INVESTIGATION OF CALIBRATION-FREE RESOLUTION TECHNIQUES AND INDEPENDENT COMPONENT ANALYSIS

S.Triadaphillou, I.Wells*, A. J. Morris and E. B. Martin

Centre for Process Analytics and Control Technology,,
School of Chemical Engineering and Advanced Materials,
University of Newcastle, Newcastle upon Tyne, NE1 7RU, UK
**Avecia, Process Technology Department, Grangemouth, FK3 8XG, UK*

Abstract: Calibration-free resolution techniques provide an alternative approach to the development of a calibration model. These combine spectroscopic measurement coupled with mathematical and statistical assumptions and give spectral profiles and non-quantitative concentration profiles for the unknown mixture. In this paper, a number of calibration free techniques including VARIMAX, ITTFA, EFA, FSWEFA, SIMPLISMA are described and applied to a synthetic spectral data set and the results are compared with the complementary technique of Independent Component Analysis (ICA) in particular FastICA and JADE. The results were comparable in all cases with ICA separating the signal from the constituent components successfully.

Keywords: Reaction Monitoring and Control, Spectral Conjunction, Spectral Deconvolution, Multivariate Calibration, Iterative Resolution Methods

1. INTRODUCTION

A number of issues are associated with the development of calibration models to predict the concentration of a product in a reaction. For example their development in terms of data generation and collection can be time consuming, the model will be sensitive to changes in process conditions and it only provides quantitative information about the property of interest with no information about side reactions and intermediates. An alternative approach is the family of calibration free resolution techniques. These enable the analyst to make full use of time resolved spectra for the determination of both qualitative and quantitative information, i.e. pure spectra and concentration profiles over the course of a reaction. In addition, on-line analysis of laboratory reactions can markedly improve both the timeliness and quality of information regarding mechanisms and kinetics, compared to the more traditional approaches of extractive sampling. Thus the application of calibration-free methods for on-line analysis can result in major advantages in terms of the understanding of a process.

Most calibration-free resolution techniques are based on the assumption that the instrumental response in a mixture is an additive linear combination of the signals from individual species, the pure components. Consequently it obeys Beer's law, i.e. the spectral response of the components is independent of time and concentration (Miller and Steele, 1990). In the case of reaction monitoring, the spectroscopic response, $\mathbf{R}$ ($I \times J$) is a function of time, t, and spectral wavelength, l. A mixture of K components gives a response, $\mathbf{R}$:

$$r_{i,j}(t,l) = \sum_{k=1}^{K} \mathbf{c}_k(t)\mathbf{s}_k(l) \tag{1}$$

where $\mathbf{c}_k(t)$ is the concentration of component k at time, t, and $\mathbf{s}_k(l)$ is the spectral response of component, k, for wavelength, l.

In this paper, a number of calibration free techniques are investigated including VARIMAX, ITTFA, EFA, FSWEFA and SIMPLISMA. These are compared with Independent Component Analysis (ICA). ICA performs a similar function to the calibration free techniques. It is a separation method that has been applied in speech, biomedical signal processing, financial time series, wireless communications and image feature extraction.

2. SYNTHETIC DATA SET

A synthetic data set was generated from an isothermal batch reaction:

$$A \xrightarrow{k_1} B \xrightarrow{k_2} C$$

where the reaction rates take the values, k_1=0.8, k_2=0.8. The reaction is defined by the following kinetic equations:

$$[A]_t = [A]_0 \exp(-k_1 t) \tag{2}$$

$$[B]_t = \frac{[A]_0 k_1}{k_2 - k_1}(\exp(-k_1 t) - \exp(-k_2 t)) \tag{3}$$

$$[C]_t = [A]_0 - [A]_t - [B]_t \tag{4}$$

where $[A]_0$ is the initial concentration of A, and $[A]_t$, $[B]_t$ and $[C]_t$ are the concentrations of A, B and C, respectively at time t.

Calibration free techniques offer a methodology to monitor a reaction to determine the kinetic profiles of A, B and C. In the case where the components A, B and C are unknown, calibration free techniques can help identify the spectral profiles. To resolve the data set, the following techniques a) PCA (PLS Toolbox) b) EFA (PLS Toolbox) c) EFA (Tauler's Toolbox) d)FW-EFA (Tauler's Toolbox) e) SIMPLISMA f) ITTFA were investigated.

