RnavGraph: A visualization tool for navigating through high-dimensional data

Waddell, Adrian
University of Waterloo, Statistics and Actuarial Science
200 University Avenue West
Waterloo (N2L 3G1), Canada
E-mail: arwaddell@uwaterloo.ca

Oldford, R. Wayne
University of Waterloo, Statistics and Actuarial Science
200 University Avenue West
Waterloo (N2L 3G1), Canada
E-mail: rwoldford@uwaterloo.ca

Introduction

Hurley and Oldford (2011) use graph theoretic results to develop navigational graphs to track movement from a display of one set of variables to another. We implement these ideas in an `R` package called `RnavGraph`, which is built using the `tcltk` `R` package together with certain `C` extensions for `tcltk`. `RnavGraph` is available from CRAN (Comprehensive R Archive Network, http://cran.r-project.org).

In this paper, we show how `RnavGraph` can be used to explore high dimensional space by interactively walking a navigational graph (i.e. a "navGraph") whose nodes represent low-dimensional space. In particular, the canonical example will be a graph where each node represents a two dimensional space and each edge is either a 3- or 4-dimensional transition between the node spaces.

Typically, each node will be a scatterplot and the transitions a sequence of 2d projections from one node's pair of variables to the other's (when these pairs share one variable this will be a 3d-rotation, otherwise it is a 4d transition). `RnavGraph` has its own interactive scatterplot tool that can display coloured dots, images, text, or glyphs and that allows interactive colour and size changes individually or in groups (i.e. single and multiple selection). However displayed, the points may be brushed and even dragged through the display (e.g. to avoid overplotting). Other features of `RnavGraph`'s scatterplot tool are the ability to zoom in on subsets of points, to pan this zoom over the entire region of the scatterplot and to even remove points from the display. `RnavGraph`'s scatterplot tool was designed with interactive exploration and clustering in mind (see Oldford and Waddell, 2011 for further discussion).

`RnavGraph` is not, however, restricted to scatterplots or even only to 2d-node representations. In what follows, we show how `RnavGraph` may also be easily extended to accommodate any data display on a node, provided that display can be constructed from `R`. Even the meaning attached to the transition along an edge from one display to another can be changed – e.g. we show how conditional displays that slice across a variate may be constructed by changing the transition semantics.

`RnavGraph`'s principal features and its extensibility will be illustrated throughout this paper using a single data set we have compiled from Hans Rosling's gapminder data repository and from the UN Inter-agency Group on Child Mortality Estimation. The variable shortnames and descriptions are given in Table 1. (Disclaimer: At the time of this writing the data has not yet been thoroughly

Example Country dataset: from Gapminder and UN sources

Shortname	Description
Murder	Murder, age adjusted, per 100,000.
Traffic_Deaths	Traffic mortality, age adjusted, per 100,000.
Childern_w	Children per woman (total fertility).
Child_Mortality	Under-five mortality rate (per 1000 live births).
Age1stMarriage	Mean age at 1st marriage of women.
Suicide_w	Suicide among women, per 100,000 standard population.

Table 1: The data was collected for 157 countries. The variables are averages over those years from 2004 to 2009 that have values.

checked for accuracy.)

With this example, we first show navigation graphs are used in conjunction with scatterplots and then give more detail on `RnavGraph`'s base functionality. This is described in more detail in Oldford and Waddell (2011) in the context of visual cluster analysis. `RnavGraph`'s extensibility to build custom plots is then illustrated, including a change in the meaning of the navigation graph itself. We conclude with some summary remarks.

2d Scatterplot Example

As suggested by Figure 1, by default, each node of a navigation graph (or navGraph) is associated

Figure 1: In `RnavGraph`, the default meaning of a navigation graph assigns to each node a 2 dimensional scatterplot.

with a scatterplot of the data values for the variable pair defining the node; node labels are displayed as the "shortname" of each variable in the pair, separated by a colon.

For illustrative purposes, the navGraph of Figure 1 has only 4 nodes, out of a possible $\binom{6}{2} = 15$. A more complete analysis for this 6 dimensional data set, might use a fuller navGraph of 15 nodes, representing all 15 variable pairs. For many problems, the dimensionality of the data will be too large for this to be manageable, and a more targeted selection of "interesting" variable pairs would be made so as to reduce the complexity of the data exploration. For example, scagnostics (Wilkinson et al, 2005) can be used to look only at those variable pairs known to reveal certain types of 2d data structure (see Hurley and Oldford, 2011; Oldford and Waddell, 2011).

