COMMUN. STATIST. -SIMULA., 17(2), 581-607 (1988)

ON THE STRUCTURE OF PARTIAL LEAST SQUARES REGRESSION

Inge S. Helland

Department of Mathematical Sciences
Agricultural University of Norway
N-1432 Aas-NLH, Norway

Key Words and Phrases: partial least squares regression; PLS-algorithms; prediction equation; biased regression; relevant factors.

ABSTRACT

We prove that the two algorithms given in the literature for partial least squares regression are equivalent, and use this equivalence to give an explicit formula for the resulting prediction equation. This in turn is used to investigate the regression method from several points of view. Its relation to principal component regression is clearified, and some heuristic arguments are given to explain why partial least squares regression often needs fewer factors to give its optimal prediction.

1. INTRODUCTION

Useful statistical methods may often have their first development outside the stâtistical community; factor analysis and other multivariate techniques probably being the most important examples. In the present paper we will examine one particular method of analysis that has been developed in recent years mainly by chemometricians, namely the socalled partial least squares regression, or PLS-regression for short.

In spectroscopy one aim is to predict chemical composition from, e.g., near infrared reflextance spectra of meat or of cereal. If the signals for each particular wavelength are considered as explanatory variables, traditional regression methods soon run into collinearity problems, since the number of wavelengths can be up to several hundred, often exceeding the number of chemical samples (objects). Many commercial calibration instruments have used selection-of-variables-techniques to limit the number of wavelengths, but this probably represents a loss of information in many cases. Also, it is difficult to assess the accuracy of the prediction when variables have been selected from the model using the same data that are being used as basis for the prediction. From a statistician's point of view, the most satisfying solution is probably to use calibration methods based on models ("Beer's law") of the effect on the spectra caused by the chemical composition in the sample. See, e.g., Næs (1985), where assuming a factor structure for the error was essential for making the model to work. Other calibration methods for such situations have been proposed by Sundberg and Brown (1985); multivariate calibration in general is reviewed by Brown (1982).

However, regression of chemical variables on spectral variables is still being used and will probably continue to be used to a large extent, even though it to many can seem like a confusion of cause and effect. One rational behind such methods goes back to Krutchkoff (1967), who showed by simulation that a large gain in mean square error (MSE) can result from performing the regression in the direction of prediction, even though the statistical model works with a cause/effect connection in the other direction. The debate following Krutchkoff's paper is still going on, but many authors have come down by supporting his conclusion for the case where one only wants to predict over a limited range of the variables. One recent paper extending this conclusion (for a large statistical sample) to the multivariate case, is Sundberg (1985).

For finite or moderate samples there remains the collinearity problem. The two most used statistical techniques for overcoming this problem are principal component regression (PCR) and ridge regression (RR). The later method has been critized in this connection by Fearn (1983), but his critique has been countered by Hoerl et al. (1985) and Næs et al. (1986). Both methods require large amount of computation when the number of variables is large. In ridge regression there is the problem of estimating the ridge parameter (see Smith and Campbell (1980) and the discussion there), and in PCR there is the problem of choosing which principal components to delete (see Joliffe (1982), Mason and Gunst (1985) and references in these papers).

By some of its proponents, PLS-regression is claimed to overcome most, if not all of these difficulties, and, to some extent they may seem to be right in this. PLS tackles the model-problem by modelling both chemical and spectral data as functions of common latent variables (although the concept of model is rather imprecise; a new term, "soft modelling", has been put forward to motivate this lack of precision). PLS overcomes the collinearity-problem in a similar way as PCR, and simulations even tend to show that PLS reaches its minimal

MSE with a smaller number of factors than PCR. We will give some theoretical arguments to support this later. Finally, PLS gives a unique way of choosing the factors, contrary to PCR, and it requires less computation than both PCR and RR.

The PLS-algorithm has its origin in Herman Wolds general systems-analysis models (see Wold and Jöreskog,1982). As a calibration method to predict chemical variables from spectral data, it has been developed mainly by Svante Wold and Harald Martens (see Wold et al. (1983), Martens and Jensen (1983), Martens (1985), Næs and Martens (1985) and references in these papers). For chemists, a review of the method has been given by Geladi and Kowalski (1986). Outside spectroscopy, the algorithm has been used for instance in the interpretation of ANOVA interactions (Aastveit and Martens, 1986), in sensory testing (Martens et al., 1983a) and in the interpretation of computer tomograph pictures (Martens et al. 1983b). A recent review of PLS as a multivariate calibration method is given by Martens (1987).

In the literature, the PLS-method is usually presented as an algorithm. In fact, two seemingly completely different algorithms are being presented. The first one, being the algorithm usually implemented in applications, is shown by Wold et al. (1984) to be related to the conjugate gradient method for inverting matrices. This relation has been further exploited by Manne (1987). The second algorithm has been treated from a statistical point of view by Næs and Martens (1985), where many of the ideas discussed in the present paper were first introduced. A formal proof of the equivalence between the two algorithms is given in Theorem 2.1 below.

It is first in the last few years that theoretical papers connected to the PLS algorithms have appeared. In addition to those mentioned above, we should also mention Lorber et al. (1987) and Höskuldsson (1987). There are still many unresolved questions in this area. What matematical statisticians can contribute with, is first to try to set a standard of rigor. When new methods are being developed, it is often fruitful to rely on intuition and heuristic arguments. However, after some time one can usually gain much by trying to replace heuristics and empirical evidence by rigorously proved mathemathical results.

Secondly, mathematical statisticians can contribute by introducing proper probabilistic models. It is difficult to evaluate calibration methods by just looking at empirical results. Either by simulation or by algebraic results it should be possible to study the performance of different methods under different model assumptions. Here much remains to be done.

The purpose of the present paper is first of all to look at some algebraic structures related to the two PLS-algothitms, and to exploit the connection to principal component regression. The relationship between PLS and statistical models with latent variables will be studied elsewhere.

2. THE ALGORITHMS

2.1 Motivation

We will limit ourselves to the PLS1-case where the relationship between a set of variables and one single variable is studied. More general cases are treated in some of the references given above.

