An Interpretation of Partial Least Squares

Paul H. GARTHWAITE*

Univariate partial least squares (PLS) is a method of modeling relationships between a Y variable and other explanatory variables. It may be used with any number of explanatory variables, even far more than the number of observations. A simple interpretation is given that shows the method to be a straightforward and reasonable way of forming prediction equations. Its relationship to multivariate PLS, in which there are two or more Y variables, is examined, and an example is given in which it is compared by simulation with other methods of forming prediction equations. With univariate PLS, linear combinations of the explanatory variables are formed sequentially and related to Y by ordinary least squares regression. It is shown that these linear combinations, here called components, may be viewed as weighted averages of predictors, where each predictor holds the residual information in an explanatory variable that is not contained in earlier components, and the quantity to be predicted is the vector of residuals from regressing Y against earlier components. A similar strategy is shown to underlie multivariate PLS, except that the quantity to be predicted is a weighted average of the residuals from separately regressing each Y variable against earlier components. This clarifies the differences between univariate and multivariate PLS, and it is argued that in most situations, the univariate method is likely to give the better prediction equations. In the example using simulation, univariate PLS is compared with four other methods of forming prediction equations: ordinary least squares, forward variable selection, principal components regression, and a Stein shrinkage method. Results suggest that PLS is a useful method for forming prediction equations when there are a large number of explanatory variables, particularly when the random error variance is large.

KEY WORDS: Biased regression; Data reduction; Prediction; Regressor construction.

1. INTRODUCTION

Partial least squares (PLS) is a comparatively new method of constructing regression equations that has recently attracted much attention, with several recent papers (see, for example, Helland 1988, 1990; Hoskuldsson 1988; Stone and Brooks 1990). The method can be used for multivariate as well as univariate regression, so there may be several dependent variables, $Y_1, \ldots, Y_l$, say. To form a relationship between the Y variables and explanatory variables, $X_1, \ldots, X_m$, PLS constructs new explanatory variables, often called factors, latent variables, or *components,* where each component is a linear combination of $X_1, \ldots, X_m$. Standard regression methods are then used to determine equations relating the components to the Y variables.

The method has similarities to principal components regression (PCR), where principal components form the independent variables in a regression. The major difference is that with PCR, principal components are determined solely by the data values of the X variables, whereas with PLS, the data values of both the X and Y variables influence the construction of components. Thus PLS also has some similarity to latent root regression (Webster, Gunst, and Mason 1974), although the methods differ substantially in the ways they form components. The intention of PLS is to form components that capture most of the information in the X variables that is useful for predicting $Y_1, \ldots, Y_l$, while reducing the dimensionality of the regression problem by using fewer components than the number of X variables. PLS is considered especially useful for constructing prediction equations when there are many explanatory variables and comparatively little sample data (Hoskuldsson 1988).

A criticism of PLS is that there seems to be no well-defined modeling problem for which it provides the optimal solution, other than specifically constructed problems in which somewhat arbitrary criteria are to be optimized; see the contributions of Brown and Fearn in the discussion of Stone and Brooks (1990). Why, then, should one believe PLS to be a useful method, and in what circumstances should it be used? To answer these questions, an effort should be made to explain and motivate the steps through which PLS constructs a regression equation, using terminology that is meaningful to the intended readers. Also, of course, empirical research using real data and simulation studies have important roles.

The main purpose of this article is to provide a simple interpretation of PLS for people who like thinking in terms of univariate regressions. The case where there is a single Y variable is considered first, in Section 2. From intuitively reasonable principles, an algorithm is developed that is effectively identical to PLS but whose rationale is easier to understand, thus hopefully aiding insight into the strengths and limitations of PLS. In particular, the algorithm shows that the components derived in PLS may be viewed as weighted averages of predictors, providing some justification for the way that components are constructed. The multivariate case, where there is more than one Y variable, is considered and its relationship to the univariate case examined in Section 3.

The other purpose of this article is to illustrate by simulation that PLS can be better than other methods at forming prediction equations when the standard assumptions of regression analysis are satisfied. Parameter values used in the simulations are based on a data set from a type of application for which PLS has proved successful: forming prediction equations to relate a substance's chemical composition to its near-infrared spectra. In this application the number of X variables can be large, so sampling models of various sizes are considered, the largest containing 50 X variables. The simulations are reported in Section 4.

* Paul H. Garthwaite is Senior Lecturer, Department of Mathematical Sciences, University of Aberdeen, Aberdeen AB9 2TY, U.K. The author thanks Tom Fearn for useful discussions that benefited this article and the referees for comments and suggestions that improved it substantially.

Journal of the American Statistical Association
March 1994, Vol. 89, No. 425, Theory and Methods

2. UNIVARIATE PLS

We suppose that we have a sample of size n from which to estimate a linear relationship between Y and $X_1, \ldots, X_m$. For $i = 1, \ldots, n$, the ith datum in the sample is denoted by $(x_1(i), \ldots, x_m(i), y(i))$. Also, the vectors of observed values of Y and X_j are denoted by $\mathbf{y}$ and $\mathbf{x}_j$, so $\mathbf{y} = \{y(1), \ldots, y(n)\}'$ and, for $j = 1, \ldots, m$, $\mathbf{x}_j = \{x_j(1), \ldots, x_j(n)\}'$. Denote their sample means by $\bar{y} = \sum_i y(i)/n$ and $\bar{x}_j = \sum_i x_j(i)/n$. The regression equation will take the form

$$\hat{Y} = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \cdots + \beta_p T_p, \tag{1}$$

where each component T_k is a linear combination of the X_j and the sample correlation for any pair of components is 0.

