### Abstract

A partial least squares treatment of multivariate data related through a complex model allows one to evaluate the interactions between large numbers of features at once. Results where the model is of water sources flowing together, each block composed of water quality data, allow the influence of the various sources to be evaluated with respect to their importance on the resulting flow downstream.
Partial Least Squares Path Modelling with Latent Variables

Prepared for Publication

in

Analytica Chimica Acta

by

Robert W. Gerlach
Bruce R. Kowalski*
and
Herman O. A. Wold¹

Laboratory for Chemometrics
Department of Chemistry, BC-10
University of Washington
Seattle, Washington 98195

September 1979

Reproduction in whole or in part is permitted for any purpose of the United States Government

Approved for Public Release; Distribution Unlimited

¹Department of Statistics
University of Uppsala
Uppsala, Sweden
SUMMARY

A partial least squares treatment of multivariate data related through a complex model allows one to evaluate the interactions between large numbers of features at once. Results where the model is of water sources flowing together, each block composed of water quality data, allow the influence of the various sources to be evaluated with respect to their importance on the resulting flow downstream.
When the goal of a study is to understand the inter-
relationship among several parts of a complex system,
statistical procedures are often employed to analyse features
from sets of samples collectively used to represent each part.
All too often, the number of features and/or parts is larger
than the number of samples and many multivariate statistical
procedures fail to be useful. A simple example is the case
where one set of independent features is to be related to only
one dependent feature by multiple regression analysis,
represented as Model I in Figure 1. The calculation can give a
perfect but possibly meaningless fit if the number of features
is greater than the number of samples. For the establishment
of a predictive model this problem is normally overcome by the
use of stepwise regression analysis. However, in this analysis
the regression coefficients are uninformative with respect to
our understanding of the model and the results provide no
information about the utility of the omitted features, which
may be only a little less informative than those chosen to
provide the best fit.

Consider the case where multiple blocks of data, each
block consisting of several features obtained over several
samples, are to be interrelated by a complex scheme or path
model. When only one block of features is to be related to a
second block of features, shown as Model II in Figure 1, a
canonical correlation analysis [1] or target-transformation analysis [2] can be carried out. For more than two blocks of data various multidimensional scaling techniques have been developed [3] which relate blocks of features along axes preserving the maximum amount of all interblock information at once. However, when not all interconnections between blocks are desired or relevant, more flexible methodology is required.

This new methodology, herein called the PLS (Partial Least Squares) approach to Path Modelling using Latent Variables, has recently been developed by H. Wold [4-8]. This important new tool allows blocks of features to be represented by unobservable or "latent" variables indirectly observed. The latent variables are then related to one another by a path or interconnection scheme predetermined by the user. The latent variables are found by an iterative procedure involving simple and multiple regression analysis so that they simultaneously and optimally (in the PLS sense) represent the measured features and provide the best fit to the path model. The method is so general that principal component analysis, multiple regression analysis, and canonical correlation analysis are included as special cases. The first application of this method to the physical sciences, an analysis of water chemistry measurements to assess the environmental impact of mine spoils drainage, is reported here.
In order to understand the impact of coal mining on local water quality, R. Skogerboe et al. [9] monitored several water quality parameters at numerous sites on Trout Creek in Colorado. Data taken at monthly intervals from October 1973 to July 1976 were provided by Skogerboe [10] for this study. Five sites best characterized the environmental impact and were selected for our present analysis. Site 1 is upstream from runoff influenced by spoils of the Midway Edna Coal Mine, which is adjacent to the stream. Sites 2, 3, and 4 monitor the runoff from strip mine spoils representing mining activity from the 1930s to the 1940s, the 1940s to the 1950s, and the 1960s to the present, respectively. Runoff from these sites enters the stream in the order given above. Site 5 is downstream from the mine. Only 25 months of data were included in this study since occasionally several features at a site were not determined in certain months. At each site the data set was composed of eleven features, pH, Cl\(^{-}\), SO\(_4^{2-}\), Ca\(^{2+}\), Fe\(^{2+}\), K\(^{+}\), Mg\(^{2+}\), Mn\(^{2+}\), Na\(^{+}\), Zn\(^{2+}\), and HCO\(_3^{-}\), all but pH reported in mg/l. The final data set had approximately eight percent of its values missing, which we filled in so as to minimize any deviation from a particular site’s known data structure [11].

Our goal was to establish a path model using all five sites. Each site, represented by a data matrix of 11 features sampled over 25 months, was used in the model as a separate
entity. In our present case the path model is clearly that shown as Model III in Figure 1. The only relationship possible is that site 1, the upstream site, and sites 2, 3, and 4 mix to form site 5, the downstream site.

In order to consider the effect of all features at once the method forms latent variables,

\[ L_k = \sum_{i=1}^{N_k} a_{k,i} x_{k,i} \]

at each site, where \( N_k \) is the number of features being considered at site \( k \), \( x_{k,i} \) is the value of feature \( i \), and the \( a_{k,i} \)'s are coefficients determined in the course of the analysis. The \( a_{k,i} \)'s for each of the upstream sites are estimated from a multiple regression of all the features at a particular site to the downstream latent variable, \( L_5 \), as diagramed in Model III of Figure 1. All coefficients \( a_{k,i} \) are then scaled so that the latent variables \( L_k \) have unit variance.

