Mathematical chromatography solves the cocktail party effect in mixtures using 2D spectra and PARAFAC

Rasmus Bro, Nanna Viereck, Marianne Toft, Henrik Toft, Peter I. Hansen, Søren B. Engelsen

The *cocktail party effect* in analytical chemistry refers to the ability to measure a complex mixture of analytes and to resolve the signals from a single analyte. We describe a new approach that we have coined *mathematical chromatography*, where data from complex spectra are resolved into pure-analyte information by so-called three-way modeling. Three-way analysis requires the addition of an extra dimension to the signals being acquired. Here, as an example, mathematical chromatography is applied to a series of diffusion-edited 2D NMR spectra of mixtures of glucose, maltose and maltotriose to demonstrate that it is possible to identify and to resolve individual components in highly overlapping 2D NMR spectra.

Keywords: Analyte; Cocktail party effect; Complex mixture; Complex spectra; Mathematical chromatography; NMR spectroscopy; Nuclear magnetic resonance; PARAFAC; Three-way analysis; 2D spectra

Rasmus Bro,
Nanna Viereck*,
Marianne Toft,
Henrik Toft,
Peter I. Hansen,
Søren B. Engelsen
Department of Food Science,
Faculty of Life Sciences,
University of Copenhagen,
Rolighedsvej 30,
1958 Frederiksberg C,
Denmark

*Corresponding author.
Tel.: +45 3533 3502;
Fax: +45 3533 3245;
E-mail: nav@life.ku.dk

1. Introduction

Resolving complex mixture spectra into their pure-component contributions is a fundamental problem in analytical chemistry. We aim to demonstrate how it is possible to take advantage of what is termed the second-order advantage in multi-way chemometrics to resolve the spectra from a mixture into pure-analyte contributions. Pure-component resolution is an inherent property of PARAFAC (parallel factor analysis) [1] modeling of trilinear (or higher order) low-rank mixture data. In terms of spectroscopy, this translates to the requirement that the spectral intensity has to be recorded as a function of two independent variables (spectral landscape; two-dimensional (2D)) instead of the usual one (spectrum; one-dimensional (1D)).

We have previously shown how this principle applies to fluorescence excitation-emission-landscapes [2,3]. However, the principle is general and the second spectral direction does not need to have spectroscopic variables, but can be time-resolved spectroscopy, hyphenated spectroscopy or spectroscopy of systems undergoing another perturbation. In this article, we demonstrate the application of mathematical chromatography to diffusion-edited 2D nuclear magnetic resonance (NMR) spectra (DOSY – diffusion-ordered spectroscopy). We outline the mathematical principle behind the new method and demonstrate its successful use on a small mixture design. The new method, which originates from psychology, will revolutionize quantitative spectral analysis of complex samples and have immense possibilities in exploiting the explorative potential of multi-parametric NMR spectroscopy.

NMR spectroscopy has become an indispensable analytical technique (e.g., for characterization of complex biological samples). The versatility and high resolution of NMR is exploited in diverse applications (e.g., medical diagnostic body scanning, biomacromolecular structure determination and metabolite mapping in tissue and body fluids). The extraction of quantitative and qualitative information from NMR spectra of complex biological samples necessitates appropriate data-analysis tools. The first report on the use of

0165-9936/$ - see front matter © 2010 Elsevier Ltd. All rights reserved. doi:10.1016/j.trac.2010.01.008

multivariate data analysis in the form of principal component analysis (PCA) [4] on NMR spectra appeared in 1983 [5], but it was not before the advent of high-throughput metabonomics [6] that quantitative NMR became a focal point of international research. The future perspectives in phenotyping for personalized nutrition and personalized drug treatment have only just begun [7]. The mathematical basis for the metabonomic advances has been the use of multivariate models (e.g., PCA). However, these powerful data-analysis tools cannot provide direct information on the underlying chemistry, due to the inherent 'rotational ambiguity' explained below. This has important implications (e.g., in the search for biomarkers that requires causal interpretation to avoid spurious results).

