Statistical Analysis of Spectral Binning Data

Colin Vitols, David Chang, Aalim Weljie, Jack Newton

June 2007

Analyzing ^{1}H NMR spectra in metabolomics studies often requires multivariate pattern recognition techniques to extract meaningful results. Spectral binning is an effective data reduction technique commonly used to prepare spectra for multivariate analysis. Multivariate analysis software packages like Umetrics SIMCA-P can readily analyze spectral binning output from the Profiler module of Chenomx NMR Suite.

Introduction

Metabolomics researchers often face the challenge of analyzing vast amounts of data. An NMR-based metabolomics workflow typically starts with a series of acquired NMR spectra, followed by data preprocessing algorithms, and ending with powerful chemometric techniques. Spectral binning is a common technique used for high-throughput data preprocessing.

This note describes using Chenomx NMR Suite to perform data preprocessing, namely spectral binning, and Umetrics SIMCA-P to build meaningful statistical models. A set of set of 46 NMR spectra of rat brain extracts will help illustrate the process of analyzing this data. Samples for the spectra were collected from the hippocampus and the frontal, occipital and temporal cortices. The goal is to differentiate the four regions using binned NMR data.

Spectral Binning in Profiler

To perform spectral binning in Profiler, you will need to specify a number of parameters. Your selections will vary depending on the requirements of the analysis you intend to perform on the output data.

Binning Method.

You can determine the bin size either by directly specifying a size in ppm, or by specifying a total number of bins. In both cases, you can restrict binning to a particular section of the spectrum by indicating start and end points in ppm. In this example, binning was performed from 0.04 to 10 ppm with a bin size of 0.04 ppm.

Binning Target.

For traditional spectral binning, select the spectrum line (standard) binning target. The residual binning target is most useful following targeted profiling analysis of the binned spectra, and will be discussed in a later note. This examples uses the spectrum line (standard) binning target.

Dark Regions.

You can exclude regions of the spectrum from the binning process by providing a comma-delimited list of regions, in ppm. This lets you exclude signals that are unrelated to your experiment. These can include water, which will vary depending on water suppression parameters, DSS or TSP, typically added as chemical shape indicators, or other compounds. In this dataset, for example, the methanol signal was excluded, since methanol was used as an extraction solvent during sample preparation; variation in methanol levels among these samples is not necessarily intrinsic to the samples themselves.

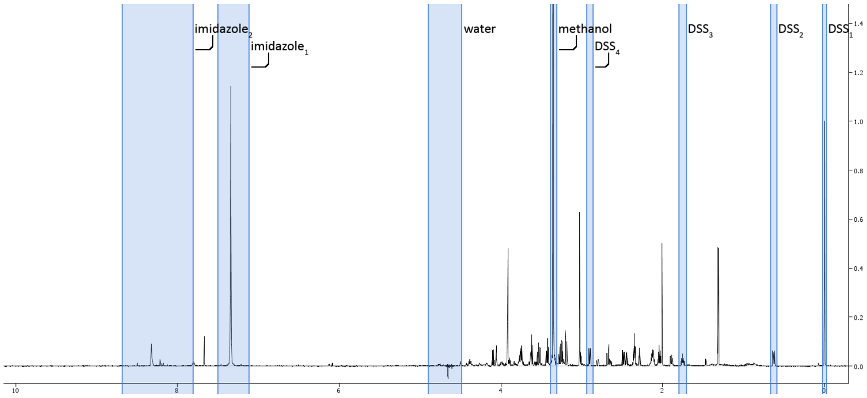


Figure 1. Dark regions are not included in the spectral binning output from Profiler

For this example, the dark regions exclude DSS at 0.552-0.695, 1.66-1.84 and 2.85-2.95 ppm, methanol at 3.22-3.47 ppm, water at 4.44-5.50 ppm and imidazole at 7.10-7.53 and 7.77-8.73 ppm. Starting the binned region at 0.04 ppm implicity excludes the DSS methyl peak occurring at 0 ppm.

Normalization.