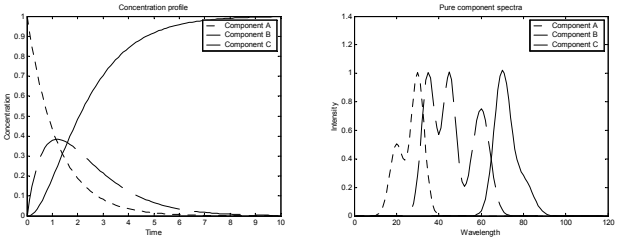


Fig. 1. Concentration profiles of A, B and C as defined by equations 2, 3 and 4.

Fig. 1 shows the concentration profiles and the pure spectra profiles for the three components of the reaction. It can be concluded that the concentration of component A, a reagent, reduces over time, while the concentration of B, an intermediate, increases and then slowly decreases and the concentration of component C, the final product increases with time. From the spectral profile, it can be observed that component A has two peaks at the 20th and 30th wavelength. Component B has three peaks at the 40th, 50th and 60th wavelength and component C has only one peak at the 70th wavelength. The concentration and pure component spectra profiles can be combined to produce a response matrix, **R**, equation 5 that defines absorbances for various wavelengths Fig. 2. This matrix is then used to reproduce the concentration and spectral profiles.

$$\mathbf{R} = \mathbf{C}^T \cdot \mathbf{S} \tag{5}$$

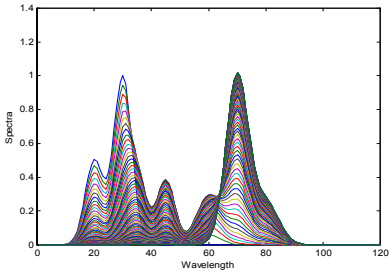


Fig. 2. Graphical representation of matrix **R**.

3. PRINCIPAL COMPONENT ANALYSIS

If the number of components in a reaction is unknown, a first estimate can be obtained through the application of Principal Component Analysis (PCA). A data matrix representing I observations on J variables can be decomposed into two matrices:

$$\mathbf{R} = \mathbf{T} \cdot \mathbf{W}^T \tag{6}$$

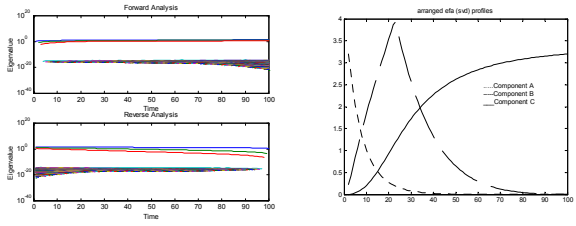
where **R** is the spectroscopic response, $(I \times J)$, **W** is the loadings matrix, ($J \times N$) and **T** is the scores matrix, ($I \times N$). For the specific reaction being considered, three components were selected from the application of PCA since the eigenvalue of the third component was still in excess of unity. This is in accord with the expected result.

4. EVOLVING FACTOR ANALYSIS

Evolving Factor Analysis (EFA) is based on the concept of sequential expanding windows, Keller and Massart (1992). A series of spectra from a reaction mixture, which contains a number of different absorbing species, are measured. As the order of the spectra in a chemical reaction provides

additional information, sub-matrices are formed by adding rows to an initial sub-matrix. By analysing the ranks of the data matrices as a function of the number of additional rows, time windows are derived. The number of species involved is equal to the number of significant eigenvalues of the second moment matrix. As new absorbing species start to become significant, new factors/eigenvalues evolve which explain the variability in the process.

EFA makes use of information in the time domain that for other approaches is ignored. In a reaction, the compound that appears first in the spectra should also be the first to disappear. Based on this concept, Tauler and Barcelo (1993) developed a technique to reconstruct the concentration profiles in reactions. For this technique, the compound windows are found by connecting the line of the compound that first appeared with the line of the last compound that appeared, both lines are then combined in a single figure from which the concentration windows are reconstructed. These profiles of the eigenvalues can be considered as a first estimate of the concentration profiles. EFA was applied to the synthetic data set, Fig. 3.



