International Symposium on Visual Analytics Science and Technology (2010)
J. Kohlhammer and D. Keim (Editors)

Stress Maps: Analysing Local Phenomena in Dimensionality Reduction Based Visualisations

C. Seifert, V. Sabol and W. Kienreich

Know-Center Graz, Austria

Abstract

Challenges in Visual Analytics frequently involve massive repositories, which do not only contain a large number of information artefacts, but also a high number of relevant dimensions per artefact. Dimensionality reduction algorithms are commonly used to transform high-dimensional data into low- dimensional representations which are suitable for visualisation purposes. For example, Information Landscapes visualise high-dimensional data in two dimensions using distance-preserving projection methods. The inaccuracies introduced by such methods are usually expressed through a global stress measure which does not provide insight into localised phenomena. In this paper, we propose the use of Stress Maps, a combination of heat maps and information landscapes, to support algorithm development and optimization based on local stress measures. We report on an application of Stress Maps to a scalable text projection algorithm and describe two categories of problems related to localised stress phenomena which we have identified using the proposed method.

Categories and Subject Descriptors (according to ACM CCS): I.3.8 [Computer Graphics]: Applications—

1. Introduction

The visual representation of large document repositories represents a frequent challenge in the field of Visual Analytics. The information landscape is a common visual metaphor capable of conveying complex relationships is the Information Landscape [KBC*07, DHJ*98]. It uses the metaphor of a geographic map to provide insight into topical clusters and employs spatial proximity in the 2D layout to represents the topical relatedness.

A plethora of projection algorithms have been developed [GKWZ08] for projecting the document set into a low-dimensional (2D) visualisation space while preserving the high-dimensional relationships as good as possible. It is obvious that complex relationships present in a very high dimensional space cannot be perfectly represented in a low dimensional visualisation space. Nevertheless, the ability of projection algorithms to preserve original relationships is crucial for visualisation users attempting to identify patterns in the data set. The goodness of fit for projection algorithms is usually evaluated by computing a global stress value which basically expresses the cumulative difference between the high-dimensional and low dimensional distances.

Each projection from a high-dimensional space to a low-dimensional one introduces an inherent error which appears in the visualisation as local phenomena. Also, in order to scale to large data sets sophisticated projection algorithms employ various optimisation techniques. These optimisations often apply neighbourhood-based strategies in order to reduce the amount of data comparisons. At the same time projection algorithms often need to produce 2D layouts which fulfil certain usability requirements. These requirements and the various optimisations can introduce further localised errors and phenomena which are cannot be properly detected by a global stress measure. For example, two projection algorithms might produce layouts with similar global stress values, where one has a uniform stress distribution and the other produces a local stress peaks. A neighbourhood-based stress measure, proposed in [CB09], focuses on local goodness of fit. While emphasizing localised quality of the projection, the measure is computed globally over the whole data set and will likely not detect isolated phenomena.

In this paper, we propose the use of Stress Maps, a combination of heat maps and information landscapes, to support algorithm development and optimization based on local stress measures. We present users with a heat map display of local stress values which mimics the topology of the in-

DOI: 10.2312/PE/EuroVAST/EuroVAST10/013-018

formation landscape. Users can easily identify high stress areas in an overview and zoom in on such areas to identify the individual information artefacts responsible for observed phenomena. We illustrate this methodology by applying it to a scalable text projection algorithm. We have been able to identify two categories of problems related to localised stress phenomena. We were able to verify both problem categories from our knowledge on the behaviour and implementation of the algorithm.

The paper is organised as follows: In Section 2 we briefly discuss relevant state-of-the art in information landscapes, dimensionality reduction techniques and stress measures. We also discuss some important, recent related work. In Section 3 we present our approach. We report on experimental results in Section 4. We draw conclusions and present future work in Section 5.

2. State-of-the-Art

In this section, we first discuss related work in information landscapes because we conducted our experiments in this application area of projection algorithms. Because our approach is, in principle, applicable to arbitrary projection algorithms, we provide a brief overview on dimensionality reduction techniques. We also discuss available stress measures which can be explored using our approach. Finally, we reference some important, recent related work.