An equation containing many parameters is typically more flexible than one containing few parameters, with the disadvantage that its parameter estimates can be more easily influenced by random errors in the data. Hence one purpose of several regression methods, such as stepwise regression, principal components regression, and latent root regression, is to reduce the number of terms in the regression equation. PLS also reduces the number of terms, as the components in Equation (1) are usually far fewer than the number of X variables. In addition, PLS aims to avoid using equations with many parameters when constructing components. To achieve this, it adopts the principle that when considering the relationship between Y and some specified X variable, other X variables are not allowed to influence the estimate of the relationship directly but are only allowed to influence it through the components T_k. From this premise, an algorithm equivalent to PLS follows in a natural fashion.

To simplify notation, Y and the X_j are centered to give variables U_1 and V_{1j}, where $U_1 = Y - \bar{y}$ and, for $j = 1, \ldots, m$,

$$V_{1j} = X_j - \bar{x}_j. \tag{2}$$

The sample means of U_1 and V_{1j} are 0, and their data values are denoted by $\mathbf{u}_1 = \mathbf{y} - \bar{y} \cdot \mathbf{1}$ and $\mathbf{v}_{1j} = \mathbf{x}_j - \bar{x}_j \cdot \mathbf{1}$, where $\mathbf{1}$ is the n-dimensional unit vector, $\{1, \ldots, 1\}'$.

The components are then determined sequentially. The first component, T_1, is intended to be useful for predicting U_1 and is constructed as a linear combination of the V_{1j}'s. During its construction, sample correlations between the V_{1j}'s are ignored. To obtain T_1, U_1 is first regressed against V_{11}, then against V_{12}, and so on for each V_{1j} in turn. Sample means are 0, so for $j = 1, \ldots, m$, the resulting least squares regression equations are

$$\hat{U}_{1(j)} = b_{1j} V_{1j}, \tag{3}$$

where $b_{1j} = \mathbf{v}_{1j}'\mathbf{u}_1/(\mathbf{v}_{1j}'\mathbf{v}_{1j})$. Given values of the V_{1j} for a further item, each of the m equations in (3) provides an estimate of U_1. To reconcile these estimates while ignoring interrelationships between the V_{1j}, one might take a simple average, $\sum_j b_{1j} V_{1j}/m$ or, more generally, a weighted average. We set T_1 equal to the weighted average, so

$$T_1 = \sum_{j=1}^{m} w_{1j} b_{1j} V_{1j}, \tag{4}$$

with $\sum_j w_{1j} = 1$. (The constraint, $\sum_j w_{1j} = 1$, aids the description of PLS, but it is not essential. As will be clear, multiplying T_1 by a constant would not affect the values of subsequent components nor predictions of Y.) Equation (4) permits a range of possibilities for constructing T_1, depending on the weights that are used; two weighting policies will be considered later.

As T_1 is a weighted average of predictors of U_1, it should itself be a useful predictor of U_1 and hence of Y. But the X variables potentially contain further useful information for predicting Y. The information in X_j that is not in T_1 may be estimated by the residuals from a regression of X_j on T_1, which are identical to the residuals from a regression of V_{1j} on T_1. Similarly, variability in Y that is not explained by T_1 can be estimated by the residuals from a regression of U_1 on T_1. These residuals will be denoted by V_{2j} for V_{1j} and by U_2 for U_1. The next component, T_2, is a linear combination of the V_{2j} that should be useful for predicting U_2. It is constructed in the same way as T_1 but with U_1 and the V_{1j}'s replaced by U_2 and the V_{2j}'s.

The procedure extends iteratively in a natural way to give components $T_2, \ldots, T_p$, where each component is determined from the residuals of regressions on the preceding component, with residual variability in Y being related to residual information in the X's. Specifically, suppose that T_i ($i \geq 1$) has just been constructed from variables U_i and V_{ij} ($j = 1, \ldots, m$) and let T_i, U_i, and the V_{ij} have sample values $\mathbf{t}_i$, $\mathbf{u}_i$, and $\mathbf{v}_{ij}$. From their construction, it will easily be seen that their sample means are all 0. To obtain T_{i+1}, first the $V_{(i+1)j}$'s and U_{i+1} are determined. For $j = 1, \ldots, m$, V_{ij} is regressed against T_i, giving $\mathbf{t}_i'\mathbf{v}_{ij}/(\mathbf{t}_i'\mathbf{t}_i)$ as the regression coefficient, and $V_{(i+1)j}$ is defined by

$$V_{(i+1)j} = V_{ij} - \{\mathbf{t}_i'\mathbf{v}_{ij}/(\mathbf{t}_i'\mathbf{t}_i)\} T_i. \tag{5}$$

Its sample values, $\mathbf{v}_{(i+1)j}$, are the residuals from the regression. Similarly, U_{i+1} is defined by $U_{i+1} = U_i - \{\mathbf{t}_i'\mathbf{u}_i/(\mathbf{t}_i'\mathbf{t}_i)\} T_i$, and its sample values, $\mathbf{u}_{i+1}$, are the residuals from the regression of U_i on T_i.

