be information in the regression of Y_2 on a different set of explanatory variables x_2 goes back to Zellner (1962) in his discussion of seemingly unrelated regressions. The connection between the efficiency of separate least squares and canonical correlations is set out by Zellner and Huang (1962). Quite generally if in multivariate normal theory linear regression

$$E(Y) = x\beta,$$

where Y is $p \times 1$, z is $p \times q$ and β is $p \times q$ if some individual elements of β are constrained to 0, the canonical statistic for the regression coefficient typically remains of dimension pq even though the canonical parameter is of reduced dimension. We are in a curved exponential family and in general efficient estimation is not based on separate least squares unless the graph theoretical structure connected with the implied independences is of very special form.

The present paper shows that in the different context of mean-square error prediction similar gains are possible without explicit assumptions about the elements of β. To understand the implicit assumptions on β, and to address the important question about when the method will not be very effective, a Bayesian formulation is probably needed at least qualitatively. What will happen if the configuration of the elements of β is such that in the canonical co-ordinates the elements for different dimensions are highly correlated? What will happen if the distribution is very long tailed? It is known (Dawid, 1973) that the commonly effective device of shrinkage towards the mean is inappropriate in the latter case.

The approach of the paper is severely empirical. In examples such as the chemometric one, leaving aside the artificiality of adding random numbers to the results of a deterministic computation, a study of non-linear and interactive effects, which must surely be present and which are sometimes crucial for interpretation, would be easier by the more conventional approach of using scientific judgment to reduce the temperature profile to a few meaningful summary statistics first. How would the authors advise introducing such effects into their general approach?

James V. Zidek (University of Kent, Canterbury): I hope that this paper will lead to increased recognition of the need to combine information across regression analyses. That need is so obvious in multivariate regression analyses that the continuing use of least squares seems inexplicable. But, even when the regression responses are independent, Bayesian and non-Bayesian considerations have shown the desirability of pooling. The Stein effect demonstrates this convincingly when many parameters are involved and it may be advantageous even in simpler situations.

Suppose, for example, that items are to be independently sampled from two populations governed by simple regression models,

$$Y_1 = X\beta_1 + \epsilon_1 \quad \textit{and} \quad Y_2 = X\beta_2 + \epsilon_2,$$

X being 'fixed'; the ϵs are independent $N(0, \sigma^2)$ residuals. If the β_i are thought similar, how might a non-Bayesian use the data from population 2 in estimating the slope parameter for population 1, β_1?

The 'relevance weighted likelihood (REWL)' proposed by Hu and Zidek (1995) provides a generalizable answer:

$$l_1(\beta_1) l_2^{p_2}(\beta_1);$$

here β_1 has replaced β_2 in the likelihood for the data from the Y_2-experiments but the result has been discounted through the 'relevance weight', $p_2 \in [0, 1]$. We readily find the maximum REWL estimator $\hat{\beta}$ to be

$$\hat{\beta} = (\mathbf{S}_1 + p_2\mathbf{S}_2)^{-1}(\mathbf{X}_1'\mathbf{Y}_1 + \mathbf{X}_2'\mathbf{Y}_2),$$

where $\mathbf{X}_i = (X_{i1}, \ldots, X_{in_i})'$, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ and $\mathbf{S}_i = \mathbf{X}_i'\mathbf{X}_i$, $i = 1, 2$.

Perhaps surprisingly, if $(\beta_1 - \beta_2)^2 < C$ for some constant C, we can select $p_2 > 0$ so that $E(\hat{\beta} - \beta_1)^2 < E(\hat{\beta}_1 - \beta_1)^2$, whatever β_1 and β_2. In fact if $(\beta_1 - \beta_2)^2 < \mathrm{cov}^{-1}(\hat{\beta}_1) + \mathrm{cov}^{-1}(\hat{\beta}_2)$ we may take $p_2 = 1$.

These results extend to the problem addressed by the paper and enable seemingly unrelated multivariate regressions to be pooled. It would be interesting to see how one could combine the two approaches.

REWL can be used more generally in the domains of spline smoothing for models like $Y = m(t) + \epsilon$, $t \in (0, \infty)$, measurements being taken at $t = t_1, \ldots, t_n$, Kalman filtering with models like $Y = X\beta(t) + \epsilon$ and non-linear regression. How could the methods suggested here be adapted for use in these broader domains as well?

These considerations lead me to three questions.

(a) Have the authors tested these linear theory methods on simulated samples from less congenial populations than the normal and if so how did they compare?
(b) What relevance have the findings discussed here after the future X-values become known and are therefore no longer random?
(c) On the basis of the lessons provided by their study, can the authors suggest an approach to combining non-linear regression results such as those obtained for independent clusters of correlated count data?

C. J. F. ter Braak (Dienst Landbouwkundig Onderzoek, Wageningen): In this very stimulating paper, the authors show large gains of their curds and whey (C&W) procedure over ordinary least squares (OLS). My impression is that some of this gain comes from the shrinkage, even if performed univariately, but that the major part comes about if the matrix of regression coefficients $\mathbf{A}$, or rather $\mathbf{A}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{A}^{\mathrm{T}}$, is 'close' to a reduced rank. This happens in the simulations with $q = 20$, because by equation (5.6) $\mathrm{rank}(\mathbf{A}) \leqslant 10$.

In contrast with the authors, I am worried about the case '$q \approx N$'. The canonical system will be very unstable; $p + q - (N - 1)$ canonical correlations will be equal to 1 and as many canonical variates have much arbitrariness. How can C&W–(G)CV then take advantage of a low rank $\mathbf{A}$? A possible remedy is to define the canonical variates by the eigenanalysis of $\mathbf{\Sigma}^{-1}\hat{\mathbf{Y}}^{\mathrm{T}}\hat{\mathbf{Y}}$, as in reduced rank regression (4.3), but with $\mathbf{\Sigma}$ replaced by $\mathbf{I}_q$ (redundancy analysis) or, if $p < N$, by a diagonal matrix with the sample error variances on the diagonal (ter Braak and Looman, 1994). The C&W method allows biplot graphical displays of $\tilde{\mathbf{A}}$ and the fitted values, as does reduced rank regression.

If $q > p$, the number of shrinkage factors in C&W–(G)CV is p, whereas in (separate) univariate shrinkages it is q. If $p = 1$, then paradoxically $\tilde{a}_i = \hat{d}\hat{a}_i$ (no contributions of other $\hat{a}_j, j \neq i$) with $\hat{d}$ defined by equation (3.15) or by equation (3.12) in C&W–GCV with $\hat{c}_1 = \hat{R}$, the multiple correlation between x and $\mathbf{y}$. If also $q \approx N$, then $\hat{R} \approx 1$, so that C&W–GCV shrinks very little. C&W–GCV is then worse than univariate shrinkage, which shrinks each $\hat{a}_i$ individually on the basis of simple correlations, and also worse than C&W–CV.

