Anal. Chem. **2008,** *80,* 4933–4944

Dimensionality Reduction and Visualization in Principal Component Analysis

Gordana Ivosev, Lyle Burton, and Ron Bonner*

MDS Analytical Technologies, Concord, Ontario, Canada, L4K 4V8

Many modern applications of analytical chemistry involve the collection of large megavariate data sets and subsequent processing with multivariate analysis techniques (MVA), two of the more common goals being data analysis (also known as data mining and exploratory data analysis) and classification. Classification attempts to determine variables that can distinguish known classes allowing unknown samples to be correctly assigned, whereas data analysis seeks to uncover and understand or confirm relationships between the samples and the variables. An important part of analysis is visualization which allows analysts to apply their expertise and knowledge and is often easier for the samples than the variables since there are frequently far more of the latter. Here we describe principal component variable grouping (PCVG), an unsupervised, intuitive method that assigns a large number of variables to a smaller number of groups that can be more readily visualized and understood. Knowledge of the source or nature of the variables in a group allows them all to be appropriately treated, for example, removed if they result from uninteresting effects or replaced by a single representative for further processing.

A consequence of the "omics" explosion is a rapid increase in the number and complexity of megavariate data sets and a corresponding need for appropriate data analysis and interpretation tools. Techniques used in proteomics [liquid chromatography mass spectrometry (LC−MS)], metabolomics (LC−MS, NMR), and transcriptomics (RNA microarrays) can generate data containing hundreds or thousands of variables for typically tens or hundreds of samples that are analyzed by multivariate analysis (MVA) tools. The data is in the form of an $m \times n$ matrix containing the responses for the n variables in each of the m samples and is processed prior to analysis[1,2] to correct the responses for individual samples (normalization, a row operation in the terminology of ref 2) and to adjust the relative importance of the variables (scaling, a column operation). Common scaling methods[1] are autoscaling (also known as standardization), where the responses for each variable are centered by subtracting its mean value and then divided by its standard deviation, and Pareto scaling, where the mean-centered values are divided by the square root of the standard deviation. Autoscaling, which was originally devised to permit the use of different variables that might have different scales, gives each variable the same variance regardless of magnitude or scale so that they are equally important. Pareto scaling reduces but does not completely remove the magnitude and seems more appropriate for chemical data from the same analytical technique where the larger peaks are often more reliable and less susceptible to noise. This has been explored by Paatero and Hopke[3] who concluded that autoscaling is inappropriate for noisy data where the contribution of weak variables can become too high.

The preprocessed data are analyzed usually to either build classification models that can be used to predict the class of unknown samples or to explore the structure present in the data (data analysis). Classification methods are said to be supervised, since knowledge of class membership is required to build the models, whereas exploratory analysis methods are usually unsupervised. Even if the goal is classification, data analysis is still valuable since it can confirm the presence of the expected sample structure (classes), may reveal other structure that should be examined, and can assist feature selection. Determining relationships among the variables, however, is often more difficult.

Here we describe a data analysis technique that focuses on the structure of the variables in the context of the samples. The technique is unsupervised, intuitive, and requires few parameters. It is based on locating correlated variables (common in chemical data) and principal component analysis (PCA) in order to analyze the sample variance (expected and unexpected) and retain the ability to visualize the behavior of the samples and sample-variable links. It is compatible with high dimensionality data and provides a means of reducing the dimensionality in a chemically meaningful way.

Data analysis is often assisted by locating correlated variables, i.e., those that show similar response profiles across all the samples, for several reasons such as: (1) In some applications these may be directly of interest, for example, because they suggest genes that are coregulated,[4] biochemical pathways, or biomarkers. (2) The group behavior may indicate that the variables are due to effects that may not be of interest, perhaps diurnal or gender-related changes, so they can be excluded from further analyses which may in turn simplify and improve sample clustering. (3) In many analytical techniques chemical components generate responses at several variables, for example, peaks in the spectra generated by NMR, IR, and MS, and identifying those that are related aids identification of the compound.

* Corresponding author. E-mail: ron.bonner@sciex.com.

(1) van den Berg, R. A.; Hoefsloot, H. C. J.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. *BMC Genomics* **2006**, *7*, 142–157.
(2) Craig, A.; Cloarec, O.; Holmes, E.; Nicholson, J. K.; Lindon, J. C. *Anal. Chem.* **2006**, *78*, 2262–2267.
(3) Paatero, P.; Hopke, P. K. *Anal. Chim. Acta* **2003**, *490*, 277–289.
(4) Heyer, L. J.; Kruglyak, S.; Yooseph, S. *Genome Res.* **1999**, *9*, 1106–1115.

10.1021/ac800110w CCC: $40.75
Published on Web 06/07/2008

Furthermore, replacing redundant, correlated variables by a single representative can simplify the analysis and reduce noise thereby improving subsequent analysis or classification. However, there have been few reports of techniques for determining correlated variables that also allow visualization of the variables and samples; below we briefly summarize some of the approaches that have been previously reported for this purpose.

Cluster analysis techniques[5] such as K-means, self-organizing (Kohonen) maps (SOM), and hierarchical clustering (HCA) are widely used to determine sample groupings for classification and require methods for measuring the similarity, or alternatively distance, between samples. In essence, each sample is represented as a single point according to its responses for the n variables, and the distance between these points is calculated; the data is autoscaled to remove the effect of response magnitude, and the classification performance is assessed by comparing the variance within a group to the variance between groups. Many distance measures are known[5] with Euclidean, city-block, and Mahalanobis being among the commonest, the latter being used if the clusters are expected to be distorted. These techniques can be used to determine correlated variables, either by using the correlation coefficient as a similarity measure or by measuring the distance between variables positioned according to their values in the samples, but both require autoscaled data and can be susceptible to artifacts. For example, Heyer et al.[4] used cluster analysis techniques to determine coexpressed genes and found many false positives caused by large, random events generating erroneous correlation coefficients. They minimized these effects by removing low-level signals and those with small variations and by developing a technique they called "jackknife correlation". In addition, they discussed problems with conventional clustering techniques that are also relevant here. K-means and SOM need an estimate of the number of clusters and SOM additionally needs an arrangement of the nodes. Both techniques are iterative and generate random initial cluster centers so that they are nondeterministic and the results may be very dependent on these initial conditions. SOM also requires a number of parameters related to training the network such as the influence of adjacent cells and how this changes as the iterations proceed. Neither technique can handle samples and variables simultaneously. HCA based on correlation can cluster the samples and variables but is susceptible to noise and outlier effects caused by autoscaling the data. Furthermore, in megavariate data sets the displays can be difficult to interpret.

Principal component analysis[6,7] is a widely used unsupervised technique that reduces high dimensionality data to a more manageable set of new variables which simplifies the visualization of complex data sets for exploratory analysis. The smaller set of new variables can be used with classification techniques that require fewer variables than samples. In PCA the new variables (principal components, PCs) are linear combinations of the original variables extracted in order of the amount of data variance that they explain. The contribution that each variable makes to a PC is called the loading, and the amount of the PC present in a particular sample is known as the score. The results of PCA are visualized as projections of multidimensional scores and loadings in two- or three-dimensional plots. Understanding the sample information (scores) can be straightforward if there are few samples and any expected separation is observed, but is more difficult if the classes are unknown or unexpected clusters are detected. Interpreting the variable behavior (loadings) and understanding the significance of the combinations of variables can be challenging,[8,9] especially if there are many significant PCs, but is important and can help explain unexpected sample groups.

The problem of interpreting the loadings plots and understanding variable behavior is not new, and a number of approaches have been described, although these have generally been applied to less complex data sets with far fewer variables than those encountered in "omics" applications. Several reported techniques exploit the fact that sample separation tends to have the same orientation as the variables responsible. This arises because for any PC the samples with the highest scores will be those that have higher responses in the variables with the highest loadings on that PC. For example, biplots[7] overlay the scores and loadings plots (appropriately scaled) for two PCs so that sample-variable relationships are more obvious. Although this can be an effective approach for small systems with few samples, variables, and PCs, it can be very complex for megavariate data sets. Furthermore, the plots are only feasible for two dimensions, or three in very simple cases, so examination of multiple pairs may be necessary, and there is no simple way to automatically identify related variables.

