Development of Chemometric Tools for 2D NMR Data Sets

Bryan A. Hanson
Dept. of Chemistry & Biochemistry, DePauw University, Greencastle IN USA
hanson@depauw.edu

FOSS Chemometric Tools

The `R` ecosystem is a free and open source (FOSS) environment for statistical computing and graphics.[1] One of its great strengths is the over 10,000 user-contributed packages. Building on the chemometrics package for spectroscopy `ChemoSpec`[2], I am currently developing a new package to handle 2D NMR spectra, described herein. In conjunction, the `readJDX` package has recently been expanded to read 2D NMR data sets in NTUPLE format.[3] The `ChemoSpec2Ddata` package serves as a repository for data sets.[4]

Disclaimer: These packages are works in progress and the details will certainly change from what is described here! There are missing features and I'm sure, errors!

`ChemoSpec2D`

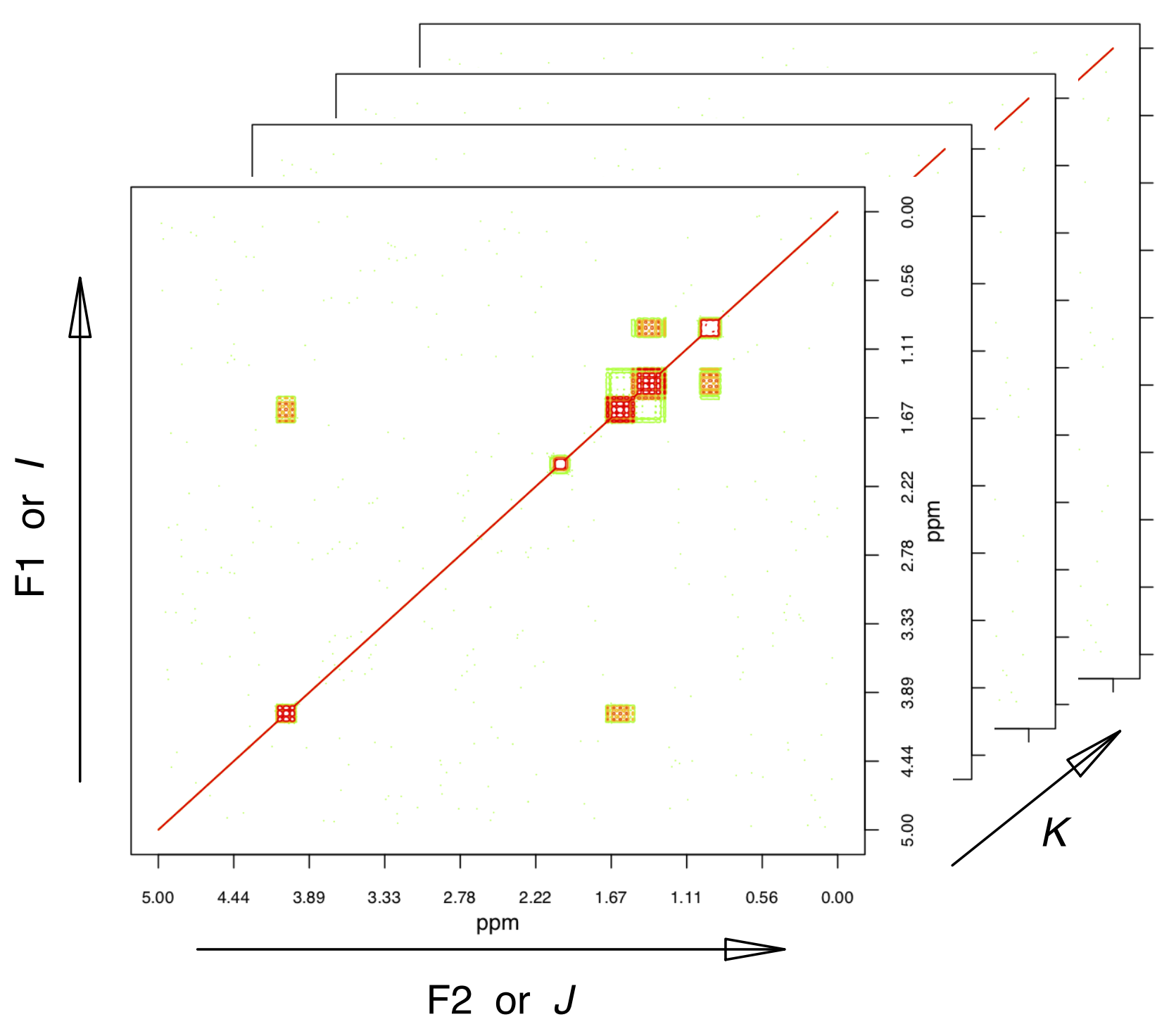
`ChemoSpec2D`[5] is designed to analyze 2D spectroscopic data such as COSY and HSQC NMR spectra using appropriate chemometric techniques. It deploys methods aimed primarily at classification of samples and the identification of spectral features which are important in distinguishing samples from each other. `ChemoSpec2D` stores and manipulates each spectrum as a data matrix, and hence a data set is a collection of 2D spectra. An entire data set is naturally visualized as a 3D array with dimensions:

