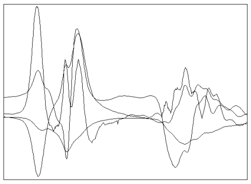

Chemometrics
Technical Note

InfoMetrix

Description of Pirouette Algorithms

This discussion introduces the three analysis realms available in Pirouette and briefly describes each of the algorithms used to execute multivariate analysis. A list of references has been included at the end of the discussion for further study on data exploration and modeling procedures.

One of the fundamental reasons for collecting data is to develop a sufficient understanding of that data to be able to use the information in characterizing future data sets. The goal of *exploratory data analysis* is to provide a quality check on the data; it does this by determining the key measurements in the data, exposing possible outliers, and indicating whether there is sufficient modeling power in the data collected to warrant further investigation. Ultimately, the purpose of most multivariate analyses will be to develop a model to predict a property of interest. That property may be categorical (*e.g.*, good or bad, species 1 or species 2 or species 3), or it may be a continuous property (*e.g.*, a concentration or bulk property) that cannot be measured directly. When the property of interest is a discrete category assignment, then *classification analysis* is the appropriate approach; continuous properties are modeled and predicted by *regression analysis* methods.

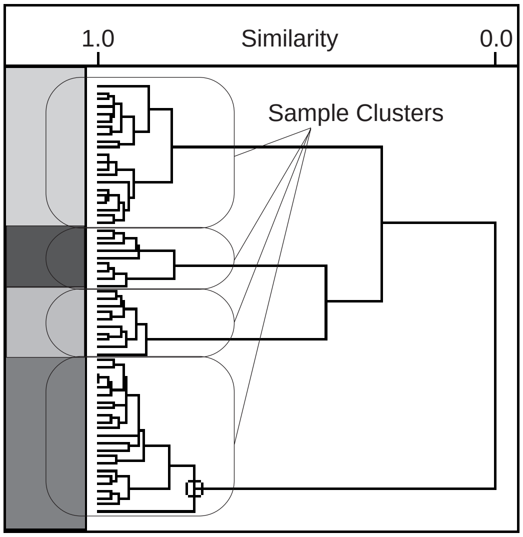

Figure 1
Example of the dendrogram results from an HCA analysis showing four distinct sample clusters

Exploratory Data Analysis

Exploratory data analysis is the computation and the graphical display of patterns of association in multivariate data sets. The algorithms for this exploratory work are designed to reduce large and complex data sets into a set of best views of the data; these views provide insight into the structure and correlations that exist among the samples and variables in your data set.

Hierarchical Cluster Analysis (HCA)

In HCA, distances between the samples (or variables) in a data set are calculated and compared. When the distances between samples are relatively small, this implies that the samples are likely similar, at least with respect to the measurements taken. Dissimilar samples will have larger relative distances. Known in biological sciences as numerical taxonomy, hierarchical cluster analysis allows the grouping of data into clusters showing similar attributes.

The primary purpose of HCA is to present data in a manner that emphasizes the natural groupings in that data set. In contrast with analytical techniques that attempt to group new samples into pre-existing categories, HCA seeks to define those categories in the first place. The presentation of HCA results in the form of a dendrogram makes it possible to visualize clustering such that relationships can be more readily seen.

Principal Component Analysis (PCA)

PCA is designed to provide you with the best possible view of the variability in a multivariate data set. This view allows you to see the natural clustering in the data, identify outliers (*i.e.*, unusual samples) and find the reasons behind any pattern that is observed. In addition, the intrinsic dimensionality of the data can be determined and, with variance retained in each factor and the contribution of the original measured variables to each, this information can be used to assign chemical meaning (or biological meaning or physical meaning) to the data patterns that emerge and to estimate what portion of the measurement space is noise.

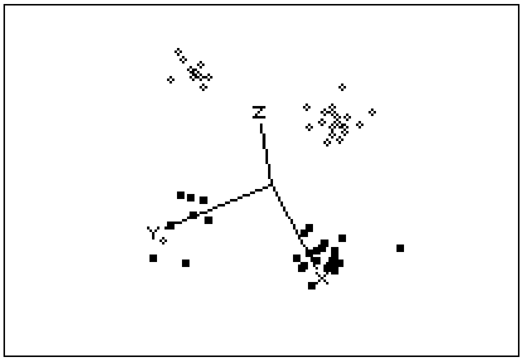

Figure 2
A 3D scores plot from PCA rotated to show the same four clusters as in Figure 1

PCA is fundamentally similar to factor analysis or eigenvector analysis (mathematics) or even perceptual mapping (marketing). It is a method of transforming complex data into a new perspective in which, hopefully, the most important or relevant information is made more obvious. This is accomplished by constructing a new set of variables that are linear combinations of the original variables in the data set. These new variables, often called eigenvectors or factors, can be thought of as a new set of plotting axes which have the property of being orthogonal (*i.e.*, completely uncorrelated) to one another. In addition, the axes are created in the order of the amount of variance in the data for which they can account. As a result, the first factor describes more of the variance in the data set than does the second factor, and so forth. The relationships between samples are not changed in this transformation, but because the new axes are ordered by their importance (*i.e.*, the variance they describe is a measure of how much distinguishing information in the data they contain), we can graphically see the most important differences between samples in a low-dimensionality plot (hopefully, three or fewer).