2.1. Information Landscape

Information landscape visualisation employs a geographic map metaphor for visual analysis of relationships in massive data sets. Relatedness in the data through is conveyed by spatial proximity in the visualisation, i.e. items which are similar and therefore close in the high-dimensional vector space are placed close to each other in the low-dimensional visualisation space. Hills (or islands) represent group s (clusters) of related documents and emerge in areas where the document count (density) is large. Hills are separated by sparsely populated flat areas which are usually represented as plains (or see). The height of a hill usually represents the local density of data points, while the area covered by the hill is an indicator of the cohesion of the corresponding data item cluster. Each Visualised item is displayed as dot or a tiny icon. Regions of the landscape are labelled with highest weight features extracted from the underlying data. The colour and/or shape of each icon can be used to encode additional information belonging to the corresponding data item, such as metadata. Interactivity of the visual component, which is often implemented using 3D rendering, typically includes navigation (zooming, panning, rotating, tilting, etc.), selection and filtering, as well as manipulations of visual properties of the data items.

Information landscapes have been routinely used for visualisation of large document sets containing millions of documents [KBC*07], where the dimensionality of the high-dimensional term space easily surpasses 10000. In [DHJ*98] information landscape has been applied on gene expression data. Application to hierarchically organised document collections has been proposed in [AKS*02], where spatial tessellations are used to reflect hierarchically organised document sets. Hierarchically organised collections (classes) are represented through nested polygonal areas, containing data items at the lowest level of the hierarchy. Dynamically changing data sets have been addressed by information landscapes with dynamic topography, where changes in the data set are represented by smoothly animated changes of the landscape topography [SKM*09, SK09].

2.2. Dimensionality Reduction

Dimensionality reduction techniques aim at mapping high-dimensional data into lower-dimensional data. Depending on the application the dimensionality of the lower dimensional space may vary. E.g., for pattern recognition tasks, one keeps a large amount of dimensions, discarding only least relevant ones. For visualisation the target space is usually very low dimensional.

Force-Directed placement (FDP) [FR91] is a method inspired by physics where points are considered as particles attracting and repulsing each other by physical forces. FDP can be seen as an MDS if the forces are calculated from the high-dimensional distances. The drawback of the global methods is that they solely optimise towards one global values, thus not reflecting local properties of the high-dimensional space in the projection.

More recently, localised, non-linear approaches have been proposed, partly as derivatives of the linear methods. Kernel PCA [SSM98] uses a kernel to apply local transformation of the high-dimensional data. Localised MDS (LMDS) [CB09] aims at preserving local distances of the data by applying a localised stress function. IsoMap [Ten00] and Local Linear Embeddings (LLE) [RS00], are further examples of non-linear dimensionality reduction methods.

Our work has been motivated by the need to optimise an existing force-directed placement algorithm. This algorithm combines clustering, force-directed placement and spatial tessellations to generate information landscapes from very large document collections [SKM*09]. It has been applied in several research and industry projects. Consequentially, we evaluated the stress map approach by trying to identify known stress-related phenomena in this algorithm. Evaluation results are outlined in section 4.2.

2.3. Local and Global Stress Measures

In this section we compare local and global stress measures focusing on global and local versions of metric multidimensional scaling. Stress is a measure of lack-of-fit between high-dimensional dissimilarities and the distance in

the layout. For the global case we focus on stress as defined in metric MDS and for the local case we focus on stress as defined in local MDS (LMDS) [CB09].

The most elementary stress definition in metric MDS is the raw stress defined by [Kru64] as the residual of sum-squares of the high-dimensional distances d_{ij} and the geometric distances g_{ij}

$$S_G = \sum_{i,j} (d_{ij} - g_{ij})^2 \quad (1)$$

Later extensions to this formula include various weighting and normalizing parameters.

In contrary to the global optimization, LMDS [CB09] includes repulsive forces between points with large distances, resulting in the stress function:

$$S_L = \sum_{i,j \in N} (d_{ij} - g_{ij})^2 - t \sum_{(i,j) \notin N} g_{gij} \quad (2)$$

$$t = \frac{|N|}{\frac{N(N-1)}{2-|N|}} \cdot median_N(d_{ij}) \cdot \tau \quad (3)$$

with N being a symmetric set of nearby pairs (i, j): $(i, j) \in N$ if j is among the K nearest neighbours of i, or i is among the K nearest neighbours of j, and t being a fixed constant depending on a tuning parameter τ, also called repulsion parameter. The stress function S_L can be optimised for a fixed τ, i.e. a fixed t. The LDMS layout depends on the choice of this parameter.

