Beyond multidimensional data in model visualization: High-dimensional and complex nonnumeric data

Nathalie Vialaneix, Anne Ruiz-Gazen

Beyond model visualization for multidimensional data: high-dimension and complex non numeric data

Nathalie Villa-Vialaneix and Anne Ruiz-Gazen

INRA, UR 0785 MIAT
BP 52627
31326 Castanet Tolosan cedex - FRANCE
e-mail: nathalie.villa@toulouse.inra.fr

Toulouse School of Economics
Manufacture des tabacs, 21 allée de Brienne
31000 Toulouse - FRANCE
e-mail: anne.ruiz-gazen@tse-fr.eu

1 Introduction

We greatly appreciate the opportunity to read and discuss the paper "Visualizing statistical models: removing the blindfold" [42]. The article manages to close the gap between statistics and visualization by combining advanced methods from both fields to visualize statistical models and methods. The article provides a very nice overview of visualization tools for statistical models, which can be used to 1) understand the model itself (*i.e.*, what the model says about the data), 2) assess its relevance (*i.e.*, if the model is accurate to describe the data, if the model has been well trained...) and 3) evaluate the variability of a family of models, use this information to select a model within a family, evaluate its robustness and combine several models in a relevant manner.

We believe that this point of view is innovative and rarely addressed, statisticians tending to rely more on graphics to visualize the data themselves and on numeric criteria and statistics to evaluate models. However, as demonstrated in the article, applied statisticians would take great advantage of using interactive visualization methods for fitting and interpreting an adequate model. Moreover, such tools are now easily accessible, using for instance the R packages **rggobi**, **classifly**, **clusterfly** and **meifly** that are described in the article. Nowadays, data analysis is a highly developing field which has applications in many disciplines such as biology, genetics, economics, marketing or meteorology. Data are also increasingly challenging: high dimensional data, "big data", complex and possibly non numeric data. In this context, visualization must be part of the standard background available for any statistician or data scientist. Combining specialized statistical methods with visualization, as described in the article, should improve the ability to understand data and design good models.

Integrating visualization and statistics/data mining methods is an emerging trend which is developing very fast. More and more R packages now include a variety of graphics for exploring the results of the analyses: **FactoMineR** [20, 26] (multivariate exploratory analysis) or the emerging package **factoextra** [22] which provides **ggplot2** [41] graphics and syntax for **FactoMineR**, **mixOmics** [13, 27] (exploratory analysis of 'omics data), **SOMbrero** [39, 29] (self-organizing maps for multivariate data, contingency tables and dissimilarity data), **VIM** [36] (visualization of missing data), **GeoXp** [2, 25] (interactive exploratory analysis of spatial data, with linked brushing), among others. Most of the developers have now understood the importance of making these tools usable by anyone and have developed graphical interfaces, taking full advantages of the emerging interactive tools available in R such as **shiny** [11] (see, for instance, the R package **Factoshiny** [38] or the shiny interface for

SOMbrero[1]).

Here, we want to discuss further two important challenges for visualization of statistical methods in modern data and problems: the first one is the issue of dimension reduction (Section 2) and the second one is the issue of non numeric data sets (Section 3).

2 High dimension

When the data are numerical and lie in a high dimensional space, the data structure is expected to be contained in a low dimension subspace and dimension reduction methods are particularly relevant. Among the unsupervised dimension reduction methods, one can quote Exploratory Projection Pursuit (EPP) with Principal Component Analysis (PCA) as a special case, the projection index being the variance. And, more recently, the Invariant Coordinate Selection (ICS) method proposed by [37] and studied in [9, 21, 1]. The PCA and ICS methods are based on a spectral decomposition and lead to some orthogonal projection matrices that define nested vector subspaces. Once chosen the subspace dimension, it is possible to use biplots in order to represent together the observations and the variables (see **FactoMineR** [20, 26]). The EPP approach is different. It is based on some projection index that measures the interest of one- or two-dimensional orthogonal projections and on an optimization algorithm that optimizes the projection index over all possible projections. The question of the existence of an underlying model for such exploratory data analysis is also of interest. In [6, 5], PCA is derived as a least squares estimation method under a model which is called a "fixed effect model" and mainly assumes that the expectations of the observations belong to a subspace. In [10], another model is introduced in the context of EPP and ICS where the structure of the data we are interested in (mainly clusters and outliers) lie also in a subspace. In the framework of the above-mentioned models, the dimension of the subspace can be estimated and even in an optimal way in the case of PCA [6]. Reducing the dimension of course simplifies the task of the data scientist who only has to explore a reduced subspace instead of the full one. However in such a context, the data are vizualized in the model space and not the contrary.

