Optimal Sets of Projections of High-Dimensional Data

Dirk J. Lehmann and Holger Theisel

Abstract—Finding good projections of n-dimensional datasets into a 2D visualization domain is one of the most important problems in Information Visualization. Users are interested in getting maximal insight into the data by exploring a minimal number of projections. However, if the number is too small or improper projections are used, then important data patterns might be overlooked. We propose a data-driven approach to find minimal sets of projections that uniquely show certain data patterns. For this we introduce a dissimilarity measure of data projections that discards affine transformations of projections and prevents repetitions of the same data patterns. Based on this, we provide complete data tours of at most $n/2$ projections. Furthermore, we propose optimal paths of projection matrices for an interactive data exploration. We illustrate our technique with a set of state-of-the-art real high-dimensional benchmark datasets.

Index Terms—Multivariate Projections, Star Coordinates, Radial Visualization, High-dimensional Data

1 Introduction

To a large extent, Information Visualization deals with high-dimensional datasets, i.e., data that can be described as point sets in a high-dimensional space. Finding appropriate projections into 2D (or 3D) is a standard problem in Information Visualization for which a variety of approaches have been proposed. Traditional data projection strategies often provide a complete tour through the space of all data projections, that quadratically or even exponentially grows with the dimensionality n of the data. Due to the large number of projections, such techniques tend to be exhausting for the user, even if n is rather small. Further approaches aim to reduce the number of dimensions, bearing the risk to overlook and lose important data patterns. Beyond that, no approach avoids the repetitive view on similar data patterns.

Finding good data projections is a non-trivial problem due to the following two reasons. Firstly, every projection discards information about the data while introducing distortions. Secondly, the space of all possible projections is large. To evaluate how useful a certain projection is, a variety of quality measures have been proposed. They describe the quality of a particular projection by a certain number.

In this paper, we propose a new approach to find relevant projections. Instead of evaluating the quality of a single projection, we introduce a simple measure of how much more insight is provided by a new projection if a number of other projections are already presented. Our main assumption here is that a new projection does not provide new insight if it can be obtained by a linear combination of optimal affine transformations of the already existing projections. Figure 1 illustrates the concept: the two projections $\mathbf{p}_1$ and $\mathbf{p}_2$ of a high-dimensional point set are considered similar because $\mathbf{p}_2$ can be obtained from $\mathbf{p}_1$ by an affine map. Contrary, if $\mathbf{p}_1$ is given, the new projection $\mathbf{p}_3$ gives new insight because it cannot be obtained by an affine map from $\mathbf{p}_1$. In fact, $\mathbf{p}_3$ shows that the data consists of (at least) two clusters which could not be seen in $\mathbf{p}_1$. Finally, if $\mathbf{p}_1$ and $\mathbf{p}_3$ are given, $\mathbf{p}_4$ still gives additional information about the data because no linear combination of affine transformations of $\mathbf{p}_1$ and $\mathbf{p}_3$ can give $\mathbf{p}_4$.

The discarding of affine transformations for comparing projections is justified in the following observation: one of the most common research questions is to find patterns and clusters in the data. If a projection reveals e.g. two clusters, the same clusters are usually visible in an affine transformation of the projection. Moreover, if two clusters in the high-dimensional data space are projected to the same location in 2D (i.e. they cannot be distinguished in the projection), our approach prefers new projections that distinguish the two clusters. In detail, we make the following contributions:

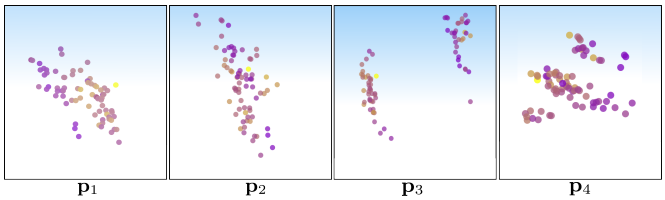

Fig. 1. Four projections of a real high-dimensional benchmark dataset: $\mathbf{p}_1$ and $\mathbf{p}_2$ are similar; $\mathbf{p}_1$ and $\mathbf{p}_3$ are different; $\mathbf{p}_4$ gives new information if $\mathbf{p}_1$ and $\mathbf{p}_3$ are known.

- We introduce a mathematical formulation for a dissimilarity function of a new projection that encodes how much new insight the projection contributes in relation to a certain number of already present projections. Section 3 introduces the measure.
- Based on this, we apply our mathematical approach to propose a greedy approach to find a low number of projections describing the dataset completely. The main idea is to insert new projections with a maximal distance to the projections being already present. We use this to define short and complete data tours. See Section 4.
- We introduce a mathematical approach to interactively explore the data by smoothly changing the projections in such a way that either maximal new insight is gained by a small change of the projection, or that the result of the projection is kept as constant as possible while changing the projection parameters. See Section 5.