The "residual variability" in Y is U_{i+1} and the "residual information" in X_j is $V_{(i+1)j}$, so the next stage is to regress U_{i+1} against each $V_{(i+1)j}$ in turn. The jth regression yields $b_{(i+1)j} V_{(i+1)j}$ as a predictor of U_{i+1}, where

$$b_{(i+1)j} = \mathbf{v}_{(i+1)j}'\mathbf{u}_{i+1}/(\mathbf{v}_{(i+1)j}'\mathbf{v}_{(i+1)j}). \tag{6}$$

Forming a linear combination of these predictors, as in Equation (4), gives the next component,

$$T_{i+1} = \sum_{j=1}^{m} w_{(i+1)j} b_{(i+1)j} V_{(i+1)j}. \tag{7}$$

The method is repeated to obtain T_{i+2}, and so on. After the components are determined, they are related to Y using the regression model given in Equation (1), with the regression coefficients estimated by ordinary least squares.

A well-known feature of PLS is that the sample correlations between any pair of components is 0 (Helland 1988; Wold, Ruhe, Wold, and Dunn 1984). This follows because (a) the residuals from a regression are uncorrelated with a regressor so, for example, $V_{(i+1)j}$ is uncorrelated with T_i for all j; and

(b) each of the components $T_{i+1}, \ldots, T_p$ is a linear combination of the $V_{(i+1)j}$'s, so from (a), they are uncorrelated with T_i. A consequence of components being uncorrelated is that regression coefficients in Equation (1) may be estimated by simple one-variable regressions, with $\hat{\beta}_i$ obtained by regressing Y on T_i. Also, as components are added to the model, the coefficients of earlier components are unchanged. A further consequence, which simplifies interpretation of U_{i+1} and $V_{(i+1)j}$, is that $\mathbf{u}_{i+1}$ and $\mathbf{v}_{(i+1)j}$ are the vectors of residuals from the respective regressions of Y and X_j on $T_1, \ldots, T_i$.

Deciding the number of components (p) to include in the regression model is a tricky problem, and usually some form of cross-validation is used (see, for example, Stone and Brooks 1990 and Wold et al. 1984). One cross-validation procedure is described in Section 4. After an estimate of the regression model has been determined, Equations (2), (5), and (7) can be used to express it in terms of the original variables, X_j, rather than the components, T_i. This gives a more convenient equation for estimating Y for further samples on the basis of their X values.

To complete the algorithm, the mixing weights w_{ij} must be specified. For the algorithm to be equivalent to a common version of the PLS algorithm, the requirement $\sum_j w_{ij} = 1$ is relaxed and w_{ij} is set equal to $\mathbf{v}'_{ij}\mathbf{v}_{ij}$ for all i, j. [Thus $w_{ij} \propto \text{var}(V_{ij})$, as the latter equals $\mathbf{v}'_{ij}\mathbf{v}_{ij}/(n-1)$.] Then $w_{ij}b_{ij} = \mathbf{v}'_{ij}\mathbf{u}_i$ and, from Equation (7), components are given by $T_i = \sum_j (\mathbf{v}'_{ij}\mathbf{u}_i)V_{ij} \propto \sum_j \widehat{\text{cov}}(V_{ij}, U_i)V_{ij}$. This is the usual expression for determining components in PLS. A possible motivation for this weighting policy is that the w_{ij}'s are then inversely proportional to the variances of the b_{ij}'s. Also, if $\text{var}(V_{ij})$ is small relative to the sample variance of X_j, then X_j is approximately collinear with the components $T_1, \ldots, T_{i-1}$, so perhaps its contribution to T_i should be made small by making w_{ij} small. An obvious alternative weighting policy is to set each w_{ij} equal to $1/m$, so that each predictor of U_i is given equal weight. This seems a natural choice and is in the spirit of PLS, which aims to spread the load among the X variables in making predictions.

In the simulations in Section 4, the weighting policies $w_{ij} = 1/m$ (for all i, j) and $w_{ij} \propto \text{var}(V_{ij})$ are examined. The PLS methods to which these lead differ in their invariance properties. With $w_{ij} \propto \text{var}(V_{ij})$, predictions of Y are invariant only under *orthogonal* transformations of the X variables (Stone and Brooks 1990), whereas with $w_{ij} = 1/m$, predictions are invariant to changes in scale of the X variables.

3. MULTIVARIATE PLS

In this section the case is considered where there are l dependent variables, $Y_1, \ldots, Y_l$, and, as before, m independent variables, $X_1, \ldots, X_m$. The aim of multivariate PLS is to find one set of components that yields good linear models for all the Y variables. The models will have the form

$$\hat{Y}_k = \beta_{k0} + \beta_{k1}T_1 + \cdots + \beta_{kp}T_p \tag{8}$$

for $k = 1, \ldots, l$, where each of the components, $T_1, \ldots, T_p$, is a linear combination of the X variables. It should be noted that the same components occur in the model for each Y variable; only the regression coefficients change. Here the intention is to construct an algorithm that highlights the similarities between univariate and multivariate PLS and identifies their differences.

