Predicting Multivariate Responses in Multiple Linear Regression

Leo Breiman* Jerome H. Friedman†

November 24, 1995

Abstract

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 as compared to 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 & 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 to several previously proposed methods for predicting multiple responses (including PLS) and exhibits superior accuracy. One version can be easily implemented in the context of standard statistical packages.

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

*Department of Statistics, University of California, Berkeley, CA 94720. Work partially supported by the National Science Foundation, Grant No. DMS-9212419.

†Department of Statistics and Stanford Linear Accelerator Center, Stanford University, Stanford, CA 94305. Work partially supported by the Department of Energy, Contract No. DE-AC03-76SF00515.

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 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 is of the form $\{y_{n1}, y_{n2}, \mathbf{x}_n\}_1^N$ where each $\mathbf{x}_n = (x_{n1}, \cdots, x_{np})$ is a p - vector of predictor variables and there are two responses y_1 and y_2. Taking the usual path, we get predictors for y_1, y_2 by doing separate regressions on $(x_1, \cdots, x_p)$. That is, the estimated regression coefficients $\hat{\mathbf{a}}_1 = (\hat{a}_{11}, \cdots, \hat{a}_{1p})$ and $\hat{\mathbf{a}}_2 = (\hat{a}_{21}, \cdots, \hat{a}_{2p})$ are solutions to

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

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

where all variables have been centered. The prediction equations for y_1 and y_2 are $\hat{y}_1(\mathbf{x}) = \bar{y}_1 + \hat{\mathbf{a}}_1^t(\mathbf{x} - \bar{\mathbf{x}})$ and $\hat{y}_2(\mathbf{x}) = \bar{y}_2 + \hat{\mathbf{a}}_2^t(\mathbf{x} - \bar{\mathbf{x}})$ where $(\bar{y}_i, \bar{\mathbf{x}})$ are the corresponding sample means (before centering). Now suppose further that the (unknown) truth happens to be $y_{n1} = b_{10} + \mathbf{b}^t \mathbf{x}_n + \varepsilon_{n1}$ and $y_{n2} = b_{20} + \mathbf{b}^t \mathbf{x}_n + \varepsilon_{n2}$ where $\{\varepsilon_{n1}\}_1^N$ and $\{\varepsilon_{n2}\}_1^N$ are independent i.i.d. $N(0, \sigma^2)$. Here y_1 and y_2 are correlated because they have the same dependence on the predictor variables, $\mathbf{b}^t \mathbf{x}$. It is also clear that accuracy is improved for *each* of the two responses by using the predictors $\tilde{y}_i = \bar{y}_i + \frac{1}{2}(\hat{y}_1 - \bar{y}_1) + \frac{1}{2}(\hat{y}_2 - \bar{y}_2)$ $(i = 1, 2)$, instead of $\hat{y}_1$ and $\hat{y}_2$ respectively.

1.1. The curds and whey (C&W) procedure.

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

of each y_i by using a linear combination

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

of the ordinary least squares (OLS) predictors

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

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

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

To simplify notation in all derivations that follow, we assume that the response and predictor variables are all centered 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 centered 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 centered predictor variables.

Assuming that (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 (centered) quantities

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

(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 coordinates (see Section 2.2) and $\mathbf{D} = diag(d_1, \cdots, d_q)$ is a diagonal matrix. Two prescriptions are derived for calculating $\{d_k\}_1^q$. A generalized cross-validation approach (Section 3.1) yields a simple formula (3.12) (3.13). This works surprisingly well. Using regular (5 or 10 - fold) cross-validation (Section 3.2) to obtain the $\{d_k\}_1^q$ gives slightly better prediction.

1.2. Statistical background.

The curds and whey (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 coordinate system (canonical coordinates), and can be viewed as a multivariate generalization of proportional shrinkage based on cross-validation [Stone (1974)].

In the case of 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)], 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 the field of chemometrics. C&W 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 to which we have compared it.

1.3. Outline of paper.

In Section 2 we assume (centered) 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, due to the idealized setting, the matrix $\mathbf{D}$ derived there underestimates the amount of shrinkage 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 in the literature for estimating multiple responses, and these are compared to C&W in the simulation study covered in Section 5. In some fields, chemometrics for example, it is not unusual to have fewer observations than predictor variables ($N \prec p$).