Another difference between EPP, PCA and ICS concerns the standardization of the data. While ICS is affine invariant in the sense that the obtained scatterplots do not depend on the affine transformation of the raw data, PCA is only orthogonally invariant meaning that the scale of the variables has an impact on the results. Concerning EPP, it is well known that the results differ depending on whether the data are made spherical or not and it very often advisable to make the data spherical [16].

As stated in the paper under discussion, for EPP, the optimization algorithm may lead to some local optima and the user may have to try several starting points and a dynamic approach as the grand tour (see the R package **tourr**, [43]) to gain more insights in the data structure. Our experience with such dynamic tools is that their use may be tedious especially when the number of observations and the dimension are high and when there is no prior information concerning the data. Another possibility that follows the idea that "collections are more informative than singletons" is to use first a non-dynamic approach but with many starting points that will lead to a collection of potentially interesting projections. Then, one can either analyze the different selected projections or average the corresponding orthogonal projection matrices as proposed in [28] and implemented in the R package **LDRTools**. Figure 1 gives an application of the method on the wine data set which contains 178 observations and 13 variables. One hundred one-dimensional orthogonal projections are first obtained by minimizing the kurtosis projection index which is an index adapted to cluster detection [31]. The projections are then averaged in a orthogonal projection with a rank equal to three which gives some insight in the data set cluster structure. Note that the choice of the rank of the projection matrix gives the dimension of the projection subspace and is integrated in the averaging process which is based on a distance minimization criteria between orthogonal projection matrices. Different distances between projection matrices can be defined based on the Frobenius norm and this choice has some impact on the obtained results exactly as in the following section when the results depend on the choice of the dissimilarity.

[1] http://shiny.nathalievilla.org

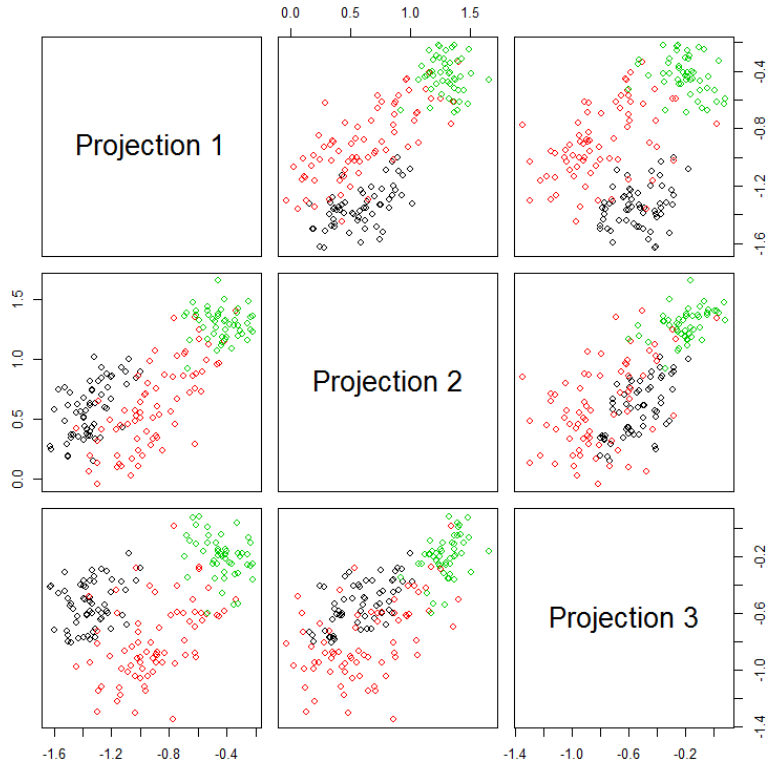

Figure 1: Average orthogonal projection of the wine data set observations obtained from 100 starting points using the kurtosis projection index

An alternative is to average projections obtained using different projection indices and several optimization algorithms as detailed in [4]. In a recent Java implementation of one-dimensional Exploratory Projection Pursuit called EPPLab[2], a plot for monitoring the convergence of the proposed (biologically inspired) optimization algorithms is included, together with a visualization of the ordered values of the projection index and cosines of the angles between projection vectors in order to check if they correspond to the same or to different projections. In this case, looking at the way the algorithm converges is helpful for tuning some of the algorithms parameters when such parameters exist. This Java program has been interfaced recently with R in the **REPPlab** package [15] and is described in [4].