We discuss parameters of the approaches and test them on high-dimensional benchmark datasets in Section 6.

2 Related Work

Related work stems from the area of multivariate *projections*, *data tours*, and *quality metrics*.

Affine and Projective Projections: A family of multivariate embeddings have been introduced as RadViz [12, 19, 8, 7] and Star Coordinates [15, 16]. The approaches define a multivariate projection from nD data space to the 2D visualization space. They introduce additional distortion which lead to confusion during a visual search.

Orthographic Projections: The multivariate Orthographic Star Coordinates [17] generalize the concept of bivariate orthographic projections, such as scatterplots. They prevent distortions by maintaining a set of orthography-preserving constraints. However, the mentioned

- *Dirk J. Lehmann, University of Magdeburg, E-mail: dirk@isg.cs.uni-magdeburg.de.*
- *Holger Theisel, University of Magdeburg, E-mail: theisel@ovgu.de*

Manuscript received 31 Mar. 2015; accepted 1 Aug. 2015; date of publication xx xxx 2015; date of current version xx xxx 2015.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

multivariate projections do not consider the data itself, even though the inherent structure of data needs to be considered for the selection of a good projection. We introduce a data-driven strategy for choosing a small set of optimal projections. Regarding this, the *ProjInspector* [20] proposes an interactive exploration technique for a set of basic projections in order to find interesting combinations of them. Our approach does not require an interactive stage to find interesting projections. In addition, the set of projections our approach produces can be utilized as such basic projections, and thus our approach can be well combined with the ProjInspector.

Distance-based Projection Techniques: The Multidimensional scaling (MDS) [27] preserves distances between the data records under projection via the spectrum of a data-dependent centered distance matrix. PCA-based techniques also belong to this family of techniques. With Glimmer [13], a high-performance approach for multilevel MDS on graphic processing units is known. The large amount of distance information required to build up a projection can be reduced by part-linear multidimensional projection (PLMP) [21] to a small number of pairwise distances between a number of representative data samples, which substantially increase performance of the projection process. Local affine multivariate projection (Lamp) [14] provides a local data projection technique by minimizing the distances of the projected data points with the aid of (interactively) initialized seed or control points in the visualization space. Our approach does not optimize data-based distances to find a good projection. Instead, it optimizes a measure between different projections in order to discard affine transformations. In fact, it could be combined with distance-based projection techniques.

Data Tours: A data tour is given by a set of (relevant) projections being a subset of the projection space, which can be investigated by the user for the purpose of visual data analysis. A time sequence of a set of projections is provided for conducting a visual data exploration. The projection pursuit [11, 6] and the grand tour [3] provide a greedy tour of (bivariate) projections, which exponentially grows with the number n of data dimensions. They allow to intuitively detect patterns of interest in the data, but they are time consuming, especially with growing n. Our concept provides a smart tour with a lower and optimal number of projections that is guaranteed to be free of redundancies, but still visits all important views of the data.

Quality Metrics: Their basic idea is to map a quality (correlation, cluster, trends) of a projection onto a real number. With this filtering tool, a set of good projections might be identified. For this, a collection of precomputed projections is rated and the worst ones are rejected. A set of metrics are available and established, such as [28, 22, 24, 1, 2, 26, 23]. We refer to [5] for further details. Quality metrics are useful to find good projections but they have a computational overhead regarding the number of required projections. Clearly, the vast majority of precomputed projections will be rejected. We introduce an alternative concept that avoids the computational overhead of quality metrics.

In the following, we establish a dissimilarity measure for projections.

3 A Dissimilarity Measure for Projections

The n-dimensional dataset is given as m data points $\mathbf{d}_j = (d_{1,j},...,d_{n,j})^T$ for $j = 1,...,m$, resulting in an $n \times m$ data matrix

$$\mathbf{Data} = (\mathbf{d}_1,...,\mathbf{d}_m). \tag{1}$$

In this paper, we restrict ourselves to 2D Star Coordinates, i.e., linear projections that are defined by a $2 \times n$ matrix $\mathbf{A}$. Then the projection of a point $\mathbf{d}_j$ is $\mathbf{A} \cdot \mathbf{d}_j$, and the matrix of all projected points is the $2 \times m$ matrix $\mathbf{A} \cdot \mathbf{Data}$. Note that $\mathbf{A}$ can be interpreted and visualized as the projection of the high-dimensional coordinate axes: for $\mathbf{A} = (\mathbf{x}_1,...,\mathbf{x}_n)$, we have $(\mathbf{x}_i - \mathbf{0}) = \mathbf{A} \cdot \mathbf{i}_i$ where $\mathbf{0}$ is the 2D origin and $\mathbf{i}_i = (\underbrace{0,...,0}_{i-1},1,\underbrace{0,...,0}_{n-i})^T$ is the i^{th} coordinate axis for $i = 1,...,n$.

