Visual Explanation of Multidimensional Projections

Renato R. O. da Silva[1,2], Paulo E. Rauber[1], Rafael M. Martins[2]

supervisor: Alexandru C. Telea[1]

1 - Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen

2 - Institute of Mathematical and Computer Sciences, University of São Paulo

Motivation

Multidimensional Projections (MPs) are key tools used to support the analysis of multidimensional data. MPs can project data to a low dimensional representation, typically visualized as a 2D scatterplot where similar elements are conveniently positioned in close neighborhoods. However such visualizations tell us which points are similar, but not why. Our aim is, thus, to enrich 2D MP scatterplots with explanatory information telling users which key dimensions make closely-projected points similar.

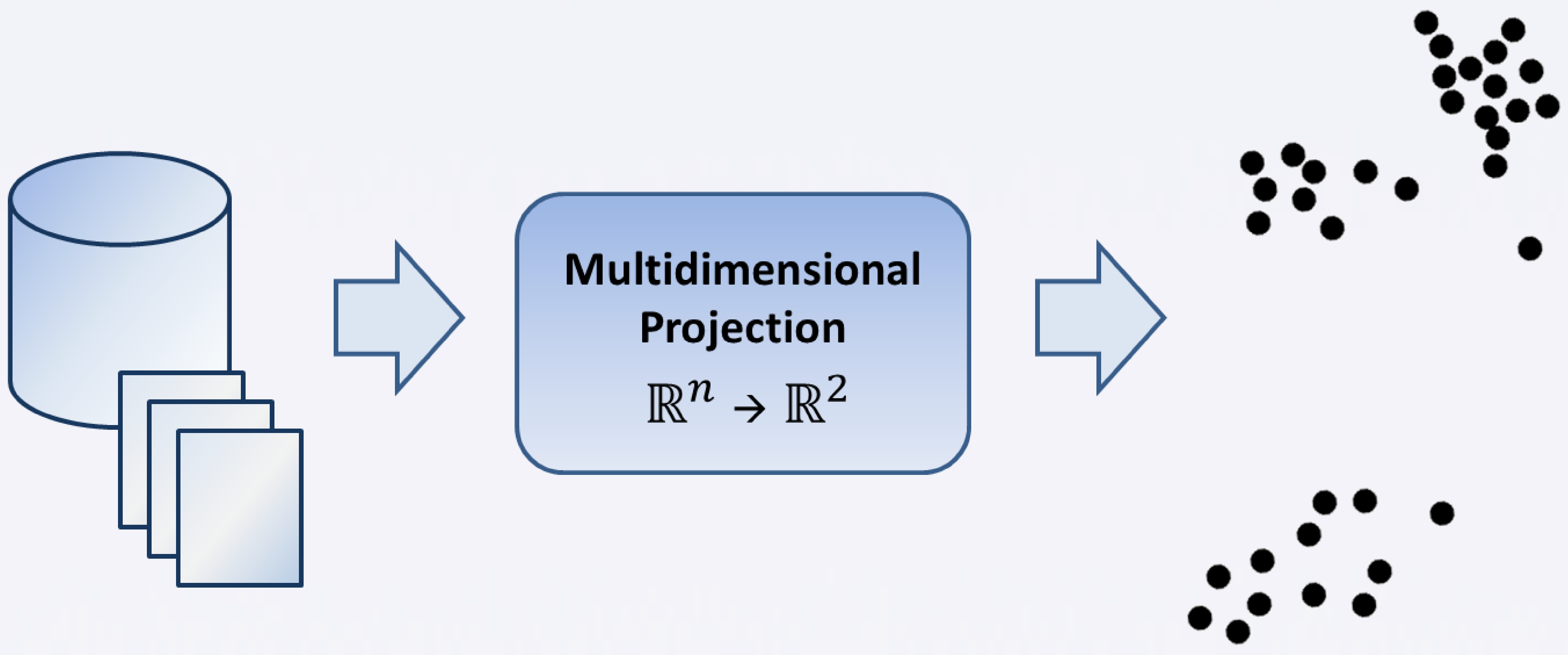

Figure 1: *Multidimensional Projection pipeline: A MP technique aims to project data into a low dimensional space based on the nD distances. The projection is typically visualied as a 2D scatterplot, whose points positions depict their similarity of the original nD space.*

Explanation maps

To explain a projection we identify which dimensions are more important to define similarity among close points. First assign a **rank** to each of the n dimensions for each projected point, based on increasing order of data variance in that dimension over that neighborhood. We also define the ranking **confidence** for each point based on how mixed are the top-ranked dimensions on their neighborhood points.