In standard terminology (e.g. Dillon and Goldstein (1984)), the rows of $\mathbf{T}$ are termed canonical weights or canonical coefficients; the columns of $\mathbf{T}^{-1}$ are termed canonical loadings.

In the multivariate linear model, errors will often be correlated because of 'omitted' predictor variables. How would such omissions, or non-diagonal $\mathbf{\Sigma}$ in general, influence the expected gain of the C&W method compared with OLS? What is known about the standard errors of estimates of the shrunken regression coefficients and predictions?

Mervyn Stone (Ruislip): The miraculous results in this paper all stem from the authors' intuition expressed as prescription (1.1). My intuition likes to view equation (1.1) as a flexible covariance adjustment of a (uniformly) flattened ('proportional shrinkage') least squares predictor, i.e.

$$\tilde{y}_i = [(1 - b_{ii})\bar{y}_i + b_{ii}\hat{y}_i] + \sum_{k \neq i} b_{ik}(\hat{y}_k - \bar{y}_k). \tag{2}$$

Even though the y_k-values are not available for the adjustment, their surrogates $\hat{y}_k$ are — and this paper decisively shows how they can be used to reduce, on average, the error in the predictor $[\cdot]$. Indeed, the value of the adjustment may lie more in the reduction of (conditional on x) bias in $[\cdot]$ than of variance. The optimal choice of $\{b_{ik}\}$ is conditional on the prescription — whose optimality itself remains an open question. In equation (2) we can see that two distinct processes are at work. Although the optimizations of b_{ii} and $\{b_{ik}, k \neq i\}$ will interact, some idea of their relative importance in any particular situation might be gleaned from additional comparison of the 'curds and whey' approach with the special case of 'no covariance adjustment' ($\{b_{ik} = 0, k \neq i\}$).

The paper cleverly reveals how, when $\mathbf{Q}$ is unknown, cross-validation generates a realistic simulacrum of the canonical representation of the optimization for known $\mathbf{Q}$. But is that representation more than descriptive? For the simplest case, $q = 2$, we find, reassuringly, that $b^*_{12} = 0$ (i.e. $\tilde{y}_1$ receives no support

from $\hat{y}_2$) if and only if $f_{12}/f_{22} = \sigma_{12}/\sigma_{22}$, which is equivalent to $y_2 = \alpha + \beta y_1 + u$ where u is uncorrelated with everything else (i.e. y_1 is *linearly sufficient* for y_2). Less accessible are the sufficient conditions $f_{12}\sigma_{12} < f_{11}\sigma_{22}$ for $b_{11}^* > 0$ and $f_{12}\sigma_{12} < f_{22}\sigma_{11}$ for $b_{11}^* < 1$ that together ensure 'flattening', i.e. $0 < b_{11}^* < 1$.

As the authors recognize, the general question of whether responses should be *married* (polygamously if $q > 2$) or stay *single* goes well beyond the framework of their paper. Brooks *et al.* (1994) addressed the marital question and found only small, somewhat inconclusive differences.

Primitive mathematical examples may also widen understanding of the possible gains of marriage. In the following two examples — one discrete, the other continuous — $p = 1$, $q = 2$ and x generates y_2, which in turn generates y_1 (the response of primary interest).

(a) $x \sim \text{Bernoulli}(\frac{1}{2})$, $y_2 \sim \text{Bernoulli}\{\alpha x + \beta(1-x)\}$, $y_1 \sim \text{Bernoulli}\{\delta y_2 + \gamma(1-y_2)\}$. Taking $\hat{y}_1$ to be the maximum likelihood estimator of $\Pr(y_1 = 1|x)$, we can use the (x, y_1, y_2) data when 'married' but only the (x, y_1) data when 'single'. The asymptotic relative efficiency (ARE) of married over single depends on α, β, γ and δ, but, if these are random uniform(0, 1), the AREs are found by simulation to be flatly spread from 1 upwards to just over 2.

(b) $x \sim$ normal, $y_2 \sim N(\alpha + \beta x, \sigma^2)$, $y_1 \sim N(\gamma + \delta y_2, \tau^2)$. Taking $\hat{y}_1$ to be the maximum likelihood estimator of $E(y_1|x)$, the ARE has a minimum of 1 when $\text{corr}(x, y_2) \to 1$ and a maximum of 2 when $\text{corr}(x, y_2) \to 0$ and $\text{var}(y_1|x)/\text{var}(y_1|y_2) \to 1$.

Qiwei Yao (University of Kent, Canterbury): An idea prompted by this paper is to consider predicting multivariate responses in a more general set-up such as $\mathcal{Y}_i = f_i(\mathcal{X}_i) + \epsilon_i$, $i = 1, \ldots, q$, where $\mathcal{X}_i$ may be a vector and $\mathcal{Y}_1, \ldots, \mathcal{Y}_q$ are probably dependent on each other. Suppose that for each i an estimator $\hat{f}_i(\cdot)$ has been formed based on the observations from the ith model only. A natural prediction for $\{y_i = f_i(\mathbf{x}_i) + \epsilon_i, i = 1, \ldots, q\}$ is $\hat{\mathbf{y}} = (\hat{f}_1(\mathbf{x}_1), \ldots, \hat{f}_q(\mathbf{x}_q))^{\mathrm{T}}$. Let

$$\mathbf{B}^* = \arg\min_{\mathbf{B}}(E||\mathbf{y} - \mathbf{B}\hat{\mathbf{y}}||^2) = E(\mathbf{y}\hat{\mathbf{y}}^{\mathrm{T}})\, E(\hat{\mathbf{y}}\hat{\mathbf{y}}^{\mathrm{T}})^{-1}. \tag{3}$$

Is $\tilde{\mathbf{y}} = \mathbf{B}^*\hat{\mathbf{y}}$ a better prediction? Intuitively, the existence of the dependence among different models would ensure that $\tilde{\mathbf{y}}$ is different from $\hat{\mathbf{y}}$. It is plausible that $E(\hat{\mathbf{y}}\hat{\mathbf{y}}^{\mathrm{T}}) - E(\mathbf{y}\hat{\mathbf{y}}^{\mathrm{T}})$ is a positive semidefinite matrix, in which case $\tilde{\mathbf{y}}$ is a shrinkage of $\hat{\mathbf{y}}$. The cross-validatory estimation (3.1) can be directly applied in the above general set-up, ignoring the extensive calculation involved. This idea seems worthy of further exploration in such contexts as *seemingly unrelated linear or non-linear systems*.

My second point relates to criterion (2.3) or its equivalent form (3) above. If $\mathbf{y}$ has a non-zero mean which will enter the right-hand side of equation (3), do we need to modify the criterion? If not, $\tilde{\mathbf{y}}$ will be different from $\hat{\mathbf{y}}$ even though $\mathcal{Y}_1, \ldots, \mathcal{Y}_q$ are independent. A relevant case is linear regression models with non-stochastic design matrices, in which of interest is the accuracy of the prediction at a given $\mathbf{x}$-value rather than at a kind of (probability) average of $\mathbf{x}$.