The projection matrices $\mathbf{A}_1,...,\mathbf{A}_r$ define a number r of projections. To define the dissimilarity of a new $2 \times n$ projection matrix $\mathbf{B}$ to $\mathbf{A}_1,...,\mathbf{A}_r$, we consider an affine transformation of each projection that is given by a 2×2 matrix $\mathbf{Q}_i$ and a translation vector $\mathbf{r}_i$. We define

$$\mathbf{E} = \mathbf{B} \cdot \mathbf{Data} - \frac{1}{r} \sum_{i=1}^{r} (\mathbf{Q}_i \cdot \mathbf{A}_i \cdot \mathbf{Data} + (\underbrace{\mathbf{r}_i,...,\mathbf{r}_i}_{m})) \tag{2}$$

and search for the $\mathbf{Q}_i$ and $\mathbf{r}_i$ that minimize the Frobenius norm of $\mathbf{E}$. This gives the dissimilarity of $\mathbf{B}$ to $\mathbf{A}_1,...,\mathbf{A}_r$:

$$d(\mathbf{B},\mathbf{A}_1,...,\mathbf{A}_r) = \frac{1}{m} \min_{\mathbf{Q}_1,...,\mathbf{Q}_r,\mathbf{r}_1,...,\mathbf{r}_r} \|\mathbf{E}\|_{Fr}^2. \tag{3}$$

Figure 2 illustrates the dissimilarity function for $n = 3, m = 65, r = 1$.

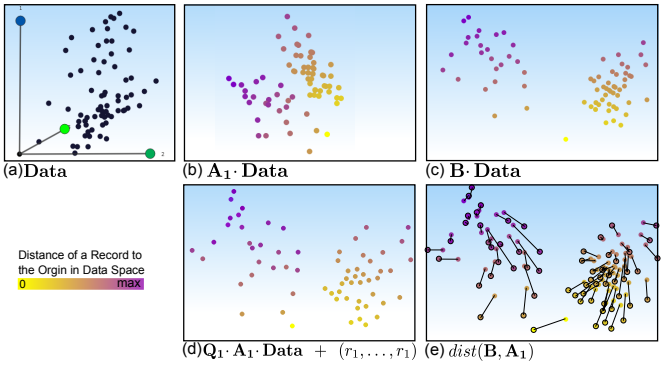

Fig. 2. Dissimilarity function for $n = 3, m = 65, r = 1$; a) n-dimensional dataset **Data**; b) projection by $\mathbf{A}_1$; c) projection by $\mathbf{B}$; d) best affine transformation of projection $\mathbf{A}_1$; e) distance of $\mathbf{B}, \mathbf{A}_1$.

Given $\mathbf{Data}, \mathbf{A}_1,...,\mathbf{A}_r$ and $\mathbf{B}$, (3) is a quadratic minimization problem with the unknowns $\mathbf{Q}_i, \mathbf{r}_i$. To formulate its closed-form solution, we consider the problem in homogenous coordinates:

$$\overline{\mathbf{Data}} = \begin{pmatrix} \mathbf{d}_1 & ... & \mathbf{d}_m \\ 1 & ... & 1 \end{pmatrix}, \overline{\mathbf{A}} = \begin{pmatrix} \mathbf{A}_1 & 0 \\ \vdots & \vdots \\ \mathbf{A}_r & 0 \\ 0...0 & 1 \end{pmatrix}, \overline{\mathbf{B}} = \begin{pmatrix} \mathbf{B} & \mathbf{0} \\ 0...0 & 1 \end{pmatrix}$$

where $\overline{\mathbf{Data}}$ is the homogeneous data matrix, $\overline{\mathbf{A}}$ is a $(2r+1) \times (n+1)$ matrix of all known projection matrices $\mathbf{A}_i$, and $\overline{\mathbf{B}}$ is the new projection matrix in homogenous coordinates. From this we compute a solution of this minimization problem as

$$\overline{\mathbf{D}} = \overline{\mathbf{Data}} \cdot \overline{\mathbf{Data}}^T \tag{4}$$

$$\overline{\mathbf{H}} = \left(\overline{\mathbf{I}} - \overline{\mathbf{D}} \cdot \overline{\mathbf{A}}^T \cdot \left(\overline{\mathbf{A}} \cdot \overline{\mathbf{D}} \cdot \overline{\mathbf{A}}^T \right)^{-1} \cdot \overline{\mathbf{A}} \right) \cdot \overline{\mathbf{Data}} \tag{5}$$