Procrustes rotation[10] is often used to help interpret the results of factor analysis, which is related to PCA, by rotating and scaling the factor axes to generate a simpler structure where each variable contributes ideally to one factor and each factor is made up of just a few variables, i.e., the loadings are ideally 1, 0, or −1. The technique can be very successful at simplifying the loadings information but eliminates variables and distorts the correlation structure of the data.

Determining related variables has been addressed in mixture analysis where MVA techniques are applied to sets of spectra with the goal of determining the underlying spectra of the pure components and their concentrations. Windig and Meuzelaar[11] developed a technique they called "VarDia" (for variance diagram) to analyze two-dimensional (2D) loadings plots obtained by performing PCA on the spectra of mixtures obtained by pyrolysis-MS. In addition to the similar sample-variable orientation mentioned above, they noted that the individual components, or factors, have particular orientations in the PC1/PC2 loadings plots and that masses associated with the factors have similar orientations. This arises because of correlation and the cosine of the angle between the orientations of a mass and a factor corresponds to the correlation coefficient of the two. To determine these masses and factors, the 2D loadings space was divided into sectors and the variance in each sector calculated by summing the loadings for the individual masses in the sector projected onto a line

(5) Theodoridis, S.; Koutroumbas, K. *Pattern Recognition*; Academic Press: New York, 1999.
(6) Jolliffe, I. T. *Principal Component Analysis*, 2nd ed.; Springer: New York, 2004.
(7) Jackson, J. E. *A User's Guide to Principal Components*; Wiley and Sons: Hoboken, NJ, 2003.
(8) van der Greef, J.; Smilde, A. K. *J. Chemom.* **2005**, *19*, 376–386.
(9) Westerhuis, J. A.; Derks, E. P. P. A.; Hoefsloot, H. C. J.; Smilde, A. K. *J. Chemom.* **2007**, *21*, 474–485.
(10) Andrade, J. M.; Gómez-Carracedo, M. P.; Krzanowski, W.; Kubista, M. *Chemom. Intell. Lab. Syst.* **2004**, *72*, 123–132.
(11) Windig, W.; Meuzelaar, H. L. C. *Anal. Chem.* **1984**, *56*, 2297–2303.

through its center. The variances were plotted in polar coordinates, and the directions of the factors and corresponding masses were determined from the sectors having the largest variances. They proposed extensions of the technique for three dimensions by successive use of two-dimensional VarDia, and also used color coding to distinguish masses.[12] More recently the approach has been made fully three-dimensional[13] by transforming 3D loadings plots into spherical coordinates and calculating the variance in pyramidal windows of equal volume. The results were displayed as two contour plots each representing the surface of a hemisphere, and hence the techniques was called "ContVarDia"; related variables were determined by locating areas of high density. Unlike biplots, this technique does not readily allow the consideration of sample-variable relationships, which is perhaps less important in mixture analysis than in exploratory data analysis. Autoscaled data is preferred, since large peaks will otherwise dominate, but since this equalizes the variance of variable responses it increases the possibility that the random behavior of small, noisy peaks will affect the results.

Here we describe a technique we call principal component variable grouping (PCVG) that is applicable to exploratory data analysis and addresses a number of the issues described above. Following PCA we analyze the loadings to identify variables that show the same response patterns across all samples; these variables are assigned to a group and represented by a unique symbol in loadings plots. Members of the same group comprise peaks that have comparable behavior including those that are from the same compound. If the data originates from LC−MS analyses the retention time provides a further aid to interpretation since chemically related ions will have the same retention time and can be examined to determine the molecular ion, fragments that may be present, and possibly the identity of the compound.

There are several advantages to this approach:

- Grouping is performed in the context of PCA, is unsupervised, and requires few parameters; the ability to visualize sample separation and sample-variable relationships is retained and enhanced.
- The technique is applicable to any number of PCs and can be applied to complex data sets with an inherently high degree of structure (groups) although limiting the number of PCs considered effectively smoothes the data by removing noise. The appropriate number of PCs can be determined using conventional methods.[6,7]
- Pareto scaling can be used and is preferred since it minimizes the effects of small noisy variables. Centered data can also be used, but smaller peaks will tend to be obscured.
- A group comprises variables that show similar behavior and can be represented by a single variable such as a member of the group or a combination (mean, sum, etc.). This also allows the generation of new variables, such as group ratios, that are still interpretable since the group is generated in the original data space. This is only feasible with a small number of variables.
- Group response patterns (profiles) can be examined to determine the nature and origin of the group behavior. This can quickly identify variables arising from analytical artifacts or uninteresting behavior that the analyst may choose to ignore in subsequent analyses.
- Related variables, even if unexpected, are determined and can be used for interpretation.

We note that it may be possible to identify related variables prior to MVA based on knowledge of the analytical technique. For example, the mass spectrum of even a pure compound contains peaks that are due to isotopes, adducts, fragments, different charge states, etc.,[14] and so the number of variables can be reduced by assigning those that are related to individual components[15] that are represented by a single value. This is straightforward for expected peaks, for example those due to isotopes, but others such as unexpected fragments may be much harder to identify especially if the related ions have been removed. Since the ability to correctly associate fragments with precursors is important for interpretation, we prefer to process all peaks and identify those that are related, deferring simplification and the choice of a group representative to later in the processing.

PCVG is a general technique that can be used in any PCA analysis. Here we illustrate the approach with peak lists generated from a small set of LC−MS analyses, but it is equally applicable to any spectroscopic or hyphenated technique and can use the data as acquired or following binning or peak picking.

EXTRACTING CORRELATED VARIABLES

Theoretical Basis. PCA generates new components that are linear combinations of the original variables with each PC having a loading vector that indicates the contribution of each variable to that PC. PCs are extracted in order of the amount of variance in the original data set that they explain. The number of PCs cannot exceed the smaller of the number of samples or variables, but in practice the bulk of the variance is often carried by the first few PCs with the remainder being due to noise. Thus, the data can be represented by a small number (n) of PCs resulting in a significant reduction of the dimensionality of the data. Each variable has a position in the n-dimensional space defined by the loadings of the selected PCs, and as noted previously, those that are correlated will have the same orientation with the angle between them reflecting the correlation coefficient. Hence, to find correlated variables we seek those that have similar orientations in the loadings space. In previous reports[11,13] the loadings space was systematically examined to find directions that contained significant variance, i.e., a group of variables. Although this approach is feasible for two or three dimensions it does not easily scale to larger dimensions. In addition the data was autoscaled so that all variables tended to make equal contributions regardless of their actual magnitude.

We note that the angle between two vectors $\mathbf{x}$ and $\mathbf{y}$, is given by

$$\theta = \arccos\left(\frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{x}||\mathbf{y}|}\right) \quad (1)$$

(12) Windig, W.; Lippert, J. L.; Robbins, M. J.; Kresinske, K. R.; Twist, J. P.; Snyder, A. P. *Chemom. Intell. Lab. Syst.* **1990**, *9*, 7–30.
(13) Statheropoulos, M.; Mikedi, K. *Anal. Chim. Acta* **2001**, *446*, 353–368.
(14) de Hoffmann, E.; Stroobant, V. *Mass Spectrometry Principles and Applications*, 3rd ed.; Wiley and Sons: Chichester, England, 2007.
(15) Warrack, B.; Hnatyshyn, S.; Zhang, H.; Sanders, M. *Proceedings, 53rd ASMS Conference on Mass Spectrometry and Allied Topics*, San Antonio, TX, June 5−9, 2005.

and is independent of the length of the vectors and, hence, any scaling used. Here the dot product $\mathbf{x}\cdot\mathbf{y}$ is given by

$$\mathbf{x}\cdot\mathbf{y} = \sum_{i=1}^{n} x_i y_i \tag{2}$$

Where x_i and y_i are coordinates of vectors **x** and **y**, respectively, and the length $|\mathbf{x}|$ is given by

$$|\mathbf{x}| = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{3}$$

Futhermore, addition of two vectors

$$\mathbf{x} + \mathbf{y} = \mathbf{z} \tag{4}$$

where coordinates of **z** are $z_i = x_i + y_i$, generates a new vector with an orientation that *does* depend on the magnitude of the individual vectors, i.e., the orientation is effectively weighted by the values of the components of the individual vectors.