For the X variables, we use the same notation as before for the sample data and adopt similar notation for the Y's. Thus for $k = 1, \ldots, l$, the observed values of Y_k are denoted by $\mathbf{y}_k = \{y_k(1), \ldots, y_k(n)\}'$, and its sample mean is $\bar{y}_k = \sum_i y_k(i)/n$. We define $R_{1k} = Y_k - \bar{y}_k$, with sample values $\mathbf{r}_{1k} = \mathbf{y}_k - \bar{y}_k \cdot \mathbf{1}$, and V_{1j} again denotes X_j after it has been centered, with sample values $\mathbf{v}_{1j}$.

To construct the first component, T_1, define the $n \times l$ matrix $\mathbf{R}_1$ by $\mathbf{R}_1 = (\mathbf{r}_{11}, \ldots, \mathbf{r}_{1l})$ and the $n \times m$ matrix $\mathbf{V}_1$ by $\mathbf{V}_1 = (\mathbf{v}_{11}, \ldots, \mathbf{v}_{1m})$. Let $\mathbf{c}_1$ be an eigenvector corresponding to the largest eigenvalue of $\mathbf{R}'_1\mathbf{V}_1\mathbf{V}'_1\mathbf{R}_1$ and define $\mathbf{u}_1$ by $\mathbf{u}_1 = \mathbf{R}_1\mathbf{c}_1$. Then T_1 is constructed from $\mathbf{u}_1, \mathbf{v}_{11}, \ldots, \mathbf{v}_{1m}$ in precisely the same way as in Section 2. Motivation for constructing $\mathbf{u}_1$ in this way was given by Hoskuldsson (1988), who showed that if $\mathbf{f}$ and $\mathbf{g}$ are vectors of unit length that maximize $[\widehat{\text{cov}}(\mathbf{V}_1\mathbf{f}, \mathbf{R}_1\mathbf{g})]^2$, then $\mathbf{R}_1\mathbf{g}$ is proportional to $\mathbf{u}_1$.

To give the general step in the algorithm, suppose that we have determined T_i, V_{ij} for $j = 1, \ldots, m$ and R_{ij} for $k = 1, \ldots, l$, together with their sample values, $\mathbf{t}_i$, $\mathbf{v}_{ij}$, and $\mathbf{r}_{ik}$. We must indicate how to obtain these quantities as $i \to i+1$. First, $V_{(i+1)j}$ is again the residual when V_{ij} is regressed on T_i, so $V_{(i+1)j}$ and $\mathbf{v}_{(i+1)j}$ are given by Equation (5). Similarly, $R_{(i+1)j}$ is the residual when R_{ij} is regressed against T_i, so

$$R_{(i+1)k} = R_{ik} - \{\mathbf{t}'_i\mathbf{r}_{ik}/(\mathbf{t}'_i\mathbf{t}_i)\}T_i \tag{9}$$

and $\mathbf{r}_{(i+1)k}$ are its sample values. (From analogy to the X's, it is clear that $\mathbf{r}_{(i+1)k}$ is also the residual when Y_k is regressed on $T_1, \ldots, T_i$.) Put $\mathbf{R}_{i+1} = (\mathbf{r}_{(i+1)1}, \ldots, \mathbf{r}_{(i+1)l})$, $\mathbf{V}_{i+1} = (\mathbf{v}_{(i+1)1}, \ldots, \mathbf{v}_{(i+1)m})$, and let $\mathbf{c}_{i+1}$ be an eigenvector corresponding to the largest eigenvalue of $\mathbf{R}'_{i+1}\mathbf{V}_{i+1}\mathbf{V}'_{i+1}\mathbf{R}_{i+1}$. The vector $\mathbf{u}_{i+1}$ is obtained from

$$\mathbf{u}_{i+1} = \mathbf{R}_{i+1}\mathbf{c}_{i+1}, \tag{10}$$

and then T_{i+1} and $\mathbf{t}_{i+1}$ are determined as in Section 2, using Equations (6) and (7).

After $T_1, \ldots, T_p$ have been determined, each Y variable is regressed separately against these components to estimate the β coefficients in the models given by (8). Cross-validation is again used to select the value of p.

It is next shown that the preceding algorithm is equivalent to a standard version of the multivariate PLS algorithm. For the latter we use the following algorithm given by Hoskuldsson (1988), but change its notation. Denote the centered data matrices, $\mathbf{V}_1$ and $\mathbf{R}_1$, by $\mathbf{\Omega}_1$ and $\mathbf{\Phi}_1$, and suppose that $\mathbf{\Omega}_i$ and $\mathbf{\Phi}_i$ have been determined.

1. Set $\underline{\phi}$ to the first column of $\mathbf{\Phi}_i$.
2. Put $\underline{\psi} = \mathbf{\Omega}'_i\underline{\phi}/(\underline{\phi}'\underline{\phi})$ and scale $\underline{\psi}$ to be of unit length.
3. $\underline{\tau} = \mathbf{\Omega}_i\underline{\psi}$.
4. Put $\underline{\zeta} = \mathbf{\Phi}'_i\underline{\tau}/(\underline{\tau}'\underline{\tau})$ and scale $\underline{\zeta}$ to be of unit length.
5. Put $\underline{\phi} = \mathbf{\Phi}_i\underline{\zeta}$ and if there is convergence go on to step 6; otherwise return to step 2.
6. $\underline{\theta} = \mathbf{\Omega}'_i\underline{\tau}/(\underline{\tau}'\underline{\tau})$.
7. $\lambda = \underline{\tau}'\underline{\phi}/(\underline{\tau}'\underline{\tau})$.
8. Residual matrices: $\mathbf{\Omega}_{i+1} = \mathbf{\Omega}_i - \underline{\tau}\underline{\theta}'$ and $\mathbf{\Phi}_{i+1} = \mathbf{\Phi}_i - \lambda\underline{\tau}\underline{\zeta}'$.