Let the basic data be given by

$$\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_K) \text{ and } \mathbf{y},$$

where each of the vectors $\mathbf{x}_1,\ldots,\mathbf{x}_K$ and $\mathbf{y}$ are N-dimensional, corresponding to observations on N units (i.e. chemical samples.) In the spectroscopy-situation, the $\mathbf{y}$ will represent some chemical variable and the $\mathbf{x}_k$'s will be measurements at different wavelengths, but we will have in mind any regression-type situation where the number K of variables is fairly large.

Until further notice we will suppose that the means $\overline{x}_1,...\overline{x}_K, \overline{y}$ have been subtracted from the variables $\mathbf{x}_1,...,\mathbf{x}_K, \mathbf{y}$. Like PCR and RR, the analysis that we will describe is not scale-invariant, so the $\mathbf{x}_k$-variables will usually be scaled in some way before the analysis (see Marquardt (1980) for the corresponding argument in the RR-case). Often each variable is scaled to unit variance, but note that such a scaling implies difficulties with the population interpretation of the covariance matrix.

Data matrices such as $\mathbf{X}$ can often be described in a meaningful way in a bilinear (factor-) form

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2' + ... + \mathbf{t}_A\mathbf{p}_A' + \mathbf{E}_A \tag{2.1}$$

where the scores $\mathbf{t}_a$ are N-vectors ("latent variables"), the loadings $\mathbf{p}_a$ are K-vectors and the residual matrix $\mathbf{E}_A$ is "small" in some sense. For a review of bilinear models, see Kruskal (1978). For an interesting argument why most reasonable matrices can be decomposed in bilinear form with few terms, see Wold (1974). An alternative argument can be based upon the singular value decomposition of $\mathbf{X}$.

The basis for the PLS-method is that the relation between $\mathbf{X}$ and $\mathbf{y}$ is conveyed through the latent variables. That means that one also has a decomposition

$$\mathbf{y} = \mathbf{t}_1 q_1 + \mathbf{t}_2 q_2 + ... + \mathbf{t}_A q_A + \mathbf{f}_A \tag{2.2}$$

for scalars q_a and with the same scores.

The question now is how to calculate the scores and the loadings. (I will not use the word estimate here, for I have not yet introduced any statistical models and hence no parameters.) Both because of the well-known indeterminancy caused by $\mathbf{TP} = (\mathbf{TC})(\mathbf{C}^{-1}\mathbf{P})$, and because we have not said anything yet about the residuals, there are many ways of doing this. To get a higher degree of uniqueness, one can impose various conditions upon the $\mathbf{t}_a$'s and $\mathbf{p}_a$'s. One common set of conditions is to force the scores to be mutually orthogonal in $\mathcal{R}^N$ or the loadings to be mutually orthogonal in $\mathcal{R}^K$. If both these requirements are imposed (and one also assumes orthogonality to rows/ columns of $\mathbf{E}_A$), it is easy to see from (2.1) that each $\mathbf{t}_a$ must be an eigenvector to $\mathbf{XX}'$, and that each $\mathbf{p}_a$ must be an eigenvector to $\mathbf{X}'\mathbf{X}$. Hence all the vectors are essentially determined from the **X**-data.

If one then wants to use both (2.1) and (2.2) to get a good fit, one is thus forced to relax upon the orthogonality requirements. This can be seen as the reason why one is lead to two different, but equivalent algorithms. In the first one, the scores are orthogonal in $\mathcal{R}^N$, in the second one, the loadings are orthogonal in $\mathcal{R}^K$. The first algorithm turns out to be the easiest computationally, and the scores and loadings from this algorithm are probably the simplest to find practical interpretations for. So from a practical point of view, this algorithm is all one need to be familiar with. To find a mathematical interpretation of the resulting prediction equation, however, and to investigate its relationship to principal component regression, it turns out to be far easier to use the second algorithm. Hence both will be introduced below, and they will be shown to be equivalent in the sense that they give the same prediction equation.

2.2 The original PLS-algorithm

The aim is to find representations of the form (2.1) and (2.2) for each A up to a maximal number. Hence, writing $\mathbf{E}_0 = \mathbf{X}$ and $\mathbf{f}_0 = \mathbf{y}$, one must have

$$\begin{aligned}\mathbf{E}_a &= \mathbf{E}_{a-1} - \mathbf{t}_a\mathbf{p}_a' \\ \mathbf{f}_a &= \mathbf{f}_{a-1} - \mathbf{t}_a q_a\end{aligned} \qquad a = 1,2,... \quad . \tag{2.3}$$

To fit into these equatons, $\mathbf{t}_a$, $\mathbf{p}_a$ and q_a are determined by induction.

The basic point now is that each $\mathbf{t}_a$ is determined as a linear combination of the $\mathbf{x}$-residuals from the previous step. In particular, for a=1 one wants

$$\mathbf{t}_1 = \sum_{k=1}^{K} \mathbf{x}_k w_{k1} = \mathbf{X}\mathbf{w}_1 \quad , \tag{2.4}$$

where $\mathbf{w}_1$ is a K-dimensional weight-vector. It is desired that $\mathbf{t}_1$ should be highly correlated with $\mathbf{y}$, and a reasonable choice is to make each component w_{k1} proportional to the covariance between $\mathbf{x}_k$ and $\mathbf{y}$. We will take

$$w_{k1} = \mathbf{x}_k'\mathbf{y} \quad , \quad \text{i.e.:} \quad \mathbf{w}_1 = \mathbf{X}'\mathbf{y} \ . \tag{2.5}$$

Different normalizations are used here in different papers on this subject. As long as the products $\mathbf{t}_a\mathbf{p}_a'$ and $\mathbf{t}_a q_a$ are conserved, it is easy to see that this normalization is of no consequences; we will use the one that gives the simplest formulas in most cases. Some of the formulas below would have been slightly simpler if we had used the normalization $\mathbf{w}_1'\mathbf{w}_1 = 1$ etc., but this would lead to unnecessary computations, and we would not so easily see the point that the algorithm in principle can stop automatically after a certain number of factors.

