Statistical Analysis of Targeted Profiling Data

Colin Vitols, David Chang, Jack Newton, Aalim Weljie

April 2006

Analyzing ^1H NMR spectra in metabolomics studies often requires multivariate pattern recognition techniques to extract meaningful results. Targeted profiling offers the ability to analyze spectra based on the identity and quantities of individual compounds. Targeted profiling results from Chenomx NMR Suite can be used as input in statistical software packages such as Umetrics SIMCA-P. Performing PCA on Chenomx targeted profiles yields information-rich results that allow rapid biological interpretation, since group separation may be directly correlated to variations in specific metabolite concentrations.

Introduction

Two key challenges face metabolomics researchers when they analyze NMR data. First, researchers must determine whether there are meaningful statistical patterns in their data. Then, they must translate these patterns into biologically relevant information in the form of metabolite identity and quantity.

In this note we demonstrate how to resolve both of these challenges using Chenomx NMR Suite and Umetrics SIMCA-P. In this approach, Chenomx NMR Suite extracts targeted profiling information from the NMR spectra (see [2]), and Umetrics SIMCA-P performs statistical analysis of this data.

A set of 46 NMR spectra of rat brain extracts will help illustrate the process of analyzing these data. Samples for the spectra were collected from the hippocampus and the frontal, occipital and temporal cortices. The goal is to determine whether it is possible to differentiate the four regions based on their metabolic profiles and, if so, which metabolites are important in differentiating the regions.

Exporting Results from Profiler

In order to perform a statistical analysis of results obtained via targeted profiling in Chenomx NMR Suite, the results from a set of analyzed spectra must first be compiled into a single file. The Profiler module allows you to generate the compiled results by exporting compound concentrations from these analyzed spectra.

The spectra for which you would like to export compound concentrations must all be in the same folder. Profiler stores exported compound concentrations as a simple tab-delimited text file, readable by most spreadsheet and statistical analysis software, including Umetrics SIMCA-P.

Data Normalization

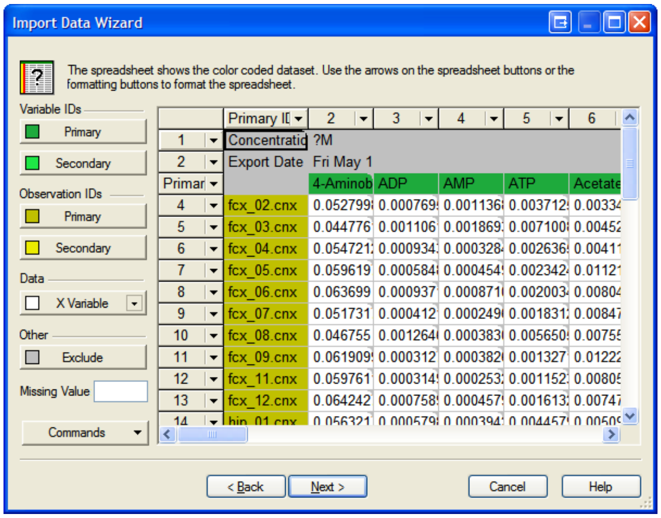
Data normalization is an important step for any statistical analysis. The objective of data normalization is to allow meaningful comparisons of samples within the dataset. When analyzing NMR spectra, dilution effects often make meaningful comparisons of data without normalization impossible; by normalizing the data you can eliminate systematic variation introduced by such effects. In the case of spectral binning (also known as "spectral bucketing"), for example, integral regions are often normalized to the total spectrum area.

To analyze targeted profiling data, two types of normalization are possible. The concentrations of all metabolites in a sample can be normalized to an endogenous metabolite that is estimated to be present at a relatively constant concentration level. Creatinine often serves this purpose in urine studies, for example, as it is a commonly used indicator of kidney function. For targeted profiling data, in a method analogous to the total area normalization used in spectral binning, metabolite concentrations can be normalized to the total concentration of all endogenous metabolites. This latter normalization method provided the basis for the analysis described in this note.

