DOI: 10.1111/cgf.12366
Eurographics Conference on Visualization (EuroVis) 2014
H. Carr, P. Rheingans, and H. Schumann
(Guest Editors)

Volume 33 (2014), Number 3

Distortion-Guided Structure-Driven Interactive Exploration of High-Dimensional Data

S. Liu[1], B. Wang[1], P.-T. Bremer[2] and V. Pascucci[1]

[1]Scientific Computing and Imaging Institute, University of Utah
[2]Lawrence Livermore National Laboratory

Abstract

Dimension reduction techniques are essential for feature selection and feature extraction of complex high-dimensional data. These techniques, which construct low-dimensional representations of data, are typically geometrically motivated, computationally efficient and approximately preserve certain structural properties of the data. However, they are often used as black box solutions in data exploration and their results can be difficult to interpret. To assess the quality of these results, quality measures, such as co-ranking [LV09], have been proposed to quantify structural distortions that occur between high-dimensional and low-dimensional data representations. Such measures could be evaluated and visualized point-wise to further highlight erroneous regions [MLGH13]. In this work, we provide an interactive visualization framework for exploring high-dimensional data via its two-dimensional embeddings obtained from dimension reduction, using a rich set of user interactions. We ask the following question: what new insights do we obtain regarding the structure of the data, with interactive manipulations of its embeddings in the visual space? We augment the two-dimensional embeddings with structural abstractions obtained from hierarchical clusterings, to help users navigate and manipulate subsets of the data. We use point-wise distortion measures to highlight interesting regions in the domain, and further to guide our selection of the appropriate level of clusterings that are aligned with the regions of interest. Under the static setting, point-wise distortions indicate the level of structural uncertainty within the embeddings. Under the dynamic setting, on-the-fly updates of point-wise distortions due to data movement and data deletion reflect structural relations among different parts of the data, which may lead to new and valuable insights.

1. Introduction

High-dimensional data arise naturally in many scientific applications and real-world phenomena. For instance, in a jet flame combustion simulation, half a million samples of chemical composition are extracted point-wise in space and time. These samples can be modeled as a high-dimensional point cloud where the chemical species involved in the simulation correspond to the dimensions of the data. In nuclear reactor safety analysis, complex simulator and controller codes are coupled to model system dynamics in the case of an accident scenario (e.g., a plane crashing into a power plant), where hundreds of deterministic and stochastic elements are encoded in the simulation. In order to consider the complete system dynamics, time evolution data of core temperature are collected under various uncertain conditions, and such temperature profiles can be modeled as a high-dimensional point cloud for transient analysis. In e-commerce, online browsing records and purchase transactions from millions of users are collected to predict consumer behavior and market trends. Such data are modeled as a point cloud in high dimensions where categorical attributes may represent the various dimensions.

Dimension reduction (DR) techniques, combined with analysis and visualization, are among the most common approaches to explore high-dimensional data. Under the general setting, DR techniques transform point cloud data in high dimensions into their low-dimensional representations (typically 2D and 3D for visual mapping), while striking a balance between structural preservation and computational efficiency. For example, principal component analysis (PCA) minimize the cost of structural transformation via projection, measured by sum of squared errors. Classic multidimensional scaling (cMDS) [Tor52] and Isomap [TDSL00] encode Euclidean (for cMDS) or geodesic (for Isomap) distance proximities among pairs of points through inner product matrices and minimizes their dissimilarities

between high-dimensional and low-dimensional spaces. Locally liner embedding (LLE) [RS00] and Laplacian Eigenmaps (LE) [BN03] both focus on preserving local neighborhood structures, by exploring linear or spectral properties in matrices that encode pair-wise relations.

However, DR techniques are typically used as black box solutions in data exploration and their results can be hard to interpret. Users typically face several challenges in applying DR in practice: (a) how to access the quality of results obtained by a DR technique; (b) how to choose among multiple DR techniques; (c) and for a fixed DR, how to choose its appropriate parameters. These challenges are partially addressed by the introduction of quality measures, such as rank-based criteria [LV09], to quantify the extent of structural preservation during the DR process. Global distortion measures could be adapted to access the quality of the embeddings, across different DR techniques, or among various parameter settings of a single DR technique. Their point-wise extensions are further computed and visualized to highlight erroneous regions of the data [MLGH13].

Ultimately, we are interested in the following question: How do we obtain insights regarding the structures of the data via explorations of its low-dimensional embeddings? We impose structural context onto the embeddings via point-wise quality measures and hierarchical clusterings. Point-wise distortion measures not only assess the fine-grained quality of the DR techniques but also highlight potential interesting regions of the data. Regions with high distortions across multiple metrics are worth further investigation as they correspond to regions with large structural uncertainty, which potentially reflect nontrivial structures of the data. On the other hand, low-dimensional point representations alone lack structural context as they are typically visualized as an unstructured point cloud; and point occlusion commonly occurs in practice. One would want to obtain a structural abstraction in high-dimensional space that summarizes the data in a certain way, e.g., via hierarchical clusterings, while exploring and exploiting such an abstraction via its embeddings.

Contributions. We provide an interactive visualization framework for exploring high-dimensional data via its low-dimensional embeddings. We ask the following question: what new insights do we obtain regarding the structure of the data, with interactive manipulations of its embeddings in the visual space? Our core contributions are:

- We augment two-dimensional embeddings with structural abstractions obtained from classical and topological hierarchical clusterings, to help users navigate and manipulate subsets of the data.
- We use point-wise distortion measures to highlight interesting regions in the domain, and further to guide our selection of the appropriate level of clusterings that are aligned with the regions of interest.
- Most importantly, our system allows users to move and delete subsets of the data in the visual space, where on-the-fly updates of point-wise distortion measures reflect structural relations among different parts of the data and potentially lead to new insights.

First, we give a systematic overview of global and point-wise distortion measures for several popular DR techniques. To the best of our knowledge, we introduce, for the first time, DR-dependent point-wise distortion measures derived from the cost of structural transformations. We review DR-independent distortion measures based on distance distortions and ranking discrepancies. In addition, we introduce two new distortion measures based on robust distance and kernel density estimate. Second, we focus on our motivation for distortion-guided, structure-driven data exploration. Third, we provide descriptions of design choices and implementation details. Finally, we showcase the utility of our framework through case studies involving real-world datasets. Our framework is highly modular and easily extensible to incorporate new DR techniques, distortion measures and interaction/visualization components, according to user demand, making it a robust tool for data exploration.