Note also that all edges of the navigation graph in Figure 1 connect nodes whose variable pairs have one variable in common. The edge in this graph can be thought of as representing a 3d space through which transition from one pair of variables to another occurs. This kind of navigation graph is therefore called a *3d transition graph*. By default, the transition along an edge in such a graph (i.e. moving the "You are here" bullet along an edge) will effect a 3d rigid rotation as in Figure 2 – note

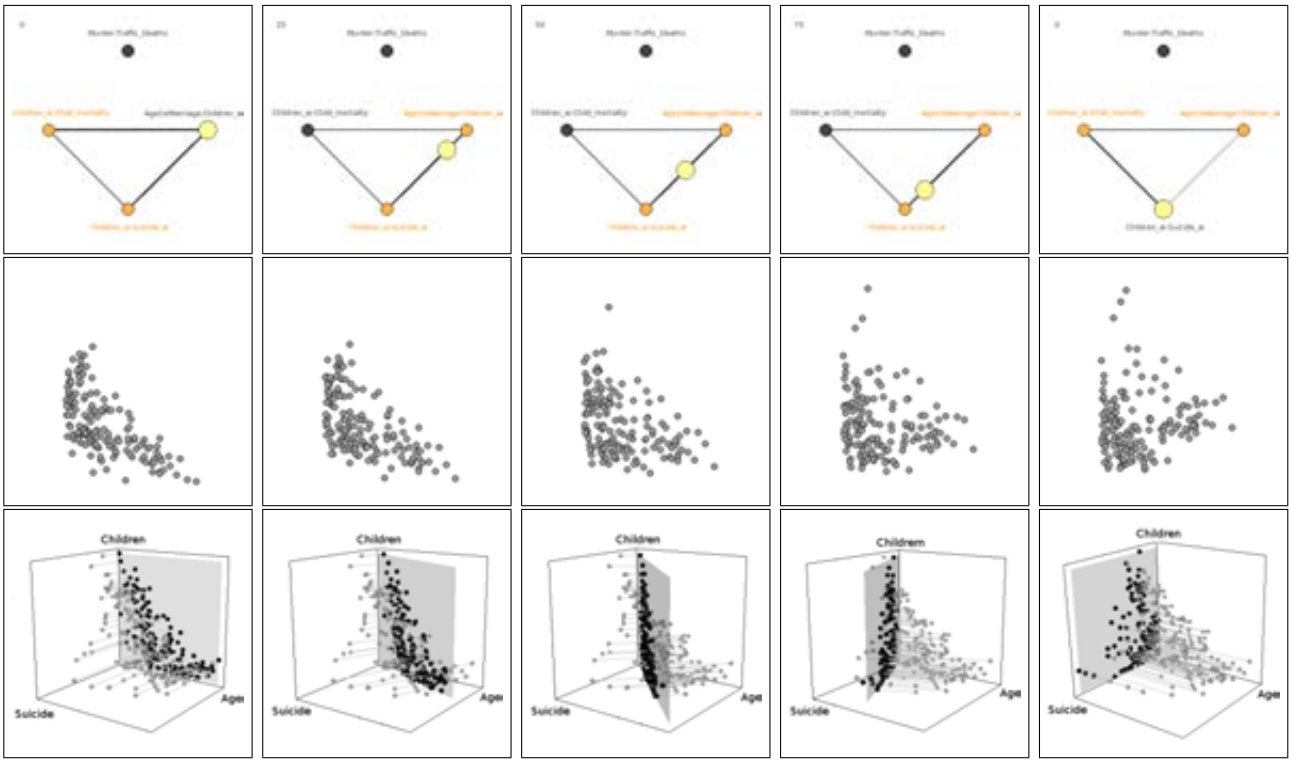


Figure 2: Moving the bullet along a 3d transition edge results with a 3d rigid rotation from one scatterplot into another.

that the bullet in is distinguished from a node by its larger diameter and brighter colour (yellow).