The C&W method can be extended to these under-determined 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 (centered) proportional shrinkage estimate $\widetilde{y}$ can be expressed as

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

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

$$E[y - \widetilde{y}]^2 < E[y - \widehat{y}]^2, \tag{2.2}$$

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

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

Here (2.3) 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 (1.5) will result in reduced mean-squared prediction error for *each* response

$$E[y_i - (\mathbf{B}^*\widehat{\mathbf{y}})_i]^2 \leq E[y_i - \widehat{y}_i]^2, \quad i = 1, \cdots, q, \tag{2.4}$$

with equality in (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 of the responses (such as weighted average quadratic loss).

2.1. Optimal proportional shrinkage.

In order 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 (i.i.d.) error

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

with

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

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

$$F(\mathbf{x}, \boldsymbol{\varepsilon}) = F_x(\mathbf{x}) F_\varepsilon(\boldsymbol{\varepsilon}); \tag{2.7}$$

that is, the errors are independent of the predictor variables. Let

$$E(\mathbf{x}) = E(\boldsymbol{\varepsilon}) = 0, \;\; E(\mathbf{x}\mathbf{x}^t) = \mathbf{V} \in R^{p \times p}, \;\; E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t) = \boldsymbol{\Sigma} \in R^{q \times q} \tag{2.8}$$

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

$$\mathbf{B}^* = \left[E(\widehat{\mathbf{y}}\widehat{\mathbf{y}}^t)\right]^{-1} E(\mathbf{y}\widehat{\mathbf{y}}^t). \tag{2.9}$$

In order 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

$$\mathbf{F} = E[\mathbf{f}(\mathbf{x})\mathbf{f}^t(\mathbf{x})] = \mathbf{A}\mathbf{V}\mathbf{A}^t \tag{2.10}$$

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), one has

$$E[\hat{\mathbf{y}}\hat{\mathbf{y}}^t] = \mathbf{F} + r\mathbf{\Sigma},\ E[\mathbf{y}\hat{\mathbf{y}}^t] = \mathbf{F}. \tag{2.11}$$

where

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

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

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

where

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

is the "noise / 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 coordinates 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 centered) 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 $\mathbf{t}^t\mathbf{y}$ and $\mathbf{v}^t\mathbf{x}$ 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

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

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

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

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 to all have unit variances

$$E(\mathbf{t}_k^t \mathbf{y})^2 = E(\mathbf{v}_k^t \mathbf{x})^2 = 1, \quad k = 1, \cdots, K. \tag{2.17}$$

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

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

(Although $\mathbf{Q}$ is not symmetric, it is the product of two symmetric matrices, so that the eigen-decomposition (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 coordinates $\{\mathbf{t}_k\}_1^q$ and the diagonal matrix

$$\mathbf{C}^2 = diag\left\{c_1^2, ..., c_K^2\right\} \tag{2.19}$$

contains the respective squared canonical correlations (2.16). The $\mathbf{x}$ - canonical coordinates 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 the case of our regression model (2.5) (2.6) (2.7) it provides a means for obtaining the optimal shrinking matrix $\mathbf{B}^*$ (2.13). Under that model $\mathbf{Q}$ (2.18) becomes

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

so that

$$\mathbf{B}^* = \left[(1 - r)\mathbf{I}_q + r\mathbf{Q}^{-1}\right]^{-1} \tag{2.21}$$

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

$$\mathbf{B}^* = \mathbf{T}^{-1}\mathbf{D}\mathbf{T}, \ \mathbf{D} = diag(d_1, ..., d_q) \tag{2.22}$$

with

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

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

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

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

1. Transform $\mathbf{y}$ to the canonical coordinate system, $\mathbf{y}' = \mathbf{T}\mathbf{y}$.
2. Perform a separate OLS regression of each y_i' on $\mathbf{x}$, $(i = 1, \cdots, q)$, obtaining $\{\hat{y}_i'\}_1^q$.
3. Separately scale (shrink) each $\hat{y}_i'$ by the factor d_i (2.23), obtaining $\tilde{\mathbf{y}}' = \{d_i \hat{y}_i'\}_1^q$.
4. Transform back to the original $\mathbf{y}$ - coordinate system, $\tilde{\mathbf{y}} = \mathbf{T}^{-1}\tilde{\mathbf{y}}'$.

Figure 1 shows graphs of the canonical coordinate 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 coordinate 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, $0 \leq d_i \leq 1$.

