APPLICATION OF INDEPENDENT COMPONENT ANALYSIS TO CHEMICAL REACTIONS

S.Triadaphillou, 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

Sophia.Triadaphillou@ncl.ac.uk, julian.morris@ncl.ac.uk, e.b.martin@ncl.ac.uk

ABSTRACT

The analysis of kinetic data monitored using spectroscopic techniques and its resolution into its unknown components is described. Independent Component Analysis (ICA) can be considered a calibration free technique with the outcome of the analyses being the spectral profiles of the unknown species. This enables the realisation of qualitative information concerning the identification of the number and type of components present within the reaction mixture over time. The ICA approaches of FastICA and JADE and the calibration free technique of multivariate curve resolution-alternating least squares were applied to the mixture spectra of a first order synthetic reaction. For all approaches the signal was successfully separated from the constituent components.

1. INTRODUCTION

Reaction monitoring is a major challenge across the process industries. This form of monitoring typically involves the measurement and prediction of the concentration of a number of components in a chemical reaction and the determination of the number of components in the mixture. Furthermore the qualitative information extracted from the spectral analysis is of importance in terms of process understanding.

Traditionally calibration models are built to predict the property of interest. However the success of the approach depends on a number of factors. Calibration modelling is time consuming with the final model being sensitive to changes in process conditions. In addition it only provides quantitative information about the property of interest with no information about side reactions and intermediates [1]. Moreover in industrial processes a number of unknown components may be present in the mixture because of operational disturbances or a lack of detailed understanding of the reaction mechanisms. Thus by-products may be produced that are unknown and information on their existence cannot be determined using existing spectral interpretation methods. There is consequently a need for more advanced methods for the resolution of mixtures of chemical reactions.

The paper investigates a number of calibration free methods for the resolution of mixtures. A simulated data set from a first order reaction was generated for different reaction rates. Techniques included Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) [2] and Independent Component Analysis (ICA) [3]. Specifically two ICA algorithms were investigated, FastICA [4] and Joint Approximate Diagonalization of Eigenmatrices (JADE) [5] with the results of JADE being used as an initial estimate in the MCR-ALS algorithm. Applications of ICA have previously been reported in the areas of voice and sound separation, biomedical signal processing, financial time series, wireless communications and image feature extraction.

2. DETERMINATION OF THE NUMBER OF COMPONENTS

The first step is the estimation of the number of components in the mixture. This factor is related to the amount of variance present as a result of sources such as noise, background and baseline changes. One method that has been applied for the identification of the number of latent variables [6] is Principal Component Analysis (PCA). PCA has been widely applied for the decomposition of the covariance matrix. A data matrix representing I observations on J variables can be decomposed as follows:

$$\mathbf{D} = \mathbf{T} \cdot \mathbf{V}^{\mathrm{T}} \qquad (1)$$

where **D** is the spectral response, $(I \times J)$, **V** is the loadings matrix, $(J \times K)$ and **T** is the scores matrix, $(I \times K)$. Each principal component is a linear combination of the *J* variables and accounts for the main sources of variation. The current principal component is mutually orthogonal to the set of principal components previously calculated.

Evolving Factor Analysis (EFA) [7] is an alternative approach to the estimation of the number of components in a mixture. It uses the idea of the sequential expanding window. A series of spectra from a reaction mixture are measured for different wavelengths and these are arranged into a data matrix. The order of the spectra provides additional information on the behaviour of a chemical reaction. This is taken into account through the formation of sub-matrices by adding rows to an initial top sub-matrix, top down, or by adding rows to an initial bottom sub-matrix, bottom up. The rank of the matrix increases every time a row is added and by analysing these ranks as a function of the number of added rows, time widows are derived where a specific number of principal components are present. This is the first step in EFA. The number of species involved is equal to the number of significant eigenvalues of the second moment matrix. Thus as new absorbing species start to become significant, new factors/eigenvalues evolve. Hence as a new component elutes in the overlapping peak, the presence of an additional eigenvalue is required to explain this variability. EFA thus takes advantage of information, which is sometimes unused in the time domain. In a reaction, the compound that appears first in the spectra should also be the first to disappear.

