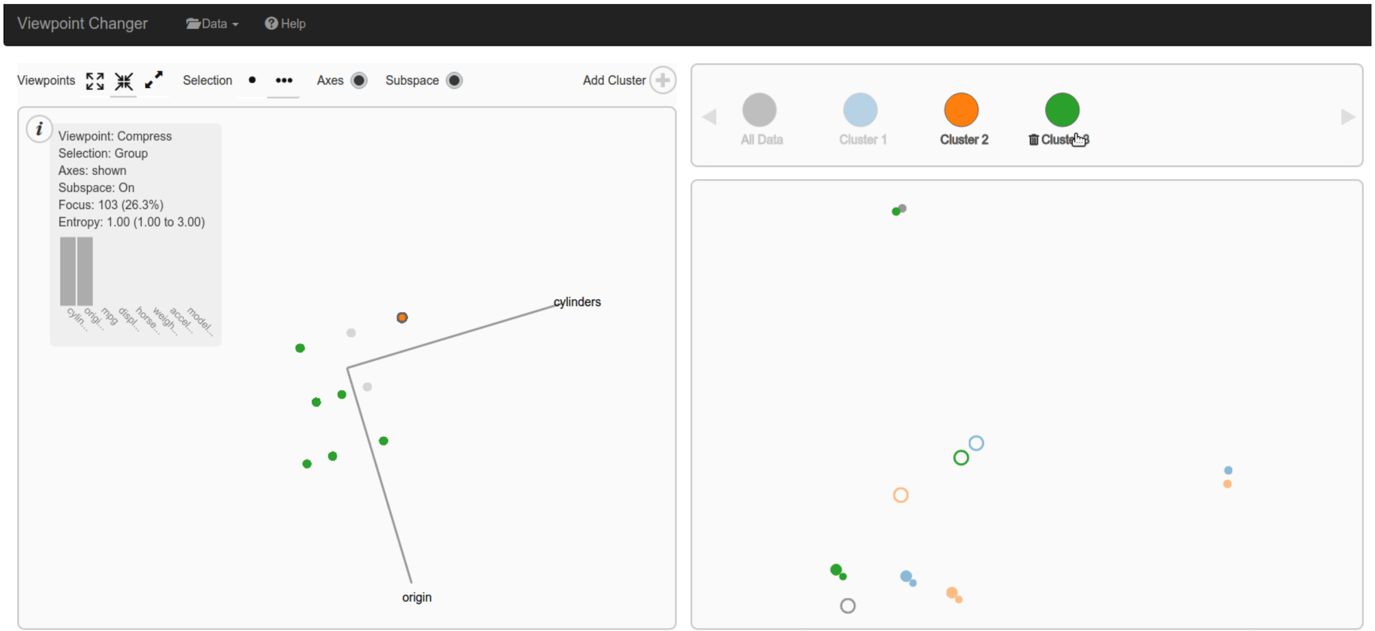
Exploring High Dimensional Data Through Locally Optimized Viewpoint Selection



Viewpoint: Compress
Selection: Group
Axes: shown
Subspace: On
Focus: 103 (26.3%)
Entropy: 1.00 (1.00 to 3.00)

Fig. 1. Teaser here.

Abstract—Dimension reduced projection is usually obtained via a global optimization. It gives a good overview of the data, but cannot satisfy different users with different focuses. That's because any local data could be distorted and misleading due to projection errors. To address this problem, we propose an interactive visualization method, to customize projections for better local analysis. First, we allow users to define their point of interest (POI) data. Then we generate local projections to minimize distortions regarding the POI. Multipe optimizations are provided for different analyses. We also reveal relationships among different POIs, by comparing their local projections. At last, our method is proved effective via case studies with a real-world dataset.

