TechnicalNOTE

Waters

AN EXPLANATION OF PRINCIPAL COMPONENTS ANALYSIS (PCA) FOR METABONOMICS

Robert S. Plumb, Jennifer H. Granger, and Chris L. Stumpf
Waters Corporation, Milford, MA, USA

STATISTICAL ANALYSIS OF LC/MS DATA

When investigating the changes in biological or chemical systems as a result of a chemical or toxic event, the comparison of control and sample LC/MS chromatograms is often insufficient to completely describe the effect. This may be due to differences in the metabolism of the test animals, fast responders or poor absorbers, or the fact that there may be more than one effect occurring, e.g., pharmacodynamic and toxic. Therefore, a more appropriate data reduction technique to elucidate variations in an LC/MS data set is to use a multivariate statistical approach; the simplest of which is Principal Components Analysis (PCA).

PCA enables complex data to be reviewed in a graphical manner to easily identify similar and dissimilar samples, thus highlighting the variability in a multivariate data set. This data viewing approach has been successfully employed in several industrial applications, including food and beverage analysis, gasoline analysis, and process control monitoring. The power of pattern recognition methods is that they reduce a complex data set to two- or three-dimensional scores maps. These maps allow the visualization of intrinsic patterns in the data sets, which indicate any relationships between the samples, identify outliers, and point toward the reason behind any pattern that is observed.

HOW PCA WORKS

An LC/MS data set is multi-dimensional, comprised of sample identification, retention time, mass, and intensity data. The Waters® MarkerLynx™ Application Manager for MassLynx™ Software effectively reduces this matrix to a two-dimensional data set of sample identifiers and peak intensity.

The resulting data has variance equal to the number of peaks detected in the LC/MS analysis, as can be seen in Figure 1. Principal Components Analysis attempts to reduce the variance in the data such that it can be viewed in two or three dimensions. In order to explain how PCA achieves this, we must first consider a much simpler data set where there are only three variables and *n* samples. If we plot the coordinates for the first sample in three-dimensional space, the graph in Figure 2 is obtained.

Retention Time_m/z pair →

Samples ↓	2.24_318.0634	5.46_317.1806	6.05_317.1722	2.97_317.1714
RAT 1	0	0	0	2.15351
RAT 2	0	0	0	2.10822
RAT 3	1.63034	0	0	0
RAT 4	1.62986	0	0	0
RAT 5	4.70965	0	0	0.730389
RAT 6	0	1.03318	0	1.83726
RAT 7	0	0	0	0
RAT 8	2.83714	0.947788	0	0.919644
RAT 9	5.23023	0	0	0.956396
RAT 10	0	0.843124	0	0

Figure 1.

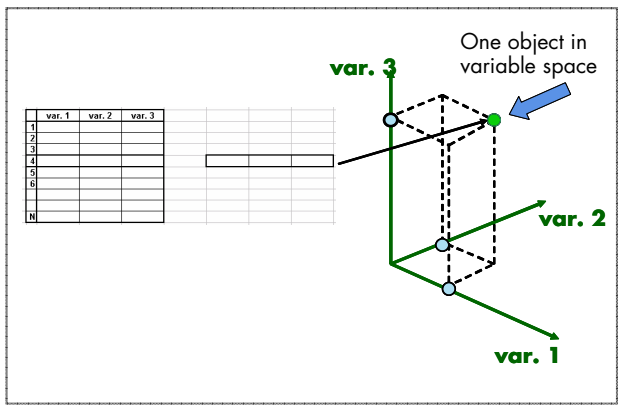

Figure 2.

Following this, we then perform the same operation on the rest of the samples, yielding the data shown in Figure 3. The mean of the swarm of data points produced is calculated, and the center of the swarm is moved to the origin of the x,y,z plot. Next, a line is constructed through the data swarm which best describes the difference in the data set (the variance), as seen in Figure 4.

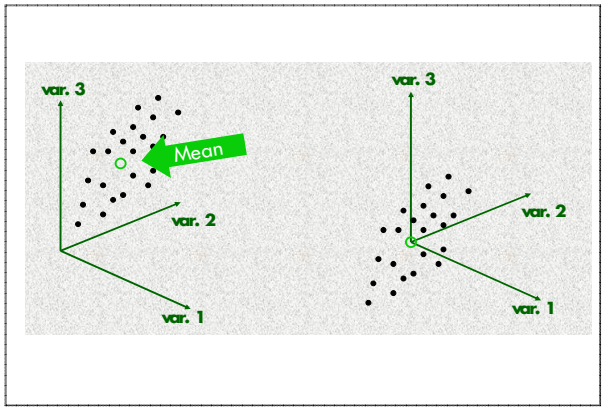

Figure 3.