In order to estimate $\mathbf{B}^*$ (2.21) one needs a sample based estimate of $\mathbf{Q}$ (2.18). A natural choice would be the "plug-in" estimate

$$\widehat{\mathbf{Q}} = (\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} \tag{2.25}$$

where

$$\mathbf{Y} = [y_{ni}] \in R^{N\times q} \text{ and } \mathbf{X} = [x_{nj}] \in R^{N\times p}, \tag{2.26}$$

are the respective (centered) data matrices. Although this choice does improve the OLS estimates, it does not provide enough shrinkage, and more improvement is possible. The reason is that the sample canonical correlations $\{\widehat{c}_i\}_1^q$ overestimate their corresponding population values $\{c_i\}_1^q$ (2.19) so that using these sample based estimates in (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) $\{\widehat{c}_i\}_1^q$ values. This is a common problem in model selection. In order 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 (C&W).

The optimal shrinking matrix $\mathbf{B}^*$ (2.3) is obtained by a regression of the responses $\{y_i\}_1^q$ on the (sample based) OLS estimates $\{\widehat{y}_i\}_1^q$ over (all future) data 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) analog to (2.3) then becomes

$$\{b_{ik}\}_{k=1}^q = \arg\min_{\{\beta_k\}_1^q} \sum_{n=1}^N \left[y_{ni} - \sum_{k=1}^q \beta_k \widehat{y}_{\setminus nk} \right]^2, \quad i = 1, \cdots, q, \tag{3.1}$$

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

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

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

$$g_n = \frac{1}{1 - h_{nn}} \tag{3.3}$$

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

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

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

3.1. GCV based multivariate shrinking (C&W-GCV).

To simplify this 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

$$h_{nn} \approx \overline{h} = \frac{1}{N} \sum_{m=1}^N h_{mm} = \frac{1}{N} trace\mathbf{H} = r \tag{3.5}$$

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

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

where

$$g = \frac{1}{1-r}. \quad (3.7)$$

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

$$[(1-g)\mathbf{Y}^t + g\widehat{\mathbf{Y}}^t][(1-g)\mathbf{Y} + g\widehat{\mathbf{Y}}]\mathbf{B} = (1-g)\mathbf{Y}^t\mathbf{Y} + g\mathbf{Y}^t\widehat{\mathbf{Y}} \quad (3.8)$$

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

$$[(1-g)^2\mathbf{I}_q + (2g-g^2)\widehat{\mathbf{Q}}]\mathbf{B} = (1-g)\mathbf{I}_q + g\widehat{\mathbf{Q}} \quad (3.9)$$

where $\widehat{\mathbf{Q}}$ is the sample canonical correlation matrix (2.25). This (3.9) shows the solution $\mathbf{B}$ is a diagonal matrix in the same coordinate system that diagonalizes $\widehat{\mathbf{Q}}$,

$$\widehat{\mathbf{Q}} = \hat{\mathbf{T}}^{-1}\hat{\mathbf{C}}^2\hat{\mathbf{T}}, \;\; \hat{\mathbf{C}}^2 = diag\{\hat{c}_1^2, \cdots, \hat{c}_q^2\}. \quad (3.10)$$

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

$$\mathbf{B} = \hat{\mathbf{T}}^{-1}\widehat{\mathbf{D}}\hat{\mathbf{T}}, \; \widehat{\mathbf{D}} = diag\{\hat{d}_1, \cdots, \hat{d}_q\} \quad (3.11)$$

with

$$\hat{d}_i = \frac{(1-r)(\hat{c}_i^2 - r)}{(1-r)^2\hat{c}_i^2 + r^2(1-\hat{c}_i^2)}, \;\; i = 1, \cdots, q. \quad (3.12)$$

Examination of (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

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

in (3.12).

Comparing these results (3.11 - 3.13) with those of (2.22 - 2.24), one sees 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 (3.12) (3.13) in place of (2.23) for the shrinking factors in the (sample) canonical coordinate system. Figure 2 shows graphs of $\widehat{d}_i$ (3.12 - 3.13) as a function of the corresponding (sample) squared-canonical correlations $\widehat{c}_i^2$, for the same values of r as in Fig. 1. The GCV canonical shrinkage factors $\widehat{d}_i$ 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 $\widehat{c}_i^2$ and r. This compensates for the upward bias in the estimates $\{\widehat{c}_i\}_1^q$ 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 i.i.d. assumption required for cross-validation. The validity of the GCV result rests on suitability of (3.1) as an estimate of (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 $\mathbf{B}$.

