CHAPTER 3

METHODOLOGY OF CORRELATION ANALYSIS

Despite the lack of a general formula relating retention parameter to structure, many attemps have been taken to detect any existing regularities of retention behavior among selected subgroups of solutes chromatographed. For the more diverse sets of solutes, many empirical or theoretical solute parameters have been investigated. The relationship between the retention parameter and these parameters usually have been doubtful and required statistical evaluation in order to check the significance of the resulting correlations. The statistical techniques that are regularly used, are multiparameter regression analysis, principal component analysis and partial least-squares analysis, which will be described in the following parts.

3.1 Multiparameter Regression Analysis

Multiparameter regression analysis was used to describe a relationship between the retention index, which is the dependent variable, and the solute parameters, which are the independent variables. A stepwise regression procedure was used to find the parameters that account for the largest amount of variance in the dependent variable.The correlation coefficients between the different independent variables were calculated to find the relationship between them.

The regression equations obtained for the retention indices on the three GC columns are of the general form :

$$I = a_0 + a_1x_1 + a_2x_2 + \ldots + e_i \qquad (3.1)$$

where I is the retention index, x_i are the independent variables, a_i are the calculated coefficients and e_i is the residual error.

The choice of the final models was based on the following three criteria :

1. The model that accounts for the largest amount of variance in the dependent variable.

2. A total standard error or standard deviation as small as possible.

3. A t-test value for each independent variable in the model higher than or equal to a defined t-value for the 95% confidence interval

The t-value for each parameter can be calculated by the following equation :

$$t = \frac{\text{estimated value}}{\text{standard error of est. value}} = \frac{a_i}{s_i} \tag{3.2}$$

This value will be compared with a theoretical value which depends on the degree of freedom of the model and the chosen confidence interval. The parameter will be rejected if t-value is lower than the theoretical value.

The criterion indicating whether a model is reliable or not is based on the difference between calculated and experimental retention indices of the products. For this purpose, a distinction between the training, test and predictive sets of products will be made. The training set and the test set are two groups of products that are available as standards. The test set consists of those products which are arbitrarily chosen out of the assembling of available products to test the model. The predictive set is the group of products that are not available at the laboratory.

The training and test sets are used to detect "outlier" in the data, which are then rejected. The training set serves as the basis for the multiparameter regression analysis, then the derived models are used for the prediction of the retention parameters of the test set. If the difference between the calculated and experimental results is not higher than 10% of the calculated result, the model is accepted and used for the prediction of data of the predictive set.

Multiple regressions were calculated using GENSTAT 5, a computer program, developed at the Rothamsted Experimental Station, Harpenden, GB and run on a VAX computer, VAX 780/8600 digital.

3.2 Principal Component Analysis (PCA)

The "true" descriptors that exactly describe the interaction between solutes and stationary phase are unknown. So, because of the need for a reliable correlation to predict the retention index of solutes, numerous descriptors have been investigated. Information about the intercorrelation between the retention indices and the descriptors or variables can be obtained with PCA. This multivariate statistical technique is found to be suited particular to this problem of categorizing information about the solutes, gives the evidence of chemical relevance for extracted information and help for gaining insight into the physical background of the system.

The mathematical operations in PCA can be performed by using eigenvalue-eigenvector extraction algorithms. The raw data matrix, in which each row concerns a particular compound and each column concerns a particular parameter (retention indices and descriptors), will be first converted into a covariance matrix (or a correlation matrix in case of normalized data). This matrix will then be transformed by the "eigenanalysis" method (of linear algebra) into a new diagonalised matrix (containing nonzero elements only on the diagonal). These are the abstract "eigenvalues" of the matrix and each abstract eigenvalue is associated with an abstract eigenvector and measures its relative importance. The first eigenvector is computed such that the sum of the magnitudes of the projections of all points on that vector is maximum. In other words, as much variation in the data as possible lies along the direction of the first vector. The second eigenvector is chosen, orthogonal to the first, so that as much of the remaining variations lies along this vector, and so on. The characteristic feature of the eigenvectors is that the correlation coefficient between them is exactly zero, which makes them completely orthogonal. Because of experimental error in the data matrix, PCA will determine a number of eigenvectors equal to the number of columns in the data matrix. However, only the eigenvectors associated with the highest eigenvalues have physical meaning. The factor model has to be compressed to incorporate only the physical significant factors. The number of factors involved gives an idea about the complexity of the data. A stepwise reproduction procedure is used to define the correct number of factors. In the first step, only the first and most important eigenvector is used to reproduce the data matrix. In the second step the next eigenvector is added and so on until the data matrix is reproduced within experimental error. Extra factors will reproduce experimental error and therefore can be dropped. In this way the data matrix is decomposed into the product of an abstract row matrix (also called scores matrix) and an abstract column matrix (also called loading matrix).

To attach chemical significance to the abstract factors, the principal factors are transformed into recognizable parameters.

Target testing and abstract rotation are two transformation procedures that can be used. With rotation, the abstract matrices are transformed into other abstract matrices. The results can be used to identify the objects and the variables of the original data matrix. If certain variables have relatively the same cofactors for the same abstract factors, they form a cluster and can be thought of containing analogous information. But rotation cannot isolate individual real factors.

Potential real factors can be tested with target transformations. Real factors can be basic vectors, which describe the properties of the objects or variables, or can be typical vectors, which are rows or columns of the original data matrix. Basic factors describe the physical background of the matrix and are the most important ones. Typical key factors can reduce the data matrix and represent empirically the information they contain. In this way useful predictions of new or missing data can be made. This target transformation technique uses the least-squares method with the principal factors and the target tester. If the predicted vector is reasonably similar to the target vector, the test vector is assigned a real factor. Once all the real factors are known,the new or missing data can be predicted .[3]

PCA was performed by using software "TARGET 90" copyrighted by E.R. Malinowski (34) and run on an IBM compatible personal computer.

The results obtained from PCA must be judged as a source of information about the complexity of the data, not as an end in itself.

[3] see the summary of mathematical procedures in Appendix C or on the ref. 34

3.3 Partial Least-Squares Analysis (PLS)

PCA works on one data matrix containing several measurements on several objects. All the data (retention index and descriptors) are tabled together in this one matrix. In the case where the data can be divided into two matrices X (containing the descriptors) and Y (containing the retention indices) and Y is to be predicted from X, PLS can be applied. PLS models were developed by H.Wold as extensions of PCA models (28). In PLS the data in the X and Y matrices are modeled by separate PCA development creating PC like models for X and Y. X and Y are connected by a diagonal matrix. PLS simultaneously makes the model both fit X and Y, to make X predict Y. The dimension is estimated by cross validation. By this technique, some data elements are kept out of the class data set, fitting the model to the remaining data with different dimensions and calculating the predicted values of the kept out data from the different models. Then other data elements are kept out and the same procedure is repeated until each data element has been kept out once. The dimension corresponding to the best prediction of kept out elements is taken as the number of significant components.

The separate model of each class of subsets provides several advantages. With PLS the data can be analyzed when the number of objects are less than the number of variables. As long as the number of principal components is fewer than about one-third of the number of objects, the component scores and residual standard deviations are well-defined and stable even with many variables compared to the number of objects. Second, the concept of the model allows the detection of outliers and abnormal objects, which otherwise will misrepresent the class patterns.

PLS was performed using software "UNSCRAMBLER" of CAMO at Solvay-Duphar, The Netherlands and run on an IBM compatible personal computer.