MULTIVARIATE CURVE RESOLUTION

TASKS

1) Building up the experimental data matrix.

D (N_{sol}, N_{wave}) which contains the experimental data (UV, CD or fluorescence spectra) at the N_{wave} wavelengths of the N_{sol} mixtures obtained either at successive titration points in the acid-base titrations (i.e., each solution is characterized by a defined pH value) or at successive temperature values in the melting experiments.

When, for a single titration, spectra are measured with two different spectrometric techniques, the two data matrices obtained can also be analyzed simultaneously. These two matrices have the rows (pH values) in common, but the columns (spectrometric channels) are different. In this case, an augmented data arrangement can be set, keeping the row space in common and expanding the column space. A row-wise augmented data matrix is constructed and described as:

$$D_{aug} = [D_1, D_2].\ \underline{\text{Scheme 2}}$$

2) Estimation of the number of different species/components (N_s)

This number can be obtained from the 'chemical' rank of the matrix, i.e., from the more significant singular values associated with the chemical changes in the matrix D, which can be estimated from visual inspection of the plots of the larger singular values. If no important contributions of baseline drift or interferences are present, each significant singular value can be related to a different chemical species or conformation. A reproduced data matrix can now be calculated for the preselected number of components:

$$D = U \cdot S \cdot V^T + E = D* + E$$

where U, S and V^T are respectively the scores, singular value and loading matrices of D for the preselected number of components, E is the residual error matrix containing the variance not explained by U, S and V^T, and $D*$ is the reproduced data matrix. When the correct number of components is chosen, the residual error matrix E is close to the noise or experimental error.

3) Local rank analysis and initial estimates.

Once the number of species is estimated, the data structure of the individual matrices is analyzed using evolutionary factor analysis methods like Evolving Factor Analysis (EFA) or Evolving Factor Analysis with a Fixed Size Moving Window (FSMW).

When analyzing data matrices from acid-base titrations, EFA provides estimations of the pH regions of existence of each species and also of their evolution during the experiment. The performance of EFA can be summarized as follows: in a first step, a singular value

decomposition (SVD) of the submatrix containing only the two first spectra of the original data matrix **D** is made.

Once the singular values are calculated for this submatrix, the third spectrum of the data matrix is added to the initial submatrix and the SVD is then made in this new submatrix. This process is repeated until SVD is made over the whole data matrix. EFA can be performed in both pH directions: from acidic to basic pH values (which is called forward analysis) and from basic to acidic pH values (backward analysis). Forward analysis provides information about the appearance of the different components along the experiment, whereas the backward analysis provides information about their disappearance. An "abstract distribution plot" can be deduced from the correct junction of the forward and backward lines related to the different singular values for the **Ns** preselected number of components.

FSMW gives information about the pH range where a species is present and also allows the estimation of the number of components that coexist in a certain pH range. Within each run of FSMW a moving submatrix (window) embracing I successive spectra is defined. In the first step of a FSMW analysis, SVD is applied in a window containing the fisrt I spectra. In the second step, the window has shifted one spectrum. Thus, this second window contains spectrum 2 to I+1. The window shifts until it contains the last I spectra. Each window is decomposed by SVD, and the I singular values are plotted as function of the window number or as function of the pH.

4) Alternating Least Squares Optimization.

Linear model (generalized Beer law) equation in matrix form is solved iteratively by an alternating least squares ALS method to obtain the matrices of individual pure spectra S^T and of concentration profiles **C** which best fit the experimental data. In matrix form the lineal model can be written as

$$D^* = C \cdot S^T + E$$

where **E** is the matrix of the residuals not explained by the chemical species in **C** and S^T; residuals in **E** should be close to the experimental error. The use of matrix **D*** instead of the experimental data matrix **D** provides more stability in the calculations since **D*** is noise-filtered for the particular number of components.

When the row-wise augmented D_{aug} matrix, containing the data obtained by two spectrometric techniques, is analyzed, **C** is an individual matrix having the concentration profiles of the absorbing species, which are common to both D_1 and D_2 matrices, and **S** is an augmented matrix $[S_1, S_2]$ which contains the pure species spectra spanning the two spectrometric detection methods:

$$D_{aug} = [D_1, D_2] = C [S_1, S_2]^T + E = C S_{aug}^T + E$$

The ALS iterative method starts with an initial estimation of the **C** matrix obtained from EFA profiles of the experimental data matrix . A first constrain is imposed here if, from EFA and FSMW, selective regions have been detected, i.e. regions where a single species is present. In these regions the concentration of the other species is forced to be equal to zero. The ALS iterative method consists of two parts:

a) In the first part an estimation of the unknown species spectra is performed from the linear equation by least squares $S^T = C^+ D^*$

where C^+ is the pseudoinverse of **C**. The matrix S^T gives the current least-squares estimation of the pure spectra. The UV absorptivities or fluorescence intensities must be positive, whereas CD absorptivities can be positive or negative. This constraint is applied accordingly during the least-squares optimization.

b) In the second stage, a new estimation of the concentration profiles is obtained by least squares using the equation $C = D^* (S^T)^+$

where $(S^T)^+$ now is the pseudoinverse of the S^T matrix. In this case the concentrations derived from the equation are constrained not only to be positive but also to give unimodal concentration profiles (i.e., profiles without double peaks). As the total concentration of the absorbing species is known, normalization of the concentration profiles by closure is also applied at this stage.

c) Steps a and b are repeated until the data matrix D^* is well explained within experimental error. Convergence is usually achieved in a few number of iterations of the ALS regression method.

The ALS iterative step can also start with the initial estimation of the S^T matrix. Initial spectra are usually derived from techniques based on the detection of "purest variables".

Concentration profiles and pure spectra obtained after the ALS analysis of different spectroscopic titrations of the same system can be somewhat discordant because of the underlying presence of certain factor analysis ambiguities (rotational and intensity ambiguities).

Whether the final solutions are equal to the true ones depends on the selectivity (pH or temperature regions where only one species is present or wavelengths where only one species absorbs) and local rank (number of species simultaneously present in a pH or temperature region) of the analyzed systems. When a high degree of selectivity is present, the recovered solutions are close to the true ones.

Moreover, the numerical solutions can be highly constrained after the performance of the simultaneous analysis of different data matrices; if the rank conditions of resolution are achieved for one species in one matrix, then the resolution is also obtained for the same species in the other matrices (even if resolution conditions did not hold for this species in those matrices).

Another advantage of the proposed soft-modeling procedure is the use in the iterative optimization process of first estimations that have chemical sense, i.e. distribution profiles estimated from EFA or "pure" spectra obtained from techniques based on the detection of "purest" variables, compared with those soft-modeling procedures that perform the iterative process from the pure SVD results.

This fact, together with the use of the different constraints and the analysis of augmented matrices is of great help for the correct resolution in the analysis of complex systems.

comments to: Romà Tauler, Anna de Juan   **Top of the page**
Main Page   **updated: February 2005**