For general a we now take

$$\mathbf{t}_a = \mathbf{E}_{a-1}\mathbf{w}_a \tag{2.6}$$

$$\mathbf{w}_a = \mathbf{E}_{a-1}'\mathbf{f}_{a-1} \tag{2.7}$$

and $\mathbf{p}_a$ and q_a are then determined such that one gets a best possible fit in (2.3). That is, for a=1, the best fit to $\mathbf{y} = \mathbf{t}_1 q_1 + \mathbf{f}_1$ is given by the regression coefficient $q_1^* = \mathbf{y}'\mathbf{t}_1/\mathbf{t}_1'\mathbf{t}_1$, and similarly $\mathbf{x}_k = \mathbf{t}_1 p_{k1} + \mathbf{e}_{k1}$ gives $p_{k1} = \mathbf{x}_k'\mathbf{t}_1/\mathbf{t}_1'\mathbf{t}_1$ (k=1,...,K) or $\mathbf{p}_1 = \mathbf{X}'\mathbf{t}_1/\mathbf{t}_1'\mathbf{t}_1$. For general a, then

$$\mathbf{p}_a = \mathbf{E}_{a-1}'\mathbf{t}_a/\mathbf{t}_a'\mathbf{t}_a \qquad [= \mathbf{X}'\mathbf{t}_a/\mathbf{t}_a'\mathbf{t}_a] \tag{2.8}$$

$$q_a = \mathbf{f}_{a-1}'\mathbf{t}_a/\mathbf{t}_a'\mathbf{t}_a \qquad [= \mathbf{y}'\mathbf{t}_a/\mathbf{t}_a'\mathbf{t}_a] \tag{2.9}$$

The new residuals $\mathbf{E}_a$ and $\mathbf{f}_a$ are found from (2.3).

The number of factors to retain in the final equation is usually determined by a crossvalidation procedure: The data set is divided into G parts, calibration is done in turn with one part removed and validated on this last part. The number of factors is chosen so that the estimated error of prediction is minimized. Crossvalidation in general is discussed by Stone (1974), in a principal component context by Wold (1978). Crossvalidation in PLS is described by several authors, e.g., Wold et al. (1984). An alternative criterion - leverage corrected mean square error - which requires less calculation than crossvalidation, is proposed by Martens and Næs (1987).

The one-dimensional regression-coefficients found at each step (2.8)-(2.9) have given origin to the term "partial least square". In some papers, (2.6) and (2.7) are also motivated as least squares fits, but this does not make much sense. The choice of the weight-factors (2.5) and (2.7) is probably the point that is most poorly motivated, but we will show later that it does lead to fairly reasonable results. Also, of course, it gives simple computations. By modifying these weights in various ways, one can introduce a rich class of calibration methods, which deserve closer study. In this paper we will stick to the choice (2.7).

If now $\mathbf{x}_0 = (x_{01}, x_{02},\ldots,x_{0k})'$ is a set of x-measurement on a new unit, one defines $\mathbf{e}_0 = \mathbf{x}_0-\overline{\mathbf{x}}$ with $\overline{\mathbf{x}} = (\overline{x}_1,\ldots,\overline{x}_K)'$ and then new scores and residuals consecutively by

$$t_{a0} = \mathbf{e}_{a-1}'\mathbf{w}_a\,, \quad \mathbf{e}_a = \mathbf{e}_{a-1} - t_{a0}\mathbf{p}_a \quad . \tag{2.10}$$

Then the corresponding y_0-value is predicted in step A by

$$\hat{y}_{A0} = \overline{y} + \sum_{a=1}^{A} t_{a0}q_a = \overline{y} + \sum_{a=1}^{A} t_{a0}(\mathbf{t}_a'\mathbf{t}_a)^{-1}\mathbf{t}_a'\mathbf{y} \tag{2.11}$$

Simulation on real and artificial data have shown that this predictor performs similarly to the PCR-predictor (see Næs et al (1986) and references there); and it even tends to get its smallest estimated prediction error after fewer terms than PCR. This, together with the computational simplicity, represents the attractive feature of the procedure. Its main unattractive characteristic is its complete lack up to now of known distributional properties under any reasonable model.

2.3 An alternative algorithm and the equivalence between the two

This second algorithm for PLS was introduced by H. Martens (see Martens, 1985), and is presented in several of his papers. In particular it is the basis for the theoretical discussion in Næs and Martens (1985). It differs from the first algorithm in that one has to use multiple regression to find the loadings, and that new q-loadings are defined in each step.

Again we use the normalization that gives the simplest formulas. As a start, put $\mathbf{E}_0^* = \mathbf{X}$, $\mathbf{f}_0^* = \mathbf{y}$, and then determine $\mathbf{p}_a^*$, $\mathbf{t}_a^*$, $\mathbf{T}_a^*$, $\tilde{\mathbf{q}}_a^* = (q_{a1}^*,\ldots,q_{aa}^*)'$, $\mathbf{E}_a^*$ and $\mathbf{f}_a^*$ consecutively by the formulas

$$\mathbf{p}_a^* = \mathbf{E}_{a-1}^{*\prime}\mathbf{f}_{a-1}^* \tag{2.12}$$

$$\mathbf{t}_a^* = \mathbf{E}_{a-1}^*\mathbf{p}_a^*/\mathbf{p}_a^{*\prime}\mathbf{p}_a^* \quad [= \mathbf{X}\mathbf{p}_a^*/\mathbf{p}_a^{*\prime}\mathbf{p}_a^*] \tag{2.13}$$

$$\mathbf{T}_a^* = (\mathbf{t}_1^*,\dots,\mathbf{t}_a^*) \tag{2.14}$$

$$\tilde{\mathbf{q}}_a^* = (\mathbf{T}_a^{*\prime}\mathbf{T}_a^*)^{-1}\,\mathbf{T}_a^{*\prime}\mathbf{y} \tag{2.15}$$

$$\mathbf{E}_a^* = \mathbf{E}_{a-1}^* - \mathbf{t}_a^*\mathbf{p}_a^{*\prime} \tag{2.16}$$

$$\mathbf{f}_a^* = \mathbf{y} - \sum_{k=1}^{a} \mathbf{t}_k^* q_{ak}^* \quad . \tag{2.17}$$

The orthogonality-properties mentioned earlier are easily seen to follow from the way the vectors are determined as projections of earlier residuals: For the first algorithm, different $\mathbf{t}_a$ are orthogonal because of (2.8) and the first part of (2.3). In the second, different $\mathbf{p}_a^*$ are orthogonal because of (2.13) and (2.16).

