Eurographics Conference on Visualization (EuroVis) 2016
K.-L. Ma, G. Santucci, and J. van Wijk
(Guest Editors)

Volume 35 (2016), Number 3

The Grassmannian Atlas: A General Framework for Exploring Linear Projections of High-Dimensional Data

S. Liu[1], P.-T Bremer[2], J. J. Jayaraman[2], B. Wang[1], B. Summa[3] and V. Pascucci[1]

[1]Scientific Computing and Imaging Institute, University of Utah
[2]Lawrence Livermore National Laboratory
[3]Department of Computer Science, Tulane University

Abstract

Linear projections are one of the most common approaches to visualize high-dimensional data. Since the space of possible projections is large, existing systems usually select a small set of interesting projections by ranking a large set of candidate projections based on a chosen quality measure. However, while highly ranked projections can be informative, some lower ranked ones could offer important complementary information. Therefore, selection based on ranking may miss projections that are important to provide a global picture of the data. The proposed work fills this gap by presenting the Grassmannian Atlas, a framework that captures the global structures of quality measures in the space of all projections, which enables a systematic exploration of many complementary projections and provides new insights into the properties of existing quality measures.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Line and curve generation

1. Introduction

Understanding high-dimensional data has become a central problem in a wide variety of applications ranging from the physical sciences and engineering to business and the social sciences. Among the large numbers of available techniques, linear projection (which produce 2D embeddings) remains the most popular approach as it is relatively cheap to compute and easy to interpret. One fundamental challenge is how to identify interesting and informative linear projections from all the possible projection directions. Even for datasets with moderate dimensions, exploring all possible axis-aligned projections, let alone all linear ones, becomes impractical. A method such as PCA (principal component analysis) can be used to obtain one optimal (maximizing variance) linear projection, however, high-dimensional datasets likely contain complex structures that cannot be adequately captured by a single linear projection.

Therefore, a common strategy is to search through a large number of potentially interesting projections and select a small set based on a ranking quality measure computed from the projections. The large number of candidates is usually generated by all axis-aligned projections [EDF08] or random sampling with dimension composition [Asi85, STBC03]. User-defined quality measures, such as the projection pursuit index [FT74], the rank-by-feature framework [SS05], and graph-theoretic scagnostics [WAG05, WAG06], are used to rank the candidates.

However, few techniques explicitly consider diversity when choosing representative projections. As a result, multiple highly ranked but redundant (similar) projections may be selected. At the same time, lower ranked ones are discarded even though they may contain complementary information. On the other hand, each quality measure is designed to capture some aspects of the data, yet little is known regarding the properties of the measure. For example, understanding the *smoothness* of a measure and the distribution of its *local maxima* is crucial in choosing the right representative projections. In particular, we demonstrate in our study that for some datasets, many quality measures contain a single maxima globally that may not be suitable for finding multiple projections.

We introduce the *Grassmannian Atlas*, a new framework to analyze, compare, and explore the space of all linear projections based on different quality measures. Rather than working with a few selected projections, the space of linear projections is modeled by the so-called *Grassmannian* [Har92], which abstracts the space of linear subspaces in a data-independent manner and compensates for affine transformations of the projections. The Grassmannian is approximated by connecting a set of *sampled* points (each corresponding to a subspace) on the surface of the manifold with a neighborhood graph based on well-defined geodesics. We then analyze a given quality measure as a scalar function defined on the Grassmannian and introduce the notion of *locally optimal* projections: the local maxima of the quality measure that are robust to small perturbations of the function. Consequently, using tools from scalar field topology, we extract a topological skeleton that describes the number, locations, and relationships among optimal projections, visualized by the topological spine [CLB11]. The topological spine captures the global structure of the quality measure across multi-

ple scales and provides important insights into its properties. It also leads to a visual map for exploring the space of projections in an intuitive manner.

Our key contributions are summarized below:

- We model the space of all linear projections based on a Grassmannian that parameterizes all linear subspaces of a high-dimensional dataset, and provide a sampling strategy to approximate the Grassmannian in any dimension;
- We construct a given quality measure as a scalar function on the Grassmannian and compute, analyze, and simplify its topological structure via the notion of topological spine;
- We provide a linked view interface based on topological spines to study locally optimal projections and explore the global structure of the space of all projections across different quality measures.

2. Related Work

Quality measures. A large number of quality measures have been proposed due to their practicality and simplicity for selecting interesting linear projections based on dimension selection. Tukey proposed a set of measures, coined *scagnostics*, to identify interesting axis-aligned scatterplots. The set of measures includes the area of the peeled convex hull, a modality measure of the kernel densities, a nonlinearity measure based on principal curves fitted to the scatterplots, etc. [WAG05]. This idea was extended by Wilkinson et al. [WAG05, WAG06] to include nine quality measures that capture properties such as outliers, shape, trend, and density. Guo [Guo03] introduced an interactive feature selection method for evaluating the maximum conditional entropy of all plots in a scatterplot matrix. Similarly, the rank-by-feature framework [SS04, SS06] allowed users to choose a ranking criterion for axis-aligned projections, such as histogram characteristics and correlation coefficients between axes.

In addition to measuring the quality of dimension selection among scatterplots, a large class of work is dedicated to assessing and measuring the quality of dimensionality reduction based on dimension composition (i.e., create new dimensions by combining existing ones). *Projection Pursuit* [FT74], one of the early approaches, defined the *interestingness* of a projection as its amount of deviation from a normal distribution. Mokbel et al. [MLGH13] used a pointwise *co-ranking* measure, which calculates the average number of neighbors that agree in high and low dimensions. Liu et al. [LWBP14] introduced a set of measures derived from objective functions that dimensionality reduction techniques aimed to minimize, such as *s*tress and *s*train [BG05]. Other criteria include measurements of distance distortions, density differences, or ranking discrepancies. Their system allowed direct manipulation of low-dimensional embeddings, guided by pointwise quality measures that get updated interactively to resolve structural ambiguities. An excellent survey that offered a comprehensive summary of various quality measures for visualizing high-dimensional data was provided by Bertini et al. [BTK11]. Our proposed framework is general enough to utilize any quality measure, but we focus on analyzing the global properties of a few popular ones, including the *projection pursuit* index, *scagnostics*, and *stress*.

