SIRIUS: Dual, Symmetric, Interactive Dimension Reductions Supplementary Material

Michelle Dowling, John Wenskovitch, *Student Member, IEEE*, J.T. Fry, Scotland Leman, Leanna House, and Chris North

Index Terms—Dimension reduction, semantic interaction, exploratory data analysis, observation projection, attribute projection

1 Introduction

In Sect. 5 of the SIRIUS paper, we mentioned comparisons between SIRIUS and existing techniques for visualizing and interacting with high-dimensional data. With Table 1, we highlight the novelty of SIRIUS, showing that no other technique addresses all three goals described in Sect. 3 of the SIRIUS paper. We also acknowledge that SIRIUS is not a comprehensive technique in that it does not explicitly afford other types of insights, such as those that have been added to the bottom of Table 1. However, it does not prohibit the addition of these insights either.

2 Comparing SIRIUS with Other Techniques

To directly compare SIRIUS with existing visual analytics techniques for high-dimensional data, we compare the capabilities of SIRIUS and the techniques exemplified by Andromeda [8], Dust & Magnet [10], Star Coordinates [5], Dis-Function [2], LAMP [4], Dimension Projection Matrix/Tree [11] (shortened to "DP Matrix/Tree in Table 1), the visualization proposed by Turkay et al. [9], Data Context Map [3], Intent Radar [6], and Doc-Function [1]. These comparisons are summarized in Table 1, which emphasize how these other techniques meet the three Goals described in Sect. 3 of the SIRIUS paper. Note that these comparisons are based on how the visualization is presented in their perspective publications and whether the visualization directly enables the given task or directly provides the given information. For example, Andromeda by default provides a similarity-based projection of the observations of high-dimensional data. Although a projection of the attributes could be achieved by using a transpose of the original data matrix as input, this additional projection is not automatically given as part of the visualization. Therefore, we consider Andromeda to *not* provide a similarity-based projection of the attributes. The following subsections provide further details on why we filled each cell of Table 1 in the manner presented.

2.1 Goal 1: Similarity-Based Projections

As discussed in Sect. 2 of the SIRIUS paper, most visual analytics techniques, including Andromeda, Dis-Function, and LAMP, focus on the observations. Thus, only a similarity-based projection is provided for the observations in many of the techniques in Table 1. However, Doc-Function provides a similarity-based projection of the attributes, and the Intent Radar maps the similarity of the attributes to the angle around the radar. As discussed in Sect. 2.3 of the SIRIUS paper, the Data Context Map projects the observations and the attributes into the same space using MDS, which necessarily distorts at the projection of the observations, the projection of the attributes, or both. In the visualization from Turkay et al., three scatterplots are provided: one that visualizes the observations using one attribute for each of the axes, one that visualizes the observations using two principle components that eliminate outliers, and one that visualizes the attributes based on their mean and standard deviation. While each of these scatterplots could arguably be a visualization that uses a simplified definition of similarity to produce the scatterplot, we consider these scatterplots to be too simple to be classified true similarity-based projections, as they don't use all the observations or attributes in a similar manner to MDS or PCA. Therefore, we propose the visualization described by Turkey et al. only partially supports this goal.

- *Michelle Dowling, John Wenskovitch, and Chris North are with the Virginia Tech Department of Computer Science. E-mail: dowlingm — jw87 — north@cs.vt.edu.*
- *J.T. Fry, Scotland Leman, and Leanna House are with the Virginia Tech Department of Statistics. E-mail: fryjt1 — leman — lhouse@vt.edu.*

2.2 Goal 2: Exploring the Projections

2.2.1 Manipulating Importances to Explore Similarities

Given the general focus on projections of observations and interactions therein as opposed to that of attributes, it is perhaps expected that none of our selected comparison techniques enable this manipulation of observation importances. However, Andromeda (via PaI), Dust & Magnet (via increasing the magnitude of a magnet), and Star Coordinates (via increasing the length of an attribute's axis) enable manipulation of attribute importances to explore observation similarities.

2.2.2 Manipulating Similarities to Explore Importances

Techniques such as Andromeda (via PrI) and Dis-Function (via dragging and dropping nodes) enable direct manipulation of the observation similarities to explore the attribute importances. In Andromeda, the result of this interaction is reflected in the position of the attribute sliders, whereas Dis-Function portrays this information in a bar graph. While LAMP affords a similar interaction, the importance given to each of the attributes is not portrayed to the analyst. Similarly, Doc-Function enables analysts to use a similar interaction technique on a projection of attributes, but the importance given to each observation is not available to the analyst. We therefore argue that LAMP and Doc-Function both only partially support this goal.

2.3 Goal 3: Relating Importances to Each other

While no other techniques related observation importances to attribute importances, the Intent Radar is the only other technique in our list that relates attribute importances to observation importances. This is accomplished by visually encoding each attribute's importance as its distance from the center of the radar. This information is then used to determine the importance of each document, with the documents provided to the right of the radar visualization.

2.4 Other Mechanisms to Generate Insight

Although our implementation of SIRIUS provides a unique interface that enables powerful interactions and insights, it is not a comprehensive implementation; there are other insights commonly afforded in other exploratory data analysis techniques for high-dimensional data. For example, distributions show how a particular piece of data compares to all others or how common certain values are. This helps analysts understand the given dataset at a high level as analysts generally have low cognitive dimensionality, as described by Self et al. [7]. Similarly, clustering data helps analysts be able to automatically group data together, which also helps give a high level overview of the dataset. We discuss each of these types of common insights in detail with regards to our selection of techniques.