2. Related Work

Quality assessment for DR. Various quality assessments of DR have been proposed primarily in the machine learning community, for both labeled and unlabeled data. For labeled data, quality measures that focus on *classification error* (e.g., [SR03]) or group memberships [GZ10] seem to be obvious choices. For instance, *quality of group compactness* [GZ10] measures consistency among group memberships in a local neighborhood of a point-based on labeled information. For unlabeled data, some criteria for evaluation relate pair-wise distances through direct comparison between high- and low-dimensional space. For example, *quality of distance mapping* [GZ10] computes the correlation coefficient between the pair-wise distance matrices before and after DR. Measurements such as *strain* [Tor52] and *stress* [BSL*08] (described in Section 3) capture absolute differences between distance matrices. Other criteria do not directly compare lengths but rather ranks of pair-wise distances. Criteria such as *precision and recall* [VPN*10], *co-ranking* [LV09], *quality of point neighborhood preservation* [GZ10] and *agreement rate* [FC07] all focus on calculating the average number of neighbors that agree in high and low dimensions. Such rank-based criteria are typically scale-independent in the sense that they are invariant under linear transformations of distances. Specific measurements of geometrical and topological distortions due to manifold compression, stretching, gluing and tearing, have been proposed and visualized in [Aup07]. For a recent survey of quality measures, see [MLGH13]. As pointed out in [LV09], a simplistic way to assess the quality of DR is to look at the value of the objective function after optimization. We adapt this idea and derive both global and local distortion measures from a for-

malized objective. We also implement a scale-independent distortion measure based on co-ranking [LV09, MLGH13].

Interactive DR. A number of interactive DR systems have been introduced to fill the gap between visual perception of the data and in-depth understanding of its underlying structure. Besides relying on existing DR techniques as black box solutions, some recent methods use customized projections updated by user interactions for feature discovery. Interactive PCA (iPCA) [JZF*09] provides a rich set of interactions to help users better understand relationships between the data and the calculated Eigenspace. The system introduced in [JJ09] combines user-defined quality metrics to preserve important features during DR, and offers automatic ordering of variables to enhance perception of patterns selected by the user. The dimension projection matrix (tree) [YRWG13] demonstrates an interactive framework that allows users to hierarchically split both data points and dimensions, to do subspace visual exploration. In [BLBC12], a concept is coined V2PI (visual to parameter interaction), where the underlaying statistical model is updated when the user changes point positions in the projected view. Similar interactive projection designs can be found in [CLKP10, PEP*11, Gle13]. In our work, instead of focusing on a single DR technique such as iPCA, or data-dependent customized projections, we propose a general framework that is applicable to any existing DR techniques and is easily extendible with various distortion measures. Part of our interaction design resembles those found in iPCA, although core contributions of our work arise from coordinated interplay among DR, data manipulation, structural skeleton and on-the-fly update of distortion measures.

Structural abstractions. One approach for understanding high-dimensional data is to generate some form of structural abstraction. Topology-based hierarchical clustering [CGOS11, SMC07, LSL*13, CBL11, GBPW10a] have been proposed to construct useful combinatorial representations for the analysis and visualization of high-dimensional datasets. Such techniques could be integrated into interactive visual environments (e.g., Mapper [SMC07] and HD-Viz [GBPW10a]). Our framework is designed to employ various structural abstractions preprocessed or obtained from both classical (e.g., average-linkage) hierarchical clustering and topological methods based on Morse-Smale decompositions [GBPW10a]. Such structural summaries augmented with DR results help users better navigate and manipulate the points within the embeddings.

3. Point-wise Distortion Measures

Point-wise (local) distortion measures provide the foundations for our interactive method. In this section, we give a systematic overview of global and point-wise distortion measures for several popular DR techniques. The first type of distortion measures quantifies the cost on structural transformation from high-dimensional to low-dimensional spaces. It is derived from the particular objective function a given DR technique is formulated to optimize; thus it is DR-dependent, as described in Section 3.1. The second type of distortion measures is DR-independent and focuses on computing distance distortions, density differences or ranking discrepancies [MLGH13], applicable across DR techniques, as described in Section 3.2.

The basic setting for DR is as follows: given a set of n points $X = \{x_1, x_2, ..., x_n\}$ in $\mathbb{R}^l$, find a set of points $Y = \{y_1, ..., y_n\}$ in $\mathbb{R}^m$ where $m \ll l$, such that Y represents X by preserving certain structural properties of X. For the purpose of our visualization tool, $m = 2$, with possible extension to $m = 3$. For a given DR technique, a global distortion measure assigns a real-valued number to the pair (X, Y), which gives an overall, coarse quality assessment, whereas a point-wise distortion measure is a function that maps points in X to $\mathbb{R}$, which provides localized, fine quality assessment.

3.1. DR-Dependent Distortion Measures

Most DR techniques can be formulated as optimization problems formalized with objectives. For the popular DR techniques described below, optimizing the objectives is typically formulated as minimizing certain cost functions. A cost function incorporates a natural quality measure that assesses how much structure, in terms of relations among data points in high dimensions, stays consistent with the one inferred by the low-dimensional embedding; or alternatively, how much cost is needed in transforming one to another. Such a cost function gives rise to a natural global distortion measure $\mathcal{E}$ to assess the overall quality of the DR, and its point-wise derivation leads to a local distortion measure $\varepsilon : X \to \mathbb{R}$ that captures how much a point contributes to the global distortion and how well it agrees with its neighbors. We further enforce $\mathcal{E} = \sum_i \varepsilon(x_i)$.

Principle Component Analysis. PCA finds the directions of projection such that the squared distance of the points to these directions is minimized. Let $\mu : \mathbb{R}^l \to \mathbb{R}^l$ be a certain projection map. PCA seeks to minimize the global cost over μ, $\mathcal{E} = \sum_i ||x_i - \mu(x_i)||^2$, and the corresponding local cost ε is defined as, $\varepsilon(x_i) = ||x_i - \mu(x_i)||^2$.

The map μ is defined by the orthogonal direction with respect to a hyperplane defined by a collection of orthogonal basis $\{u_1, u_2, ..., u_m\}$ (where $u_i \cdot u_i = 1$ and $u_i \cdot u_j = 0$ for $i \neq j$). The projection $\hat{x}_i := \mu(x_i) \in \mathbb{R}^l$ of a given point $x_i \in X$ under μ could be written as $\hat{x}_i = \bar{x} + \sum_{j=1}^m z_j^i u_j$, where the mean $\bar{x} = \frac{1}{m}\sum_i x_i$, and $z_j^i = (x_i - \bar{x}) \cdot u_j$. Now the global cost can be written as $\mathcal{E} = \sum_i ||x_i - \hat{x}_i||^2$ and the local cost $\varepsilon(x_i) = ||x_i - \hat{x}_i||^2$.

Classic Multidimensional Scaling. MDS is commonly referred to as a class of techniques rather than a specific algorithm. cMDS [Tor52], also known as Principle Coordinate Analysis (PCoA) or Torgerson Scaling, is closely related to PCA. In cMDS, the distance is converted to inner production

dissimilarity and *strain* is optimized though an Eigenvalue decomposition.

