Interactive Visual Clustering of High Dimensional Data by Exploring Low-Dimensional Subspaces

Adrian Waddell R. Wayne Oldford

University of Waterloo

ABSTRACT

The structure of a set of high dimensional data objects (e.g. images, documents, molecules, genetic expressions, etc.) is notoriously difficult to visualize. In contrast, lower dimensional structures (esp. 3 or fewer dimensions) are natural to us and easy to visualize. A not unreasonable approach then is to explore one low dimensional visualization after another in the hope that together these will shed light on the higher dimensional structure.

In our poster, we describe the graph theoretic structure recently proposed in [3] that represents low-dimensional spaces as graph nodes and transitions between spaces as edges. Of interest are walks along these graphs that reveal meaningful structure. If the nodes are two dimensional and edges exist, say, only between 2d spaces which share a variate, then the walk could be represented dynamically as a series of scatterplots, one transitioning into the next via a 3d rigid transformation. We demonstrate how these graphs are constructed and dynamically explored via our open source R package, `RnavGraph`.

Index Terms: I.5.5 [Pattern Recognition]: Implementation—Interactive systems

1 INTRODUCTION

The purpose of cluster analysis is to conjecture plausible differences in kind amongst a given collection of instances. This is also what our human visual system excels at; it has evolved to facilitate quick and considered detection of the visually like and unlike through a wide variety of cues – e.g. location and relative proximity, movement, shape, colour, texture and matching against predetermined patterns. Consequently, visualization is a natural and powerful resource for cluster analysis; it is especially valuable in identifying unanticipated structures.

Unfortunately, the same evolutionary path has meant our visual system is poorly equipped to be of much help in identifying high dimensional structure. And most data these days are of high, and ever increasing, dimensionality. Consequently, automated methods of pattern recognition and cluster analysis have seen increasing recent use and development; even so, intuition as to what constitutes a "cluster" in high dimensions remains largely, though by no means exclusively, based on our experience with our own visual perception – e.g. near neighbours, *k*-means, local density modes, etc.

Automated and purely visual methods for cluster detection are largely complementary in the circumstances in which they have most value. Automated methods may be routinely applied to data of many more dimensions than three, where our visual experience and ability necessarily end. Unfortunately, to do so, automated methods rely (at least implicitly) on determining pre-defined patterns in data configurations 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. Following [3], we do this by introducing a graph structure, called a *navigational graph*, or *navGraph*, whose vertices represent a unique pair of variates. When we add only edges between vertices which share a variate, the edge itself represents a three dimensional space formed by the union of those variates. Such a *navGraph* is called a *3d-transition graph* in [3].

For example, the Olive data in [2] records the percentage composition of the following eight fatty acids in 572 different Italian olive oils: arachidic (a), eicosenoic (e), linoleic (l1), linolenic (l2), oleic (o), palmitic (p1), palmitoleic (p2) and stearic (s). One possible 3d transition graph is shown at the centre of Figure 1. The

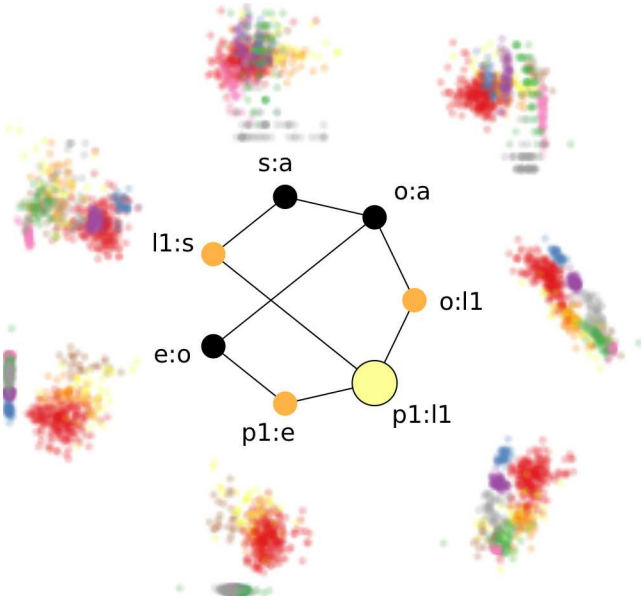
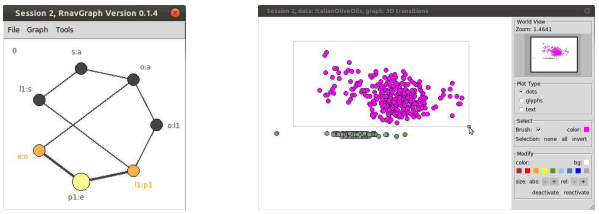


Figure 1: A 3d-transition graph for selected pairs of variates from the Olive data. The "bullet" is the large (yellow) vertex at p1:l1. Vertices connected to the bullet vertex are coloured differently (orange) from those which are not (black). Outside each node, the point cloud of the 572 data points is shown for that variate pair.

2d point clouds associated with each node use a color key that corresponds with the geographic regions of olive oils, i.e. North-Apulia, South-Apulia, Calabria, Sicily, East-Liguria, West-Liguria, Umbria, Coastal-Sardinia, and Inland-Sardinia. If these geographic regions are the "true" cluster structure, it is possible to recover much of this structure in `RnavGraph` simply through the spatial structure of the data in low dimensions, as we will show partly in the next section and in more detail in our poster.

2 VISUALLY CLUSTERING THE OLIVE DATA USING RNAVGRAPH

The `RnavGraph` interface has two major pieces – the navigation graph, or navGraph, and an interactive 2d scatterplot. The two displays are shown side by side in Figure 2 as they might appear on a



(a) The navGraph window. (b) The interactive scatterplot window.

