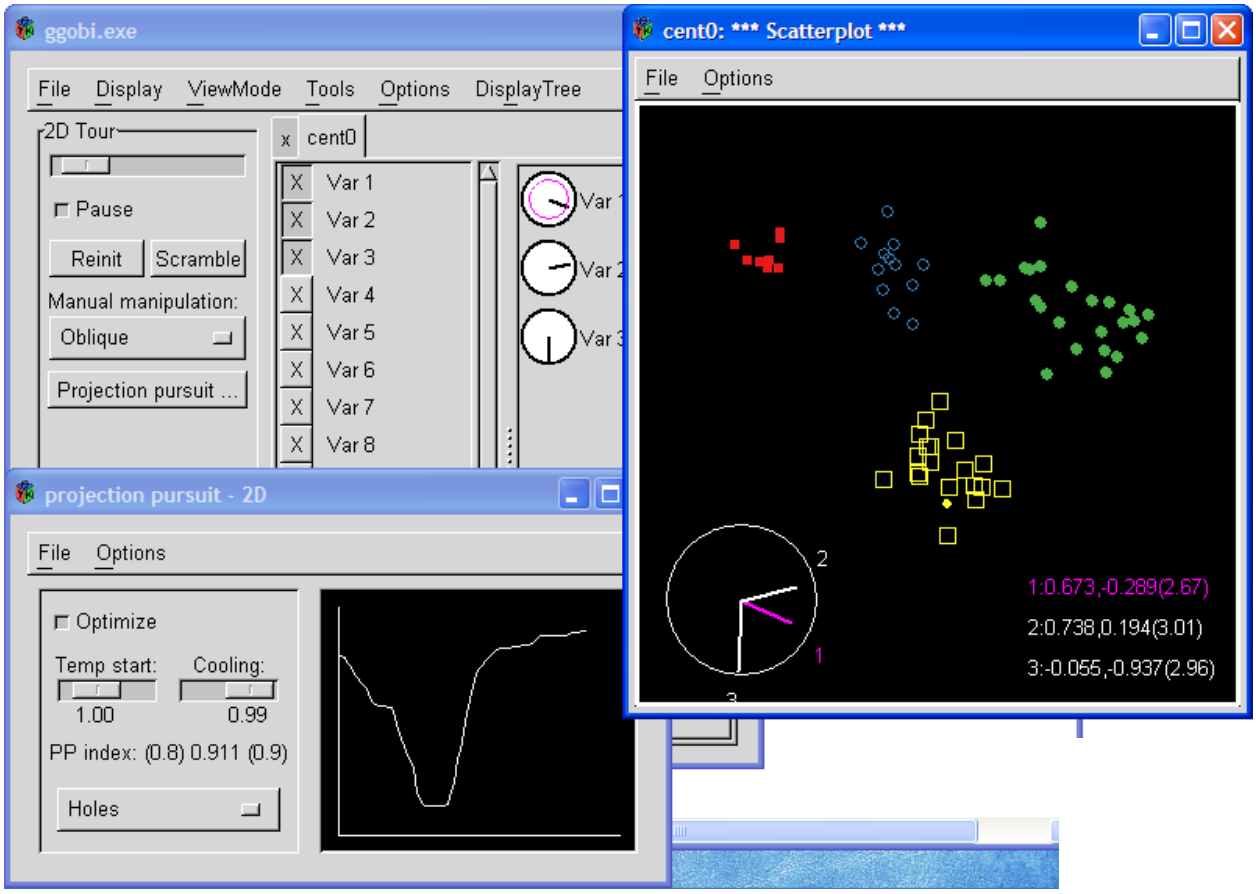
Data visualization of Multivariate Data

http://www.ggobi.org/

Ggobi display finding four clusters of tumors using the PP index on the set of 63 cases. The main panel shows the two dimensional projection selected by the PP index with the four clusters in different colors and glyphs. The top left panel shows the main controls and the left bottom panel displays the controls and the graph of the PP index that is been optimized. The graph shows the index value for a sequence of projection ending at the current one.