(a) PLS Toolbox (b) Tauler's Toolbox

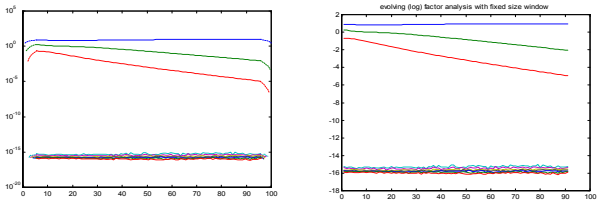
Fig 3. Results from application of EFA.

From the results of the application of the algorithm from the PLS Toolbox, it can be observed that the forward analysis indicates that 3 independent factors have evolved. One factor appears at the onset of the reaction, a second soon after the first and a third after the second. It is clear that these three factors correspond to the reagent, the intermediate and the final product. The backward analysis suggests that there is only one factor remaining at the end of the reaction with the two other factors disappearing. Once again the results confirm what is known about the reaction.

Application of EFA using the approach in Tauler's Toolbox, which involves a combined analysis of the data matrix, provides an initial estimate of the concentration profiles. However for this data set the concentration profile of the intermediate appears to have shifted from the baseline and the concentration at time point 22 is larger than expected.

5. FIXED WINDOW EVOLVING FACTOR ANALYSIS

A method that is similar to EFA is that of Fixed-Size Window Evolving Analysis (FSWEFA), Cuesta Sanchez *et al*, (1997). In FSWEFA the idea of the fixed-size window is introduced. A small 'window' of rows is selected that is moved over the data set. Analogous to the EFA plots, the eigenvalues of the fixed window (or their log) is plotted against analysis time. In some situations, it is possible to calculate the singular value decomposition at each window position and the associated values are plotted as a function of the window position. The main advantage of FWEFA over EFA is that it is able to detect low concentrations of impurities even at low separations. This is the situation in this example where the second eigenvalue corresponds to the impurity and the impurity under the main component can be localised.



(a) PLS Toolbox (b) Tauler's Toolbox

Fig. 4. Results from application of FSWEFA.

For the fixed window method both the approach described in the PLS Toolbox and Tauler's Toolbox were used with a fixed window size of 10 (Tauler, 2002). As can be seen from Fig. 4, both approaches give similar results, i.e. both identify three factors.

6. SIMPLISMA – PURE VARIABLES

SIMPLISMA is a method that identifies pure variables (Vandeginste *et al*, 2002; Gourvenec *et al*, 2002). It is based on the evaluation of the relative standard deviation of each column of data matrix, **R**. The idea is that a large relative standard deviation is indicative of high purity. Once the pure variables have been identified, the data set can be resolved into the pure spectra. The iterative algorithm for pure variables, for the SIMPLISMA method is as follows. Suppose that the inverse of the data matrix is represented by **D**, ($J \times I$) with elements $d_{j,i}$, where J is the number of variables and I is the number of spectra. First the length λ_j for variable j is calculated:

$$\lambda_j = \left(\left(\frac{1}{I} \right) \sum_{i=1}^{I} (d_{j,i})^2 \right)^{\frac{1}{2}} \qquad (7)$$

where $\lambda_j^2 = \mu_j^2 + \sigma_j^2$, μ_j is the mean of variable j, and σ_j is the standard deviation of variable j. The next step is the calculation of the first relative standard deviation (first purity) for variable j, $p_{j,1}$:

$$p_{j,1} = \frac{\sigma_j}{\mu_j} \tag{8}$$

For variables with low noise range intensity, problems can arise. This occurs because the value of μ_j approaches zero so the value of $p_{j,1}$ will be large. To address this, the purity and length are re-defined and a noise correction term is added. The next pure variable is then determined as the one most independent of the first pure variable. The data matrix is scaled by its length:

$$\delta_{j,i} = \frac{d_{j,i}}{\lambda_j} = \frac{d_{j,i}}{(\mu_j^2 + (\sigma_j + a)^2)^{1/2}} \tag{9}$$

where a is the correction factor for low intensity variables. The correlation about the origin matrix, $\mathbf{\Gamma}$ is defined as follows:

$$\mathbf{\Gamma} = \left(1/I\right)(\lambda)\mathbf{D}(\lambda)^T \tag{10}$$

and the determinant is calculated for variable i

$$\omega_{j,2} = \begin{vmatrix} \gamma_{j,j} & \gamma_{j,p_1} \\ \gamma_{p_1,j} & \gamma_{p_1,p_1} \end{vmatrix} \tag{11}$$

