Resolution of Two-way Data from Spectroscopic Monitoring of Reaction Systems by Parallel Vector Analysis (PVA) and Window Factor Analysis (WFA) with Inspection of the Effect of Mass Balance

JIAN-HUI JIANG¹ and YUKIHIGO OZAKI, ¹†

¹† Department of Chemistry, School of Science, Kwansei-Gakuin University, Uegahara, Nishinomiya 662-8501, Japan (E-mail: ozaki@kwansei.ac.jp)

This paper reviews recent progress in self-modeling curve resolution of two-way data from spectroscopic monitoring of reaction systems. Particular emphasis is put on the effect of mass balance on the resolution of two-way data. Two slightly modified principal component analysis procedures are introduced for extracting spectral subspaces and determining the number of reactions in time domains. A new resolution method, parallel vector analysis (PVA), is presented.

(Received on August 10, 2001; Accepted on September 13, 2001)

Resolution of the two-way signals from multi-component mixtures into factors for single species is a useful tool for exploring the behavior in complex chemical systems. Potential applications include chromatography hyphenated with multi-channel detection, multivariate monitoring of kinetic reactions or processes as well as multivariate measurements of perturbation-dependent equilibrium systems. A variety of chemometric techniques have been proposed to resolve this kind of two-way data. These techniques, in principle, require no prior knowledge concerning the data. The only assumption is a certain bilinear model for the data, which is generally satisfied for practical multi-channel measurements on mixtures with varying compositions.

In terms of the uniqueness of the solution, two-way resolution techniques mainly fall into two types. Techniques of the first type aim at finding a unique resolution in which the factors for single species are uniquely defined according to the mathematical principle involved. Provided the mathematical principle is compatible with the chemical or physical model for the system under study, the unique solution is generally consistent with the true profiles. Routine techniques for unique resolution comprise evolving factor analysis (EFA),¹ window factor analysis (WFA),² heuristic evolving latent projection (HELP),³ subwindow factor analysis (SFA)⁴ and parallel vector analysis (PVA).⁵ To combat the so-called rotational ambiguity intrinsic in two-way resolution, these techniques have to inspect the evolving behavior of the components using some evolutionary rank analysis methods and utilize certain feature regions such as selective regions or zero-concentration regions. In the setting of hyphenated chromatography, conditions of unique resolution have been discussed.⁶

Techniques of the second type attempt to find a rational resolution in which the factors for single species do not violate any a priori knowledge such as non-negativity and unimodality. In principle, rational resolution produces a set of feasible solutions. The accuracy of the solution depends on the correlation among the pure profiles underlying the two-way data, and the accuracy can be evaluated by computing the range of feasible solutions⁸ or resolving the solution using different starts. In situations where the correlation among the pure profiles is not very severe, rational resolution can yield solutions approximating well to the true profiles. Typical approaches for rational resolution consist of orthogonal projection analysis (OPA),⁹ simple to use interactive self-modeling mixture analysis (SIMPLISMA),¹⁰ iterative target transformation factor analysis (ITTPFA),¹¹-¹² positive matrix factorization (PMF),¹³ alternating least squares (ALS)¹⁴ and elementary matrix transformation-based procedure.¹⁵ These approaches are distinguished from each other in the start estimates and the iteration procedures. A good start approximating to the true profiles as well as possible used to yield improved resolution while an efficient optimization algorithm is expected to show fast convergence.

Most of these two-way resolution techniques have been proposed in the setting of hyphenated chromatography. A special advantage with the use of two-way resolution in hyphenated chromatography is that under optimized conditions, the separation capability of chromatography splits along the chromatographic direction the mixture system of complicated compositions into a series of subsystems only containing very few components such that the rotational ambiguity is removed largely and a unique resolution can thus be achieved.

Spectroscopic monitoring of reactions and processes is of great importance in reaction kinetics and industrial chemistry. The resulting data are usually analyzed using hard-modeling approaches in which the data are fitted to a model pre-specified for the reaction under investigation as well as multivariate calibration. In situations where no exact information about the model and the compositions of the system is available, these approaches will be exposed to a substantial risk of giving erroneous solutions. As a result, two-way resolution methods as a type of soft-modeling techniques have recently attracted increasing interest in the studies of chemical reactions and processes using multivariate spectroscopic monitoring. These methods guarantee a resolution robust to possible model deviations in cases where no exact information or no prior
knowledge is available concerning the reactions. Moreover, they enable one to investigate the reactions in ‘real-world’ surroundings where interferents may be present in the system. Nevertheless, most of these two-way resolution techniques have been proposed in the setting of hyphenated chromatography in which the concentration profiles are not constrained by mass balance. As far as the data from multivariate monitoring of kinetic reactions or processes are concerned, it is essential for the data analysts to take into account the particularity of reaction systems in which the mass balance always holds for the concentration profiles.