Let b_{ij} be the inner product between a pair of points x_i, x_j in $\mathbb{R}^l$ and $\hat{b}_{ij}$ be the corresponding inner product in $\mathbb{R}^m$. That is, treating points as vectors, $b_{ij} = x_i \cdot x_j$ and $\hat{b}_{ij} = y_i \cdot y_j$. The relationship between distance matrix and inner product matrix can be defined as, $d_{ij}^2 = b_{ii} - 2b_{ij} + b_{jj}$, where d_{ij} corresponds to the Euclidean distance between x_i and x_j . We define the global cost to be equal to the strain, that is, $\mathcal{E} = \frac{\sum_{i,j}(b_{ij}-\hat{b}_{ij})^2}{\sum_{i,j} b_{ij}^2}$. The local cost corresponds to the point-wise strain, $\varepsilon(x_i) = \frac{\sum_j (b_{ij}-\hat{b}_{ij})^2}{\sum_{i,j} b_{ij}^2}$.

Laplacian Eigenmap. LE [BN03] seeks to minimize a global cost function, $\mathcal{E} = \sum_{i,j} ||y_i - y_j||^2 w_{ij}$, under appropriate constraints. The corresponding local cost is $\varepsilon(x_i) = \frac{1}{2}\sum_j ||y_i - y_j||^2 w_{ij}$. The algorithm proceeds by first constructing an adjacency graph on X based on either k-nearest neighbor (KNN) graph or ε-neighborhood. If x_i and x_j are connected by an edge, the weight w_{ij} is either defined as a heat kernel, that is, $w_{ij} = \exp(-||x_i - x_j||^2/t)$ (with diffusion parameter t), or simply defined as $w_{ij} = 1$; otherwise $w_{ij} = 0$.

Isomap. Isomap [TDSL00] is a nonlinear DR technique based on cMDS. In Isomap, the distance between pairs of points is geodesic distances approximated by the shortest paths between pairs of points in a neighborhood graph. Therefore the cost function is the same as cMDS except the Euclidean distance matrix is replaced by an approximated geodesic distance matrix.

Locally Linear Embedding. LLE [RS00] represents each point (in $\mathbb{R}^l$) as a weighted linear combination of its neighbors and tries to preserve this linear relationship in the reduced dimension $\mathbb{R}^m$. It optimizes the following global cost, $\mathcal{E} = \sum_i ||y_i - \sum_j W_{ij} y_j||^2$, where W_{ij} is the weight matrix that stores such a linear relationship. The local cost can be written as, $\varepsilon(y_i) = ||y_i - \sum_j W_{ij} y_j||^2$.

3.2. DR-Independent Distortion Measures

DR-independent criteria, on the other hand, can be applicable to a collection of DR techniques, and are inspired by measurements of distance distortions, density differences or ranking discrepancies. Some nonlinear DR techniques, such as LE, use constraints in their algorithms to remove an arbitrary scaling factor in the embedding. Points in the reduced dimension are therefore computed under a fixed scale, which means that ranges of values in $\mathbb{R}^l$ and $\mathbb{R}^m$ differ drastically, rendering the scale-dependent distortion measures such as local stress, robust distance distortion and kernel density estimate distortion meaningless. To address this issue, we use two types of scaling factors. The first one computes the ratio between the radiuses of minimum enclosing balls [GÖ99] of the data in $\mathbb{R}^l$ and $\mathbb{R}^m$ to rescale the embedding. The second type, which is also less sensitive to outliers, computes the ratio of average distances to the centroid.

Kernel Density Estimate distortion. We introduce a novel class of distortion measures based on a kernel density estimate (KDE). Each of these measures (based on a chosen kernel) quantifies differences in densities among local neighborhoods. In addition, a multiscale version of the measure is easily attainable by varying the parameters associated with a given kernel; thus it allows adaptive data explorations. A kernel is a non-negative similarity measure $K : \mathbb{R}^l \times \mathbb{R}^l \to \mathbb{R}^+$ where more similar points have higher value. We consider a Gaussian kernel here, where $K(p,x) = \exp(-||p-x||^2/2\sigma^2)$. A KDE is a way to estimate a continuous distribution function over $\mathbb{R}^l$ for a finite point set $P \subset \mathbb{R}^l$. Specifically, $KDE_P(x) = \frac{1}{|P|}\sum_{p\in P} K(p,x)$. The distortion function measures differences between KDE in $\mathbb{R}^l$ and KDE in $\mathbb{R}^m$. That is, the global KDE distortion, $\mathcal{K} = \sum_i |KDE_X(x_i) - KDE_Y(y_i)|$. and the local KDE distortion, $k(x_i) = |KDE_X(x_i) - KDE_Y(y_i)|$.

Stress. This distortion measure is based upon an objective function used in a distance scaling version of MDS, referred to as *stress*. We use the stress to measure distance distortions. Let d_{ij} be the distance between a pair of points i, j in $\mathbb{R}^l$ and $\hat{d}_{ij}$ be the corresponding distance in $\mathbb{R}^m$. Global stress is defined as, $\mathcal{S} = \frac{\sum_{i,j}(d_{ij}-\hat{d}_{ij})^2}{\sum_{i,j} d_{ij}^2}$. Local stress is, $s(x_i) = \frac{1}{2}\cdot \frac{\sum_j (d_{ij}-\hat{d}_{ij})^2}{\sum_{i,j} d_{ij}^2}$.

Robust distance distortion. We also introduce a distortion measure inspired by robust MDS (rMDS) [APV10,CD06]. It shares similarities with stress but is proved to be more robust with respect to noise and outliers. The global robust distance distortion is defined as, $\mathcal{R} = \frac{\sum_{i,j}|d_{ij}-\hat{d}_{ij}|}{\sum_{i,j}|d_{ij}|}$. The local robust distance distortion is, $r(x_i) = \frac{\sum_j |d_{ij}-\hat{d}_{ij}|}{\sum_{i,j}|d_{ij}|}$.

Co-ranking distortion. For completeness, we include in our system a rank-based, scale-independent criterion derived from co-ranking matrices [LV09, MLGH13]. Let d_{ij} be the distance between a pair of points x_i, x_j in $\mathbb{R}^l$ and $\hat{d}_{ij}$ be the corresponding distance between y_i, y_j in $\mathbb{R}^m$. The *rank* of x_j with respect to x_i is $\rho_{ij} = |\{k \mid d_{ik} \le d_{ij} \quad or \quad (d_{ik} = d_{ij} \quad and \quad 1 \le k < j \le N)\}|$. Similarly, the rank of y_j with respect to y_i is $\gamma_{ij} = |\{k \mid \hat{d}_{ik} \le \hat{d}_{ij} \quad or \quad (\hat{d}_{ik} = \hat{d}_{ij} \quad and \quad 1 \le k < j \le N)\}|$, where $|\cdot|$ denotes set cardinality. The difference $R_{ij} = r_{ij} - \rho_{ij}$ is considered *rank errors*. The co-ranking matrix C is defined by $C_{kl} = |\{(i,j) \mid \rho_{ij} = k \quad and \quad \gamma_{ij} = l\}|$. A DR with no errors would produce a diagonal co-ranking matrix.

