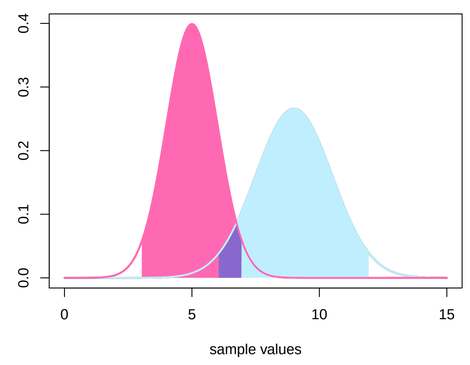
Implementation of ANOVA-PCA in R for Multivariate Data Exploration

Matthew J. Keinsley & Bryan A. Hanson

Dept. of Chemistry & Biochemistry
DePauw University, Greencastle Indiana USA

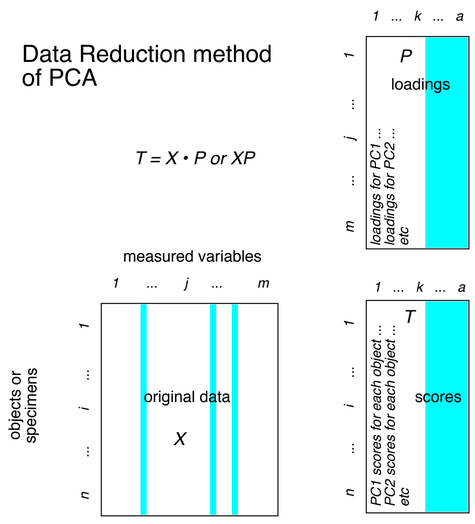
Analysis of Variance

Analysis of Variance (ANOVA) is a significance test which considers whether or not two samples come from the same population. In the diagram below, each curve represents a series of measurements on two samples, with each described by a mean and standard deviation. A null hypothesis is defined which typically states that the two samples are from the same population. ANOVA can be thought of as measuring how much the two samples can drift apart and still come from the same population (i.e. the null hypothesis is true). The key values from an ANOVA calculation are the test statistic and the p-value. The p-value represents the probability of achieving a test statistic more extreme than the one observed if the two samples are from the same population. The smaller the p-value, the more confident one is about rejecting the null hypothesis at a selected confidence level, usually 95%. Internally, the calculation compares the between sample variation to the within sample variation to generate the test statistic. This is referred to as "partioning the variance."



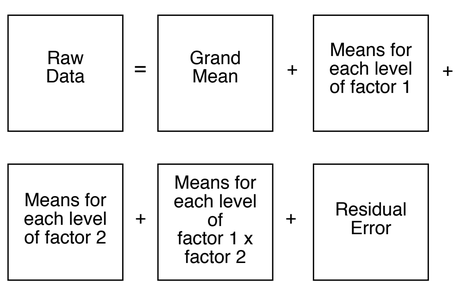
Principal Component Analysis

Principal Component Analysis (PCA) is a common multivariate data analysis method which eliminates uninformative variables (noise) and re-expresses the data in terms of abstract "principal components." The diagram below illustrates the matrix algebra used in the computation. The PCA results are scores and loadings. Scores represent the samples in the new data space, while loadings are the weights which each variable should be multiplied to obtain the score. While score plots are usually color-coded by group membership, *PCA is blind to group membership*.



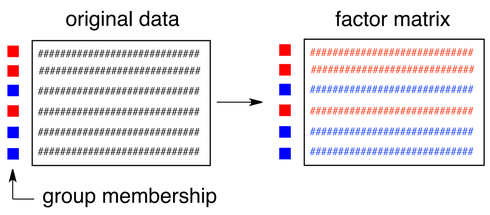
ANOVA-PCA

ANOVA-PCA is a combination of both methods developed by Harrington. The data is partitioned into submatrices (shown below) corresponding to each experimental factor in a manner reminiscent of ANOVA. The submatrices are then separately subjected to PCA after adding back the residual error. If the effect of a factor is large compared to the residual error, separation along the 1st PC in the score plot should be evident. With this method, the significance of a factor can be visually determined. *ANOVA-PCA is not blind to group membership.*



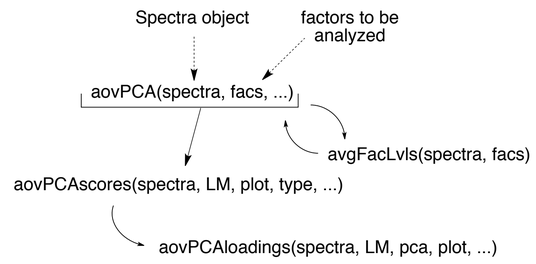
ANOVA-PCA con't

The method to create the submatrices is shown below. Data for all samples belonging to each level of a particular factor are replaced by their group averages.