Rodney Brooks (University College London): Using the method of *joint continuum regression* (JCR) (Brooks and Stone, 1994) on the chemometrics example discussed in Section 7.1, we obtain the following results corresponding to those reported in Table 2 for the cross-validatory assessment criterion, the average of the leave-one-out predictive squared errors:

Response	1	2	3	4	5	6	Average
Predictive squared error	0.089	0.170	0.233	0.102	0.166	0.155	0.153

Leave-one-out cross-validation suggested a large number of components (constructed regressors) for responses 1, 3 and 4 but even if we restrict the number of components to a smaller number, say 6, this only increases the predictive squared errors for responses 1, 3 and 4 to 0.119, 0.253 and 0.119 respectively. The above results for JCR were obtained by only considering the 11 values 0, 0.1, 0.2, . . ., 1 of the control parameter α in JCR. Small improvements in the results may have been obtained by considering further values of α in the interval (0, 1).

Compared with the authors' C&W method, we see that there is a substantial improvement in the predictive squared errors for responses 1 and 2 using JCR. For the remaining four responses, the results are fairly similar, with small decreases for JCR over C&W in three cases.

There seems to be an error in equation (3.4). Starting with the standard downdating formula for the raw one-out predictive value and then centring by subtracting the mean vector $\bar{\mathbf{y}}$ from each side would give equation (3.2), but this would require N^{-1} to be added to the elements of $\mathbf{H}$ in equation (3.4).

Constantinos Goutis (Universidad Carlos III de Madrid), **Tom Fearn and David Brown** (University College London): Our understanding is that traditional methods of combining multiple responses exploit the (conditional on $\mathbf{X}$) correlation between the responses within the same observation. Stein notwithstanding, there is little to gain by combining uncorrelated ys linearly. In the authors' method, this correlation plays little role as their argument works even if $E(\epsilon\epsilon^{\mathrm{T}}) = \mathbf{I}$. It seems, however, that the assumption of random explanatory variables is important. We wonder whether the driving force is the marginal correlation of y_is. If so, we worry about the consequences of the technical fudging in the derivation of the optimal $\mathbf{B}^*$. More importantly, questions about fixed *versus* random design matrix at the data collection stage become crucial and the usual approach of ignoring such questions or bypassing them too quickly by conditioning on $\mathbf{X}$ might not be the best strategy. It would be interesting to see the method's performance for fixed $\mathbf{x}$.

The authors assume that shrinkage is a good thing. This need not be always the case since whether we have an improved expected mean-squared error depends on the parameter values. Excepting the always good James–Stein estimator, other shrinkage methods like ridge or principal components regression must shrink in the right direction. The present method shrinks in a subspace determined by the data in a complicated way. It is difficult to get a feeling for the effect of shrinkage on the parameter values or the predictions, and hence to make a reasonable guess about when the method is likely to work. We have some idea of the authors' ideas about suitable data structures from the design of the simulation study, where the parameters are drawn from short-tailed distributions with mean 0. However, it would be nice to have some explicit comments on this, nicer to have some optimality results indicating for which parameter values inequality (2.2) holds and nicest to see the prior distributions for which this methodology provides sensible posterior estimates.

A final point concerns the examples of Section 7. It would be preferable to compare the authors' joint approach not with ordinary least squares but with some equation-by-equation shrinkage method. The question is not whether it is possible to beat ordinary least squares in situations with many highly correlated variables — we know the answer to that — but whether 'jointness' helps. In the chemometrics example, using either principal components regression or partial least squares with around six factors reduces the predictive squared error for responses 1 and 2 to 0.2.

D. V. Lindley (Minehead): Here is an alternative treatment of the situation discussed in the paper. The part of the model defining the likelihood is $\mathbf{y}_i \sim N(\mathbf{X}\boldsymbol{\theta}_i, \mathbf{I}\sigma^2)$ where $\mathbf{y}_i$ is the vector of observations on the ith response and $\boldsymbol{\theta}_i$ the corresponding vector of regression coefficients θ_{ij} on predictor j. (The convention of Greek letters for parameters and Roman for data seems useful, so the authors' a has been changed to θ.) It remains to specify the distribution of the $\boldsymbol{\theta}$s. Like the choice of the likelihood, this will depend on the practical situation, remembering that the xs and ys are measurements of real quantities and not just symbols. One possibility, that combines the within- and between-regression ideas of Lindley and Smith (1972) and the treatment of the two-way lay-out of Lindley (1974), is to suppose

$$\theta_{ij} = \mu + \alpha_j + \beta_j + \gamma_{ij},$$

reflecting differences of response quantities α_i, prediction quantities β_j and interactions γ_{ij}. Adding exchangeability and normality, these can be independent and identically distributed with zero means and standard deviations σ_{a}, σ_{b} and σ_{c} respectively. It does no harm to suppose μ uniform. For this modelling to be reasonable it will usually be necessary to scale the response quantities appropriately. As in the paper, the data are supposed centred at their means. The above references, improved by the methods of Dawid (1977), show how the distribution of the θs, and hence the prediction of future responses, may be found. For example, the posterior means of $\boldsymbol{\theta}_i$ satisfy the equations, given for simplicity for two predictors and fixed variances,

$$\mathbf{X}^{\mathrm{T}}\mathbf{y}_1 = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{E})\boldsymbol{\theta}_1 + \mathbf{F}\boldsymbol{\theta}_2,$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{y}_2 = \mathbf{F}\boldsymbol{\theta}_1 + (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{E})\boldsymbol{\theta}_2,$$

where $\mathbf{E}$ and $\mathbf{F}$ depend on σ_{a}, σ_{b}, σ_{c} and σ. These standard deviations have a posterior distribution which can be found, at least approximately. I am not an expert in the necessary numerical work, but some Markov chain Monte Carlo procedure looks promising. The effect of $\mathbf{E}$ and $\mathbf{F}$ is to provide shrinkage of a form apparently different from that proposed by the authors. As Lindley and Smith (1972) showed, with a single response variable it produces an improved form of ridge regression. Notice

that this procedure avoids the *ad hoc* nature of the curds and whey procedure. It does involve thinking about the practical reality but then, having done this and formulated a model, leaves the rest to computation by coherent procedures. It also avoids canonical quantities which are often nothing more than mathematical artifices. What does $37 \times \text{height} - 62 \times \text{weight}$ mean?

The following contributions were received in writing after the meeting.

Alison J. Burnham, J. F. MacGregor and Roman Viveros (McMaster University, Hamilton): We congratulate Professor Breiman and Professor Friedman for their proposed curds and whey method. We have concerns, however, regarding the broad conclusions that they draw on the method's performance relative to other methods. Specifically, their simulation does not resemble any real physical situation that we are familiar with, particularly those for which partial least squares (PLS) is recommended.

PLS is designed for data with a strong latent variable structure. The underlying rank usually is much lower than that observed. Measurement and sampling errors are usually responsible for the apparent full rank nature of the observed data. The linear relationship between the predictors and responses is assumed to lie in the space defined by the latent variables. An examination of the authors' simulation reveals that the data simulated have none of these characteristics.