The bullet's progression is shown in the upper left corner as a percentage along the transition. From left to right in Figure 2: the first row shows the bullet moving from the node labelled "Age1stMarriage:Children_w" along the 3d transition edge towards the node "Children_w:Suicide_w"; the second row shows the linked data display where, simultaneously as the bullet moves, a 3d point cloud rotates in the space defined by the three variables; the last row shows how this rotation is done – by rotating a plane around the common variable axis, continually projecting all points onto it and updating the corresponding 2d scatterplot of all 157 countries.

The graph complement of a 3d transition graph will be a 4d transition graph – edges connect only nodes with no variables in common. Figure 3 shows a transition along the edge of a 4d transition graph. The resulting scatterplot sequence is no longer a 3d rigid rotation and the third row of Figure

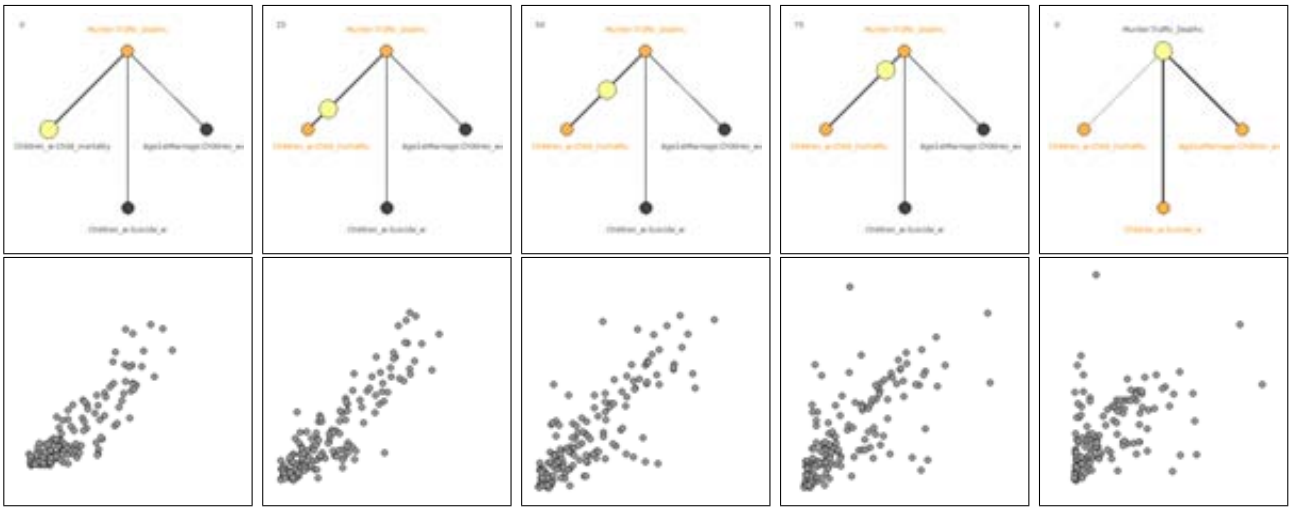


Figure 3: Moving the bullet along a 4d transition edge results with a smooth transition from one scatterplot into another.

2 cannot be reproduced here. Instead of the intermediate projection planes rotating around a single axis in 3d space, in `RnavGraph` as in the grand tour (e.g. Buja et al, 1988, Cook et al, 1995), they are now a sequence of planes in 4d space ordered along a geodesic path from the source plane of the initial variable pair to the destination plane of the target variable pair. The result is a smooth 4d transition between scatterplots.