Index Terms—Dimension-reduced projection, local analysis, high-dimensional data

1 Introduction

Dimension-reduced projection is widely used for high-dimensional data analysis. It seeks to approximate the original distribution in a low-dimensional space. Such approximation is often globally optimized to make a good overview of the data. But a sole overview cannot show all aspects. Detail analysis is a necessary supplement, as suggested by Shneiderman in his information seeking mantra [26]. However, due to approximation errors, data relationships will inevitably be distorted. The distortions are hard to ignore when it comes to local data. They largely harm the perception of local structures, yet are often transparent to users. Even if distortions are shown in the projection [28] [2], users have no means to control their distribution. In general, globally optimized projections make good overviews, but are not suitable for local data analysis.

One way to alleviate this problem, is to observe the data in a different perspective. Some previous works [12] [21] [17] allow users to change the projection by adjusting dimension weights. It helps to clarify detailed structures and find out more useful information. To help build a targeted analysis, featured clusters are often provided beforehand [21] [19]. But users know nothing about the given clusters. They have to solve their puzzles by manually searching the enormous data space. It's a blinded and exhausting process, where users have no clue how to steer the dimensions to get a better view. Even if interesting projections are found, it's hard to explain or assess them without distortion information. There is no guarantee that local structures will be shown more precisely in the new perspective.

The low-efficiency of dimension-based exploration is caused by three facts. First, dimension weights are too abstract for users to understand and control. It's hard to explain something like $'25\%Weight+75\%Height'$. Second, users don't know about the interplay between dimension weights and the projection. They cannot foresee the effects of changing weights, which directly leads to a try-and-error process. Third, the exploration is blinded without a clear target. It somehow depends on luck to get informative findings.

Compared to dimensions, it's more reasonable and more efficient to explore projections via data features. It is the core spirit behind

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org.
Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx/

feature-based mining techniques. Projection pursuit [9] searches for projections to optimize predefined indices. The rank-by-feature framework [25] ranks a series of projections by their interestingness. In more recent works, users are involved to specify desired features [13] and relationships [11] [10]. These methods, though effective, largely depend on predefined metrics or users' prior knowledge. It makes them unsuitable for an interactive exploration starting from scratch. In fact, such techniques have the power to reduce local distortions in a linear projection. But little attention had been paid in this aspect.

Inspired by previous works, we wonder if we could introduce projection pursuit into interactive explorations, to solve the local distortion problem. Users can specify their point of interest (POI) data. Then we return projections with the least local distortion for a better analysis. There are good reasons for such kind of exploration. First, data relationships are easier to perceive, understand, and control. Users may not be familiar with dimensions, but they are interested in, and also familiar with featured relationships like clusters and outliers. They can decide which part of data should be studied in detail. Second, projection pursuit can be used to reduce distortion of local data. It's far more efficient than manual controls. Users can be free from parameter toning, and focus more on data features. Third, the exploration is more targeted with a POI. Besides, users don't have to explain projections by their dimensions. It makes more sense to explain them in the context of data features.

These thoughts were pushed forward into the method we'd like to present in this paper. It is an interactive approach to steer high-dimensional data exploration, through consecutive local analyses in locally optimized projections. To be specific, the exploration contains four steps. First, for any given projection, we help the user find a piece of interesting local data. Then for some chosen data, we provide distortion-free projections to enhance its features for a better perception. In addition, subspaces related to the projections can reveal causes of the features. Since the chosen data could be a false cluster or missing some important pieces, we also help to shape it into a more consistent and complete cluster. At last, whenever some valuable local data is found, the user can store it for further analyses. We provide a 'projection map' containing all featured projections to support the analysis. It helps to compare different pieces of data, and organize the high-dimensional exploration. More details will be introduced in the following sections.

In summary, our contributions include:

(1) We help users customize dimension-reduced projections for a targeted and distortion-free local data analysis.

(2) We propose a data-based interactive exploration method. Users are able to steer the exploration efficiently, by focusing on local data analysis, rather than dimension toning.