$$\text{F2} \times \text{F1} \times \text{no. samples} \tag{1}$$

or

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

where `F2` and `F1` are the x- and y-axes/dimensions. We will refer to this array as $\underline{\mathbf{X}}$.



Configuration of data array $\underline{\mathbf{X}}$
I, J and K are array indices.

`ChemoSpec2D` Con't

`ChemoSpec2D` treats each spectrum 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 comparison, in a weak analysis, the 3D data set is unfolded into a series of contiguous 2D matrices and analyzed using methods typical for any 2D data set (which are fundamentally bilinear)[6]. 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, strong analysis, while trilinear, has fewer parameters to estimate so it is simpler (but computationally more demanding). The interpretation is also more straight-forward.

What is PARAFAC?

PARAFAC is "parallel factor analysis" and is the means of implementing the strong analysis described above. This is a statistical technique that is conceptually analogous to principal components analysis (PCA). It is also referred to as CANDECOMP. PCA decomposes a 2D data set into scores and loadings, and is *bilinear*:

$$\mathbf{X}^{(n\,\times\,p)} = \mathbf{C}^{(n\,\times\,R)} \times \mathbf{S}^{(R\,\times\,p)} + \mathbf{E} \tag{3}$$

Where $\mathbf{X}$ is the raw data, composed of n samples $\times\, p$ frequencies, $\mathbf{C}$ are the scores representing the samples, and $\mathbf{S}$ contains the loadings. R is the number of principal components selected by the analyst. Typically $R << p$, since noise and correlating variables have been eliminated. 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. 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, yielding a *trilinear* relationship. Because the data is 3D, standard matrix algebra cannot be applied to $\underline{\mathbf{X}}$. However, the relationship can be expressed as a summation:

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + \epsilon_{ijk} \tag{4}$$

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 an error term.

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

$$\mathbf{X} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T + \mathbf{E} \tag{5}$$

PARAFAC Con't

Here, $\odot$ represents the column-wise Kronecker product, a matrix multiplication variant needed in this situation. $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are the component matrices as above ([7, 8]).

Regardless of the mathematical representation, the algorithm provides matrices $\mathbf{A}$, $\mathbf{B}$ 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 which show how samples cluster (this is because the samples are in the K^{th} dimension of $\underline{\mathbf{X}}$). Standard matrix multiplication of $\mathbf{A} \times \mathbf{B}^{\mathrm{T}}$ 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 ([9]). Keirs provides discussion and suggestions on terminology best practices for these mathematical systems.[10]

User-Facing Functions

Utility Functions

files2Spectra2DObject Imports 2D data sets. The format options are currently rather limited!

chkSpectra2D Checks the integrity of a `Spectra2D` object. This can be used directly and is also called by nearly every other function.

sumSpectra2D Prints a short summary of the `Spectra2D` object.

sumGroups2D Prints a short summary of the group membership of the spectra in a `Spectra2D` object.

removeGroup2D Remove an entire group from a `Spectra2D` object.

removeSample2D Remove one or more samples from a `Spectra2D` object.

removeFreq2D Delete selected frequencies (on either dimension).

removePeaks2D Set selected peaks to NA (on either dimension).

plotSpectra2D Plots one 2D spectrum stored in a `Spectra2D` object, as a contour plot. Serious plotting and exploration is probably better done on the spectrometer. This function is for quick checks and also publication-quality plots.

normSpectra2D Normalizes the 2D spectra stored in a `Spectra2D` object.