Next we display the top ranks and their confidences over the projection using a dense map technique based on nearest-neighbor (Voronoi) interpolation, with top-ranks encoded by a categorical **colormap** and confidences by **brightness**.

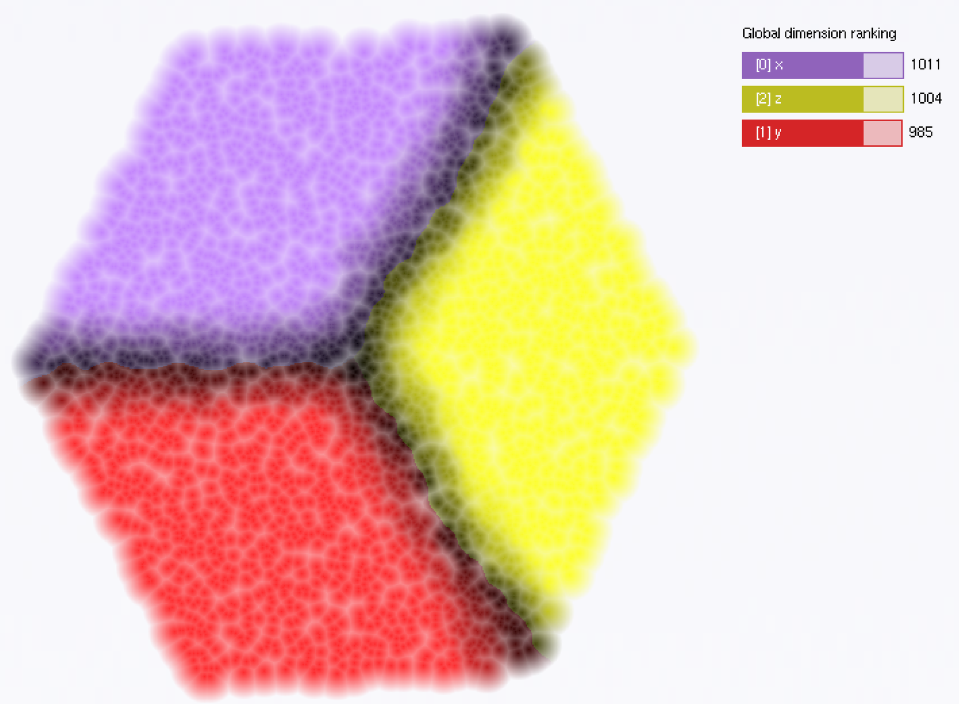

Figure 2: *Explanation map of projected points sampled from three cube faces. Points on the same cube-face have the same value for one dimension d, so they have assigned the same color. Color brightness indicate the confidence ranking, from hight (bright) to low (dark). A histogram in the right indicates the top-ranked dimensions frequencies and their labels.*

Map color coding scheme

Many datasets might have more dimensions than the available colors in a categorical colormap. In that case we only define colors to the C more frequent top-ranked dimensions of the projected points. Dimensions having top ranks (for fewer points) which do not get mapped to colors, due to the colormap's limited size C, are mapped to a reserved color.

Single dimension explanation

Clusters created by the projection layout can be explained using the top-ranked dimension for each point. The regions defined by areas of same color can create clusters sub-partitions, grouping elements based on the dimension that best explains their similarities.