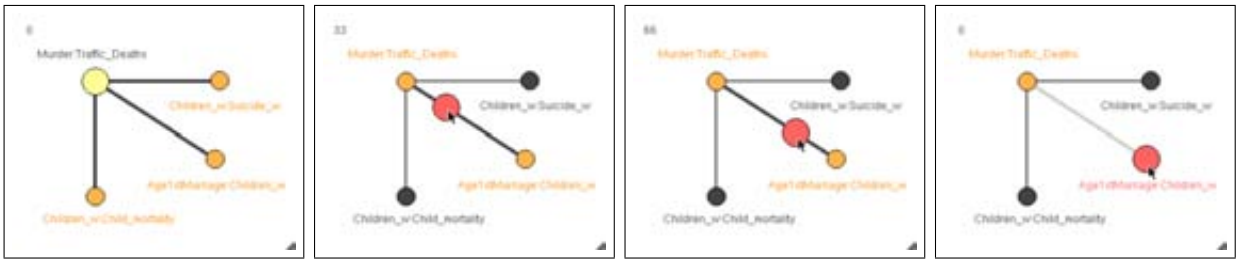
`RnavGraph` Functionality

The navigation graphs in Figure 1, 2 and 3 are screen shots from an `RnavGraph` session. Nodes and node labels can be interactively relocated in the display. The bullet can be interactively moved along the edges using mouse drag and drop, or by identifying the target node and scrolling, or simply and automatically by double-clicking on the target node. Also, through multiple selection of connected nodes, entire paths can be selected and the corresponding sequence of transitions are animated. Interesting paths can be saved, edited, annotated, and rewalked once saved. Every change of the bullet state updates the data displays connected with the navigation graph. Finally, multiple displays may be connected with a navigation graph in order to have different views on the data update simultaneously.

We have implemented a scatterplot display specifically for `RnavGraph` and its data exploration and clustering purposes; its use is described in more detail in Oldford and Waddell (2011). This scatterplot tool double buffers every screen update to make the changes in its display look smooth. The plotting symbol can be a dot, an image, a star glyph, or simply text; all can be assigned colours and all but the text (at time of writing) can be resized.

Figures 4(a)-(d) show a 4d transition, where the bullet is moved with the mouse pointer from the node "Murder:Traffic_deaths" to the node "Age1stMarriage:Children_w"; the matching sequence of scatterplots, with each country's flag as its plotting symbol, is shown in Figures 5(a)-(d), respectively, At the right of each scatterplot display is a smaller scatterplot of all points as dots; this is the "world view" and the white rectangular region within it shows the portion that appears in the large scatter plot as the "main view".

Figure 6 shows the same scatterplots as in Figure 5, but with different point symbols. In addition, Figure 6(c) shows an example of using the brush and Figure 6(a), 6(b) and 6(d) display only



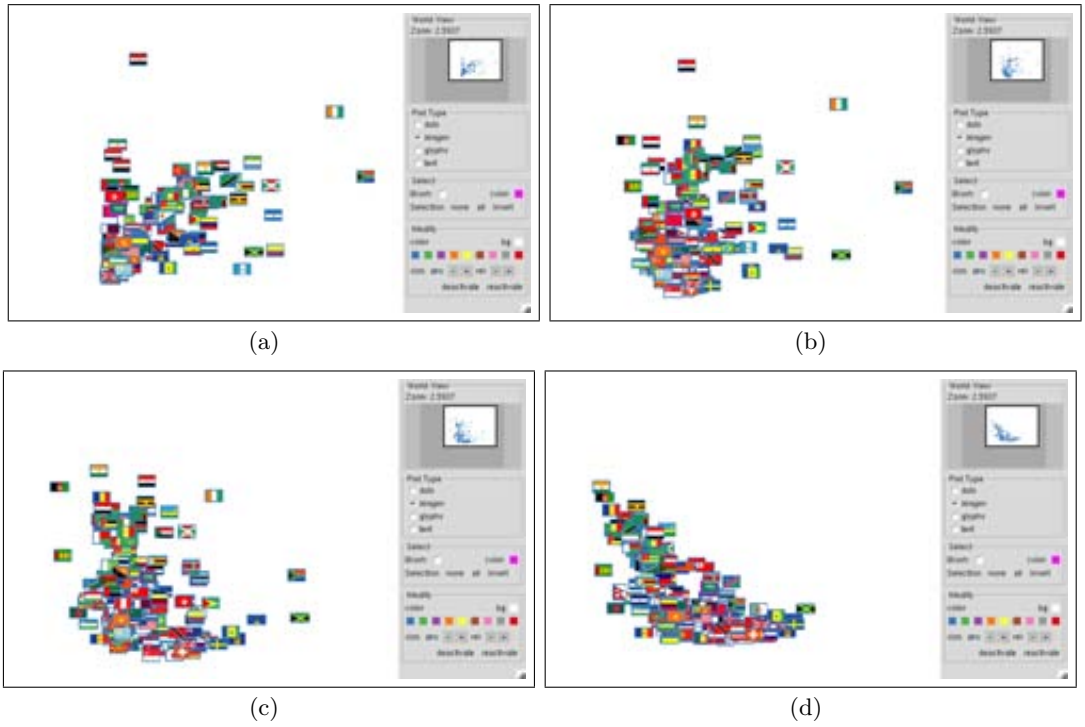
(a) Nodes adjacent to the bullet are highlighted with a different colour.