In [LV09], a quality for dimension reduction is proposed as a sum of partial entries in the co-ranking matrix, $Q = \frac{1}{Kn}\sum_{k=1}^{K}\sum_{l=1}^{K} C_{kl}$, where K corresponds to the number of neighbors under consideration. Therefore every co-ranking ma-

trix C can be decomposed into a per-point permutation matrix C^i for every point x_i, with $C = \sum_{i=1}^{N} C^i$ and $C^i_{kl} = |\{j \mid \rho_{ij} = k \quad and \quad r_{ij} = l\}|$. The point-wise contributions is $Q_i = \frac{1}{K} \sum_{k=1}^{K} \sum_{l=1}^{K} C^i_{kl}$, where $Q = (\sum_{i=1}^{N} Q_i)/N$. For a given point, a larger Q_i corresponds to less local distortion. Therefore, we define global co-ranking distortion as $\mathcal{Q} = -Q$ and local co-ranking distortion as $q = -Q_i$.

4. Motivation and Data Exploration Pipeline

We highlight our motivation for distortion-guided, structure-driven data exploration. We discuss the rationale for (a) transitioning from a static setting to a dynamic setting in exploring the data via point-wise distortion measures, and (b) using hierarchical clusterings for data abstractions. We then describe a typical interactive data exploration pipeline, as illustrated in Figure 1.

Distortion measures under the dynamic setting. Visualizing point-wise distortions under the static setting illustrates the qualitative disparities among different regions of the embedding, which in turn, reflects structural discrepancies within the original data. Regions with higher distortions correspond to areas with more structural uncertainty (and equivalently, less structural preservations) in their embeddings. We ask the following questions: (a) Why do certain areas of the original data have higher point-wise distortions? (b) Are such distortions due to the structures of the original data that are hidden in its embedding? (c) Is it possible for us to manipulate the locations of some points in the embedding in order to achieve better point-wise distortions locally, and what would such a manipulation tell us about the original data? These questions motivate us to compute and visualize distortion measures under the dynamic setting, where on-the-fly updates of point-wise distortions due to data movement and data deletion reflect structural relations among different parts of the data. Such data manipulations in the visual space do not trigger a new DR optimization process but result in updates of relevant distortion measures, which offer valuable feedback as to how much the manipulated results deviate from the original embedding. By moving subsets of points, an increase (or decrease) in distortion measures indicates structural dependencies (or independencies respectively) among different parts of the data, which may lead to new and valuable insights.

Structure-driven manipulation. Meaningful data manipulations (e.g., data movement and data deletion) in the visual space should be structure-driven, that is, the selected points should respect certain structures of the original high-dimensional data. We impose structural context onto the embeddings via hierarchical clusterings, which serve as structural abstractions of the data at multiple scales. Our framework currently allows users to choose from two classes of built-in clustering methods: classical (e.g., single- or average-linkage) hierarchical clustering [Def77] and topological hierarchical clustering based on Morse-Smale complexes [GBPW10b]. In addition, the users can also directly import existing hierarchical clustering results or class labeling of the data using a simple file format. Such clusterings help users navigate and manipulate subsets of the data at an appropriate level of abstraction.

Data exploration pipeline. We illustrate a typical interactive exploration pipeline in Figure 1. (a) We apply a certain DR technique to the high-dimensional dataset and obtain its initial embedding, where global distortion measures such as co-ranking could be employed to select a suitable DR and its optimal parameter setting. (b) We visualize point-wise distortions on the embedding. Regions with high distortions across multiple measures (for example) are identified as regions of interest for further investigation. (c) We apply hierarchical clustering of the data. (d) We use point-wise distortions to guide our clustering selection, where the appropriate level of clustering is chosen based on its agreement with the region of interest. (e) We allow users to move and/or delete a subset of data that belongs to a targeted cluster in the visual space, where on-the-fly updates of point-wise distortion measures reflect structural relations among different parts of the data. A decrease/increase in distortion measure of the targeted cluster typically indicates structural independencies/dependencies among the target and its neighboring clusters. (f) In addition, with detailed parameter analysis across each cluster, we obtain further insights regarding differentiating factors among different regions of the data. Finally, we obtain a collection of structural insights.

5. Design and Implementation

In this section, we describe components of the user interface, user interaction design and system implementation.

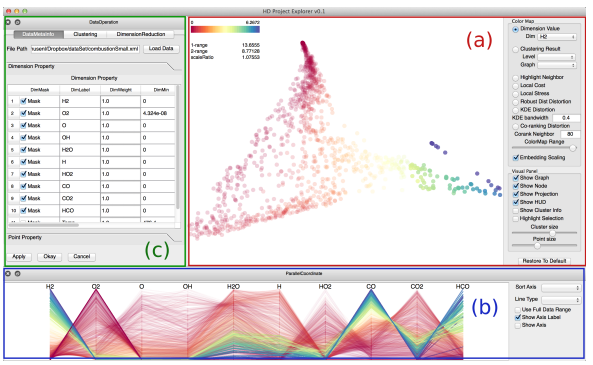

Figure 2: *A system overview showing two views and one control panel. (a) Embedding view. (b) Parallel coordinates view. (c) Data panel.*

5.1. Interface Design

A system overview is shown in Figure 2. The overall interface consists of two views and one data operation panel.