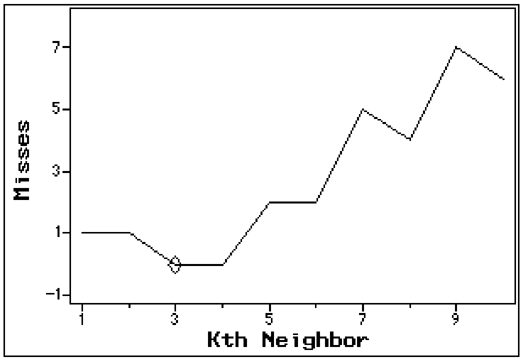

Figure 3
A 3 nearest neighbor model minimizes the number of misclassifications in this KNN example

Classification Analysis

Classification in the Pirouette analysis system is the computation and the graphical display of class assignments based on the multivariate similarity of one sample to others. The algorithms for this classification work are designed to compare new samples against a previously-analyzed experience set.

In contrast with KNN, which is based simply on Euclidean distances among sample points, SIMCA develops principal component models for each category in the training set. When classifications are attempted for unknown samples, a comparison is made between the unknown's data and each class model. The model which best fits the unknown, if any, is the class assigned to that sample.

K-Nearest Neighbors (KNN)

Classification with KNN is based on a distance comparison among samples: an N–dimensional distance between all samples in the data set is calculated, where N is the number of variables in the measured data. The predicted class of a test sample is then determined based on the identity of those samples closest to the unknown sample. This is accomplished in a fashion analogous to voting: each of the K nearest samples votes once for its class; the class receiving the most votes is assigned to the test sample.

KNN is tolerant of sample-poor situations and is the only method which works well even when categories are strongly subgrouped. SIMCA (see next algorithm) assumes class homogeneity and does not work as well when strong sub-groupings are present.

Soft Independent Modeling of Class Analogy (SIMCA)

Reliable classification of unknown samples is the ultimate goal of a SIMCA analysis. However, there are other desirable questions which can be answered by SIMCA as well. For example, we can use SIMCA to help distinguish among the most important variables to retain in the training set and to ascertain the presence of likely outliers. By examining the variance structure within each class, we can come to understand the complexity of a category, and use this information to further refine the effectiveness of the training data.

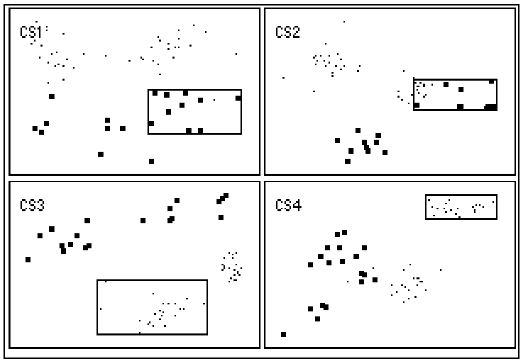

Figure 4
A SIMCA display showing sample fit for each of four separately modeled classes

Another important aspect for classification with SIMCA is the ability not only to determine whether a sample does belong to any of the predefined categories, but also to determine that it does not belong to any class. This is also in contrast to KNN, which will yield a class prediction regardless of whether the prediction is reasonable or not. Class predictions from SIMCA fall into three possible outcomes:

1. *The sample is properly classified into one of the pre-defined categories*
2. *The sample does not fit any of the categories*
3. *The sample properly fits into more than one category*

We can place confidence limits on any of the outcomes as well, because these decisions are made on the basis of statistical tests.

Comparison of Discriminant Analysis to KNN and SIMCA

A discriminant analysis function was not added to Pirouette, because the KNN technique largely duplicates the function of discriminant analysis, resulting in models that are very similar in predicting power and easier to conceptualize. The situations where discriminant analysis and KNN could give different results are only where the answer is highly uncertain (an extreme extrapolation of a model or an area where the distinction between two groups is not clear). In addition, the discriminant analysis technique does not have a good diagnostic for unusual samples, a problem shared by KNN but solved by the SIMCA algorithm.

Regression Analysis

When the property of interest is difficult to measure directly, regression methods can be used to predict the value of that property based on related properties which are easier to measure. The goal of a regression analysis is to develop a model (called a calibration model) which correlates information in a set of measurements to some desired property (or properties).

To fully test a model created in the calibration step requires a validation procedure. In general, validation entails the application of a model to test samples for which the properties are already known. Thus, by comparing the predicted values to the known, we can establish a measure of reliability for the model.

