International Workshop on Machine Learning in Visualisation for Big Data (2020)
D. Archambault, I. Nabney and J. Peltonen (Editors)

Improving the Sensitivity of Statistical Testing for Clusterability with Mirrored-Density Plots

M. C. Thrun

[1] Databionics AG, Mathematics and Computer Science, Philipps-University Marburg, Germany
[2] Dept. of Hematology, Oncology and Immunology, Philipps-University Marburg, Germany

Abstract

For many applications, it is crucial to decide if a dataset possesses cluster structures. This property is called clusterability and is usually investigated with the usage of statistical testing. Here, it is proposed to extend statistical testing with the Mirrored-Density plot (MDplot). The MDplot allows investigating the distributions of many variables with automatic sampling in case of large datasets. Statistical testing of clusterability is compared with MDplots of the 1st principal component and the distance distribution of data. Contradicting results are evaluated with topographic maps of cluster structures derived from planar projections using the generalized U-Matrix technique. A collection of artificial and natural datasets is used for the comparison. This collection is specially designed to have a variety of clustering problems that any algorithm should be able to handle. The results demonstrate that the MDplot improves statistical testing but, even then, almost touching cluster structures of low intercluster distances without a predominant direction of variance remain challenging.

CCS Concepts

• ***Information systems*** → ***Clustering;***

1. Introduction

One type of cluster analysis can be desribed as the search for subsets of objects such that the members of each subset look like each other but do not look much like objects outside the cluster [Bon64], [BHV12]. Then the question arises if a dataset has the appropriate tendency for such clustering. A clustering algorithm may provide a result in which the grouping is also homogenous between the clusters meaning the objects were arbitrarily mapped into different groups. The dataset in the shape of an empty sphere called GolfBall [Ult05a] serves as an example because a hierarchical clustering algorithm will provide a dendrogram that proposes a clustering that does not reflect the entirely homogenous structure of the data [Thr18]. In such a case, the dataset does not possess cluster structures, and the clustering is potentially misleading for many applications [AAB19]. The property of a dataset having cluster structures is sometimes called clusterability [AAB19]. Typically, clusterability can be investigated visually with projections methods (e.g. [Sam70] [VPN*10], heatmaps (e.g. [HB11]), or cluster structure dependent with dendrograms (c.f. [Mur04]). This work investigates the clusterability on several datasets with well-known cluster structures that depict typical challenges arising in cluster analysis. The investigation is performed with statistical testing procedures for clusterability evaluated by [AAB19]. The two most-promising statistical tests evaluated are the multimodality tests for either distance distributions [KL12] or the 1st principal component [AW12]. In this work, the investigation of clusterablity is extended by the adaption of the visualization technique called Mirrored-Density plot [TGU20] to either distance distributions or the 1st principal component of data. The clusterability of contradicting results is evaluated further using the generalized Umatrix technique [UT17] resulting in topographic maps [TLLU16] of cluster structures.

2. Clusterability

This section describes the statistical testing procedures, the MDplot and the datasets used. Statistical testing for clusterability is performed with the R package ‘clusterability’ available on CRAN. The package provides statistical testing in order to investigate if a dataset possesses cluster structures [AAB19]. Dimensionality reduction with PCA (optionally scaled and standardized) can be performed for the datasets and then the 1st component is statistically tested for unimodality with either dip test [Hartigan/Hartigan, 1985] or Silverman test [Sil81]. Here, Hartigans’ dip test is used because it has the highest sensitivity in distinguishing unimodality from non-unimodality compared to other approaches [FD13]. If a distance matrix is given, then the dip test is applied to the vector of distances [KL12].

2.1. Mirrored-Density plot (MDplot)

The Mirrored-Density plot (MDplot) introduced in [TGU20] visualizes a density estimation in a similar way to the violin plot