Assume that $\mathbf{\Omega}_i = \mathbf{V}_i$ and $\mathbf{\Phi}_i = \mathbf{R}_i$ and that w_{ij}, the weights in Equation (7), are chosen so that $w_{ij} = \mathbf{v}'_{ij}\mathbf{v}_{ij}$. It must be shown that (a) $\underline{\tau} \propto \mathbf{t}_i$, (b) $\mathbf{\Omega}_{i+1} = \mathbf{V}_{i+1}$, and (c) $\mathbf{\Phi}_{i+1} = \mathbf{R}_{i+1}$.

Proof of (a). Hoskuldsson showed that when there is convergence at step 5, then $\underline{\zeta}$ is an eigenvector corresponding to the largest eigenvalue of $\mathbf{\Phi}'_i\mathbf{\Omega}_i\mathbf{\Omega}'_i\mathbf{\Phi}_i$. By assumption, $\mathbf{\Phi}'_i\mathbf{\Omega}_i\mathbf{\Omega}'_i\mathbf{\Phi}_i = \mathbf{R}'_i\mathbf{V}_i\mathbf{V}'_i\mathbf{R}_i$, so $\underline{\zeta}$ is proportional to $\mathbf{c}_i$. Hence, from step 5 and Equation (10), $\underline{\phi} \propto \mathbf{u}_i$. After convergence, repeating steps 2–5 has no effect, so, from step 2, $\underline{\psi} \propto \mathbf{\Omega}'_i\underline{\phi} \propto \mathbf{V}'_i\mathbf{u}_i$, and, from step 3, $\underline{\tau} \propto \mathbf{V}_i\mathbf{V}'_i\mathbf{u}_i$. From (6), the jth component of $\mathbf{V}'_i\mathbf{u}_i$ is $w_{ij}b_{ij}$ (by assumption, $w_{ij} = \mathbf{v}'_{ij}\mathbf{v}_{ij}$); so from (7), $\mathbf{t}_i = \mathbf{V}_i\mathbf{V}'_i\mathbf{u}_i$. Hence $\underline{\tau} \propto \mathbf{t}_i$.

Proof of (b). From steps 6 and 8, $\mathbf{\Omega}_i - \mathbf{\Omega}_{i+1} = \underline{\tau}\underline{\theta}' = \underline{\tau}\underline{\tau}'\mathbf{\Omega}_i/(\underline{\tau}'\underline{\tau}) = \mathbf{t}_i\mathbf{t}'_i\mathbf{V}_i/(\mathbf{t}'_i\mathbf{t}_i)$, because $\underline{\tau} \propto \mathbf{t}_i$ and $\mathbf{\Omega}_i = \mathbf{V}_i$. The jth column of $\mathbf{t}_i\mathbf{t}'_i\mathbf{V}_i/(\mathbf{t}'_i\mathbf{t}_i)$ is $\mathbf{t}_i\mathbf{t}'_i\mathbf{v}_{ij}/(\mathbf{t}'_i\mathbf{t}_i) = \mathbf{t}_i(\mathbf{t}'_i\mathbf{v}_{ij})/(\mathbf{t}'_i\mathbf{t}_i)$. From (5), the latter term equals $\mathbf{v}_{ij} - \mathbf{v}_{(i+1)j}$. Hence $\mathbf{\Omega}_i - \mathbf{\Omega}_{i+1} = \mathbf{V}_i - \mathbf{V}_{i+1}$.

Proof of (c). Let $\underline{\zeta} = \kappa\mathbf{\Phi}'_i\underline{\tau}/(\underline{\tau}'\underline{\tau})$, where κ is a constant for which $\underline{\zeta}'\underline{\zeta} = 1$ (step 4). Then from steps 5 and 7, $\lambda = \underline{\tau}'\underline{\phi}/(\underline{\tau}'\underline{\tau}) = \underline{\tau}'\mathbf{\Phi}_i\underline{\zeta}/(\underline{\tau}'\underline{\tau}) = \underline{\zeta}'\underline{\zeta}/\kappa = 1/\kappa$, so $\lambda\underline{\zeta}' = \underline{\zeta}'/\kappa = \underline{\tau}'\mathbf{\Phi}'_i/(\underline{\tau}'\underline{\tau})$. From step 8, $\mathbf{\Phi}_i - \mathbf{\Phi}_{i+1} = \lambda\underline{\tau}\underline{\zeta}'$, so $\mathbf{\Phi}_i - \mathbf{\Phi}_{i+1} = \underline{\tau}\underline{\tau}'\mathbf{\Phi}_i/(\underline{\tau}'\underline{\tau}) = \mathbf{t}_i\mathbf{t}'_i\mathbf{R}_i/(\mathbf{t}'_i\mathbf{t}_i)$. From (9), the latter term also equals $\mathbf{R}_i - \mathbf{R}_{i+1}$.