Partial Least Squares (PLS) and Principal Component Regression (PCR)

Pirouette includes two of the most effective multivariate regression methods, both of which employ factor analysis principles: Principal Component Regression (PCR) and Partial Least Squares (PLS) regression. The utilization of the two methods is identical, but in practice you would want to ascertain which is the more appropriate method for your data, then stick with the preferred method for future work with that type of data.

These multivariate regression methods achieve their goals in basically the same manner, with one important difference. PCR uses steps similar to those used in PCA to decompose the data matrix (of independent variables) into principal components, then relates the calculated objects from the decomposition to the dependent variable(s). The relationship which is created is reduced to a regression vector, which can be used subsequently to predict a value of the dependent variable for new, test samples.

PLS approaches the decomposition of the independent variables in a similar way, but with a twist: during the steps of the decomposition, information extracted from the independent variable matrix is passed to the dependent variable vector and vice versa. The result from PLS is also a regression vector, but one in which correlations between the independent block of data (often referred to as the X block) and the dependent block (the Y block) are included.

Comparison of Multiple Linear Regression (MLR) to PCR and PLS

MLR is a common modeling technique applied to various types of data, in which the responses for each measured variable are expected to behave more or less independently. Thus, with MLR, you are restricted to cases in which the number of variables does not exceed the number of samples. As a result, MLR suffers from a tendency to overfit the data if there is any noise in the measurements.

If you have a specific need, you can do MLR in Pirouette by performing a PCR analysis and having the algorithm compute all components. The model with all components retained is the MLR solution. In order to see if MLR is an appropriate technique, you would compare the validation errors of MLR against the validation errors for PCR or PLS.

References

Exploratory data analysis

Massart, D.L.; Vandeginste, B.G.M.; Deming, S.N.; Michotte, Y.; and Kaufman, L; Chemometrics: a textbook, (Elsevier: Amsterdam, 1988).

Classification analysis

Forina, M. and Lanteri, S.; "Data Analysis in Food Chemistry" in B.R. Kowalski, Ed., Chemometrics. Mathematics and Statistics in Chemistry (D. Reidel Publishing Company, 1984), 305-349.

Sharaf, M.A.; Illman, D.L.; and Kowalski, B.R.; Chemometrics (Wiley: New York, 1986).

Regression analysis

Martens, H. and Naes, T.; Multivariate Calibration (Chichester: John Wiley & Sons, 1989).

Thomas, E.V. and D.M. Haaland; "Comparison of Multivariate Calibration Methods for Quantitative Spectral Analysis", Anal. Chem. (1990) 62: 1091-1099.

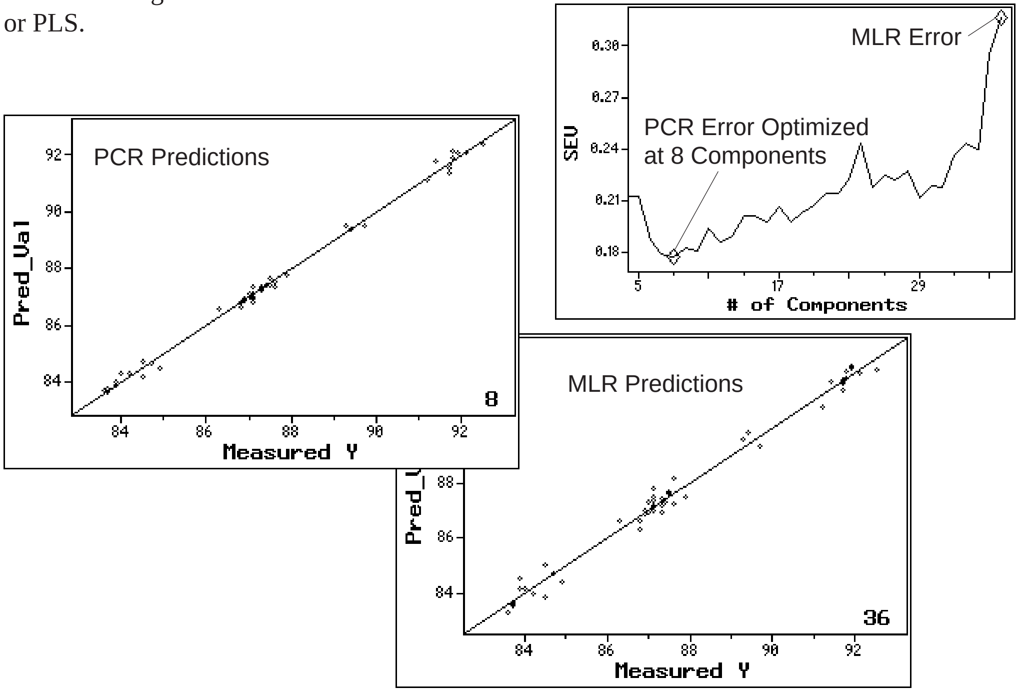

Figure 5
Error and prediction plots for PCR and MLR show the model optimum when an eight principal component (PCR) model is employed