1

Fundamentals

1.1 Multilinear and Nonlinear Regression Analyses

Traditional simple linear regression analysis calculates the correlation between one dependent (i.e. log k′ value of a solute) and one independent variable (i.e. concentration of organic modifier in the eluent C vol%). The general formula can be described by:

$$Y = a + b_1 \cdot X_1 \tag{1.1}$$

where $Y = \log k'$, $a =$ constant, $X_1 = C$ and b_1 is the coefficient of regression.

However, when the retention time of a set of solutes has been determined on more than one HPLC column, in mobile phases containing various organic modifiers at different concentrations and at various temperatures and we are looking for the influence of each chromatographic parameter on the log k′ value, each of them has to be included in equation (1.1), which modifies to:

$$Y = a + b_i \cdot X_i + \cdots + b_i \cdot X_i + \cdots + b_k \cdot X_k \tag{1.2}$$

Multivariate (multiple) linear regression (MLR) analysis deals with the solution of equations containing minimally two independent variables. Moreover the parameters included in equation (1.2) software generally calculate some additional mathematical-statistical values such as standard deviation of regression coefficients (s_{bi}, where 'i' indicates any of the independent variables included in MLR), the path coefficients (b_i'), calculated F values, and the coefficient of determination (r^2). The path coefficients are dimensionless numbers indicating the relative impact of a given independent variable (stationary phase characteristics, mobile phase composition, column temperature, etc.) on the dependent variable (generally retention time of the set of analytes). A higher path coefficient value (close to 1) indicates a higher influence of the given independent variable on the dependent one. The calculated F value shows the fit of the equation to the experimental data. When it is higher than the tabulated value corresponding to the same number of observations and independent variables, the relationship between the variables is significant. Calculated F values can be

Multivariate Methods in Chromatography: A Practical Guide Tibor Cserháti

found in every statistical handbook. The coefficient of determination indicates the ratio of the change of the dependent variable explained by the change of the independent variables. In other words, $r^2 = 0.9243$ indicates that 92.43% of the change of the dependent variable can be explained by the change of the independent variables.

The number of papers dealing with the application of multiple nonlinear regression (MNLR) is fairly low. This may be due to the fact that linear models generally adequately describe the relationship between dependent and independent variables making it unnecessary to use the more complicated nonlinear (mainly exponential and logarithm) regression analysis.

1.2 Stepwise Regression Analysis and Partial Least Squares Method

Besides the traditional MLR technique some new computerized regression analysis methods have been developed based on theoretical considerations.

It is well known that in the traditional multivariate regression analysis the presence of independent variables (chromatographic parameters or physicochemical characteristics) that exert no significant influence on the dependent variable (retention behaviour) lessens the significance level of the independent variables that significantly influence the dependent variable. To overcome this difficulty, stepwise regression analysis (SRA) automatically eliminates from the selected equation the insignificant independent variables increasing in this manner the information power of the calculation. The form and the statistical parameters computed by SRA are identical to those discussed in Section 1.1 and their statistical meaning is the same.

The partial least squares (PLS) technique is actually a method of preference for carrying out regression analysis. The advantage of the method is that it can be successfully employed when the independent variables are highly intercorrelated. PLS relates two different sets of variables using one data set as predictor variables. The dimensionality of the original data set is reduced by estimating one or more underlying background variables. PLS forms the score vectors as linear combinations of the original sets of variables. The theory of PLS has been previously discussed in detail [1]. PLS has been readily accepted and extensively used in up-to-date mathematical-statistical studies. The theory of various regression analysis models is discussed in detail in Mager [2].

1.3 Two- and Three-dimensional Principal Component Analysis, Various Factor Analytical Techniques

In many cases the chromatographer is not interested in the dependence of one retention parameter on the chromatographic or physicochemical characteristics (see above multiple regression methods), but rather wishes to find the relationship between all parameters without one being the dependent variable. Both PCA and factor analysis (FA) comply with this requirement. The main advantages of these computation techniques in chromatography are:

- elucidating the similarity and dissimilarity among the variables (clustering chromatographic systems or solutes according to their retention behaviour);

- the possibility of the extraction of one or more background (theoretical) variables having concrete physicochemical meaning for the theory and practice of chromatography;
- making it possible to reduce the number of variables (chromatographic systems or solutes) to the minimum necessary for the solution of a problem.