The remainder of this paper is structured as follows. In the next section, we briefly review the related literature. Section 3 gives an overview on the proposed method based on the exploration process. Then we elaborate each part of our method in detail in Section 4. Section 5 presents case studies to demonstrate the effectiveness of our method. In session 6, we have a discussion about weaknesses and potential improvements of our method. At last, we end this paper with the conclusions.

2 Related Work

2.1 Dimension-Driven Projection Exploration

Dimension Manipulation and Grand Tour [21] [17] [12]

2.2 Data-Driven Projection Pursuit

Projection Pursuit [9] [7] [25] [13] [22]

Projection Pursuit for Classification [27] [6] [29] [24] [1]

Targeted Projection Pursuit [14] [4] [10] [11]

2.3 High-Dimensional Local Data Analysis

Distortion Analysis [20] [18] [28]

Subspace Cluster Estimation [5] [16] [31] [19]

3 Overview

In this work, we propose to steer high-dimensional data exploration via local data analysis. We facilitate local analysis by reducing local projection distortions. In this section, we first clarify the concept of local distortion reduction. Then we demonstrate how it supports an efficient data exploration. At last, we give an overview of the interactive exploration process supported by our method.

3.1 Local Distortion Reduction

Distortion usually refers to the gap between original data distances and the projected distances. We call it the *distance distortion*. Our approach helps to reduce such distortions locally for a more faith perception. However, it may not be enough to support data feature analyses. There have been lots of dimension reduction techniques [15] [3] [30] [23], aiming to reduce distance distortions globally. But a more recent research [8] has shown that, those projections cannot guarantee a good performance for certain analytic tasks. The main cause, in our opinion, is the existence of relationship distortions.

By 'relationship', we refer to a relative concept of distance, i.e. 'close' or 'far'. In the high-dimensional space, relationship should be defined in a certain data range and dimensional subspace. Assume that there are four data items distributed in a two-dimensional plane, as shown in Figure 2. (Figure to be added.) When talking about data A, B and C, we can say that 'C is far away from B (compared to A)'. But when talking about data B, C and D, C seems close to B (compared to D). On the other hand, C is closer to B than A in dimension X, while the opposite happens in dimension Y. When combing all data and dimensions, weaker relationships give way to the stronger ones. The weak ones (e.g. C is closer to B than A) can no longer be perceived. Even the strong ones are not as obvious as in the original context. We call it the *relationship distortion*. The situation is alike in more complex real-world datasets. Integrated measurements cannot reflect local relationships precisely. That's why approximating the overall distances cannot guarantee enhanced data features.

In contrast, reducing distortion of the interested relationship can support a more targeted analysis. For example, when we talk about two objects being similar, we tend to ignore their differences temporally and vice versa. We believe that distortion reduction should be built upon certain analytic targets, namely the local data and relationships. It is the key to revealing hidden local features. That's the reason we provide two different types of distortion reduction, for different purposes. In fact, we will prove that distance distortion is a special case of relationship distortion. Both goals can actually be achieved in the same framework. More details will be introduced in Section 4.

3.2 Workflow

Our method promotes an efficient high-dimensional data exploration, by facilitating distortion-free local analysis. It not only helps to maintain a faithful perception, but aims to enhance local features for better information mining. To be specific, the exploration contains four steps (see Figure 2):

Step. 1: First, we present a globally optimized projection as an overview of the data. For the given projection, we help the user find a piece of interesting local data. Projection distortion and cluster suggestions are displayed to indicate potential outliers and clusters. The data chosen by user is called a *focus*, meaning that it's the current focus in local analysis.

Step. 2: After some focus is chosen, we find its most featured projections for a targeted analysis. By 'features', we refer to three kinds of local data relationships we defined.