2. Quantitative spectroscopy

In quantitative spectroscopy, a set of I samples is measured as a function of J spectral variables (e.g., chemical shifts in NMR spectroscopy). Each spectrum may be contained in one row of an $I \times J$ matrix, $\mathbf{X}$. If each spectrum is a sum of contributions from F underlying spectra (i.e. F chemical analytes) held in a matrix $\mathbf{B}$ ($J \times F$), then $\mathbf{X}$ can be modeled as a bilinear model as the product of the sample-specific concentrations, a_{if} (held in $\mathbf{A}$ of size $I \times F$) and the pure-analyte spectra [Equation (1)]:

$$\mathbf{X} = \mathbf{A}\mathbf{B}^{\mathbf{T}} + \mathbf{E} \quad (1)$$

where $\mathbf{E}$ contains measurement noise. This model may also be written in scalar notation, which will be useful in the following [Equation (2)]:

$$X_{ij} = \sum_{f=1}^{F} a_{if} b_{ij} + e_{ij} \quad (2)$$

The problem, in practice, is to determine $\mathbf{A}$ and $\mathbf{B}$ from the measured $\mathbf{X}$. A natural proposition is to estimate these parameters from $\mathbf{X}$ using a least squares or similar fitting procedure. However, because [Equation (3)]:

$$\hat{\mathbf{X}} = \mathbf{A}\mathbf{B}^{\mathbf{T}} = \mathbf{A}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{B}^{\mathbf{T}} = \mathbf{C}\mathbf{D}^{\mathbf{T}} \quad (3)$$

for any non-singular $F \times F$ matrix, $\mathbf{Q}$, then any estimate may equally well lead to estimates of $\mathbf{C}$ and $\mathbf{D}$ instead of the true $\mathbf{A}$ and $\mathbf{B}$ (i.e. the solution may be any one of a class of alternative solutions due to this mathematical rotational ambiguity). As a result, bilinear models (e.g., PCA) cannot provide pure-analyte spectra.

In 1944, Cattell [8] described a fundamental way to solve the rotational indeterminacy. This idea, called parallel proportional profiles, became operational in 1970 when Harshman [1] developed the PARAFAC model. The principle in PARAFAC is to use several sets of matrices that provide information on the same basic entities (e.g., concentrations and spectral profiles) but in different proportions [9]. Such different proportions may arise by measuring the basic matrix data, $\mathbf{X}$, as a function of a new dependent spectral variable (e.g., analyte diffusion in NMR spectroscopy). The basic two-way data tables are thereby extended to three-way tables. A three-way table may be described as an $I \times J \times K$ array, $\mathbf{X}$, with typical elements x_{ijk} containing the measurement of sample i at variable j on occasion k. The PARAFAC model decomposes such three-way data into a set of trilinear terms and residuals. The algebraic model structure is [Equation (4)]:

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijf} \quad (4)$$

For a given choice of F, the number of chemical analytes, the three-way array is decomposed into so-called scores $\mathbf{A}$ ($I \times F$) with elements a_{if}, loadings $\mathbf{B}$ ($J \times F$) with elements b_{jf} and loadings $\mathbf{C}$ ($K \times F$) with elements c_{kf}. Unlike PCA, the PARAFAC solution is unique and directly estimates the underlying physical features of the data, if these approximately follow the model [10,11]. Hence, there is no rotational ambiguity.

Multivariate curve-resolution (MCR) is a two-way alternative to PCA that sometimes allows unique solutions similar to PARAFAC [12,13]. However, the model as such is not unique in general, so there is no guarantee that a meaningful solution can be obtained with MCR. For certain types of data, it is highly unlikely that unique solutions are obtained unless there are truly selective variables (i.e. distinct variables where each chemical analyte has a signal without any overlap from other analytes). Such prerequisites are not needed in PARAFAC. If the data set follows the model and the underlying profiles different for each analyte, the solution is unique.

3. 2D NMR spectroscopy

Using diffusion-edited 2D NMR spectroscopy, it is possible to obtain data that follow the PARAFAC model [14,15]. The signal intensity is recorded as a function of chemical shift and of the squared field-gradient strength following a low-rank trilinear structure [Equation (5)]:

$$I_{\Omega\delta k} = \sum_{f=1}^{F} S_{\Omega f} A_{\delta f} C_{kf} \quad (5)$$

where $S_{\Omega f}$ denotes the spectral intensity at chemical shift Ω for compound f, and where $A_{\delta f} = \exp[-Df\gamma^2\delta^2\Delta' - R_f]$ denotes the attenuation due to diffusion at gradient strength δ for compound f. Parameter C_{kf} denotes the concentration in sample k of compound f. When comparing the above model with the PARAFAC model, it follows that the diffusion-edited 2D NMR spectra can be