Visual Clustering of High-dimensional Data

Adrian Waddell and Wayne Oldford
University of Waterloo

GOAL
VISUAL CLUSTERING

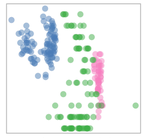
Automated and purely visual methods for cluster detection are complementary in the circumstances in which they have most value.

Automated methods may be routinely applied to data of more than three dimensions, where our visual experience and ability necessarily end.

Unfortunately, automated methods rely (implicitly) on pre-defined data patterns and so different methods can produce different clusterings.

The point of visual clustering is to use interactive data visualization tools in concert with automated methods so as to take best advantage of both.


True clusters


Mixture model based clustering

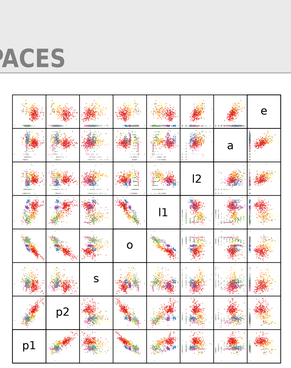

K means

CHALLENGE
VISUALIZATION AND HIGH DIMENSIONAL DATA SPACES

A common approach is to lay out small dimensional structures, either spatially or temporally, in such a way that they may be visually linked by the data analyst.

By alternately focussing on low dimensional structures and then linking these together, it is hoped that higher dimensional structure might be revealed.

A serious challenge is to determine the low-dimensional spaces worth visiting.



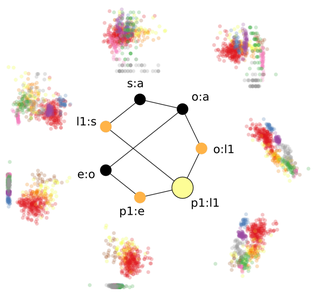
APPROACH
CONNECTING LOW DIMENSIONAL SUB-SPACES

Following Hurley and Oldford (2011), we use graphs as a navigational infra-structure to track movement from a display of one set of variables to another.

On a navigation graph each vertex represents a low dimensional space. Edges represent a transition from one space into another.

When the nodes are 2d spaces, defined by two variables, the data at each node may be displayed as a 2d scatterplot. Edges are then either 3d or 4d spaces connecting the 2d spaces and can be displayed as either 3d rigid rotations or as a sequence of smoothly interpolated 2d projections through the 4d space along a geodesic.

A *3d transition graph* has only edges between vertices which share one variable; a *4d transition graph* has only edges between vertices which share no variable.



A 3d transition graph

3d rigid rotation

DATA
ITALIAN OLIVE OILS

This data set records the percentage composition of 8 fatty acids found in the lipid fraction of 572 Italian olive oils.

The fatty acids are:
palmitic (p1), palmitoleic (p2), stearic (s), oleic (o), linoleic (l1), linolenic (l2), arachidic (a), eicosenoic (e).

The oils are samples taken from nine different areas:
North-Apulia, South-Apulia, Calabria, Sicily, East-Liguria, West-Liguria, Umbria, Coastal-Sardinia, Inland-Sardinia.

GRAPH CONSTRUCTION
VARIABLE GRAPHS, LINE-GRAPHS, SCAGNOSTICS

Transition graphs are easily constructed from a set of variables together with a set of interesting variable pairs.

At right, the *variable graph* G shows four variables (A,B,C,D) and all pairs as being of potential interest. The *3d transition graph* is simply the line-graph of G, L(G), and the *4d transition* its complement, $\overline{L(G)}$.

For even 8 variables, if all pairs are of interest, the number of nodes and edges in a 3d transition graph can be quite large (28 nodes, 168 edges; 210 edges for the 4d transition graph).

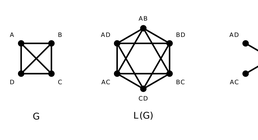
To reduce these numbers, we pre-calculate weights for the edges on the variable graph, and then remove all edges below some threshold.

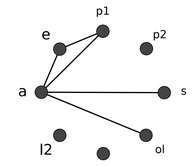
One collection of interesting weights are *scagnostic* weights (Wilkinson et al. 2005).

