ChemoSpec2D: Chemometric Workflows for 2D NMR

Bryan A. Hanson

Freelance Software Developer & Prof. Emeritus

Dept. of Chemistry & Biochemistry, DePauw University, Greencastle IN USA

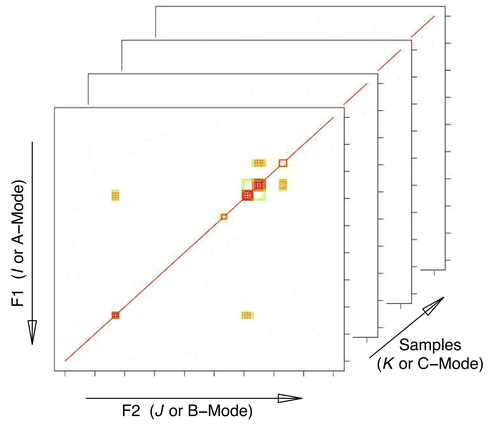
hanson@depauw.edu

ChemoSpec2D

ChemoSpec2D[1] is an **R** package[2] designed to analyze 2D spectroscopic data such as COSY and HSQC NMR spectra using appropriate chemometric techniques. It includes 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. Chemometric analysis of 2D NMR data sets is of interest for quality control of protein biosimilar products, for monitoring process chemistry, and can be applied to low field instruments. A 2D NMR data set is naturally visualized as a 3D array with dimensions:

$$\mathbf{F1} \times \mathbf{F2} \times \text{no. samples} = \text{2D Spectrum} \times \text{no. samples}$$

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



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[3]. 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. *All analyses described here operate directly on spectra – no manual peak curation is necessary.*

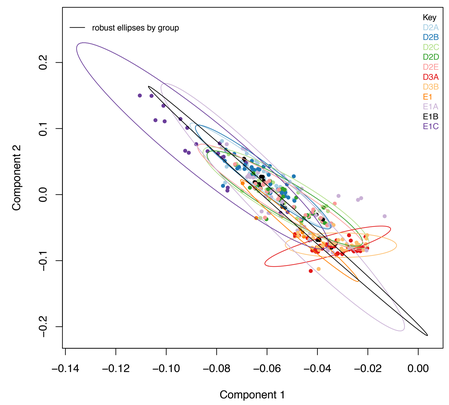
To demonstrate some of the features of **ChemoSpec2D** we will use subsets of the NIST mAb data set[4]. These are ^{1}H, ^{13}C gHSQC spectra. Spectra were collected on several instruments at several field strengths, with a variety of acquisition schemes. The data includes both the NIST Fab and the SSS, which is an isotopically enriched version of the NIST Fab. Data set NIST 344 has 344 spectra from groups D2A, D2B, D2C, D2D, D2E, D3A, D3B, E1, E1A, E1C and E1B. Data set NIST 175 is composed of spectra from groups D2A, D2B, D3A, D3B, and E1B (codes from Table 1[4]). All spectra were collected at 37C except those in group E1B.

Linearized PCA

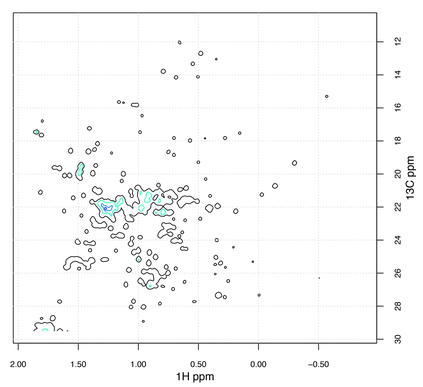
To date, nearly all of the chemometric explorations of 2D NMR data have linearized the data and carried out principal component analysis (PCA). Linearizing refers to the process of taking a single 2D spectrum, and concatenating each F2 slice one after the other to give a "long" 1D spectrum that represents the original data. Each 2D spectrum is linearized, and the resulting "long" 1D spectra are stacked to give a matrix of data which is then subjected to PCA. **ChemoSpec2D** can carry out the linearization and the results can be transferred to **ChemoSpec** for further analysis[5]. However, as this method does not keep the 2D data set intact, no results will be shown here other than timings. In addition, the loading results are difficult to interpret.

Multivariate Image Analysis

Multivariate Image Analysis (MIA) is a technique from the field of image analysis which in the current context treats each 2D spectrum as an image. The data is reduced *only* along the samples dimension. The next figure shows the score plot for the NIST 344 data set.

