Perspectives in Drug Discovery and Design, 1 (1993) 269–278
ESCOM

PD3 018

Partial Least Squares (PLS): Its strengths and limitations

Richard D. Cramer III

Tripos Associates, 1699 S. Hanley Road, St. Louis, MO 63144, U.S.A.

Received 31 August 1993
Accepted 6 September 1993

Key words: Partial Least Squares; PLS; CoMFA; QSAR; Chance correlation; Multiple regression

SUMMARY

For structure–activity correlation, Partial Least Squares (PLS) has many advantages over regression, including the ability to robustly handle more descriptor variables than compounds, nonorthogonal descriptors and multiple biological results, while providing more predictive accuracy and a much lower risk of chance correlation. The major limitations are a higher risk of overlooking 'real' correlations and sensitivity to the relative scaling of the descriptor variables.

INTRODUCTION

Because most sets of biological observations, particularly those produced by testing different chemical structures in the same biological system, cannot be adequately described by existing theory, researchers often seek semiempirical models in which the changes in observed values are predicted as a mathematical function of properties which are better understood.

Partial Least Squares (PLS) is an important new technique, introduced by Hermann [1] and Svante Wold [2,3], for producing a linear equation to describe or predict differences in the values of one or more properties from differences in the values of other properties. In structure–activity studies, such a linear equation is usually called a Quantitative Structure–Activity Relationship or QSAR. The described or predicted properties are called the 'dependent variables' or, in the PLS literature, the 'Y-block'. The describing or predicting properties are called the 'independent variables', or the 'X-block'. Partial Least Squares can be regarded as a major generalization of the more familiar technique of multiple regression (MR), since for identical problem formulations PLS and MR produce identical answers. Most of the initial applications of PLS have been in analytical chemistry, particularly food chemistry, but the technique has had a substantial impact on QSAR practice within the last few years.

COMPARISON OF PARTIAL LEAST SQUARES WITH MULTIPLE REGRESSION

The major advantages PLS offers over MR in structure–activity studies are:

(1) The ability to produce useful, robust equations even when the number of 'independent

variables', or coefficients to be evaluated, vastly exceeds the number of experimental observations. The popular new 3D QSAR technique Comparative Molecular Field Analysis (CoMFA) [4], which explains differences among a small number of biological test results in terms of differences in the fields exerted by the tested molecules at thousands of points in space, depends upon this property of PLS for its success.

(2) Predictions from PLS-derived models tend to be more accurate than those from MR-derived models [5] (provided the problem setups are formulated optimally for each method).

(3) PLS models are much more stable when the sets of independent variable values are correlated rather than orthogonal, the most common situation in structure–activity studies.

(4) A PLS study can simultaneously derive models for more than one dependent variable, for example for results from multiple receptor assays or against multiple microorganisms.

The differences in behavior between MR and PLS result from their different strategies for identifying a linear relationship. Put as simply and qualitatively as possible, MR treats independent variables as independent entities, scaling and offsetting each variable separately to obtain the best overall relationship with the dependent variable. In contrast, PLS considers all the independent variables together, as a 'block'. In an iterative process, PLS repeatedly transforms both 'dependent' and 'independent' variable blocks so that their commonality is maximal (see Appendix for the PLS algorithm in one widely used program). Indeed, the more familiar method of principal components (simple factor) analysis (PCA) proceeds identically [6], except that the PLS transformation objective is maximum overlap between the variable blocks, as opposed to the maximum within-block coverage desired in PCA, and the numerical method NIPALS [7] usually used in PLS differs from the numerical method of matrix inversion usual in PCA. Both in PLS and PCA, each iteration of the method extracts a new 'component'. When the number of components extracted by PLS equals the number of columns used in MR, MR and PLS produce the same QSAR from the same data.

However, PLS and MR also differ somewhat in the philosophy underlying their use, which becomes important operationally in their respective criteria for 'best relationship'. Of course, in QSAR applications both methods seek primarily an equation which can predict the biological properties of molecules not used to derive the model. However, establishing a 'best relationship' criterion a priori is chancy, because increasing the descriptors in a model will always improve its fit to the existing data but often degrade its predictive accuracy for new data.

The 'least squares' criterion of MR originates with Gauss, who showed that the best fit among a set of experimental observations (planetary movements), whose relative differences were caused only by 'independent and normally distributed' errors of measurement, to a single theoretically exact predictor (Newton's laws), would minimize the sum of squared differences between predicted and observed values. To make the chancy decision whether some additional independent variable improves an MR model, usually Fisher's F-ratios with and without the variable are compared. (The F-ratio is the ratio of model-explained to model-unexplained variance in the dependent variable, each weighted by the number of degrees of freedom.)