3.2. Fully cross-validated multivariate shrinking (C&W-CV).

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

$$\mathbf{B} = \widehat{\mathbf{T}}^{-1}\mathbf{D}\widehat{\mathbf{T}}, \;\; \mathbf{D} = diag\{d_1, \cdots, d_q\}, \tag{3.14}$$

with $\widehat{\mathbf{T}}$ being the sample $\mathbf{y}-$ canonical coordinate transformation matrix (3.10), and $\mathbf{D}$ the solution to

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

Here (3.15) the subscript $\backslash n$ on a quantity refers to that quantity calculated with the nth observation removed. Note that (3.15) is a purely quadratic criterion in

$\mathbf{\Delta} = diag\{\delta_1, \cdots, \delta_q\}$ so that the solution for $\mathbf{D}$ is unique and can be obtained by straightforward linear algebra given the other quantities appearing in (3.15).

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

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

in (3.14).

This (3.14 - 3.16) generalizes the GCV approach by removing the approximation (3.5), and accounting for the variability in the estimate of the sample canonical coordinate transformation $\widehat{\mathbf{T}}$ (3.10), in the estimation of the canonical coordinate 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 $\widehat{\mathbf{T}}_{/n}$ and $\widehat{\mathbf{y}}_{/n}$ must be calculated for each observation ($n = 1, \cdots, N$) 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 $\widehat{\mathbf{T}}$ and $\widehat{\mathbf{y}}$, 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 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. That is, a separate least-squares regression (1.2) (1.3) of each response y_i on the predictor variables $\mathbf{x}$. 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 to 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 in the past 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 (below) we include a brief description of some of these and examine their relationship to our procedures. In Section 5 we compare performance 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

$$\{\hat{a}_{ij}\}_{j=1}^{p} = \arg\min_{\{a_j\}_1^p} \sum_{n=1}^{N} [y_{ni} - \sum_{j=1}^{p} a_j x_{nj}]^2 + \lambda_i \sum_{j=1}^{p} a_j^2, \quad i = 1, \cdots, q. \tag{4.1}$$

This (4.1) biases the coefficient estimates toward 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

$$\hat{\lambda}_i = \arg\min_{\lambda} \sum_{n=1}^{N} [y_{ni} - \hat{y}_{\setminus ni}]^2, \;\; i = 1, \cdots, q, \tag{4.2}$$

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 (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 the regression model (2.5) (2.6) and suppose one wishes to find the coefficient matrix $\tilde{\mathbf{A}}_r \in R^{q \times p}$ of rank $r \leq \min(p, q)$ that minimizes

$$\tilde{\mathbf{A}}_r = \arg\min_{rank(A)=r} E(\mathbf{y} - \mathbf{A}\mathbf{x})^t \mathbf{\Sigma}^{-1} (\mathbf{y} - \mathbf{A}\mathbf{x}) \tag{4.3}$$

with Σ given by (2.8). The solution to (4.3) is

$$\tilde{\mathbf{A}}_r = \mathbf{B}_r \hat{\mathbf{A}} \tag{4.4}$$

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 coordinate matrix (2.18) and

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

In applications of reduced rank regression the sample canonical coordinates $\hat{\mathbf{T}}$ (2.25) (3.10) are taken as estimates of the corresponding population quantities in (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 (analog of (4.2)). This estimate (2.25) (3.10) (4.5) (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. FICYREG.

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 (4.8) $\hat{\mathbf{T}}$ is the sample canonical coordinate matrix (2.25) (3.10) and

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

with

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

and

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

The $\{\hat{c}_i^2\}_1^q$ in (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 (4.10) and (3.12) is that the canonical coordinate shrinkage factors $\{f_i\}_1^q$ (4.10) depend on the number of responses q as well as 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, (4.10) (4.11) shrink less for a larger number of responses. The corresponding C&W-GCV factors $\{d_i\}_1^q$ (3.12) (3.13) depend only on $\{\hat{c}_i^2\}_1^q$ and p irrespective of the number of responses. For all values $q \geq 1$ one has