In situations where there are several Y variables and multivariate PLS could be used, an alternative is repeated application of univariate PLS. Each Y variable would be taken in turn and a regression equation determined from just its sample values and the explanatory variables. To compare univariate and multivariate PLS, suppose that a regression equation is being determined for one of the dependent variables, Y^* say, and consider the way in which the component T_{i+1} is constructed after the components $T_1, \ldots, T_i$ have been determined. With both PLS methods, T_{i+1} is determined from $\mathbf{u}_{i+1}$ and the $\mathbf{v}_{(i+1)j}$'s where, for $j = 1, \ldots, m$, $\mathbf{v}_{(i+1)j}$ is the residual from a multiple regression of X_j on $T_1, \ldots, T_i$. The *only* difference between the methods is in the way $\mathbf{u}_{i+1}$ is formed. With univariate PLS, $\mathbf{u}_{i+1}$ is the residual when Y^* is regressed on $T_1, \ldots, T_i$, whereas with multivariate PLS, each Y_k is regressed separately against $T_1, \ldots, T_i$, and $\mathbf{u}_{i+1}$ is a linear combination of the residual vectors; compare Equation (10). Choosing between multivariate and univariate PLS is equivalent to deciding the way to form $\mathbf{u}_{i+1}$ and, although one might expect multivariate PLS to use more information than univariate PLS, they actually use identical amounts in other stages of the algorithm.

To discuss the question of which PLS method is expected to give the more accurate prediction equation, three hypothetical examples are considered. In each, chemical characteristics of samples must be predicted from their near-infrared spectral readings at different wavelengths, using prediction equations derived from calibration samples for which both chemical values and spectral readings are available.

- *Example 1.* Three Y variables: concentrations of protein, starch, and sugar. An equation for estimating protein concentration is required.
- *Example 2.* Two Y variables: baking quality of wheat and its protein content. Baking quality is to be predicted.
- *Example 3.* The same as Example 2, except protein content is to be predicted.

Multivariate PLS aims to find components that are good predictors of all Y variables, but for Example 1, this aim seems inappropriate. For predicting protein, components preferably should be sensitive to protein concentration and reasonably insensitive to starch and sugar concentrations, so that only changes in the protein level affect predictions. In contrast, Example 2 is a case where it might be advantageous to seek components that are good predictors of both the Y variables. The baking quality of wheat is highly dependent on its protein content, and protein content can be measured much more accurately. Hence for predicting baking quality, protein might provide a useful guide to suitable components. In Example 3, clearly a different weighting policy from that in Example 2 should be used, because protein content is the variable of interest. Indeed, because protein can be measured more accurately than baking quality, for Example 3 it seems reasonable to give very little weight to baking quality readings.

PLS methods have been used mostly for problems similar to Example 1, so it is perhaps not surprising that univariate PLS has been found to generally perform better than multivariate PLS. Examples 2 and 3 illustrate that if multivariate PLS is used, then the weight placed on the different Y variables should reflect which variable is to be predicted. That is, although more than one Y variable might influence the construction of components, it can be preferable to construct a separate set of components for predicting each Y. This differs from the way that multivariate PLS is normally used; a single set of components for predicting all the Y variables has generally been advocated (Hoskuldsson 1988; Sjostrom, Wold, Lindberg, Persson, and Martens 1983). Changing the relative importance of the Y's is not difficult and can be achieved simply by rescaling them, but an appropriate scaling is difficult to decide and theoretical results to guide its choice are lacking. In practice, Y variables that are not closely related to the one to be predicted should probably be ignored and cross-validation used to compare different scalings of those Y variables thought relevant.

4. SIMULATION COMPARISONS

4.1 Model and Parameter Values

In this section the performance of PLS and other methods of forming prediction equations are compared. Rather than analyze real data sets, simulation was used so that models could be controlled, enabling the standard assumptions of regression analysis to be satisfied fully and the model parameters to be varied systematically. The intention is to identify situations where PLS performs well, so parameter values were based on a set of near-infrared (NIR) data, a type of data for which PLS has proved useful.

For the simulations, explanatory variables were given a joint multivariate normal distribution, $(X_1, \ldots, X_m)' \sim MVN(\underline{\mu}, \mathbf{\Gamma})$. When these variables have the value $\mathbf{x} = (x_1, \ldots, x_m)'$, Y is given by the regression equation,

$$Y = \alpha_0 + \underline{\alpha}'\mathbf{x} + \varepsilon, \quad (11)$$

where α_0 and the vector $\underline{\alpha}$ are unknown constants and $\varepsilon \sim N(0, \sigma^2)$.

A feature of NIR data is that the number of explanatory variables is large, commonly equaling 700, and the number of sample points is much smaller. To widen the scope of results, such extreme cases were not used, and models contained 8, 20, or 50 explanatory variables. The simulated data sets contained 40 more observations than the number of explanatory variables, so OLS methods could be applied straightforwardly.