Such situations as Gauss's, where there is a reliable theory linking one set of observations to another, are called 'hard models' in the PLS literature. In contrast, PLS is recommended for 'soft models', those describing many sets of observations containing much correlated error. In such cases, the existence of any relationship, far from being theoretically proven, is in fact the desired result of the investigation. Here the PLS literature advocates that the

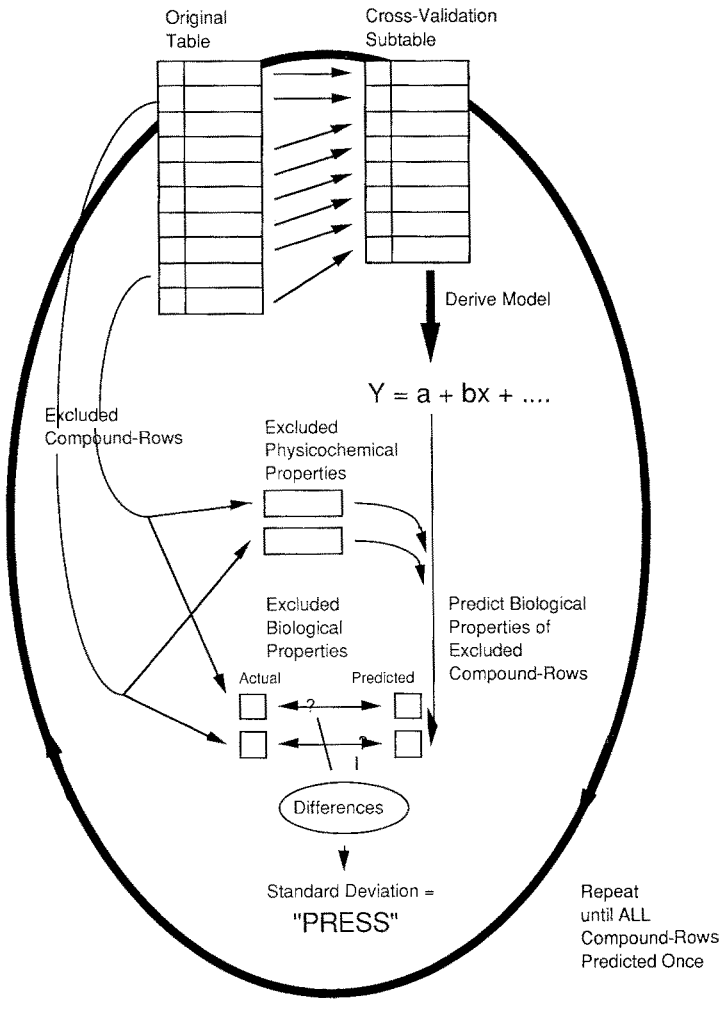

Fig. 1. Schematic representation of the cross-validation process.

investigator search for 'latent variables', factors hidden within the actual observed independent variables, which are hopefully causative or at least correlative with similar latent variables within the dependent variables. In principle, each additional PLS component identifies another such latent relationship and in practice improves the fit between X- and Y-blocks. To address the chancy question, how many components are then 'best', PLS takes advantage of cross-validation, (aka jack-knifing, leave-1-out, leave-n-out). Cross-validation is a recent statistical development, general in applicability, that substitutes today's speed of computation for theoretical assumptions about data distributions. Assuming that the optimal QSAR is the one which best predicts, in cross-validation one pretends that one or more randomly chosen objects are unknown. The entire QSAR is rederived with those objects completely omitted and then used to predict the dependent values of the omitted objects. The cross-validation cycle is repeated until all objects have been omitted, so all dependent values have been 'predicted', the errors of prediction being accumulated, as shown in Fig. 1. Using cross-validation, the 'optimal number of

components' in a PLS analysis is that for which the error of prediction in a cross-validation test is least. A final PLS model is then derived, using all the data and the 'optimal number of components'.

Some of these rather abstract concepts about data analysis may be clarified somewhat by a concrete example. Suppose an investigator is empirically searching for the best lubricant by simple measurements of various properties of a set of unlabeled organic liquid samples. He might soon decide that apparent viscosity, high boiling point, low melting point, high refractivity of light, and low odor intensity (low volatility), solubility in gasoline, and perhaps high color intensity are promising predictors. If a PLS 'soft model' were formulated from a combination of these data, someone who knew the chemical structures of the samples might identify its latent variables [8] with 'molecular size', 'molecular stickiness' and perhaps 'molecular flexibility'.