To assess the local quality of a given layout Chen et al. [CB09] propose a LC-Meta criterion. The LC-Meta criterion measures the preservation of local structures in terms of overlap of set of nearest neighbours in the high-dimensional space and and set of nearest neighbours in the low-dimensional space. The parameter number of neighbours has to be set beforehand, values of 6 or 8 seem to be good choices [CB09]. The LC-Meta criterion is not smooth and can not be subjected to optimization, but can be used to select among various parameter configurations. The LC-meta criterion can be calculated point-wise and globally. The point-wise version can be used for evaluating stress on a local level. The global version gives an idea of the average local quality of the layout.

This localised stress measure does not differentiate between different forms of projection errors. In our experiments, we were interested in errors introduced by differences in both high-dimensional and low-dimensional distances. Therefore, we introduce an alternative local stress measure, as outlined in section 3.1.

2.4. Related Work

In early 2010, Schreck et al. [SvLB10] described a methodology for the visual assessment of projection precision which, in large parts, antedates our stress map approach (This paper was submitted in early 2010. We were not aware of the work of Schreck et al. and would like to thank the reviewers for pointing it out.). The visualisation and integration strategy is in fact very similar. However, the stress evaluation function we propose features a novel weighting term which differentiates between types of projection errors (compare section 3.1).

3. Our Approach

We create a stress map from a given layout based on a regular grid which covers the layout area. Each grid cell is first assigned the (normalised) stress value computed from the chosen stress function for the cell's position. The grid is then interpreted as a height map and cell values are used as interpolation support points to generate landscape geometry. A heat map is created from the grid by mapping cell values to a colour palette. The resulting stress map is composed by applying the heat map as a texture to the landscape geometry.

The stress map reflects the stress function values in both colour and height. However, the location of individual items is the same in the stress map and in the information landscape. Furthermore, the metaphor of the information landscape is fully retained in the stress map. It is therefore possible to switch between information landscape and stress map without loss of visual context. We expect this property of stress maps to ease interpretation of stress-related phenomena.

In our experiments, we employed a non-linear colour palette which represents low to high stress values as a smooth transition from blue to red and very high stress values as yellow (compare scale at lower right of figures 1(c) and 1(d)). The resulting pop-out effect enables the pre-attentive detection of regions having very high stress. Individual items were represented as coloured dots. The blue to red range of the described colour palette was mapped to the full range of stress values in assigning item colours. Therefore, items remained visible especially in high-stress regions. The colour coding of items facilitates assessment of stress on a single item level.

3.1. Adapted Local Stress Measure

The stress map visualisation is independent of the applied stress-measures. In general, two types of errors (stress) may occur when mapping high-dimensional data to a lower dimension. The first type of error is if two items with a large high-dimensional distance are placed nearby in the layout. We refer to this kind of error as $E_{l \to s}$ (l stands for large and s for small, the first index represents the high-dimensional distance). The second type of error is $E_{s \to l}$ occurring when items that are nearby in the high-dimensional space are mapped to locations with large distances in the layout. No error occurs in the other cases. Table 1 shows an overview over the error types.

		high-d distance	
		large	small
low-d distance	large	no error	$E_{s \to l}$
	small	$E_{l \to s}$	no error

Table 1: *Types of errors in projection algorithms*

To be able to identify both types of errors we propose the following stress function s to visualise local phenomena in an projection based layout.

$$s_{ij} = w^l_{ij} \cdot w^h_{ij} \cdot (d_{ij} - g_{ij})^2 \quad (4)$$

$$w^l_{ij} = (1 - g_{ij})^a \quad (5)$$

$$w^h_{ij} = (1 - d_{ij})^b \quad (6)$$

where w^h reflects the influence of the high-dimensional distance (the size of the neighbourhood in high-d) and w^l reflects the influence of the low-dimensional distance (to which extend nearby positioned items contribute to the stress). The exponents a and b define the size of the neighbourhood. The formula assumes normalized distances, i.e. $d_{ij} \in [0,1]$ and $g_{ij} \in [0,1]$.