Rather than systematically investigate the loadings space and look for summed variance, we first find the variable with the longest loadings vector given by eq 3. Since we are using Pareto scaled data, which measures the variance while still considering the magnitude of the original variable, this can be considered to be the most significant. We then calculate the angles between this target vector and all others and retain those that are within a defined angle, α; these are assigned to the first group. An alternative approach would be to use the correlation coefficient, but we prefer the angle since this retains a visual link with the loadings plots. Having determined the first group we consider only the remaining variables and use the one with the longest vector as the new target. This process is repeated until all the variables have been assigned.

In practice, once we have determined variables that are part of a group, we calculate a new target vector by summing the group members with length greater than 50% of the original target vector length according to eq 4 and re-examine all variables using this new target. The 50% threshold is used to ensure that a large number of small variables do not unduly influence the calculation. As mentioned, the direction of the new target will be weighted by the lengths of the individual components and it can be used directly since the angle calculation is independent of length. This refinement is especially useful if the initial target vector has a slightly different orientation to the other members of the group. This situation can occur in mass spectrometry, for example, if the target variable is saturated so that its intensity and orientation are not consistent with its related peaks.

Algorithm. The basic algorithm can be considered to be an extension of manual interpretation of the loadings plot and requires only two parameters, the number of PCs to use (n) and the angle for group membership. We have, however, found it useful to add additional steps to optionally exclude small peaks and filter the groups generated. The most important of the latter is to reject groups that do not have a certain number of members; since spectroscopic data often has correlated variables, groups with one or two members are likely artifacts. We collect these in to a single group so that they can still be examined. In addition to ignoring small peaks (defined by a minimum vector length requirement) we can set a minimum length for the target vector so small variables that are randomly aligned do not start new groups.

The algorithm is implemented as follows:

1. Perform PCA using Pareto scaling and determine the number of PCs to use. Although it is possible to use all of the PCs, and completely reconstruct the original data, usually only the first few contain important information and the rest are mainly noise. The use of a smaller number of PCs effectively smoothes the data, improves the performance of the algorithm, and generally enhances visualization.

2. Ignore variables close to the origin. This optional step removes variables that generate short vectors with little variance that may arise from noise.

3. Find the variable with the longest vector calculated according to eq 3. This defines the target vector, **t**, in the selected PC loadings space.

4. Find all variables within a given spatial angle of the target vector. Each variable defines another vector, **x**, in the PC subspace. We calculate the angle between **x** and **t** using eq 1 and retain those where the calculated angle is smaller than the user specified parameter, α.

5. Sum all members of the group with vector lengths greater than 50% of the length of the target vector to generate a "recentered" target vector.

6. Use the recentered vector as the new target vector and repeat step 4.

7. If the number of variables assigned to the group is smaller than a user-defined minimum, assign them to a special group; otherwise assign the retained variables to a new group and remove them from further consideration.

8. Repeat from step 3 until all variables of sufficient magnitude are assigned.

Implementation and Usage. The algorithm has been implemented in a small program that interacts with the MarkerView software (Applied Biosystems | MDS Sciex, Toronto, Ontario, Canada) to retrieve the loadings values and determine the groups according to user-defined parameters. The program interacts with the MarkerView software so that group members are assigned a symbol in the loadings plot which is immediately updated. This allows the user to experiment with the parameter settings and observe the results.

In addition to the parameters and operation described above, the program also provides two convenience features: (1) The program can use the active PCs, i.e., those used in the current scores and loadings plots, a fixed number of PCs, or the number that explain a given percentage of the total variance. (2) Optionally, groups can contain correlated and anticorrelated variables together. If this option is selected the program will consider variables within a given spatial angle on either side of the origin, otherwise the two sides are treated separately.

Although there are guidelines for selecting the number of PCs,[6,7] the choice of α depends on the complexity of the data and the desired group correlation. In practice we interactively vary these parameters while observing the effect on the loadings plots and select values that best reflect the loadings structure. For our example mass spectrometry data we expect correlated peaks, and

so we also use the "minimum number of variables" parameter to avoid groups containing just one or two variables.

The resulting groups contain variables that have similar expression profiles across all samples, and some will arise from the same compound. In the case of LC−MS data these will have the same retention time, i.e., they are isotopes, adducts, etc., and can be used to indicate the molecular ion and generate elemental compositions. Group members with the same retention time that are not related in these known ways may be fragments that can be used to aid structure identification. Group members that have different retention times arise from different compounds with similar behavior and depending on the experiment it may be possible to interpret these further, for example, to determine if they are part of a common metabolic pathway.

Further analysis can be simplified if desired by replacing the entire group, or members with the same retention time, by a single group representative that could be the most intense variable, a weighted average, etc. This dimensionality reduction is performed in the original data space and is much easier to interpret than the PCs, especially if used as the input for techniques such as linear discriminant analysis (LDA) or to generate new variables.

EXPERIMENTAL SECTION

Data Processing. Data processing was performed using the MarkerView software and the tool described above. Briefly, the MarkerView software performs all of the steps needed to analyze MS or LC−MS data with PCA, i.e., feature extraction (peak finding), alignment, normalization, and scaling. LC−MS peak finding[16] seeks clusters of ions that satisfy minimum peak width criteria in mass and time, and these are then processed using a spectral peak finder that is also used to process individual mass spectra. Alignment is performed using mass and retention time windows; if two peaks (from the same sample or different samples) are in the same window they are assumed to arise from the same peak.

Our preference is to use narrow windows for peak finding, so that small peaks are more likely to be found, but larger windows for alignment since intersample changes are often greater, especially for retention times. For example, for QqTOF data we tend to use mass windows of 2−5 ppm for peak finding and 10−20 ppm for alignment. Minimum LC peak width criteria for peak finding depend on the chromatography but are typically equivalent to 2−8 scans, depending on the expected peak widths. LC alignment tolerances are usually 30−60 s and depend largely on the total analysis time.

For LC−MS data the peak label contains m/z, retention time, and an ordinal number. In the interest of conserving screen space so that more labels can be displayed, the mass and retention times used in the labels are rounded to one decimal place, but the peaks table retains the data with full precision, and the ordinal number assists in locating them.

The results of PCA are evaluated by examining scores and loadings plots and by generating profile plots that show variable response for all samples. The display order used for profile plots can correspond to the sample class, the acquisition order, or some other user specified factor, and it is useful to quickly switch between these modes so that class-related changes and systematic effects can be distinguished and identified. Visualization is further enhanced by assigning separate symbols to sample classes and variable groups and using these in the various plots. This is especially powerful for loadings plots that depict the projection of hundreds or thousands of variables in a low dimensionality display and where the groups have been determined automatically as described here.

Finally, the MarkerView software can link profile plots to the raw data so that putative changes can be reviewed.

Example Data Set. Our approach to finding and visualizing correlated variable groups is illustrated using a set of LC−MS data acquired on an QSTAR XL (Applied Biosystems | MDS Sciex, Toronto, Ontario, Canada) from pre- and postdose urine samples from three rats that were given a single 20 mg/kg dose of vinpocetin,[17] an over-the-counter medication widely available in Europe. Following a 4 h fasting period, urine samples were collected at three time periods: 0−8, 8−16, and 16−24 h; vinpocetin was then administered, and samples were collected for the same time periods. Samples were diluted 10×, spun for 10 min at 4 °C, and kept frozen at −20 °C until analyzed.

Data was acquired in a random order using positive electrospray ionization at a rate of 5 scans/s for 30 min; the declustering potential (DP) was set to 90 V in order to induce fragmentation. Peaks were found using the background subtraction feature of the MarkerView software, a minimum intensity of 1 count, a minimum peak width of 2 ppm, and a minimum LC peak width of 20 scans for peaks with a retention time greater than 1 min. Alignment used 25 ppm and 0.4 min tolerance windows. This resulted in 209 peaks which were reduced to 180 by further requiring that variables be present in at least two samples.

In comparison to a real, long-term metabolomics study or one designed to find biomarkers, this is a small sample set, but the number of variables is still greater than the number of samples. It demonstrates several effects and illustrates the use of our approach without introducing great complexity. We have used the approach for the detailed examination of more complex data sets, but these are outside the scope of the current report and will be described elsewhere.