Traditional two-dimensional principal component analysis (2D-PCA), a versatile and easy-to-use multivariate mathematical-statistical method has been developed to contribute to the extraction of maximal information from large data matrices containing numerous columns and rows. PCA makes possible the elucidation of the relationship between the columns and rows of any data matrix without one being the dependent variable. 2D-PCA is a so-called projection method representing the original data in smaller dimensions. It calculates the correlations (similarities and dissimilarities) between the columns of the data matrix and classifies the variables according to the coefficients of correlations taking into consideration simultaneously the magnitude and sign of the coefficients of correlation. As the resulting matrices of principal component loadings and variables (scores) are also multidimensional, the visual evaluation of such matrices is cumbersome or even impossible.

The plot of principal component loadings and/or scores of the first versus the second principal component has been frequently used for the evaluation of the similarities and differences among the observations. This method takes into consideration only the variance explained in the first two principal components and ignores the impact of variances explained by the other principal components on the distribution of the matrix elements. The use of this approximation is only justified when the first two principal components explain the overwhelming majority of variance, which is not probable in the case of large original data matrices. Theoretically PCA can be used for the analysis of any data matrices; however, its inadequate application may lead to serious misinterpretation of the results. As it was previously mentioned, PCA calculates the similarities and dissimilarities among the variables and observations according to the differences among the coefficients of correlation. This means that the distribution of principal component loadings and scores will be similar in the theoretical cases when each coefficient of correlation is in the range of 0.1, 0.5 or 0.9. While the scattering of points calculated from the coefficient of correlation 0.1 does not contain any useful (significant) information, a similar distribution of points marks significant relationships when it is calculated from the correlation matrices of 0.9. The publication of a table containing each coefficient of correlation overcomes the difficulty emerging from the similar scattering of points.

PCA software generally calculates the so-called ‘eigenvalue’, the ratio of variance explained by the individual principal component loadings and scores, and the numerical values of principal component loadings and scores in the principal components. The total variance explained by PCA has to be previously fixed and arriving at this limit value the computation stops. Depending on the character of the original data matrix the total variance explained is generally fixed at 95 or 99%. Eigenvalues indicate the ratio of variance explained by the individual principal components; their value decreases monotonously with increasing number of principal components. It is generally assumed that only eigenvalues more than 1 contain valuable information; eigenvalues less than 1 only refer to the error of measurement. Similarly to eigenvalues, the ratio of variance also decreases with increasing number of principal components. The numerical values of principal component loadings

and scores indicate the impact of a given point of observation on the individual principal component.

Traditional PCA is a typical multivariate two-dimensional statistical method unsuitable for the evaluation of three(or more)-dimensional data matrices. A three-dimensional PCA (3D-PCA) model has been developed (Tucker model) to overcome this difficulty. However, 3D-PCA, a very elegant and easy method to carry out, has not found frequent application in the evaluation of chromatographic data matrices.

For the exact mathematical treatment of these methods see references [1], [2], [5] and [6] in the Introduction and Malinowski and Howery [3].

As a result of its versatility PCA has been extensively used not only in chromatography but also in other fields of research.

Interestingly, FA suitable for similar calculations has been less frequently employed in chemometrics studies than PCA.

1.4 Canonical Correlation Analysis

Sometimes analytical chemists are interested in the simultaneous dependence of more than one retention characteristic or more than one chromatographic parameter on a considerable number of chemical (eluent composition, eluent pH, etc.), physical (temperature, flow rate, etc.) and physicochemical (connectivity and sterical indices, etc.) variables. As the number of dependent variables is higher than one, traditional MLR models cannot be applied in these instances. Canonical correlation analysis (CCA) has been developed for the solution of this type of computational problem; it calculates the correlation between two different data sets. CCA computes linear correlations between the variables of the two data sets, both of them including more than one variable. The variables in the data set with lower number of variables (matrix II, i.e. more than one retention parameter) are considered as dependent variables, while the variables in the data set with higher number of variables (matrix I, i.e. physicochemical parameters of solutes) are considered as independent variables. The correlation coefficients of the linear relationships are computed in such a manner that they explain the maximal information in the data set containing the lower number of variables (matrix II) using the information in the other matrix (matrix I). The number of relationships explaining 100% of the variance in matrix II is equal to the number of original variables in matrix II. The theoretical background of CCA is given in detail in Orloci *et al*. [4].