As an illustration, results from the analysis of four real data sets are given in Table 6. These include the two data sets from the paper, a quantitative structure activity relationship (QSAR) data set from chemistry and a data set from an industrial mineral processing plant (mineral process). For comparison only, an indication of the effective rank (ER) of $\mathbf{X}$ is given by the number of principal components required to explain 95% of the variability in $\mathbf{X}$. The ER for Breiman and Friedman's simulation was obtained by averaging over 1000 simulations. PRESS is the prediction residual error sum of squares and is defined as the numerator of the quantity $A(m)$ in the paper.

Clearly PLS gives results similar to or better than ordinary least squares (OLS) and C&W–GCV for the average PRESS on these four examples. Except for one case, PLS shows a smaller 'Robin Hood' effect than C&W–GCV, as judged from the $I(m)$-values. These results bring into question the strong conclusions given in the paper on the comparative merits of the various methods.

In summary, the area requires far more work to determine which methods are appropriate for any given data structure. It is possible that PLS is popular in chemometrics and process engineering because it is well suited to the data structures encountered in those fields. It would be interesting to find out for which *real* applications Breiman and Friedman's methods are well suited.

Trevor Hastie (Stanford University) **and Robert Tibshirani** (University of Toronto): Another common example of multivariate responses occurs in classification and discrimination problems. If G is a J-level categorical response, then $G = j$ can be coded as a J-vector Y of all 0s with a 1 in position j. Note that $E(Y_k|X) = P(G = k|X)$, the critical ingredient for good (Bayes optimal) classification.

Several multiresponse regression approaches to classification are possible.

(a) Regress Y on X directly to estimate $P(X)$, the vector of conditional probabilities. This works

TABLE 6
Comparisons between Breiman and Friedman's simulation and four real data sets

Data set	*Structure details*				*PRESS*			*I(m)*	
	n	*No. of Y*	*No. of X*	*ER of X*	*OLS*	*C&W–GCV*	*PLS*	*C&W–GCV*	*PLS*
QSAR	15	8	8	4	141	127	35	0.91	0.28
Mineral process	230	6	12	5	25	25	24	1.00	0.95
Chemometrics	56	6	22	5	0.42	0.33	0.25	0.83	1.08
Scottish elections	30	4	7	3	1.19	0.93	0.91	0.89	0.83
Breiman and Friedman's simulation	100	5	50	32					

satisfactorily for regression estimates that are averages or local averages (such as in classification trees). Problems occur with highly structured approaches such as linear regression, since estimates can stray from [0, 1], and severe masking can occur (see Hastie *et al.* (1994)).

(b) Model $P(X)$ via multiple logistic regression (to overcome the problems above). Here form instead $J-1$ logits $\eta_j(X) = \log\{P_j(X)/P_J(X)\}$, and proceed by maximum multinomial likelihood.

(c) Use Fisher's linear disriminant analysis via optimal scoring.

These approaches can potentially be enhanced by the 'borrowing strength' techniques described in this paper. In addition to the negative correlation in the responses Y_j due to their construction, correlations in the regression functions can be exploited.

Expanding on (c), Hastie *et al.* (1994) (following Breiman and Ihaka (1984)) exploited the connections between Fisher–Rao discriminant analysis, canonical correlation analysis and optimal scoring. The following connections are worth making here.

(a) Gaussian classification using a common covariance matrix amounts to nearest centroid classification in the canonical co-ordinates derived from the canonical correlation model between Y and X.

(b) These canonical co-ordinates also provide a hierarchy of subspaces for classification and amount to confining the class centroids to optimally spread-out lower dimensional subspaces.

(c) There is a connection here to reduced rank regression — the identical recipe is used: a singular value decomposition of $\hat{Y}$ with *right* metric $Y^{\mathrm{T}}Y/N = \mathrm{diag}(\pi_1, \ldots, \pi_J)$, the sample priors.

These subspaces provide useful informative views of the data and their classes. When classification performance is the goal, this paper suggests that shrinking may be better than truncation as a means of regularization. This provides a kind of continuous rank reduction for linear discriminant analysis.

There are increasingly more interesting problems where X and/or Y represent sampled analogue signals, e.g. classification problems where the objects are digitized images (Y discrete, X image), functional magnetic resonance imaging time series of response images (Y image, X time and treatment stimulus) and the gait curves of Leurgans *et al.* (1993) where both X and Y are sampled analogue functions. In cases like this there is a great potential for increased precision by borrowing strength and adopting some shrinking scheme as described in this paper. Also the domains of X and/or Y can call for additional spatial smoothing via 'penalized' metrics $\Sigma_X + \Omega_X$ and $\Sigma_Y + \Omega_Y$ (Leurgans *et al.*, 1993; Hastie *et al.*, 1995).

Inge S. Helland (University of Oslo): The linear model of Section 2.1 is sufficiently rich and relevant to be developed further. Including centring in the model, we assume $\mathbf{y} = \boldsymbol{\mu}_y + \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}_x) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ has expectation $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, whereas $\mathbf{x}$ has expectation and covariance matrix $\mathbf{V}$. Assume that the N units of the training sample are independent with the same distribution as the units to be predicted, and replace equation (2.9) by a version with mean-centred variables. Taking centring into account in the same way as in Helland and Almøy (1994), theorem 1, we find that equation (2.13) still holds, but now with equation (2.12) replaced by $r = E\,\mathrm{trace}\{(\mathbf{X}^{\mathrm{T}}(\mathbf{I} - N^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}})\mathbf{X})^{-1}\mathbf{V}\} = p/(N-p-2)$. (The last equality is exact under multinormality, an approximation otherwise.)

This gives another demonstration that the assumption $p < N$ (in fact $p < N-2$) is necessary for this theory to make sense. More importantly, it gives a larger shrinkage than is found from equation (2.23) with $r = p/N$, not as large an increase as is implied by equation (3.12), but of the same order of magnitude. This may mean an improvement of the method of Section 2; further improvements may perhaps be found by trying to compensate for the bias in the estimated squared canonical correlation coefficients.

I have some further comments. It does not surprise me that two-block partial least squares (PLS) does so poorly in simulations; I have always had difficulty in understanding the rationale for that method, and I shall be glad if someone can find better alternatives. We should bear in mind, however, that chemometricians like to have the possibility of combining prediction with interpretations of the factor analytic type, preferably by using the same method. This seems to be difficult to achieve by basing the solution on ridge regression. Also, the univariate picture as far as prediction is concerned is not simple either. Even though the simulations in Frank and Friedman (1993) seem to indicate that ridge regression in many cases gives better prediction than univariate PLS, one can easily find specifications of parameters such that PLS clearly dominates ridge regression asymptotically. The idea of combining univariate PLS with canonical rotation of the type proposed by Breiman and Friedman seems to be worth investigating.