3. MULTIVARIATE CURVE RESOLUTION-ALTERNATE LEAST SQUARES (MCR-ALS)

MCR-ALS is an iterative resolution method developed by Tauler. It has been applied to mixture dynamic processes that have been monitored through spectroscopic techniques and also to other chemical data whose instrumental responses obey Beer-Lambert law. Beer's law states that the spectral response of the components is independent of time and concentration [8]. In the case of reaction monitoring, the spectroscopic response, **R**, is a function of two variables, the time, t, and the spectral wavelength, l. As a result, a mixture of K components gives a response:

$$\mathbf{R}(t,l) = \sum_{i=1}^{K} \mathbf{c}_i(t)\mathbf{s}_i(l) \tag{2}$$

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

An advantage of this method is that knowledge of the concentration of all the compounds at the beginning of the kinetic process is not required. Tauler et al [9] developed a technique to reconstruct the concentration profiles in reactions. For this approach, the compound windows were found by connecting the line of the compound that first appears with the line of the last compound that appears. Both lines are combined in a single figure from which the concentration windows can be reconstructed. These profiles of the eigenvalues can be considered as a first rough estimate of the concentration profiles.

Based on the Beer Lambert law, the aim of MCR-ALS is the optimal decomposition of the data matrix **D** into the product of two smaller matrices, **C**, that relates to the concentrations and $\mathbf{S}^T$ that denotes the spectral profiles:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \tag{3}$$

E is the error-related matrix. The decomposition of data matrix **D** is performed by iterative optimization. The error in the raw data set is minimised using the following equations under suitable constraints for **C** and $\mathbf{S}^T$:

$$\min_{\mathbf{S}^T} \left\| \mathbf{D} - \mathbf{C}\mathbf{S}^T \right\| \tag{4}$$

$$\min_{\mathbf{C}} \left\| \mathbf{D} - \mathbf{C}\mathbf{S}^T \right\| \tag{5}$$

A description of the MCR-ALS algorithm can be found in the work of De Juan et al, [10], [11]

4. INDEPENDENT COMPONENT ANALYSIS

An alternative calibration free resolution method that is considered is Independent Component Analysis (ICA). ICA can be used to identify the spectral profile of each species in a mixture, i.e. to 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

mixture observations. ICA can be considered as an extension to PCA in that while PCA finds principal components that are uncorrelated and that are linear combinations of the observed variables, ICA is designed to extract components that are independent and that constitute the observed variables.
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 the recorded signals, s_i. The signals s_i are statistically mutually independent by definition and are called independent components. The basic problem is:

$$d_i = a_{i1}s_1 + a_{i2}s_2 + \ldots + a_{in}s_n \quad \forall \quad i = 1,\ldots,n \qquad (6)$$

where d_i are the observed random variables that are modeled as a linear combination of n random variables, s_i, and $a_{ij}, i, j = 1,\ldots,n$ are real coefficients that are assumed to be unknown. It is also assumed that each mixture d_i and each independent component s_i are random variables and not true time signals or time series. Equation 6 can be rewritten in vector matrix notation:

$$\mathbf{d} = \mathbf{A}\mathbf{s} \qquad (7)$$

where $\mathbf{d}$ is a column random vector whose elements are d_i, $\mathbf{s}$ is a column random vector whose elements are s_i and $\mathbf{A}$ is a matrix with elements a_{ij}.
The statistical estimation problem concentrates on two aspects, first under what conditions can the model be estimated and secondly what can be estimated. In practice both the mixing coefficients, a_{ij}, and the independent components, s_i, could be estimated using the observed variables d_i. For simplicity it is assumed that $\mathbf{d}$ is a pre-whitened vector, i.e. all its components are uncorrelated and their variances are equal to unity. An alternative way to express the ICA model is:

$$\hat{\mathbf{s}} = \mathbf{W}\mathbf{d} \qquad (8)$$

where $\hat{\mathbf{s}}$ is the estimate of $\mathbf{s}$, d_i is the observed random variables and $\mathbf{W}$ is a separating matrix which has to be estimated. $\mathbf{W}$ can be defined as the weight matrix of a two-layered feed-forward network where $\hat{\mathbf{s}}$ is the output and d is the input. The network is constrained to have statistically independent elements of $\hat{\mathbf{s}}$, i.e. they have non-Gaussian distributions. Non-gaussianity can be measured by either kurtosis or negentropy.
Two ICA methodologies were evaluated, FastICA and Joint Approximate Diagonalization of Eigenmatrices (JADE). JADE is a cumulative-based batch algorithm for source separation. Specifically it is a method that uses higher-order cumulative tensors that are a generalisation of the covariance matrix. For this family of methods, the fourth-order cumulative tensor is used to make the fourth-order cumulants zero or as small as possible. JADE computes the eigenvalue decomposition of a symmetric matrix.

5. EXPERIMENTAL GENERATING THE SYNTHETIC DATA SET

A first-order synthetic data set was generated by considering a starting material A to which a reagent was added. As a result A is converted to product B with a specific rate constant, k_1, and B is converted to product C with reaction rate, k_2, i.e. $A \xrightarrow{K_1} B \xrightarrow{K_2} C$.
Several experiments were performed for different rate constants. For the data set generated from the first order synthetic reaction, the reaction rates took the values:

k_1=0.8	k_2=0.8	Experiment 1
k_1=0.8	k_2=0.08	Experiment II
k_1=0.8	k_2=0.008	Experiment III

The data was obtained from the following kinetic equations:

$$[A]_i = [A]_0 \exp(-k_1 t_i) \qquad (9)$$

$$[B]_i = \frac{[A]_0 k_1}{k_2 - k_1}(\exp(-k_1 t_i) - \exp(-k_2 t_i)) \qquad (10)$$

$$[C]_i = [A]_0 - [A]_i - [B]_i \qquad (11)$$

where $[A]_0$ is the initial concentration of A, and $[A]_i$, $[B]_i$ and $[C]_i$ are the concentrations of A, B and C respectively at time point i.
Fig. 2 shows the concentration profiles of the three components for the reaction for experiments I, II and III respectively. The spectral profiles for the components are the same for all experiments since the same data was used in all three experiments and are presented in Fig. 1 (lower right hand-side).
Based on the data decomposition, it is expected that the pure spectra of the components should be the

same for all experiments, although the concentration profiles in the different experiments need not have a common shape. The varying shape of the kinetic profiles in the different experiments can be due either to different underlying kinetic models, different initial concentrations or to different reaction constants. In this case the different reaction constants produce the different concentration profiles for each experiment.

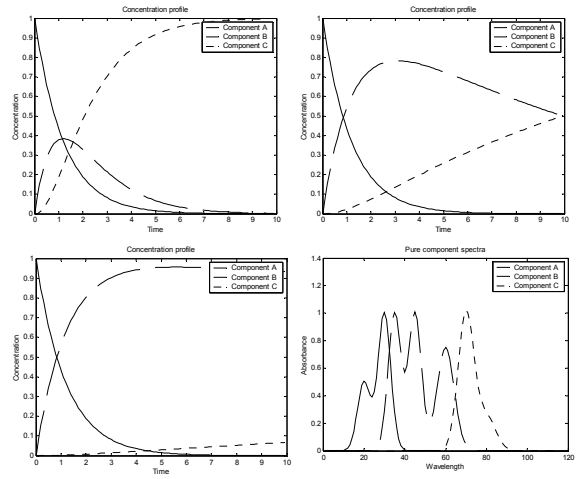

Fig. 1: Concentration profiles of A, B and C for experiment I, II and III respectfully as defined by equations 9, 10 and 11