Statistical Analysis with Umetrics

With the concentration data exported from Profiler and normalized, you can now start the statistical analysis of the data using SIMCA-P. The version used in this application note is Umetrics SIMCA-P 11.0.

You must first create a new SIMCA-P project. The data source is the concentration set exported and normalized previously.



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	0.00411
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	bin_01.cnx	0.056321	0.000579	0.000394	0.004457	0.0050

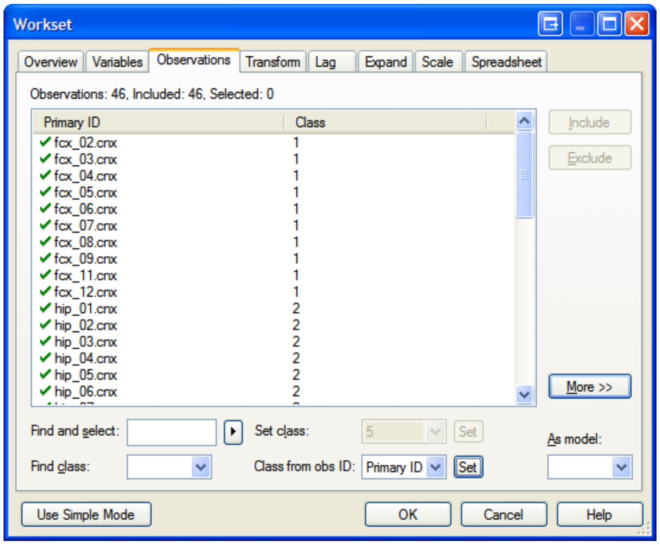
Figure 1. The Import Data Wizard

After importing the targeted profiling data into SIMCA-P, you can begin analysis of 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).

One of the goals in this area of research is to use data obtained from your samples to build an appropriate model for classification and, if possible, determine the factors leading to differences among the samples. PCA and PLS-DA are often used for this purpose, as these techniques determine orthogonal latent variables that describe the input data and classify the data based on these variables. The primary advantage of using targeted profiling as an input to PCA or PLS-DA is that the resulting variables are combinations of measured metabolite concentrations. As such, these variables are easier to interpret as factors in the underlying classification model. Thus, targeted profiling provides meaningful and interpretable factors describing the input data.

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.



Observations: 46, Included: 46, Selected: 0

Primary ID	Class
✓ fcx_02.cnx	1
✓ fcx_03.cnx	1
✓ fcx_04.cnx	1
✓ fcx_05.cnx	1
✓ fcx_06.cnx	1
✓ fcx_07.cnx	1
✓ fcx_08.cnx	1
✓ fcx_09.cnx	1
✓ fcx_11.cnx	1
✓ fcx_12.cnx	1
✓ hip_01.cnx	2
✓ hip_02.cnx	2
✓ hip_03.cnx	2
✓ hip_04.cnx	2
✓ hip_05.cnx	2
✓ hip_06.cnx	2

Find and select:

Set class: 5 Set

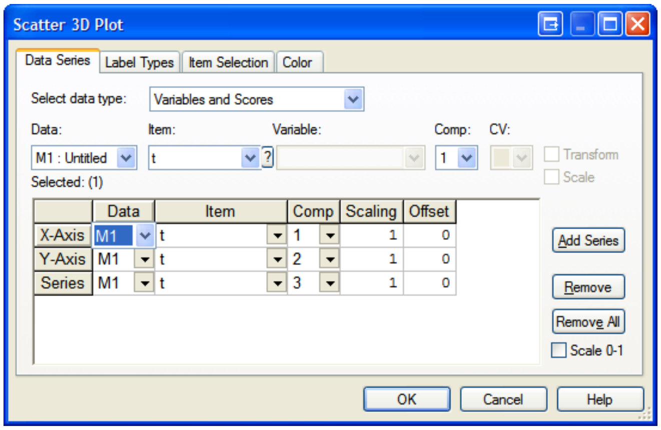
As model:

Find class:

Class from obs ID: Primary ID Set

Figure 2. Setting classes in the Workset Wizard

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 at the start of each file name to define classes for the PLS-DA model (fcx = frontal cortex, hip = hippocampus, ocx = occipital cortex, tcx = temporal cortex). 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.