where $\overline{\mathbf{I}}$ is the $(n+1) \times (n+1)$ unit matrix and $\overline{\mathbf{H}}$ is an $(n+1) \times m$ matrix with a vanishing last row. Note that $\left(\overline{\mathbf{A}} \cdot \overline{\mathbf{D}} \cdot \overline{\mathbf{A}}^T \right)$ is a symmetric quadratic $(2r+1) \times (2r+1)$ matrix, depending on r. Since $r < n << m$ usually applies, the calculation of the inverse performs well and is only weakly affected by the curse of dimensionality. Further we get

$$\overline{\mathbf{E}} = \overline{\mathbf{B}} \cdot \overline{\mathbf{H}} \tag{6}$$

where $\overline{\mathbf{E}}$ is a $3 \times m$ matrix with a zero third row. $\overline{\mathbf{E}}$ is the homogenous version of (2) with optimal $\mathbf{Q}_i, \mathbf{r}_i$, i.e.,

$$d(\mathbf{B},\mathbf{A}_1,...,\mathbf{A}_r) = \frac{1}{m} \|\overline{\mathbf{E}}\|_{Fr}^2. \tag{7}$$

The proof of (7) is provided in Appendix 1. The behavior of d under scaling of the projection matrices is given by

$$d(\beta\mathbf{B}, \alpha_1\mathbf{A}_1,..., \alpha_r\mathbf{A}_r) = \beta \; d(\mathbf{B},\mathbf{A}_1,...,\mathbf{A}_r) \tag{8}$$

for any real β and real non-zero α_i. The α_i have no influence because of the discarding of affine transformations, the linear behavior in β is due to the fact that d essentially adds up Euclidean distances of the projected points.

4 Gradient Ascent for Optimal Projections

Based on the dissimilarity measure for projections, we present an algorithm to find a finite (low) number of projections that represent the high-dimensional data best. The main idea is to find projections that have a large dissimilarity to each other. We propose a greedy algorithm: starting with a projection $\mathbf{A}_0$, we repeatedly find new projections $\mathbf{A}_1, \mathbf{A}_2, ...$ until a new projection does not give new insight into the data. Given $\mathbf{A}_0, ..., \mathbf{A}_i$, we search for $\mathbf{A}_{i+1}$ such that it has maximal dissimilarity to $\mathbf{A}_0, ..., \mathbf{A}_i$. For this, we apply a gradient ascent of d:

$$\begin{aligned} \mathbf{B}_0 &= \mathbf{A}_i \\ \mathbf{B}_{j+1} &= \text{orth}(\, \mathbf{B}_j + \lambda \, \nabla_{\mathbf{B}} \, d(\mathbf{B}, \mathbf{A}_1, ...\mathbf{A}_i) \,) \end{aligned} \tag{9}$$

and stop if $\|\mathbf{B}_{j+1} - \mathbf{B}_j\|_{Fr}^2 < \rho$. The convergence parameter ρ, as a numerical parameter, steers the smallest dissimilarity that has to be reached to stop the algorithm. It influences the performance of the ascent and the final number of projections. See Sec. 6.4 for details. Then $\mathbf{A}_{i+1} = \mathbf{B}_{j+1}$. The whole algorithm stops if $\mathbf{A}_0, ..., \mathbf{A}_i$ are complete, i.e., for any new projection $\mathbf{B}$ we have $d(\mathbf{B}, \mathbf{A}_0, ..., \mathbf{A}_i) = 0$. In (9), the function orth() computes a matrix orth($\mathbf{A}$), by applying a Gram-Schmidt orthonormalization to the row vectors of $\mathbf{A}$, which guarantees an orthographic projection of the data to the visualization space [17]. Due to the scaling behavior of d described in (8), it is required to restrict the length of row vectors in $\mathbf{B}_{j+1}$ to one, which is done by this orthonormalization. (6), (7) give that gradient $\nabla_{\mathbf{B}} d$ of d in the variables $\mathbf{B}$ can be computed as

$$\nabla_{\overline{\mathbf{B}}} \, d(\mathbf{B}, \mathbf{A}_1, ..., \mathbf{A}_r) = \frac{2}{m} \; \overline{\mathbf{B}} \cdot \overline{\mathbf{H}} \cdot \overline{\mathbf{H}}^T \tag{10}$$

being a $3 \times (n+1)$ matrix where both the last row and the last column are zero.

Our algorithm has the following parameters: the start projection $\mathbf{A}_0$, the step size λ for the gradient ascent, and the convergence parameter ρ. While choosing $\lambda = 1$, the other parameters are discussed in Section 6.4 and 6.5.