Also, when examining plots obtained through EPP or ICS, the question of whether the obtained views reveal a significant structure or are only spurious is of course crucial. In [34], approximate p-values are derived by looking at the tail probability of the maximum of a Gaussian random field associated with a particular projection pursuit index. In [9], the choice of the dimension is based also on some kind of p-values. In [8], possible protocols are described in order to get inferential validity from visual discovery using exploratory data analysis. This point however deserves more attention.

3 Visualizing models for non numeric data

One of the major challenges of the modern data analysis is that data appear less and less in the form of standard multivariate vectors. Examples of such complex data include graphs (or networks, *i.e.*, graphs with additional information describing the nodes and/or edges), functional data, categorical time series, strings, ... To deal with such data in statistical models (clustering, classification, regression), a common approach is to rely on a numeric description of the pairwise relations between observations (*e.g.*, between two nodes in a graph, between two textual documents,...). These descriptors can be kernels [19], that are positive definite similarities which allow to embed the data into a Hilbert space [3] or any other (dis)similarities. Many statistical methods have been adapted to deal with such data (SVM [7] or LS-SVM for regression [35], kernel k-means [14], kernel PCA [32], relational topographic methods [18]...).

In these cases, the model cannot be described anymore by the three levels of specificity given in Section 2.1 of [42]: an additional level exists even before the model family, which corresponds to the choice of the (dis)similarity and is included in the model as a way to capture some features in the raw complex data. The choice of the (dis)similarity has to be designed with care, as a relevant way to summarize the data. Using the same principles as the ones described in [42] can help evaluating the relevance of a set of dissimilarities, from the data point of view and from the model point of view. An example of such an approach is provided in the sequel with a graph. However, we barely scratch the surface of this issue: such data could be the object of very specific treatments and innovative visualization techniques.

3.1 Visualization for dissimilarity evaluation

Graphs are used to represent relational data, *i.e.*, data in which entities (nodes of the graph) are described by their relations (edges of the graph). Common problems associated with these data are visualization and node clustering. As they are no "natural" Euclidean embedding of a graph, most methods that aim at displaying a graph are intrinsically very different from multivariate data visualization methods. A very popular family of algorithms is based on a reference to physics and mimics electric and spring forces to position the nodes in a 2D space: these are called Force Directed Placement (FDP) [17].

However, using more sophisticated statistical methods and models for graphs might be useful for gaining knowledge or using the graph for predictions. In this case, kernel and (dis)similarity methods have been proven useful (see [33] for examples in the field of biology) but the choice of the kernel/(dis)similarity is left to the user and can be critical to obtain meaningful results. Standard examples of common kernel/(dis)similarities include: 1) the Euclidean distance between the K eigenvec-

[2] https://github.com/fischuu/EPP-lab

tors associated with the K smallest eigenvalues of the graph Laplacian: this dissimilarity is the one used when performing spectral clustering [40]; 2) the heat kernel [24]; 3) the shortest path length between two nodes along the edges of the graph. Following the ideas of [42], Figure 2 shows an example of what can be d-in-ms plots and ms-in-d plots for (dis)similarity diagnostic in this case with a simple graph based on the novel "Les Misérables" from the French author Victor Hugo[3]. The first row of this figure uses a FDP approach to set the positions of the nodes: this can be seen as a visualization in the data space. At this stage, the model corresponds to the way the data are summarized through a given (dis)similarity: the colors of the nodes thus correspond to the (dis)similarity measure from one of the nodes' point of view (here, "Jean Valjean", the main character of the novel). The second row of the figure uses a MDS (but a projection based on a grand tour technique could have been used for a more interactive or original visualization) to display the nodes in a 2D space, similarly to what they are in the embedding (pseudo)-Euclidean space induced by the (dis)similarity. These plots can thus be seen as d-in-ms plots: the graph represented in the model space, which is the space induced by the (dis)similarity. The information provided by the color is the same as in the latter visualization. While the first row is easier to read, it is very limited for understanding the dissimilarity. It is restricted to one node's point of view and 77 such graphics must be used to have a complete visualization of all the (dis)similarity measures between nodes. The second row provides additional information: the heat kernel is the (dis)similarity that produces the most compact view while the shortest path length tends to give a more uniform view. The distance between the 8 eigenvectors of the Laplacian associated with the smallest eigenvalues is between the other two. From a clustering perspective, it is thus expected that the same model behaves very differently with these various summaries of the graph. For instance, Figure 3 shows that a simple hierarchical clustering of the nodes provides highly unbalanced clusters with the heat kernel and much more uniform cluster sizes with the shortest path length.