Subspace clustering. Various subspace clustering methods provide interesting alternatives for selecting multiple linear projections to understand different aspects of high-dimensional data. These subspaces can be constructed either by selecting different subsets of the dimensions (subspace search) or by grouping subsets of the data that occupy common linear subspaces (subspace clustering). The former class of methods (e.g., [CFZ99]) was adopted for visualization [TMF*12] to capture complex multivariate structures. However, these subspace search methods are limited to finding axis-aligned subspaces only. In a recent work [LWT*14], the later class of subspace clustering methods [Vid11] is introduced, which identify views that focus on various subsets of data points that share a subspace. By assuming that, the high-dimensional dataset can be represented by a mixture of low-dimensional linear subspaces with mixed dimensions, this approach can produce views that focus on a specific region of the space spanned by each of the clusters. Our proposed framework is different in that we not only care about selecting multiple interesting linear projections, but also aim to gain a holistic understanding of their relationships. Such an understanding is made possible by first viewing quality metrics as a function on the *Grassmannian* and then summarizing the function and visualizing its abstraction.

Recently, Lehmann et al. introduced an interesting approach [LT16] to capturing the optimal set of linear projections. The proposed method adopted a dissimilarity measure, which produces a set of linear projections optimized for differences among them in order to remove duplicated data patterns. Compared to our approach, which is optimized for obtaining locally optimal views based on quality metrics, their method is tuned to maximizing dissimilarity, which may not guarantee the "quality" of the selected views.

3. Method

As mentioned above, the Grassmannian Atlas is designed to provide a more intuitive and reliable approach to select a set of 2D linear projections for visualization of a given high-dimensional point cloud. The challenge is that there exist an infinite number of possible projections, and the top ranked ones according to some quality measure may not be the most informative ones. In particular, similar projections are likely to have similar quality measures. Consequently, a cluster of very similar projections will be chosen over a potentially very different and more informative projection with slightly lower ranking. Instead, we propose to select a set of locally optimal projections as representatives based on computing the high-dimensional topological structure of the chosen quality measure. Figure 1 provides an overview of the approach. First, we randomly choose a (large) set of linear projections represented as linear subspaces and together with their neighborhood graph use them as a discrete approximation of the Grassmannian manifold, which defines the space of all possible linear projections (see Section 3.1). We then evaluate the chosen quality measure on the Grassmannian and compute its topological spine (Section 3.2). The local maxima of the topological spine then indicate locally optimal projections (with respect to the given measure), i.e., those that cannot be improved with incremental changes. Finally, the topological spines

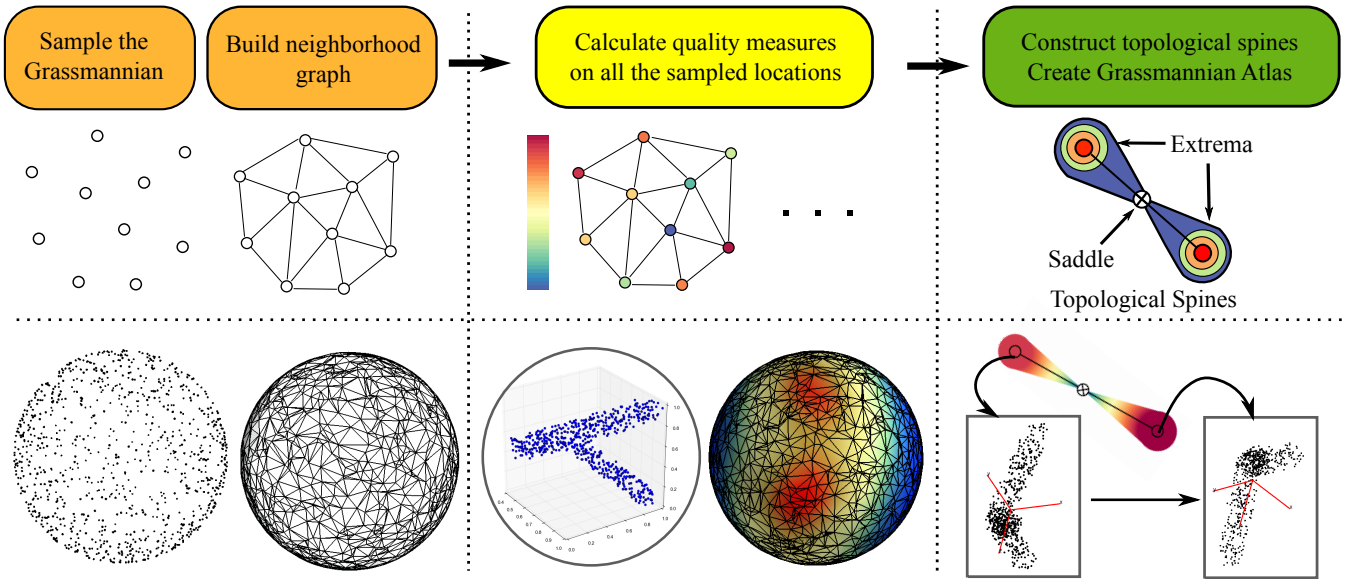

Figure 1: First row: the three steps (marked with different colors) for constructing the *Grassmannian Atlas*. Bottom row: examine the space of linear projections involving a 3D example. For illustration purposes, the left panel displays point cloud samples representing projections rather than subspaces as the Gassmannian has no intuitive embedding.

also serve as a convenient and intuitive interface to navigate between different projections.

3.1. Grassmannian Manifold

Grassmannian. We are interested in understanding the structure of quality measures on the space of projections. In a visualization setting, one typically can consider a set of projections to be *equivalent* if they produce the same scatterplots under affine transformations. Therefore, a somewhat simpler yet equivalent approach is to directly consider the space of 2D linear subspaces rather than the space of projections. Since projections that transform the data into the same subspaces produce equivalent scatterplots under affine transformations, the space of linear subspaces is much smaller than the space of projections, does not suffer from redundancies, and most importantly, admits a well-known geodesic distance metric among the subspaces.

The space of r-dimensional linear subspaces of $\mathbb{R}^n$ is called the Grassmannian, denoted by $Gr(r,n)$, and is known to be an embedded manifold of dimension $r(n-r)$ [Har92]. Each point on $Gr(r,n)$ represents a linear subspace typically encoded by its orthonormal basis. Given two subspaces with orthonormal basis $\mathbf{A}$ and $\mathbf{B}$, their geodesic distance on the manifold can be computed by decomposing $\mathbf{A}^T\mathbf{B}$ using its SVD (singular value decomposition) and obtaining $\sum_{i=1}^{r}\left(\theta_i^2\right)^{\frac{1}{2}}$. Each $\theta_i = \cos^{-1}\sigma_i$ denotes a principal angle, where σ_i is the corresponding singular value. In our context, we have $r = 2$ and study the space of 2D linear subspaces $Gr(2,n)$ for an n-dimensional (i.e., nD) dataset.