The total stress of an item i is then defined by

$$s_i = \sum_j s_{ij} \quad (7)$$

Note that s_{ij} depends on w_{ij} which allow to reduce the items taken into consideration to a local neighbourhood (either high- or low-dimensional).

4. Experiments

In the following experiments we are interested in errors of type $E_{l \to d}$, i.e. item pairs with large distance in the high-dimensional space mapped nearby in the low-dimensional space. Therefore we set a we set $a = 20$ and $b = 0$ in equations 5 and 6. For our experiments we used the Reuters-21578 text collection.

4.1. Algorithm

For information landscape computation of a text document data set we employ an algorithm combining clustering, force-directed placement and spatial tessellations [SKM*09]. We first recursively apply a k-means clustering algorithm to create a hierarchy of topical clusters. A cluster split-and merge strategy attempts to determine the optimal amount of children at each hierarchy level and prevents the degeneration of the cluster hierarchy. The recursive, hierarchical projection algorithm starts with top level clusters and projects their centroids into a rectangular area using a force-directed placement (FDP) method. A polygonal area is assigned to each cluster by applying Voronoi area subdivision on the projected centroids. Sub-clusters are recursively projected in the same manner and inscribed within the areas of their parent clusters producing a hierarchy of nested polygonal areas. At the bottom of the hierarchy the documents (leafs) are projected within their parent-cluster's area using the same FDP method. Clusters (as well as sub-clusters on all hierarchy levels) are labelled with the highest frequency terms of the centroid vector providing orientation at any required level of detail. The described projection method is fast and scales with the time and space complexity of O(n*log(n)), n being the number of clustered documents: 10000 vectorised text abstracts can be processed in about 10 seconds on a 2.8 GHz Core i7 860 processor using 64bit Java VM (1.6.0_18), while over 300000 abstracts can be clustered and projected in slightly over five minutes using less than 6GB memory.