Despite its obvious advantages, CCA has found only limited application in chromatography [5–8].

1.5 Discriminant Analysis

Discriminant analysis (DA) can be used for the computation of a hyperplane in the input space minimizing the within class variance and maximizing the between class distance. It is one of the supervised pattern recognition methods frequently used in various fields of chemometric calculations. It finds explicit boundaries between given classes, in order to discriminate them. The latent (combined) variable is the linear combination of the original variables.

1.6 Spectral Mapping

The majority of multivariate methods classify the chromatographic retention data taking into consideration simultaneously the strength and selectivity of the retention, so cannot be applied when the separation of the strength and selectivity of the effect is required. The spectral mapping (SPM) technique, another multivariate mathematical-statistical method, overcomes this difficulty [9]. The method divides the information into two matrices using the logarithm of the original data. The first matrix is a vector containing the potency values related to the overall effect. The second matrix (selectivity map) contains information concerning the spectra of activity (the qualitative characteristics of the effect) [10]. SPM first calculates the logarithm of the members of the original data matrix facilitating the evaluation of the final plots in terms of log ratios. Consecutively, SPM subtracts the corresponding column-mean and row-mean from each logarithmic element of the matrix calculating potency values. The source of variation remaining in the centred data set can be evaluated graphically (selectivity map). SPM has been previously employed for the characterization of stationary phases in HPLC [11].

1.7 Nonlinear Mapping

As mentioned in Section 1.3, the principal components and scores computed by PCA are generally multidimensional, which makes the visual evaluation cumbersome, even impossible. Nonlinear mapping (NLM) has been developed for the reduction of the dimensionality of complicated matrices consisting of numerous columns and rows. The method can reduce the dimensionality of the data matrices to two in such a manner that the distances between the points on the projection plane approximate the distances on the multidimensional space. This means that points (i.e. chromatographic systems) near to each other are similar and points situated far away from each other are markedly different.

1.8 Cluster Analysis

Similarly to NLM various cluster analysis (CA) techniques have been developed and successfully employed for the easy visualization of multidimensional data matrices by reducing dimensionality. Variables with similar characteristics are near to each other on the CA dendograms, while variables with different characteristics are far away from each other. On account of the good visualization of the results, CA is generally combined with other multivariate methods (mainly with PCA). The principles of CA are discussed in detail in Willett [12].

1.9 Other Multivariate Techniques

In addition to the mathematical-statistical methods discussed above, many other computational techniques have been developed and applied in the chemometrical evaluation of

chromatographic data sets. However, they have not been frequently used so that their contribution to mathematical-statistical evaluation in chromatography is fairly low.

1.10 Measured and Calculated Physicochemical Parameters of Chromatographic Systems and Analytes

A high number of physical, physicochemical and biophysical parameters have been applied in the chemometrical investigation of chromatographic data. The majority of these characteristics are calculated and not measured values. This can be explained by the fact that the advent of rapid computers with high calculation capacity and complicated software make possible the rapid calculation of numerous parameters, while the measurement of any concrete parameter is time-consuming and sometimes expensive compared with computational results.