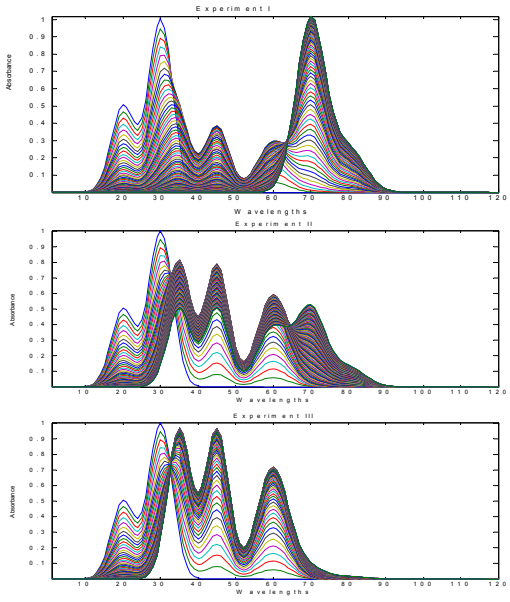

Fig. 2: Graphical representation of matrix D

The concentration and pure component spectra profiles can be combined to produce a matrix, **D** that consists of a number of spectra for different wavelengths. The plots of this matrix can be seen in Fig. 2.

$$\mathbf{D} = \text{conc}^{T} \cdot \text{spectra} \tag{12}$$

The matrix **D** is then used to resolve the spectral profiles of each species. The aim of this paper is to show that although different experiments are performed, in each case the results are the same and the components are identified successfully.

6. RESULTS AND DISCUSSION

After creating the experimental data sets, PCA was initially applied for the estimation of the number of components. For the specific reaction being considered, and for all three experiments, three components were selected since the eigenvalue of the third component was still in excess of unity. This is in accord with the expected result.

EFA was also used to estimate the number of components and to define the reaction process. The results of the first experiment are plotted in Fig.3. Similar results were obtained for the other two experiments. It can be observed that the forward analysis indicates that three 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 other two factors disappearing. Once again the results confirm what is known about the reaction.

The next step was to define the spectral profiles for each of the experiments. 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. FastICA was used for the separation of the spectral profiles (Fig.4). ICA was performed with 'tanh' non-linearity and the independent components were estimated in parallel by FastICA. This approach is similar to the maximum likelihood estimation for supergaussian data. The results from the application of the JADE algorithm can be seen in Fig. 5.

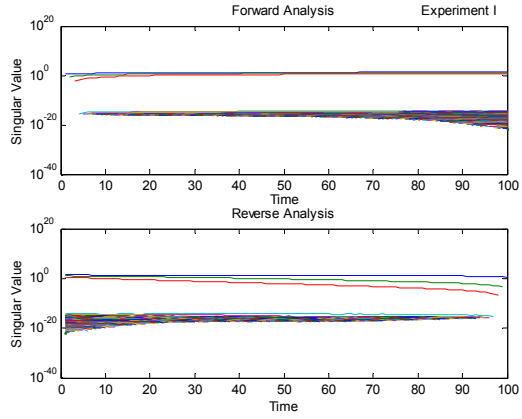

Fig. 3: EFA plots for the first experiment. The thick line defines the noise level and the solid lines above the noise level show the number and possible location of components appearing and disappearing during the kinetic process.