5 Interaction Concepts

We describe an approach for an interactive analysis of a dataset by smoothly changing the projection matrix $\mathbf{A}$. This means that we consider a time-varying projection matrix $\mathbf{A}(t)$ where we use our dissimilarity measure to compute its path from an initial projection $\mathbf{A}(t_0)$ and some user input. For this, we propose two strategies: *maximal deformation* or *minimal deformation* of the projection. For maximal deformation, the path should consist of a sequence of projections that are maximally distant to their neighbors. In other words: for maximal deformation, the projection $\mathbf{A}(t) \cdot \mathbf{Data}$ should have maximal changes under minimal changes of $\mathbf{A}(t)$. Contrary, for minimal deformation, the projection $\mathbf{A}(t) \cdot \mathbf{Data}$ should have minimal changes under maximal changes of $\mathbf{A}(t)$. This strategy aims to provide information on which coordinate axes are dependent on each other. For both strategies, we apply an Euler integration of $\mathbf{A}(t)$:

$$\mathbf{A}(t_{i+1}) = \mathbf{A}(t_i) + (t_{i+1} - t_i) \, \dot{\mathbf{A}}(t_i) \tag{11}$$

where the time derivative $\dot{\mathbf{A}}$ of $\mathbf{A}$ is unknown. For the strategy of maximal deformation, $\dot{\mathbf{A}}(t_i)$ is chosen to maximize $d(\mathbf{A}(t_{i+1}), \mathbf{A}(t_i))$ for $(t_{i+1} - t_i) \to 0$. This is an eigenproblem: setting $r = 1$, we consider the eigenvector $\overline{\mathbf{e}}_{n+1}$ corresponding to the largest eigenvalue of $\overline{\mathbf{H}} \cdot \overline{\mathbf{H}}^T$. Note that $\overline{\mathbf{e}}_{n+1}$ forms both the first and second optimal row of the projection matrix $\mathbf{A}$. Since $\overline{\mathbf{e}}_{n+1}$ is an eigenvector, its length is undefined, gives us two degrees of freedom α, β for scaling $\overline{\mathbf{e}}_{n+1}$ in each row of $\mathbf{A}$. This gives:

$$\dot{\overline{\mathbf{A}}} = (\alpha \, \overline{\mathbf{e}}_{n+1} \; , \; \beta \, \overline{\mathbf{e}}_{n+1} \; , \; \overline{\mathbf{0}}_{n+1})^T \tag{12}$$

where $\overline{\mathbf{0}}_{n+1}$ is the $(n+1)$-dimensional zero vector. Then α, β are subject of user interaction: the user can draw the 2D path of the projection of a coordinate axis $\mathbf{x}_i(t)$ of $\mathbf{A}(t)$ from which we get is tangent $\dot{\mathbf{x}}_i(t)$. This gives the parameters α, β by $\dot{\mathbf{x}}_i = (\overline{\mathbf{e}}_{n+1}).i \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ where $(\overline{\mathbf{e}}_{n+1}).i$ is the i^{th} component of $\overline{\mathbf{e}}_{n+1}$.

For the strategy of minimal deformation, we consider the third-smallest eigenvector $\overline{\mathbf{e}}_3$ of $\overline{\mathbf{H}} \cdot \overline{\mathbf{H}}^T$. Note that $\overline{\mathbf{H}} \cdot \overline{\mathbf{H}}^T$ has at least two vanishing eigenvalues, reflecting the discarding of the affine transformations of the projections. Then we get

$$\dot{\overline{\mathbf{A}}} = (\alpha \, \overline{\mathbf{e}}_3 \; , \; \beta \, \overline{\mathbf{e}}_3 \; , \; \overline{\mathbf{0}}_3)^T \tag{13}$$

with a similar treatment of α, β as above.

6 Experiments

As proof of concept, we present a set of approach-related experiments. Our experiments run on a mobile workstation with a 2.4 GHz 64 Bit Intel CPU with 8 cores, 12 GB RAM, and WIN 7 OS in single-core and single-thread mode.

We introduce the used benchmark data in Sec. 6.1, we illustrate the interaction tool in Sec. 6.2 and the optimal set of projections in Sec. 6.3, which are compared with the commonly used PCA-based approach for visual data exploration. In Sec. 6.4, we investigate the stability of the gradient ascent regarding the influence of convergence parameter ρ and, in Sec. 6.5, regarding the initial projection $\mathbf{A}_0$ (cf. Sec. 4).

6.1 The High-Dimensional Test Datasets

Five high-dimensional test datasets are used from the UCI data base [4]: *Iris* [9], *Yeast* [18], *Wine* [10], *Wdbc* [25], and *Cars* [4]. Table 1 points the data characteristics.

Dataset	Dimensions	Records	Classes
Iris	5	150	3
Yeast	10	1484	10
Wine	14	178	3
Wdbc	32	569	2
Cars	33	7755	52

Table 1. Characteristics of the benchmark test datasets.