Uniform sampling. To obtain an approximation of the Grassmannian $Gr(2,n)$, we generate a discrete point cloud sample of the manifold and construct a neighborhood graph based upon the geodesic distances on the manifold. Ideally, the sample should be uniformly random and dense to adequately capture the structure of the manifold, as well as the structure of a reasonable function defined on the manifold. We first discuss how to construct an approximately uniform sampling of a given size, and later in this section, we provide experiments for understanding the relationships among input data dimension, sample size, and sample density. The sampling quality is evaluated in Section 4.

A random sample on the Grassmannian $Gr(2,n)$ can be generated by constructing uniformly distributed random rotation matrices [Mez06]. More specifically, we use the QR decomposition [JM92] of a Gaussian random matrix S (i.e., a matrix that contains random numbers with a Gaussian distribution) to compute a random rotation matrix T, that is, $T = Q \cdot \mathrm{diag}(\mathrm{sign}(\mathrm{diag}(R)))$ where $S = QR$. A random sample on the Grassmannian therefore corresponds to a 2D subspace generated by applying a random rotation matrix to a pair of standard basis in $\mathbb{R}^n$. To ensure the set of rotation matrices is approximately uniformly distributed, we can resample the initial points using the k-means++ seed point initialization algorithm [AV07], which maximizes the spread of points by selecting points away from already selected samples. Finally, we construct a neighborhood graph connecting the sampled points using geodesics. Since the sample is approximately uniform, a k-nearest neighbor graph (kNN) is sufficient (with an appropriately chosen k). Such a graph is a discrete approximation of $Gr(2,n)$ that supports the subsequent topological analysis.

Sampling experiments. In practice, for a given data dimension n, the choice of the number of samples is crucial for reliable analysis of the data. To this end, we study the relationships among the number of samples (m), the data dimension (n), and the sampling density defined by the average nearest neighbor distance (d_{ann}). In

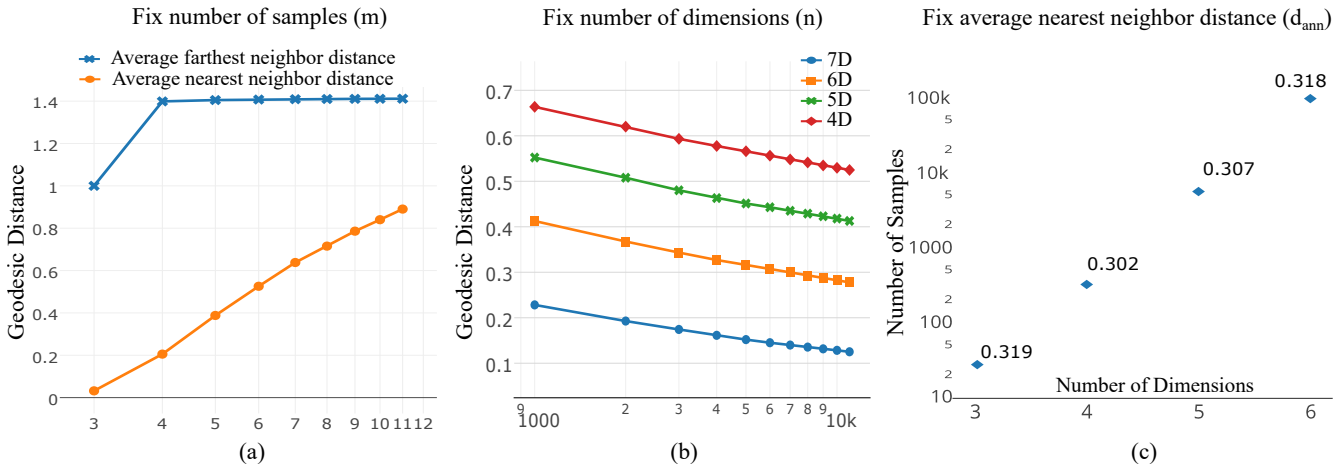

Figure 2: Sampling experiments. Let m be the sample size, d_{ann} be the average nearest neighbor distance, and n be the data dimension. (a) For a fixed $m = 1500$, d_{ann} increases with an exponential increase of n (x-axis, log-scale). (b) For a fixed n ($4 \leq n \leq 7$), d_{ann} (y-axis) decreases with an exponential increase of m (x-axis, log-scale). (c) To maintain a fixed density $d_{ann} \approx 0.3$, m (y-axis, log-scale) scales exponentially with n (x-axis).

Figure 2(a), for a fixed $m = 1500$, we vary the data dimension n where $3 \leq n \leq 10$, and compute d_{ann}. We observe that d_{ann} increases as n grows exponentially (notice that x-axis is log-scale), indicating increasing sparsity in higher dimensions. In Figure 2(b), for a fixed n ($4 \leq n \leq 7$), we observe that d_{ann} decreases with the exponential increase of m (notice that the x-axis is log-scale). Finally in Figure 2(c), we illustrate that for an approximately fixed $d_{ann} \approx 0.3$, the required number of samples m increases exponentially with the number of dimensions n (notice that the y-axis is log-scale).

3.2. Quality Measures

Our framework applies to any quality measure; in this work, we focus on three categories: *scagnostics* [WAG05, WAG06], *projection pursuit indices* [CBC93, LCKL05], and the measures derived from objective functions of dimensionality reduction methods [LWBP14].

The graph-theoretic scagnostics comprises a set of nine measures describing the shape, trend, and density of points from linear projections: *outlying*, *skewed*, *sparse*, *clumpy*, *striated*, *convex*, *skinny*, *stringy*, and *monotonic*. These measures help to automatically highlight interesting or unusual scatterplots from a scatterplot matrix. Scagnostics computation relies on graph-theoretic measures such as the convex hull, alpha hull, and minimal spanning tree of the points. Take the *skinny* measure for example, $c_{skinny} = 1 - \sqrt{4\pi area(A)}/perimeter(A)$, where A indicates an alpha hull of the points in the projection.

Projection pursuit indices are quality measures developed on the basis of the original projection pursuit approach [FT74] to capture various features in a projection. In particular, we include *gini*, *entropy* [LCKL05] (highlighting class separation), *central mass*, and *hole* [CBC93] measures in this study. Finally, the objective functions of dimensionality reduction methods are also used for identifying interesting projections. Linear Discriminant Analysis (LDA) can be adopted to measure the amount of class separation. *Stress*, which is the objective function in the distance scaling version of Multidimensional Scaling (MDS), measures the quality of distance preservation. Let d_{ij} be the distance between a pair of points i, j in $\mathbb{R}^n$ and $\hat{d}_{ij}$ be the corresponding distance in $\mathbb{R}^k$, where $k < n$. Stress is defined as $\sum_{i,j}(d_{ij} - \hat{d}_{ij})^2 / \sum_{i,j} d_{ij}^2$ [BSL*08].