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

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

4.4. Two-block partial least squares.

Partial least squares ("PLS") regression [Wold (1975)] is very popular in the field of 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 $cov(\mathbf{t}^t\mathbf{y}, \mathbf{v}^t\mathbf{x})$ replacing $corr(\mathbf{t}^t\mathbf{y}, \mathbf{v}^t\mathbf{x})$ in (2.15), and the constraints in (2.17) replaced by $\{\mathbf{t}_k^t\mathbf{t}_k = \mathbf{v}_k^t\mathbf{v}_k = 1\}_1^q$. The (ordered) set of canonical covariance $\mathbf{x}$ - linear combinations

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

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

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

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

This (4.14) (4.15) is a separate OLS regression of each response y_i on the first K $\mathbf{x}$ - canonical covariance linear combinations (4.13). The 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) (4.15) is a regularization parameter of the procedure; its value is determined through cross-validation (analog of (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 trying to understand it statistically. Frank and Friedman (1993) provide some insight by comparing its results to that of a particular formulation of multivariate ridge regression 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 ridge regressions, 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) suggest a variety of multivariate generalizations of ridge regression along the lines of FICYREG. The four competitors described above have seen use on data and two (separate ridge regressions and two-block PLS) are very popular.

Of the six procedures described above, four (C&W-CV, reduced rank regression, separate ridge regressions, 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 of the procedures but two (separate ridge regressions 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 ridge regressions are clearly equivariant under response scale changes but not under rotations in the response space. They are equivariant under (rigid) rotations of the $\mathbf{x}$ - coordinates, 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 $\mathbf{x}$ spaces, but not equivariant under scale changes in either space. Both ridge 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 coordinates (2.25) (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 coordinate 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) (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 assumes that the data $\{\mathbf{y}_n, \mathbf{x}_n\}_1^N$ are an i.i.d. 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 (4.16) where the elements of $\mathbf{G}$ depend only on the sample canonical correlations. Zidek (1978) then shows that for the particular loss function defined, the form of $\mathbf{G}$ used in FICYREG (4.10) (4.11) gives estimates dominating OLS. It is perhaps no surprise that many multivariate multiple regression procedures involve canonical coordinates at a basic level, since, as shown in Sections 2.1 and 2.2, the canonical coordinate system emerges as the natural one for optimal proportional shrinkage (2.22) (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 ridge) to justify their consideration as viable alternatives. And, among those that do, which ones provide the best trade-off between accuracy improvement and increased complexity, both in terms of implementation and computation. The answers to these questions may well 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 (below) 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 (2.5) (2.6). The errors $\{\varepsilon_i\}_1^q$ were generated from a normal distribution with zero mean and covariance matrix $\mathbf{\Sigma}$ (2.8)

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

Two covariance structures among the errors were studied:

$$\mathbf{\Sigma} = \sigma^2 \cdot \mathbf{I}_q \text{ and } \mathbf{\Sigma} = \sigma^2 \cdot diag\{i^2\}_1^q. \tag{5.5}$$

In the first, the error variance associated with each response is the same, whereas in the second they are very different. More complicated (nondiagonal) 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/signal matrix $\mathbf{R}$ (2.14). Two values of σ^2 (5.5) were studied. They were chosen so that (on average) signal/noise ratios of 1.0 and 3.0 respectively 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

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

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 (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 $g(j,k)$ (5.7), when viewed as a function of the predictor variable index j, represents a (normalized) "bump" centered at j_k with support (nonzero 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 centered 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 is comprised of two sample sizes ($N = 100, 400$), three values for the number of responses ($q = 5, 10, 20$), five values for the average correlation among the response functions (2.6)

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

(controlled by ρ (5.10)), two error covariance structures (5.5), and two signal to noise ratios $(1.0, 3.0)$. A complete factorial design over all of 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 ridge, 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

$$e_i^2(m) = \int [(\mathbf{a}_i - \tilde{\mathbf{a}}_i(m))^t \mathbf{x}]^2 p(\mathbf{x}) d\mathbf{x}$$

$$= (\mathbf{a}_i - \tilde{\mathbf{a}}_i(m))^t \mathbf{V} (\mathbf{a}_i - \tilde{\mathbf{a}}_i(m)) \tag{5.12}$$

where $\mathbf{a}_i = \{a_{i1}, \cdots, a_{ip}\}$ is the "true" coefficient vector (2.6) (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) (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