RESULTS AND DISCUSSION

Figure 1 shows the PC1/PC2 scores and loadings plots obtained by performing PCA with Pareto scaling on the LC−MS data set. Symbols in the scores plot indicate pre- and postdose samples for each of the three time points using filled and open circles to indicate dose status and color to indicate time. In addition, the symbols in the scores plot have labels indicating the individual animal (R2, for example), the time point (0−8), and the dose status (post or pre). The scores plot shows that PC1 (58% of the total variance) separates the pre- and postdose samples and PC2 (18.2%) seems to reflect the sample collection time point with the 0−8 h samples having the largest negative loadings. The greatest vinpocetin induced difference is observed for the 0−8 h samples with the 8−16 and 16−24 h samples returning toward the corresponding predose values.

As mentioned, the loadings plot is typically harder to interpret than the scores plot although the use of Pareto scaling indicates

(16) Sangster, T. P.; Wingate, J. E.; Burton, L.; Teichert, F.; Wilson, I. D. *Rapid Commun. Mass Spectrom.* **2007**, *21*, 2965–2970.

(17) Vereczkey, C. *Eur. J. Drug Metab. Pharmacokinet.* **1985**, *10*, 89–103.

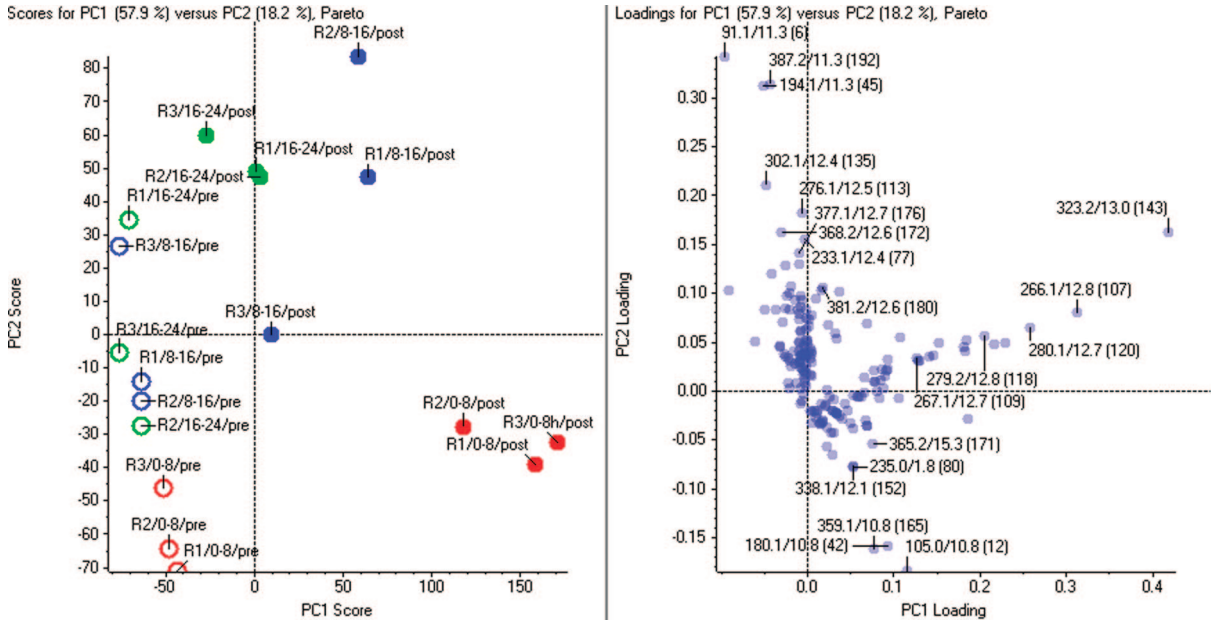

Figure 1. Results of performing PCA with Pareto scaling on the LC−MS data set. In the scores plot (left), the symbol indicates predose (open circles) and postdose (filled circles) and the color indicates the time point (red = 0−8, blue = 8−16, green = 16−24 h). The labels also indicate the individual animal (R1 = rat 1 etc.). Labels in the loadings plot (right) contain the *m*/*z* value, the retention time, and an ordinal number.

structure in the form of radial clusters of variables. In general these are correlated, but since the plot corresponds to the projection of a multidimensional data set on to two dimensions, variables that appear to be part of the same cluster may actually be separated in another dimension. In this case, variables with positive PC1 loadings, particularly those in the direction of the variable designated "323.2/13.0 (143)", appear largely responsible for the separation on PC1.

Prior to grouping the variables we examined the percent variance explained by each PC to determine the appropriate number to use. For this data the first four PCs explain 90% of the total variance suggesting that four or five is a good starting point. To illustrate how the grouping algorithm functions, we applied the algorithm using three, four, and five PCs, with angles ranging from 10° to 90°, with and without requiring a minimum of two variables per group, and without length filtering (i.e., using all variables); the results are shown in Table 1. In all cases the number of groups decreases dramatically as the angle is increased seeming to level in the range of 4−10 groups before a final decrease at 90°. Since there are only 180 variables in total, the large number of groups produced with very small angles clearly indicates that these values are too small; many of the 106 groups produced for five PCs with a 10° angle can contain only a single variable. Reasonable numbers for the angle seem to be in the 40−50° range, but the fact that the number of groups is roughly constant over a wide range of parameters suggests that, for this data at least, the values are not critical. We have observed similar behavior in data sets of varying complexity and believe that if number of PCs is chosen appropriately,[6,7] suitable angles will generally be in the 30−50° range. As can be seen, the effect of introducing the requirement for a minimum number of group members is dramatic, especially for small angle values, and again indicates that many of the groups have single members. As mentioned previously, when groups are rejected for this reason their variables are assigned to a special "too few peaks" group; the asterisks indicate the presence of this group and show the range of parameters for which all variables are assigned to a group.

Table 1. Number of Groups as a Function of the Number of PCs and the Angle, with and without a Requirement That a Group Have a Minimum Number of Members[a]

	min peaks per group = 0			min peaks per group = 2		
	Number of PCs			Number of PCs		
	3	4	5	3	4	5
angle (deg)						
10	49	77	106	17*	19*	14*
20	21	37	53	10*	13*	14*
30	13	23	31	9*	10*	15*
40	9	12	18	5*	9*	9*
50	5	9	11	5	7*	9*
60	5	6	7	5	5	6
70	5	5	5	5	4	5
80	5	3	5	4*	3	5
90	3	3	4	3	2*	3*

[a] Asterisks indicate groups that additionally have a "too few peaks" group.

To illustrate the technique we chose to initially use four PCs, an angle of 40°, all peaks (no minimum length criteria), and not restrict the number of members in a group. This resulted in 12 groups being generated, as in Table 1, and so 12 symbols were assigned to the variables and used to color the loadings plots from Figure 1; the result is shown in Figure 2. The left panel of Figure 2 shows the PC2/PC1 loadings and is identical to the right panel of Figure 1. The coloring immediately shows that there is a significant group of correlated peaks with positive PC1 loadings (group 2, green squares) and another with negative PC2 loadings (group 3, dark blue squares) but that the variables with large positive PC2 loadings