M. C. Jones (The Open University, Milton Keynes): There may be better ways of estimating the mean-squared error risk than by cross-validation. Professor Friedman mentioned in his talk that cross-validation was asymptotically unbiased for estimating the risk. This is what worries me. There has remained an obsession with unbiased estimates of risk in related contexts, but this seems inappropriate: if one is happy to use biased estimates of quantities of primary interest, then surely one should equally contemplate biased estimates of their risk (and expect to do better). We certainly can do better in this way for selecting the bandwidth in kernel density estimation (e.g. Jones *et al*. (1996)) but it is not clear whether the major gains to be made in that situation transfer to other contexts. That problem may be special in that density and risk estimation problems require different orders of magnitude of smoothing; other problems may not. Perhaps the asymptotic nature of cross-validation's unbiasedness in the current context disguises a helpful finite sample bias! (I really mean a lower finite sample variance.) But the general point (made also by Johnstone (1988) and Foudrinier and Wells (1995)) of exploring alternative biased risk estimators remains valid; with colleagues at the Open University, I am exploring the possibilities in the (simpler) contexts of shrinkage parameter selection in univariate response ridge regression and for unstructured multinomial data.

Secondly, I have a serious complaint about the paper which is the 'curds and whey' name nonsense. Our discipline has its problems with outreach to workers in other disciplines as well as to the public at large, and giving supposedly practical techniques silly meaningless names sets us back even further. Witty names or acronyms that can at least be explained to users are perhaps not so terrible. But curds and whey seems to be based only on sorting out the signal from the noise and could equally well apply to almost any statistical technique (perhaps I should call my next contribution the 'signal and noise' method, or even the 'statistics' method). Professor Friedman mentioned the more sensible 'multivariate flattening' alternative, and I would not be unhappy with that.

Samuel D. Oman (Hebrew University, Jerusalem): The proposed shrinkage estimator is both interesting and novel. I shall concentrate on the direction, as opposed to the norm, of the shrinkage. In particular, consider Sclove's (1968) modification of the Stein estimator for the univariate linear model $\mathbf{y} = \mathbf{X}\mathbf{a} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, namely

$$\mathbf{a}^* = \mathbf{a}_0 + \left(1 - \frac{c(p-k)\{\mathrm{SSE}/(n-p+2)\}}{(\hat{\mathbf{a}} - \mathbf{a}_0)'(\mathbf{X}'\mathbf{X})^{-1}(\hat{\mathbf{a}} - \mathbf{a}_0)}\right)^+ (\hat{\mathbf{a}} - \mathbf{a}_0) \tag{4}$$

where $0 < c \leqslant 2$, SSE is the usual residual sum of squares, $\hat{\mathbf{a}}$ is the maximum likelihood estimator (MLE) and $\mathbf{a}_0$ is the MLE under the assumption

$$\mathbf{a} \in \mathcal{L} \tag{5}$$

for $\mathcal{L}$ a k-dimensional ($k < p - 2$) subspace of span($\mathbf{X}_{n\times p}$). Shrinking towards $\mathbf{a}_0$ may be motivated by preliminary testing or empirical Bayes considerations (Sclove *et al*., 1972; Efron and Morris, 1973); moreover, $\mathbf{a}^*$ in equation (4) dominates $\hat{\mathbf{a}}$ with respect to prediction mean-squared error, with a substantial improvement if $\mathcal{L}$ is appropriately chosen (Jennrich and Oman, 1986; Oman, 1991; Oman *et al*., 1993).

The trick is in choosing $\mathcal{L}$. I shall illustrate with the chemometrics data analysed by the authors (and kindly sent to me by them), treating each of the six response variables separately. The explanatory variables ($X_1, \ldots, X_{20}$) comprise a temperature profile at points 1, . . ., 20 along the length of the reactor, and all these profiles are increasing in the range 1–4 and decreasing in the range 16–20. This suggests spanning $\mathcal{L}$ by X_{21} and X_{22}, together with principal components extracted separately from the vectors $\mathbf{T}_1 = (X_1, \ldots, X_4)$, $\mathbf{T}_2 = (X_5, \ldots, X_{15})$ and $\mathbf{T}_3 = (X_{16}, \ldots, X_{20})$. From an eigenanalysis it is clear that one component is sufficient for each of $\mathbf{T}_1$ and $\mathbf{T}_3$. For $\mathbf{T}_2$, I simply computed the shrinkage factor in equation (4) for each choice of k_2 ordered components, and then chose k_2 to obtain maximum shrinkage, subject to being small. This gave $k_2 = 3$ for all responses except the third, for which $k_2 = 1$ appeared more reasonable. Since k_2 is based on the observed data $\mathbf{y}$, $\mathbf{a}^*$ is no longer guaranteed to dominate $\hat{\mathbf{a}}$, but this seems no worse in principle than choosing a regression submodel with a small number of variables and low C_p. (Alternatively, we could use the multiple-shrinkage approach of George (George, 1986a, b; George and Oman, 1996).) One at a time cross-validation then gave Table 7, analogous to the authors' Table 2. On the whole, the Stein–Sclove estimator did much better than the

TABLE 7
Predictive squared error for the chemometrics example

Response	*Results from the following methods:*	
	Ordinary least squares	*Stein*
1	0.5662	0.3727
2	1.1709	0.4383
3	0.2593	0.2104
4	0.1228	0.1163
5	0.2771	0.2202
6	0.1897	0.1681
Average	0.4310	0.2543

curds and whey, even though it made no attempt to share information across regressions. This suggests that we should use subspaces in problems where they naturally present themselves. For example, perhaps the curds and whey method could be improved by somehow using Stein predictions instead of ordinary least squares as building blocks in the authors' equation (1.5).

Finally, a technical comment: if $p < q$ then $\mathbf{B}^*$ in equation (2.13) is not invertible, $\mathbf{R}$ in equation (2.14) is not defined and equation (2.21) should read $\mathbf{B}^* = \{r\mathbf{I} + (1-r)\mathbf{Q}\}\mathbf{Q}^{-1}$.

P. D. Sasieni (Imperial Cancer Research Fund, London): In Section 3.2, why not solve the constrained optimization problem directly? That is, find $\mathbf{D}$ to be the diagonal matrix $\boldsymbol{\Delta}$ that minimizes

$$\sum_{i=1}^{q}\sum_{n=1}^{N}\{y_{ni} - (\hat{\mathbf{T}}_{\backslash n}^{-1}\boldsymbol{\Delta}\hat{\mathbf{T}}_{\backslash n}^{-1}\hat{\mathbf{y}}_{\backslash n}^{-1})_i\}^2$$

subject to $\boldsymbol{\Delta} = \text{diag}\{\delta_1, \ldots, \delta_q\}$ and

(a) $\delta_i \geqslant 0$ for $i = 1, \ldots, q$, and
(b) $\{d_i\}_1^q$ are monotone in the sample canonical correlations $\{\hat{c}_i\}_1^q$.

All the constraints implied by (a) and (b) are linear, so the problem is one of quadratic programming.