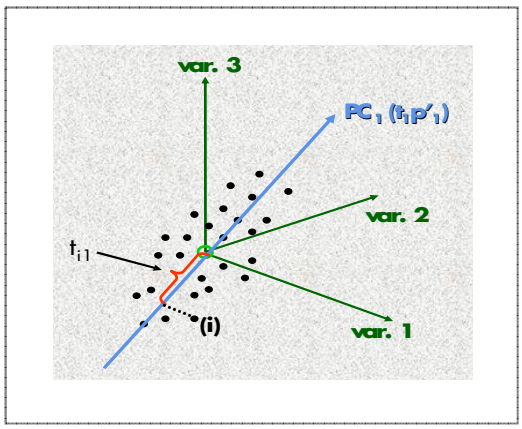

Figure 4.

This line is termed principal component one, PC_1, and describes the largest variation in the data, or, the direction in which the points spread most in the variable space. The score value, t_{i1}, is obtained by the projection of the point (i) onto the principal component line, and is the distance from this projection point to the origin. The second principal component, PC_2, is obtained by plotting a line through the data which next best describes the variance in the data; this must be orthogonal to the first principal component. The t_{i2} value is obtained by the projection of the point back onto the PC_2 line (Figure 5). The two principal components make up a plane in the variable space. The points are then projected down on the plane that can be lifted out and viewed as a two-dimensional plot describing the object's relationships, known as a *scores plot* (t_{i1}/t_{i2}), Figure 6. In this plot, similarities and dissimilarities between objects (samples) can be viewed. The perpendicular distance from the object to the projection on the plane is the residual, or the variation not described by the two PC values.

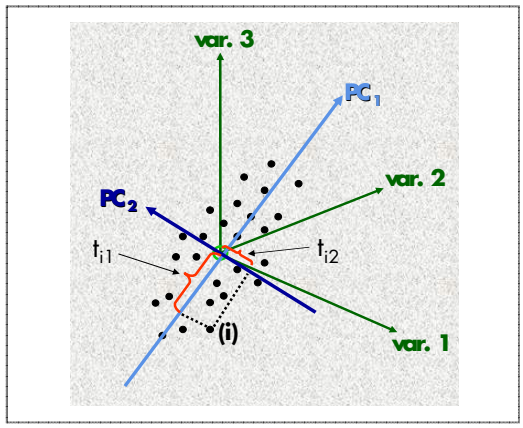

Figure 5.

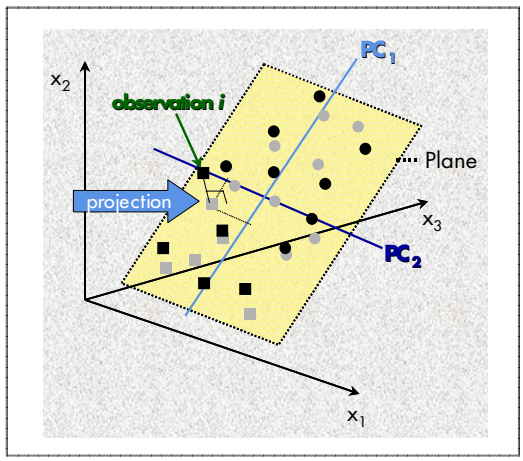

Figure 6.

The corresponding *loadings plot* (PC_1/PC_2) describes the variables' relationships, and is also a means of interpreting the scores plot by indicating which variables are responsible for differences or similarities between objects. Once this process has been completed, the final plot generated allows a graphical visualization of data as shown in Figure 7, a scores plot generated from MarkerLynx.

The scores plot is partnered by the loadings plot which identifies the points, in our case m/z and retention time pairs, that are responsible for the variance in the data. In this plot, the ions at the greatest distance from the origin contribute most to the observed group clustering.

EFFECT ON METABONOMICS DATA PROCESSING

The usefulness of this type of multivariate statistical data processing in metabonomics lies in the fact that the samples are treated as a population. Therefore, any outliers due to poor adsorption or fast metabolizers can be either identified and followed, or altogether excluded from the data analysis. The approach of PCA has been exploited in the field of metabonomics to identify the changes in endogenous metabolite profiles due to a toxic effect. One great advantage of this type of analysis, as compared to simple chromatogram-to-chromatogram comparisons, is that time-related data can be graphically represented, as shown in Figure 8.