Since different principles are involved in the various chromatographic techniques, the impact of parameters used in the calculations may also be different. Thus, the importance of a given molecular characteristic may be high in GC and negligible in CZE. The parameters used more frequently in GC are: boiling point, molar volume, molecular mass, molar refraction, octanol-water partition coefficient, various indicator variables (i.e. no side chain = 0, one side chain = 1, two side chains = 2), AM1 total energy, Randic molecular profile, Randic shape profile, mean electrotopological state, number of rotatable bonds, Gutman molecular topological index by valence vertex degree, third component accessibility directional WHIM index/unweighted, H autocorrelation of lag 5/weighted by atomic masses, R maximal autocorrelation of lag 2/weighted by atomic Sanderson electronegativities, H autocorrelation of lag 2/weighted by atomic Sanderson electronegativities, R maximal autocorrelation of lag 4/weighted by atomic van der Waals volume, Balaban-type index from van der Waals weighted distance matrix, Balaban-type index from electronegativity weighted distance matrix, average valence connectivity index chi-2 and chi-0, valence connectivity index chi-1, polarity index, magnitude of dispersive interactions between a methylene group and the stationary phase, partial molar excess Gibbs free energy of solution per methylene group, cohesive energy, solubility parameter, enthalpy of vaporization distance edge vector, standard molecule chemical potential, energy of the lowest unoccupied molecular orbital (LUMO), dipole moment (DIP), the maximum of the net atomic charge on the C atom (QMAX), the sum of positive charge on C atoms (QTOT), topological indexes (CHI-2, CHI-0AV, CHI-2AV, CHI-0a, CHI-2A), wiener index (WA), heat of formation (HOF), total energy (TE), the maximum of the net atomic charge on the H atoms (Q+), binding energy (BE), core–core interaction (CCIE), solvent accessible surface area (SASA), polar solvent accessible surface area (pSASA), and apolar solvent accessible surface area (apSASA).

Quantitative structure–retention relationship (QSRR) studies carried out in subcritical chromatography used excess molar refraction, dipolarity/polarizability, hydrogen bond acidity and basicity, and McGowan's characteristic volume as molecular parameters.

The principle of separation in TLC and HPLC is similar, so the physicochemical parameters employed are also similar. Thus, the following molecular characteristics have been extensively applied: Hansch-Fujita's substituent constants characterizing hydrophobicity (π); indicator variables for proton acceptor and proton donor properties (H-Ac and H-Do,

respectively); molar refractivity (M-RE); Swain and Luton's electronic parameters characterizing the inductive and resonance effects (F and R, respectively); Hammett's constant characterizing the electron-withdrawing power of the substituents at meta and para+ortho positions (σ_m and σ_{p+o}, respectively); Taft′s constant characterizing the steric effects of substituents (Es); Sterimol width parameters determined by distance of substituents at their maximum point perpendicular to attachment (B_1 and B_4). These molecular parameters were calculated according to the additivity rule from the fragmental constants. Fragmental constants are parameters characterizing elementary molecular substructures. However, it has to be borne in mind that the application of the additivity rule does not take into consideration the possible intramolecular interactions among the substructures in the molecule which may result in inadequate results.

References

[1] Geladi, P., and Kowalski, B. R. Anal. Chim. Acta 185 (1986) 1–17.
[2] Mager, H. Moderne Regressionsanalyse, Salle, Sauerländer, Frankfurt am Main, 1982.
[3] Malinowski, E. R., and Howery, D. C. Factor Analysis in Chemistry, John Wiley & Sons, Ltd, New York, 1980.
[4] Orloci, L., Rao, C. R., and Stitiler, W. M. Multivariate Methods in Ecological Work, International Cooperative Publishing House, Fairland, MD, 1979.
[5] Forgács, E., Cserháti, T., and Bordás, B. Chromatographia 36 (1993) 19–26.
[6] Forgács, E., Cserháti, T., and Bordás, B. Anal. Chim. Acta 279 (1993) 115–122.
[7] Cserháti, T., and Forgács, E. Chem. Intell. Lab. Syst. 28 (1995) 305–313.
[8] Forgács, E., and Cserháti, T. J. Liq. Chrom. Rel. Technol. 19 (1996) 1849–1858.
[9] Lewi, P.J. Arzneim. Forsch. 26 (1976) 1295–1300.
[10] Lewi, P. Chemom. Intell. Lab. Syst. 5 (1989) 105–116.
[11] Hamoir, T., Cuaste Sanchez, F., Bourguignon, B., and Massart, D. L. J. Chromatogr. Sci. 32 (1994) 488–498.
[12] Willett, P. Similarity and Clustering in Chemical Information, Research Studies Press, New York, 1987.