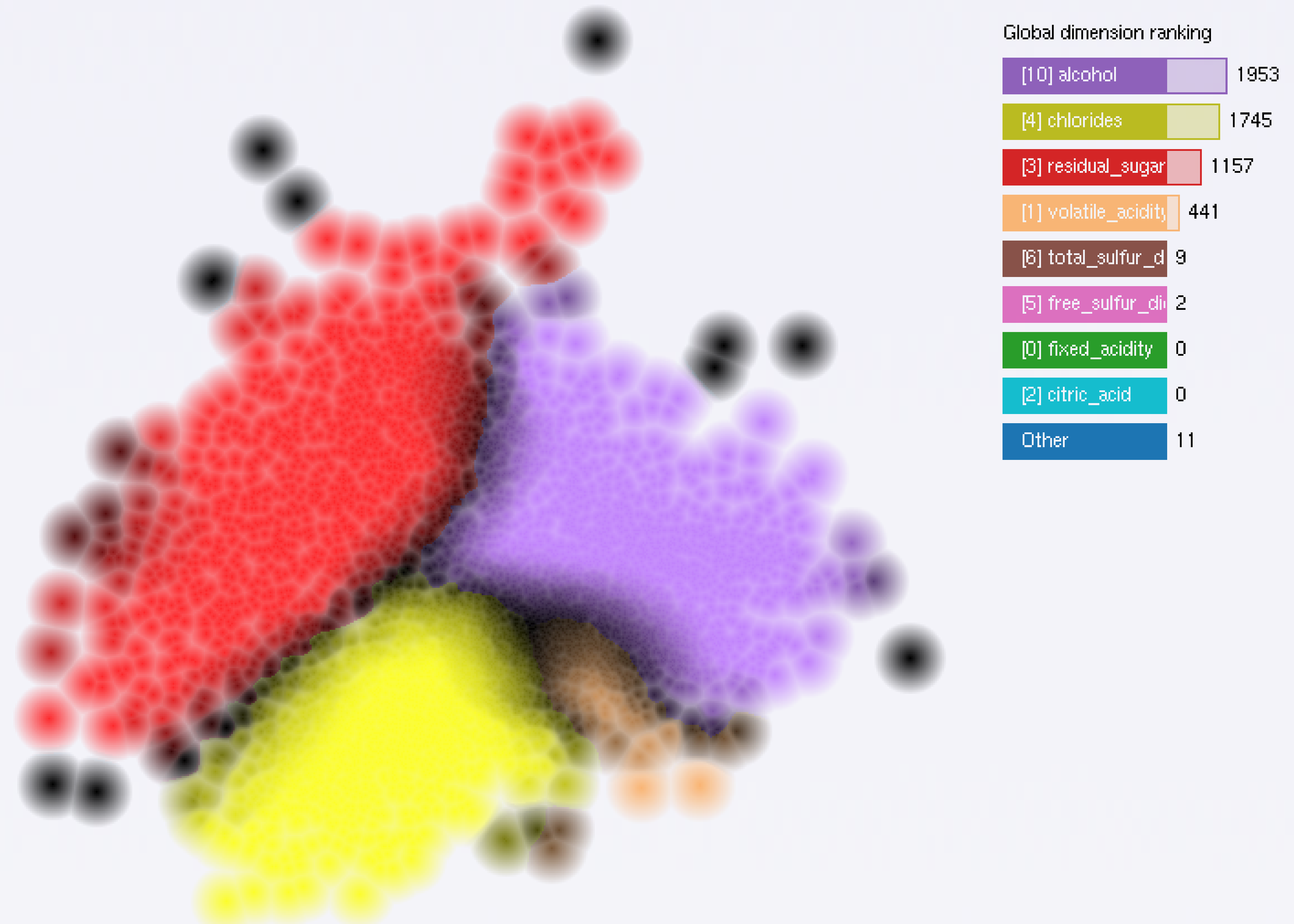

Figure 3: *Projected clusters can be explained and further partitioned by compact same-color areas, which are defined by the same top-ranked dimension among points neighborhoods.*

Dimension set map

To explain groups in a deeper level of detail we can utilize more than one dimension per point. Based on an user-defined threshold τ, we can select the dimensions whose ranks sum up to at most τ and create a set of dimensions per point. Next we assign colors to the most frequent dimension sets in the map, creating more specialized sub-partitions.