In detail, the Fisher's *Iris* plants data base consist of 5 dimensions with 150 records. It gives measurements of the sepal as well as the petal length and width for three iris species. An amount of protein localization sites is given in the *Yeast* dataset, with 10 dimensions and 1484 records. It is usually used to develop probabilistic classifications systems in order to predict properties of proteins. The *Wine* data consist of 14 dimensions with 178 records. It stems from a chemical analysis of three cultivars of wine which have grown in the same Italian region. Thus, wine-specific characteristics are summarized, such as the level of alcohol, the amount of phenols, or the color intensity. The *Wisconsin Diagnostic Breast Cancer* aka *Wdbc* consists of 569 records with 32 quantitative dimensions each. It contains a set of attributes of cell nucleus measurements that are obtained from breast cancer patients. It turned out that a linear separation by a 2D classifier based on the attributes area, texture, and smoothness allows to diagnose benign and malignant cancer cells. The Cars data base contains 33 dimensions and 7755 records. A broad parameter set for different car models is provided, which encompasses attributes, such as the number of cylinders, the maximum velocity, or the power of a car.

Note that a potential a priori classification within the data is not within the focus of our approach or even required. Thus such cases are treated as usual dimensions. Furthermore, to guarantee a fair comparison between outcomes, to avoid numerical influence, and to reduce scaling effects, we linearly normalized the data within the interval $[0, 1]$. Since affine transformations are removed, the normalization does not negatively affect the quality of the optimal set of projections.

6.2 Path-based Interaction

We illustrate the interaction concept of Sec. 5. Figure 3 presents a representative coordinate axis interaction for the *Wine* (Figure 3 (top)) and *Yeast* dataset (Figure 3 (bottom)): A coordinate axis is moved along a (green colored) path to yield time-varying projections $\mathbf{A}(t)$, shown by Figure 3 (middle). We present the projections $\mathbf{A}(t)\cdot\mathbf{Data}$ at time-points $t_i, i = 1,\ldots,4$ for both the minimal deformation (Figure 3 (left)) and maximal deformation case (Figure 3 (right)). It can be seen that the maximal deformation projections are different to each other, reflecting the maximization of the dissimilarity measure during the interaction. In contrast to that, the minimal deformation projections produce similarly shaped outcomes and are similar to the initial projection.

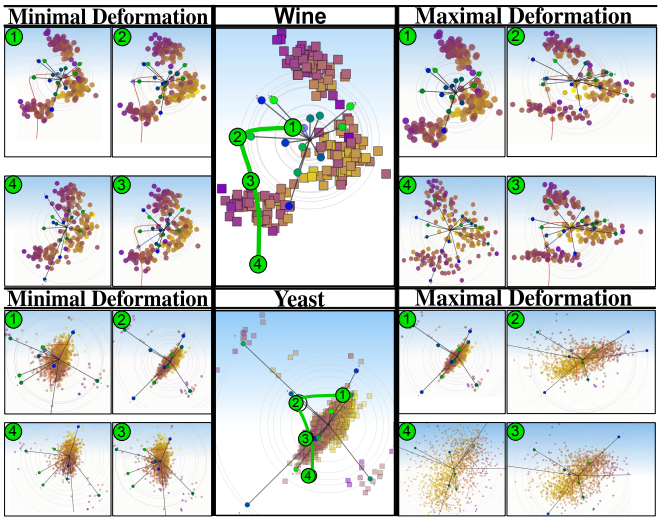

Fig. 3. Minimal (left) and maximal (right) deformation of data patterns during identical interaction in the *Wine* (top) and *Yeast* (bottom) dataset along an interaction path (green).

To preserve minimal or cause maximal dissimilarity between projections might lead to fluctuations of eigenvectors (cf. Sec. 5), even though the function of eigenvalues itself is smooth over the interaction. This effect is caused by data characteristics and might lead to jitter of $\mathbf{A}(t)$. Thus, it provides additional structural data insight. In the following, we construct optimal sets of projections of the test data.

6.3 Optimal Set of Projections for Test Data

The gradient ascent of Sec. 4 is applied to the test data. For this, the question of an appropriate initial projection arises: An established initial standard configuration $\mathbf{A}^\pi$ of a multivariate projection is the radial layout [17, 15], given by

$$\mathbf{A}^\pi = \begin{pmatrix} x_0,\ldots,x_{n-1} \\ y_0,\ldots,y_{n-1} \end{pmatrix} \text{ with } (x_i,y_i)^T = b\cdot(\sin(i\cdot\alpha),\cos(i\cdot\alpha))^T$$

and $i = 0,\ldots,n-1$ whereas $\alpha = \frac{2\cdot\pi}{n}$ and $b = \sqrt{2/n}$. Following [17], a construction scheme of an orthonormalization (cf. Sec. 4) is given by using a radius $b = \sqrt{2/n}$, meaning that $\mathbf{A}^\pi$ becomes an orthographic projection. It is an appropriate candidate for the initial projection $\mathbf{A}_0$ of our gradient descent. Thus, our approach defines $\mathbf{A}_0 = \mathbf{A}^\pi$ as the initial projection for the gradient ascent with the convergence parameter $\rho = 0.1$.