Let again $\mathbf{x}_0 = (x_{01},\dots,x_{0K})'$ be a new set of x-values, and put $\mathbf{e}_0^* = \mathbf{x}_0 - \bar{\mathbf{x}}$. Then, as in (2.10), new scores and residuals for this point are determined in the same way as for the other points. Thus, as in (2.13) and (2.16), for a=1,2,... we let

$$t_{a0}^* = \mathbf{e}_{a-1}^{*\prime}\mathbf{p}_a^*/\mathbf{p}_a^{*\prime}\mathbf{p}_a^* \tag{2.18}$$

$$\mathbf{e}_a^* = \mathbf{e}_{a-1}^* - t_{a0}^*\mathbf{p}_a^* = \mathbf{e}_0^* - \sum_{r=1}^{a} t_{r0}^*\mathbf{p}_r^* \ , \tag{2.19}$$

and predict y_0 in step A by

$$\hat{y}_{A0}^* = \bar{y} + \sum_{a=1}^{A} t_{a0}^*\, q_{Aa}^* \quad . \tag{2.20}$$

The formulas given here can be simplified in various ways; later we will give a very simple formula for the prediction (2.20). But first we will show that the two algorithms are equivalent. In fact the loadings $\mathbf{p}_a^*$ are equal to the weights $\mathbf{w}_a$ from the first algorithm.

Theorem 2.1

With notation as above we have for a=1,2,...

a) $\mathbf{p}_a{}^* = \mathbf{w}_a$,

b) $\{\mathbf{t}_1{}^*,\ldots,\mathbf{t}_a{}^*\}$ *span the same space in* $\mathcal{R}^N$ *as* $\{\mathbf{t}_1,\ldots,\mathbf{t}_a\}$,

c) $\mathbf{f}_a{}^* = \mathbf{f}_a$,

d) $\widehat{y}_{a0}{}^* = \widehat{y}_{a0}$.

Proof

Let $\mathcal{P}_{ta}{}^*$ and $\mathcal{P}_{ta}$ be the projections upon the two spaces described in b), and let $\mathbf{I}$ be the identity matrix of size N by N. Then using the orthogonality of $\mathbf{t}_1$, $\mathbf{t}_2$,..., we find from (2.3), (2.8) and (2.9)

$$\mathbf{E}_a = (\mathbf{I} - \mathcal{P}_{ta})\mathbf{X}$$
$$\mathbf{f}_a = (\mathbf{I} - \mathcal{P}_{ta})\mathbf{y} \quad ,$$

and (2.15) - (2.17) give

$$\mathbf{E}_a{}^* = \mathbf{X} - \sum_{r=1}^{a} \mathbf{t}_r{}^* \, \mathbf{p}_r{}^{*\prime}$$
$$\mathbf{f}_a{}^* = (\mathbf{I} - \mathcal{P}_{ta}{}^*)\mathbf{y}$$

We prove a) and b) simultaneously by induction in the parameter a. They are trivial for a=1. Assume them to be true up to a-1. Then the formulas above show that

$$\mathbf{E}_{a-1}'\,\mathbf{f}_{a-1} = \mathbf{E}_{a-1}^{*'}\,\mathbf{f}_{a-1}^{*} = \mathbf{X}'(\mathbf{I} - \mathcal{P}_{t,a-1})\mathbf{y} \ ,$$

giving a). But the two sets of scores are given by

$$\mathbf{t}_a = \mathbf{E}_{a-1}\mathbf{w}_a = (\mathbf{I}-\mathcal{P}_{t,a-1})\mathbf{X}\mathbf{w}_a = \mathbf{X}\mathbf{w}_a - \mathcal{P}_{t,a-1}\mathbf{X}\mathbf{w}_a$$

$$\mathbf{t}_a^* \ \alpha \ \mathbf{E}_{a-1}^*\mathbf{p}_a^* = (\mathbf{X} - \sum_{r=1}^{a-1} \mathbf{t}_r^*\mathbf{p}_r^{*'}\ )\ \mathbf{p}_a^* = \mathbf{X}\mathbf{p}_a^* = \mathbf{X}\mathbf{w}_a \ .$$

Since, by the induction hypotheses, the vector subtracted from $\mathbf{X}\mathbf{w}_a$ in $\mathbf{t}_a$ belongs to the span of $\mathbf{t}_1^*,\dots,\mathbf{t}_{a-1}^*$, this proves b).

By the formulas above for $\mathbf{f}_a$ and $\mathbf{f}_a^*$, c) follows from b).

Taking the two sets of scores as columns in matrices, it follows from b) that $\mathbf{T}_a^* = \mathbf{T}_a\mathbf{D}$ for some non-singular matrix $\mathbf{D}$. (An explicit formula for $\mathbf{D}$ will be given in Section 3.3 below.) From the way the scores t_{r0}^* and t_{r0} are constructed, one must then also have $[t_{10}^*,\dots,t_{a0}^*] = [t_{10},\dots,t_{a0}]\ \mathbf{D}$. Thus we have a simple linear change of variable between the two sets of scores, and since prediction in both cases is based upon multiple linear regression on the scores, d) follows.

Essentially the same proof works for the corresponding algorithms in the case (PLS2) with several $\mathbf{y}$-variables; this is shown in an unpublished note by Chris Rogers.

3. THE VARIABLES GIVEN BY THE ALGORITHMS

3.1. The weight factors

A recurrence relation for the weights $\mathbf{w}_a$ is most easily found using $\mathbf{p}_a^* = \mathbf{w}_a$ in the second algorithm. In the proof of

Theorem 2.1, we showed that $\mathbf{t}_a*$ is proportional to $\mathbf{X}\mathbf{w}_a$, or properly normed (cf. (2.13)): $\mathbf{t}_a* = \mathbf{X}\mathbf{w}_a/\mathbf{w}_a'\mathbf{w}_a$. Thus $\mathbf{T}_a* = (\mathbf{t}_1*,...,\mathbf{t}_a*) = \mathbf{X}\mathbf{W}_a\mathbf{C}_a$, where we define $\mathbf{W}_a = (\mathbf{w}_1,...,\mathbf{w}_a)$, and let $\mathbf{C}_a$ be the normalizing matrix $diag(\|\mathbf{w}_1\|^{-2},..., \|\mathbf{w}_a\|^{-2})$.