where the index p_1 represents the index for the first pure variable. The determinant is used as a weighting function and as a consequence the elements of the second purity spectrum become:

$$p_{i,2} = (\sigma_i / (\mu_i + a)) \cdot \omega_{i,2} \tag{12}$$

and the equation for the standard deviation spectrum is given by:

$$\sigma_{j,i}^s = \sigma_j \omega_{j,i} \tag{13}$$

For the general case, where j>2 the determinant is:

$$\omega_{j,i} = \begin{vmatrix} \gamma_{j,j} & \gamma_{j,p_1} & \cdots & \gamma_{j,pi-1} \\ \gamma_{p_1,j} & \gamma_{p_1,p_1} & \cdots & \gamma_{p_1,p_{i-1}} \\ \cdots & \cdots & \cdots & \cdots \\ \gamma_{p_{i-1},j} & \cdots & \cdots & \gamma_{p_{i-1},p_{i-1}} \end{vmatrix} \tag{14}$$

and in a similar manner to equation (8), the general formulation for the purity spectrum is $p_{j,i} = (\sigma_j / (\mu_j + a)) \cdot \omega_{j,i}$ and with the correction factor a included, the values for $\omega_{j,1}$ become:

$$\omega_{j,1} = \lambda_j^2 / (\mu_j^2 + (\sigma_j + a)^2) \tag{15}$$

For the identification of pure variables, the number of components was set to three and the noise allowed was 5%. The results can be seen in Fig. 5. The final profiles in Fig. 5 can be compared with the expected results in Fig. 1. After comparison, it can be concluded that the results are similar.

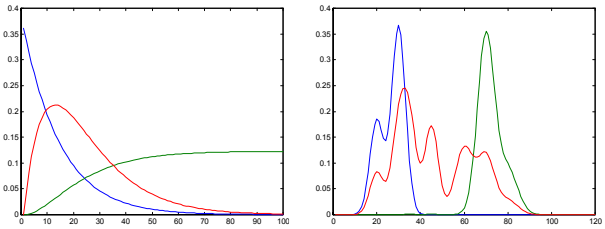


Fig. 5. Concentration profiles and spectral profiles extracted by SIMPLISMA.

7. VARIMAX AND ITTFA

In this section, an ITTFA algorithm in combination with VARIMAX is investigated. The principle on which ITTFA is based is that an initial target is defined and updated until specific criteria are satisfied, Vandeginste *et al*, 1998. The main criteria for success are that appropriate constraints are formulated for updating the targets with realistic initial targets being identified. Targets are adapted by replacing negative values that are produced in the estimated concentration and spectral profiles by zero. Thus for this application, non-negative constraints for the spectra and the concentration profiles are imposed. To select a target, different methods can be used for the initial profiles for each factor including VARIMAX rotation. VARIMAX rotation is based on the principle that the principal components axes can be rotated:

$$\mathbf{F} = \mathbf{V}^T\mathbf{O} \tag{16}$$

where the columns of **V** are the abstract factors of **R** that require to be rotated into real factors.

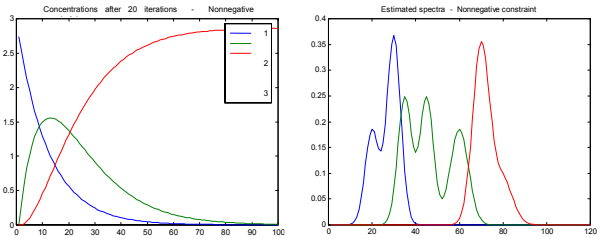


Fig. 6. Final estimates for the concentration and spectral profiles.

The matrix $\mathbf{V}^T$, is rotated by the orthogonal rotation matrix **O** so that the resulting matrix **F** fulfils the criterion that **F** has maximum row simplicity. A measure of simplicity of a vector is the variance of the square of the p elements that are to be maximised. After 20 iterations, the results for the specific reaction can be seen in Fig. 6. These results are compared to the expected results, Fig. 1. From the comparison, it can be concluded that the concentration and spectral profile plots appear to match the expected concentrations

8. INDEPENDENT COMPONENT ANALYSIS

An alternative calibration free resolution method that can be considered is Independent Component Analysis (ICA). ICA can be used to identify the spectral profile of each species in a mixture, i.e. identify the unknown components. ICA is a method designed to offer a solution to the Blind Source Separation problem, i.e. separate the source signals from the observations of their mixtures. ICA can be considered as an extension of PCA in that while PCA identifies principal components that are uncorrelated and that are linear combinations of the observed variables, ICA extracts components (IC's) that are independent and that constitute the observed variables, Hyvarinen *et al*, (2001).