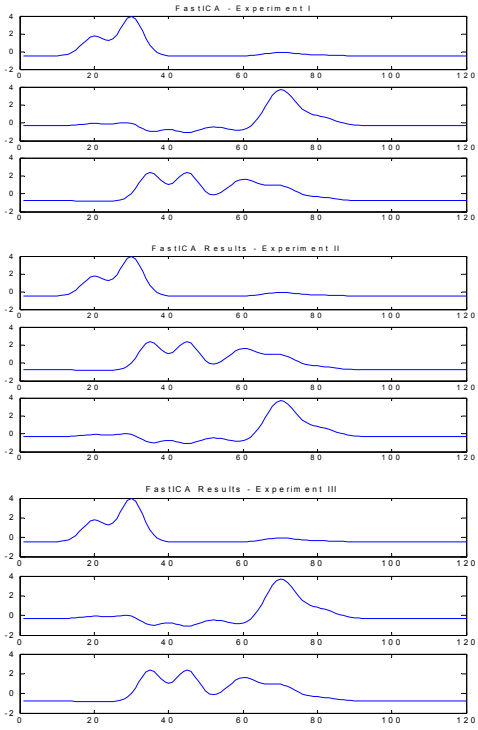

Fig. 4. Estimated spectral profiles by the FastICA algorithm for all experiments

These results indicate that ICA is very 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 componets are easily recognisable.

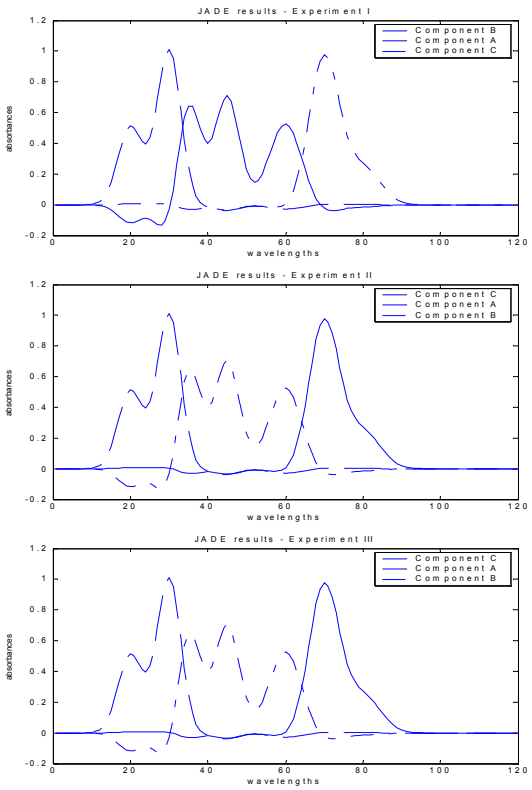

Fig. 5: Estimated spectral profiles from the JADE algorithm for all the experiments

Although the results appear almost perfect, further improvement can be achieved. This can be done by applying constraints to the data. As mentioned before MCR-ALS is a method used for the improvement of an initial estimate. During the procedure, the initial estimates of the concentration profiles, or of the species spectra, are given and new concentration profiles are calculated by least-squares. Normally for MCR-ALS, the results from EFA are used. However in this application the results from the JADE algorithm were used as an initial estimate of the spectral profiles. A comparison of the results of this procedure for the three experiments and of what was expected can be seen in Fig 6. A number of constraints such as unimodality and non-negativity were also imposed. Once the concentration profiles and the pure spectra become stable, the resulting data matrix can be resolved.

From Fig.6 it can be observed that the results are extremely promising. In all three experiments, the spectra were resolved identically, showing that the different reaction rates and mixture spectra do not influence the hidden information relating to the identification of the species. Component A had an

overlap between what was expected and what was estimated. Component's C predicted and real values were almost identical. Finally component's B estimated peaks were situated at the same wavelengths as the real peaks have to be situated making component B easily recognisable.