Then (2.15) gives

$$\mathbf{q}_a* = \mathbf{C}_a^{-1}(\mathbf{W}_a'\mathbf{S}\mathbf{W}_a)^{-1}\mathbf{W}_a'\mathbf{s} \ , \tag{3.1}$$

where the following fundamental matrices are introduced

$$\mathbf{S} = \mathbf{X}'\mathbf{X} \ , \ \mathbf{s} = \mathbf{X}'\mathbf{y} \ . \tag{3.2}$$

Now from (2.12) for $\mathbf{w}_{a+1} = \mathbf{p}_{a+1}*$ and the relations in the proof of Theorem 2.1 we have

$$\mathbf{w}_{a+1} = \mathbf{E}_a*'\mathbf{f}_a* = \mathbf{X}'(\mathbf{I} - \mathcal{P}_{ta}*)\mathbf{y} \ .$$

Using again $\mathbf{T}_a* = \mathbf{X}\mathbf{W}_a\mathbf{C}_a$ in $\mathcal{P}_{ta}* = \mathbf{T}_a*(\mathbf{T}_a*'\mathbf{T}_a*)^{-1}\mathbf{T}_a*'$, we get

$$\mathbf{w}_{a+1} = \mathbf{s} - \mathbf{S}\mathbf{W}_a(\mathbf{W}_a'\mathbf{S}\mathbf{W}_a)^{-1}\mathbf{W}_a'\mathbf{s} \ . \tag{3.3}$$

This recurrence relation will be the basis for much of the following, and it will be written in the form

$$\mathbf{w}_{a+1} = (\mathbf{I} - \mathbf{S}\mathbf{H}_a)\mathbf{s} \ , \tag{3.4}$$

where now $\mathbf{I}$ has dimension K by K, and

$$\mathbf{H}_a = \mathbf{W}_a(\mathbf{W}_a'\mathbf{S}\mathbf{W}_a)^{-1}\mathbf{W}_a' \ . \tag{3.5}$$

Note that any matrix of the form $\mathbf{W}_a\mathbf{C}$ with $\mathbf{C}$ nonsingular can replace $\mathbf{W}_a$ in the definition of $\mathbf{H}_a$. The important thing is that the columns form a basis for the space

$\mathcal{S}_a$ spanned by $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_a$. A particular basis is of special interest.

Proposition 3.1

As long as $\mathbf{w}_A$ *is nonzero, an alternative basis for* $\mathcal{S}_A$ *is given by the vectors*

$$\mathbf{s}, \mathbf{S}\mathbf{s}, ..., \mathbf{S}^{A-1}\mathbf{s} \ .$$

Proof

We have $\mathbf{w}_1 = \mathbf{s}$. Use induction in a together with (3.3) to show that each $\mathbf{w}_a$ can be written as a linear combination of $\mathbf{s}, \mathbf{S}\mathbf{s}, ..., \mathbf{S}^{a-1}\mathbf{s}$. (Note that the last term in (3.3) is a linear combination of the columns of $\mathbf{S}\mathbf{W}_a$.). But since $\mathbf{w}_1, ..., \mathbf{w}_A$ form a basis for $\mathcal{S}_A$, and all thus are linear function of A new vectors, the latter must be linearly independent, and the Proposition follows.

The alternative basis described in Proposition 3.1 is central to the connection between PLS and the conjugate gradient method from numerical analysis (see also Wold et al., 1984 and Manne, 1987; this relationship has also recently been commented upon by Schweder, 1987). In the numerical litterature; $\mathbf{s}, \mathbf{S}\mathbf{s}, ..., \mathbf{S}^{a-1}\mathbf{s}, ...$ is called a Krylov sequence. The dimension of the space spanned by this sequence will be the maximal number of factors that the PLS-algorithm can give. This maximal space is also spanned by the <u>relevant factors</u> of $\mathbf{S} = \mathbf{X}'\mathbf{X}$, i.e., the eigenvectors of $\mathbf{S}$ with non-zero components along $\mathbf{s} = \mathbf{X}'\mathbf{y}$, one for each eigenvalue in case these should be degenerate. These results and the corresponding population model will be further discussed elsewhere.

3.2 Prediction equation

From (2.20) and Theorem 2.1d we get

$$\hat{y}_{A0} = \bar{y} + (\mathbf{t}_{10}^*,\ldots,\mathbf{t}_{A0}^*)\,\tilde{q}_A^* \ .$$

Similarly to the relation $\mathbf{T}_a^* = \mathbf{X}\mathbf{W}_a\mathbf{C}_a$, we find

$$(t_{10}^*,\ldots,t_{A0}^*) = (\mathbf{x}_0 - \bar{\mathbf{x}})'\mathbf{W}_a\mathbf{C}_a \ .$$

Using (3.1), we then get one of our main results:

Theorem 3.1

For both PLS-algorithms, the prediction at step A *is given by*

$$\hat{y}_{A0} = \bar{y} + (\mathbf{x}_0 - \bar{\mathbf{x}})'\mathbf{b}_A \tag{3.6}$$

where

$$\mathbf{b}_A = \mathbf{H}_A\mathbf{s}, \tag{3.7}$$

$\mathbf{s} = \mathbf{X}'\mathbf{y}$ *and* $\mathbf{H}_A$ *is given by (3.5) with* $\mathbf{S} = \mathbf{X}'\mathbf{X}$.

Note again, that from Proposition 3.1, we can replace $\mathbf{W}_A$ in the definition of $\mathbf{H}_A$ by

$$\mathbf{V}_A = (\mathbf{s}, \mathbf{S}\mathbf{s}, \ldots, \mathbf{S}^{A-1}\mathbf{s}) \tag{3.8}$$

so

$$\mathbf{H}_A = \mathbf{V}_A(\mathbf{V}_A'\mathbf{S}\mathbf{V}_A)^{-1}\mathbf{V}_A' \ . \tag{3.9}$$

Also note that one can multiply $\mathbf{S}$ and $\mathbf{s}$ by arbitrary scalars and still get the same space spanned by the columns of

$\mathbf{V}_A$. Hence if we use the same scalar multiplying both $\mathbf{S}$ and $\mathbf{s}$, we get the same regression vector $\mathbf{b}_A = \mathbf{H}_A\mathbf{s}$ from (3.8) and (3.9). Up to now we have taken $\mathbf{S} = \mathbf{X}'\mathbf{X}$ and $\mathbf{s} = \mathbf{X}'\mathbf{y}$. However, if one wants to treat the prediction in the framework of a population model, it is more meaningful to use covariance-matrices and -vectors, therefore, dividing both $\mathbf{S}$ and $\mathbf{s}$ by N or N-1.