Basically an ICA model is a "statistical latent variable model" in the sense that it describes how the observed data are generated by a process of mixing a number, n, of recorded signals θ. The signals θ are statistically mutually independent by definition and are called independent components (ICs). The basic problem is:

$$\eta_m = a_{m1}\theta_1 + a_{m2}\theta_2 + \ldots + a_{mn}\theta_n, \ \forall \ \ m=1,..,n \qquad (17)$$

where η_m are the observed random variables that are modeled as linear combinations of n random variables θ_m and the a_{ij}, $i, j = 1,\ldots,n$ are real coefficients that are assumed to be unknown. It is also assumed that each mixture η_m and each independent component θ_m are random variables and not time signals or time series. Equation 17 can be rewritten as:

$$\boldsymbol{\eta} = \mathbf{A}\boldsymbol{\theta} \qquad (18)$$

where $\boldsymbol{\eta}$ is a column random vector whose elements are η_m, i.e. if **R** is the data matrix, then n corresponds to each row of **R**, $\boldsymbol{\theta}$ is a column random vector whose elements are θ_m and **A** is a matrix with elements a_{ij}. The statistical estimation problem concentrates on two aspects, under what conditions can the model be estimated and what can be estimated. The answer is that the mixing coefficients a_{ij}, and the ICs, θ_m, must be estimated using the observed variables η_m. For simplicity it is assumed that $\boldsymbol{\eta}$ is a pre-whitened vector, i.e. all its components are uncorrelated and their variances are equal to unity. An alternative way to describe ICA is:

$$\hat{\boldsymbol{\theta}} = \mathbf{M}\boldsymbol{\eta} \qquad (19)$$

where $\hat{\boldsymbol{\theta}}$ is the estimate of $\boldsymbol{\theta}$, η_m is the observed random variable and **M** is a separating matrix which has to be estimated. Matrix **M** can be defined as the weight matrix of a two-layer feed-forward network where $\hat{\boldsymbol{\theta}}$ is the output and $\boldsymbol{\eta}$ is the input. The network is constrained to have statistically independent elements of $\hat{\boldsymbol{\theta}}$, i.e. they have non-Gaussian distributions. Non Gaussianity can be measured by either kurtosis or negentropy.

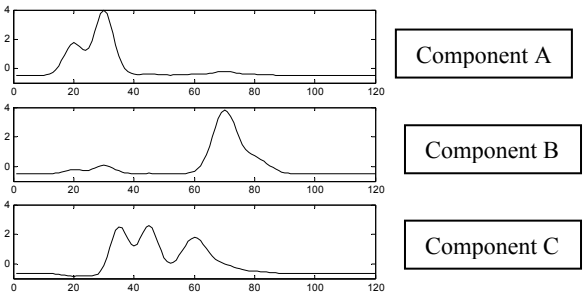


Fig. 7. Estimated spectral profiles using ICA.

The problem of spectral analysis in chemical mixtures represents a very similar problem to that of ICA since it is assumed in spectral analysis that the components of interest are strongly related to the data of the mixture through Beer Lambert's law. Hyvarinen and Oja (1997) have developed an algorithm, FastICA that is used in this paper for the separation of the spectral profiles. Non-gaussianity was a main characteristic of the spectral for this example. The results can be seen in Fig. 8.

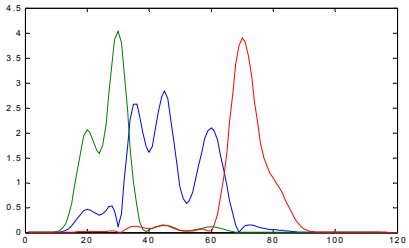


Fig. 8. Results of the application of JADE.

Another ICA algorithm that was also evaluated was the "Joint Approximate Diagonalization of Eigenmatrices" (JADE) (Cardoso, 1999). It is a cumulative-based batch algorithm for source separation. The results can be seen in Fig 8. ICA is

shown to be effective for the analysis of spectral data. The difference in scaling does not affect the qualitative information gained. The main peaks are situated where expected and the components are easily recognisable.