[3]This graph has been first used in [23] and is available in the R package **SOMbrero**.

3.2 Visualization for node clustering diagnostic

Using the shortest path length as a descriptor of the graph "les Misérables", The R package **SOMbrero** allows us to perform the relational version of the self-organizing map (SOM) algorithm [29]. This version extends the standard SOM method to data described by a dissimilarity matrix. The results commented in this section are the ones described in the package's vignette. The SOM parameters are set to obtain a net having dimensions 5×5 and the algorithm is trained with 385 iterations (default values in **SOMbrero** which corresponds to 5 times the number of observations).

Again, the recommendations made in [42] can be applied to this case: in relational SOM, prototypes[4] are expressed as symbolic convex combinations of the nodes of the graph

$$\text{prototype} \sim \sum_i \beta_i \times \text{node}_i$$

(with $\beta_i \geq 0$ and $\sum_i \beta_i = 1$). A dissimilarity induced by the dissimilarity chosen to summarize the graph can be computed for all pairs of prototypes, as explained in [29]. Using this trick, Figure 4 shows a typical d-in-ms view which is suggested in [12] and is implemented in the package: this view displays the neurons of the net as octagons. The dimensions of each side of the octogon is proportional to the dissimilarity between the prototype and its neighboring prototype in this direction. The color level indicates the number of nodes of the graph classified inside this neuron. The center of the net is composed of empty neurons (white neurons) and the top left hand side of the map (neurons 4 and 5) is rather separated from the rest.

Using the coordinates of the nodes obtained by FDP, a plot similar to the m-in-ds plot of Figure 17 in [42] can be obtained: each prototype being displayed at coordinates $\sum_i \beta_i \times (x_i, y_i)$ where (x_i, y_i) are the coordinates of node_i in FDP. In Figure 5, this technique is used to visualize the evolution of

[4]called "nodes" in the article [42]; we chose not to use this name in our article in order to differentiate between the representer of the clusters and the nodes of the original graph. The entities that compose the net are also called "nodes" in [42], we will call then "neurons" in order to differentiate them from their prototypes that have values in the data space.