This example can be extended to crudely illustrate the virtues of cross-validation as a 'best model' sector. Consider also the possibility that among the unlabeled bottles there is one unlabeled can, which happens to contain high-quality motor oil. The decision model that would probably then be optimal according to an F-test — look for a good lubricant in a metal container — would work out badly if the next set of candidates included canned soft drinks. However, a cross-validation test, used with either MR or PLS [9], would at least identify a suspicious data distribution, by the discontinuity occurring whenever the properties of the canned liquid were omitted from the model development.

EVALUATING THE PLS MODEL

The most widely used metric for evaluating a QSAR, whether from MR or PLS, is its r^2 value, defined as:

$$r^2 = (SD_{original} - SD_{remaining\ error})/SD_{original}$$

where $SD_{original}$ is the sum of the squared deviations of the original dependent values from their mean and $SD_{remaining\ error}$ is the sum of squared differences between original and 'predicted dependent values' after the QSAR has been derived. Qualitatively, r^2 indicates the 'predictive power' of the QSAR, because an r^2 of 0.0, resulting when $SD_{original}$ equals $SD_{remaining\ error}$, implies that model predictions have an average error no better than uselessly 'predicting' any unknown dependent value to be the average of the known values, while an r^2 of 1.0, resulting from an $SD_{remaining\ error}$ equal to 0.0, implies that the average model-prediction errors are negligible compared to $SD_{original}$. (The only virtue in reporting r values rather than r^2, sometimes seen in the older QSAR literature, seems to be that larger r values give the impression of a better QSAR.)

Because MR tries to fit the data optimally whereas PLS tries to obtain the most predictive relation by using cross-validation, each emphasizes a different definition of 'predicted dependent values' and hence of $SD_{remaining\ error}$ and of r^2. The MR generates each predicted dependent value from the equation which best fits all dependent values together, while PLS generates each predicted dependent value from a cross-validation model, derived while omitting that value. Therefore, the PLS 'cross-validated' r^2 can never be greater than the 'classical' r^2 of MR, and in fact can be negative if the sum of squared prediction errors is larger than $SD_{original}$. However, the cross-validated r^2 seems a much better estimate of the accuracy of 'true' predictions, those for molecules

not considered at any stage of the QSAR development. To avoid confusion, Clementi has recommended that the PLS cross-validated r^2 be renamed q^2 [10].

RISK OF CHANCE CORRELATION WITH PLS

The derivation of a QSAR involves implicit decisions about which kinds of structural change do or do not have an effect on the biological response. Thus, like any other decision based on statistical analysis of data collections, a QSAR might be subject to two kinds of error: omitting structural factors which in fact are related to response (Type I errors); and reporting structural factors which in fact are not related to response (Type II errors). In practice, classical QSAR methodology with stepwise MR has been mainly subject to Type II errors. There are many structural factors which cause any particular biological response, and if enough combinations of independent structural factors are independently compared to a few biological responses, sooner or later the numbers will agree by chance. This possibility was quantified by Topliss and coworkers [11], who applied stepwise MR to sets of random numbers, representing both biological response and structural factors. The probability of 'chance correlation' in these studies proved very high. From this and other work, a 'rule-of-thumb' evolved that in QSAR there should be at least four compounds studied for every independent variable (four times as many rows as columns) for the chance correlation risk to be acceptably low.

This high risk with MR of reporting as true chance correlations among unrelated numbers has clearly hindered the development of new QSAR descriptors. To contribute to molecular discovery, a structure–activity relationship must be formulated early in a research program, when only a few compounds have been studied and there are many different ways of explaining the results. Of course, this is exactly when the risk of chance correlation is highest (whether the SAR is a formal QSAR or an informal 'intuition'). When only a few structural descriptors could be safely considered anyway, there was little incentive to move beyond the 'classic QSAR substituent trio' of lipophilic, electronic and steric effects. The chance correlation limits on descriptors imposed by MR have a delicious irony — the more you know about the structural differences among your tested molecules, the less you can rely on any particular relationship. Too much knowledge about a problem seemed to prevent its useful solution!