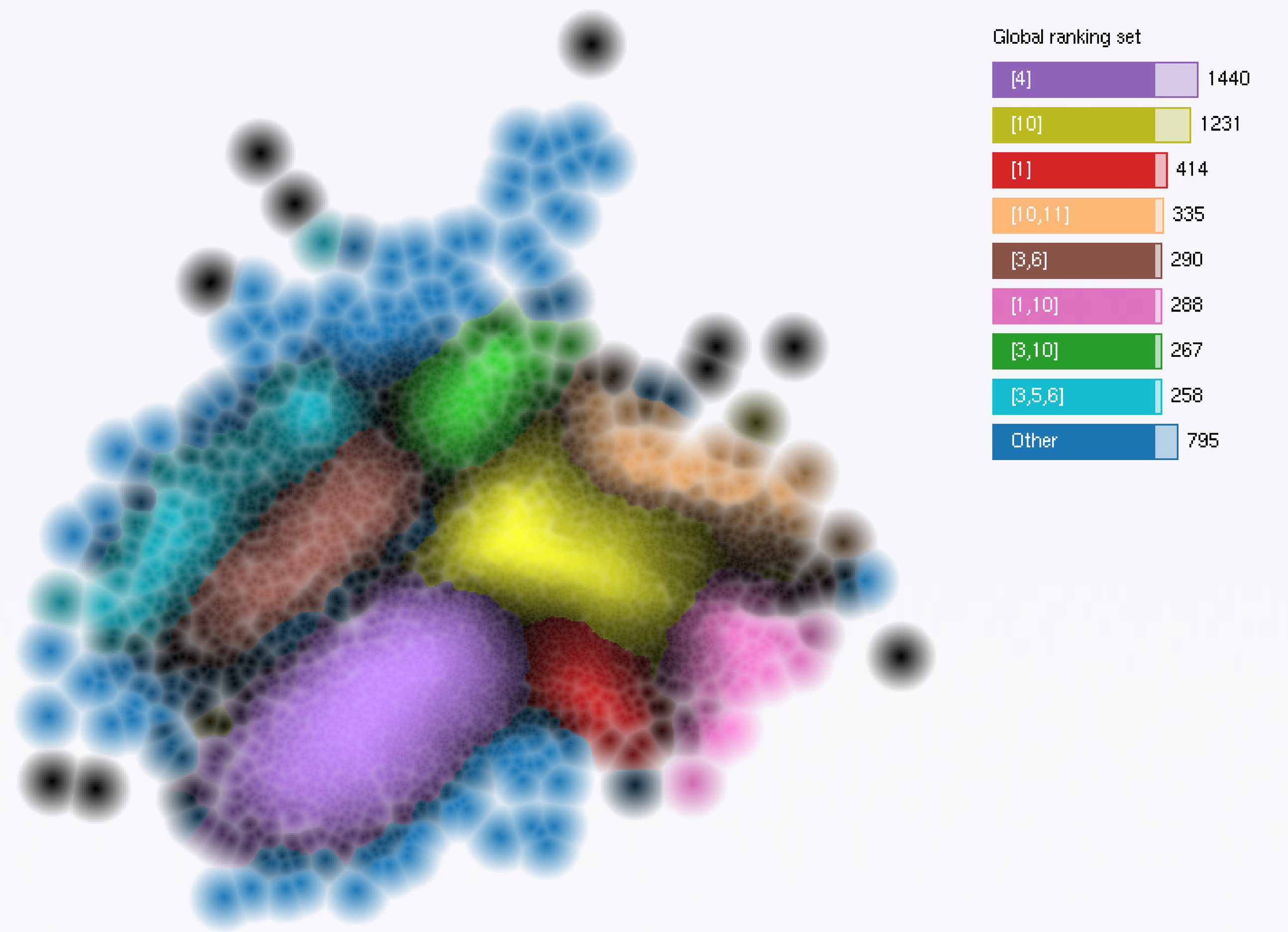

Figure 4: *Based on an user-defined similarity threshold, each point will define a set of more important dimensions. The more frequent ones are visualized as specialized sub-partitions of the points clusters.*

Future work

Ongoing work targets better visual encodings that map dimension names atop of projection point-groups. We also want to explaining regions by both dimensions and dimension-values, thereby leading to more refined explanations.

Contact

Renato Rodrigues Oliveira da Silva
Johann Bernoulli Institute for Mathematics and Computer Science
University of Groningen
Nijenborgh 9
9747 AG Groningen
e-mail: **rros@usp.br**

FAPESP

This work was supported by **FAPESP**, process nº 2014/01692-6