Given an approximation of the Grassmannian, we consider various quality measures of interest as scalar functions on the Grassmannian, and calculate their values on all the sampled locations.

3.3. Topological Summaries of Quality Measures

Given the list of subspace (samples) with the corresponding quality values, the tradition approach simply selects the highest ranking views and presents them to the user. However, as discussed above, some of these views may be similar and thus redundant. Consider the 1D example of Figure 3(a). The two highest ranking samples are close together, i.e., represent a very similar projection, but the second peak is ignored, even though in practice it may provide a very different view and thus likely more information. Treating the samples as individual points, these relationships between subspaces are difficult to consider. Exploiting the underlying manifold structure, however, leads to an intuitive definition of locally optimal subspace. Given both the samples and their neighborhood relations it is natural to consider only those subspaces that have no neighbor with a higher metric value. Intuitively, we prefer views where no small adjustment could lead to a higher quality value. Such a tendency naturally leads to the concepts of topology and in particular the Morse complex of the metrics.

Morse complex and persistence. We use the topological notions of *Morse complex* to identify local maxima of a function and *persistence* to quantify their robustness.

Given an Morse function defined on a smooth manifold, $f : \mathbb{M} \rightarrow$

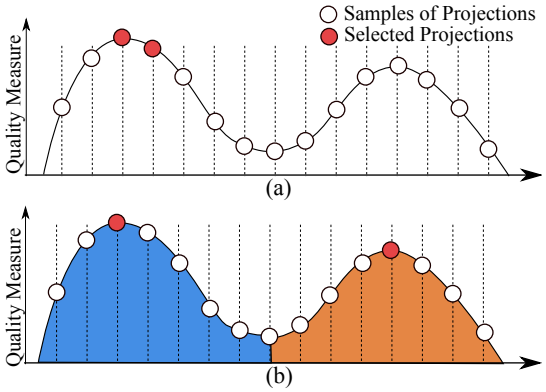

Figure 3: Selecting projections based purely on the ranking of a quality measure, (a) fails to identify structurally distinct projections as those obtained via topological analysis (b).

$\mathbb{R}$, An *integral line* of f is a path in $\mathbb{M}$ whose tangent vector agrees with the gradient of f at each point along the path. An integral line starts at a local minimum and ends at a local maximum of f. *Descending manifolds* (surrounding local maxima) are constructed as clusters of integral lines that have common destinations. The descending manifolds form a cell complex that partitions $\mathbb{M}$, referred to as the *Morse complex*.

In our context, $\mathbb{M}$ is the Grassmannian, a smooth manifold without a boundary, and f is a quality measure of interest. We identify local maxima of f based on the Morse complex, and they correspond to structurally distinct regions within the landscape of f. To further quantify the robustness of a local maximum, we use the notion of topological persistence. The *persistence* of a local maximum is defined to be the minimum amount of perturbation to the function that removes it. In Figure 3, for example, the right peak is less persistent than the left peak, since it can be removed with a nearby critical point (e.g., a local minimum) with a smaller amount of perturbation. We use the discrete algorithm of [GBPW10] to approximate the Morse complex of a measure, given a sampling and neighborhood graph as discussed in Section 3.1.

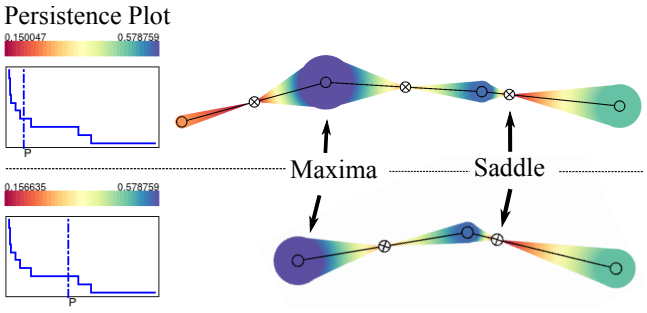

Figure 4: Multiscale topological spine representations. The persistence plots are shown on the left: the x-axis corresponds to the persistence threshold, and the y-axis is the number of current cells in the simplification. The long plateau in the persistence plot (bottom) corresponds to a stable topological structure.

Topological spines. The Morse complex provides a structural summary of the topology of a function, and is well defined in any dimension, but it is not easy to visualize. Instead, we use the concept of topological spines [CLB11] to visualize both the space of projections and an intuitive interface for users to select and explore various projections (see Figure 4)

The topological spine adapts a terrain metaphor, as shown in Figure 4, that connects local maxima whose corresponding descending manifolds have shared boundaries. Intuitively, these connections can be interpreted as the ridge-lines between neighboring peaks of a terrain. The topological spine uses two parameters to simplify its structure. First, *persistence* is used to remove noise and artifacts to construct a simplified dual complex. Second, a *variation* threshold is provided to determine which of the remaining connections should be considered "ridge-like", and only those above the threshold are visualized. Furthermore, the size of each cell in the Morse complex, i.e., the number of samples it contains, is encoded by the width of the topological spine. The persistence plot (see Figure 4) is essential for understanding the distribution of robust features in the function: a long flat plateau indicates the existence of multiple robust peaks that are good candidates for selection, whereas a descending slope suggests excessive noise and the lack of robust structures.

3.4. Interactive User Interface

Apart from the automatic selection of locally optimal views, our system also allows users to interactively explore the different view points using the topological spine as a selection interface. In particular, the system allows selection of the simplification (persistence) levels that automatically updates the spine and provide dynamic transitions between maxima/projections. The interface consists of two linked views, the topological spine panel and the dynamic projection panel. The former displays the topological spine of the chosen quality measure at the selected persistence set directly via the embedded persistence plot (see Figure 4). The projection panel displays the dataset using the currently selected linear projection (local maxima). To better understand the relationships between projections we use the dynamic projection approach [STBC03] to create animated transitions between projections by displaying a set of intermediate linear projections (see the supplementary video for more details).