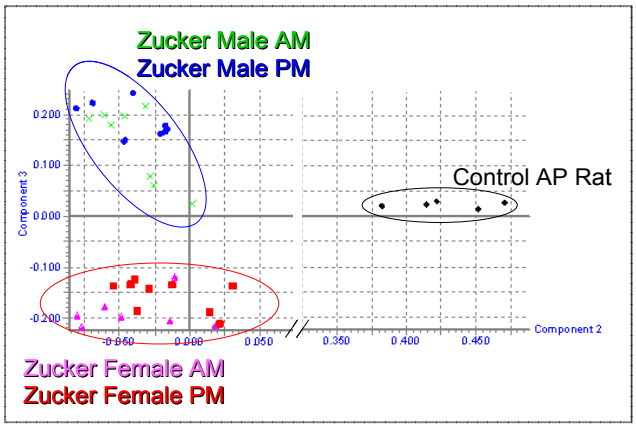

Figure 7.

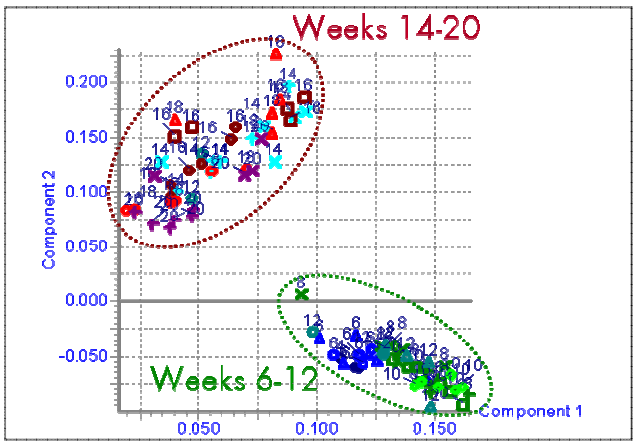

Figure 8.

CONCLUSION

The identification of biomarkers of toxicity and disease are essential to the development of new medicines, especially in the field of personalized medicine. The discovery of the these new biomarkers requires the generation and processing of information-rich LC/MS data, to elucidate these complex patterns which describe the underlying biological process. The benefit of using a multivariate statistical approach to the analysis of this data is that complex biological patterns can be deconvoluted. Principal components analysis not only facilitates the visualization of these complex patterns, but also identifies the signals responsible for the variance in the data. This makes PCA the obvious initial statistical approach for the analysis of mass spectrometric data for metabonomics.

TechnicalNOTE

Waters

Sales Offices:

AUSTRIA 43 1 877 18 07
AUSTRALIA 61 2 9933 1777
BELGIUM AND LUXEMBOURG 32 2 726 1000
BRAZIL 55 11 5543 7788
CANADA 800 252 4752 X2205
CZECH REPUBLIC 420 2 617 11384
DENMARK 45 46 59 8080
FINLAND 358 9 506 4140
FRANCE 33 1 3048 7200
GERMANY 49 6196 400600
HONG KONG 852 29 64 1800
HUNGARY 36 1 350 5086
INDIA 91 80 2837 1900
IRELAND 353 1 448 1500
ITALY 39 02 27 4211
JAPAN 81 3 3471 7191
KOREA 82 2 820 2700
MEXICO 52 55 5524 7636
THE NETHERLANDS 31 76 508 7200
NORWAY 47 6 384 6050
PEOPLE'S REPUBLIC OF CHINA 86 10 8451 8918
POLAND 48 22 833 4400
PUERTO RICO 787 747 8445
RUSSIA/CIS 7 095 931 9193
SINGAPORE 65 6278 7997
SPAIN 34 93 600 9300
SWEDEN 46 8 555 11 500
SWITZERLAND 41 62 889 2030
TAIWAN 886 2 2543 1898
UK 44 208 238 6100
US 800 252 4752

WATERS CORPORATION
34 Maple St.
Milford, MA 01757 U.S.A.
T: 508 478 2000
F: 508 872 1990
www.waters.com

Waters

For Complete Confidence

Nov. 2004 720001055EN KJ-PDF

LLOYD'S REGISTER QUALITY ASSURANCE · ISO9001

UKAS QUALITY MANAGEMENT 001

IT