Table 1. A summary of the comparisons between SIRIUS and existing visual analytics techniques for high-dimensional data. “O” or “A” denotes that the given technique has the specified ability, whereas “o” or “a” denotes that the specified ability is only partially supported or only supported under certain circumstances. A more thorough description of why each cell contains it perspective marks is provided in the supplementary materials.

		SIRIUS	**Andromeda** [8]	**Dust & Magnet** [10]	**Star Coordinates** [5]	**Dis-Function** [2]	**LAMP** [4]	**DP Matrix/Tree** [11]	**Turkay et al.** [9]	**Data Context Map** [3]	**Intent Radar** [6]	**Doc-Function** [1]
SIRIUS Goals	Goal 1: Similarity-based projection of observations (O or o) and attributes (A or a (Sect. 2.1)	OA	O	O	O	O	O	OA	oa	oa	A	A
	Goal 2: Manipulate attribute importance (A or a) or observation importance (O or o) to explore observation similarities or attribute similarities (e.g. PaI on the attributes or observations), respectively; Sect. 2.2)	OA	A	A	A							
	Goal 2: Manipulate observation similarities (O or o) or attribute similarities (A or a) to explore attribute importances or observations importances (e.g. PrI on the observations or attributes), respectively; Sect. 2.2)	OA	O			O	o					a
	Goal 3: Relate attribute importances to observation importances (O or o) or vice versa (A or a) (Sect. 2.3)	OA									O	
Other	Distribution of observations across attributes (O or o) or attributes across observations (A or a)		O	oa	o	Oa		o	Oa	o		OA
	Clustering of observations (O or o) or attributes (A or a)			o			O	OA			A	

The most common insight afforded by our selection of techniques is the distribution of observations across attributes, which can be seen in Andromeda (seeing the raw data values for selected nodes along the attribute sliders), Dis-Function (the parallel bars view), the visualization provided by Turkey et al. (by manipulating the axes of the first scatterplot), and Doc-Function (searching and the Highlight feature). The Data Context Map only partially supports this insight by allowing analysts to manipulate the ranges for the contour lines, through which analysts can eventually learn the distribution. Star Coordinates also partially supports this insight by allowing analysts to select value ranges of interest for each axis, which can reveal the distribution of observations. Alternatively, analysts can manipulate the size and orientation of the attribute axes to view the distribution of observations across a single attribute. Similarly, the Dust & Magnet visualization partially supports viewing distributions of observations across attributes by having observations move faster towards a moved attribute if it has a higher value for that attribute. However, Dust & Magnet and Dis-Function also partially support insights regarding the distribution of attributes across observations through visual encodings of node color and size, and the raw data matrix (respectively). Dimension Projection Matrix/Tree partially supports this insight by allowing analysts to refine the attributes used in a single projection of observations in the matrix to one attribute for each axis. Doc-Function directly affords analysts this insight by hovering over or clicking keywords.

Finally, LAMP, and Dimension Projection Matrix/Tree show clustering of observations. LAMP determines clusters both via *k*-nearest neighbors and via a silhouette coefficient. Dust & Magnet partially supports this insight by coloring the observations by a analyst-selected categorical attribute. Dimension Projection Matrix/Tree directly supports automatic clustering of observations or of attributes by clicking on a corresponding button in a toolbar which performs the clustering using a kNN graph. However, the Intent Radar clusters attributes by mapping the results from agglomerative clustering to both color and position.

REFERENCES

[1] E. T. Brown. *Learning from Users Interactions with Visual Analytics Systems*. PhD thesis, Tufts University, 2015.

[2] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang. Dis-function: Learning distance functions interactively. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 83–92, Oct 2012. doi: 10.1109/VAST.2012.6400486

[3] S. Cheng and K. Mueller. The data context map: Fusing data and attributes into a unified display. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):121–130, 2016.

[4] P. Joia, D. Coimbra, J. A. Cuminato, F. V. Paulovich, and L. G. Nonato. Local affine multidimensional projection. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2563–2571, 2011.

[5] E. Kandogan. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pp. 107–116. ACM, New York, NY, USA, 2001. doi: 10.1145/502512.502530

[6] T. Ruotsalo, J. Peltonen, M. Eugster, D. Głowacka, K. Konyushkova, K. Athukorala, I. Kosunen, A. Reijonen, P. Myllymäki, G. Jacucci, et al. Directing exploratory search with interactive intent modeling. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pp. 1759–1764. ACM, 2013.

[7] J. Z. Self, M. Dowling, J. Wenskovitch, I. Crandell, M. Wang, L. House, S. Leman, and C. North. Observation-level and parametric interaction for high-dimensional data analysis. *ACM Transactions on Interactive Intelligent Systems*.

[8] J. Z. Self, R. Vinayagam, J. T. Fry, and C. North. Bridging the gap between user intention and model parameters for data analytics. In *SIGMOD 2016 Workshop on Human-In-the-Loop Data Analytics (HILDA 2016)*, p. 6, 06/2016 2016.

[9] C. Turkay, P. Filzmoser, and H. Hauser. Brushing dimensions-a dual visual analysis model for high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2591–2599, 2011.

[10] J. S. Yi, R. Melton, J. Stasko, and J. A. Jacko. Dust & magnet: multivariate information visualization using a magnet metaphor. *Information Visualization*, 4(4):239–256, 2005.

[11] X. Yuan, D. Ren, Z. Wang, and C. Guo. Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2625–2633, 2013.