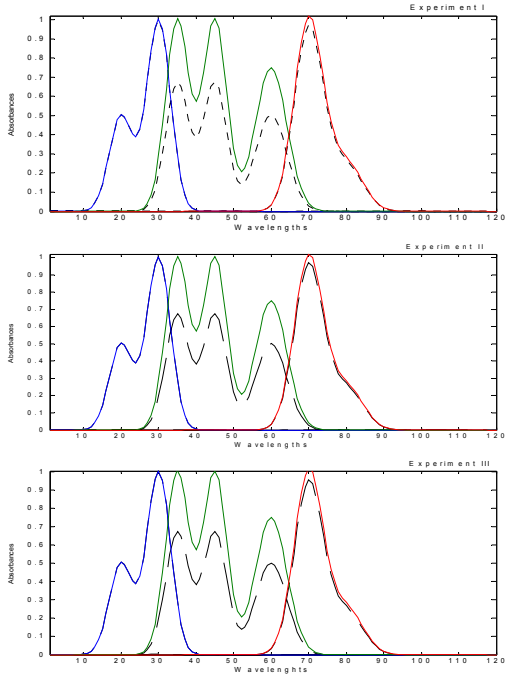

Fig. 6: Results of MCR-ALS with an initial estimate by JADE. Solid lines are the true profiles. Dotted lines are the estimated profiles

7. CONCLUSIONS

The ability of ICA to handle component spectra was examined. The application of ICA to an artificially generated spectral data set for different reaction rates has demonstrated that it is an effective approach to its resolution. Both FastICA and JADE can be regarded as another method for the resolution of chemical mixtures since they both extracted recognizable spectra. The combination of MCR-ALS and JADE also gave good results. ICA has shown that unknown components in a mixture can be identified by the spectra of separated independent components. A further advantage of ICA is that it enables the implementation of the resolution of data in limited time. ICA can also be applied in process monitoring and control. This area is now under consideration.

8. ACKNOWLEDGEMENTS

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

9. REFERENCES

1. Vandeginste, B.G.M., Massart, D.L, Buydens, L.M.C., Jong S.De, Lewi, P.J., Smeyers-Verbeke, J., *Handbook of Chemometrics and Qualimetrics: Part B*. 1998: Elsevier.
2. Tauler, R., *MCR-ALS*. 2002, http://www.ub.es/gesq/mcr/theory.htm, http://www.ub.es/gesq/roma/roma.htm
3. Hyvarinen, A., Karhunen, J., Oja, E., *Independent Component Analysis*. Adaptive and Learning Systems for Signal Processing, Communications and Control, ed. S. Haykin. 2001: John Willey and Sons.
4. Hyvarinen, A., Oja, E., *A Fast-point algorithm for Independent Component Analysis.* Neural Computation, 1997. 9: p. 1483-1492
5. Cardoso, J.-F., *High-order contrasts for independent component analysis.* Neural Computation, 1999. 11(1): p. 157-192
6. Miller, J.C., Miller, J.N., *Statistics for Analytical Chemistry*. Third ed. 1993: Ellis Horwood Ltd.
7. Keller, H.R., Massart, D.L., *Evolving factor analysis.* Chemometrics and Intelligent Laboratory Systems, 1992. 12(3): p. 209-224.
8. Muller A., Steele, D., *On the Extraction of Spectra of Components from Spectra of Mixtures. A Development in Factor Theory.* Spectrochimica Acta, 1990. 46(5): p. 817-842.
9. Tauler, R., Barcelo, D., *Multivariate Curve Resolution Applied to Liquid Chromatography-diode Array Detection.* Trends in Analytical Chemistry, 1993. 12(8): p. 319-327.
10. De Juan, A., Maeder, M., Martinez, M., Tauler, R., *Application of a Novel Resolution Approach Combining Soft- and Hard-modelling Features to Investigate Temperature-dependent Kinetic Processes*, Analytica Chimica Acta, 442, p. 337-350, 2001
11. De Juan, A., Maeder, M., Martinez, M., Tauler, R., *Combining Hard- and Soft-Modelling to Solve Kinetic Problems*, Chemometrics and Intelligent Laboratory Systems, 54: p. 123-141, 2000.