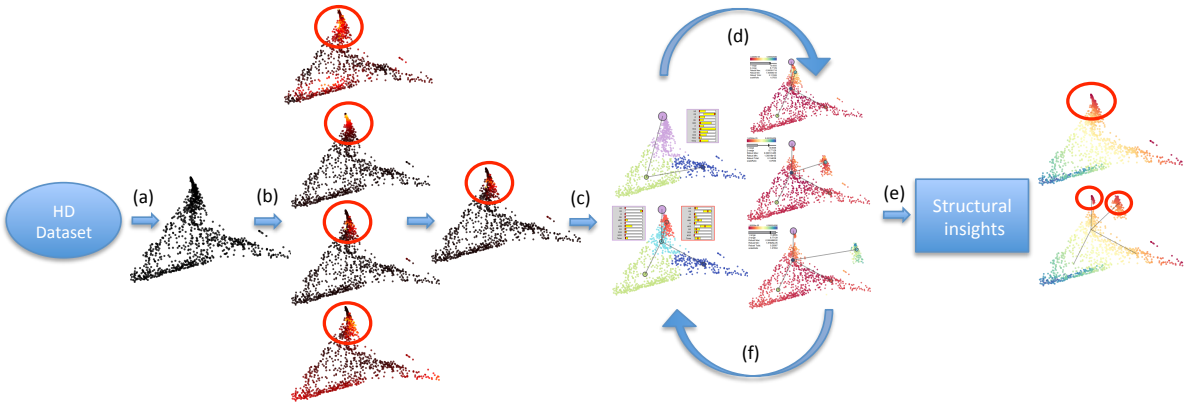

Figure 1: *A typical interactive data exploration pipeline. We could apply different DR for an additional round of analysis, as well as different distortions inside each analysis cycle. (a) DR; (b) Distortion-guided selection of region of interest; (c)-(d) Hierarchical clustering of the data and distortion-guided clustering selection. (e) Data manipulations with on-the-fly update of distortion measures reveal structural insights of the data. (f) Parameter differentiations across different clusters for additional structural insights.*

These visual components are coordinated to provide a comprehensive view of the data by highlighting its various aspects. They are interconnected such that selections and changes made in one component will be reflected in others. The system is highly modular and is easily extendable to include additional visual components.

Embedding view. This view is the main canvas of the interface where the results of DR, points embedded in 2D, are visualized. It contains a rich set of user interactions for data exploration. One could apply different colormaps to visualize points by values of a particular dimension, clustering labels or point-wise distortion measures.

Parallel coordinate view. This view displays the original data with each of its dimensions as a vertical axis and each point as a line drawing through each of the axes. A normalization of the range for each axis is optional to increase readability of the data.

Data panel. This panel contains various data operations such as DR and clustering. The panel is part of the interlinked system so that changes made to the dataset are instantly reflected through other views. The panel consists of three sub-panels. The meta-information panel gives a direct view of the data, in terms of its dimensions and statistics, and includes the ability to filter (hide) certain dimensions for analysis; the clustering panel allows the user to select distance metrics, data standardization schemes (see supplemental material) and hierarchical (e.g., classical single-, average-linkage, topology-based) clustering methods, while also allowing loading of existing clustering; and the DR panel enables the user to choose DR techniques and specify their parameters in an online fashion.

5.2. Interaction Design

The fundamental principle behind our interaction design is to obtain fresh insights regarding the structure of the data via distortion-guided, structure-driven, interactive manipulations. We provide a list of interaction semantics in the embedding view to aid our manipulations and explorations.

View interactions. Interactions in this category do not cause re-calculation of distortion measures. Typical operations include, point selection through the Lasso tool or cluster-level selection; view zooming and panning; filtering of data points; and selection highlighting. We provide some details regarding the structure-driven cluster-level operations. In the embedding view, a solid circle represents each cluster center, whose radius scales with the size of the cluster. *Cluster selection* allows the user to select points in a cluster in the view through selection of the cluster center. *Cluster expansion* enables the user to expand a selected cluster on-the-fly to reveal its child clusters. *Cluster compression* merges selected child clusters into their shared parent cluster. A neighborhood graph could also be constructed connecting cluster centers based on their distance proximities, which functions as a structural skeleton.

Data interactions. To visually assist the user to obtain new insights, we introduce a set of data manipulations that cause re-computation of distortion measures, namely, data movement and data deletion. *Data movement* changes the location of selected points via mouse movement. Upon releasing the mouse, both global and point-wise distortion measures are re-calculated and visualized. The increase or decrease of global distortion measure informs the user of the amount of global structural change, while on-the-fly updates of point-wise distortion measures provide valuable information to users regarding structural relations among different parts of the data. *Data deletion* allows users to remove points from the dataset and re-run DR and clustering. Data deletion can remove outliers affecting the DR quality, points with high/low distortions, or hidden/occluded clusters and allow focused analysis of subsets of the data.

5.3. System Implementation

We would like to provide an easily extensible framework that allows additions of new DR techniques, distortion measures and interaction/visualization components by following the standard Model-View-Controller paradigm. For DR, we use an open source C++ library named Tapkee [LWG13]. This template-based, easily extensible library provides more than a dozen commonly known DR techniques. We modify this library to incorporate point-wise distortion calculations so they fit seamlessly in our modular design. We chose Cluster3.0 library for the hierarchical clustering. Cluster3.0 is implemented in ANSI C and provides fast routines to calculate hierarchical clustering with different distance metrics. Qt is used for general GUI design and drawing functionalities in views. In addition, we provide topological hierarchical clustering based on approximated Morse-Smale segmentation [GBPW10b]. Both clustering and DR modules are based on APIs that are oblivious to the underlying implementation, and as a result the library implementations could be easily updated or replaced. For interactive applications, responsiveness is essential to the usability of the tool; therefore, we have recorded the detailed interaction performance information in the supplementary material.

6. Results

We showcase the utility and effectiveness of our framework through case studies involving real-world datasets from combustion and nuclear simulations, see the supplementary video for interactive details.

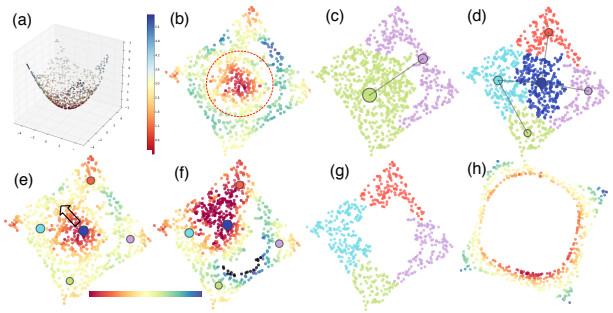

Figure 3: *Parabola. (a) 3D embedding colored by z-coordinate. (b) 2D embedding colored by KDE distortion. (b)-(d) Distortion-guided clustering selection. On-the-fly update of distortion measures for data movement (e)-(f), and data deletion (g)-(h). Distortion measures adopt spectral colormap.*

Synthetic Dataset: Parabola. We first demonstrate, via a synthetic dataset, distortion-guided clustering selection, data movement and data deletion in combination with an on-the-fly update of point-wise distortion measures. We use a parabola dataset as a proof-of-concept example, which contains trivial structural information that is easily interpretable in the embedding view. We follow our pipeline illustrated in Figure 1. Step (a)-(c): We apply PCA to the data and obtain a 2D embedding colored by KDE distortions (Figure 3(b)). Both KDE distortion and local cost (not shown here) identify a central region of interest (enclosed by the red circle) with low distortion. Step (d): We use point-wise distortion to guide our clustering selection where we arrive at a configuration with five clusters after cluster expansions (Figure 3(b)-(d)). Step (e): We allow the user to move points that belong to the blue (central) cluster and update the distortion on-the-fly (Figure 3(e)-(f)). A drastic increase in distortion along its boundary indicates a structural dependency among the blue cluster and its neighbors. Finally, through deletion of the blue cluster (Figure 3(g)-(h)), we could re-apply DR on the remaining points for focused structural analysis.