Contrast the 'four rows for every column' stricture of classical QSAR with the situation of CoMFA, with its hundreds of columns for every row. The least-squares algorithm of MR cannot even be applied to a CoMFA data table, because there are an astronomical number of combinations among thousands of independent variables that would fit a few dozen dependent variable values perfectly. What confidence can then be placed in the single PLS-generated CoMFA combination, whether cross-validated or not?

To answer this question, several variations of the Topliss experiment have been performed [12,13]. One of these, in which only the biological data values underlying a successful CoMFA were scrambled, either by repeatedly interchanging random pairs of values or by replacing them with random values, offers the most direct reassurance to CoMFA practitioners. For example, with 21 compounds the probability of obtaining a q^2 of 0.25 or higher after the biological data are scrambled is less than 5%.

The more generally revealing experiments, however, were those exactly like Topliss's, in that all the data were random numbers and the numbers of rows and columns were systematically varied.

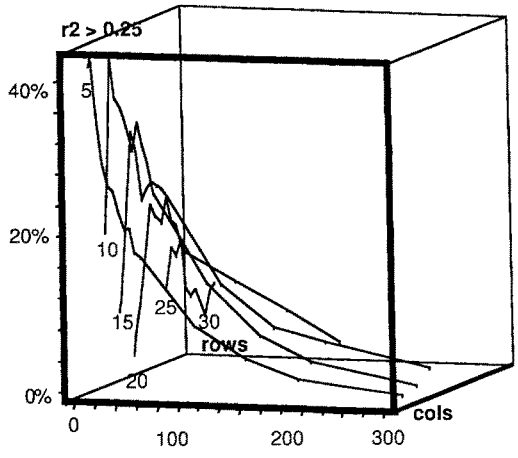

Fig. 2. Frequency of a chance correlation with a q^2 greater than 0.25, as a function of the number of rows and columns containing random data, using PLS.

Here the intuitively very unexpected result shown in Fig. 2 was obtained. The 3D graph plots the frequency of observing a q^2 of 0.25 or higher with PLS in a thousand trials for a particular number of rows and columns. The jagged lines connect experiments involving the same number of rows but different numbers of columns. Each jagged line exhibits a clear maximum, which, within each sequence, occurs when the number of columns is equal to the number of rows. As the number of columns increases beyond the number of rows, the frequency of chance correlation decreases asymptotically to zero.

This result is surprising, because surely the probable number of opportunities for chance correlation steadily increases as the number of columns of random values increases, as verified in the Topliss studies. What these experiments must mean is that PLS can no longer detect correlations involving only a few columns as the number of columns increases. To confirm this inference, random variables were added to tables containing a perfect correlation, and indeed it was found that, as the number of random variables increased, the probability of observing a q^2 of 0.25 or higher decreased indefinitely from 100%. Thus, PLS can clearly fail to discover good correlations involving only a small fraction of the independent variables under consideration. Of course, a stepwise MR will always find small numbers of highly correlated columns.

These empirical results comparing MR and PLS are consistent with general statistical expectations. If PLS yields less Type II errors (chance correlation) than MR for a given set of data, then it is expected that PLS should be more subject to Type I errors (overlooking 'true' correlations). In QSAR practice, however, chance correlation has certainly been a problem whereas overlooking true correlations has not.

SIGNAL-TO-NOISE RATIO IN PLS

From these experiments with PLS on random and correlated numbers, one can imagine that there are two competing factors that determine whether or not a particular correlation is observed by PLS:

(1) The strength of the 'signal', i.e., the proportion of the variance in all the independent variables together which correlates with variance in the dependent variable(s).

(2) The strength of the 'noise', i.e., the proportion of the variance in all the independent variables together which does **not** correlate with variance in the dependent variable(s).

PLS will succeed in finding a relationship with a usefully high q^2 only if this signal/noise ratio is high enough. This behavior is quite unlike experiences with MR, wherein variables that are not explicitly related to the dependent variable have no influence at all on the goodness of the resulting model.

Since q^2 is the primary indication of whether or not a PLS has any useful predictive utility, techniques for raising q^2, necessarily by improving the signal-to-noise ratio within the independent variables, are of great current interest. The GOLPE (Generating Optimal Linear PLS Estimations) technique [10] of Clementi and co-workers has yielded superior values in several 3D QSAR studies, although not always superior 'true predictions'. GOLPE preselects the most relevant independent variables, using a D-optimal design in the variable loadings space (i.e., variables are sought which both initially tend to correlate with the dependent variable(s) and also provide the most independent information about chemical structure). In terms of the signal/noise ratio concept, GOLPE drops variables which are most likely to include noise, i.e., that are unrelated to the dependent variable(s).