Next, \( L_5 \) is regressed upon the upstream latent variables to estimate the \( P_{k,5} \)'s in the expression

\[ L_5 = \sum_{k=1}^{4} P_{k,5} L_k \]

Using the \( P_{k,5} \)'s and \( L_k \)'s to estimate \( L_5 \) we perform a multiple regression of the features of site 5 on it in order to estimate
the $a_{5,1}'s$. From the newly found $a_{5,1}'s$ we form a new $L_5$ which is scaled to unit variance and the entire procedure is repeated until all $a_{k,1}$ and $P_{k,5}$ converge. All calculations were initiated with all $a_{k,1}$ and $P_{k,5}$ set to one. A similar series of path models can be developed to analyze any number of blocks of variables connected by any set of paths.

Using all 11 features in each block, the calculation of Model III converged with an overall fit of 0.99. The square of the fit correlation coefficient, $R^2$, gives the relative amount of information at $L_5$ accounted for by the other four latent variables and is calculated from

$$R^2 = \sum_{k=1}^{4} P_{k,5} R_{k,5}$$

where $R_{k,5}$ is the correlation between $L_k$ and $L_5$. The site contributions to $R^2$ are given in Table 1. We note that the good fit is primarily due to a strong relation between sites 4 and 5. The contributions of each individual feature to the fit were calculated and showed that the high correlation was due largely to a fit between HCO$_3^-$ at site 4 and Ca$^{2+}$ and Mg$^{2+}$ at site 5. Although only a small amount of the total variance in all of the data is accounted for by this relationship, it is a rather striking one as HCO$_3^-$ introduced by site 4 strongly buffers the Ca$^{2+}$ and Mg$^{2+}$ concentration.
A principal component analysis of the features at site 5 yielded two readily interpretable components. The first component represented the major salt load $\text{Ca}^{2+}$, $\text{Mg}^{2+}$, $\text{Na}^+$, $\text{K}^+$, $\text{SO}_4^{2-}$, and $\text{Cl}^-$ on the creek and the second component represented primarily the trace metals zinc and manganese. Thus, a more directed analysis targeting on the principal components was suggested. Results of Model III calculations where $L_5$ is represented by an individual principal component are also shown in Table 1. The first component is modeled by the upstream values of site 1 and the first source of mine drainage represented by site 2. These results indicate that site 2 has by far the most dramatic effect on water quality. Similar results were obtained for the second principal component with an additional smaller contribution from site 4.

We have also performed Model III calculations when $L_5$ represents only one of the features from site 5, a non-iterative calculation. An example using $\text{Cl}^-$ is also shown in Table 1. Though the concentrations of $\text{Cl}^-$ and the other major species at sites 2, 3, and 4 are comparable in magnitude [9], drainage from site 2 is obviously the dominant influence on the downstream $\text{Cl}^-$ concentration. Drainage represented by site 4 also perturbs the downstream $\text{Cl}^-$ concentration, most likely because it represents flow from the newest spoils, which have a greater concentration of the more soluble salts. The lack of
influence from site 3 shows that drainage by this site is not
different enough or large enough to alter the Cl\(^{-}\) composition
set at site 2.

From the above it is clear that quantitative estimates of
the effect of stream components contributing to the load at the
downstream site can be made. In addition, detailed information
can be obtained on each component. For example, for many
species which have a high concentration at an upstream site but
fail to be used in modelling the downstream site, we believe
some form of buffering or precipitation action may be taking
place. In these cases the PLS analysis show where more
extensive investigation should be directed if the stream
chemistry is to be fully understood. Conclusions we have
arrived at using the PLS path modelling scheme are compatible
with those obtained in our laboratory using a battery of
standard multivariate techniques on a more extensive data set
of which the present data is a subset.

The above results show how PLS path modelling using latent
variables can provide insight into the interrelationships
between groups of features. It is especially important to note
that the treatment of groups of features as a unit allows one
to include many more features in the analysis than would
normally be allowed by more conventional techniques when one is
confronted with limited quantities of data. In all the above
calculations we have considered 44 features in sites 1 through 4 and obtained consistently interpretable results with only 25 sets of data. This form of analysis can be a powerful aid to anyone confronted with blocks of features which are related to one another along a set of logical paths.

This work was partially supported by the Office of Naval Research.
REFERENCES


11. S. Wold, Pattern Recognition, 8 (1976) 127.
Fig. 1. Model I represents a multiple regression analysis of one matrix onto a single feature, Model II depicts two matrices of features related to one another, and Model III shows the particular multi-matrix path model dealt with through a partial least squares analysis. In Model III the 4 matrices on the left represent sources of flow in a watershed which combine to form the flow represented by the fifth matrix.
Table 1. $(P_{k,5}X(R_{k,5})$ values for sites 1 through 4 and the corresponding $R^2$ for models where $L_5$ is described in column 1. PCs are principal components.
<table>
<thead>
<tr>
<th></th>
<th>Site 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 Features</td>
<td>0.02</td>
<td>-0.04</td>
<td>0.06</td>
<td>0.93</td>
<td>0.97</td>
</tr>
<tr>
<td>PC 1</td>
<td>0.35</td>
<td>0.69</td>
<td>-0.16</td>
<td>0.03</td>
<td>0.91</td>
</tr>
<tr>
<td>PC 2</td>
<td>0.21</td>
<td>0.59</td>
<td>0.00</td>
<td>0.11</td>
<td>0.91</td>
</tr>
<tr>
<td>Cl$^{-}$</td>
<td>0.09</td>
<td>0.58</td>
<td>-0.08</td>
<td>0.29</td>
<td>0.88</td>
</tr>
</tbody>
</table>