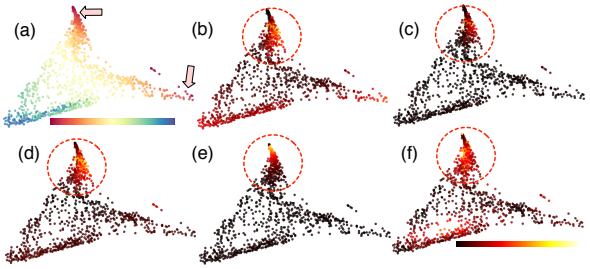

Figure 4: *Combustion. (a) Points colored by temperature. (b)-(f) All five distortion measures (local cost, local stress, robust distance distortion, KDE distortion and co-rank distortion) indicate an interesting region with high distortion around a temperature minima. Temperature image uses spectral colormap and distortion measure images adapt hot colormap.*

Combustion Simulation. This dataset consists of 2.8K samples of chemical composition and temperature extracted point-wise from time-varying jet simulations of turbulent CO/H_2-air flames [HSPC06]. The simulation records 10 chemical compounds: H_2, O_2 (Oxygen gas / Oxidizer), O (Oxygen), OH (Hydroxide), H_2O (Water), H (Hydrogen), HO_2, CO (Carbon monoxide), CO_2 (Carbon dioxide) and HCO. The dataset can be modeled as a 10D point cloud with temperatures as observations. The domain scientists are interested in understanding conditions that trigger extinction and re-ignition phenomena, which correspond to points (parameter settings) with minimal temperatures.

Our interactive data exploration process follows a typical pipeline illustrated in Figure 1. Step (a): We apply cMDS to the dataset, and color the points by temperature. The result is shown in Figure 4(a), where two areas are visible with minimal temperatures (marked by arrows), which may correspond to extinction scenarios. Step (b): In order to better understand the DR result and identify the area of interest for further analysis, we visualize various point-wise distortion measures (Figure 4(b)-(f)). All five of our distortion measures indicate that relatively large distortion exists among points near one of the temperature minima (top area enclosed by the red circle). Such a region becomes our primary target for further investigation. Steps (c)-(d) We apply classi-

cal (average-linked) hierarchical clustering to the data. As illustrated in Figure 5(a)-(b), we use point-wise distortions to guide our clustering selection, where the appropriate level of clustering is chosen based on its agreement with the region of interest. Through cluster expansion, we arrive at a resolution with five clusters (Figure 5(b)), where the red cluster (pointed by red arrow) agrees well with the region of interest (area enclosed by the red circle in Figure 4(b)). Steps (e): We allow the user to move a subset of the data that belongs to the red cluster away from its neighboring clusters, as illustrated in Figure 5(c)-(e). We observe a drastic decrease of point-wise distortion in the area of interest under moderate movement (Figure 5(d)). This indicates a certain level of structural independencies between the red cluster and its neighborhood points. Therefore, the points in the red cluster may potentially correspond to a distinct extinction phenomenon that is different from its nearby cluster. However, further data movement substantially increases the distortion measure (Figure 5(e)), which indicates that the red cluster is not completely separated from the rest of the data. Step (f): To further investigate the nearby red and purple clusters that both contain points with local minimal temperatures, we display summary statistics of parameters associated with each cluster in Figure 5(g) (where the red and yellow bars correspond to the mean values and the data range of the labeled parameters). Such summary statistics indicate that the differentiating factor between those two clusters is the vastly different HO_2 concentration (marked by pink arrows). In addition, our tool provides alternative topological hierarchical clustering results to further validate the separation of these local minima, as illustrated in Figure 5(f) where the blue cluster (pointed by blue arrow) is a topologically different region (based on the Morse-Smale segmentation) with respect to its neighbors, see [GBPW10a] for details. Finally, it turns out that the red cluster in Figure 5(b) represents an independent temperature local minima that correspond to parameter configurations of a special extinction condition (previously unknown to domain scientists as described in [GBPW10a]), where the mixing of fuel and oxidizer is highly turbulent and blows the flame out, resulting in a large amount of HO_2.

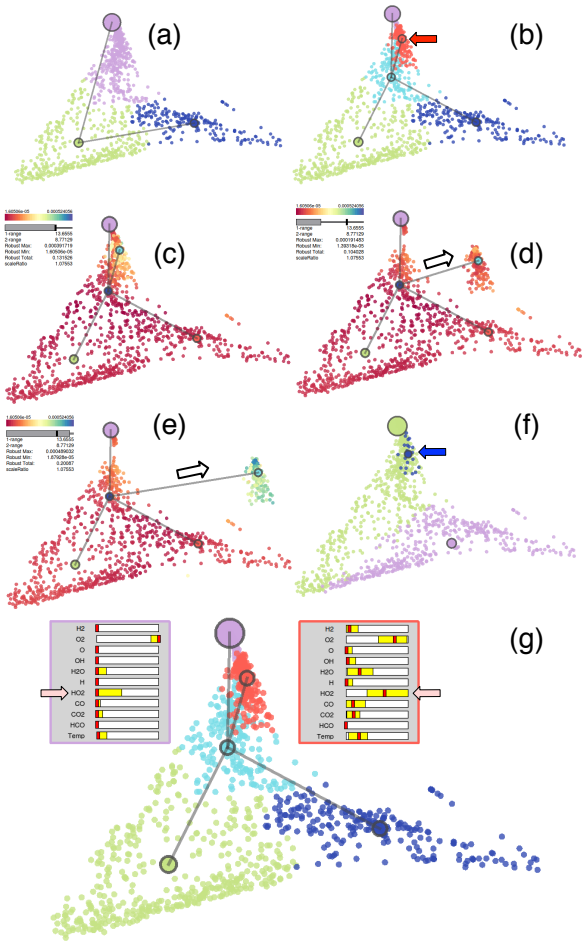

Figure 5: *Combustion. (a)-(b) Distortion-guided cluster selection. (c)-(e) On-the-fly updates of point-wise distortion measure (local stress) reflect structural relations between different parts of the data. (f) Validation of two overlapped temperature minima based on topological clustering. Distortion is colored by spectral colormap. The parameter boxes in (g) contain summary statistics of parameters in the clusters.*

Nuclear Reactor Safety Analysis. This dataset simulates an accident scenario when a plane crashes into a sodium-cooled fast reactor power plant and destroys three of the four cooling towers [MYA*13], and, thus, the reactor core cooling capabilities are disabled. A recovery crew then arrives at the site and attempts to re-establish the cooling of the reactor by restoring the damaged towers one by one, during which time the core temperature keeps increasing if the cooling system is disabled. When the reactor reaches a maximum temperature of 1000K the simulation is considered a system failure scenario; otherwise it is a system success. A set of stochastic parameters, such as crew arrival time and tower recovery time, influence how the core temperature changes over time. An ensemble of 609 transient simulations has been generated, each consisting of a time-varying core temperature profile corresponding to a single simulation. We sample each profile at 100 time steps and map it to a 100D space. The domain scientists are interested in studying the structure of this dataset and understanding characteristics associated with system failures and system successes, for nuclear reactor safety analysis.