(b) Edge nodes marked; bullet dragged 33% of the way along.

(c) 66% of the way along.

(d) All the way; the traversed edge changes colour.

Figure 4: The same navigation graph with four different bullet states. Each node represents a display of some variable pair, and each edge some transition from one display into another.



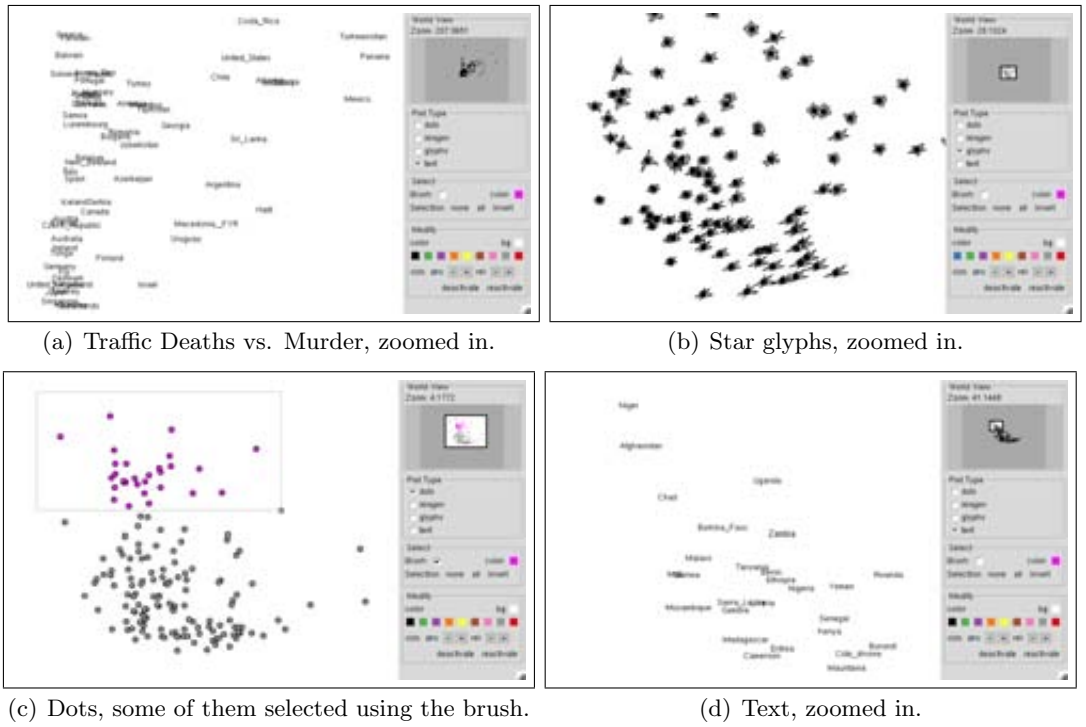
(a) (b)

(c) (d)

Figure 5: The four bullet states of the navigation graph in Figures 4(a)-(d) are linked to these four scatterplot displays in order.

a subset of the data points because the "main view" is zoomed in. Moving the small rectangle around in the "world view" is one way to pan the data which appear in the "main view".

The $2d$ scatterplot framework with $3d$ rotations and $4d$ transitions can be useful for interactive clustering and exploring data. For our Country data, it is also interesting to highlight some countries and compare them to each other or to the rest of the world. We have also found, that looking at the results of dimensionality reduction algorithms (such as LLE by Roweis 2000 and Isomap by Tenenbaum et al 2000) can be very insightful. When clustering data, clusters are either perceived directly through the spatial location of the points, or interactively by looking at which points travel together. Again, see Oldford and Waddell (2011) for more detail.



(a) Traffic Deaths vs. Murder, zoomed in.

(b) Star glyphs, zoomed in.

(c) Dots, some of them selected using the brush.

(d) Text, zoomed in.

Figure 6: Same scatterplots as in Figure 5; however, with different point objects (glyphs, images and text) and different plotting regions.