DOI: 10.2312/mlvis.20201102

[HN98]. The MDplot uses for density estimation the Pareto density estimation (PDE) approach [Ult05b]. It can be shown that comparable methods have difficulties in visualizing the probability density function in case of uniform, multimodal, skewed, and clipped data if density estimation parameters remain in a default setting [TGU20]. In contrast, the MDplot is particularly designed to discover interesting structures in continuous features and can outperform conventional methods [TGU20]. The MDplot does not require any adjustments of parameters of density estimation, which makes the usage compelling for non-experts. In this work, the MDplot technique is adapted in the R package "FCPS" on CRAN with the goal to either visualize the 1st principal component of PCA or the vector consists of the elements of the upper triangle of the distance matrix. No prior knowledge or assumptions about the data are necessary. If multimodalities are visible, it can be assumed that the data possesses cluster structures suitable for conventional clustering algorithms [AAB19]. As an alternative to the investigation of clusterability, the unsupervised projection and visualization method of the Databionic swarm (DBS) [TU20b] is used to evaluate the datasets for which MDplot contradicts statistical testing. The first two modules of the parameter-free swarm algorithm require either a distance matrix or a data matrix and result in the topographic map [TLLU16] of the generalized U-Matrix [UT17] of projected points. Cluster analysis is performed in the third module. If valleys are visible, then the data possess cluster structures and the number of clusters is the number of valleys. However, in the third module, the purpose of DBS to provide a clustering is not relevant for this work. The algorithm is available in the R package "DatabionicSwarm" on CRAN.

2.2. Datasets

To investigate the clusterability of statistical and visualization methods, datasets are used which exploit fundamental clustering problems [TU20a]. Table 1 gives an overview of the challenges. Detailed descriptions and displays of the datasets and their challenges are presented in [TU20a], projections of datasets are investigated in [Thr18]. Additionally, 500 random 3D points are uniformly drawn between zero and one for which the Euclidean distance is computed. This last dataset is called “UnitSquare”.

3. Results

In the first part, all thirteen datasets are compared to statistical testing. In the second subsection the Euclidean distance distributions of artificial datasets are investigated. Selected datasets are evaluated with planar projections using the topographic map [TLLU16] based on the generalized U-Matrix [UT17].

3.1. Combining Statistical Testing with adapted MDplots

The MDplot of Clusterability is presented in Figure 1 and 2. On the x-axis, the name of each dataset and the p-value of statistical testing for each dataset are displayed. On the y-axis, the range of the mirrored density is shown. If multimodality is detected via statistical testing, then a dataset possesses cluster structures [AAB19]. The density is either estimated for the upper triangle of the distance matrix or the 1st component of the PCA (c.f. [AAB19]) which in this section was neither centered nor scaled. Nine out of thirteen datasets have either a significant p-value indicating cluster structures or a p-value near 1 indicating no cluster structures (GolfBalll data, UnitSquare distances). Statistical testing for clusterability shows a not significant p-value for Target, Hepta, Wingnut, Tetra in Figure 2. These datasets have no predominant direction of variance. The MDplot show mulitmodalities in the 1st principal component of Target and Hepta in Figure 2. The MDplot shows slight multimodalities in WingNut but agrees with statistical testing for Tetra, meaning that at least the cluster structures of Tetra are undetectable through the MDplot or statistical testing.

Name of Dataset	Challenge
Atom(1)*	Completely overlapping convex hull
Chainlink(1)	Linear nonseparable entanglements
EngyTime(1)	Overlapping clusters separable only by density
GolfBall(1)*	No distance-based cluster structures
Hepta(1)*	Nonoverlapping convex hulls with varying intracluster distances
Lsun3D(1)	Varying geometric shapes with noise defined by outliers
Target(1)*	Overlapping convex hulls combined with noise defined by four groups of outliers
Tetra(1)*	Low intercluster distances
TwoDiamonds(1)	Identification of the weak link in chain-like connected clusters
WingNut(1)*	Low intercluster distances versus large intracluster distances
Tetragonula	Smooth transition between clusters and outliers
Leukemia	Highly unbalanced cluster sizes

Table 1: *Summary of the challenges of the 12 datasets for cluster analysis [TU20a]. (1) Low-dimensional datasets were generated under the hypothesis that humans are most often able to group objects in two- or three-dimensional plots by eye [TU20a]. (*) without predominant direction of variance.*