Once again, we following the data exploration pipeline illustrated in Figure 1. Step (a): We apply cMDS to obtain a 2D embedding. Step (b): Both local stress and robust distance distortion visualizations (Figure 6(a)-(b)) identify an interesting region in the lower part of the embedding (enclosed by the red circle) with relatively high distortions. Step (c)-(d): We apply classical hierarchical clustering on the data. Through cluster expansion and compression (Figure 6(c)), we obtain a hierarchical clustering with four clus-

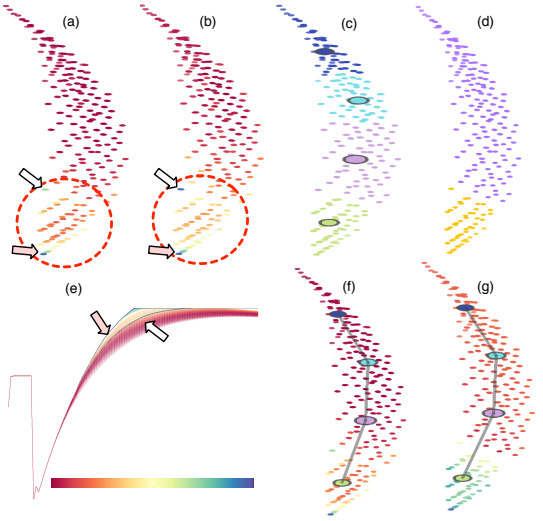

Figure 6: *Nuclear. (a) Local stress; (b) Robust distance distortion; (c) Distortion-guided cluster selection; (d) Points colored by their labels: system failure (yellow) and system success (purple); (e) Plot of 609 time-varying core temperature profiles in the parallel coordinate plots where x-axis is time, y-axis is temperature. (f)-(g) On-the-fly update of local stress before (f) and after (g) movement of points belonging to the bottom cluster. The embedding views are re-scaled in the paper due to space constraints.*

ters where the green cluster agrees almost perfectly with the region of interest. Step (e): We allow the user to move the points associated with the green cluster away from its neighbors in the visual space, and a small movement increases the distortion measure drastically (Figure 6(f)-(g), distortions before and after data movement). This change of distortion indicates that the green cluster is structurally dependent on the rest of the data. Step (f): Now we visualize the embedding with known labels of the data, as illustrated in Figure 6(d), where points are colored by their labels of success (purple) or failure (yellow). We observe that the green cluster in Figure 6(c) agrees almost perfectly with the the yellow cluster (failure cases) in Figure 6(d). This offers validation that our distortion-guided clustering selection captures some inherent structure of the data. By further investigating the local stress and robust distance distortion (Figure 6(a)-(b)), we notice there are two points with the highest distortions. These points are marked by arrows in Figure 6(a), (b) and (e), where Figure 6(e) illustrates all the time-varying core temperature profiles in the parallel coordinate plot. The point marked by white arrow corresponds to a boundary scenario that separates system failures from system successes, and the other marked by pink arrow corresponds to a limiting scenario that reaches failure temperature at the earliest simulation time. These distortion-guided observations again offer valuable information of the data. Furthermore, we could investigate just the system success scenarios by removing all the failure cases. As shown in Figure 7(a), we delete all the failure cases and re-apply cMDS. Through local distortion visualizations (Figure 7(b)-(c)), we can identify a point with high distortion that corresponds to a boundary scenario among the success cases (Figure 7(d)).

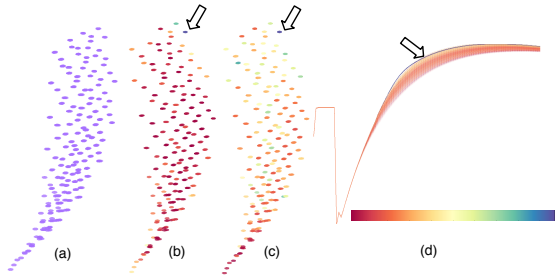

Figure 7: *Nuclear. (a) Interactive deletion of failure cases; (b)-(c) re-apply DR and visualize by local cost (b) and KDE distortion (c). Both visualizations reveal a point (indicate by white arrow) with high distortion that corresponds to a boundary scenario for the success cases. (d) Success scenarios in parallel coordinate plots. Embedding views are rescaled in the paper due to space constraints.*

7. Conclusions and Future Work

We propose a distortion-guided and structure-driven interactive framework for high-dimensional data exploration via its visual embeddings, such that: (a) The structural abstractions obtained through hierarchical clusterings allow multi-scale data manipulations, even with hidden or occluded data points; (b) Point-wise distortion measures are used to guide the cluster expansion and compression process to select the appropriate level of clustering and help users explore meaningful subregions of the data; (c) Combining interactive data manipulations in the embedding view with on-the-fly updates of distortion measures provides new insights regarding structural relations among different parts of the data. We rely on the clustering algorithms to provide approximated structural representations of the data for our interactive process, therefore the accuracy of our inferred results depend on the inherent characteristics of any chosen clustering method. Currently several clustering and DR algorithms used in our tool have a time complexity of $O(n^2)$. Therefore, main challenges for future research include system scalability (e.g. implementations of scalable PCA [GP14, Lib13]), and distortion approximations with respect to large datasets with millions of points.

Acknowledgments

We thank James C. Sutherland and Diego Mandelli for the combustion and nuclear dataset respectively. This work was performed in part under the auspices of the US DOE by LLNL under contract DE-AC52-07NA27344. This work is also supported in part by the NSF, DOE, NNSA, SDAV SciDAC Institute and PISTON, award numbers NSF 0904631, DE-EE0004449, DE-NA0002375, DE-SC0007446, and DE-SC0010498, respectively.