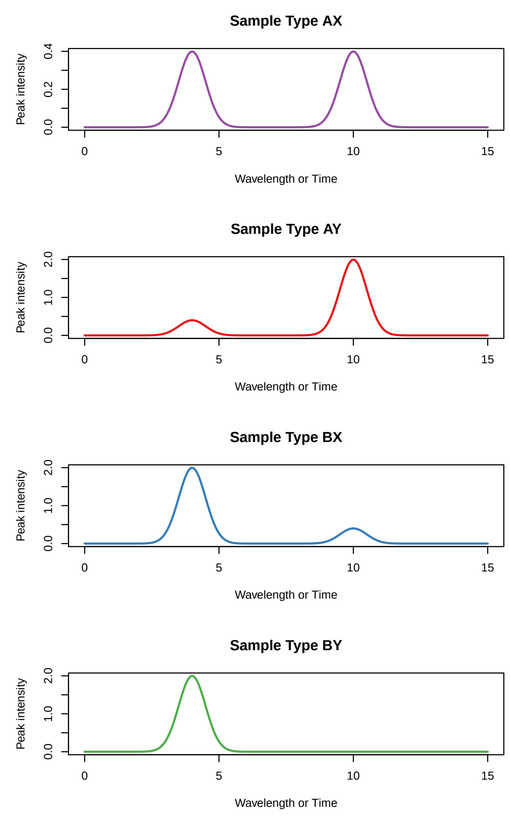
Computational Strategy

We implemented the ANOVA-PCA concept using a series of R functions as shown below. The functions were integrated into the package ChemoSpec, which uses a Spectra object to store the spectral data and associated information.



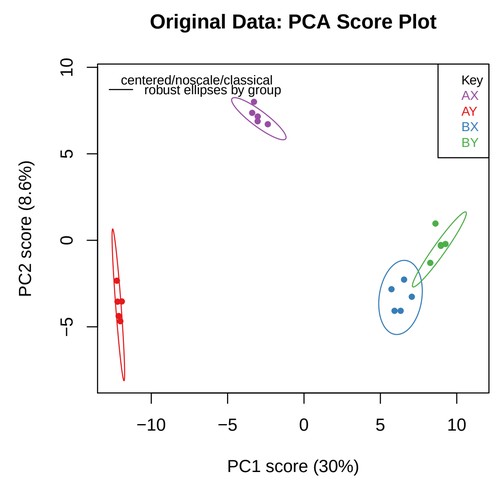
Simulated Data

For testing purposes, a simulated data set was created by generating data which simulate UV-Vis spectra or chromatograms. Four different sample types were created in such a way as to represent group membership and the effects of the factors. The levels of the first factor are A and B; the second factor has levels X and Y. The entire data set contains 5 of each sample type, and noise was added to the spectra in order to make them more realistic. The prototype samples are shown below; type BY represents a interaction factor as a peak is missing entirely compared to the other sample types.



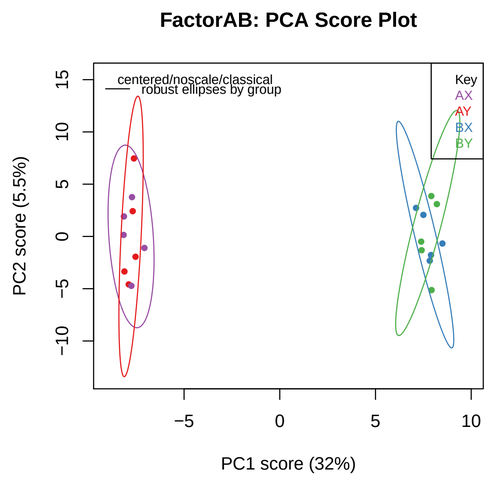
Standard PCA

The results of a traditional PCA analysis on these simulated data are shown below. As expected, this data set does not present a challenge and each group is clustered away from the others.

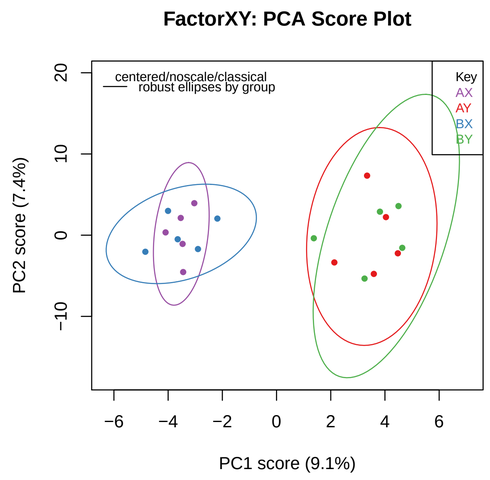


aovPCA Score Plots

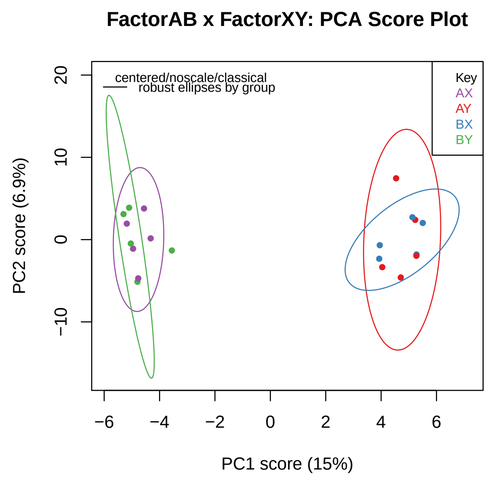
The following figures show the results of aovPCA on the simulated data. Factor 1 has levels A and B. Because the 1st factor is significant, there is separation between levels along PC1.