Custom Plots

Using our own 2d scatterplot tool is a good staring point to explore data. `RnavGraph`, however, can be connected to any data display that uses a graphics system which is accessible by `R`.

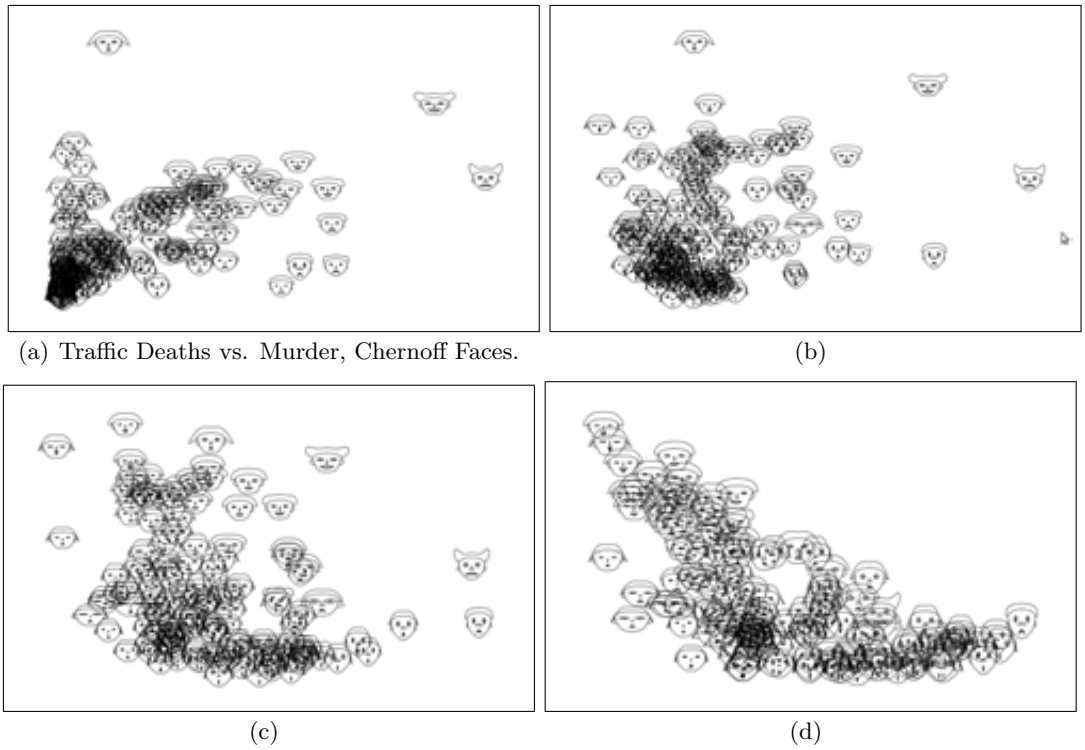
Figures 7(a)-(d) show the same scatterplot sequence as Figures 5(a)-(d); both follow the bullet states of Figures 4(a)-(d). Figure 7, however, displays Chernoff faces (Chernoff 1973) as plotting symbols. This was implemented simply by having bullet changes update a visualization in the usual `R` plotting system (making use also of the Chernoff faces implemented in the `TeachingDemos` `R` package). The plotting is done using `R`'s base graphic engine.

In Figures 8(a)-(d), we again follow the sequence of Figures 5(a)-(d). This time, however, instead of a scatterplot, a density estimate of the bivariate data is computed and then plotted as a surface using an `rgl` 3d display. Note that, at any point in a navigational graph transition (3d or 4d) this 3d density can be grabbed and moved around at will, just as any 3d display can be manipulated in `rgl`.

Slicing

`RnavGraph`, like the `PairViz` `R` package (Hurley and Oldford, 2011b), shows that graphs can be useful to structure data analytic tasks. Navigational graphs, however, entail meaning – the relation between nodes, edges, and the data display must all be defined and well understood. Beyond the semantics defined for 2d scatterplots (Hurley and Oldford, 2011), there is room for much creativity and further research (see also Hurley and Oldford, 2011b, for some unusual graphs).