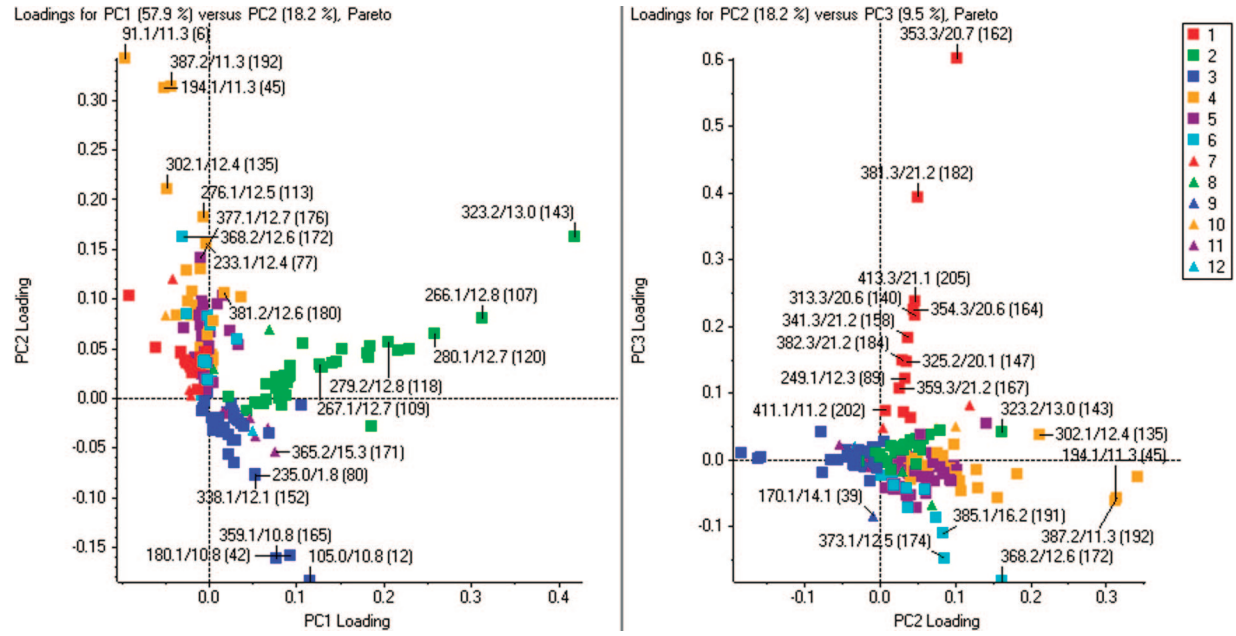

Figure 2. Loadings plots for PC1/PC2 (left) and PC2/PC3 (right) after assigning groups and symbols using four PCs and an angle of 40°. The legend in the right panel depicts the symbols used for both plots.

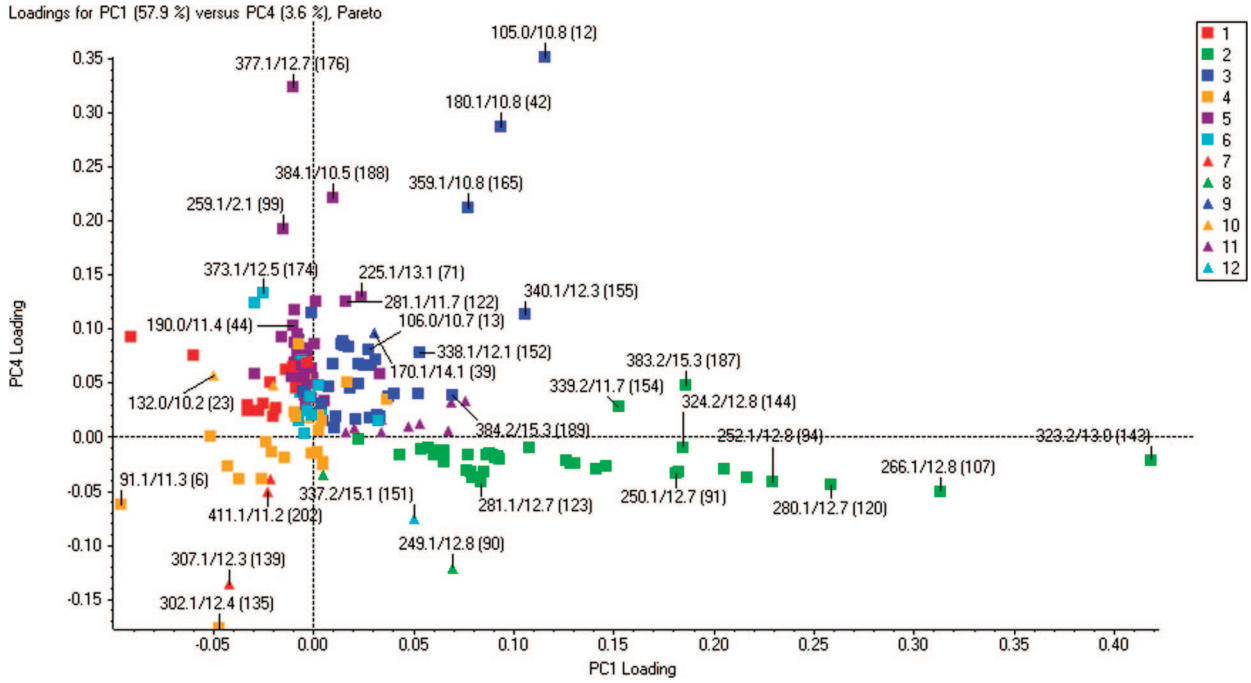

Figure 3. PC1/PC4 projections for the loadings from Figure 2.

appear to come from several groups (4, 5, 6, and perhaps others). Furthermore, group 1 (red squares), which is initiated by the variable with the longest loading vector in the entire data, does not have the largest loading on either PC1 or PC2. These observations indicate that there are significant effects in higher PCs as can be seen in the right-hand panel of Figure 2 where it is clear that group 1 has major loadings on PC3 (and its most intense variable has a larger loading than any of the variables in the left-hand panel) and also that groups 4 and 6 are now separate. It must be stressed that these observations are not apparent without grouping and subsequent symbol assignment and that techniques such as VarDia and biplots would not detect the nuances since they are restricted to two dimensions. In fact for these data the PC1/PC4 projection is one of the clearest at revealing the behavior of the variable groups as shown in Figure 3.

Reviewing group profiles can help determine the origin or nature of the groups, and flexibility in the displays can reveal effects that are due to instrument changes during the analyses. The latter point is illustrated in Figure 4 which shows profile plots for group 1 from Figure 2 plotted in group order (postdose samples in sample collection time order before the predose samples) in the lower panel and in acquisition order in the upper panel. Clearly there is no

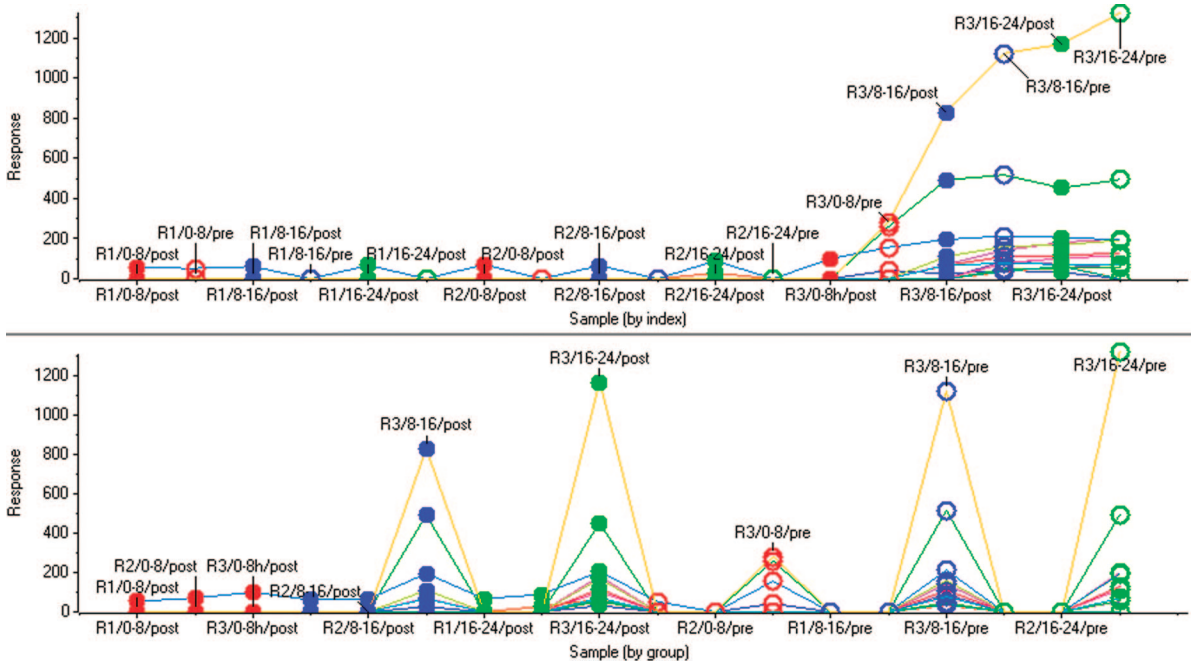

Figure 4. Profile plots for group 1 (from Figure 2) displayed in group order (lower panel) and acquisition order (upper) using the same symbols as Figure 1. The acquisition order plot indicates that these variables are associated with a systematic variation that is not sample related.

systematic pattern in the lower panel, but the upper panel shows that these ions are due to some instrumental change, such as contamination buildup, that occurred later in the data acquisition. Displaying profiles in this way is very powerful since it allows related ions to quickly be detected, their group or time behavior assessed, and appropriate action taken. In this example, these ions are not relevant to the analysis and can be excluded from subsequent processing. It should be noted that the data was not normalized prior to this analysis since doing so might have obscured this pattern. Sample-related normalization can be performed if required once the artifacts have been removed.