This paper reviews the effect of mass balance on the resolution of two-way data discussed in our previous study. Two slightly modified PCA procedures are discussed for the ascertainment of the number of reactions in different time domains and the extraction of spectral subspaces essential for resolution. A new two-way resolution approach, parallel vector analysis (PVA) is also reviewed. The idea of PVA is to construct a set of subspaces comprising only one common (spectral) component and then find a vector that is in parallel with a series of vectors coming from different subspaces. With suitably constructed subspaces, the PVA procedure offers a versatile approach to the unique resolution of spectral profiles. A special advantage of PVA is that one can selectively resolve the components of interest without regard to the global bilinear model and other interferents. The data obtained from an on-line Fourier-transform Raman spectroscopic monitoring of the anionic dispersion polymerization of styrene and 1,3-butadiene has been analysed using PVA and WFA. The results demonstrate that the PVA method and the modified PCA procedures are useful approaches for the resolution of two-way data from multi-channel spectroscopic monitoring of reactions.

Theory

Chemical Rank Estimation in Reaction Systems

Suppose a system in which K independent reactions or processes take place is monitored at J spectroscopic channels at I time moments, and the resulting data matrix can be described by the following bilinear model

\[
X = CS + E
\]

where \(X\) is the \(I \times J\) matrix of measurement data, \(C\) designates the \(I \times N\) matrix of concentration profiles for the \(N\) spectrally active species in the system, \(S\) is the \(J \times N\) matrix of spectroscopic profiles for the \(N\) species, the superscript \(^T\) symbolizes matrix or vector transposition, and \(E\) is the matrix of measurement errors. In most applications it is assumed that both \(S\) and \(C\) have full column rank. Then one immediately has the conclusion that the chemical rank or pseudo-rank of measurement data \(X\) equals the number of components \(N\) in the system. Actually, this conclusion is the theoretical foundation for determining the number of components in a system and is generally true in situations where the full-rank assumption holds for both \(S\) and \(C\). However, as far as a system containing \(K\) reactions is concerned, it is known that the mass balance condition must be kept for each reaction. That is, the net variations of concentrations for every reactant and product in one independent reaction must be proportional to each other. Such kind of proportional variations in concentrations will cause a concentration profile matrix with deficient column rank.

Suppose the stoichiometry of the \(K\) reactions in the system under study are symbolized by the following balanced chemical equations:

\[
a_{12} R_{12} + a_{23} R_{23} + \ldots = b_{11} P_{11} + b_{21} P_{21} + \ldots \quad (k = 1, 2, \ldots, K)
\]

where \(a_{ij}\) and \(b_{ij}\) denote the number of moles of reactants \(R_{ij}\) and \(R_{k}\) that participate in the \(k\)th reaction to yield \(P_{ij}\) and \(P_{k}\) moles of products \(P_{ij}\) and \(P_{k}\). One can obtain the extent of the \(k\)th reaction at the \(i\)th moment is

\[
\xi_{ij} = -\frac{1}{a_{ij}} \Delta c_{R_{ij}}, \quad \xi_{ik} = -\frac{1}{a_{ik}} \Delta c_{R_{ik}} \quad (k = 1, 2, \ldots, K)
\]

where \(\Delta c_{R_{ij}}\) denotes the net variation of concentration between the \(i\)th moment and a specified zero time resulted from the \(k\)th reaction.

Therefore, the total concentration variation vector for species \(A\), \(\Delta c_A\), which may participate in several reactions in the system, is given by

\[
\Delta c_A = \sum_{k=1}^{K} w_{Ak} \xi_{ik}
\]

Then, the concentration profile of species \(A\), \(c_A\), is

\[
c_A = c_{A0} I + \sum_{k=1}^{K} w_{Ak} \xi_{ik}
\]

where \(c_{A0}\) is the initial concentration for species \(A\) and \(I_j\) is an \(I \times 1\) vector with all entries being one. Letting \(\Xi = (\xi_{11}, \xi_{21}, \ldots, \xi_{K1})\), \(u\) be an \(N \times 1\) vector with the \(n\)th element being the initial concentration of the \(n\)th spectroscopic-active species, and \(W\) be an \(N \times K\) matrix with the \((n, k)\)th \((n = 1, 2, \ldots, N; k = 1, 2, \ldots, K)\) entry being \(w_{nk}\), one can rewrite eq (8) in matrix notation as

\[
C = \Xi W^{T} + I_u u^{T}
\]

Based on eq (9), it can be reasoned that the rank of \(C\) is the smaller one of \(K + 1\) and \(N - M + F\). For \(F = 1\) if the initial concentration of a certain interferent is not zero, and \(F = 0\) otherwise.