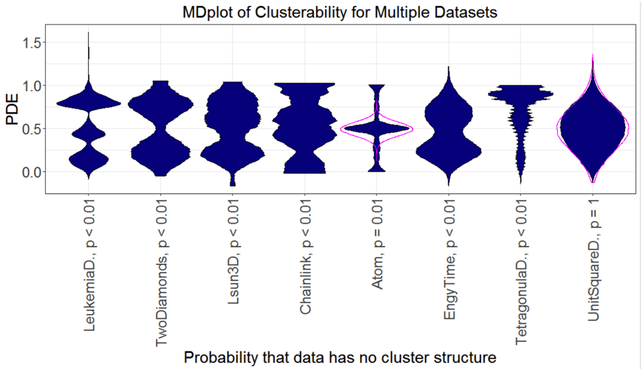

Figure 1: *Adapted MDplot for Clusterability visualize the density estimation of either the 1st principal componentif data matrices are given or the upper triangle of the distance matrix. Congruent p-values of statistical testing [AAB19] and the names of the datasets are written on the x-axis. Abrr.: D. = Distances*

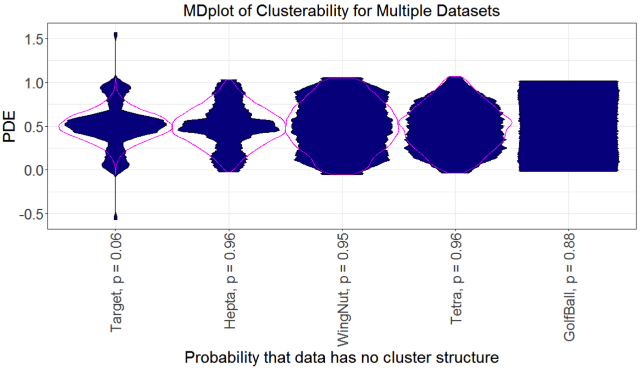

Figure 2: *Adapted MDplot for Clusterability visualize the density estimation of the 1st principal component and the names of the datasets are written on the x-axis. Statistical testing does not agree with the MDplot.*

Dataset	PCA (2)	Distances	Multimodality Visible
Atom	$p = 0.45$	$p = 0.01$	Yes, Yes
Chainlink	$p = 0.03$	$p = 0.01$	Yes, Yes
EngyTime	$p = 0.02$	$p = 1$	Yes, No
GolfBall	$p < 0.01$	$p = 0.01$	No, No
Hepta	$p < 0.01$	$p = 0.01$	Yes, Yes
Lsun3D	$p < 0.01$	$p = 0.01$	Yes, Yes
Target	$p = 0.06$	$p = 0.01$	Yes, Yes
Tetra	$p = 0.99$	$p = 0.63$	No, Slightly
TwoDiamonds	$p = 0.99$	$p = 0.07$	No, Yes
WingNut	$p = 0.27$	$p = 0.11$	No, No

Table 2: *Results of the dip test for the centered and scaled PCA (2) as well as the distribution of euclidean distances for artificial datasets. MDplots are only described due to space restrictions.*

3.2. Variants of of Statistical Testing Compared to MDplot

Table 2 presents the alternative approaches of testing for clusterability. Here, the data is centered and scaled prior to the PCA or the multimodality of the Euclidean distance distributions is tested with the dip-test. Descriptions of MDplots are in the same order as statistical testing. Results for Tetragonula, Leukemia, and UnitSquare are presented in the subsection above because only distance matrices were available for these three datasets.

3.3. Investigating Selected Cases with Planar Projections

In the second part, the cluster structures of the datasets Target, Hepta, Tetra, Wingnut are examined with planar projections of DBS, and compared to the reference of the GolfBall dataset, which does not possess cluster structures. The visualizatons presented here are topographic maps with hypsometric tints which correspond to high-dimensional distance and density structures. Hypsometric tints are surface colors that represent ranges of elevation. The contour lines are combined with a specific color scale. The colour scale is chosen to display various valleys, ridges, and basins: blue colours indicate small distances (sea level), green and brown colours indicate middle distances (low hills), and shades of white colours indicate vast distances (high mountains covered with snow and ice). Valleys and basins represent clusters, and the watersheds of hills and mountains represent the borders between clusters. In this 3D landscape, the borders of the visualisation are cyclically connected with a periodicity. Here, projected points are depicted in magenta. The topographic map of the generalized U-Matrix show clear cluster structures for Hepta, Target and Tetra because a number of valleys is presented in Figures 3, 4, 5. Still, the topographic only indicates cluster structures for WingNut in Figure 6, contrary to GolfBall for which no cluster structures are visible in Figure 7.