After excluding the ions from group 1, PCA was repeated and PCVG performed under the same conditions (four PCs, 40°) but additionally requiring that groups have at least two members; this resulted in five groups as shown in Figure 5. Figure 6 shows the profiles for each of these groups; for clarity only the most intense member of each group is shown. Visual inspection combined with examination of the peak information (not shown) indicates how many ions have the same retention time and are likely to belong to the same compound and how many compounds share the same profile, for example:

• Group 1 appears to be due to diurnal changes since the intensities show similar variations with sample collection time regardless of dose state; the compounds in this group are most intense in the 0−8 h samples and then decrease. The group contains 29 ions with several different retention times.

As an example of using group information to aid interpretation, this group contains a number of ions with a retention time of 10.8 min and rounded m/z values of 180.1, 181.1, 202.1, 359.1, 360.1, 105.0, and 106.0. Since these are related by behavior as well as retention time, it seems likely that they are formed from a compound with a molecular ion (MH^+) at 180.1, with 181.1 being the ^{13}C isotope, 202.1 being the $(M + Na)^+$ adduct, 359.1 the proton-bound dimer, and 360.1 its ^{13}C isotope, and 105.0/106.0 a fragment and isotope peak. The accurate mass of the MH^+ ion, available in the MarkerView software peak table, is 180.0614 which matches the elemental composition of hippuric acid within 4 mmu, and the presence of the fragment at 105.0297 is consistent with the loss of a glycine moiety, also within 4 mmu.

• Group 2 contains 33 ions of which 31 have a retention time around 12.8 min. These correspond to a vinpocetin metabolite,[17] since they are absent from the predose samples, that forms quickly as the intensity is greatest in the 0−8 h samples. The accurate mass of the most intense ion in this group corresponds within 4 mmu to the loss of ethylene from vinpocetin.

• Group 3 is the inverse of group 1, i.e., it corresponds to diurnal compounds that are lowest in the 0−8 h samples. There is some difference between the post- and predose samples, and between individual animals, which might correspond to drug-related compounds, but this cannot be confirmed without further work. This group has 24 ions having several retention times.

• Group 4 has 46 relatively weak ions at a variety of retention times. The behavior of the weakest ions is hard to determine due to noise, but the profile shown suggests some diurnal variation that may also distinguish individual animals since rat 1 (sample names include R1) are always higher.

• Group 5 has 26 ions, predominantly at two retention times, and is also probably due to the presence of one or more metabolites since the ions are observed only in the postdose samples. The profile differs from group 2 in that the intensity is much lower in the 8−16 and 16−24 h samples. This behavior can arise for two reasons: (1) these are weak peaks related to those in group 2 that are not detected in later time point samples where the amount is lower, and (2) they arise from compounds that are in fact depleted (excreted or have reacted further) in the later samples. The first hypothesis is somewhat supported by the presence of peaks from around 12.8 min as in group 2.

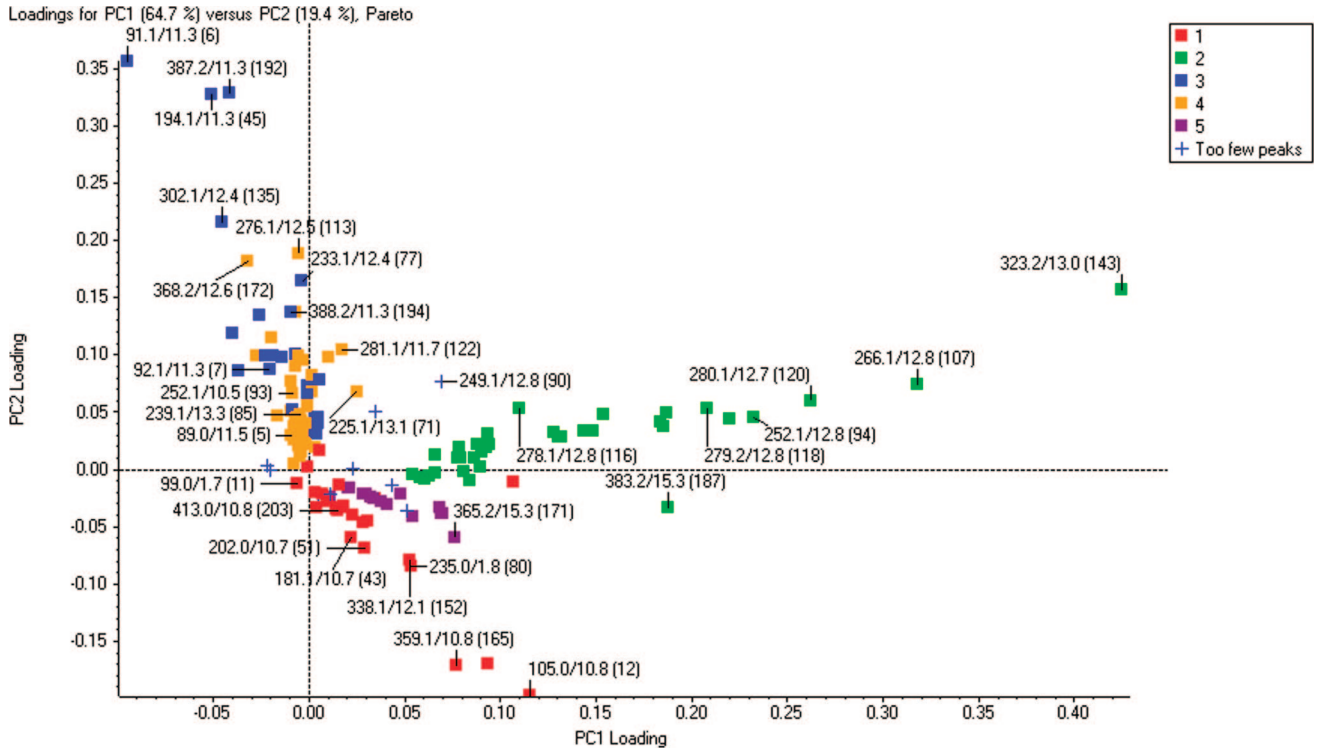

Figure 5. Loadings plot for PC1 and PC2 after removal of the contamination group and reassignment of the groups using four PCs, 40°, and additionally requiring that a group have at least two members.

Although it is common to visualize the scores and loadings for two or three PCs, it is sometimes useful to generate loadings plots for a single PC but label the individual variable values based on the groups determined from a larger number of PCs. The visualization that can be achieved is shown in Figure 7 which depicts the loadings for PC1 and PC2 individually with the same group symbols as in Figure 5 added. Some care must be taken in interpreting these plots since it is tempting to regard them as line spectra, whereas they are in fact representations of the underlying peak (variable) list; that is, the retention time, which is part of the peak label, must also be considered. Nevertheless the display can be useful since it shows the results in a familiar form and can help clarify the relationship between variables. Here, for example, the dominance of group 2 (green circles) in PC1 is apparent, as is the fact that many of the peaks have a retention time around 12.8 min with the peak at 383.2/15.3 being a clear exception. Likewise, the negative loadings of PC2 are dominated by group 1 (red squares) with a main component around 10.8 min; the relationship between the molecular ion, dimer and fragments, and their ^{13}C isotope peaks is prominent and can be tentatively assigned to hippuric acid as described above. The positive PC2 loadings are dominated by peaks with two retention times, 11.3 min (m/z 91.1, 194.1, 195.1, 387.2, 388.2) and 12.4 min (m/z 233.1, 302.1, and 404.2 unlabeled). The peaks at 11.3 min appear to have a molecular ion (194.1/195.1), dimer (387.2/388.2), fragment (91.1/92.1) relationship that is consistent with phenylacetylglycine; the accurate mass of the molecular ion matches this composition within 4 mmu.