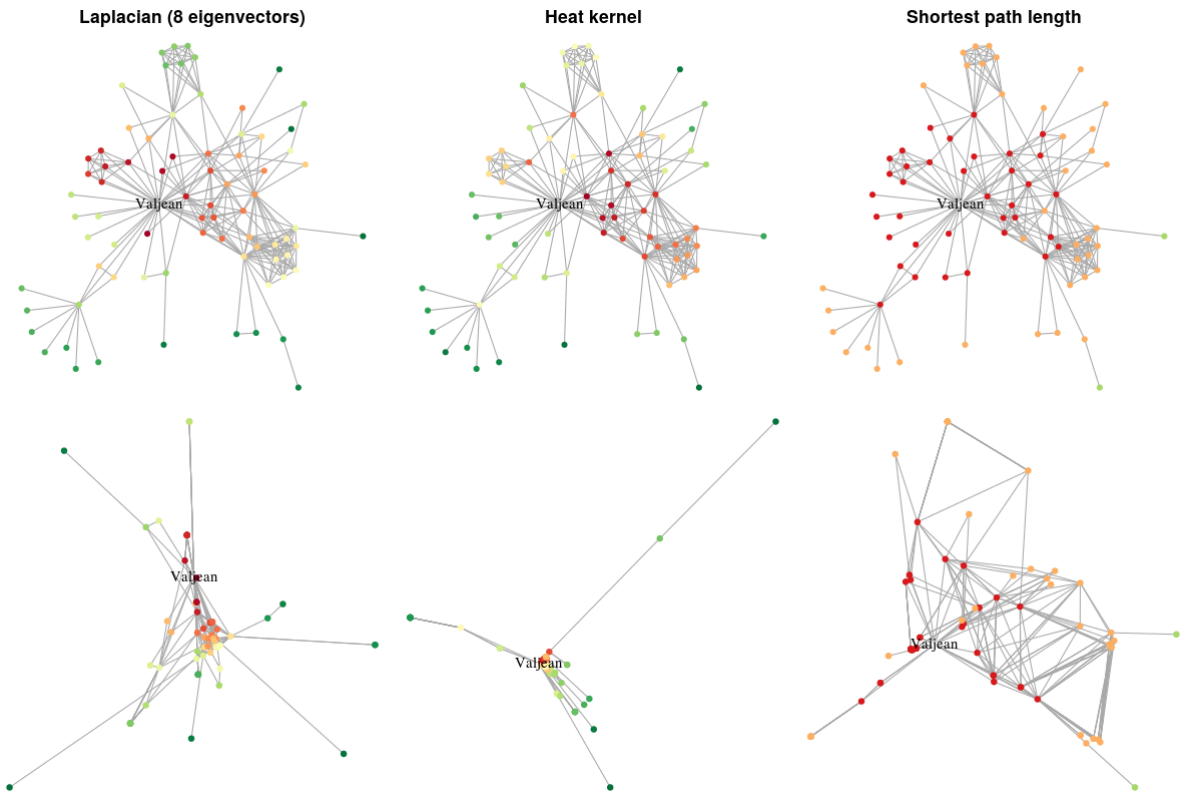

Figure 2: Comparison of different dissimilarities for the data set "Les Misérables" from an ms-in-d point of view (first row) and a d-in-ms point of view (second row). In the first approach, the nodes are positioned with a FDP algorithm. In the latter, they are positioned using the embedding in the (pseudo)-Euclidean subspace induced by the (dis)similarity thanks to Multi Dimensional Scaling (MDS). The colors of the nodes are obtained according to their distance to the main character of the novel, "Jean Valjean".