R. Southworth and C. C. Taylor (University of Leeds): It seems that this problem may be closely related to that of predicting 'compositional' data. In this case we have a p-vector of predictor variables and a q-vector of response variables, but with the added constraints that the response variables must sum to 1, and that each response must lie somewhere in the interval [0, 1]. This can occur when each response is a mixture — for example, a pixel in a remotely sensed image can be classified as 10% water, 40% grass and 50% trees — or more generally viewed as 'fuzzy' classification. We have been exploring the use of neural networks in the context of classification of Meteosat images which have been classified as six-dimensional mixtures. So, the question is: can the proposed methods of this paper be extended (or restricted) to allow this constraint on the response vector?

It also would be of interest to see how nonparametric methods, such as multilayered perceptrons and Friedman's smooth multiple additive regression technique — which are seemingly tailor made for modelling this type of problem — compare with the new method.

Finally, it might be interesting to see some sort of cluster analysis on the results of each method to interpret better their similarities or differences perhaps.

Rolf Sundberg (Stockholm University): Initially my attitude was sceptical, but after reading the paper I must admit that the methods appear to make sense, even though I might prefer to see the main principle in the light of Copas (1983) of literal regression of y on $\hat{y}$, rather than general shrinkage. Many questions pop up, however.

I find it difficult to judge to what extent the simulation results are of predictive value for other situations. If the purpose is response prediction the relative effects shown for estimation exaggerate to a varying but considerable degree. How much does the success depend on the assumption that future x be random, with the same distribution as the sample x-vectors? I am not convinced that the data structures simulated could be expected to bring out the best sides of the reduced rank regression (RRR) and partial least squares PLS2 methods, which select a limited number of components of y. Nevertheless, am I right in believing that the most important feature of the curds and whey (C&W) methods is to shrink near-collinearities in Y to (almost) 0, like the ambitions of RRR and PLS2?

Until Section 6 collinearity is no problem. The first simulation study appears to involve near-collinear data, but this is much an illusion — on average the minimal eigenvalue of V as given by expressions (5.2)–(5.3) is no less than 0.4, with an x-dimension of $p = 50$! In contrast, the 'real' (in fact, simulated) data of example 1 are highly collinear in x, but to my surprise I find only methods based on ordinary least squares compared in this example.

Since near-collinearities are typical for situations with many x-variables, the C&W–RR technique of Section 6.1 seems more interesting for applications. I would try one-factor continuum regression rather than ridge regression, however, since the former shrinks only to compensate for collinearity (Sundberg, 1993). It would have been natural and interesting to see the joint continuum regression of Brooks and Stone (1994) included in the study. Also, the PLS2 method used is not the improved version of de Jong (1993), is it?

Edward V. Thomas (Sandia National Laboratories, Albuquerque): The use of correlated responses to improve the prediction of one or more of the responses is an intriguing prospect. The authors propose an interesting procedure for carrying this out. The variants of this procedure appear to be much more effective than the two-block partial least squares procedure which, although highly promoted, has been used with little success in chemical problems.

The general situation where the procedures proposed (or any other multiple-response procedure) will most probably be effective is where the separate models for each response are weak relative to the correlation between the responses. Conversely, these methods will probably not have much additional benefit in cases where the separate models for each response are strong relative to the correlation between the responses. Figs 9(c) and 9(d) bear this out. For the weakest models ($SS = 100$, $S/N = 1$) there were significant gains relative to ordinary least squares, whereas for the strongest models ($SS = 400$, $S/N = 3$) the gains were less substantial.

The authors find 'little risk of making things worse' by combining multiple responses. When using a predictor based on a single-response empirical model, an important requirement is that the relationship between the response and the regressors is stationary in time. By design, this requirement is met in the simulation studies. In practice, however, the greatest risk of a poor prediction in the future might be that the relationship between the response and the regressors is not stationary in time. In the case of a multiple-response predictor, the assumption of stationarity would be broader as it would also include the correlation structure of the responses. Thus, there might be more risk of making things worse with respect to predicting a particular response by using a multiple-response predictor as opposed to a single-response predictor when

(a) the relationship between the response of interest and the regressors is stationary in time and
(b) the correlation structure among all responses is not stationary.

With this in mind, I am very eager to try out the 'curds and whey' method in several applications where the separate models for each response are weak and where I am confident that stationarity will hold.

Howell Tong (University of Kent, Canterbury): I have one question and three comments on this interesting paper. The methods proposed seem to be mainly driven by linearity and the simulations concentrate on normality. How robust are they against departures from these? It seems to me that the shrinkage technique of Akaike (1978) might provide one possible approach to these problems. He based his methodology on the updating of a prior distribution $\pi(k)$ on the class of models appropriately indexed. He introduced the notion of the *likelihood of a model*, denoted by $P_k(x)$. Here, x denotes the data. In its crudest form, the posterior is given by

$$\pi(k|x) = P_k(x)\,\pi(k) \Big/ \sum_j P_j(x)\,\pi(j),$$

where

$$P_k(x) = C\,\exp\{-\tfrac{1}{2}\mathrm{AIC}(k)\}.$$

(Akaike (1978) has given more refined forms of the posterior.) As an example, the method gives an alternative to the James–Stein estimator for the mean vector, say $(\theta_1, \ldots, \theta_L)$, of an L-dimensional normal:

$$\theta_{\mathrm{VS}i} = \left\{ \sum_{k \in I_i} \pi(k|x) \right\} x_i,$$

where I_i indexes those models whose mean vector has non-zero ith component and VS denotes 'variable selection'. One interesting feature is that it distinguishes zero-mean components from non-zero-mean components. This then raises the question whether we should also distinguish between 'proven' from 'unproven' covariates in the context of Breiman and Friedman's paper. Finally, in the context of non-linear time series prediction it is the mean-squared error *conditional* on the (current) 'state' that is more relevant (e.g. Tong (1995)). Similarly, it seems to me that, in the context of Breiman and Friedman's paper, it is the mean-squared error conditional on the covariates that would be more relevant.

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

We wish to thank all the discussants for stimulating remarks. It is an honour for us that so many important contributors to our field found our work worthy of comment. We would especially like to thank Stone and Yao for pointing out several errors in the original version of our manuscript. Although these errors did not change any results or conclusions correctness is important, and we thank them for their careful reading. Space allotted to us precludes a discussion of all the important points raised, so we shall restrict our comments to the several general themes running through the contributions.

Some of the discussants (Wold, Burnham, MacGregor and Viveros, and Sundberg) were surprised at the relatively poor performance of two-block partial least squares (PLS) compared with the other methods in our simulation studies. However, others (Helland and Thomas) were not surprised. A variety of explanations were offered to mitigate the poor showing of PLS. Wold suggests that the version of two-block PLS used in our comparison is considered obsolete and has not been used in chemometrics since around 1987. This may be so but that is by no means clear. Garthwaite (1994) described two-block PLS as the same method we used, as do Frank and Friedman (1993). Wold was a discussant on the latter paper and no remarks about its obsolescence were made at that time. In any case, by the time that Wold informed us of the alternative version, our simulation study had already been completed. We urge Wold (and others) to subject different versions of PLS, as they come along, to extensive thoroughly designed simulation studies like those presented here. We look forward to the results.