As we have noted, identifying groups allows dimensionality reduction by selecting a chemically meaningful representative either of the entire group or of the individual components as determined by the retention times. To illustrate this we performed PCA using only the most intense variable in the first three groups to produce the scores and loadings plots shown in Figure 8. Comparison with Figure 1 shows that the sample separation is very similar. In addition, group representatives provide good compounds for use in a targeted assay that could generate high-quality data for use in classification.

CONCLUSIONS

Data analysis is important in processing the megavariate data sets that can be rapidly generated by modern analytical tools such as NMR, LC−MS, and microarrays. PCA is a popular processing tool since it analyzes the variance present in the data in an unsupervised manner, i.e., considers both the expected and unexpected variance, and can work with high dimensionality data that contains many variables. PCA results are commonly plotted in two or three-dimensional plots that reflect the behavior of the samples (scores plot) or variables (loadings plot). In general, interpretation of the scores plot is fairly straightforward, since an expected sample separation is either observed or not, but the loadings plot can be much harder to interpret. Visualization of the loadings is often further hampered by autoscaling, which can cause small peaks to appear important even if they are not. For mass spectrometry data, the situation can be somewhat alleviated by identifying ions that are known to be related to a single component and replacing them with a single variable which represents that component, but this relies on expected behavior (isotopes, adducts, etc.) and can be difficult when unexpected fragments are generated; detecting the latter is particularly valuable for spectral interpretation.

We have described an approach that uses PCA to generate loadings values that are analyzed to find variables that are correlated across the samples—in essence we are using the

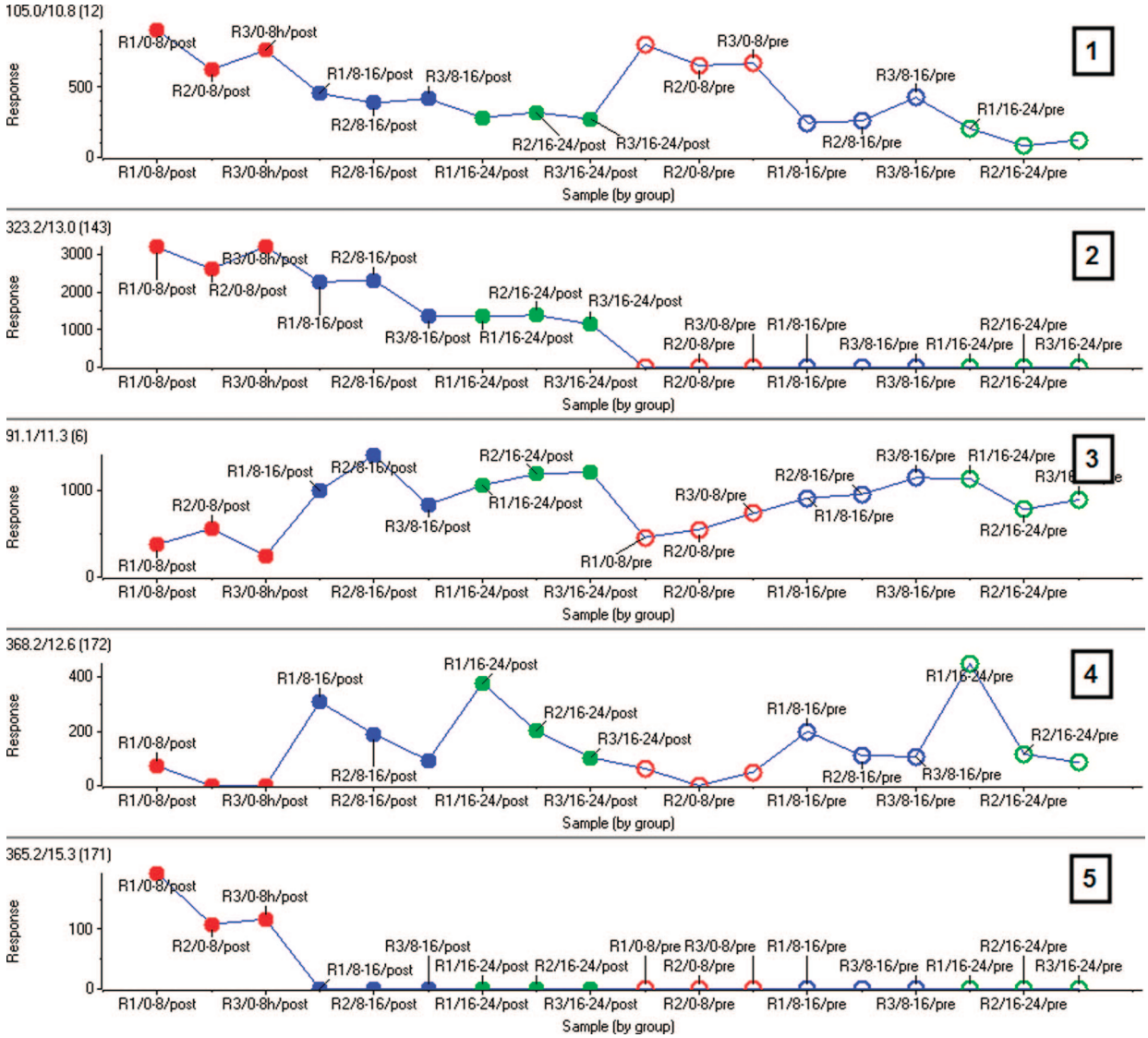

Figure 6. Profiles for the most intense ion in each of the five groups (bold numbers) depicted in Figure 4; the symbols used for the samples are as in Figure 1.

samples to group the variables in addition to the converse. Once the groups have been identified in *n*-dimensional PC space the corresponding variables are given a symbol that is used in projections of the loadings to fewer dimensions and greatly assists visualization and interpretation of the variable structure. The approach has some similarities to existing techniques but also has several distinct differences and advantages, for example:

• The underlying PCA analyzes the variance in the data and depicts the relationship between the samples and variables, whereas K-means, SOMs, and HCA do not analyze the variance and K-means and SOMs deal with the samples and variables separately.

• The technique is simple, intuitive, and requires just two parameters that control grouping and variable visualization without affecting the PCA results. It does not require an estimate of the expected number of clusters, unlike K-means and SOM.

• Correlation is measured directly via the angle between vectors representing the variables; scaling methods other than autoscaling can be used and are preferred so that small noisy variables do not have an undue effect. HCA can use correlation which requires that the data be autoscaled.

• Any number of PCs can be used and visualized, unlike biplots, VarDia, and ContVarDia, so it is compatible with large, complex data sets with many response patterns.

Identifying variable groups in this way has a number of other advantages:

• Interpretation: Groups indicate variables that have similar behavior across the samples and other analysis-specific information may allow groups to be subdivided or aid interpretation. For example, in LC−MS data, group members that have the same retention time most likely originate from the same compound, and well-known relationships can be applied to identify features such as molecular ions, adducts, dimers, fragments, etc. Knowing that different compounds show similar behavior may also allow higher level interpretation, for example it may be possible to postulate that such compounds are part of the same biological pathway.