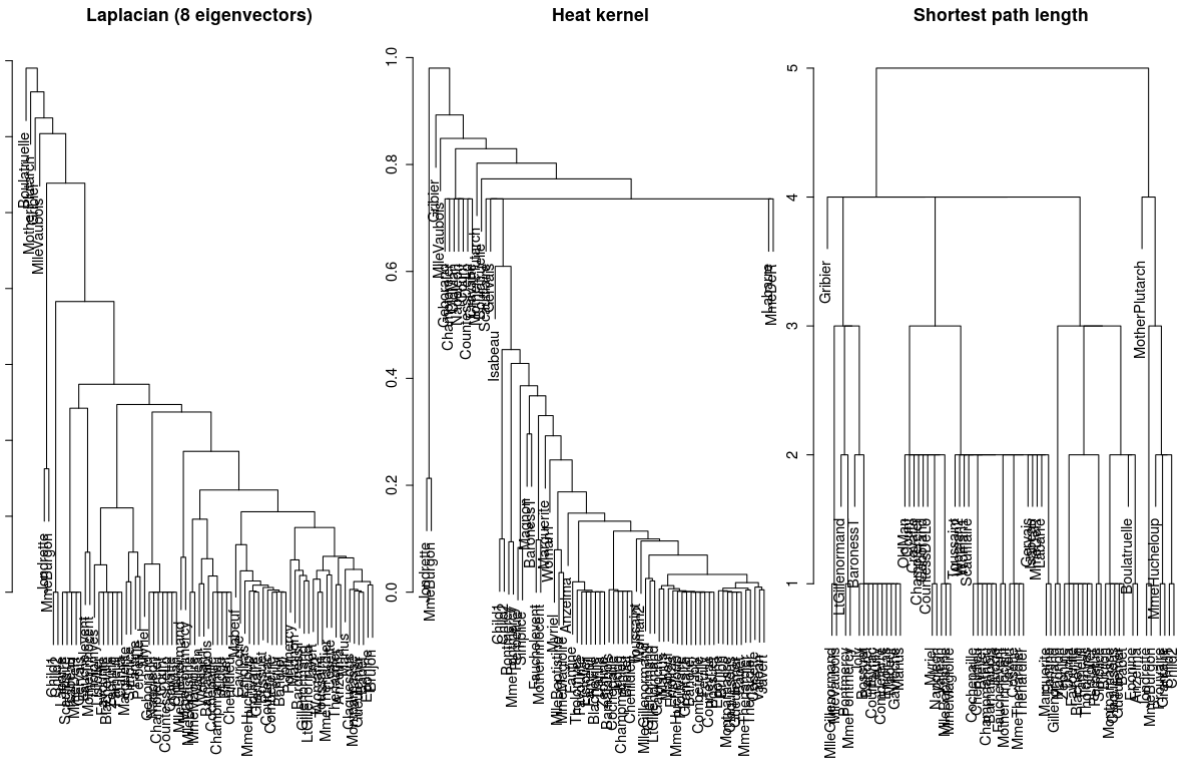

Figure 3: Comparison of the result of hierarchical clustering with various dissimilarities for the data set "Les Misérables".

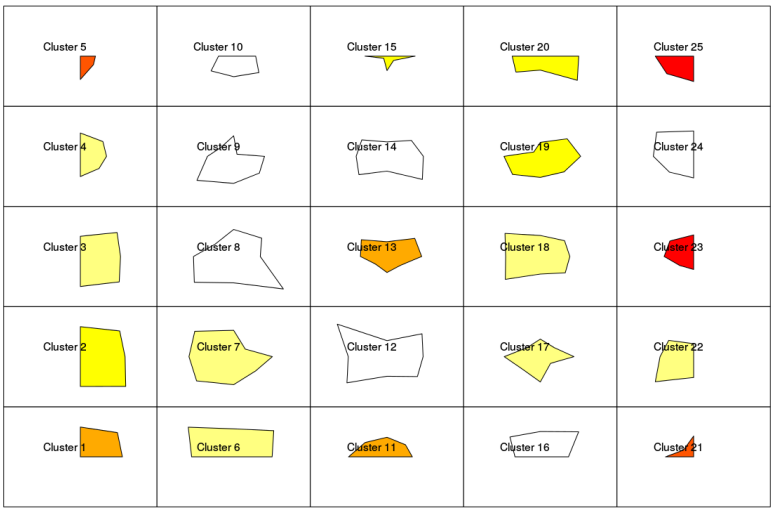

Figure 4: A visualization of the model space: each neuron is represented by an octogon whose dimensions are proportional to its distances with its neighboring prototypes.

the prototypes during the training of the net[5]. This figure shows that, starting from prototypes that have a rather central "position" in the graph, the SOM slowly organizes and stabilizes. The last iteration shows that neurons 4 and 5 that were discussed earlier correspond to nodes located at the bottom left hand side of the FDP representation. Actually, these nodes are those related to father Myriel, a bishop who is helping Jean Valjean at the beginning of the novel.

Further discussions about graph visualization using SOM are provided in [30].

4 Conclusion

In conclusion, the article [42] is a very interesting overview on statistical models visualization. We believe that this field, which integrates statistics and visualization, will be an increasing part of the standard statisticians' toolkit, and will facilitate the exploration and selection of models suited to the data and problems at hand. Modern data are a source of important challenges for this topic, that will have to handle very high dimensional data, complex (and possibly non vectorial) data and big data. We have not addressed this latter issue in our discussion but it is clearly an important and emerging topic (see for instance, the R package **bigvis** [44]).

[5] **SOMbrero** contains an option that allows the user to save a given number of intermediate states (prototypes and clustering) during the training.

References

[1] F. Alashwali and J. Kent. The use of a common location measure in the invariant coordinate selection and projection pursuit. arXiv preprint arXiv:1501.07240, 2015.

[2] Y. Aragon, T. Laurent, L. Robidou, A. Ruiz-Gazen, and C. Thomas-Agnan. *GeoXp: Interactive exploratory spatial data analysis*, 2013. R package version 1.6.2.

[3] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[4] A. Berro, D. Fischer, K. Nordhausen, and A. Ruiz-Gazen. Repplab: detecting groups and outliers using exploratory projection pursuit. Submitted, 2015.

[5] P. Besse. Models for multivariate data analysis. In R. Dutter and W. Grossmann, editors, *Proceedings of 11th Symposium on Computational Statistics (COMPSTAT94)*, pages 271–285, Vienna, Austria, 1994. Physica Verlag.