The other major argument put forward is that our simulation studies did not cover the range of situations for which PLS is especially appropriate. Burnham, MacGregor and Viveros point out that PLS had been motivated by situations with strong 'latent variable' structure, and Wold suggests that PLS is especially appropriate when the main canonical axes of the responses are preferentially aligned with those of the predictor variables. These points apply equally well to single-response PLS and were discussed at length in Frank and Friedman (1993). There single-response PLS was seen to be quite competitive in simulation studies like those presented here, so that these issues do not explain the poor showing of two-block PLS observed in this study.

Another set of situations where it is claimed that PLS works well and is not covered in our simulations is where there is a high degree of collinearity among the predictor variables. However, we focused on such situations in Section 6. There we see (Fig. 11(a)) that the competitive advantage of the C&W–RR method over two-block PLS increases with increasing collinearity among the predictor variables.

It was pointed out by Sundberg, and Burnham, MacGregor and Viveros, that the examples we used

TABLE 8
Predictive squared error PSE for the chemometrics example, using C&W–RR

Response	1	2	3	4	5	6	Average
PSE	0.12	0.18	0.22	0.18	0.20	0.20	0.18

TABLE 9
Predictive squared error PSE for the Scottish elections example, using C&W–RR

Response	1	2	3	4	Average
PSE ($\times 1000$)	1.02	0.41	0.28	1.74	0.86

for illustration in Section 7 do exhibit a high degree of predictor collinearity. We agree, and it was an oversight on our part that we did not report the results of applying C&W–RR to these examples. Table 8 shows these results for the chemometrics example (Section 7.1) and Table 9 for the Scottish elections example (Section 7.2).

We see that indeed C&W–RR is generally superior to C&W–GCV for these two examples. This is especially so for the first two responses of the chemometrics example and the second and third responses in the Scottish elections example. We note that the results Wold reports for (new) PLS2 are an average of 0.26 for the chemometrics data and 0.91 for the elections data. This compares with 0.18 and 0.86 for the C&W–RR method. These gains over two-block PLS are consistent with those observed in the simulation study of Section 6.

We also note that our results were obtained by *complete* cross-validation, i.e. each case in turn was held out of the training set, the predictor constructed using only the remaining cases and then the error of this predictor on the left-out case computed and averaged. The paper on joint continuum regression (Brooks and Stone, 1994) indicates that there cross-validation was used to optimize the values of various parameters, and then the value of the resulting (optimized) cross-validated error was reported. If so, this error estimate is biased downwards.

Although C&W–RR performs quite well on these examples, this proves little. As discussed in the first paragraph of Section 7 a ('real') data set represents one random replication of one situation. As such it provides at best anecdotal evidence. As statisticians we advise our clients to avoid anecdotal evidence in favour of carefully designed experiments. It is often said that statisticians are the last to use statistical principles in their own work. In our simulation experiments we found that even the worst performing method (on average) would have the smallest error on some (perhaps many) of the individual replications. We included the examples in Section 7 because it is customary to do so (and the editors insisted on it). Although basing conclusions on a few illustrative examples is common practice, we feel that such evidence at best is inconclusive, and often misleading.

Besides the accusation of being generally unfriendly to PLS there were several other criticisms of our simulation studies. Many centred on their not being sufficiently extensive either in the scope of situations covered or in the list of competitors included. The former criticism is always present in any such study. With the possible exception of the Princeton robustness study (Andrews *et al.*, 1972) the studies presented here are to our knowledge the most comprehensive yet reported, requiring over 1000 h of central processor unit time on SUN workstations to complete. Hand suggests that they were too extensive in that we averaged over a wide variety of regression functions for each set of factor values in our design. He suggests selecting a few regression functions for each and averaging only over the distribution of error (and/or predictor variable) values. This represents an alternative approach, but it would be open to criticism concerning the particular choices for the regression functions. Of course this could represent yet another factor level of the design, but that would involve computation beyond our present capability. As to the list of competitors, we would have included joint continuum regression (Brooks and Stone, 1994) in our study but by the time that it appeared our simulation study had already been completed.

Some discussants (Goutis, Fern and Brown, and Sundberg) are concerned about the validity of our approach to controlled experiments where the predictor design is fixed rather than being random. The

theoretical development of Section 2 seems applicable to this case but problems can arise in the use of cross-validation (or generalized cross-validation) to estimate the diagonal matrix **D** (3.12) (3.15). In this case the estimate of **D** could be accomplished by using methods for fixed designs such as the little bootstrap (Breiman, 1992).

The situation where $q \simeq N$ is raised by ter Braak. Although this case seems somewhat unusual, some comments can be made. Assume that $q = N$ and that $p < N$. ter Braak states that the canonical system will be very unstable. However, the canonical system is a convenient and interesting device for solving equation (3.9) and this equation is stable with a unique solution. ter Braak rightly notes that C&W gives only p shrinkage factors whereas doing individual shrinking gives $q > p$ factors. He points out, for example, that in the case $p = 1$ there is only a single shrinkage factor with value close to 1, whereas doing individual regressions allows q shrinkage factors to be computed. This is certainly true, although each of the q factors will also have value close to 1. Therefore, no matter which approach is adopted, there will be very little shrinkage. If p is small compared with N then all shrinkage factors, both univariate and C&W, will have values close to 1. Furthermore, it is not at all clear that estimating q individual shrinkage factors will do as well in reducing prediction error as simply using the p shrinkage factors produced by C&W.

Several discussions (Zidek, Cox, Tong and Goutis, Fern and Brown) question the robustness of our procedures to violations in the assumptions concerning the structural model (linearity) or to long-tailed distributions (outliers). Our procedures combine either ordinary least squares (C&W–GCV or C&W–CV) or ridge regression (C&W–RR) so that they necessarily inherit the robustness properties of those procedures. However, as noted by Garthwaite, Yao and Hastie and Tibshirani our results provide a general framework for combining multiple-response estimates of any type. The basic paradigm that leads to C&W–RR (Section 6) is employed replacing $\hat{\mathbf{Y}}_\lambda$ with the corresponding $\hat{\mathbf{Y}}$ derived from the selected (single-response) method. One then estimates (through cross-validation) either the shrinkage parameter r to be used in equation (3.12) or the diagonal matrix **D** in equation (3.15).

If robustness to outliers is desired then one can combine correspondingly robust regression procedures. For non-linear structural models one of the popular flexible modelling procedures can be used to obtain the response estimates $\hat{\mathbf{Y}}$. We have used this approach (Breiman and Friedman, 1994) to develop a multiple-response version of multivariate adaptive regression spline (MARS) modelling (Friedman, 1991). Here the MARS method is applied to each response separately and then these response estimates are combined using the C&W procedure. Substantial improvements in prediction accuracy, over using the single-response estimates separately, were observed.