The formulae (3.6) - (3.7) with $\mathbf{H}_A$ given by (3.5) are true even if arbitrary orthogonal weight vectors $\mathbf{w}_1,\ldots,\mathbf{w}_A$ should be chosen in the PLS-algorithm (2.3) - (2.11). This can be seen by inspecting the above proof (taking $\mathbf{p}_a{}^* = \mathbf{w}_a$ as a definition in Theorem 2.1). However, the alternative formula (3.9) for H_A is crucially dependent upon the special weights chosen in the PLS-algorithm.

When the number N of units is about 3 times the number of variables or greater, our explicit formulae can be shown to give faster computation than the original algorithms. Much more important, however, is that the formulae make it possible to give explicit interpretations of the prediction resulting from the algorithms.

3.3 An alternative formula for $\mathbf{b}_A$.

Consider first the matrix $\mathbf{D}$ in the relationship $\mathbf{T}_A{}^* = \mathbf{T}_A\mathbf{D}$ mentioned in section 2.3 above. Let $\mathbf{D} = \{d_{ij}\}$. Then, since the vectors $\mathbf{t}_i$ are orthogonal, we get

$$d_{ij} = (\mathbf{t}_i'\mathbf{t}_i)^{-1}\mathbf{t}_i'\mathbf{t}_j{}^* = (\mathbf{t}_i'\mathbf{t}_i)^{-1}\mathbf{t}_i'\mathbf{X}\mathbf{w}_j/\mathbf{w}_j'\mathbf{w}_j = \mathbf{p}_i'\mathbf{w}_j/\mathbf{w}_j'\mathbf{w}_j\,,$$

so $\mathbf{D} = \mathbf{P}_A'\mathbf{W}_A\mathbf{C}_A$ in the notation defined before (3.1). This gives

$$\mathbf{X}\mathbf{W}_A\mathbf{C}_A = \mathbf{T}_A{}^* = \mathbf{T}_A\mathbf{P}_A'\mathbf{W}_A\mathbf{C}_A\,,$$

and hence $\mathbf{T}_A = \mathbf{X}\mathbf{W}_A(\mathbf{P}_A'\mathbf{W}_A)^{-1}$. Similarly for new observations

$\mathbf{x}_0$, we get

$$(t_{10},\dots,t_{A0}) = (\mathbf{x}_0 - \overline{\mathbf{x}})'\, \mathbf{W}_A(\mathbf{P}_A'\mathbf{W}_A)^{-1} ,$$

and from (2.11) we have

$$\hat{y}_{A0} = \overline{y} + (\mathbf{x}_0 - \overline{\mathbf{x}})'\mathbf{b}_A$$

with

$$\mathbf{b}_A = \mathbf{W}_A(\mathbf{P}_A'\mathbf{W}_A)^{-1}\mathbf{q}_A \qquad [\ \mathbf{q}_A = (q_1,\dots,q_A)'\ ]. \tag{3.10}$$

This formula was suggested by Martens and Næs (1987) (the first version was from 1981), and it can be proved in several ways. The above proof was suggested by Chris Rogers.

Several authors have pointed out that the matrix $\mathbf{P}_A'\mathbf{W}_A$ is triangular with all diagonal-elements equal to 1. Manne (1987) proves that the matrix also is bidiagonal, and gives the resulting simple inversion formula.

Note that even though (3.10) looks slightly simpler than (3.7)-(3.9), the latter formulas are expressed directly in terms of covariances of the observed variables and hence allows simpler interpretations.

4. INTERPRETATION OF THE PREDICTION EQUATION

4.1 Interpretation via a transformation

Under the usual regression model $\mathbf{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, the distribution of $\mathbf{s} = \mathbf{X}'\mathbf{y}$ will be $\mathrm{N}(\mathbf{S}\boldsymbol{\beta}, \sigma^2\mathbf{S})$, where again $\mathbf{S} = \mathbf{X}'\mathbf{X}$. Assuming $\mathbf{S}$ is invertible, let $\sqrt{\mathbf{S}}$ be the positive definite, symmetric square root of it, and define $\mathbf{u} = (\sqrt{\mathbf{S}})^{-1}\mathbf{s}$. Then $\mathbf{u}$ will be K-dimensional,

$$\mathbf{u} \sim \mathrm{N}(\gamma, \sigma^2\,\mathbf{I}), \quad \gamma = \sqrt{\mathbf{S}}\boldsymbol{\beta} \ . \tag{4.1}$$

If we transform the estimated regression coefficient vector $\mathbf{b}_A$ similarly into $\mathbf{g}_A = \sqrt{\mathbf{S}}\,\mathbf{b}_A$, Theorem 3.1 together with (3.9) gives

Proposition 4.1

The PLS regression vector at step A *is given by*

$$\mathbf{b}_A = (\sqrt{\mathbf{S}})^{-1}\mathbf{g}_A, \quad \mathbf{g}_A = \mathbf{G}_A\mathbf{u}\ , \tag{4.2}$$

where $\mathbf{G}_A = \sqrt{\mathbf{S}}\,\mathbf{H}_A\sqrt{\mathbf{S}}$ *is the projection upon the space spanned by the vectors* $\{\mathbf{Su}, \mathbf{S}^2\mathbf{u}, ..., \mathbf{S}^A\mathbf{u}\}$.

Note that this Proposition is true irrespectibly of which distributional assumptions we make on the variables. If the distribution is given by (4.1), then of course the natural estimator of γ is $\hat{\gamma} = \mathbf{u}$. What PLS does , is instead to project $\mathbf{u}$ upon an A-dimensional space which again depends upon $\mathbf{u}$. This may seem strange, but perhaps it is easier to accept such non-linear estimators for those knowing the James-Stein regression estimator and similar shrinkage estimators (see Copas (1983) for a discussion.) In the present context, a partial motivation can be made as follows: The data are not only given by $\mathbf{u}$, but also by $\mathbf{S}$, which gives the covariance structure of the original $\mathbf{x}$-variables. If these variables are highly collinear, then transforming back to $\hat{\beta} = (\sqrt{\mathbf{S}})^{-1}\mathbf{u}$ gives an unstable estimator. This is the usual collinearity-problem of the multiple regression estimator. No such problem arises, however, if γ is estimated by $\mathbf{g}_A$ in (4.2) with small or moderate A, since the estimator then is a linear combination of $\mathbf{Su}, \mathbf{S}^2\mathbf{u}, ..., \mathbf{S}^A\mathbf{u}$, cancelling the factor $(\sqrt{\mathbf{S}})^{-1}$.