One possibility is to use the navigation graph framework to accommodate data-conditional slicing across a variable. This is most naturally done when the navigational graph is a 3d transition graph. The two variables not shared by an edge define the x and y location in a scatterplot, and the common



(a) Traffic Deaths vs. Murder, Chernoff Faces. (b)

(c) (d)

Figure 7: Custom plotting with Chernoff faces as point symbols using `R`'s base graphic engine.

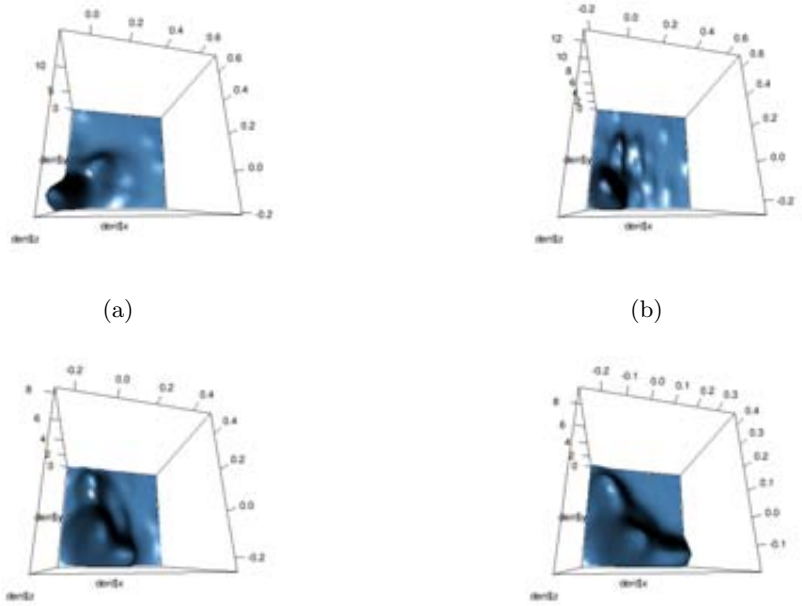
variable, in combination with the progression of the bullet transition, determine the conditioning.

We implement this with another example of density estimation. Figure 9 shows a dynamic hexbin example. The data points are binned into hexagons, and the colour saturation of each hexagon represents the bin count (see Carr 1987). The display and binning algorithm are implemented using Java and Java2d.

The scatterplots in Figure 9(a) and 9(e) are identical and plot joint (hexbin) density estimate of the average age of women at their first marriage versus the under-five child mortality estimates over all countries. That is, when the bullet's location is on top of a node, we don't condition our display.

Moving the bullet towards the other node results in changing the countries that are being selected for the binning algorithm. In Figure 9(b), 9(c) and 9(d), we select the countries that fall into the children per women quantile interval of $p \pm 20$, where p is the proportional progression of the bullet along that variable.

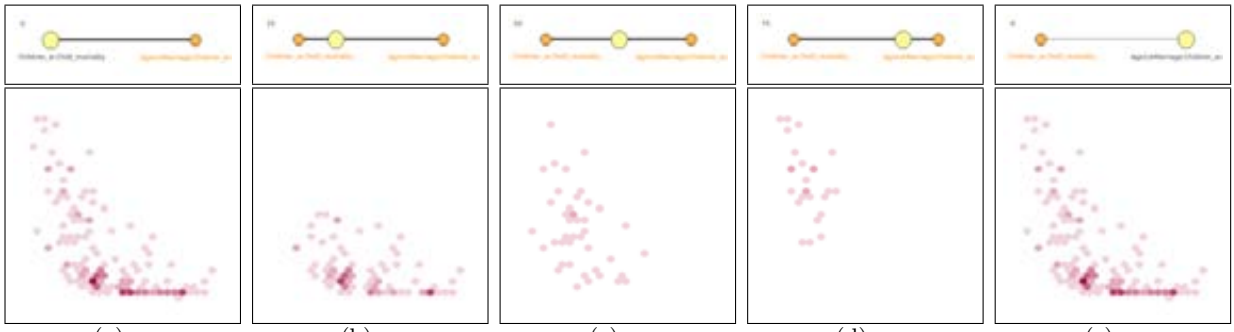
Hence, Figure 9(b) shows the hexbin plot of average age of women at their first marriage versus the under-five child mortality estimates for countries where women have relatively few children compared to the rest of the world. Figure 9(c) shows the same hexbin plot for countries where the number of children per women is near the world average; and Figure 9(d) shows the hexbin plot for countries where women have a lot of children, relative to the rest of the world.