One other counterintuitive property of PLS is the fact that results depend on the relative 'scaling' of the independent variables. With MR, if one decides to reexpress a column of distances in Angstroms instead of centimeters, the only effect on the resulting QSAR is that the 'distances' coefficient will be multiplied by 10^{-7}. With PLS, the result will be that no variables other than distances will appear in the (cross-validated) QSAR. This behavior is another manifestation of the sensitivity of PLS to signal/noise ratio. If one multiplies all the values in a column by 10^7, the magnitude of the resulting values, whether signal or noise, will overwhelm any possible influence from any other column.

There is no obvious way to avoid the scaling dependencies of PLS. 'Autoscaling' before analysis, i.e., rescaling all independent variables to the same (unit) variance, does give each variable the same initial opportunity to influence the PLS result. However, autoscaling is itself a scaling choice, which in some situations yields physically implausible results.

Many of the apparent anomalies experienced in applying PLS to the 3D molecular field samples that are the CoMFA descriptors are easily understood, once the fundamental influence of signal-to-noise ratio on PLS results is fully appreciated [14]. Autoscaling works badly within molecular fields, because its effect is to say that 'a receptor atom distant from a ligand atom is just as likely to strongly interact as one touching a ligand atom'. Electrostatic interactions were negligible compared to steric interactions in early CoMFA studies, because the units of steric interaction happened to yield steric fields of variance about three times that of electrostatics. Similarly, efforts to include single 'scalar' variables, such as logP in 3D QSAR with CoMFA fields, to model transport effects, were ineffective because the variance in a single logP column was inconsequential compared with the hundreds or thousands of field variables. These difficulties were overcome by the introduction of 'CoMFA Standard' scaling, a variant of autoscaling which gives each field overall the same variance as each scalar variable, but retains the same relative scaling of individual values within a field. A more subtle anomaly is the failed expectation that the q^2 from PLS on 3D molecular field samples should improve as the field sampling density

is increased. However, because most regions of 3D space have little relevance to the dependent variable, a uniform increase in sampling density increases the 'noise' in the independent variables at least as rapidly as the 'signal', so improvement in q^2 will not often occur.

CONCLUSIONS

Although the dependence of PLS results on the overall signal-to-noise ratio among the independent variables and on the scaling of individual variables is often surprising to the QSAR researcher when first encountered, it is a minor inconvenience compared to the risk of chance correlation when MR is used with many variables, and the impossibility of using MR successfully when the candidate variables outnumber the observations. Given the large number and the colinearity of the variables which are physicochemically plausible descriptors for almost any set of structure–activity data, and the improved predictivity which is observed for PLS in comparative studies of data analysis methods, PLS seems the natural choice of the informed researcher for data analysis in any QSAR situation.

REFERENCES

1 Joreskog, K. and Wold, H. (Eds.) Systems Under Indirect Observation: Causality, Structure, Prediction, North-Holland, Amsterdam, 1982. PLS was developed in response to a request by the Swedish government for an objective method of handicapping state-run horse races (S. Wold, private communication).
2 Wold, S., Martens, H. and Wold, H., In Kagstrom, B. and Ruhe, A. (Eds.) Lecture Notes in Mathematics, Matrix Pencils, Vol. 973, Springer, Berlin, 1983, p. 286.
3 Wold, S., Ruhe, A., Wold, H. and Dunn III, W.J., SIAM J. Sci. Stat. Comput., 5 (1984) 735.
4 Cramer III, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
5 Lorber, A., Wangen, L.E. and Kowalski, B.R., J. Chemometrics, 1 (1987) 19.
6 Hoskuldsson, A., J. Chemometrics, 2 (1988) 231.
7 Wold, H., In Krishnaiah, P. (Ed.) Multivariate Analysis, Academic Press, New York, NY, 1966, p. 391. For one of several more efficient methods, see Lindgern, F., Geladi, P. and Wold, S., J. Chemometrics, 7 (1993) 45.
8 Cramer III, R.D., J. Am. Chem. Soc., 102 (1980) 1837.
9 Cramer III, R.D., Bunce, J.D., Patterson, D.E. and Frank, I.E., Quant. Struct.–Act. Relatsh., 7 (1988) 18.
10 Baroni, M., Costantino, G., Cruciani, G., Riganelli, D., Valigi, R. and Clementi, S., Quant. Struct.–Act. Relatsh., 12 (1993) 9.
11 a. Topliss, J.G. and Edwards, R.P., J. Med. Chem., 22 (1979) 1238.
b. Topliss, J.G. and Costello, R.J., J. Med. Chem., 15 (1971) 1066.
12 Clark, M. and Cramer III, R.D., Quant. Struct.–Act. Relatsh., 12 (1993) 137.
13 Wakeling, I.N. and Morris, J.J., J. Chemometrics, 7 (1993) 291.
14 Cramer III, R.D., DePriest, S.A., Patterson, D.E. and Hecht, P., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, 1993, pp. 443–485.