PARAFAC-Related Functions

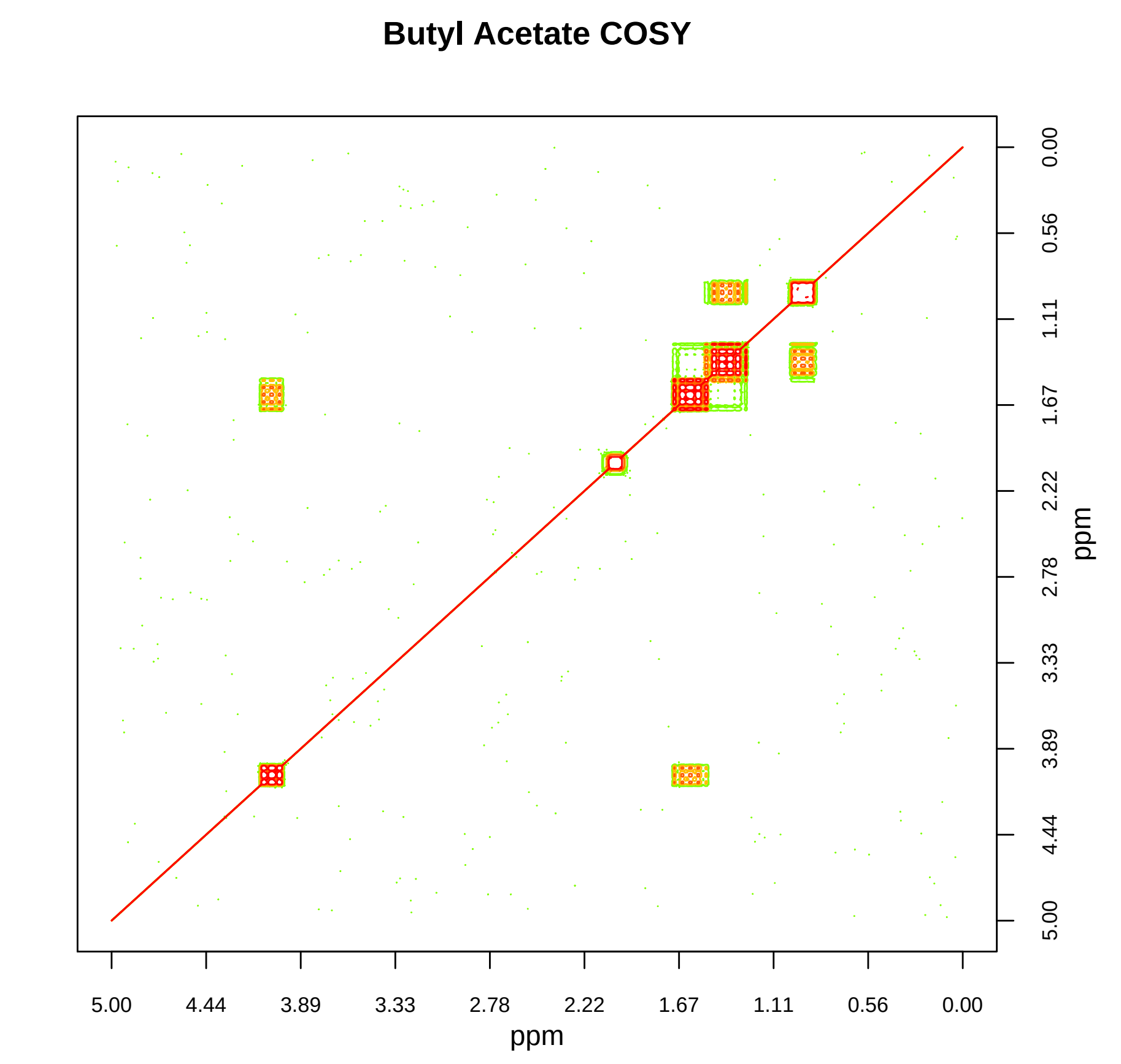
pfacSpectra2D Carries out PARAFAC analysis of the `Spectra2D` object.

pfacScores Plots the scores from a PARAFAC analysis. Useful for looking at how the samples cluster.