4.2 Spectral representation and connection to PCR

Looking at the first PLS-algorithm, we see that it stops in a natural way at the step A = M, where M is the first integer where $\mathbf{w}_{A+1} = \mathbf{E}_A'\mathbf{f}_A = \mathbf{0}$. For $A \leq M$ the matrix $\mathbf{W}_A'\mathbf{S}\mathbf{W}_A$ will have full rank and has a spectral decomposition

$$\mathbf{W}_A'\mathbf{S}\mathbf{W}_A = \sum_{a=1}^{A} \varphi_{aA}\mathbf{v}_{aA}\mathbf{v}_{aA}' \tag{4.3}$$

where $\mathbf{v}_{1A}, ..., \mathbf{v}_{AA}$ are orthonormal eigenvectors with eigenvalues $\varphi_{1A}, ..., \varphi_{AA}$, all nonzero. Using this and (3.5) in (3.7), we find

$$\mathbf{b}_A = \sum_{a=1}^{A} (\varphi_{aA})^{-1} \ \mathbf{W}_A\mathbf{v}_{aA}\mathbf{v}_{aA}'\mathbf{W}_A'\mathbf{s} \ , \tag{4.4}$$

this being equivalent to formula (3.3) in Næs and Martens (1985).

The formula (4.4) suggests a close analogy with the predictor given by principal component regression. This has the same form as the PLS-predictor (3.12), but with $\mathbf{b}_A$ replaced by

$$\mathbf{b}_{A,PCR} = \sum_{a=1}^{A} (\lambda_a)^{-1}\mathbf{z}_a\mathbf{z}_a'\mathbf{s} \ ,$$

where

$$\mathbf{S} = \sum_{k=1}^{K} \lambda_k\mathbf{z}_k\mathbf{z}_k' \tag{4.5}$$

is the spectral decomposition of $\mathbf{S} = \mathbf{X}'\mathbf{X}$ and $\mathbf{z}_1,...,\mathbf{z}_A$ are the selected principal components (usually with large or moderate eigenvalues.)

This analogy becomes an identity if we let A take its maximal value M in (4.4). One can show that M is the number of different eigenvalues λ_k such that $\mathbf{z}_k$'s $\neq 0$ for at least one $\mathbf{z}_k$ corresponding to this eigenvalue.

Theorem 4.1

If A = M, *the PLS-predictor is* $\hat{y}_M = \bar{y} + (\mathbf{x}_0 - \bar{\mathbf{x}})'\mathbf{b}_M$ *with*

$$\mathbf{b}_M = \sum_{a=1}^{M} (\lambda_a)^{-1} \mathbf{z}_a \mathbf{z}_a'\mathbf{s} \tag{4.6}$$

for a suitable choice of and ordering of the eigenvectors $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K$.

The proof of this Theorem will be given elsewhere.

The interpretation of (4.6) is that PLS with A=M gives the principal component solution with all nonzero eigenvalues, and, that is important: it is found with the minimal number of terms. All terms with $\mathbf{z}_a$'s = 0 are automatically neglected, and for multiple eigenvalues, the algorithm picks just the right eigenvectors.

Perhaps more important than the precise statement of Theorem 4.1, is what it seems to imply when the conditions are approximately fulfilled. Remember that both the λ_k and the $\mathbf{z}_k$ depend on the data through $\mathbf{S} = \mathbf{X}'\mathbf{X}$, so all the variables in (4.6) must be thought of as having some random noise attached to it. Imagine that the complete spectral decomposition of $\mathbf{S}$ has several terms with λ_k and/or $\mathbf{z}_k$'s "nearly" zero. Then, since the considerations of the previous section seem to imply certain continuity properties of $\mathbf{b}_A$, the formula (4.6) suggests that $\mathbf{b}_A$ will be "nearly" equal to the PCR-predictor after some terms. Furthermore, if some sets of eigenvalues are "nearly" coinciding, it seems likely that the PLS-predictor will stabilize after fewer terms than the PCR-predictor does. This discussion can be made more precise, but the matter will not be pursued further here. At least, the results are consistent with the simulations showing that PLS tends to achieve its minimal MSE after fewer terms than PCR. However, the discussion also indicates that PLS and PCR, each with an optimal number of factors, not only should give similar results in terms of MSE, but they should also give similar predictions. This also seems to be the case on practice.

Note that the reduction in PLS in the number of terms because of multiple eigenvalues can be considerable. An extreme case is when **S** is proportional to the identity matrix. Then PLS gives the least squares predictor in the first term, while a principal component representation of the forms (4.5) will need A = K in most cases. Thus reduction in the number of terms in PLS has little to do with collinearity, it is primarily connected to the equality or near equality of the eigenvalues of **S**.

5. DISCUSSION

Simulations have indicated that PLS often reaches its minimal mean square error after fewer factors than PCR does. Formulas for the mean square error of prediction of PLS can be studied to a certain extent, especially if one integrates over future observations under different weight functions as in Gunst and Mason (1979). Such calculations seem to confirm that the mean square error of PLS usually is lower than that of PCR after the first few factors. The more interesting problem, but one that probably is impossible to solve analytically, would be to try to compare the minimal mean square errors for the two methods. Simulations may also be difficult here, since the difference seems to be very small.

If more was known about the prediction error of PLS, this quantity could perhaps be estimated in concrete cases, thus making approximate confidence statements available.

Another open problem in connection with the PLS-algorithm is to find a test for the number of factors to include in the algorithm. Today the number of factors is usually determined by crossvalidation.