Figure 2: On right, the brush has been used to highlight the top group in the point cloud.

data analyst's screen. The positions of the points in the scatterplot display are determined by the position of the bullet in the navGraph display. Our 2d 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, link data between multiple displays and to analyze a subset of the data. In Figure 2b we show that the analyst has selected a brushing operation and highlighted all points in the top group by sweeping out a rectangular area. These selected points may be "deactivated", causing them to disappear, so as to allow the analyst to focus on the remaining data. We show the remaining data in Figure 3 as a point cloud of oelic vs. arachidic. The three different colourings

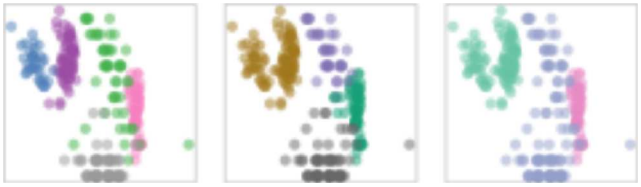


Figure 3: Closer examination of the `o:a` space. The colouring represent from left to right: true region, *k*-means clustering `kmeans(data,k=9)`, and model based clustering `Mclust(data,1:20)`.

correspond to the true geographic regions and to the outcome of automated cluster methods, i.e. k-means and model based clustering. Surprisingly, neither method separates the "Inland" from the "Coastal" areas of Sardinia (top left corner). Overall, the *k*-means seems to be doing a better job than model based clustering.

The separation of "Inland" from the "Coastal" Sardinia olive oils, however, can be shown with `RnavGraph` (and hence visually) quite well. In Figure 4, we show four states of a 3*d*-transition from `p1:l1` to `l1:s`. While the bullet is dragged along the edge, the scatterplot dynamically displays the a 3d rigid rotation from one scatterplot into the other. Of course, this separation of "Inland" vs.

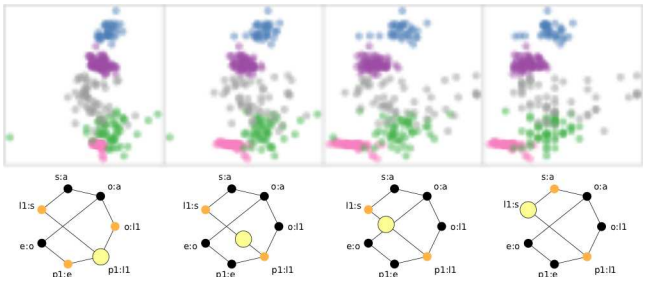


Figure 4: The same navigation graph is shown with four different bullet positions. The point clouds resemble a 3d rigid rotation.

"Coastal" olive oils is not evident on every 3d transition, but interactively exploring all transitions along the graph strongly suggest this separation. More exploration on the complete data set is shown on our poster.

3 Graph Construction

A serious challenge is to determine the low-dimensional spaces worth visiting. For p variables, there are $\binom{p}{2}$ possible nodes in a navGraph. In [3], Hurley and Oldford describe a variety of methods for construction of graphs so as to focus only on those subspaces that have interesting data structure. The resulting graph provides a small navigational structure to explore, an important advantage over other structures (e.g. a scatterplot matrix as navigation as in [1]) which would be overwhelmed by large numbers of variables. Experience to date suggests that scagnostic measures [6] are particularly valuable in identifying interesting subspaces. All such methods from [3] are available in the `RnavGraph` package.

For very high dimensions, when the context does not naturally produce a graph with small numbers of vertices and/or edges, some dimensionality reduction should be pursued before building the navGraph, see [3]. Figure 5 is an example of a data set of images

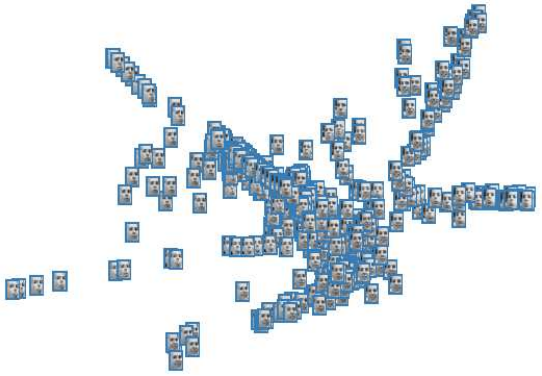


Figure 5: Image cloud of the Frey faces. The associated navGraph bullet location is 46% along the way from `i2:i3` to `i3:i5`.

taken from a movie. Each image is an array of 28×20 greyscale pixels – a point in $p = 560$ dimensions! This dimensionality was reduced to 5 (`i1` to `i5`) by local linear embedding (LLE [4]) and the navGraph constructed. Clearly, there is considerable structure in this data and it is not restricted to the first two dimensions. By using a navGraph to explore the reduced dimension set of variates, the target number of dimensions can be considerably larger than usual, e.g. 10 or 20.

Any 3d- or 4d-transition graph can be viewed through the `navGraph(...)` function, with an unlimited variety of visualizations beyond point clouds, see [5].

References

[1] N. Elmqvist, P. Dragicevic, and J.-D. Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *Visualization and Computer Graphics*, 14(6):1539 –1148, 2008.

[2] M. Forina, C. Armanino, S. Lanteri, and E. Tiscornia. Classification of olive oils from their fatty acid composition. *Food Res. and Data Analysis*, pages 189–214, 1983.

[3] C. Hurley and R. Oldford. Graphs as navigational infrastructure for high dimensional data spaces. *Comp'l. Stats.*, 26(4):585–612, 2011.

[4] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[5] A. Waddell and R. Oldford. Visual clustering of high-dimensional data by navigating low-dimensional space. *Proc. ISI*, 58(STS 57), 2011.

[6] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *Proc. IEEE Symp. on Info. Vis.*, pages 157–164, 2005.