3.5. Computation Complexity

Since the sampling of Grassmannian $Gr(2,n)$ and the construction of neighborhood graphs are independent from the actual dataset as well as the quality measures, the sampling process need to be computed for each dimension n only once. Let m be the number of data points, n the number of data dimensions, and k the number of samples on the Grassmannian. Evaluating the quality measures for each linear projection takes between $O(mn^2)$ (Scagnostics with binning optimization) and $O(m^2n)$ (Stress). The algorithm used to construct the topological spine from the samples of a given quality measure has a complexity of $O(k\log k)$. Therefore, the overall computation complexity for a given data with a selected quality measure is $O(m^2nk+k\log k)$. The theoretical relationship between

the number of samples k and data dimension n is examined in Section 4. Quality measures and their corresponding topological spines are pre-computed to support interactive exploration. For the examples in this paper, the computation time varies between 2 to 30 minutes, depending on the data dimension, sample size, and the number of quality measures. The test setup consists of a machine with Intel Core i5 2.8GHz processor running Linux. The software framework is written in C++/Qt and compiled with GCC 4.8.

4. Results and Evaluation

In this section, we first evaluate our sampling procedure by showing that our approach samples the Grassmannian evenly and completely. Subsequently, we show that the topological structure is stable for different sampling sizes and neighborhood graphs. We then compare the topological structures of various metrics on different datasets to better understand the behavior of each metric. Finally, we show our collaboration with domain experts for applying the proposed framework to the Word2Vec dataset.

4.1. Parameter Validation: Sampling Density and Sampling Size

To reliably represent functions defined on the Grassmannian, we require a uniformly distributed sample that covers the entire manifold. For moderate input dimensions, the Grassmannian has comparatively low dimensions, and creating sufficient samples, especially during offline pre-processing, is straightforward. If the data dimension becomes too large for the available resources, the Grassmannian has been shown to be amenable to dimension reduction, i.e., a PCA [Jol05].

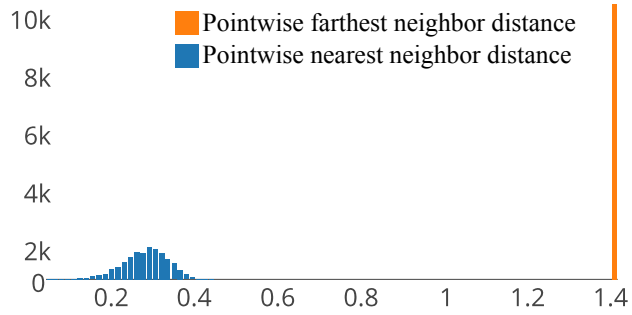

Figure 5: A histogram showing the distribution of pointwise nearest (blue) and farthest (orange) neighbor distances for $Gr(2,5)$ with $10K$ samples.

To validate our results, Figure 5 shows the histogram of nearest neighbor distances and farthest neighbor distances for $10k$ samples from $Gr(2,5)$. As expected, the nearest neighbor distances are tightly clustered, indicating a nearly uniform distribution. Similarly, the farthest neighbor distances indicate that the entire manifold has a "diameter" of 1.4. As the sample is random and/or resampled, the uniform farthest neighbor distance makes it unlikely (though not impossible) that the manifold is not completely covered. However, a high-quality sample of the Grassmannian does not necessarily guarantee that a given metric defined on the Grassmannian is well sampled.

Figure 6 shows the persistence plots and topological spines for the two-planes dataset (see below) for different numbers of samples and different neighborhood sizes for graph construction. All results are stable, indicating that at least for this dataset the Grassmannian is sufficiently sampled and our approach is numerically stable. We have performed similar parameter studies for all experiments in the paper and found similar results.

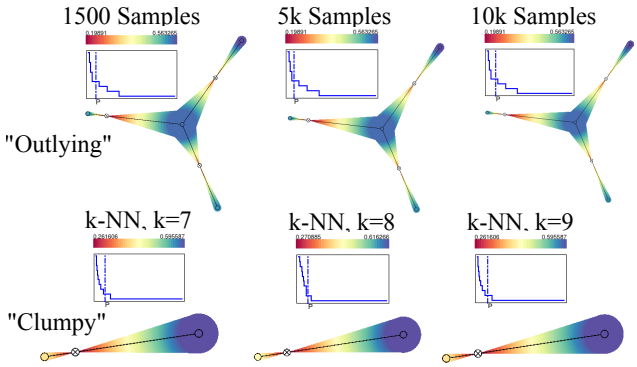

Figure 6: Validating the stability of topological spines by varying the number of samples and the number of neighbors for the k-NN graph.

4.2. Validation With Synthetic Two-Planes Dataset

To evaluate the effectiveness of our approach we analyze a synthetic dataset containing samples from two 2D planes embedded in $\mathbb{R}^3$ that intersect with a 75-degree angle (see Figure 7(a)). The *scagnostics skinny* measure (Figure 7(b)) identifies the head-on projection in which both planes are skinny as the main mode and various other projections where only a single plane is "skinny" as alternatives. The *Stress* measure (Figure 7(c)) finds only a single, stable maximum, which identifies an average view in which both planes are equally distorted. The projection pursuit index *central mass* (Figure 7(d)), on the other hand identifies good projections for both planes as local maxima. These experiments demonstrate that the Grassmannian Atlas not only is able to identify good projections but also provides insights into the measure itself. A measure with only a single stable maximum likely produces some globally average view whereas multiple maxima indicate several complementary views emphasizing different, local aspects of the data.

4.3. Quality Measure Comparisons

The Grassmannian Atlas not only helps to identify complementary projections and summarize the structure of quality measures, but also provides an avenue for examining and comparing high-level structures of quality measures in general. In particular, the persistence plot encodes a number of interesting properties in a concise and intuitive manner. As discussed in Section 3.3, the persistence plot records the number of salient local maxima depending on the simplification threshold. In general, the most interesting feature in a persistence plot is the number and width of stairs. Multiple stairs indicate several sets of complementary projections, and the width encodes how stable these features are.

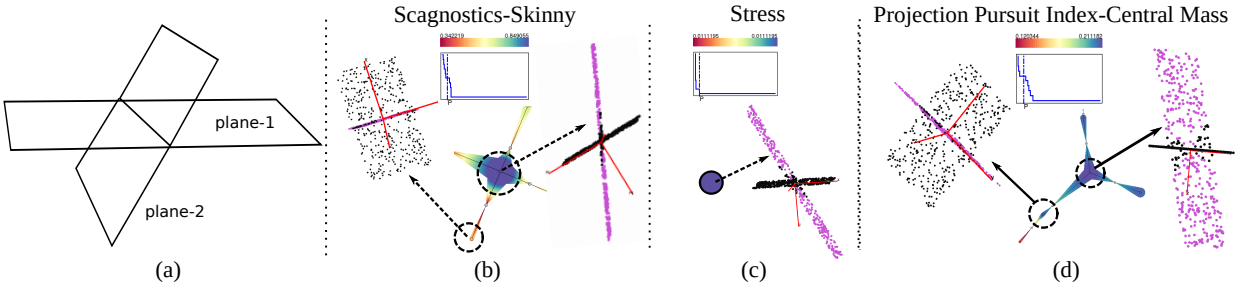

Figure 7: Validate the *Grassmannian Atlas* framework on a synthetic two-planes dataset. The dataset is sampled from the space illustrated in (a). In (b), the two maxima within the topological spine correspond to the projections where one or both planes are at the "skinniest". In (c), the (global) stress measure captures one only interesting projection at its global maxima. In (d), the projection pursuit index *central mass* measure captures the two projections where one of the two planes becomes "skinny".