As a general attitude towards the PLS-method one can of course question the value of introducing another biased regression method as competitor to PCR and RR. What PLS offers in comparison with these methods, is first some gain in computation time, which may not be too important with the use of modern computers (though certainly of value when microcomputers are being used). Next it seems to get its optimal prediction with fewer factors than PCR, which, in addition to minimize computing, may help interpretation in some cases. Thirdly, unlike PCR, the PLS method gives a unique way of choosing which factor to include next. Finally, the PLS-method does not only look at the conditional distribution of **y** given **x**, but treats both **x** and **y** as random variables, connected through the latent variables **t**.

The main disadvantages of PLS are still its lack of known distributional properties, together with the fact that the method has been derived in a rather ad hoc way, not from any welldefined optimization principle.

ACKNOWLEDGEMENTS

Tormod Næs and Harald Martens have been the main inspirators behind the present paper. Also, discussions with Kim Espensen, Graham Wetherhill, Chris Theobald, Rolf Manne, Jan Søreng, Ragnar Winther and Emlyn R. Williams are appreciated, and the Department of Statistics of the University of Edinburgh is acknowledged for its hospitality during the time when part of this work was being done.

BIBLIOGRAPHY

Aastveit, A.H. and Martens, H. (1986). ANOVA interactions interpreted by Partial Least Squares regression. Biometrics, 42 829-844.

Brown, P.J. (1982). Multivariate calibration. J.R. Statist. Soc. B, 44 287-321.

Copas, J.B. (1983). Regression, prediction and shrinkage. J.R. Statist. Soc. B, 45 311-354.

Fearn, T. (1983). Misuse of ridge regression in the calibration of near infrared reflectance instrument. Applied Statistics, 32 73-79.

Geladi, P. and Kowalski, B.R. (1986). Partial least-squares regression: A tutorial. Analytica Chimica Acta, 185 1-17.

Gunst, R.F. and Mason, R.L. (1979). Some considerations in the evaluation of alternate prediction equations. Technometrics, 21 55-63.

Hoerl, A.E., Kennard, R.W. and Hoerl, R.W. (1985). Practical use of ridge regression: A challenge met. Applied Statistics, 34 114-120.

Höskuldson, A. (1987). PLS methods. Submitted to J. Chemometrics.

Joliffe, I.T. (1982). A note on the use of principal components in regression. Applied Statistics, 31, 300-303.

Kruskal, J.B. (1978). Factor analysis and principal components I. Bilinear methods. In: International Encyclopedia of Statistics. Collier Macmillan Publishers, London.

Krutchkoff, R.G. (1967). Classical and inverse regression methods of calibration. Technometrics, 9 425-439.

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

Manne, R. (1987). Analysis of two partial-least-squares algorithms for multivariate calibration. Chemometr. & Int. Lab. Syst., 2 187-197.

Marquardt, D.W. (1980). Comment to Smith and Campbell (1980).

Martens, H. (1985). Multivariate Calibration. Dr. techn. Thesis. Technical University of Norway, Trondheim.

Martens, H. (1987). Multivariate calibration: Combining harmonies from an orchestra of instruments into reliable predictors of chemical compositions. 46th Session of the ISI, Invited paper.

Martens, H. and Jensen, S.A. (1983). Partial least squares regression: A new two-stage NIR calibration method. In: Progress in Cereal Chemistry and Technology 5a (Proceedings, 7th world Cereal and Bread Congress, Prague, June 1982, J. Holas, J. Kratochvit, eds.) Elsevier Publ., Amsterdam, 607-647.

Martens, M., Lea, P. and Martens, H. (1983a). Predicting human response to food quality by analytical measurements: the PLS regression method. Proc. Nordic Symposium on Applied Statistics (O.H.J. Christie, ed.), June 1983. Stokkand Forlag Publ., Stavanger, Norway.

Martens, H. and Næs, T.(1987). Multivariate calibration by data compression. In: Near-Infraread Technology for the Agricultural and Food Industries. (P. Williams and K. Norris, eds.) Amer. Ass. of Cereal Chemists. St. Paul.

Martens, H., Vangen, O. and Sandberg, E. (1983b). Multivariate calibration of an X-ray computer tomograph by smoothed PLS regression. Proc. Nordic Symposium on Applied Statistics (O.H.J. Christie, ed.), June 1983, Stokkand Forlag Publ., Stavanger, Norway.

Mason, R.L. and Gunst, R.F. (1985). Selecting principal components in regression. Statistics & Probability Letters, 3 299-301.

Næs, T. and Martens, H. (1985). Comparison of prediction methods for multicollinear data. Commun. Statist. - Simula. Computa., 14 545-576.

Næs, T. (1985). Multivariate calibration when the error covariance matrix is structed. Technometrics, 27 301-311.

Næs, T., Irgens, C. and Martens, H. (1986). Comparison of linear statistical methods for calibration of NIR instruments. Applied Statistics, 35 195-206.

Schweder, T. (1987). Canonical regression. Report no 804. Norwegian Computing Center, Oslo.

Smith, G. and Campbell, F. (1980). A critique of some ridge regression methods. J. Amer. Statist. Ass., 75 74-103.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. J.Roy. Statist. Soc. 36 111-133.

Sundberg, R. (1985). When is the inverse regression estimator MSE-superior to the standard regression estimator in multivariate controlled calibration situations? Statistics & Probability Letters, 3 75-79.

Sundberg, R. and Brown, P. (1985). Multivariate calibration with more variables than observations. Research report no. 139, Institute of Actuarial Mathematics and Mathematical Statistics, University of Stockholm.

Wold, H. and Jøreskog, K.G., eds. (1982). Systems under indirect observation. Causality - structure - prediction. Contributions to Economic Analysis, 139, parts I and II. North-Holland Publ. Co. Amsterdam.

Wold, S., Martens, H. and Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. Proc. Conf. Matrix Pencils (A. Ruhe, B. Kågstrøm, eds.), March 1982, Lecture Notes in Mathematics, Springer Verlag, Heidelberg, 286-293.

Wold, S. (1974). A theoretical foundation of extra thermodynamic relationships (linear free energy relationships). Chimica Scripta, 5 97-106.

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

Wold, S., Wold, H., Dunn, W.J. and Ruhe A., (1984) The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. Siam J. Sci. Stat. Comput., 5 735-743.

Received by Editorial Board member December, 1986; Revised February, 1988.

Recommended by Anant M. Kshirsagar, University of Michigan, Ann Arbor, Michigan.

Refereed by Robert L. Mason, Southwest Research Institute, San Antonio, Texas.