Now, one can reach an important conclusion concerning the systems of reactions or processes. That is, the pseudo-rank of the measurement matrix \(X\) for a system containing \(K\) independent reactions or processes is the smaller one of \(K + 1\) and \(N - M + F\). Because the common practice always involves the case where \(N - M + F \geq K + 1\), one then obtains that in common situations the pseudo-rank of the measurement matrix for a system containing \(K\) reactions is \(K + 1\).

Since the measurement matrix \(X\) for a system containing \(K\) independent reactions only gives a chemical rank of \(K + 1\), one can convert the component-related rank-deficient bilinear model \((1)\) to a reaction-related full-rank bilinear model as follows:

\[
X = \Xi W + I_u x_0 \times E\]

where \(R = X W\) and \(x_0 = S u\). Obviously, \(x_0\) is the spectral response measured at the specified zero time. The \(k\)th column \(r_k\) of \(R\) is

\[
r_k = Sw_k = \sum_{s=1}^{S} w_{sk} s_s
\]
where \( w_k \) is the \( k \)-th column of \( W \), and \( s_i \) is the pure spectra of the \( i \)-th species. One can conclude that \( r_i \) is actually the spectral variation purely caused by the \( i \)-th reaction when this reaction progresses by a unit extent. Because \( r_i \) is unique for each reaction, we will refer to \( r_i \) as the pure spectra of the \( i \)-th reaction. The pure spectra are a linear combination of the pure spectra of the species participating in the reaction.

Extracting spectral subspaces and determining the number of reactions in time domains

In the systems comprising multiple reactions or processes, different reactions proceed actively in varied time domains. Consequently, in some time domains the products of reactions having completed will be present in the system as interfering species that are not involved in the reactions taking place in these time domains. The concentration profiles of these interferents are thereby extended to a very wide time domain such that many species are embedded by each other, inducing the problem that the resolution of these species cannot be obtained uniquely with standard two-way resolution techniques. The effect of such interferents can be removed using mean-centered principal component analysis (MCWPCA), which is a procedure for extracting the spectral subspaces for the active reactions.

Consider a small time window \( X \) that contains the \( i \)-th to the \((i + H - 1)\)-th spectra, i.e., \( X \) is an \( H \times \lambda \) submatrix of \( X \) which is composed of the \( i \)-th to the \((i + H - 1)\)-th rows of \( X \), then one can obtain based on eq (10) that \( X_i = \Xi_i \cdot \mathbf{I}_K + \mathbf{E}_i \) \( (11) \)

where \( \Xi_i \) is an \( H \times K \) matrix whose columns are the extent vectors of the \( K \) reactions taking place in the window with the zero time specified as the starting position of the window, \( R \) is a \( J \times K \) matrix with the \( k \)-th \((k = 1, 2, \ldots, K) \) column being the pure spectra of the \( k \)-th reaction in the time window, and \( x_i \) is the spectra at the starting position of the window, i.e., the transpose of the \( i \)-th row of \( X \). In MCWPCA, mean-centering is firstly applied to the rows of \( X_i \):

\[
P_B X_i = P_B \Xi_i + P_B E_i
\]

where \( P_B = (\mathbf{I}_H - \mathbf{L}_0^H / H) \) is the mean-centering operator. It is easy to show that the pseudo-rank of \( P_B X_i \) is \( K_i \), the number of reactions occurring in the time window. Secondly MCWPCA computes the singular value decomposition (SVD) of \( P_B X_i \) by

\[
X_i = U D V^T = P_B \Xi_i + P_B E_i
\]

where \( \Xi_i \) is an \( H \times K_i \) matrix whose columns are the extent vectors of the \( K_i \) reactions taking place in the time window \( X_i \), and the first \( K \) singular vectors \( V \) offer an estimate of the subspace of \( R_i \), the subspace generated by the pure spectra of the \( K \) reactions proceeding actively in the time window.

Based on the conclusion that MCWPCA indicates the number of reactions in the time windows, one can implement it in an evolving manner. This results in a slightly modified evolving factor analysis procedure, mean-centering window evolving factor analysis (MCWFEA). MCWFEA yields a number of lines of singular values, which is called local reaction map. The number of significant singular lines at different time domains indicates the number of reactions taking place in the corresponding time domains, thereby suggesting the starting point and the ending point for each reaction.