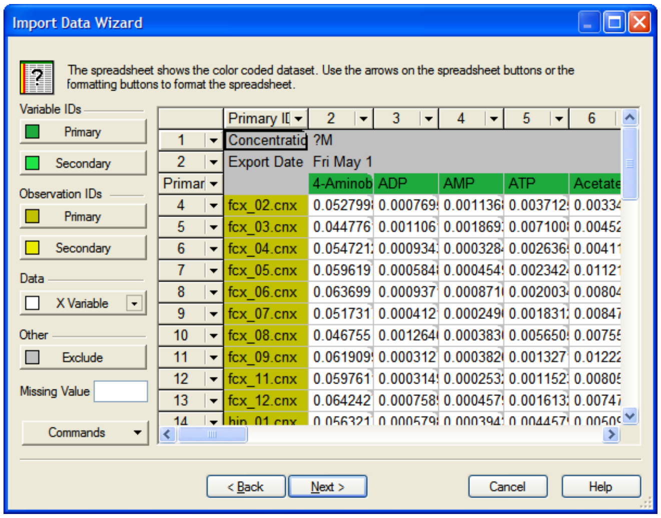
To normalize your binning data, you can choose one of two methods. Total area (%) records bin values as a fraction of the total spectrum area, excluding the area in the specified dark regions. Standardized bins (sa) records bin values as a fraction of the CSI area expressed in units of standard area (sa).

The standard area of a CSI peak is its experimental area as a fraction of the area under a theoretical DSS methyl peak at 0.5 mM. Generally, you should only use standardized bins in datasets where you can assume dilution effects to be insignificant. This example uses total area normalization.

Statistical Analysis with Umetrics SIMCA-P Software

With the spectral binning data exported from Profiler, you can now start the statistical analysis of the data using Umetrics SIMCA-P software. The version used in this application note is Umetrics SIMCA-P 11.5.

You must first create a new SIMCA-P project. The data source is the spectral binning data exported from Profiler.



The spreadsheet shows the color coded dataset. Use the arrows on the spreadsheet buttons or the formatting buttons to format the spreadsheet.

	Primary ID	2	3	4	5	6
1	Concentratio	?M				
2	Export Date	Fri May 1				
Primar		4-Aminob	ADP	AMP	ATP	Acetate
4	fcx_02.cnx	0.052799	0.000769	0.001136	0.003712	0.00334
5	fcx_03.cnx	0.044776	0.001106	0.001869	0.007100	0.00452
6	fcx_04.cnx	0.054721	0.000934	0.000328	0.002636	0.0041
7	fcx_05.cnx	0.059619	0.000584	0.000454	0.002342	0.0112
8	fcx_06.cnx	0.063699	0.000937	0.000871	0.002003	0.00804
9	fcx_07.cnx	0.051731	0.000412	0.000249	0.001831	0.00847
10	fcx_08.cnx	0.046755	0.001264	0.000383	0.005650	0.00755
11	fcx_09.cnx	0.061909	0.000312	0.000382	0.001327	0.01222
12	fcx_11.cnx	0.059761	0.000314	0.000253	0.001152	0.00805
13	fcx_12.cnx	0.064242	0.000758	0.000457	0.001613	0.00747
14	hip_01.cnx	0.056321	0.000579	0.000394	0.004457	0.0050

Figure 2. The Import Data Wizard

With the spectral binning data imported into SIMCA-P, you can begin analyzing the data. The most common statistical models applied to metabolomic datasets are *Principal Components Analysis* (PCA) and *Partial Least Squares for Discriminant Analysis* (PLS-DA).

PCA is a data reduction technique used to reduce the dimensionality of a multi-dimensional dataset while retaining the characteristics of the dataset that contribute most to its variance.

PLS-DA is a regression extension of PCA that takes advantage of *class information* to attempt to maximize the separation between groups of observations.

In this dataset, you can use the available class information, specifically, the brain region that a particular sample came from, to establish a PLS-DA model. You can extract this information from the spectrum file names.

First, edit the workset by clicking on "Workset > Edit > 1", and clicking on the "Observations" tab. Select "Primary ID" from the "Class from obs ID:" drop-down. Click on "Set", enter a start position of 1 and a length of 3, then click "OK". This will use the regions indicated in the file names to define classes for the PLS-DA model. If you do not see the "Observations" tab, you may need to switch to Advanced Mode by clicking the appropriate button in the bottom left corner of the workset wizard.

