An Introduction to ChemoSpec2D

Bryan A. Hanson[a]

[a]Dept. of Chemistry & Biochemistry, DePauw University; hanson@depauw.edu

This version was compiled on November 30, 2018

A collection of functions for exploratory chemometrics of 2D spectroscopic data sets such as COSY and HSQC NMR spectra. `ChemoSpec2D` **deploys methods aimed primarily at classification of samples and the identification of spectral features which are important in distinguishing samples from each other. Each 2D spectrum (a matrix) is treated as the unit of observation, and thus the physical sample in the spectrometer corresponds to the sample from a statistical perspective. In addition to chemometric tools, a few tools are provided for plotting 2D spectra, but these are not intended to replace the functionality typically available on the spectrometer.** `ChemoSpec2D` **takes many of its cues from** `ChemoSpec` **and tries to create consistent graphical output and to be very user friendly.**

This vignette is based upon `ChemoSpec2D` version 0.2.0.

Background

`ChemoSpec2D` is designed to analyze 2D spectroscopic data such as COSY and HSQC NMR spectra using appropriate chemometric techniques. It deploys methods aimed primarily at classification of samples and the identification of spectral features which are important in distinguishing samples from each other. `ChemoSpec2D` stores and manipulates each spectrum as a matrix of data, and hence a data set is a collection of 2D spectra. Thus the entire data set is naturally visualized as a 3D array with dimensions:

$$\text{F2} \times \text{F1} \times \text{no. samples}$$

or

$$\text{2D Spectrum} \times \text{no. samples}$$

where F2 and F1 are NMR-speak for the x- and y-axes/dimensions. We will refer to this array as $\underline{\mathbf{X}}$. See Figure 1.

`ChemoSpec2D` treats each spectrum/matrix as the unit of observation, and thus the physical sample that went into the spectrometer corresponds to the sample from a statistical perspective. Keeping this natural unit intact during analysis is referred to as a *strong* multi-way analysis. In a weak analysis, the 3D data set is unfolded into a series of contiguous 2D matrices and analyzed using methods suitable for any 2D data set (such methods are fundamentally bilinear) (Huang *et al.* (2003)). In the weak approach, each slice of a 2D spectrum becomes just another 1D spectrum, and the relationship between the slices in a single 2D spectrum is lost. Oddly enough, the trilinear/strong analysis has fewer parameters to estimate so it is simpler, but computationally more demanding. The interpretation is also more straightforward. These techniques are akin to PCA, and seek to *reduce the data to a limited number components represented by scores and loadings*. Noise in the data set is reduced, and correlating variables are collapsed.

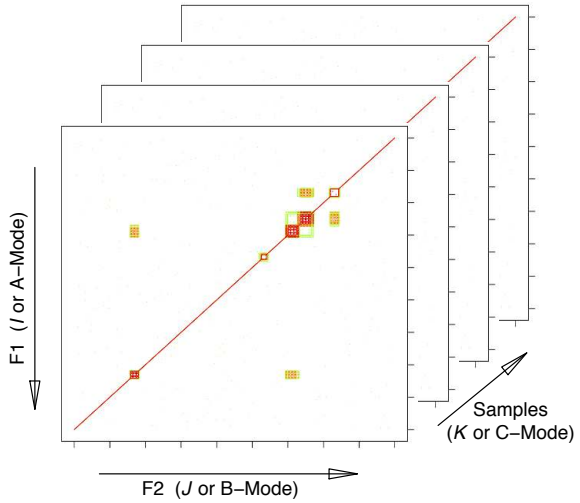

Fig. 1. Configuration of data array $\underline{\mathbf{X}}$. I, J and K are array indices; F2 and F1 are standard terms for NMR dimensions. The mode terminology is typically used in the PARAFAC literature. This structure is also sometimes referred to as a data cube.

The literature on the topics of chemometric analysis of 3D data sets uses a wide variety of terminology to describe analysis options, and the same mathematical analysis can be found under different guises in different fields. Here are some of the more relevant and frequently encountered terms:

- Multivariate Image Analysis (MIA). This term is typically used when the "images" are photographs for example, and the term covers a lot of ground, such as finding particular objects within a photograph. While 2D NMR spectra are typically plotted as contours, there is no reason why they cannot be plotted as an image or heat plot, which is essentially a photographic image. Classification using MIA methods involves a TUCKER1 analysis in which each image is the input, and the sample mode (only) is reduced to a requested number of components. Scores are the output. `ChemoSpec2D` can carry out MIA.
- TUCKER3. TUCKER3 is an operation in which all three dimensions of a 3D data set are reduced, and each dimension can have a different number of components. `ChemoSpec2D` does not carry out TUCKER3 analysis.
- PARAFAC. PARAFAC, or parallel factor analysis, is very similar to TUCKER3 except that each dimension is reduced to the *same* number of components.

ChemoSpec2D can carry out PARAFAC and it is discussed in greater detail below.

- N-way Image Analysis. In the case of 2D NMR spectra N would be three. This is a general term term which includes processes such as the TUCKER variations and PARAFAC.
- Principal Tensor Analysis (PTA). A tensor is just another term for an array of any dimension, so in the case of a collection of 2D NMR the data is a 3-way tensor. PTA is not a statistical method, but rather an algorithmic option to carry out the statistical computations.

Keirs provides discussion and suggestions on terminology best practices (Kiers (2000)).