Select data type: Variables and Scores

Data: M1 : Untitled

Item: t

Comp: 1

Selected: (1)

	Data	Item	Comp	Scaling	Offset
X-Axis	M1	t	1	1	0
Y-Axis	M1	t	2	1	0
Series	M1	t	3	1	0

Figure 3. Setting up the Scatter 3D Plot

Once you have filtered the data and 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 properly fit the model.

When the fit of the PLS-DA model is complete, you can use a 3D plot to visualize the results. To generate a 3D plot of the PLS-DA scores, start a new 3D scatter plot from the Plot/List menu, and select the Variables and Scores data type, then select t[1] for the X-Axis, t[2] for the Y-Axis, and t[3] for the Series. Figure 4 shows the resulting PLS-DA scores plot. Each data point represents a particular sample, and is automatically color-coded according to its class. This plot shows a very good separation between the various regions of the brain.

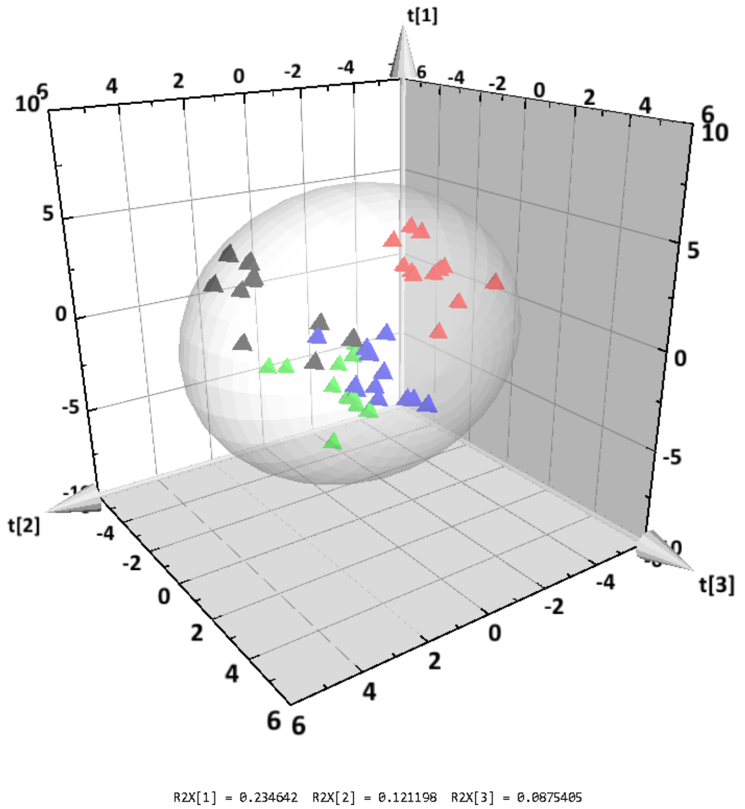


Figure 4. The PLS-DA 3D Scores Plot

Having determined that there is a clear pattern in the data, it is natural to ask which metabolites are responsible for separating the various elements of the brain. In other words, what is the biological significance of the data? The loadings plot can provide an answer to this question, by describing the weighting coefficients for each metabolite.