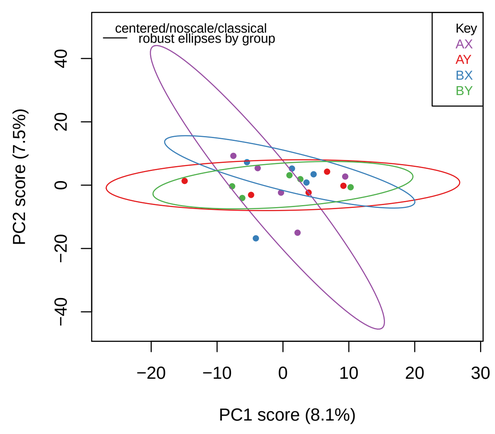
Factor 2 has levels X and Y. As before, separation is observed along PC1, but it is not as good as for Factor 1. Note the groups here are reversed compared to the 1st factor.



The following figure shows the interaction between Factors 1 and 2. In this case, AX and BY are grouped separately from AY and BX. This separation therefore shows that the interaction between the two factors is significant as expected.

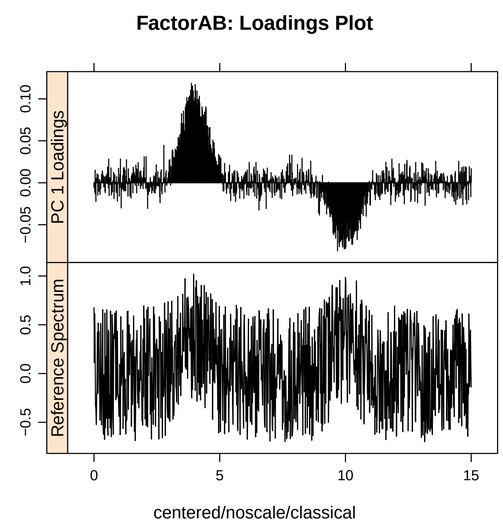


Finally, aovPCA on the residual error is shown. The residual error is the unexplained variance, so there should be no separation on PC1 nor any clustering. This is exactly what is observed.



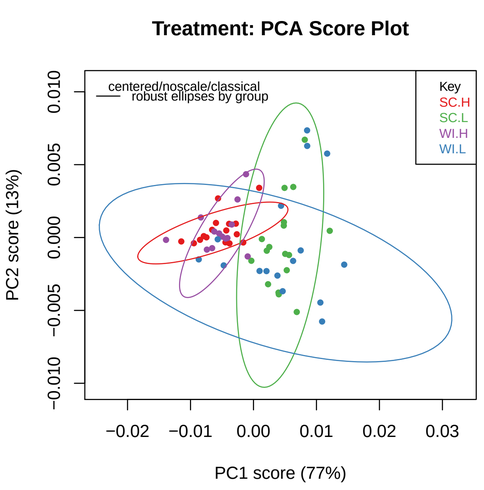
aovPCA Loadings Plots

The loading for Factor AB is shown below. Peaks on the PC1 loadings show which wavelengths contribute the most to the separation in the data. The reference spectrum clearly shows the noise that was added to the data.

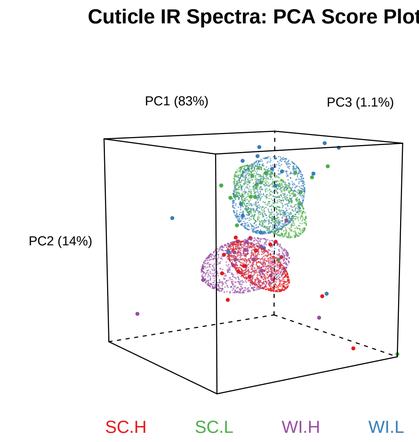


aovPCA of Real Data

The data in this example is composed of infrared (IR) spectra of the leaf surface of the common "weed" purslane (*Portulaca oleracea*). Varieties collected in South Carolina and Wisconsin were grown in low and high water conditions as a part of a larger study on climate change. The aovPCA score plot with Treatment as the factor is shown below; aovPCA is not able to separate the groups based upon this factor.



However, if the first 3 PCs from standard PCA are plotted, we can see separation between the treatments. With this difficult data set standard PCA outperforms aovPCA.



References

Pinto, Bosc, Nocairi, Barros, and Rutledge. "Using ANOVA-PCA for Discriminant Analysis..." *Analytica Chimica Acta* 629.1-2 (2008): 47-55.

Harrington, Vieira, Espinoza, Nien, Romero, and Yergey. "Analysis of Variance–Principal Component Analysis..." *Analytica Chimica Acta* 544.1-2 (2005): 118-27.

Software: R packages ChemoSpec and HandyStuff github.com/bryanhanson

Acknowledgements

Support was provided by the Science Research Fellows program and Chemistry Department research funds.