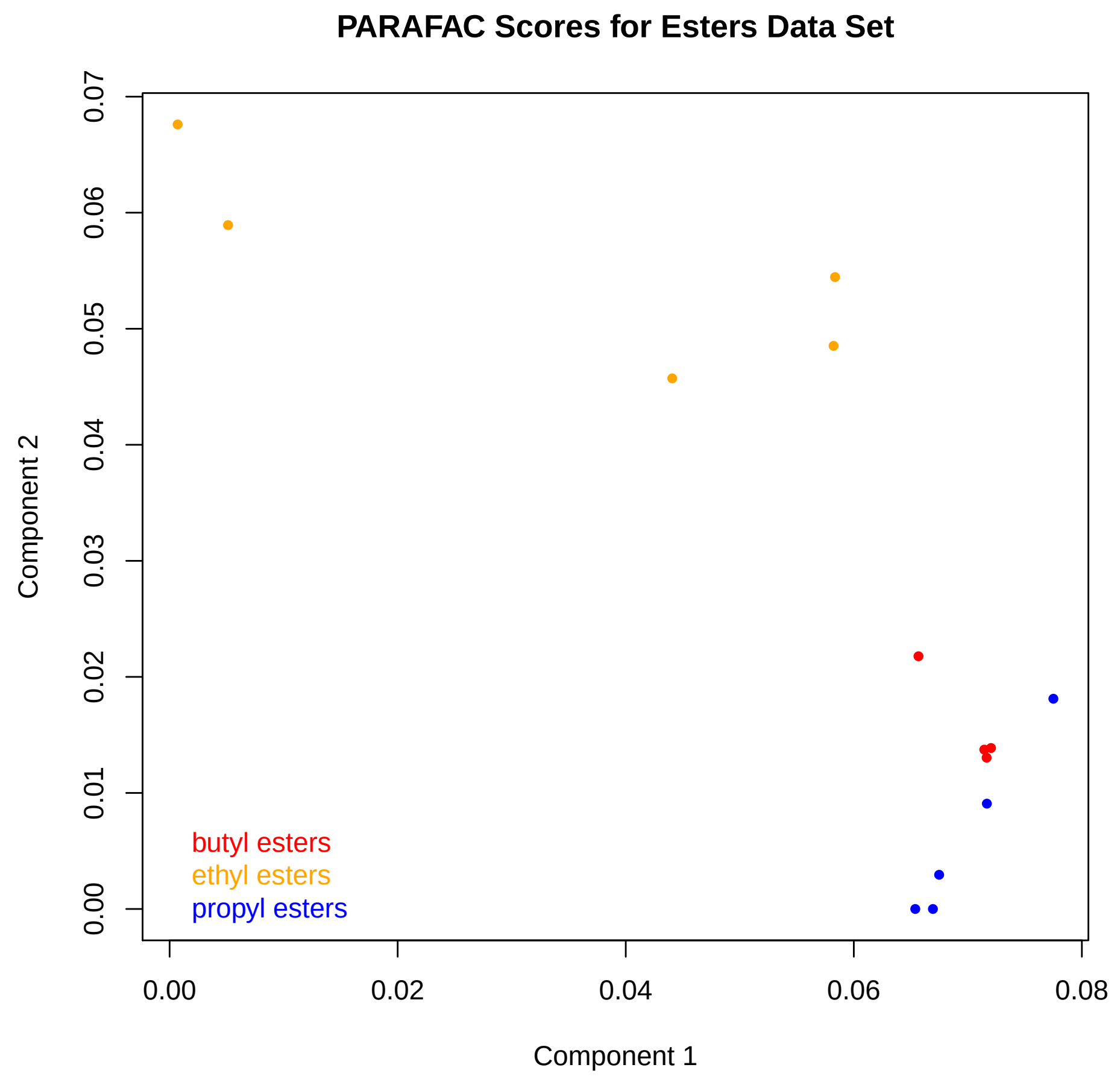
pfacLoadings Plots a 2D pseudo-spectrum showing which peaks contribute to each component.

Esters Data Set

Supporting package `ChemoSpec2Ddata` provides a set of simulated 300 MHz COSY spectra of 14 esters of alkanoic acids. The COSY spectrum of butyl acetate is shown below as an example.