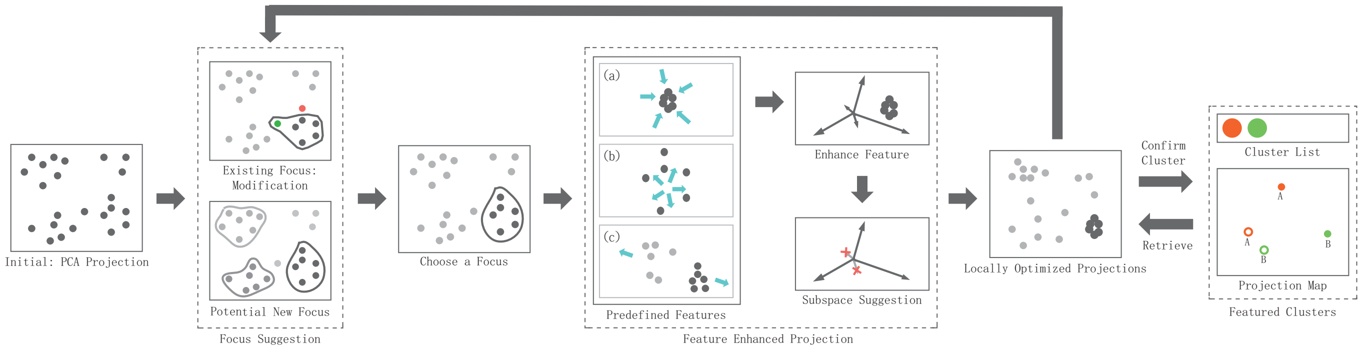


Fig. 2. The overview of delivery system.

Projections are optimized to show these relationships with the least distortion. Based on the projection, we suggest a dimensional subspace that's most related to the feature. Its dimensions help to explain causes of the relationships.

Step. 3: Since the focus is chosen in a projection, it could be a false cluster or missing some important pieces. We provide suggestions to help shape the focus into a more consistent and complete cluster. Whenever the focus is changed, the featured projections will also be updated. The process then returns to Step 2. This loop goes on until the user fully understands structures of the focus, and acquires a satisfactory cluster. After that, he can save the results and transfer his attention to another focus, returning to step 1.

Step. 4: When some informative local data is found, the user can store it in the focus list for a further analysis. A 'projection map' is provided for featured projections of all focuses. It helps to compare different focuses, and navigate the high-dimensional exploration.

The whole process is all about clarifying data relationships and figuring out the causes. Users never have to worry about configuring the projection. They are able to transfer seamlessly between different local analyses. In the next section, we'll introduce in detail how we support such exploration in each step.

4 High-dimensional Data Exploration Guided by Distortion-Free Local Analysis

As shown in the workflow, the proposed method supports a four-step exploration. In this section, we'll elaborate details of our method in each step of the exploration process.

4.1 Discovering Interesting Local Data

Following Shneiderman's suggestion [26], we first provide the PCA projection as an overview of the data. Then we help the user find an interesting subset as the focus of subsequent local analysis.

In a projection, there are two situations where some local data is considered interesting. The first case is related to distance distortion. Incorrect distances result in false neighborhood relationships. Closely distributed data may be far away in the original space and vice versa. Data involved in a distorted local area is regarded informative in the projection. It's also the basic idea in previous works concerning about local errors [20] [28]. But such analysis can only focus on each datum at a time. It's hard to describe group relationships in this context. In the second case, the data is involved in some featured relationships, like being an outlier or a cluster. The relationship may have been distorted, but it's still strong enough to dominate the current projection. Hence, it makes a reasonable focus for a further study. Besides, it's suitable to describe a group of data in the context of relationships, rather than distance errors.

To put it simply, distance distortion analysis focuses more on the neighborhood of a single datum. Relationship distortion analysis helps to study a group of data. Regarding the two cases, we adopt different means to help user find an interesting local focus.