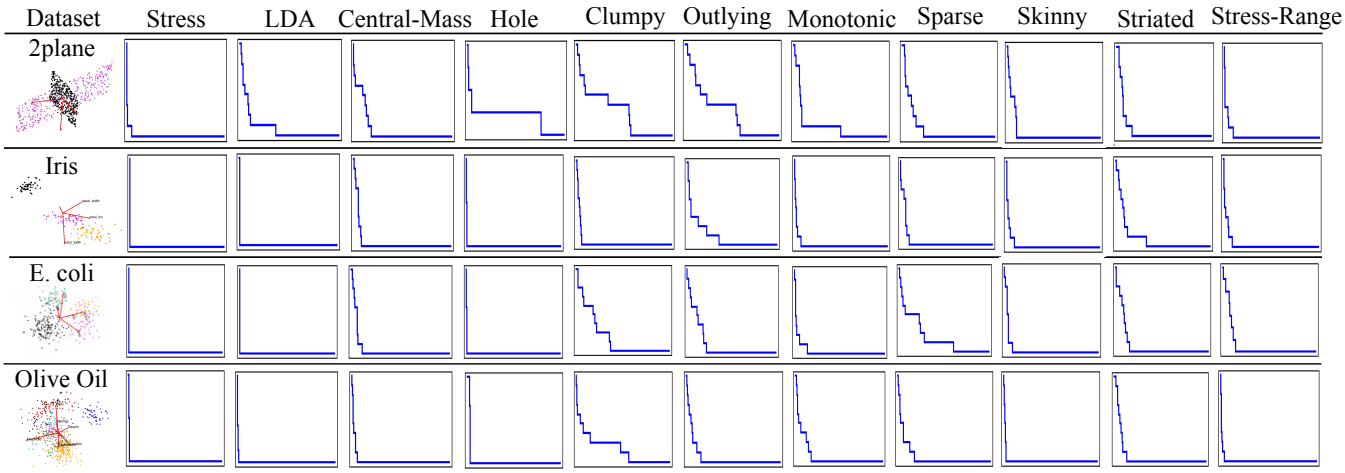

Figure 8: Quality measures comparison by evaluating their respective persistence plots, which provide concise summaries of the multiresolution topological structure. Only four datasets are shown here due to space constrains.

We compute the persistence plots for all 16 quality measures (9 scagnostics, 3 projection pursuit indices, 4 based on objective functions of dimension reduction techniques) and include 11 of these in Figure 8. For each measure we evaluate its behavior for five datasets: (i) 2-planes synthetic dataset (3D), (ii) UCI Iris dataset (150 samples in 4D), (iii) UCI E. coli dataset (332 samples in 6D, a subset of the original 336 samples in 8D), (iv) olive oil dataset (572 samples in 8D), and (v) housing dataset (506 samples in 14D). The details for each dataset can be found in the UCI machine learning repository (`http://archive.ics.uci.edu/ml/`).

As shown in Figure 8, surprisingly few measures ever show more than two or three complementary projections based on the number of wide stairs in their persistence plots, and the *stress* measure captures a single robust projection in most cases. Such an observation has important implications for ranking-based projection selection - selecting more projections would most likely result in information redundancy. The significant discrepancies among the topological structures of different quality measures can be explained by their formulations and design goals. The *stress* measure originates from the objective function of MDS [BSL*08], and is designed to create a single embedding that best preserves the pairwise distances. Therefore the *stress* measure typically produces a single projection that is optimal on average. On the other hand, quality measures that focus on evaluating the quality of projections based on local structure preservation typically provide multiple, complementary projections. As shown in Figure 8, the *clumpy*, *outlying* measures are some of the more effective ones for identifying complementary projections.

In general, given an appropriate quality measure, the Grassmannian Atlas can reliably identify potentially diverse and locally optimal projections. Compared to conventional rank-based approaches, our framework summarizes the structural relationships among projections according to the topology of the quality measure, and provides a more reliable and locally optimal set of projections for visualization.

For example, as shown in Figure 9, based on the *clumpy* quality measure, our framework identifies multiple interesting projections for the E. coli dataset that capture meaningful biological relationships.

The data points (corresponding to different E. coli strains) in the two highlighted projections form clear clusters that are well

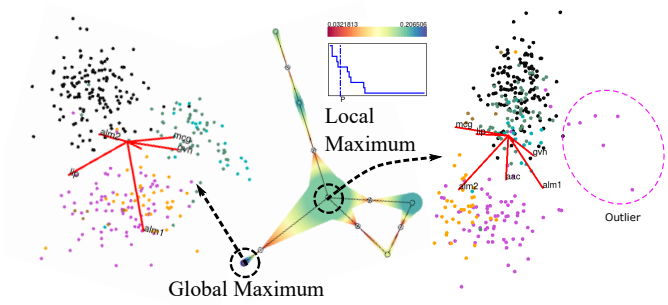

Figure 9: The complementary projections captured by *Grassmannian Atlas* using the scagnostics *clumpy* measure for the E. coli dataset.

aligned with the localization site classification labels (see details in [HN96]). The black corresponds to the *cytoplasm* localization site, which comprises *cytosol* (the gel-like substance enclosed within the cell membrane) and the *organelles* (the cell's internal sub-structures); the purple represents inner membrane without signal sequence; the orange contains inner membrane with uncleavable signal sequence; the light green corresponds to outer membrane; the brown (with only 5 points) is the outer membrane lipoprotein; and the dark green corresponds to *perisplasm*, a concentrated gel-like matrix in the space between the inner cytoplasmic membrane and the bacterial outer membrane. The projection at the global maxima captures clear separation between the black, and the (light and dark) green points, separating materials from the inner membrane to the ones from or close to the outer membrane. On the other hand, the projection at the local maxima merges the black with the green points. Both projections group the purple and orange points into one cluster that contains information regarding the inner membrane.