The Figures 4 and 5 illustrate the optimal set of projections produced by our approach (top) in comparison to the same number of best PCA-based projections (bottom), w.r.t. our benchmark datasets (cf. Sec 6.1). PCA is given by the eigenvectors $\mathbf{e}_i, i = 1,\ldots,n$ of the data's covariance matrix, which minimizes correlation and maximizes variance. Pairwise eigenvectors define a $2\times n$ projection $\mathbf{A}_{ij} = \mathbf{e}_{i/j} = (\mathbf{e}_i,\mathbf{e}_j)^T$. Note that the complete number of PCA projections grows quadratically in the dimension number n, while the number of our set of optimal projections grows linear in n. For instance, the *Wine* dataset with $n = 14$ dimensions has a total number of 91 PCA-based projections (which can be found in the supplemental material), while our optimal set only requires 7 projections. However, in order to provide a fair comparison that reflects the use of PCA in practice, a subset of the largest pairwise eigenvalues is presented for each case, which has the same number of projections as our optimal set.

For our optimal sets in the figures, the annotated dissimilarity label d of a projection $\mathbf{A}_i$ describes the dissimilarity to the subset of predecessor projections $\{\mathbf{A}_o,\ldots,\mathbf{A}_{i-1}\}$ referring to (7) as $d(\mathbf{B},\mathbf{A}_o,\ldots,\mathbf{A}_{i-1}) = d(\mathbf{A}_i)$ with $\mathbf{B} = \mathbf{A}_i$. Consequently, we treat the PCA projections similarly: the dissimilarity label d of a PCA projection $\mathbf{e}_{i/j}$ also describes the dissimilarity to the subset of predecessor PCA-based projections. This comparison setup facilitates an empirical comparison of the dissimilarity behavior for both techniques. Keep in mind that a larger dissimilarity means that more data insight is given with a certain projection. Finally, the dissimilarity behavior is summarized by a graph at the end of each projection sequence for each dataset and projection technique.

For our optimal sets, it can be seen that the dissimilarity rapidly decreases with growing index i, i.e., $d(\mathbf{A}_i) > d(\mathbf{A}_{i+1})$ with $i \geq 1,\ldots r-1$. This appears to be plausible, since the degree of freedom to find a projection that cannot be generated as affine transformation gets small if a sufficient number of projections is available. In fact, only the first two, three or occasionally four projections of the optimal projection set show relevant data patterns. Clearly, stopping the ascent in early stages would still lead to projections showing the most important patterns. Beyond that, our experiments illustrate that the number r of projections is optimal with $r \leq \frac{n}{2}$.

In comparison to that, each new PCA-based projection only provides little additional insight compared to the first PCA projection $\mathbf{e}_{1/2}$ for each case: The dissimilarity values are much smaller and almost negligible compared to those of our optimal set of projections. On the other hand, relevant patterns that are shown by the PCA-based projections can also be seen in the set of optimal projections. Thus, our experiments empirically illustrate the advantages of our optimal set of projections compared to PCA.

6.4 Influence of Convergence Parameter

The convergence parameter ρ influences both the dissimilarity between successively selected projections during the gradient ascent and the algorithm's performance and convergence behavior. In fact, a too large value of parameter ρ would cause projections that have a small dissimilarity to each other. Thus, the parameter should be rather small in order to facilitate projections that have a large dissimilarity and thus provide new data insights. On the other hand, if ρ is chosen too small, then the algorithm's performance decreases. In order to find a good choice of the convergence parameter ρ, we investigate the algorithm's behavior for a set of small values, such as $\rho = 0.1$, $\rho = 0.01$, and $\rho = 0.001$ (with $\mathbf{A}_0 = \mathbf{A}^\pi$). Figure 6 illustrates the results.

Intra Set Differences: Figure 6 (left) shows column-wise the intra set differences $d(\mathbf{A}_i)$ of the projections $\mathbf{A}_i, i = 1,\ldots,r$ for each value of ρ and row-wise for the test data. It can be seen that the patterns are comparable and only weakly dependent on ρ. Furthermore, the calculation time grows approximately logarithmically in ρ.

