

## **Between principal axes analysis and clustering: the missing links.**

Ludovic Lebart  
Telecom-ParisTech, CNRS

### **Abstract**

Most practitioners consider clustering methods and principal axes techniques (principal components analysis, two-way and multiple correspondence analysis, etc.) as complementary approaches in the exploration of multivariate data sets. As far as visualizations of data are concerned, the enrichment resulting from the simultaneous use of both families of methods is widely recognized. However, the attempts to theoretically justify these pragmatic attitudes are rather scarce and scattered in the statistical literature. A review of works at the intersection of these two fields of research reveals yet a wealth of algorithms often adapted to various empirical contexts. At the outset, back to the first half of the twentieth century, the rotations using some specific criteria in the framework of factor analysis could be viewed as the first attempts to find clusters of variables... Let us mention as an illustration some techniques of divisive clustering that make use of a principal axis at each step to split the sample and build a descending tree. Others statisticians come up with the simultaneous visualization of both clusters and individuals onto the same display or recommend clustering from principal coordinates. The popular Self Organizing Maps (or Kohonen maps) do not deal explicitly with principal axes, but strive yet to reconcile the concept of cluster with that of (non-linear) subspace. When dealing with contingency tables, theoretical links can be established in some cases in which a particular clustering technique provides hierarchical indexes that coincide with the eigenvalues from the correspondence analysis performed on the same table. More generally, some inequalities between the indexes derived from a hierarchical clustering and the eigenvalues of the correspondence analysis of the same contingency tables can be established. In a quite different context, the introduction of local metrics leads to a series of hybrid methods resembling projection pursuit algorithms that enrich the visualizations of clusters. More recently, clustering techniques using the laplacians of graphs suggest new approaches. Finally, the graph structures defined by either the series of nearest neighbors or thresholds of distances appear to have interesting spectral properties. They may provide a valuable bridge between these two large families of methods.