Each scagnostic assigns a weight between 0 and 1 that measures some geometric feature of the scatterplot that has data analytic interest.

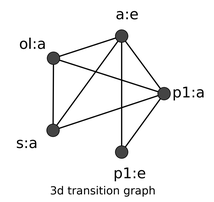
Scagnostics include the measures: *outlying, skewed, clumpy, convex, skinny, striated, stringy, straight* and *monotonic*.

For the olive oil example, we choose two navigation graphs that only contain 15% of the nodes (scatterplots) that perform best in either the not-convex or striated scagnostic measure.

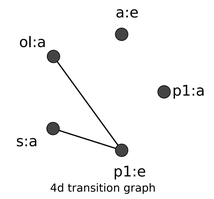




Variable graph with edges that represent the scatterplots that score highest in either the not-convex or striated scagnostic measure.



3d transition graph



4d transition graph

4d transition

INTERACTIVITY
RnavGraph

RnavGraph is an R package we wrote that implements navigational graphs.

The RnavGraph interface has two major pieces – the navigation graph, or navGraph, and an interactive 2d scatterplot.

The projection in the scatterplot display is determined by the position of the bullet (and) in the navGraph display.

Our scatterplot implementation can display points, text, images and star glyphs. In addition, the scatterplot display is completely interactive, allowing the analyst to brush, zoom, pan, subset, and link data between multiple displays.

In the first figure on the right, the top group in the point cloud has been selected via a brushing operation.

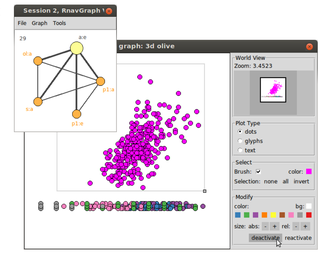
The selected points are then “deactivated”, causing them to disappear from all views in order to focus on the remaining data.

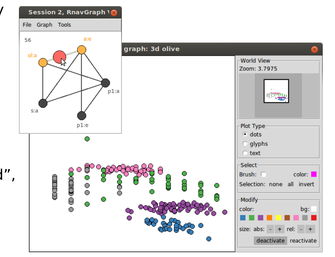
The second figure on the right shows the remaining data in a different projection.

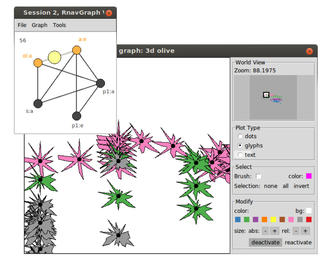
The third figure shows the data more closely zoomed in and using star glyphs.

This means that, in addition to grouping points by spatial positions, one can also compare nearby points on all other variate values simultaneously, simply by comparing the shapes of the glyphs.

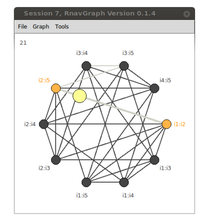
The “World view” shows the relative size of the scaling numerically and visually as the smaller white rectangle.

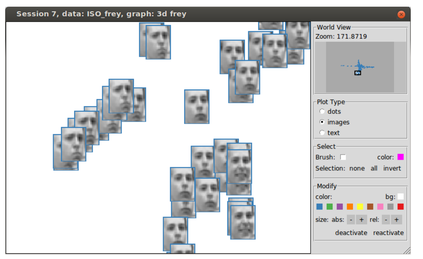






SAMPLE SESSION WITH IMAGES
FREY FACES AND LLE REDUCED IMAGE DATA





MORE INFORMATION

RnavGraph is distributed as an R package and hosted on CRAN. The graphical user interface was written with tcl and tk via the tcltk R package.

For more details see our web site: www.navgraph.com

C. Hurley and R. Oldford. Graphs as navigational infrastructure for high dimensional data spaces. Computational Statistics, 26:585–612, 2011.

L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In Information Visualization, 2005.

The color scheme used for scatterplots is according to the “Set1” from ColorBrewer.