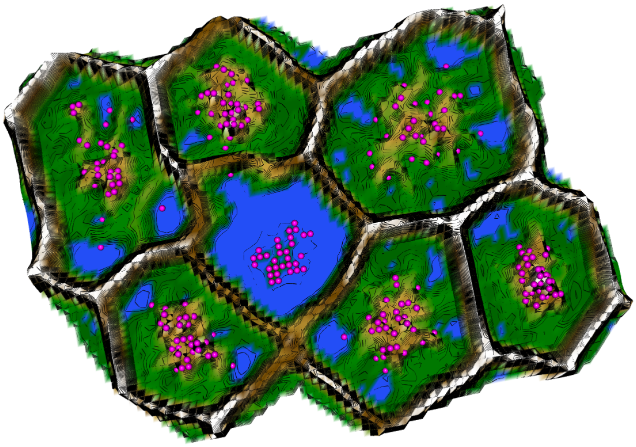

Figure 3: *Topographic map of the Hepta dataset visualizes clear cluster structures because several valleys are visible. The more dense cluster with smaller intracluster distances lies in a blue see.*

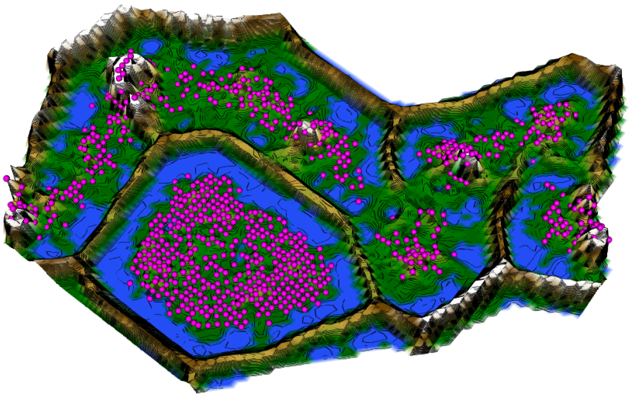

Figure 4: *Topographic map of the Target dataset visualizes clear cluster structures because two valleys are visible. Outliers are visualized on top of mountains or in vulcanos.*

4. Discussion

The results compared statistical testing to visualization approaches for clusterability. In many cases, statistical testing of multimodality of the 1st principal component seems to clearly distinct datasets that have cluster structures to datasets that have none reproducing

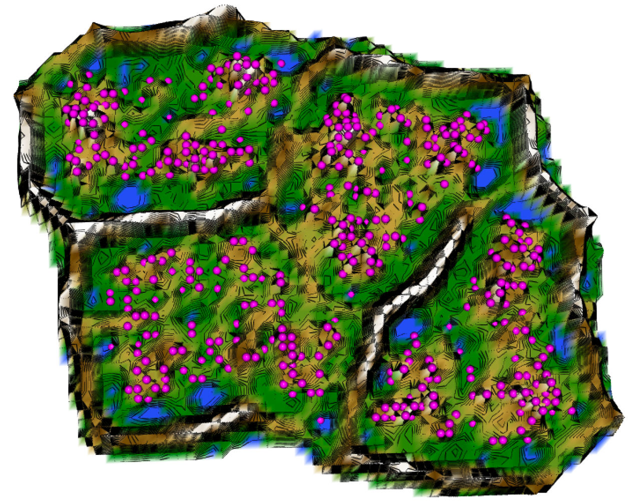

Figure 5: *Topographic map of the Tetra dataset visualizes clear cluster structures because several valleys are visible. Low intercluster distances are not challenging if the intracluster distances remain small.*

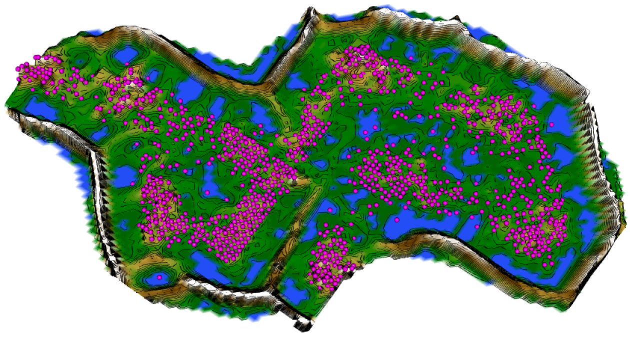

Figure 6: *Topographic map of the WingNut dataset indicates valleys and, thus, a cluster structure. Low intercluster distances become challenging for the topograpgic map if the intracluster distances get too large.*