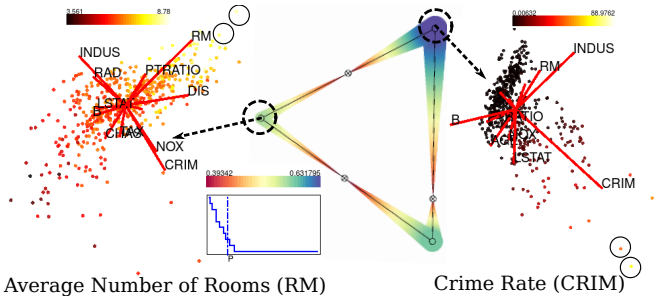

Figure 10: The different outliers captured by the *Grassmannian Atlas* using the scagnostics *outlying* measure for the housing dataset. The outliers are highlighted by small solid circles.

Figure 10 shows a set of housing data in which each entry records various property characteristics (14 in total), such as crime rate, median property value, average number of rooms per dwelling, etc. of towns in Boston area. By utilizing the proposed framework and examining the topological spine and corresponding projection computed from the *outlying* measure, we are able to identify some interesting outliers which shed light on the large socioeconomic inequality correlated with the geological separation.

As shown in the projection on the right, we are able to identify outliers that correspond to towns with a comparatively very high crime rate. The difference is so extreme that this outlying pattern is strongest among all the linear projection samples. By looking at one of the local extrema (the projection on the left), we can see the average number of rooms also are correlated with some outliers. After examining the individual data points, we can see the outliers corresponding to the towns that have around 8-9 average rooms per dwelling, while at the same time the minimal number is around 3.5.

4.4. Word2Vec Dataset

The following study of Word2Vec dataset is a collaboration with an expert in natural language processing (NLP). The popular Word2Vec algorithm [MSC*13] learns a vector space representation of words by modeling the intrinsic semantics of large text corpora. It consolidates the statistical relationships between words in an abstract high-dimensional feature space. According to our collaborator, the analysis and visualization approach for such a dataset is very limited. Often, the t-SNE [VdMH08] nonlinear projection algorithm is used for visualization, but most relationships in Word2Vec are linear in nature. He suggests a visualization tool that can produce interesting linear projections to emphasize semantic properties in different parts of the data could lead to valuable new insights.

The complete Word2Vec dataset is obtained by running the Word2Vec algorithm on corpora of news articles, containing 100 billion words. The dimension of the resulting vector representations for the words is fixed at 300. The data used in our experiment is a small subset of the Word2Vec dataset, containing 900 frequently occurring words obtained from the Google analogy task list. This list contains pairs of words with a semantic or syntactic relationship between them, e.g., (queen, king) and (man, woman). Following this, we use PCA to reduce the dimension of the word vectors to 5D in order to reduce the sampling cost. Note that subsampling and dimension reduction are both common strategies in NLP to limit the complexity of the input data without introducing significant errors. To provide a context for the visualization, we label the 900 words with 10 categories such as adjective, adverb, verb, and different groups of nouns (e.g., capitals and countries in different continents, states of the US, etc.).

As shown in our quality measure comparison analysis in Section 4.3, several measures, such as *clumpy*, *outlying*, which are more likely to identify multiple complementary projections. In addition, *clumpy* by definition will likely highlight cluster-like features. As demonstrated in Figure 11, the *clumpy* measure helps capture the projections that reveals interesting semantic relations in the analogy dataset. The largest maxima (shown on the right) correspond to a projection that clearly separates cities and countries from all other words and does well in separating their respective continents (e.g., orange for North America, dark green for Europe, and blue for South America). A second projection (shown on the left) does less well on cities and countries, but nicely separates the remaining groups of words. Our collaborator considers the left projection to be the most informative overall, yet it does not have a very high global ranking, and it would likely be ignored in a ranking-based approach.

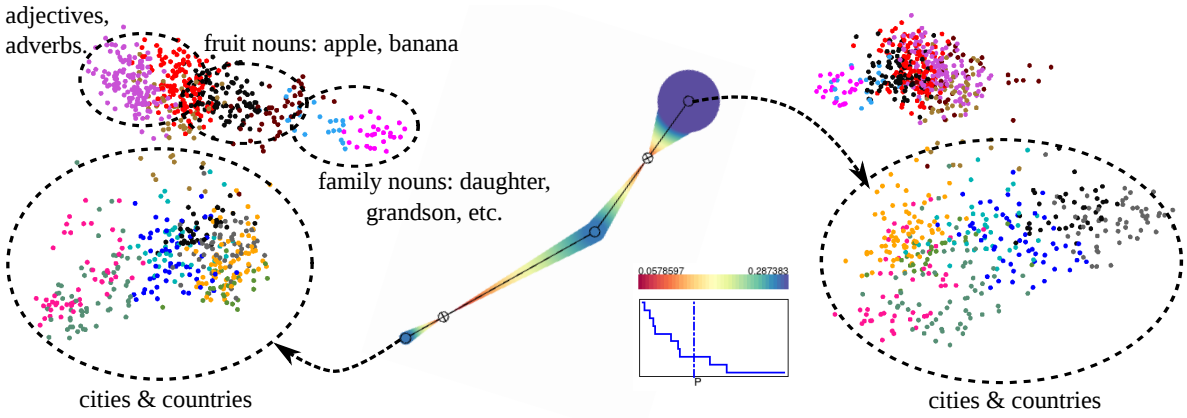

Figure 11: Word2Vect dataset. The *clumpy* measure helps to identify the two projections that highlight clear separation between cities and countries from the rest of the data points.

A one-on-one session is carried out to obtain meaningful feedback from the collaborator. First, a carefully prepared demo by the researcher is presented to the collaborator. Then the collaborator is directed to experiment with the tool to explore the various measures and projections interactively. The session is concluded by a discussion regarding the capability and usability of the tool. Our collaborator shows great interest in the capability of the proposed framework. He points out that the Grassmannian Atlas framework can be a useful tool for exploring the word feature space, especially considering it does not have any restriction on what quality measures can be adopted. For example, he suggests new measures specifically tailored towards text analysis can be designed by incorporating semantic relationships among words. Regarding the possible challenges for using the proposed tool, the collaborator points out the basic concept can be challenging to digest at first, since it approaches the problem from a fundamentally different perspective (the space of all linear projections).

5. Conclusion

The Grassmannian Atlas provides a fundamentally unique approach to exploring the space of all linear projections, the Grassmannian. By studying quality measures as functions defined on the Grassmannian, we are able to identify local optimal projections as well as obtain an intuitive understanding of the topological structures of the quality measures themselves. Our framework not only enables the comparison among multiple quality measures (Figure. 8), but also helps to guide the design of and provide benchmarks for new quality measures.