• Unsupervised pattern determination: The group profile, that is the response of group members across all of the

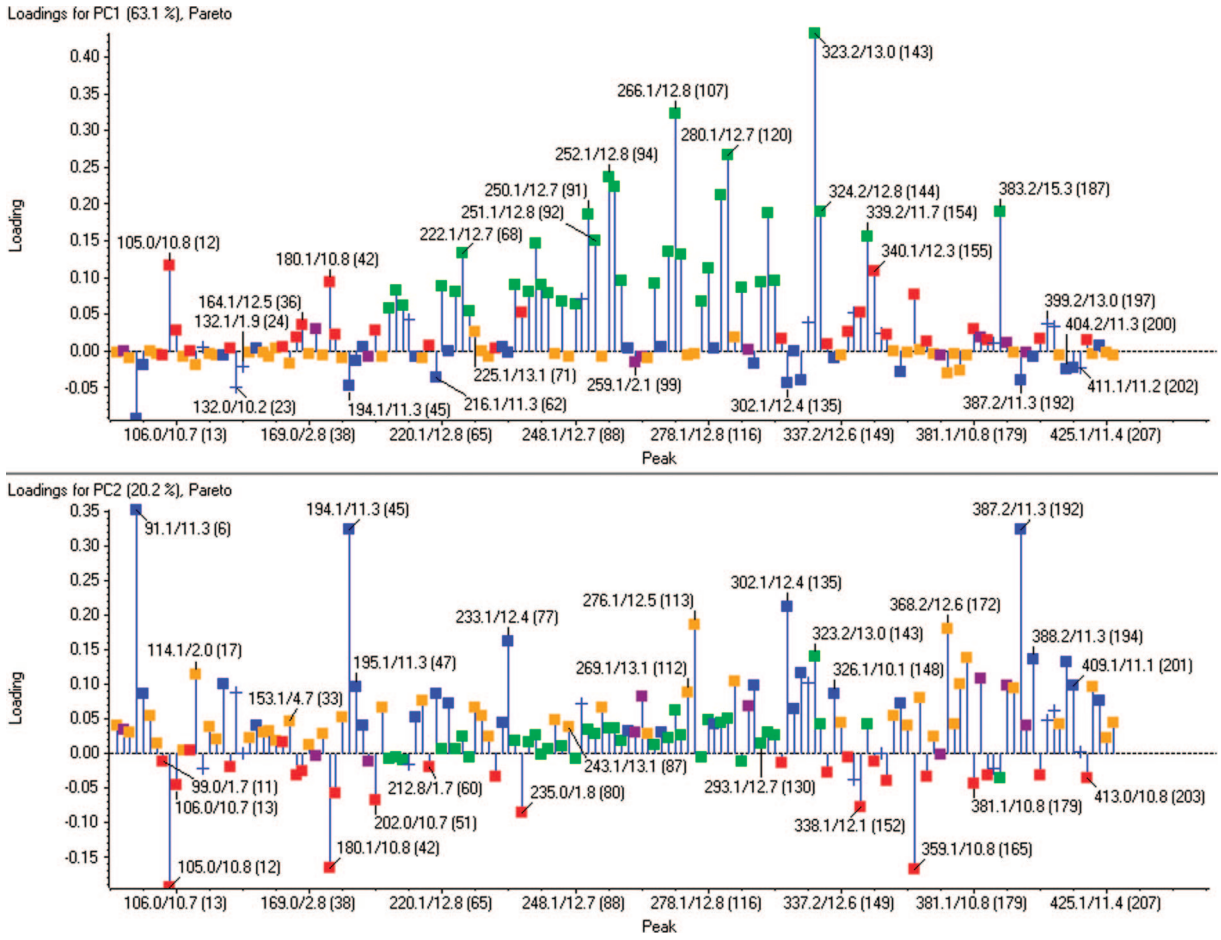

Figure 7. Comparison of loadings for PC1 (upper) and PC2 (lower) for the data of Figure 4. Individual variables are marked with the symbol corresponding to their assigned group using the symbols of Figure 4.

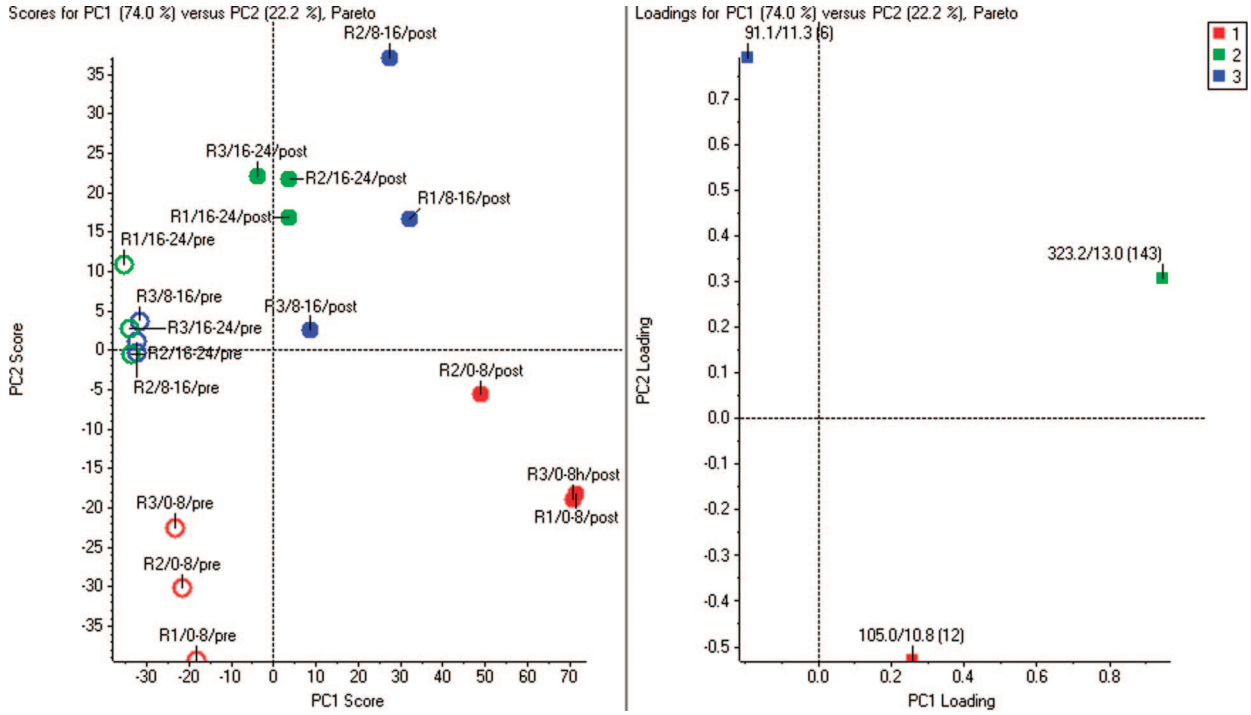

Figure 8. PCA using the largest member of the top three groups from Figure 4. Symbols used for the samples are as in Figure 1, and those for variables are as in Figure 4.

samples, can reveal the nature or origin of the variables even if their identity is unknown. Linking this to the ability to order the displays in different ways can help detect artifact changes that have experimental origin which can then be treated appropriately. In some cases the variables can be safely ignored, as here, but other circumstances may require reanalysis, redesign of the experiment, or data detrending. Other profiles may identify real variables causing effects that are not relevant to the study, such as diurnal or gender differences.

- Dimensionality reduction in the original data space: Since groups relate to the underlying chemistry, they also provide a good way to find characteristic variables that can be used to represent each group or subgroup. Here, for illustration, we have used the most abundant member of each group, but other values, such as the mean, median, or sum of the group intensities, may be more relevant or robust to noise. Group representatives could also be analyzed in targeted assays or used to build classifiers.
- Visualization: Grouping variables in *n*-dimensional PC space is a powerful tool to help visualization and interpretation of complex loadings plots representing a large number of variables. Other techniques, such as nonlinear mapping,[18] show relationships between samples, but the PCVG approach retains the ability of PCA to simultaneously depict relationships between samples and variables.

(18) Sammon, J. W., Jr. *IEEE Trans. Comput.* **1969**, *18*, 401–409.

We have applied PCVG in a variety of MS applications such as tissue imaging (where ions having similar spatial distributions are grouped) and xenobiotic metabolism and interspecies comparison studies, as well as metabolomics and peptide/protein expression analysis. Although our examples are in mass spectrometry, the approach is applicable to any data analyzed by PCA, for example, in NMR it would reveal peaks with different chemical shifts arising from the same compound and again aid interpretation. To further refine the technique, we plan to follow group assignment by a re-examination of the data to ensure that all variables, including those unassigned in the first pass, are assigned to the closest group. We also intend to augment group assignment by using additional knowledge such as retention time information and develop methods for automatically selecting the number of PCs and the angle.

ACKNOWLEDGMENT

We are grateful to Dr. Gérard Hopfgartner, University of Geneva, for the LC−MS data set and helpful discussions. We also wish to acknowledge an anonymous reviewer for helpful comments and for drawing our attention to ref 13.

Received for review January 15, 2008. Accepted April 16, 2008.

AC800110W