9. JADE AND MCR-ALS

Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) is a method developed by Tauler (2002). During the procedure, the initial estimates of the concentration profiles or the species spectra are given and then new concentration profiles are calculated by least-squares. In this application, the results from the JADE algorithm were used as an initial estimate of the spectral profiles. The results can be seen in Fig. 9. Compared with Fig. 8, the spectral profiles have clearly improved and the concentration profiles are also reproduced. The constraints of unimodality and non-negativity were imposed. Once the concentration profiles and the pure spectra became stable, the resulting data matrix was resolved.

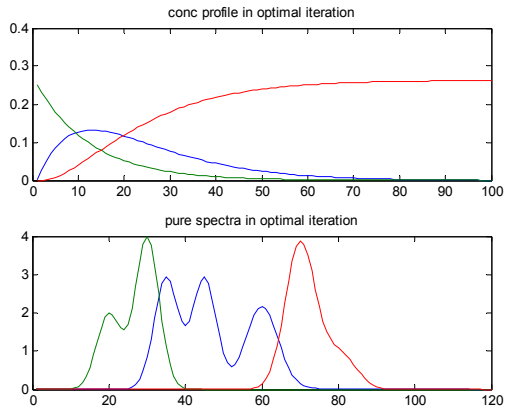


Fig. 9. MCR-ALS with initial estimate by JADE.

10. CONCLUSIONS

A number of calibration-free resolution techniques have been presented. The application of these techniques to an artificially generated spectral data set has demonstrated that they are all effective in terms of its resolution. In addition ICA i.e. both FastICA and JADE can be regarded as another method for the resolution of chemical mixtures. The combination of MCR-ALS and JADE also gave good results. Although the chemical mixture described in this application is simple, ICA has shown that unknown components in a mixture can be identified by the spectra of separated independent components. As an analyst will typically know the range of possible co-existing species in an analytical sample but not the exact number and identities, ICA could prove to be an effective technique. A further advantage of ICA is that it enables the implementation of the resolution of data in limited time. Furthermore ICA could be applied in process monitoring and control and this area is now being considered.

11. ACKNOWLEDGEMENTS

The author would like to acknowledge the EPSRC project KNOWHOW (GR/R/938010) and the EU project BATCHPRO (HPRN-CT-2000-0039) for financial support.

12. REFERENCES

Cardoso, J.F. (1999). High-order contrasts for independent component analysis. *Neural Computation*. **11(1)**, pp. 157-192.

Cuesta Sanchez, F., S.C. Rutan, M.D. Gil Garsia and D.L. Massart (1997). Resolution of multi-component overlapped peaks by the orthogonal projection approach, evolving factor analysis and window factor analysis. *Chemometrics and Intelligent Laboratory Systems*, **36**, pp. 153-164.

Gourvenec, S., D.L. Massart and D.N. Rutledge (2002). Determination of the number of components during mixture analysis using the Durbin-Watson criterion in the orthogonal projection approach and in the SIMLe-to-use interactive self-modeling mixture analysis approach. *Chemometrics and Intelligent Laboratory Systems*, **61**, pp. 51-61.

Hyvarinen, A. and E. Oja (1997). A fast-point algorithm for independent component analysis. *Neural Computation*, **9**, pp. 1483-1492

Hyvarinen, A., J. Karhunen and E. Oja (2001). *Independent component analysis. Adaptive and learning systems for signal processing, communications and contro*l, ed. S. Haykin. John Wiley & Sons.

Keller, H.R. and D.L. Massart (1992). Evolving factor analysis. *Chemometrics and Intelligent Laboratory Systems* **12(3)**, pp. 209-224.

Muller A. and D. Steele (1990). On the extraction of spectra of components from spectra of mixtures. A development in factor theory. *Spectrochimica Acta*. **46(5)** pp. 817-842.

Tauler, R. and D. Barcelo (1993). Multivariate curve resolution applied to liquid chromatography-diode array detection. *Trends in Analytical Chemistry*. **12(8)**, pp. 319-327.

Tauler, R., *MCR-ALS*. 2002, http://www.ub.es/gesq/mcr/theory.htm, http://www.ub.es/gesq/roma/roma.htm

Vandeginste, B.G.M., D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi and J. Smeyers-Verbeke (1998). *Handbook of Chemometrics and Qualimetrics: Part B*. Elsevier.