PARAFAC

Theory (Light). PARAFAC is "parallel factor analysis." This is a statistical technique that is somewhat analogous to PCA. Recall that PCA decomposes a 2D data set into scores and loadings, and is bilinear:

$$\mathbf{X}^{(n \times p)} = \mathbf{C}^{(n \times R)} \times \mathbf{S}^{(R \times p)} + \mathbf{E}$$

Where $\mathbf{X}$ is the raw data, composed of n samples $\times$ p frequencies, $\mathbf{C}$ are the scores, and $\mathbf{S}$ are the loadings. R is the number of components selected. R is very much smaller than p, as noise and correlating variables have been reduced. Matrix $\mathbf{C}$ can be thought of as "concentrations" or weights. Matrix $\mathbf{S}$ is composed of "spectra" which serve as loadings. $\mathbf{E}$ consists of residuals/error. The goal of the PCA algorithm is to solve this equation and return $\mathbf{C}$ and $\mathbf{S}$.

In comparison, PARAFAC decomposes a 3D data set into three matrices, and is trilinear. Because the data is 3D, standard matrix algebra cannot be applied. However, the math can be expressed as a summation:

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + \epsilon_{ijk}$$

Where x_{ijk} is an element of the 3D data array $\underline{\mathbf{X}}$. a_{ir} is an element of the matrix $\mathbf{A}$, and so forth for $b/\mathbf{B}$ and $c/\mathbf{C}$. ϵ is the error term.

If $\underline{\mathbf{X}}$ is flattened by taking the Kth-dimension slices and concatenating them left-to-right to give a matrix $\mathbf{X}$, then 2D matrix operations *can* provide a solution:

$$\mathbf{X} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T + \mathbf{E}$$

Here, $\odot$ represents the Khatri-Rao product, a matrix multiplication variant needed in this situation. $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are the component matrices as above (Bro and Smilde (2003); Smilde *et al.* (2004) Appendix 4.A presents a number of alternative notations for PARAFAC).

Interpretation. Regardless of the mathematical representation or algorithmic solution, the results provide $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. Interpretation of the component matrices depends upon how $\underline{\mathbf{X}}$ was constructed (i.e. which dimension represents the samples). In the case of ChemoSpec2D $\mathbf{C}$ contains values analogous to scores in that they can be used to see how samples cluster (this is because the samples are the third dimension of $\underline{\mathbf{X}}$). Standard matrix multiplication of $\mathbf{A} \times \mathbf{B}^{\mathbf{T}}$ for a particular column (component) gives a 2D loading plot (a pseudo-spectrum) showing the contributions (loadings) of each peak to the component. ChemoSpec2D uses the R package multiway to carry out PARAFAC (Helwig (2017)).

PARAFAC is also known by other terms:

- Tensor rank decomposition.
- Canonical polyadic decomposition.
- CANDECOMP (canonical decomposition); PARAFAC and CANDECOMP are the same mathematical process but were reported at about the same time and given different names by their respective discoverers.
- Tri-linear decomposition.

Functions

The list below gives each user-facing function and a brief description of what it does. Full information is of course available via the help function, e.g. ?sumSpectra2D. Note that a number of the utility functions are actually in a supporting package called ChemoSpecUtils but are loaded automatically when activating ChemoSpec2D.

- Utility Functions
  - **files2Spectra2DObject** Imports 2D data sets. The format options are currently rather limited and not completely vetted!
  - **chkSpectra** Checks the integrity of a Spectra2D object. This can be used directly and is also called by nearly every other function to ensure data integrity.
  - **sumSpectra** Prints a short summary of the Spectra2D object.
  - **sumGroups** Prints a short summary of the group membership of the spectra in a Spectra2D object.
  - **removeGroup** Remove an entire group from a Spectra2D object.
  - **removeSample** Remove one or more samples from a Spectra2D object.
  - **removeFreq** Delete selected frequencies (on either dimension).
  - **removePeaks2D** Set selected peaks to NA (on either dimension).
  - **plotSpectra2D** Plots 2D spectra stored in a Spectra2D object, as a contour plot. Detailed data exploration is probably better done on the spectrometer which is much more suited. This function is for quick checks and also publication-quality plots.

  - **plotSlice** Plots a slice of a 2D spectrum.
  - **inspectLvls** An easy way to inspect the data in order to choose appropriate contour levels.
  - **normSpectra2D** Normalizes the 2D spectra stored in a Spectra2D object.
  - **centscaleSpectra2D** Centers and scales 2D spectra stored in a Spectra2D object.

- Statistical Analysis Functions

  - **pfacSpectra2D** Carries out PARAFAC analysis of the Spectra2D object.
  - **pfacScores** Plots the scores from a PARAFAC analysis. Useful for looking at how the samples cluster.
  - **pfacLoadings** Plots a 2D pseudo-spectrum showing which peaks contribute to each component.
  - **miaSpectra2D** Carries out MIA on a Spectra2D object, equivalent to a TUCKER1 analysis.
  - **miaScores** Plots the scores from MIA.
  - **miaLoadings** Plots the loadings from MIA.
  - **plotScree** Plots scree plots from MIA or PARAFAC.

Acknowledgments. I'd like to thank Teddy Zartler of Pfizer for encouragement in developing this package, and for providing valuable test data sets.

References

Bro R, Smilde A (2003). "Centering and scaling in component analysis." *Journal of Chemometrics*, **17**(1), 16–33. .

Helwig NE (2017). *multiway: Component Models for Multi-Way Data*. R package version 1.0-3, URL https://CRAN.R-project.org/package=multiway.

Huang J, Wium H, Qvist K, Esbensen K (2003). "Multi-way methods in image analysis-relationships and applications." *Chemometrics and Intelligent Laboratory Systems*, **66**(2), 141–158.

Kiers H (2000). "Towards a standardized notation and terminology in multiway analysis." *Journal of Chemometrics*, **14**(3), 105–122.

Smilde A, Bro R, Geladi P (2004). *Multi-way Analysis: Applications in the Chemical Sciences*. Wiley.