the results of [AAB19]. Surprisingly, a 1D representation of data variance provides enough information to investigate the existence of a wide variety of cluster structures because the first two or three principal components usually do not define a subspace that is most informative about the cluster structure in data [Cha83], [DSC94], [VK01] and [AH94], if the task of clustering is defined as the grouping of similar objects. Topographic maps of planar projections were used in this work to evaluate datasets for which statistical testing contradicted the MDplot because they are a good alternative for the detection of clusterability [TU20b]. In every contradictory case, the topographic maps demonstrated that the visualized MDplot was more sensitive to the detection of clusterability than statistical testing. One of the reasons for the improvement lies in the usage of the PDE. Prior works demonstrated that the PDE is particularly suitable for the discovery of structures in continuous data and allows the discovery of mixtures of Gaussians [UTHGL15]. However, using topographic maps has the following disadvantage besides being computationally expensive. The topographic map is based on the generalized U-Matrix, which is computed toroidal for a planar projection meaning that the borders of the map are cyclically connected. The advantage is that cluster structures are not disrupted by boundaries [Thr18]. However, the usage of a toroidal map necessitates a tiled landscape display, which means every projected point and every hill or valley is shown four times. To obtain the 3D landscape shown above, an island usually is cut out manually, which can be a challenging task. The results for the dip test for distance distributions illustrated again that the MDplot is more sensitive to multimodality than statistical testing. However, it is not surprising that statistical testing and MDplot failed on the datasets of EngyTime and WingNut because both datasets are strictly based on structures defined by density. The results also illustrate that the detection of clusterability is challenging for datasets with almost touching clusters of Tetra and WingNut which have no predominant direction of variance. This insight was not reported priorly [AAB19]. Contrary to the work presented there, chaining effects in data (e.g., Chainlink) are detectable but not always statistically significant (Target). Comparing the two PCA options showed that statistical testing was dependent on preprocessing and the MDplot preferred non-scaled and non-centered data. In sum, the combination of visualization with statistical testing provides acceptable results for the decision if a dataset has cluster structures.

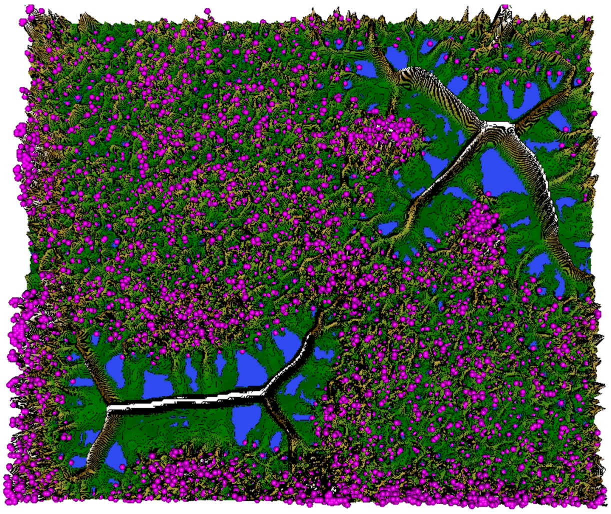

Figure 7: *Topographic map of the GolfBall dataset does not show valleys, indicating that the dataset has no cluster structures.*

5. Conclusion

This work demonstrates that the sensitivity of statistical testing can be improved by the MDplot. However, the detection of almost touching clusters with low intercluster distances an without a predominant direction of variance remain still challenging, even if such cluster structures are obviously separable by the human eye.