Resolution of spectral profiles by parallel vector analysis (PVA)

The PVA method is motivated by the theorem that, for \( M \) different subspaces \( A_1, A_2, ... \) and \( A_0 \) that are generated by a set of linear independent spectral vectors \( s_1, s_2, ..., s_n \), if \( s_i \) belongs to each subspace, and for any other spectral vector \( s (n = 2, 3, ..., N) \) these exist a subspace which does not contain \( s_n \), then, if there is a non-zero vector \( x \) belonging to all the subspaces, it must satisfy

\[ L (x, b_w) = \sum_{m=1}^{M} || A_m b_w - x ||^2 \]

where \( A_m (m = 1, 2, ..., M) \) is the matrix made up by the base vectors of the corresponding subspace, \( b_w \) is the combination coefficient vector to be optimized. Since the method attempts to seek vectors that come from different subspaces, but are in parallel with each other, this approach is then called as parallel vector analysis (PVA). It can be shown that the solution \( x \) is the first eigenvector (the eigenvector corresponding to the largest eigenvalue) for the following eigen-problem:

\[ \sum_{m=1}^{M} A_m A_m^T x = \lambda x \]

Notice that \( A_0 A_0^T \) is the projection matrix on the subspace \( A_0 \). If there is a vector common to all subspaces, the largest eigenvalue of

\[ \sum_{m=1}^{M} A_m A_m^T \]

is \( M \). Conversely, if one obtains the first eigenvector associated with an eigenvalue very close to \( M \), this eigenvector must belong to all subspaces. This conclusion is consistent with the motivation of PVA. Moreover, this conclusion reveals an important fact that one can evaluate the solution using the eigenvalues: If the largest eigenvalue is significantly smaller than \( M \), then there must not be a vector common to all subspace. If the second largest eigenvalue also shows insignificant difference between \( M \), then there must exist more than one vector belonging to all subspaces. That is, the solution is not unique.

Resolution of the extent curves of reactions by WFA

WFA is derived from another generic theorem for unique resolution that if the spectral subspace for all the interferents that have concentration profiles overlapping that of the analyte is available, then the concentration profile of the analyte is uniquely resolvable. Therefore, the fundamental principle of WFA to find a time window in which the subspace orthogonal to spectra of interfering species or reactions can be estimated. The implementation of WFA in the resolution of reaction systems is slightly different from the standard one, which is outlined as follows: First, a time window \( X_a \) where the reaction to be resolved is proceeding is defined. That is, the time window starts from the initiating point of the reaction and ends by the terminating point of the reaction. Second, one or several time domains \( X_i \) outside the time window \( X_a \) are chosen where the interfering reactions are in active progress. Third, the spectra in each time domain \( X_i \) are separately analyzed by MCWPCA to extract the spectral subspace \( V_i \) of reactions active in this time domain. Obviously, one can see that the augmented matrix \( X = (V_1, ...) \) gives an estimate of the spectral subspace for all the interfering reactions. Fourth, the spectra in the time window \( X_a \) are each subtracted by the first spectra of the window, i.e., all the rows of \( X_a \) are subtracted by the first row of \( X_a \). That is, \( Z_a = X_a - 1_1 x_1^T \), where \( x_1^T \) is the first row spectra of \( X_a \), and \( H \) is the number of rows in \( X_a \). Based on eq (11), one then has

\[ Z_a = X_a - 1_1 x_1^T = x_0 \cdot R_0^T + E_o \]

where \( E_o \) is the combination coefficient vector. Since the method attempts to seek vectors that come from different subspaces, but are in parallel with each other, this approach is then called as parallel vector analysis (PVA). It can be shown that the solution \( x \) is the first eigenvector (the eigenvector corresponding to the largest eigenvalue) for the following eigen-problem:

\[ \sum_{m=1}^{M} A_m A_m^T x = \lambda x \]

Multiplying \( Z_a \) by \((I_H - (V_i^T)^2)^T\) yields
eliminating the interference from concentration-stable reaction system. MCWPCA may be a useful approach for information provided by FSWEFA or MCWEFA.

After the extent curves of reactions have been resolved, the concentration profiles can be computed based on eq (8) and other information provided by FSWEFA or MCWEFA.

Conclusion

The review paper has reported several new methods for the resolution of two-way data from multivariate monitoring of reaction system. MCWPCA may be a useful approach for eliminating the interference from concentration-stable components, and PVA may be a very competitive procedure for self-modeling curve resolution of two-way data, especially in situations where the selective regions are not easily available in two directions. As a matter of fact, PVA and WFA are intrinsically two complementary methods: PVA is an approach for the resolution of spectral profiles, while WFA is a method for resolving the concentration profiles.

References