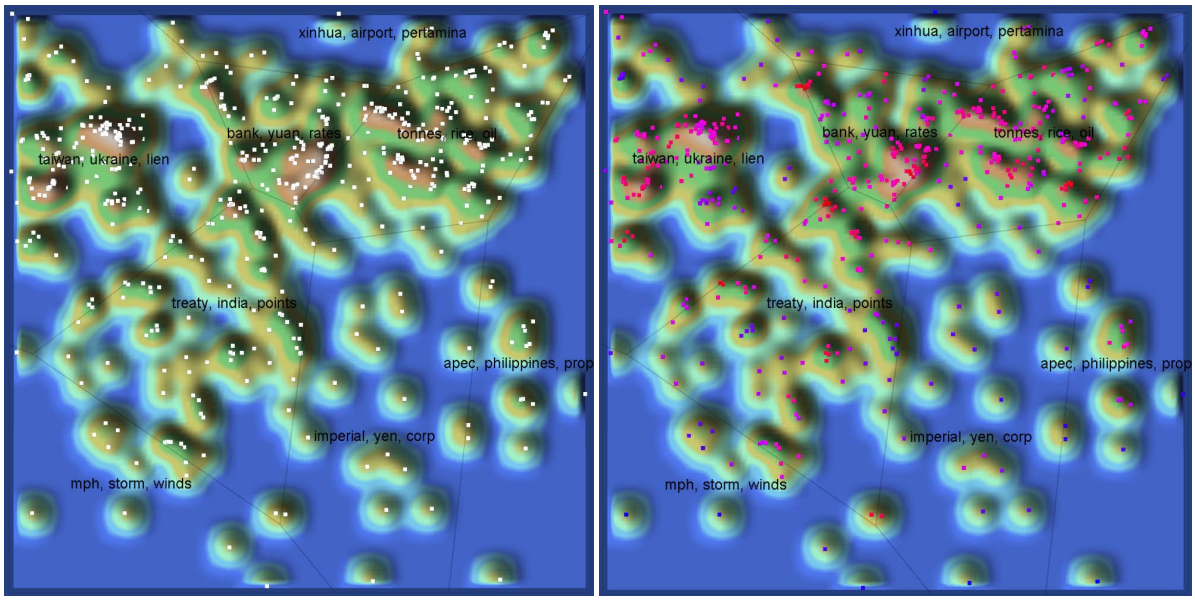
4.2. Results and Discussion

Figure 1 shows an example Information Landscape with 529 documents from the Reuters-21578 text data collection (the subset was generated by searching for "China"). Image 1(a) displays the standard landscape. In the following image 1(b) we assigned the stress values of each item to its colour (blue meaning low, red meaning high stress). The landscape texture remained unchanged so that hills appear where concentration of documents is large. In figure 1(c) we go a step further and also encode the stress value in the landscape. In the resulting heat-map hills correspond to regions of high-stress. Regions with low-stress remain flat and blue.

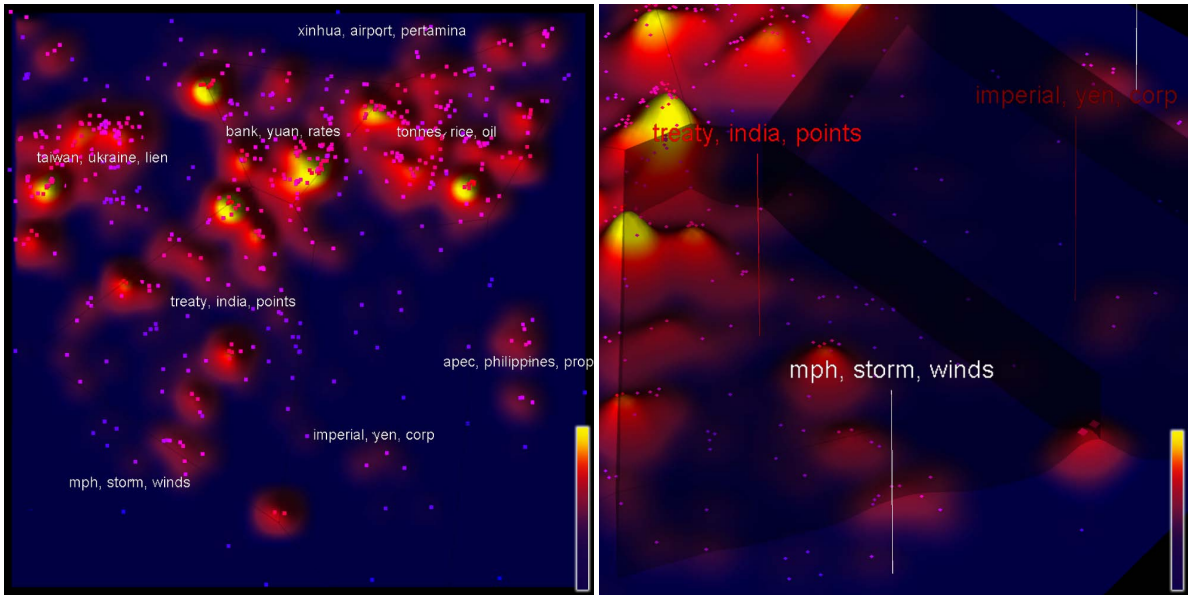
An exhaustive analysis of the correlation between visual phenomena and the computed stress properties is beyond the scope of this paper, and will be referenced in the future work section. However, we manually inspected results for a sample data set and a projection algorithm with known properties using the stress map approach. We were able to verify the occurrence of two expected phenomena.

Clusters containing a large number of documents tend to have high stress. An inspection of cluster cohesion (i.e. the inverse averaged inner cluster distance) suggests that relevant clusters feature comparably low cohesion in the high-dimensional space but comparably high cohesion in the visualisation space. We can attribute this effect to the nature of the area subdivision algorithm, which does not consider high-dimensional cohesion when assigning the amount of area to a cluster.

The force-directed projection algorithm threats items in neighbouring clusters independently. This leads to artefacts at the cluster boundaries in the visualisation. For example, the lowest peak in 1(c) is located at the boundary between the clusters "treaty, india, points" and "imperial, yen, corp". The two items have a large high-dimensional distance but are located very close to each other in the visualisation (as shown in figure 1(d)).