4.1.1 Datum Suggestion Based on Distance Distortion

For any given projection, we consider a datum interesting if its distances to other data have been severely distorted. To measure the distortion, we accumulate distance errors for each datum in the projection:

$$Error(\mathbf{x}'_i) = \sum_{j=1}^{n} (Dist(\mathbf{x}_i, \mathbf{x}_j) - Dist(\mathbf{x}'_i, \mathbf{x}'_j))^2, i = 1, 2, \cdots n$$

Here the $\mathbf{x}_i$ and $\mathbf{x}'_i$ represents the original data and the projected data respectively. Distance is measured by the Euclidean distance metric, taking into account all dimensions. We use point size to encode the accumulated distortion of each datum, as shown in Figure. Figure to be added.

On the other hand, we provide interactive hints to reveal the real distances. The approach is similar to that used in [28], but uses a different metaphor. When user hovers on the projection, we construct a so-called 'high-dimensional lantern' using interpolation. Assume that the hovered position corresponds to a two-dimensional datum $\mathbf{p}'$, we interpolate its high-dimensional counterpart as follows:

$$\mathbf{p} = \sum_{i=1}^{n} \mathbf{w}_i \cdot \mathbf{x}_i = \sum_{i=1}^{n} \frac{Dist(\mathbf{x}'_i, \mathbf{p}')^{-1}}{\sum_{j=1}^{n} Dist(\mathbf{x}'_j, \mathbf{p}')^{-1}} \cdot \mathbf{x}_i$$

The interpolation weight $\mathbf{w}_i$ of data $\mathbf{x}_i$ depends on its distance to the hovered place in the projection. Closer data get larger weights. When user hovers right on $\mathbf{x}'_i$, $\mathbf{w}_i$ equals 1 while all the other weights get 0. The result equals the original data: $\mathbf{p} = \mathbf{x}_i$.

By the interpolation, we aims to infer what kind of data is desired by the user. Then this desired point acts as a high-dimensional lantern, shedding lights on all the other data to indicate their distances. With the lighting metaphor, we encode distance information using the saturation tunnel in HSL color space:

$$Saturation(\mathbf{x}'_i) = \max\{(\alpha d_i^2 + \beta d_i + \gamma)^{-1}, 1\},$$

$$d_i = Dist(\mathbf{x}_i, \mathbf{p}), \quad i = 1, 2, \cdots n$$

The data gets high saturation, if it's close to the interpolated point in the original space. The parameters α, β and γ come from the inverse-square law of the lighting model. Empirical values are chosen to accommodate most datasets. When some datum has a large distance distortion, its lights will not be able to illuminate its neighbors in the projection. In contrast, some far away points will be highlighted as the real neighbors. Figure. demonstrates the actual effect. Figure to be added.

In summary, large data points are potentially interesting data with high distortion and inconsistent illumination. With the hints, user hovers around the projection like experiencing an adventure. He holds a lantern to explore unknown structures in the complex data space. Compared to [28], this method enables a more smooth and natural perception of distance information.

4.1.2 Cluster Suggestion

4.2 Featured Viewpoints for the Focus

4.2.1 Desirable Local Features

4.2.2 Feature Enhanced Projections

4.2.3 Subspace Suggestion

4.3 From Focus to Cluster

4.3.1 Focus Improvement

4.3.2 Cluster Comparison via Viewpoint Map

5 Case Study

6 Discussion

7 Conclusion

Acknowledgments

The authors wish to thank A, B, C. This work was supported in part by a grant from XYZ.

References

[1] G. Albuquerque, M. Eisemann, and M. A. Magnor. Perception-based visual quality measures. In *2011 IEEE Conference on Visual Analytics Science and Technology, VAST 2011, Providence, Rhode Island, USA, October 23-28, 2011*, pages 13–20, 2011.

[2] M. Aupetit. Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing*, 70(7-9):1304–1330, 2007.

[3] I. Borg and P. J. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[4] 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 2012, Seattle, WA, USA, October 14-19, 2012*, pages 83–92, 2012.

[5] K. M. Carter, R. Raich, and A. O. H. III. On local intrinsic dimension estimation and its applications. *IEEE Trans. Signal Processing*, 58(2):650–663, 2010.