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

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

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

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

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

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

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

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

The first two criteria (5.13) (5.14) provide a means of comparing each of the six methods to OLS in terms of how much average (squared) error reduction each gives relative to OLS. The third 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 less accurate than its corresponding OLS estimate. The last two measures (5.16) (5.17) allow comparisons among the six biased methods themselves. For each individual replication, the value of (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) (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. Figures 3 - 6 display box plots of the mean values of (5.13), (5.14), (5.16), and (5.17) respectively over all of the 120 situations covered by the simulation study. That is, each box plot summarizes the distribution of 120 (mean) values. Figure 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 ($A(OLS) = 1$). All of the multivariate methods, except two-block PLS, also show substantial improvement over separate (uniresponse) ridge regressions. 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 two improvement over OLS, as averaged over all 120 situations, and about a 61% improvement over separate ridge regressions.

Figure 4 shows the distribution of average individual mean-squared error ratio $I(m)$ (5.14). These distributions are fairly similar to the corresponding ones 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 in order to achieve modest (relative) improvement in those that are poorly estimated by OLS (and PLS). Comparing Figs. 3 and 4 one sees that the other methods do not exhibit the Robin Hood effect; they produce roughly equal relative improvement across all responses.

Figures 5 shows the distributions of $RA(m)$ (5.16), and Fig. 6 the logarithm of $RI(m)$ (5.17). C&W-CV is seen to have the best average performance, or within a few percent of the best, in every one of the 120 situations. C&W-GCV 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.

Figures 3 and 4 show that, averaged over all responses, all of the six biased methods considered here provide improved performance over OLS. That improvement was fairly dramatic for some of the methods. From a perspective of caution one might ask how probable it is that an individual response estimate by one of these methods will be less accurate that its OLS estimate. That is, 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. Figure 7 addresses this issue for all of the methods by showing the distribution (over all 30,000 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. One sees that the most cautious method by this measure is FICYREG. On average less than 3% of its response estimates are worse than OLS. C&W-GCV is seen to be only slightly less cautious, its estimates being worse than the corresponding OLS estimates an average of 5% of the time. C&W-CV also exhibits fairly cautious behavior 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). Figure 8 shows the distribution of the logarithm of this quantity for each method, separately for each of the two error variance structures (5.5).

The left 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, $\mathbf{\Sigma} = \sigma^2 \cdot \mathbf{I}_q$, and the right box plot is the corresponding distribution over the other 60 for which they are very unequal, $\mathbf{\Sigma} = \sigma^2 \cdot diag\{i^2\}_1^q$. One sees 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. On the other hand, the caution associated with two-block PLS is seen to dramatically depend 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 not uncommon. Frank and Friedman (1993) argued that an intrinsic (implicit) assumption associated with two-block PLS is the simple error covariance structure $\mathbf{\Sigma} = \sigma^2 \cdot \mathbf{I}_q$. 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. On the other hand, 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 at the same time providing nearly as much accuracy as the most accurate one C&W-CV (Figs. 5 and 6).

Figure 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. One sees from the upper left frame that separate ridge regressions are unaffected by the degree of correlation among the responses (5.11) whereas the multivariate methods all perform better with higher (positive or negative) correlation, as would be expected. The middle left frame shows that the performance (relative to OLS) of all methods, except two-block PLS, is better with highly unequal error variances (5.5). As one would

expect all methods improve (relative to OLS) with decreasing sample size (middle right frame) and decreasing signal to noise ratio (lower left frame), but FICYREG seems to enjoy less improvement than the others. The lower right frame shows the dependence of $A(m)$ on the number of responses q. The performance of separate ridge regressions 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 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 while 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 ridge regression). Of the methods compared here, only two-block PLS provided inferior results to separate ridge regressions. 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 would appear to be the logical choice among 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 ridge regressions. 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 the field of chemometrics. By contrast, single response PLS is competitive with other (single response) biased regression methods, performing almost as well as ridge regression [Frank and Friedman (1993)]. This together with the fact that separate ridge regressions substantially outperform two-block PLS, suggest that in environments where PLS for some reason must be used, performing separate (uniresponse) 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 ridge regressions.

It is important to note that all of 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, one cannot be sure that it is represented within the scope of our simulation study. It is possible that for factor values very different than 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. Under-determined systems.

Separate ridge regressions 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 nonsingular. 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) (3.10) is then confined to the nonzero variance subspace of the responses by using the generalized inverse of $\mathbf{Y}^t\mathbf{Y}$ in (2.25). Dealing with singular $\mathbf{X}^t\mathbf{X}$ on the other hand must be done with care.