(a) standard landscape (b) standard landscape with stress-coloured items



(c) stress map, landscape encodes items stress (d) visual explanation of a local stress phenomenon

Figure 1: *Steps of stress visualisation: 1(a) standard landscape visualisation without stress indicators, 1(b) single items in the landscape are coloured corresponding to their stress value (blue - low, red - high), 1(c) stress map: the landscape using the stress value for single items to define heights (yellow - highest stress, red - high stress, blue - low stress), 1(d) example stress peak: two items at the boundary of two clusters, which were laid out independently, showing high stress*

5. Conclusion and Future Work

We have proposed stress maps as a visualisation methodology for detecting local stress phenomena in projection based layouts. A stress map is composed as a combination of a heat map and a height map expressing stress values. It can be seamlessly integrated with an information landscape created by the projection algorithm to be evaluated.

We have defined two types of errors ($E_{l\to s}$ and $E_{s\to l}$) that occur when mapping from high-dimension space to low-dimensional space. Depending on the application one error type might be of greater interest than the other. We therefore defined a stress function for the visualisation that allows detection of both types of errors by adjusting two parameters (which could even be exposed to users through interface elements). This feature also sets our approach apart from important, recent related work. A comparative evaluation of approaches is an obvious direction of future work.

We have investigated the results obtained by the proposed methodology using a projection algorithm and test data set with known properties. We found strong visual indicators for two expected stress-related phenomena. An exhaustive analysis of the correlation between visual phenomena and the computed stress properties is a natural next step.

The current version of the stress map approach displays area stress level and item stress level. However, it does not display the extend to which other items contribute to the stress of an specific item. This information is implicitly computed during the evaluation of the stress functions and could be visualised, for example by showing the directions of the large stress components as a vector field.

6. Acknowledgement

The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency (FFG).

References

[AKS*02] ANDREWS K., KIENREICH W., SABOL V., BECKER J., DROSCHL G., KAPPE F., GRANITZER M., AUER P., TOCHTERMANN K.: The InfoSky Visual Explorer: Exploiting hierarchical structure and document similarities. *Information Visualization 1*, 3–4 (Dec 2002), 166–181.

[CB09] CHEN L., BUJA A.: Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association 104*, 485 (2009), 209–219.

[DHJ*98] DAVIDSON G. S., HENDRICKSON B., JOHNSON D. K., MEYERS C. E., WYLIE B. N.: Knowledge mining with vxinsight: Discovery through interaction. *JOURNAL OF INTELLIGENT INFORMATION SYSTEMS 11* (1998), 259–285.

[FR91] FRUCHTERMAN T. M. J., REINGOLD E. M.: Graph drawing by force-directed placement. *Software - Practice and Experience 21*, 11 (November 1991), 1129–1164.

[GKWZ08] GORBAN A. N., KÈGL B., WUNSCH D. C., ZINOVYEV A. (Eds.): *Principal Manifolds for Data Visualization and Dimension Reduction*. Springer, 2008.

[KBC*07] KRISHNAN M., BOHN S., COWLEY W., CROW V., NIEPLOCHA J.: Scalable visual analytics of massive textual datasets. In *Parallel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International* (March 2007), pp. 1–10.

[Kru64] KRUSKAL J.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*, 1 (March 1964), 1–27.

[RS00] ROWEIS S. T., SAUL L. K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science 290*, 5500 (2000), 2323–2326.

[SK09] SABOL V., KIENREICH W.: Visualizing temporal changes in information landscapes. Poster and Demo at Eurovis 2009, Jun 2009.

[SKM*09] SABOL V., KIENREICH W., MUHR M., KLIEBER W., GRANITZER M.: Visual knowledge discovery in dynamic enterprise text repositories. In *IV '09: Proceedings of the 2009 13th International Conference Information Visualisation* (Washington, DC, USA, 2009), IEEE Computer Society, pp. 361–368.

[SSM98] SCHÖLKOPF B., SMOLA A., MÜLLER K.-R.: Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput. 10*, 5 (1998), 1299–1319.

[SvLB10] SCHRECK T., VON LANDESBERGER T., BREMM S.: Techniques for precision-based visual analysis of projected data. In *IS&T/SPIE Conference on Visualization and Data Analysis (VDA 2010)* (2010).

[Ten00] TENENBAUM J. B.: A global geometric framework for nonlinear dimensionality reduction. *Science 290*, 5500 (December 2000), 2319–2323.