[6] J. Choo, H. Lee, J. Kihm, and H. Park. ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2010, Salt Lake City, Utah, USA, 24-29 October 2010, part of VisWeek 2010*, pages 27–34, 2010.

[7] D. Cook, A. Buja, J. Cabrera, and C. Hurley. Grand tour and projection pursuit. *Journal of Computational and Graphical Statistics*, 4(3):155–172, 1995.

[8] R. Etemadpour, R. Motta, J. G. de Souza Paiva, R. Minghim, M. C. F. de Oliveira, and L. Linsen. Perception-based evaluation of projection methods for multidimensional data visualization. *IEEE Trans. Vis. Comput. Graph.*, 21(1):81–94, 2015.

[9] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Computers*, 23(9):881–890, 1974.

[10] M. Gleicher. Explainers: Expert explorations with crafted projections. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2042–2051, 2013.

[11] X. Hu, L. Bradel, D. Maiti, L. House, C. North, and S. Leman. Semantics of directly manipulating spatializations. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2052–2059, 2013.

[12] D. H. Jeong, C. Ziemkiewicz, B. D. Fisher, W. Ribarsky, and R. Chang. ipca: An interactive system for pca-based visual analytics. *Comput. Graph. Forum*, 28(3):767–774, 2009.

[13] S. Johansson and J. Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Trans. Vis. Comput. Graph.*, 15(6):993–1000, 2009.

[14] P. Joia, D. Coimbra, J. A. Cuminato, F. V. Paulovich, and L. G. Nonato. Local affine multidimensional projection. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2563–2571, 2011.

[15] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[16] E. Kandogan. Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations. In *2012 IEEE Conference on Visual Analytics Science and Technology, VAST 2012, Seattle, WA, USA, October 14-19, 2012*, pages 73–82, 2012.

[17] D. J. Lehmann and H. Theisel. Orthographic star coordinates. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2615–2624, 2013.

[18] S. Liu, B. Wang, P. Bremer, and V. Pascucci. Distortion-guided structure-driven interactive exploration of high-dimensional data. *Comput. Graph. Forum*, 33(3):101–110, 2014.

[19] S. Liu, B. Wang, J. J. Thiagarajan, P. Bremer, and V. Pascucci. Visual exploration of high-dimensional data through subspace analysis and dynamic projections. *Comput. Graph. Forum*, 34(3):271–280, 2015.

[20] R. M. Martins, D. B. Coimbra, R. Minghim, and A. C. Telea. Visual analysis of dimensionality reduction quality for parameterized projections. *Computers & Graphics*, 41:26–42, 2014.

[21] J. E. Nam and K. Mueller. Tripadvisor$^{\mathrm{n\text{-}d}}$: A tourism-inspired high-dimensional space exploration framework with overview and detail. *IEEE Trans. Vis. Comput. Graph.*, 19(2):291–305, 2013.

[22] D. T. Nhon and L. Wilkinson. Scagexplorer: Exploring scatterplots by their scagnostics. In *IEEE Pacific Visualization Symposium, PacificVis 2014, Yokohama, Japan, March 4-7, 2014*, pages 73–80, 2014.

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

[24] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. *Comput. Graph. Forum*, 31(3):1335–1344, 2012.

[25] J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, 2005.

[26] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages, Boulder, Colorado, USA, September 3-6, 1996*, pages 336–343, 1996.

[27] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Comput. Graph. Forum*, 28(3):831–838, 2009.

[28] J. Stahnke, M. Dörk, B. Müller, and A. Thom. Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions. *IEEE Trans. Vis. Comput. Graph.*, 22(1):629–638, 2016.

[29] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. A. Magnor, and D. A. Keim. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, IEEE VAST 2009, Atlantic City, New Jersey, USA, 11-16 October 2009, part of VisWeek 2009*, pages 59–66, 2009.

[30] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[31] X. Yuan, D. Ren, Z. Wang, and C. Guo. Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2625–2633, 2013.