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

$$\mathbf{X}^t\mathbf{X} = \mathbf{U}\mathbf{E}^2\mathbf{U}^t,\ \mathbf{U}^t\mathbf{U} = \mathbf{U}\mathbf{U}^t = \mathbf{I}_p,\ \mathbf{E}^2 = diag\{e_1^2, \cdots, e_r^2, 0, \cdots\} \tag{6.1}$$

where $r < p$ is the rank of $\mathbf{X}^t\mathbf{X}$, and the eigenvalues $\{e_1^2, \cdots, e_r^2\}$ are in descending order. The matrix $\mathbf{Z}_r \in R^{N\times r}$ formed by 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 (2.25) in place of $\mathbf{X}$. The regression coefficient estimates associated with the last $p - r$ columns are then all defined to have zero value. This is equivalent to using the generalized inverse of $\mathbf{X}^t\mathbf{X}$ in (2.25).

A problem with this approach is that the resulting (nonzero) 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 in 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. C&W-ridge.

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 ridge regression worked consistently better. With this approach the coefficient matrix $\hat{\mathbf{A}}_\lambda \in R^{q\times p}$ is obtained from separate ridge

regressions 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 ridge regression 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 (5-fold) cross-validation

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

The C&W-ridge estimates are then given by

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

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

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 $\hat{\mathbf{Y}}_{\hat{\lambda}} \in R^{N\times q}$

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

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

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 (6.9) and with

$$\hat{r} = \frac{1}{N}trace[\mathbf{X}(\mathbf{X}^t\mathbf{X} + \hat{\lambda}\mathbf{I}_p)^{-1}\mathbf{X}^t] \tag{6.10}$$

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

Unlike C&W-GCV, C&W-ridge 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 ridge regression, 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-ridge 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 ($p/N = 1/2$ and $1/4$) C&W-ridge 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-ridge may have the best performance. This will likely depend other aspects of the problem such as sample size and (unknown) signal-to-noise ratio. A reasonable strategy would be to compare the methods using cross-validated error estimates as a guide.

6.2. Simulation study.

For $p > N$ the competitors to C&W-ridge (Section 6.1) are separate ridge regressions (Section 4.1) and two-block PLS (Section 4.4). In order 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 (5.4) (5.5) each with two values of σ^2 chosen to give (average) signal/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) (5.11) of 0.0, 0.35, and 0.70. For each replication the predictors were generated from (5.1) (5.2) with r assigned three values: $r = 0.0$, $r = 0.90$, and $r = 0.99$. The response values were computed from (2.5) (2.6) with the true coefficient values $\{a_{ij}\}$ generated in the same manner described in Section 5.1. A full factorial design over all of the above levels gives rise to 144 situations; 100 replications were performed for each one. Thus, the entire simulation study is comprised of 14400 replications.

The performance measure used to compare the three methods is

$$RA(m) = \frac{\sum_{i=1}^{q} e_i^2(m)}{\min_{k=1,3} \sum_{i=1}^{q} e_i^2(k)}, \quad m = 1, 2, 3 \tag{6.11}$$

with $\{e_i^2(m)\}_1^q$ given by (5.12). This measures the error squared (averaged over the responses) of each method relative to the corresponding minimum over all of the

methods. For each replication (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 (6.11) over the 100 replications for each of the 144 situations.

Figure 10 shows box plots for each method of the distribution of the 144 averages of (6.11) over all situations. C&W-ridge is seen to produce the best average error (squared), or within a few percent of the best, in every situation. The corresponding quantity for separate ridge regressions 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 ridge regressions.

Figure 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. One can see from the upper left frame 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 is able to 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 ridge regressions.

The relative advantage of C&W-ridge over the other two methods is seen (Fig. 11) to increase with decreasing signal to noise ratio (upper right frame), and increasing dispersion among the response error variances (middle left frame). Its competitive advantage is slightly less for more responses (middle right frame) and more predictor variables (lower right frame). The degree of correlation among the responses does not seem to strongly effect its advantage (lower left frame). The performance of C&W-ridge is seen to dominate that of separate ridge regressions and two-block PLS for every level of every factor.

7. Examples

In this section we illustrate the application of C&W to two published data sets and compare its performance to 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 one. 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.