The advantage of our approach lies in the ability to provide a holistic interpretation of the space of all linear projections for a given measure, However, this ability also leads to an unavoidable battle against the *curse of dimensionality*: the space complexity of the Grassmannian and the number of samples (based on current sampling techniques) needed for a reliable coverage grow exponentially in the number of dimensions. To mitigate this issue, we assume that our data typically has low intrinsic dimensions and apply dimension reduction as a pre-processing step, and therefore retain a balance between the sampling expense and the data accuracy. To decrease the number of samples required for accurately representing the Grassmannian, an adaptive data centric sampling approach is preferred. However, at the moment efficient sampling of an arbitrary function on the Grassmannian is still an open problem (and also an opportunity for future research).

Acknowledgments

This work was performed in part under the auspices of the US DOE by LLNL under Contract DE-AC52-07NA27344., LLNL-CONF-658933. This work is also supported in part by NSF IIS-1513616, NSF 0904631, DE-EE0004449, DE-NA0002375, DE-SC0007446, DE-SC0010498, NSG IIS-1045032, NSF EFT ACI-0906379, DOE/NEUP 120341, DOE/Codesign P01180734. Bei Wang is partially supported by NSF IIS-1513616.

References

[Asi85] ASIMOV D.: The grand tour: a tool for viewing multidimensional data. *SIAM SISC 6*, 1 (1985), 128–143. 1

[AV07] ARTHUR D., VASSILVITSKII S.: k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (2007), Society for Industrial and Applied Mathematics, pp. 1027–1035. 3

[BG05] BORG I., GROENEN P. J.: *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005. 2

[BSL*08] BUJA A., SWAYNE D. F., LITTMAN M. L., DEAN N., HOFMANN H., CHEN L.: Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics 17*, 2 (2008), 444–472. 4, 7

[BTK11] BERTINI E., TATU A., KEIM D.: Quality metrics in high-dimensional data visualization: An overview and systematization. *Visualization and Computer Graphics, IEEE Transactions on 17*, 12 (2011), 2203–2212. 2

[CBC93] COOK D., BUJA A., CABRERA J.: Projection pursuit indexes based on orthonormal function expansions. *Journal of Computational and Graphical Statistics 2*, 3 (1993), 225–250. 4

[CFZ99] CHENG C.-H., FU A. W., ZHANG Y.: Entropy-based subspace clustering for mining numerical data. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (1999), ACM, pp. 84–93. 2

[CLB11] CORREA C., LINDSTROM P., BREMER P.-T.: Topological spines: A structure-preserving visual representation of scalar fields. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 1842–1851. 1, 5

[EDF08] ELMQVIST N., DRAGICEVIC P., FEKETE J.-D.: Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE TVCG 14*, 6 (2008), 1539–1148. 1

[FT74] FRIEDMAN J., TUKEY J.: A projection pursuit algorithm for exploratory data analysis. *IEEE TC C-23*, 9 (1974), 881–890. 1, 2, 4

[GBPW10] GERBER S., BREMER P.-T., PASCUCCI V., WHITAKER R.: Visual exploration of high dimensional scalar functions. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1271–1280. 5

[Guo03] GUO D.: Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization 2*, 4 (2003), 232–246. 2

[Har92] HARRIS J.: *Algebraic geometry: a first course*, vol. 133. Springer Science & Business Media, 1992. 1, 3

[HN96] HORTON P., NAKAI K.: A probabilistic classification system for predicting the cellular localization sites of proteins. In *Ismb* (1996), vol. 4, pp. 109–115. 8

[JM92] JENNINGS A., MCKEOWN J. J.: *Matrix computation*. Wiley New York, 1992. 3

[Jol05] JOLLIFFE I.: *Principal component analysis*. Wiley Online Library, 2005. 6

[LCKL05] LEE E.-K., COOK D., KLINKE S., LUMLEY T.: Projection pursuit for exploratory supervised classification. *Journal of Computational and Graphical Statistics 14*, 4 (2005). 4

[LT16] LEHMANN D., THEISEL H.: Optimal sets of projections of high-dimensional data. *Visualization and Computer Graphics, IEEE Transactions on 22*, 1 (Jan 2016), 609–618. 2

[LWBP14] LIU S., WANG B., BREMER P.-T., PASCUCCI V.: Distortion-guided structure-driven interactive exploration of high-dimensional data. *Computer Graphics Forum 33*, 3 (2014), 101–110. 2, 4

[LWT*14] LIU S., WANG B., THIAGARAJAN J. J., BREMER P.-T., PASCUCCI V.: Multivariate volume visualization through dynamic projections. In *Large Data Analysis and Visualization (LDAV), 2014 IEEE 4th Symposium on* (2014), IEEE, pp. 35–42. 2

[Mez06] MEZZADRI F.: How to generate random matrices from the classical compact groups. *arXiv preprint math-ph/0609050* (2006). 3

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

[MSC*13] MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S., DEAN J.: Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (2013), pp. 3111–3119. 8

[SS04] SEO J., SHNEIDERMAN B.: A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on* (2004), IEEE, pp. 65–72. 2

[SS05] SEO J., SHNEIDERMAN B.: A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization 4*, 2 (2005), 96–113. 1

[SS06] SEO J., SHNEIDERMAN B.: Knowledge discovery in high-dimensional data: Case studies and a user survey for the rank-by-feature framework. *IEEE TVCG 12*, 3 (2006), 311–322. 2

[STBC03] SWAYNE D. F., TEMPLE LANG D., BUJA A., COOK D.: GGobi: evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis 43* (2003), 423–444. 1, 5

[TMF*12] TATU A., MAAS F., FARBER I., BERTINI E., SCHRECK T., SEIDL T., KEIM D.: Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on* (2012), IEEE, pp. 63–72. 2

[VdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-sne. *Journal of Machine Learning Research 9*, 2579-2605 (2008), 85. 8

[Vid11] VIDAL R.: A tutorial on subspace clustering. *IEEE Signal Processing Magazine* (2011). 2

[WAG05] WILKINSON L., ANAND A., GROSSMAN R. L.: Graph-theoretic scagnostics. In *INFOVIS* (2005), vol. 5, p. 21. 1, 2, 4

[WAG06] WILKINSON L., ANAND A., GROSSMAN R.: High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *Visualization and Computer Graphics, IEEE Transactions on 12*, 6 (2006), 1363–1372. 1, 2, 4