References

[AAB19] Adolfsson A., Ackerman M., Brownstein N. C.: To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition 88* (2019), 13–26. 1, 2, 4

[AH94] Arabie P., Hubert L.: Cluster analysis in marketing research. *Advanced Methods of Marketing Research* (01 1994). 4

[AW12] Ahmed M. O., Walther G.: Investigating the multimodality of multivariate data with principal curves. *Computational Statistics & Data Analysis 56*, 12 (2012), 4462–4469. 1

[BHV12] Bouveyron C., Hammer B., Villmann T.: Recent developments in clustering algorithms. In *ESANN* (2012), Citeseer. 1

[Bon64] Bonner R. E.: On some clustering technique. *IBM Journal of Research and Development 8*, 1 (1964), 22–32. URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5392274&isnumber=5392271, doi:10.1147/rd.81.0022. 1

[Cha83] Chang W.: On using principal components before separating a mixture of two multivariate normal distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 32*, 3 (1983), 267–275. 4

[DSC94] De Soete G., Carroll J. D.: *K-means clustering in a low-dimensional Euclidean space*. Springer, 1994, pp. 212–219. 4

[FD13] Freeman J. B., Dale R.: Assessing bimodality to detect the presence of a dual cognitive process. *Behavior research methods 45*, 1 (2013), 83–97. 1

[HB11] Havens T. C., Bezdek J. C.: An efficient formulation of the improved visual assessment of cluster tendency (ivat) algorithm. *IEEE Transactions on Knowledge and Data Engineering 24*, 5 (2011), 813–822. 1

[HN98] Hintze J. L., Nelson R. D.: Violin plots: a box plot-density trace synergism. *The American Statistician 52*, 2 (1998), 181–184. 2

[KL12] Kalogeratos A., Likas A.: Dip-means: an incremental clustering method for estimating the number of clusters. In *Advances in neural information processing systems* (2012), pp. 2393–2401. 1

[Mur04] Murtagh F.: On ultrametricity, data coding, and computation. *Journal of classification 21*, 2 (2004), 167–184. 1

[Sam70] Sammon J. W.: Interactive pattern analysis and classification. *IEEE Transactions on computers 100*, 7 (1970), 594–616. 1

[Sil81] Silverman B. W.: Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological) 43*, 1 (1981), 97–99. 1

[TGU20] Thrun M. C., Gehlert T., Ultsch A.: Analyzing the fine structure of distributions. *PLOS ONE under revision* (2020), preprint available at arXiv.or: arXiv:1908.06081. doi:arXiv:1908.06081. 1, 2

[Thr18] Thrun M. C.: *Projection Based Clustering through Self-Organization and Swarm Intelligence*. Springer, Heidelberg, 2018. doi:10.1007/978-3-658-20540-9. 1, 2, 4

[TLLU16] Thrun M. C., Lerch F., Lötsch J., Ultsch A.: Visualization and 3d printing of multivariate data of biomarkers. In *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)* (2016), Skala V., (Ed.), vol. 24, pp. 7–16. URL: http://wscg.zcu.cz/wscg2016/short/A43-full.pdf. 1, 2

[TU20a] Thrun M. C., Ultsch A.: Clustering benchmark datasets exploiting the fundamental clustering problems. *Data in Brief 30* (2020), 105501. doi:https://doi.org/10.1016/j.dib.2020.105501. 2

[TU20b] Thrun M. C., Ultsch A.: Swarm intelligence for self-organized clustering. *Journal of Artificial Intelligence in press* (2020). doi:10.1016/j.artint.2020.103237. 2, 4

[Ult05a] Ultsch A.: Clustering wih som: U* c. In *5th Workshop on Self-Organizing Maps (WSOM)* (2005), vol. 2, pp. 75–82. 1

[Ult05b] Ultsch A.: Pareto density estimation: A density estimation for knowledge discovery. In *Innovations in classification, data science, and information systems* (Berlin, Germany, 2005), Baier D., Werrnecke K., (Eds.), vol. 27 of *Proceedings o f the 27th Annual Conference of the Gesellschaf für Klassifikation*, Springer, pp. 91–100. 2

[UT17] Ultsch A., Thrun M. C.: Credible visualizations for planar projections,. In *12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM)* (Nancy, June 2017), Cottrell M., (Ed.), pp. 1–5. doi:10.1109/WSOM.2017.8020010. 1, 2

[UTHGL15] Ultsch A., Thrun M. C., Hansen-Goos O., Lötsch J.: Identification of molecular fingerprints in human heat pain thresholds by use of an interactive mixture model r toolbox (adaptgauss). *International journal of molecular sciences 16*, 10 (2015), 25897–25911. 4

[VK01] Vichi M., Kiers H. A.: Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis 37*, 1 (2001), 49–64. 4

[VPN*10] Venna J., Peltonen J., Nybo K., Aidos H., Kaski S.: Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research 11*, Feb (2010), 451–490. 1