References

[APV10] AGARWAL A., PHILLIPS J. M., VENKATASUBRAMANIAN S.: Universal multi-dimensional scaling. *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (2010), 1149–1158. 4

[Aup07] AUPETIT M.: Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing 70* (2007), 1304–1330. 2

[BLBC12] BROWN E. T., LIU J., BRODLEY C. E., CHANG R.: Dis-function: Learning distance functions interactively. *Proc. IEEE Conf. on Visual Analytics Science and Technology* (2012), 83–92. 3

[BN03] BELKIN M., NIYOGI P.: Laplacian Eigenmaps for dimensionality reduction and data representation. *Neural Computation 15*, 6 (2003), 1373–1396. 2, 4

[BSL*08] BUJA A., SWAYNE D. F., LITTMAN M. L., DEAN N., HOFMANN H., CHEN L.: Data visualization with multidimensional scaling. *J. Comp. Graph. Stat. 17*, 2 (2008), 444–472. 2

[CBL11] CORREA C., BREMER P.-T., LINDSTROM P.: Topological spines: A structure-preserving visual representation of scalar fields. *IEEE Trans. Vis. Comput. Graphics 17*, 12 (2011), 1842–1851. 3

[CD06] CAYTON L., DASGUPTA S.: Robust Euclidean embedding. *Proc. Int. Conf. on Machine learning* (2006), 169–176. 4

[CGOS11] CHAZAL F., GUIBAS L. J., OUDOT S. Y., SKRABA P.: Persistence-based clustering in riemannian manifolds. *Proc. ACM Symp. on Computational Geometry* (2011), 97–106. 3

[CLKP10] CHOO J., LEE H., KIHM J., PARK H.: iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction. *Proc. IEEE Conf. on Visual Analytics Science and Technology* (2010), 27–34. 3

[Def77] DEFAYS D.: An efficient algorithm for a complete link method. *Comput. J. 20*, 4 (1977), 364–366. 5

[FC07] FRANCE S., CARROLL D.: Development of an agreement metric based upon the rand index for the evaluation of dimensionality reduction techniques, with applications to mapping customer data. *Machine Learning and Data Mining in Pattern Recognition, Lect. Notes Comput. Sc. 4571* (2007), 499–517. 2

[GÖ99] GÄRTNER B.: Fast and robust smallest enclosing balls. *Algorithms-ESA'99, Lect. Notes Comput. Sc.* (1999), 325–338. 4

[GBPW10a] GERBER S., BREMER P.-T., PASCUCCI V., WHITAKER R.: Visual exploration of high dimensional scalar functions. *IEEE Trans. Vis. Comput. Graphics 16*, 6 (2010), 1271–1280. 3, 8

[GBPW10b] GERBER S., BREMER P.-T., PASCUCCI V., WHITAKER R. T.: Visual exploration of high dimensional scalar functions. *IEEE Trans. Vis. Comput. Graphics 16*, 6 (2010), 1271–1280. 5, 7

[Gle13] GLEICHER M.: Explainers: Expert explorations with crafted projections. *IEEE Trans. Vis. Comput. Graphics 19*, 12 (2013), 2042–2051. 3

[GP14] GHASHAMI M., PHILLIPS J. M.: Relative errors for deterministic low-rank matrix approximations. *Proc. ACM-SIAM Symp. on Discrete Algorithms* (2014). 9

[GZ10] GORBAN A., ZINOVYEV A.: Principal manifolds and graphs in practice: from molecular biology to dynamical systems. *Int. J. Neural. Syst. 20*, 3 (2010), 219–232. 2

[HSPC06] HAWKES E. R., SANKARAN R., PÉBAY P. P., CHEN J. H.: Direct numerical simulation of ignition front propagation in a constant volume with temperature inhomogeneities: II. Parametric study. *Combust. Flame 145* (2006), 145–159. 7

[JJ09] JOHANSSON S., JOHANSSON J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Trans. Vis. Comput. Graphics 15*, 6 (2009), 993–1000. 3

[JZF*09] JEONG D. H., ZIEMKIEWICZ C., FISHER B., RIBARSKY W., CHANG R.: iPCA: An interactive system for pca-based visual analytics. *Comput. Graph. Forum 28*, 3 (2009), 767–774. 3

[Lib13] LIBERTY E.: Simple and deterministic matrix sketching. *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (2013). 9

[LSL*13] LUM P. Y., SINGH G., LEHMAN A., ISHKANOV T., VEJDEMO-JOHANSSON M., ALAGAPPAN M., CARLSSON J., CARLSSON G.: Extracting insights from the shape of complex data using topology. *Sci. Rep. 3* (2013). 3

[LV09] LEE J. A., VERLEYSEN M.: Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing 72*, 7 (2009), 1431–1443. 1, 2, 3, 4

[LWG13] LISITSYN S., WIDMER C., GARCIA F. J. I.: Tapkee: An efficient dimension reduction library. *J. Mach. Learn. Res. 14* (2013), 2355–2359. 7

[MLGH13] MOKBEL B., LUEKS W., GISBRECHT A., HAMMER B.: Visualizing the quality of dimensionality reduction. *Neurocomputing 112* (2013), 109–123. 1, 2, 3, 4

[MYA*13] MANDELLI D., YILMAZ A., ALDEMIR T., METZROTH K., DENNING R.: Scenario clustering and dynamic probabilistic risk assessment. *Reliab. Eng. Syst. Safe. 115* (2013), 146–160. 8

[PEP*11] POCO J., ETEMADPOUR R., PAULOVICH F. V., LONG T., ROSENTHAL P., OLIVEIRA M., LINSEN L., MINGHIM R.: A framework for exploring multidimensional data with 3D projections. *Comput. Graph. Forum 30*, 3 (2011), 1111–1120. 3

[RS00] ROWEIS S. T., SAUL L. K.: Nonlinear dimensionality reduction by locally linear embedding. *Science 290*, 5500 (2000), 2323–2326. 2, 4

[SMC07] SINGH G., MÉMOLI F., CARLSSON G.: Topological methods for the analysis of high dimensional data sets and 3D object recognition. *Proc. Eurographics Symp. on Point-Based Graphics* (2007), 91–100. 3

[SR03] SAUL L., ROWEIS S.: Think globally, fit locally: unsupervised learning of nonlinear manifolds. *J. Mach. Learn. Res. 4* (2003), 119–155. 2

[TDSL00] TENENBAUM J. B., DE SILVA V., LANGFORD J. C.: A global geometric framework for nonlinear dimensionality reduction. *Science 290*, 5500 (2000), 2319–2323. 1, 4

[Tor52] TORGERSON W. S.: Multidimensional scaling: I. theory and method. *Psychometrika 17*, 4 (1952), 401–419. 1, 2, 3

[VPN*10] VENNA J., PELTONEN J., NYBO K., AIDOS H., KASKI S.: Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *J. Mach. Learn. Res. 11* (2010), 451–490. 2

[YRWG13] YUAN X., REN D., WANG Z., GUO C.: Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. *IEEE Trans. Vis. Comput. Graphics 19*, 12 (2013), 2625–2633. 3