To choose parameter values, data from a set of 195 hay samples were used. NIR spectra of the samples were transformed to reduce the effect of particle-size variation (as is standard practice in NIR analysis), and then the transformed values at 50 wavelengths were extracted. Their mean and variance–covariance matrix were determined and used as the values of $\underline{\mu}$ and $\mathbf{\Gamma}$ for the model containing 50 independent variables. For each smaller model, spectral values for a random subset of the 50 wavelengths were used. Measurements of neutral detergent fiber for each hay sample had been determined by chemical analysis. These were regressed against the transformed spectral values, and the estimated regression coefficients were taken as the values of α_0 and $\underline{\alpha}$ in Equation (11). For the hay data, the error variance, σ^2, equalled about 5.0. But it was thought the performance of PLS relative to other methods might be sensitive to this parameter, so values $\sigma^2 = 1.0$, 3.0, 5.0, 7.0, and 10.0 were examined.

4.2 Regression Methods

Six methods of forming prediction equations are examined. The first two are forms of PLS that differ only in the mixing weights, w_{ij}, that they use. In PLS(E), the weights are set equal to each other, and in PLS(U), they are unequal, with $w_{ij} \propto \text{var}(V_{ij})$. With both methods, the following cross-validation procedure was used to select the number of components to include in a model for a given (simulated) data set. First, the data were split into three groups. One group at a time was omitted, and data from the other groups were used to construct components and determine a prediction equation for Y. This equation was used to predict Y values for the group that was omitted, and the predictions were compared with the group's actual values. This was repeated until each of the groups had been omitted once, and then the total sum of squared errors in prediction over all groups was calculated. Components were added to the regression model until the next component would increase this total sum of squared errors. (The data could have been partitioned into any number of groups, but three groups seemed adequate and a larger number would have required more computer time.)

The third method used to form prediction equations was ordinary least squares (OLS) using all of the X variables in the regression model. The fourth method (FVS) used forward variable selection to construct regression models. Cross-validation might have been used to decide when to stop selecting variables but, in line with common practice, F test values were used instead. At each step, the "best" X variable not in the model was added to it if the partial F test value for that variable's inclusion exceeded 4.0. The fifth method is principal components regression (PCR). Principal components were computed from the sample covariance matrix of the X variables and used as the independent variables in a regression with variable selection. The dependent variable was Y and, as with FVS, a principal component was added to the regression model if the relevant F test value exceeded 4.0.

The last method we examine is a Stein shrinkage method (SSM) given by Copas (1983), who showed that it is uniformly better than OLS for the loss function used here. Suppose that we have a sample of size n and that, for simplicity, the X variables have been centered so that their sample means are 0. Let the prediction equation from an OLS regression be $\hat{y} = \bar{y} + \mathbf{a}'\mathbf{x}$ and let $\hat{\sigma}^2$ be the residual mean squared error on $\nu = n - m - 1$ degrees of freedom. Also, let the centered sample data for $X_1, \ldots, X_m$ be denoted by the $n \times m$ matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$. Then the SSM prediction equation is $\hat{y} = \bar{y} + K\mathbf{a}'\mathbf{x}$, where the shrinkage factor, K, is given by $K = 1 - (m - 2)\hat{\sigma}^2/\{(1 + 2\nu^{-1})\mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{a}\}$.

4.3 Loss Function and Simulation Procedure

Suppose that a prediction equation has the form $\hat{y} = \hat{\alpha}_0 + \underline{\hat{\alpha}}'\mathbf{x}$. Then, from Equation (11), $y - \hat{y} = (\alpha_0 - \hat{\alpha}_0) + (\underline{\alpha} - \underline{\hat{\alpha}})'\mathbf{x} + \varepsilon$. Given $\hat{\alpha}_0$ and $\underline{\hat{\alpha}}$, the expected squared error in predicting y can be determined for a future, as yet unknown, $\mathbf{x}$ value, because the distributions of the X's and ε are known. From this prediction mean squared error we subtract σ^2, the contribution of random error. This leaves the loss caused by inaccuracy in estimating the regression coefficients,

$$\text{Loss} = [(\alpha_0 - \hat{\alpha}_0) + (\underline{\alpha} - \underline{\hat{\alpha}})'\underline{\mu}]^2 + (\underline{\alpha} - \underline{\hat{\alpha}})'\mathbf{\Gamma}(\underline{\alpha} - \underline{\hat{\alpha}}), \quad (12)$$

which we take as the loss function.

In the simulations, a model size ($m = 8$, 20, or 50) and random error variance ($\sigma^2 = 1.0$, 3.0, 5.0, 7.0, or 10.0) were selected and, using the parameter values corresponding to that model size, a sample set of $40 + m$ data were simulated. Each datum consisted of values of Y and the X's. From the sample set, prediction equations were estimated using each of the six regression methods described previously, and the accuracy of the equations was measured by the loss function given in Equation (12). The procedure was replicated 500 times for each model size and error variance, and the average loss was determined for each regression method.