[6] P. Besse, H. Caussinus, L. Ferré, and J. Fine. Principal components analysis and optimization of graphical displays. *Statistics*, 19:301–312, 1988.

[7] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In 5^{th} *annual*

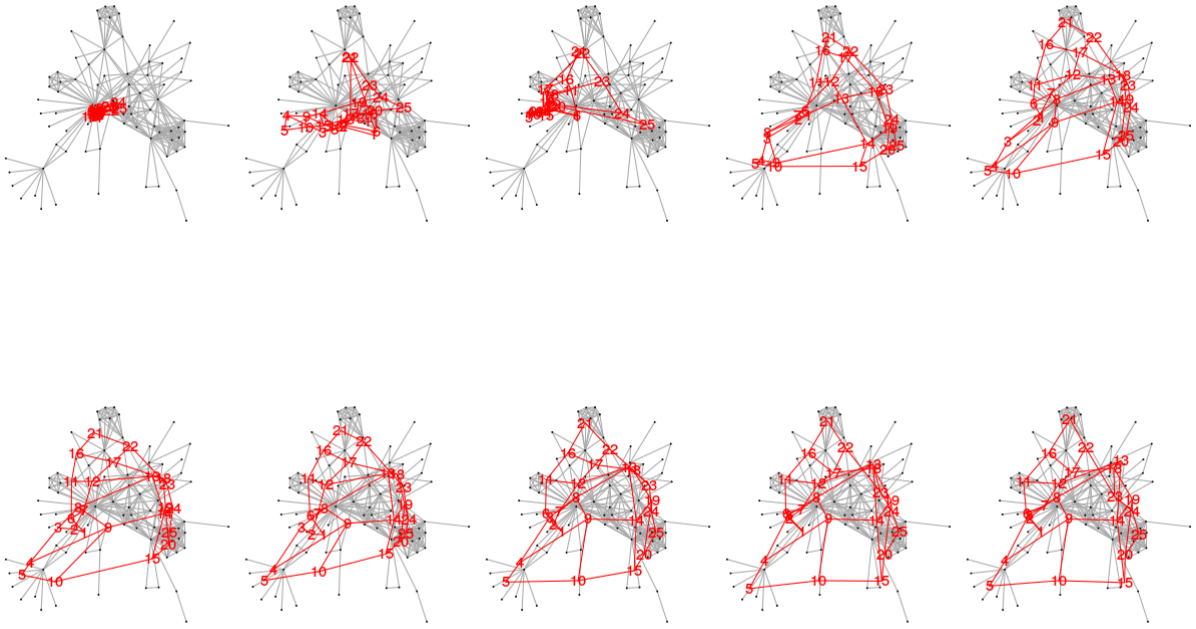

Figure 5: A visualization of the data space: evolution of the prototypes during the training of the SOM for dissimilarity data using the graph "Les Misérables" summarized through the shortest path length.

ACM Workshop on COLT, pages 144–152. D. Haussler Editor, ACM Press, 1992.

[8] A. Buja, D. Cook, H. Hofmann, M. Lawrence, E.K. Lee, D.F. Swayne, and H. Wickham. Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4361–4383, 2009.

[9] H. Caussinus, M. Fekri, S. Hakam, and A. Ruiz-Gazen. A monitoring display of multivariate outliers. *Computational Statistics & Data Analysis*, 44(1):237–252, 2003.

[10] H. Caussinus and A. Ruiz-Gazen. *Data Analysis*, chapter Exploratory projection pursuit, pages 67–92. Number 3. ISTE Ltd and John Wiley & Sons Inc, London, UK, 2009.

[11] W. Chang, J. Cheng, J.J. Allaire, Y. Xie, and J McPherson. *shiny: Web Application Framework for R*, 2015. R package version 0.11.1.

[12] M. Cottrell and E. de Bodt. A Kohonen map representation to avoid misleading interpretations. In M. Verleysen, editor, *Proceedings of the European Symposium on Artificial Neural Networks*, pages 103–110, Bruxelles, Belgium, 1996. Editions D Facto.

[13] S. Dejean, I. Gonzàlez, and K.A. Lê Cao. *mixOmics: Omics Data Integration Project*, 2014. R package version 5.0-3.

[14] I.S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means, spectral clustering and normalized cuts. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 551–556, 2004.