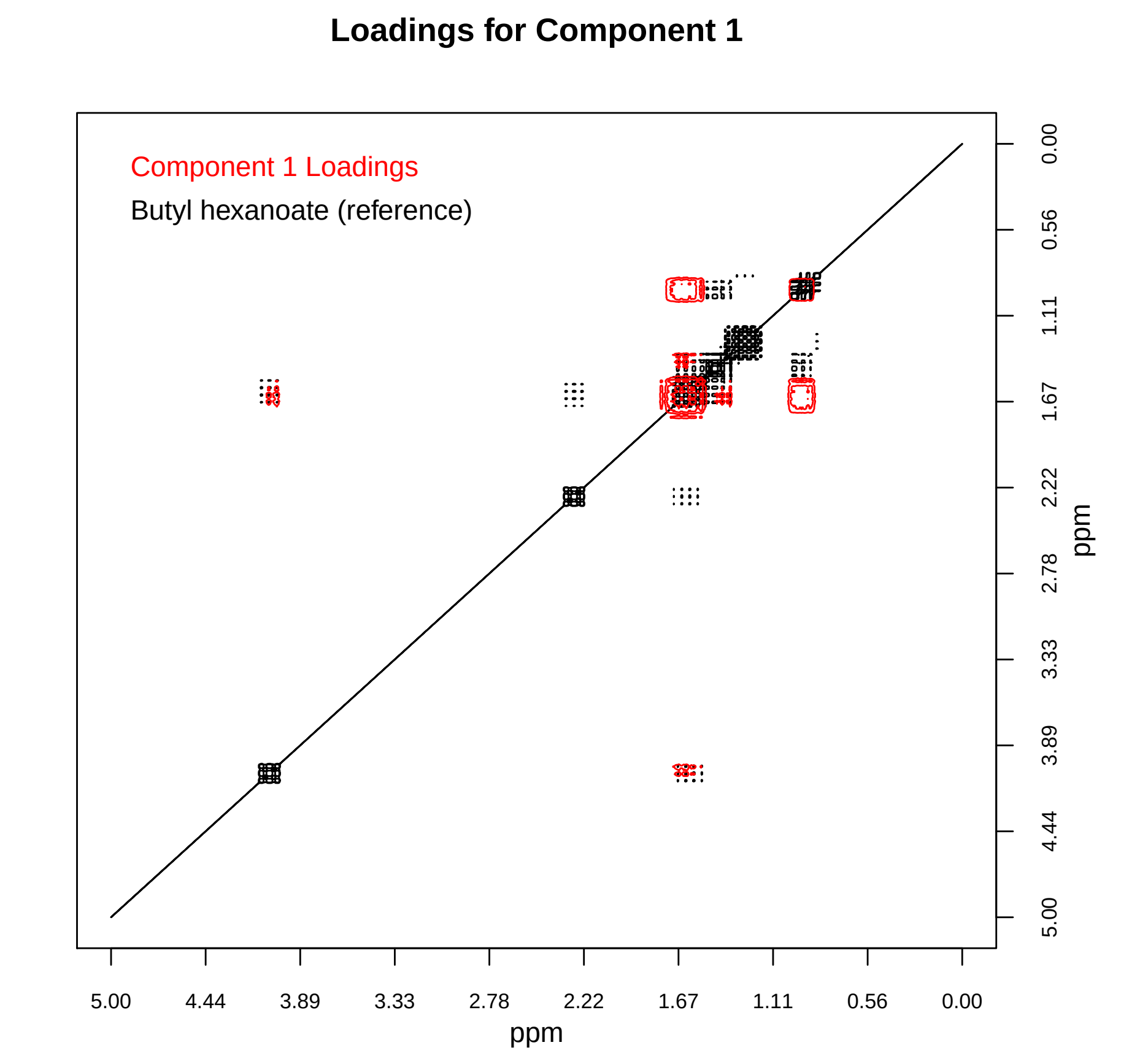


PARAFAC analysis gives a score plot which shows how the samples cluster. The interpretation of this plot is identical to the interpretation of a score plot in PCA.



Esters Data Set, Con't

The next plot shows the first loading component computed by PARAFAC, in red. This can be described as a pseudo-spectrum. A reference spectrum is shown in black for comparison.



References

[1] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018.

[2] Bryan A. Hanson. *ChemoSpec: Exploratory Chemometrics for Spectroscopy*, 2017. R package version 4.4.97.

[3] Bryan A. Hanson. *readJDX: Import Data in the JCAMP-DX Format*, 2018. R package version 0.3.225.

[4] Bryan A. Hanson. *ChemoSpec2Ddata: 2D NMR Data Sets for Use with ChemoSpec2D*, 2017. R package version 0.0.5.

[5] Bryan A. Hanson. *ChemoSpec2D: Exploratory Chemometrics for 2D Spectroscopy*, 2017. R package version 0.1.240.

[6] J Huang, H Wium, KB Qvist, and KH Esbensen. Multi-way methods in image analysis-relationships and applications. *Chemometrics and Intelligent Laboratory Systems*, 66(2):141–158, JUN 28 2003.

[7] R Bro and AK Smilde. Centering and scaling in component analysis. *Journal of Chemometrics*, 17(1):16–33, JAN 2003.

[8] A Smilde, R Bro, and P Geladi. *Multi-way Analysis: Applications in the Chemical Sciences.* Wiley, 2004.

[9] NE Helwig. *multiway: Component Models for Multi-Way Data*, 2017. R package version 1.0-3.

[10] HAL Kiers. Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics*, 14(3):105–122, MAY-JUN 2000.

Software

The software described here is currently available at `github.com/bryanhanson`

PANIC 2018 in Beautiful La Jolla!