4.4 Results

The results of the simulations are given in Table 1. Copas (1983) showed that the expected loss for OLS is $\sigma^2\{n(m + 1) - 2\}/\{n(n - m + 2)\}$. This gives theoretical values that typically differ by about 1.8% from the average losses for OLS in Table 1, indicating that an adequate number of replicates were used in simulations. In the first six rows of the table, OLS and SSM have the smallest average losses,

Table 1. Average Loss for Six Methods of Forming Prediction Equations, for Different Model Sizes and Error Variances

Model size[a]	Error variance	Model					
		OLS	SSM	FVS[b]	PCR[c]	PLS(E)[c]	PLS(U)[c]
8	1.0	.24	.24	.32 (6.1)	.31 (6.1)	.30 (4.8)	.36 (5.5)
8	3.0	.72	.71	1.19 (4.7)	.94 (5.1)	.85 (3.8)	.91 (4.5)
8	5.0	1.19	1.18	1.86 (3.9)	1.50 (4.5)	1.48 (3.4)	1.55 (3.9)
8	7.0	1.67	1.64	2.44 (3.5)	2.02 (4.2)	2.03 (3.0)	2.10 (3.6)
8	10.0	2.40	2.32	3.13 (3.2)	2.85 (3.8)	2.69 (2.6)	2.75 (3.3)
20	1.0	.54	.53	.68 (9.6)	.69 (11.2)	.73 (5.6)	.74 (6.3)
20	3.0	1.73	1.69	1.91 (7.0)	1.84 (8.3)	1.74 (3.8)	1.64 (4.7)
20	5.0	2.70	2.59	2.71 (5.9)	2.71 (7.0)	2.38 (3.2)	2.33 (4.1)
20	7.0	3.73	3.53	3.44 (5.3)	3.44 (6.3)	2.86 (2.7)	2.99 (3.7)
20	10.0	5.58	5.15	4.52 (4.8)	4.64 (5.7)	3.46 (2.5)	4.01 (3.3)
50	1.0	1.34	1.31	1.28 (13.3)	1.20 (19.4)	1.14 (6.7)	1.11 (8.1)
50	3.0	4.14	3.95	2.33 (9.5)	2.93 (14.3)	2.37 (4.1)	2.30 (5.4)
50	5.0	6.80	6.26	3.06 (8.0)	4.30 (12.2)	2.98 (3.3)	3.03 (4.5)
50	7.0	9.32	8.29	3.74 (7.2)	5.51 (10.6)	3.47 (2.9)	3.64 (3.9)
50	10.0	13.23	11.25	4.53 (6.4)	7.58 (9.8)	3.88 (2.6)	4.38 (3.4)

[a] Number of X variables.
[b] Average number of variables in the fitted regression in parentheses.
[c] Average number of components in the fitted regression in parentheses.

whereas in the last eight rows, average losses for the PLS methods are smallest, suggesting that PLS is likely to prove most useful when the number of explanatory variables and the error variance are both large. The simulations also illustrate the potential benefit of biased regression methods. OLS consistently has a slightly higher average loss than SSM, as theory predicts, and they both have losses that are substantially higher than other methods when the model size and error variance are large.

Other studies using real data from NIR applications have found that PCR generally gives poorer prediction equations than PLS methods (see, for example, Sjostrom et al. 1983). The results here tentatively suggest that this is not due to NIR data failing to satisfy the usual assumptions made in regression analysis. Table 1 also shows that PLS(E) tended to use fewer components in prediction equations than did PLS(U), but there was little to choose from between these methods in their average losses.

The strength of collinearities between explanatory variables can influence the relative performance of prediction methods (Gunst and Mason 1977). In the simulations so far, the explanatory variables have strong collinearities, as is common with NIR data. To examine the effect of weakening them, simulations were repeated with each diagonal element of Γ increased by 20%. Average losses for OLS are independent of Γ and hence were essentially unchanged from Table 1. This was also the case with SSM, but losses for other methods generally increased. For the PLS methods the changes were sometimes substantial, the greatest being from 3.9 to 7.4, and only FVS had larger increases. Despite this, the PLS methods were still the best for models containing 20 variables when $\sigma^2 = 7.0$ and 10.0 and for all models containing 50 variables, except when $\sigma^2 = 1.0$. This is consistent with the view that PLS methods are suited to models with many variables and large error variances.

[*Received January 1992. Revised March 1993.*]

REFERENCES

Copas, J. B. (1983), "Regression, Prediction, and Shrinkage" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 45, 311–354.

Gunst, R. F., and Mason, R. L. (1977), "Biased Estimation in Regression: An Evaluation Using Mean Squared Error," *Journal of the American Statistical Association,* 72, 616–628.

Helland, I. S. (1988), "On the Structure of Partial Least Squares Regression," *Communications in Statistics, Part B Simulation and Computations,* 17, 581–607.

——— (1990), "Partial Least Squares Regression and Statistical Methods," *Scandinavian Journal of Statistics,* 17, 97–114.

Hoskuldsson, P. (1988), "PLS Regression Methods," *Journal of Chemometrics,* 2, 211–228.

Sjostrom, M., Wold, S., Lindberg, W., Persson, J.-A. and Martens, H. (1983), "A Multivariate Calibration Problem in Analytical Chemistry Solved by Partial Least-Squares in Latent Variables," *Analytica Chemica Acta,* 150, 61–70.

Stone, M., and Brooks, R. J. (1990), "Continuum Regression: Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 52, 237–269.

Webster, J. T., Gunst, R. F., and Mason, R. L. (1974), "Latent Root Regression Analysis," *Technometrics,* 16, 513–522.

Wold, S., Ruhe, A., Wold, H., and Dunn, W. J. (1984), "The Collinearity Problem in Linear Regression: The Partial Least Squares (PLS) Approach to Generalized Inverses," *SIAM Journal on Scientific and Statistical Computing,* 5, 735–743.