Exploring Multivariate Data with Sparse Projection Pursuit Analysis

Peter Wentzell
Department of Chemistry, Dalhousie University
(Joint work with Yannick MacMillan and Stephen Driscoll)

Abstract

Exploratory data analysis methods have become an essential part of modern chemical research involving multivariate measurements. In some fields, techniques such as principal components analysis (PCA) and hierarchical cluster analysis (HCA) are applied in more than 50% of published research articles. The popularity of these methods can be attributed to the intuitive visualization of data and their increasing role as de facto methods to support hypotheses in experiments involving a limited number of samples. Projection pursuit analysis (PPA) is an interesting alternative to traditional exploratory tools because it is based on finding "interesting" projections of the data. Although it has been found to be extremely effective in cases where other methods fail, it has several drawbacks, including a requirement for a high sample-to-variable ratio ("skinny data") and poor interpretability of the projection vectors. Sparse projection pursuit analysis (SPPA) is variant of PPA that addresses both of these issues, combining a kurtosis-based PPA algorithm with a genetic algorithm (GA) for variable selection. In this talk, the principles of PPA and its sparse implementation will be presented, with several examples from chemistry to illustrate its effectiveness for unsupervised clustering of data.