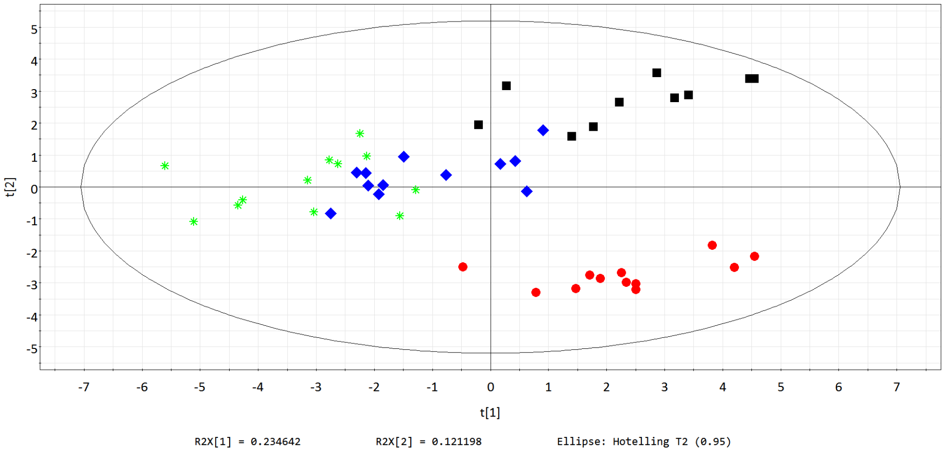


Figure 5. The PLS-DA 2D Scores Plot

Interpreting both the scores and loadings plots will be simpler with a 2D projection of the data. To generate a 2D plot of the PLS-DA scores, start a new scatter plot from the Plot/List menu, select the Variables and Scores data type, and select t[1] for the X-Axis and t[2] for the Series.(Figure 5).

Examining the loadings plot for PC 1 and PC 2 (Figure 6) will reveal which metabolites are important in separating the two samples. To generate a 2D plot of the PLS-DA loadings, start a new scatter plot from the Plot/List menu, select the Observations and Loadings data type, and select p[1] for the X-Axis and p[2] for the Series. There is a direct geometric link between the scores plot and the loadings plot. The metabolites responsible for "pulling" samples to, for example, the lower-left quadrant of the scores plot can be found in the lower-left quadrant of the loadings plot. For example, you can see that alanine, fumarate, choline, inosine, lactate and myo-inositol are significant in characterizing samples from the hippocampus (red circles), and that aspartate, hypoxanthine, acetate, O-phosphocholine, glycine and 4-aminobutyrate are significant in characterizing samples from the frontal cortex (black squares).

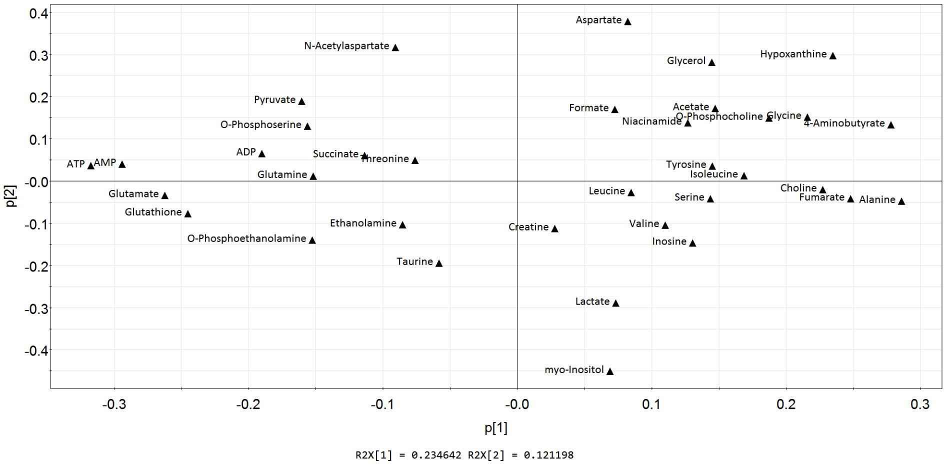


Figure 6. The PLS-DA Loadings Plot

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.

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