REFERENCES IN THE DISCUSSION

Akaike, H. (1978) Likelihood of a model. *Research Memorandum 127*. Institute of Statistical Mathematics, Tokyo.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. and Tukey, J. W. (1972) *Robust Estimates of Location: Survey and Advances*. Princeton: Princeton University Press.

ter Braak, C. J. F. and Looman, C. W. N. (1994) Biplots in reduced-rank regression. *Biometr. J.*, **36**, 983–1003.

Breiman, L. (1992) Submodel selection and evaluation in regression — the *X*-fixed case and Little Bootstrap. *J. Am. Statist. Ass.*, **87**, 734–751.

Breiman, L. and Friedman, J. H. (1994) A new approach to multiple outputs. *Wrkshp: Machines that Learn — Neural Networks for Computing, Snowbird.*

Breiman, L. and Ihaka, R. (1984) Nonlinear discriminant analysis via scaling and ACE. *Technical Report*. University of California, Berkeley.

Brooks, R. and Stone, M. (1994) Joint continuum regression for multiple predictands. *J. Am. Statist. Ass.*, **89**, 1374–1377.

Brown, L. D. (1966) On the admissibility of invariant estimators of one or more location parameters. *Ann. Math. Statist.*, **37**, 1087–1136.

Brown, P. and Payne, C. (1975) Election night forecasting (with discussion). *J. R. Statist. Soc.* A, **138**, 463–498.

——— (1984) Forecasting the 1983 British General Election. *Statistician*, **33**, 217–228.

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

——— (1982) Multivariate regression shrinkage estimators with unknown covariance matrix. *Scand. J. Statist.*, **9**, 209–215.

Copas, J. B. (1983) Regression, prediction and shrinkage (with discussion). *J. R. Statist. Soc.* B, **45**, 311–354.

Dawid, A. P. (1973) Posterior expectations for large observations. *Biometrika*, **60**, 664–667.

——— (1977) Invariant distributions and analysis of variance models. *Biometrika*, **64**, 291–297.

Dillon, W. R. and Goldstein, M. (1984) *Multivariate Analysis: Methods and Applications*. New York: Wiley.

Efron, B. and Morris, C. (1973) Stein's estimation rule and its competitors — an empirical Bayes approach. *J. Am. Statist. Ass.*, **68**, 117–130.

Foudrinier, D. and Wells, M. T. (1995) Estimation of a loss function for spherically symmetric distributions in the general linear model. *Ann. Statist.*, **23**, 571–592.
Frank, I. E. and Friedman, J. H. (1993) A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, **35**, 109–148.
Friedman, J. H. (1991) Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, **19**, 1–141.
Garthwaite, P. H. (1994) An interpretation of partial least squares. *J. Am. Statist. Ass.*, **89**, 122–127.
George, E. I. (1986a) Minimax multiple shrinkage estimation. *Ann. Statist.*, **14**, 188–205.
———(1986b) Combining minimax shrinkage estimators. *J. Am. Statist. Ass.*, **81**, 437–445.
George, E. I. and Oman, S. D. (1996) Multiple-shrinkage principal component regression. *Statistician*, **45**, 111–124.
Hastie, T., Buja, A. and Tibshirani, R. (1995) Penalized discriminant analysis. *Ann. Statist.*, **23**, 73–102.
Hastie, T., Tibshirani, R. and Buja, A. (1994) Flexible discriminant analysis by optimal scoring. *J. Am. Statist. Ass.*, **89**, 1255–1270.
Helland, I. S. and Almøy, T. (1994) Comparison of prediction methods when only a few components are relevant. *J. Am. Statist. Ass.*, **89**, 583–591.
Hu, F. and Zidek, J. V. (1995) The relevance weighted likelihood. To be published.
Jennrich, R. and Oman, S. D. (1986) How much does Stein estimation help in multiple linear regression? *Technometrics*, **28**, 113–121.
Johnstone, I. M. (1988) On admissibility of some unbiased estimators of loss. In *Statistical Decision Theory and Related Topics 4* (eds S. S. Gupta and J. O. Berger), vol. 1, pp. 361–379. New York: Springer.
Jones, M. C., Marron, J. S. and Sheather, S. J. (1996) A brief survey of bandwidth selection for density estimation. *J. Am. Statist. Ass.*, **91**, 401–407.
de Jong, S. (1993) SIMPLS: an alternative approach to partial least squares regression. *Chemometr. Intell. Lab. Syst.*, **18**, 251–263.
Leurgans, S. E., Moyeed, R. A. and Silverman, B. W. (1993) Canonical correlation analysis when the data are curves. *J. R. Statist. Soc.* B, **55**, 725–740.
Lindley, D. V. (1974) A Bayesian solution for the two-way analysis of variance. *Colloq. Math. Soc. János Bolyai*, **9**, 475–496.
Lindley, D. V. and Smith, A. F. M. (1972) Bayes estimates for the linear model (with discussion). *J. R. Statist. Soc.* B, **34**, 1–41.
Nomikos, P. and MacGregor, J. F. (1995) Multivariate SPC charts for monitoring batch processes. *Technometrics*, **37**, 41–59.
Oman, S. D. (1991) Random calibration with many measurements: an application of Stein estimation. *Technometrics*, **33**, 187–195.
Oman, S. D., Naes, T. and Zube, A. (1993) Detecting and adjusting for non-linearities in calibration of near-infrared data using principal components. *J. Chemometr.*, **7**, 195–212.
Payne, C. D. and Brown, P. J. (1981) Forecasting the British election to the European Parliament. *Br. J. Polit. Sci.*, **11**, 235–245.
Sclove, S. L. (1968) Improved estimators for coefficients in linear regression. *J. Am. Statist. Ass.*, **63**, 596–606.
Sclove, S. L., Morris, C. and Radhakrishnan, R. (1972) Non-optimality of preliminary test estimators for the multivariate normal mean. *Ann. Math. Statist.*, **43**, 1481–1490.
Sundberg, R. (1993) Continuum regression and ridge regression. *J. R. Statist. Soc.* B, **55**, 653–659.
Tong, H. (1995) A personal overview of non-linear time series analysis from a chaos perspective (with comments and discussion). *Scand. J. Statist.*, **22**, 399–445.
Wold, S. (1995) PLS for multivariate linear modelling. In *Methods and Principles in Medical Chemistry* (ed. H. van de Waterbeemd), vol. 2. Weinheim: Verlag Chemie.
Wold, S., Ruhe, A., Wold, H. and Dunn III, W. J. (1984) The collinearity problem in linear regression: the partial least squares approach to generalized inverses. *SIAM J. Sci. Statist. Comput.*, **5**, 735–743.
Wold, S., Sjöström, M. and Hellberg, S. (1987) Chemometrics: multivariate analysis and design. *Bull. Int. Statist. Inst.*, **4**, 477–495.
Zellner, A. (1962) An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Am. Statist. Ass.*, **57**, 348–368.
Zellner, A. and Huang, D. S. (1962) Further properties of efficient estimation for seemingly unrelated regression equations. *Int. Econ. Rev.*, **3**, 300–313.