(a) (b)

(c) (d)

Figure 8: Custom plotting: Density estimates shown with the `rgl` 3d display.



(a) (b) (c) (d) (e)

Figure 9: Hexbin plot of sliced data. The slicing variable is the average number of children per women.

Summary

`RnavGraph` provides a framework for exploring data interactively. Graphs are used as a "road-map" for the data. Each node of a navigation graph represents some image and the edges represent a real-time morphing from one image into another. A good starting point for exploring or clustering data is to use the default visualization method with nodes representing 2d scatterplots and edges corresponding to 3d rigid rotations or 4d transitions. Our own implementation of a 2d scatterplot display features a set of interactive utilities, useful especially for clustering. Also, our scatterplot tool can display different point symbols such as text, images or glyphs, in order to simplify the identification of an individual point. Furthermore, points from the same data may be linked between different displays of the same data. That is, `RnavGraph` allows one to use linked multiple displays of the same data within the same and/or between different `RnavGraph` sessions. The user is, however, free to use any display or even define his/her own meaning of nodes and edges of a navigation graph. The reader is encouraged to try our open source `R` package, read the package vignette, and try some of the package demos.

REFERENCES (RÉFERENCES)

Carr, D. B., R. J. Littlefield, W. L. Nicholson, and J. S. Littlefield. (1987). "Scatterplot Matrix Techniques for Large N", *Journal of the American Statistical Association* 82, no. 398 (June 1, 1987): 424-436.

Chernoff, H. (1973). "The Use of Faces to Represent Points in K-Dimensional Space Graphically." *Journal of the American Statistical Association 68*, no. 342: 361-368.

Hurley, C. B., and R. W. Oldford. (2010). "Pairwise Display of High-Dimensional Information via Eulerian Tours and Hamiltonian Decompositions", *Journal of Computational and Graphical Statistics* 19, no. 4 (December 2010): 861-886.

Hurley, C. B., and R. W. Oldford. (2011). "Graphs as navigational Infrastructure for high dimensional data spaces", *Computational Statistics* (Online First February 2011).

Hurley, C. B., and R. W. Oldford. (2011b). "Eulerian tour algorithms for data visualization and the `PairViz` package". *Computational Statistics* (Online First February 2011).

Oldford, R. W. and Waddell, A. R. (2011). "Visual Clustering of High-dimensional Data by Navigating Low-dimensional Spaces", *58th Congress of the International Statistical Institute*, *Special Topics Session 57*, Dublin, Ireland.

Rosling, H. "Gapminder", *GapMinder Foundation* `http://www.gapminder.org/`

Roweis, Sam T., and Lawrence K. Saul. "Nonlinear Dimensionality Reduction by Locally Linear Embedding." *Science* 290, no. 5500: 2323-2326.

Tenenbaum, Joshua B., Vin de Silva, and John C. Langford. (2000) "A Global Geometric Framework for Nonlinear Dimensionality Reduction." *Science* 290, no. 5500: 2319-2323.

UN Inter-agency Group for Child Mortality Estimation (2010). "Levels & Trend in Child Mortality". *Rreport 2010*. `http://www.childmortality.org`.

Wilkinson, Leland, Anushka Anand, and Robert Grossman. (2005). "Graph-theoretic scagnostics", *Proceedings - IEEE Symposium on Information Visualization*: 157-164.

RÉSUMÉ (ABSTRACT)

We discuss the implementation of the ***RnavGraph*** *R package and demonstrate its functionality on some high dimensional data.* ***RnavGraph*** *facilitates controlled exploration of high dimensional data space via (user determined) low dimensional trajectories through that space. The trajectories are paths on a navigation graph (navGraph), a graph whose nodes are plots and whose edges represent transitions from one plot to another. In* ***RnavGraph****, there are two primary display regions – the navGraph in one and the "plot" (some visualization of the data) in the other. The navGraph display drives the data visualization through the position of a "You are here" bullet on the navigation graph.* ***RnavGraph****'s features and extensibility are illustrated on the basis of an international data set we have compiled from Hans Rosling's gapminder data repository and from the UN Inter-agency Group on Child Mortality Estimation.*