[15] D. Fischer, A. Berro, K. Nordhausen, and A. Ruiz-Gazen. *REPPlab: R Interface to* **EPP-lab**, *a Java Program for Exploratory Projection Pursuit.*, 2014. R package version 0.4.2.

[16] J.H. Friedman. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82(397):249–266, 1987.

[17] T. Fruchterman and B. Reingold. Graph drawing by force-directed placement. *Software, Practice and Experience*, 21:1129–1164, 1991.

[18] B. Hammer and A. Hasenfuss. Topographic mapping of large dissimilarity data sets. *Neural Computation*, 22(9):2229–2284, September 2010.

[19] T. Hofmann, B. Shölkopf, and A.J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008.

[20] F. Husson, J. Josse, S. Lê, and J. Mazet. *FactoMineR: Multivariate Exploratory Data Analysis and Data Mining*, 2015. R package version 1.29.

[21] P. Ilmonen, H. Oja, and R Serfling. On invariant coordinate system (ICS) functionals. *International Statistical Review*, 80(1):93–110, 2012.

[22] A. Kassambara. *factoextra: Visualization of the outputs of a multivariate analysis*, 2015. R package version 1.0.1.

[23] D.E. Knuth. *The Stanford GraphBase: A Platform for Combinatorial Computing*. Addison-Wesley, Reading, MA, 1993.

[24] R.I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th International Conference on Machine Learning*, pages 315–322, 2002.

[25] T. Laurent, A. Ruiz-Gazen, and C. Thomas-Agnan. GeoXp: An R package for exploratory spatial data analysis. *Journal of Statistical Software*, 47(2):1–23, 2012.

[26] S. Lê, J. Josse, and F. Husson. Factominer: An r package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18, 2008.

[27] K.A. Lê Cao, I. González, and S. Déjean. *****Omics: an R package to unravel relationships between two omics data sets. *Bioinformatics*, 25(21):2855–2856, 2009.

[28] E. Liski, K. Nordhausen, H. Oja, and A. Ruiz-Gazen. Averaging orthogonal projectors. arXiv preprint arXiv:1210.2575, 2012.

[29] M. Olteanu and N. Villa-Vialaneix. On-line relational and multiple relational SOM. *Neurocomputing*, 147:15–30, 2015.

[30] M. Olteanu and N. Villa-Vialaneix. Using SOMbrero for clustering and visualizing graphs. In revision for publication in the Journal de la Société Française de Statistique, 2015.

[31] D. Peña and F.J. Prieto. Cluster identification using projections. *Journal of the American Statistical Association*, 96(456), 2001.

[32] B. Schölkopf, A. Smola, and K.R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[33] B. Schölkopf, K. Tsuda, and J.P. Vert. *Kernel methods in computational biology*. MIT Press, London, 2004.

[34] J. Sun. Significance levels in exploratory projection pursuit. *Biometrika*, 78(4):759–769, 1991.

[35] J. Suykens, T.V. Gestel, J.D. Brabanter, B.D. Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, 2002.

[36] M. Templ, A. Alfons, A. Kowarik, and B. Prantner. *VIM: Visualization and Imputation of Missing Values*, 2014. R package version 4.1.0.

[37] D.E. Tyler, F. Critchley, L. Dümbgen, and H. Oja. Invariant co-ordinate selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):549–592, 2009.

[38] P. Vaissie, A. Monge, and F. Husson. *Factoshiny: Perform Factorial Analysis from FactoMineR with a Shiny Application*, 2015. R package version 1.0.

[39] N. Villa-Vialaneix, L. Bendhaïba, J. Boelaert, and M. Olteanu. *SOMbrero: SOM Bound to Realize Euclidean and Relational Outputs*, 2015. R package version 1.0.

[40] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[41] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer New York, 2009.

[42] H. Wickham, D. Cook, and D. Hofmann. Visualizing statistical models: removing the blindfold. *Statistical Analysis and Data Mining*, 2015. Forthcoming.

[43] H. Wickham, D. Cook, H. Hofmann, and A. Buja. **tourr**: An R package for exploring multivariate data with projections. *Journal of Statistical Software*, 40(2):1–18, 2011.

[44] H. Wickham, Y. Hue, and R Core Team. *bigvis: Tools for visualisation of big data sets*, 2013. R package version 0.1.