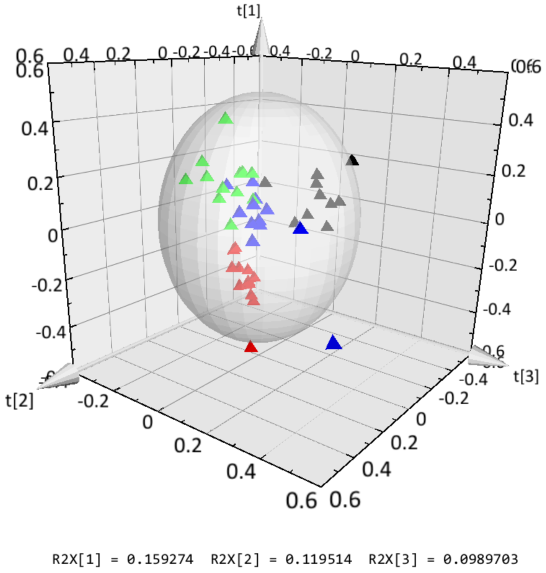


Figure 3. The PLS-DA 3D scores plot

Once you have identified class values, you can generate a PLS-DA model. Click "Analysis > Change Model Type" and select "PLS Discriminant Analysis". To fit the model to the data, select "Analysis > Autofit", which will extract the principal components needed to fit the model.

When the fit of the PLS-DA model is complete, you can use a 3D plot to visualize the results (Figure 3). To generate a 3D plot of the PLS-DA scores, click "Plot/List > Scatter 3D Plot". The default selection is to plot the first three principal components. Each data point represents a particular sample, and is automatically color-coded according to its class.

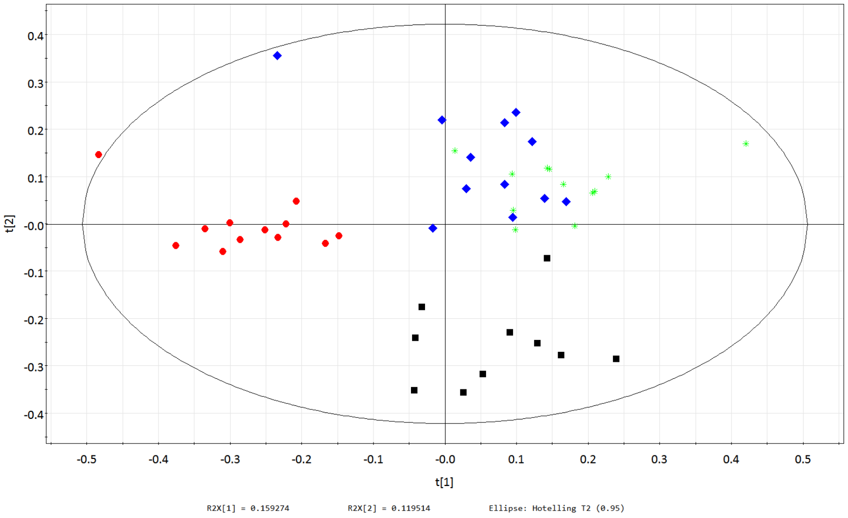


Figure 4. The PLS-DA 2D scores plot. The red dots, corresponding to the hippocampus samples, are well separated from the other samples.

The 2D scores plot (Figure 4) shows the red dots, corresponding to samples from the hippocampus, well separated from the other observations, corresponding to the frontal, occipital and temporal cortices. The loadings plot (Figure 5) indicates a range of bins, labeled by their frequency in ppm that are significant in separating the hippocampus samples from the rest.

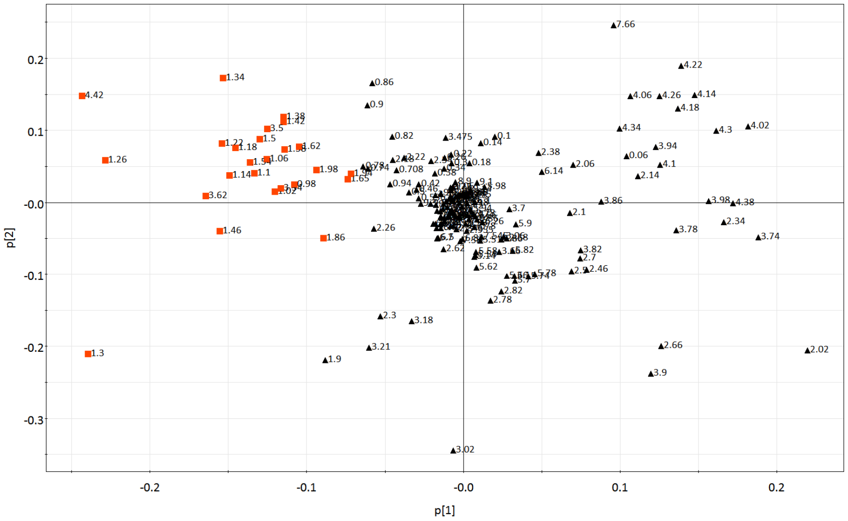


Figure 5. The PLS-DA loadings plot for components 1 and 2. Bins that are significant in differentiating the hippocampus samples are marked in red (top left and bottom left quadrants).