Projection pursuit

Data of three or more dimensions are difficult to visualize. On the other hand, two-dimensional, or even three-dimensional, views (*i.e.*, two- or three-dimensional projections) of the data are easy to visualize. In a two-dimensional graph, it is not hard to make out clusters or any other data structures. In a three-dimensional graph, we can use rotation software that enables us to visualize the data. However, as the dimension gets higher, visualization becomes difficult at best.

One solution is to look at low dimensional projections of high dimensional data but, again, we encounter a problem because, as the dimension gets higher, the number of views becomes far too large. The motivation behind *projection pursuit* (*PP*, for short) methodology (Friedman and Tukey (1974); further developed by Friedman (1987); see also Barnett (1981)) is to find a few low dimensional views of the data that describe the structure of the high dimensional dataset, such as clusters, outliers or subspaces containing the data, as they may provide interesting information about the scientific questions motivating the data analysis.

A projection is considered interesting if it shows a non-random or non-normally distributed point cloud. Projections showing a pattern of clusters or showing outliers are considered "interesting" since they differ markedly from a normal distribution. However, projections chosen at random are likely to be close to a normal distribution. This is a consequence of the Central Limit Theorem because projections are linear combinations of variables.

The method of projection pursuit finds the projections that optimize a criterion called the *projection pursuit index* that measures how interesting a structure is within a view. The most common indices are the Legendre Index and the Hermite index..

Let *Y=PX* be a one dimensional projection of our data. The *Hermite index* measures the distance from the empirical distribution of Y to a normal distribution. It was proposed by Hall (1989); Cook *et al* (1993) and Cook *et al* (1995) recommended using just two of the Hermite polynomial expansions of this distance resulting in a very simple expression that it is easily computable and hence not difficult to optimize. The two term Hermite index is:

$$I_H(P) = a_1^2 - 2^{\frac{1}{2}}\pi^{-\frac{1}{4}}a_0 + \frac{1}{2}\pi^{-\frac{1}{2}}$$

where $a_0 = Ave(\pi^{-\frac{1}{4}}e^{-\frac{Y^2}{2}})$ and $a_1 = Ave(\pi^{-\frac{1}{4}}e^{-\frac{Y^2}{2}} \times Y)$. The function *Ave* represents the average of the expression over the sample points.

In order to define the Legendre index, we transform the projection *Y* into a variable *U* in the interval *[-1,1]* by the function $U=2\Phi(Y)-1$ where $\Phi(t)$ is the normal distribution function. The *Legendre index* measures the L_2 distance between the distribution of *U* and a uniform distribution on the interval *[-1,1]*. This index was proposed by Friedman (1987). Again we use a two-term approximation based on Legendre polynomial expansion:

$$I_L(P) = a_1^2 + a_2^2$$

where $a_1 = \sqrt{\frac{3}{2}}Ave(U)$ and $a_2 = \sqrt{\frac{5}{8}}Ave(3U^2 - 1)$.

The method of projection pursuit consists of selecting projections that optimize a projection pursuit index and examining these projections graphically for interesting structures. Cook *et al* (1993, 1995) provide a detailed assessment of these and other indices.

Data visualization with the grand tour and projection pursuit

Cook *et al* (1993, 1995) describe Xgobi/Ggobi, a fascinating computer implementation of these ideas that combines the idea of a Grand Tour (essentially, a movie of data projections, a continuous sequence of two-dimensional projections of multi-dimensional data) (Asimov, 1985) with that of projection pursuit.

Example: Several datasets are included in GGobi. Open GGobi, you will see a screen of the software GGobi in action.

GGobi installation

On Windows PC:

Please download GGobi Windows 32 bit or GGobi Windows 64bit
Go to http://www.ggobi.org/downloads/ and download the corresponding exec

R Package:

From CRAN `install.packages("rggobi")`

On Mac OS X:

1. Download and Install XQuartz or another version of X11. You may already have it under Applications/Utilities. If not download it and install at: XQuartz-2.7.9.dmg
2. Download and Install GTK2 at gtk2-framework.dmg
3. Finally GGobi Install GGobi-2.1.8.dmg from Here