APPENDIX

PLS with cross-validation algorithm in Sybyl/QSAR

Algorithmically the PLS implementation can be stated as follows:

Let P_{ij} be the predictor matrix,
R_{ik} be the response matrix,
i = 1..N is row index,

j = 1..p is predictor column index,
m = 1..p as j,
k = 1..r is response column index, and
a = 1..A is component index
g = 1..G is subgroup index for cross-validation, if desired.

First perform block-based 'CoMFA standard' scaling, if desired, on X matrix.
Partition the rows into G disjoint groups by random selection.

FOR EACH CROSS-VALIDATION SUBGROUP g:
Construct the subgroup matrices by dropping nonincluded rows to form X_{ij} (predictors, from P matrix) and Y_{ik} (responses, from R matrix), with i = 1..n, n < N.
Compute mean and (if autoscaled) sigma for each column in X, and generate

$X'_{ij} = X_{ij} - \text{mean}_j$ or (if autoscaling),
$X'_{ij} = (X_{ij} - \text{mean}_j/\text{sigma}_j$ and similarly for Y'_{ik}.

FOR EACH COMPONENT a:

initialize response latent variable:
$V_{ia} = Y_{i1}$ for all i

DO THIS...

compute predictor weights:

$$W1_{ja} = \sum_j (X'_{ij} \cdot V_{ia})$$

scale predictor weights to unit length:

$$W1_{ij} = W1_{ij} / \sqrt{\sum_j W1_{ja}^2}$$

predictor latent variables:

$$U_{ia} = \sum_j (X'_{ij} \cdot W1_{ij})$$

response weights:

$$W2_{ka} = \sum_i (Y'_{ik} \cdot (U_{ia})$$

scale response weights to unit length:

$$W2_{ka} = W2_{ka} / \sqrt{\sum_j W2_{ka}^2}$$

response latent variables:

$$V_{ia} = \sum_k (Y'_{ik} \cdot W2_{ka})$$

...UNTIL CONVERGENCE OF V_{ia} OR ITERATION COUNT IS REACHED.

inner relationship:

$$RO_a = \sum_i (V_{ia} \cdot U_{ia}) / \sum_i (U_{ia}^2)$$

predictor loadings:

$$B_{ja} = \sum_i (X'_{ij} \cdot RO_a \cdot U_{ia} / \sum_i (RO_a) \cdot U_{ia})^2$$

residuals:

$$X'_{ij} = X'_{ij} - U_{ia} + RO_a \cdot B_{ja}$$

$$Y'_{ik} = Y'_{ik} - U_{ia} \cdot RO_a \cdot W2_{ka}$$

END OF COMPONENT LOOP.

Compute the regression coefficients as follows:
Set Z_{jm} to the identity matrix
Set b_{mk} to the zero matrix

FOR EACH COMPONENT a

$$b_{mk} = b_{mk} + W2_{ka} \cdot RO_a \cdot \sum_j (Z_{jm} \cdot W1_{ja})$$

$$Z_{jm} = Z_{jm} - B_{ja} \cdot RO_a \cdot \sum_n (Z_{nm} \cdot W1_{na})$$

Unscale the coefficients and obtain the offset term:

FOR EACH RESPONSE COLUMN k
Initialize the $OFFSET_k$ to be $(mean_k)$

FOR EACH PREDICTOR COLUMN j
$b_{jk} = b_{jk} \cdot (sigma_k)/(sigma_m)$
$OFFSET_k = OFFSET_k - b_{jk} \cdot (mean_j)$
END PREDICTOR LOOP

FOR EACH ROW r 'NOT' IN SUBGROUP
Compute the predicted $R_{rk} = OFFSET_k + \sum_j (b_{jk} \cdot X_{rj})$
Compute the error of prediction $e = R_{rk}$ – predicted R_{rk}
Add e^2 to SSE_{ka}, the sum of squared errors for response k at component a
END ROW PREDICTIONS

END RESPONSE LOOP

...END OF COMPONENT LOOP