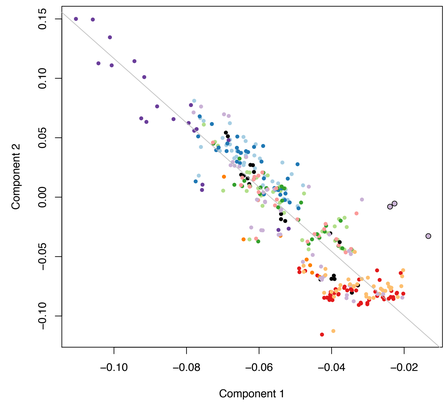


As in PCA, MIA gives loadings which show which peaks are driving the separation, as seen next. The black contours are a typical reference spectrum and the shades of blue are the loadings pseudo-spectrum.

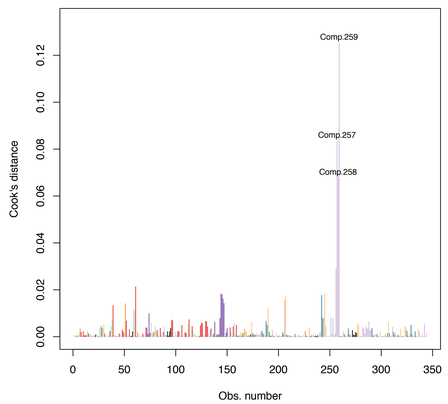


Application: Identifying Outliers

An interesting empirical observation is that the scores of similar spectra in a MIA analysis fall close to a line. This suggests a method for quality control. In the next figure the gray line represents a linear model (r^2 = 0.81). Three data points that are outliers are outlined in black.



Potential outliers can be identified automatically via a plot of Cook's distance, a well-established measure of the influence of a data point on a linear model (regression). The three circled points are readily identified as problematic spectra. This approach has a great deal of potential as a means of quality control.



PARAFAC

PARAFAC or "parallel factor analysis" is another data reduction method which reduces that data along all three dimensions[6]. PARAFAC is considerably more demanding on computational resources and results are not shown here. However, the results are similar to the MIA analysis.

Computational Performance

The performance of the functions described here is given in the following table. Time values are seconds and are the median computation time for the number of runs given in repetitions.

Calculation	*Repetitions*	*Machine*	*Parallel*	*NIST 175*	*NIST 344*
LPCA	50	Mac	no	79	277
MIA	50	Mac	no	19	64
MIA	50	AWS t2.2x	no	–	77
MIA	50	AWS m4.2x	no	–	79
MIA	50	AWS c5.4x	no	–	76
PFAC	20	AWS c5.4x	no	235	–
PFAC	1	AWS c5.4x	yes	78	yikes!

AWS refers to Amazon Web Services which provides virtual machines with various hardware configurations. The most expensive machine used here costs $0.68/CPU hour. Mac refers to MacBook Pro with 16 GB RAM running on a 2.9 GHz Intel Core i5 chip with 2 cores. All machines are running the latest software.

Free & Open Source Software

The **R** ecosystem is a free and open source (FOSS) software environment for statistical computing and graphics which runs on all common platforms[2]. One of its great strengths is the over 10,000 user-contributed packages. Both **ChemoSpec2D** and the more general purpose **ChemoSpec**[5] package for chemometric analysis of spectroscopic data sets are part of the **R** ecosystem.

References

[1] Bryan A. Hanson. *ChemoSpec2D: Exploratory Chemometrics for 2D Spectroscopy*, 2019. R package version 0.2.16.

[2] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[3] 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, 2003.

[4] Robert G. Brinson, John P. Marino, Frank Delaglio, Luke W. Arbogast, Ryan M. Evans, Anthony Kearsley, Geneviève Gingras, Houman Ghasriani, Yves Aubin, Gregory K. Pierens, Xinying Jia, Mehdi Mobli, Hamish G. Grant, David W. Keizer, Kristian Schweimer, Jonas Ståhle, Göran Widmalm, Edward R. Zartler, Chad W. Lawrence, Patrick N. Reardon, John R. Cort, Ping Xu, Feng Ni, Saeko Yanaka, Koichi Kato, Stuart R. Parnham, Desiree Tsao, Andreas Blomgren, Torgny Rundlöf, Nils Trieloff, Peter Schmieder, Alfred Ross, Ken Skidmore, Kang Chen, David Keire, Darón I. Freedberg, Thea Suter-Stahel, Gerhard Wider, Gregor Ilc, Janez Plavec, Scott A. Bradley, Donna M. Baldisseri, Mauricio Luis Sforça, Ana Carolina de Mattos Zeri, Julie Yu Wei, Christina M. Szabo, Carlos A. Amezcua, John B. Jordan, and Mats Wikström. Enabling adoption of 2D-NMR for the higher order structure assessment of monoclonal antibody therapeutics. *mAbs*, 11(1):94–105, 2019.

[5] Bryan A. Hanson. *ChemoSpec: Exploratory Chemometrics for Spectroscopy*, 2019. R package version 5.0.226.

[6] R Bro. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38(2):149–171, 1997.