A line plot of one of the samples (Figure 6), with the significant bins marked, indicates the parts of the spectrum that are most influential in differentiating the hippocampus spectra from the others. Bins centered at 1.30, 3.53, 3.62 and 4.42 ppm are of particular interest.

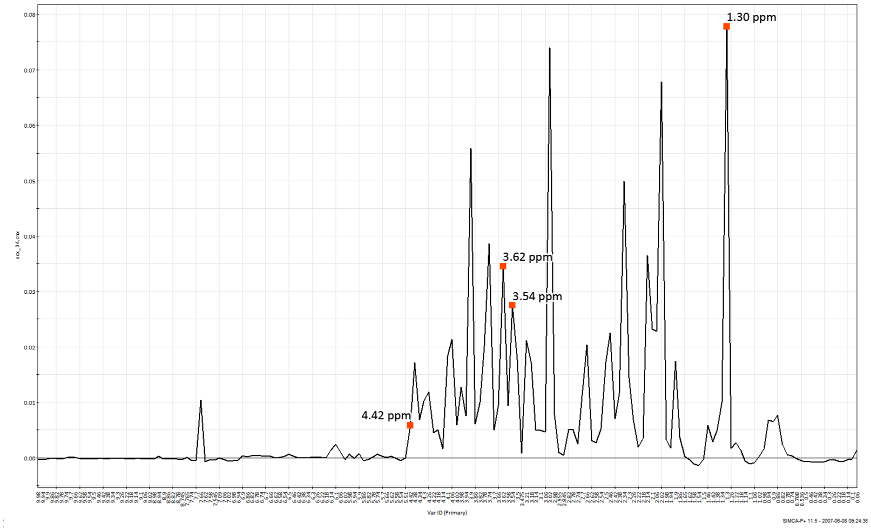


Figure 6. A line plot of a representative sample (ocx_04), with significant bins from the loadings plot marked in red.

Correlating Metabolites in Profiler

Returning to Profiler, you can open the actual spectrum corresponding to the line plot above. Based on the interesting bins identified in the line plot, targeted profiling reveals that these bins correlate to lactate at 1.3 ppm, myo-inositol at 3.54 and 3.62 ppm and inosine at 4.42 ppm (Figure 7).

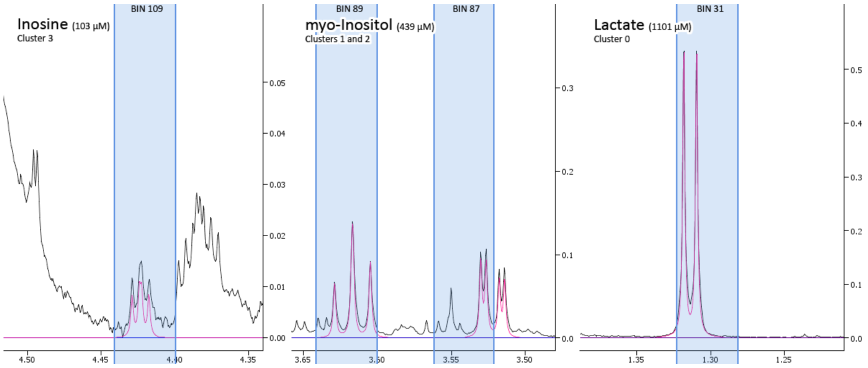


Figure 7. Targeted profiling of the bins identified in the line plot reveals metabolites that are significant in differentiating the hippocampus samples.

Acknowledgements

Chenomx would like to thank Brent McGrath and Dr. P. Silverstone in the Department of Psychiatry at the University of Alberta for providing data for this application note, and Mark Earll at Umetrics for his assistance with the statistical analysis.

References

[1] R Rosewell and C Vitols. **March 2006.** *Identifying Metabolites in Biofluids. Chenomx Inc.*

[2] AM Weljie, J Newton, PM Mercier, E Carlson, and CM Slupsky. **2006.** *Targeted Profiling: Quantitative Analysis of 1H-NMR Metabolomics Data. Anal Chem.* 78(13):4430-4442

[3] D Chang, J Newton, C Vitols, and AM Weljie. **April 2006.** *Statistical Analysis of Targeted Profiling Data. Chenomx Inc.*