7.1. A chemometrics example

This data is 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 produced polymers:

y_1: number-average molecular weight

y_2: weight-average molecular weight

y_3: frequency of long chain branching

y_4: frequency of short chain branching

y_5: content of vinyl groups

y_6: content of vinylidene groups.

Because the distributions of the values of all of 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, 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. That is, 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). C&W is seen to improve the predictive accuracy of all of the responses, with that improvement being substantial for three of them (y_2, y_5, and y_6). On the whole C&W decreased the squared-error by about 20%. The GCV shrinkage factors (3.11) (3.12) are $\hat{\mathbf{D}} = diag\{0.994, 0.973, 0.864, 0.172, 0.142, 0.000\}$. This indicates that the effective response dimension is around three.

7.2. Scottish elections

Brown (1980) lists electorial results for all 71 Scottish constituencies in two British general elections of February and October 1974. The raw data given in the article consists of the total votes for each of the four parties (Conservative, Labour, Liberal, 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 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 October and February vote for each party divided by the size of the electorate. There are $p = 7$ predictor variables. The first four are the February votes for each party divided by the size of the electorate. The next three are binary variables:

$x_5 = 0.5$ if Liberal intervenes (Lib. vote in Oct. $\succ 0$, Lib. vote in Feb.$= 0$), else $x_5 = 0$,

$x_6 = 0.5$ if constituency is in a rural area, else $x_6 = 0$,

$x_7 = 0.5$ if Labour or Nationalist won in Feb. and $|x_2 - x_4| \prec 0.2$, else $x_7 = 0$.

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 baseline, we include the predictor consisting of the average of each October response over the 30 constituencies (second column). The GCV shrinkage factors (3.11) (3.12) are $\hat{\mathbf{D}} = diag\{0.96, 0.52, 0.20, 0.00\}$ indicating an effective response dimensionality of less than two.

8. Conclusion

The results presented in this paper strongly suggest that the conventional (statistical) wisdom, that one 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 favor 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 ridge regressions.

The C&W procedure tends to improve expected prediction accuracy for every response. This suggests the intriguing prospect that even when there is only a single response of interest, if there are variables 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.

9. References.

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

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

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

Brown, P, J. and Zidek, J. V. (1982). Multivariate regression shrinkage estimators with unknown covariance matrix. *Scad. J. Statist.* **9**, 209-215.

Copas, J. B. (1983). Regression, prediction, and shrinkage (with discussion). *J. Roy. Statist. Soc.* **B45**, 311-354.

Copas, J. B. (1987). Cross-validation shrinkage of regression predictors. *J. Roy. Statist. Soc.* **B49**, 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. *Numerische Mathematik* **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. Amer. Statist. Assoc.* **89**, 122-127.

Golub, G. H. and van Loan, C. F. (1989). *Matrix Computations.* Johns Hopkins Univ. 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 *Proceedings of the Fourth Berkeley Symposium (Vol. I)*, ed. J. Neyman, Berkeley: University of Calif. Press, 361-379.

Massey, W. F. (1965). Principal components regression in exploratory statistical research. *J. Amer. Statist. Assoc.* **60**, 234-246.

Skagerberg, B., MacGregor, J., and Kiparissides, C. (1992). Multivariate data analysis applied to low - density polyethylene reactors. *Chemometrics and Intelligent Laboratory Systems* **14**, 341-356.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statis. Soc.* **B36**, 111-147.

van der Merwe, A. and Zidek, J. V. (1980). Multivariate regression analysis and canonical variates. *Canadian J. Statist.* **8**, 27-39.

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. Academic Press.

Zidek, J. (1978). Deriving unbiased risk estimators of multinormal mean and regression coefficient estimators using zonal polynomials. *Annals of Statist.* **6**, 769-782.

10. Figure captions.

Figure 1: Population canonical coordinate shrinkage factors (2.23) as a function of squared (population) canonical correlation, for various ratios of parameter to observation count.

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

Figure 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.

Figure 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.

Figure 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).

Figure 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).

Figure 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.

Figure 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, ERRVAR2, respectively).

Figure 9: Interaction of method with the other factors of the ($p \prec N$) simulation design. Ordinate is average response mean-squared error relative to OLS (5.13). (Number = average response correlation, RESP = number of responses, SS = sample size, S/N = signal to noise ratio).

Figure 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).

Figure 11: Interactions of method with the other factors of the $p \succ N$ simulation design. 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).