Inter Set Differences: We are interested in a comparison of the projections with the same index i but different values in ρ: Be $\mathbf{A}_\rho = \{\mathbf{A}_0,\ldots,\mathbf{A}_r\}$ the set of projections w.r.t. ρ, and be $\mathbf{A}_\rho(i) = \mathbf{A}_i$ a projection of it, then we define the inter set difference $d(i,\rho_k,\rho_l) = d(\mathbf{A}_{\rho_k}(i),\mathbf{A}_{\rho_l}(i))$ as the dissimilarity between projections with the same index in different sets that are based on different values of ρ. Figure 6 (right) shows the inter set differences $d(i,\rho_k,\rho_l)$ with $i = 1,\ldots,r$. The differences behave quite stable and they are just weakly dependent on the accuracy of ρ.

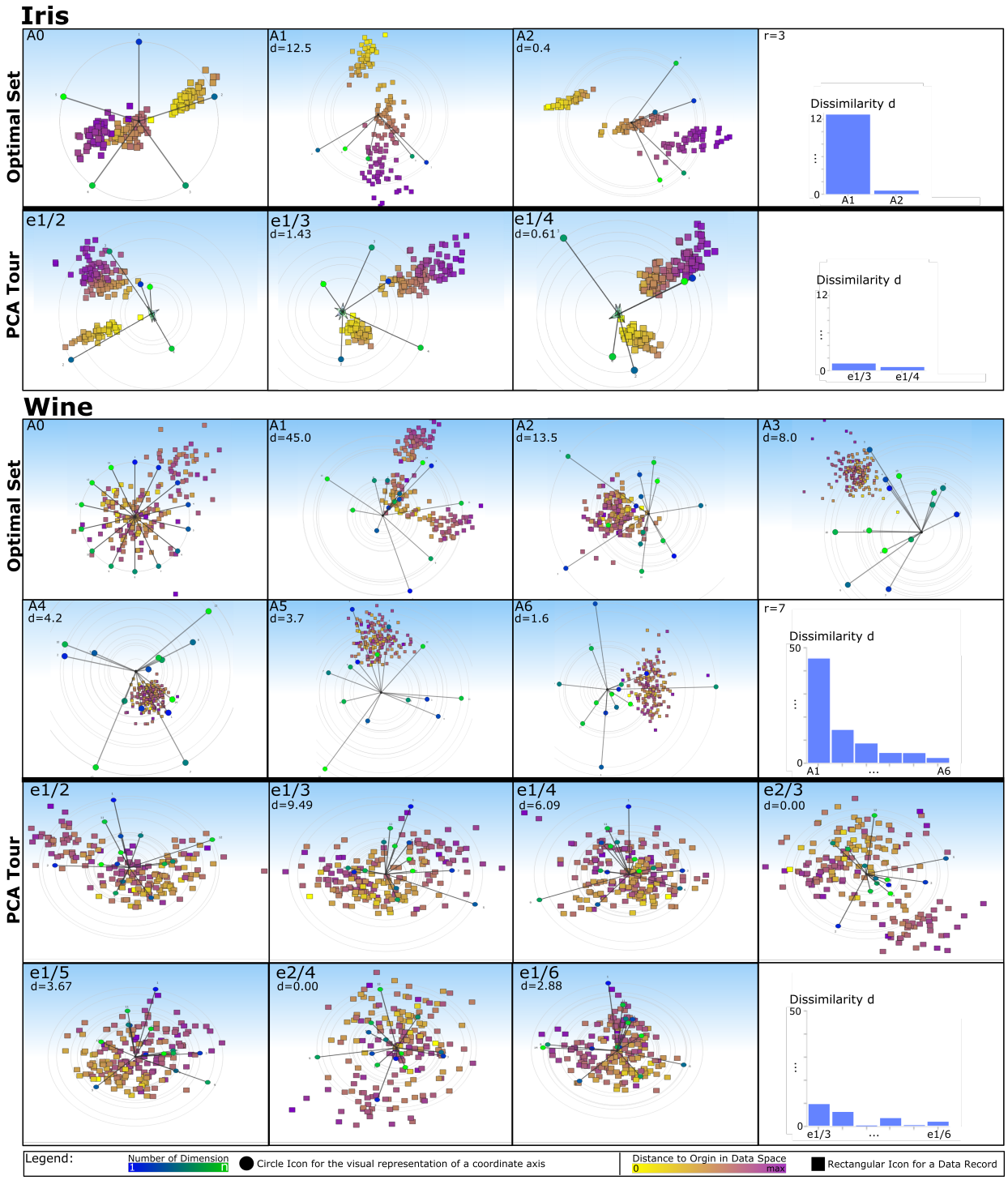

Fig. 4. Optimal set of projections compared to a subset of projections of the PCA tour for the test data *Iris* (top) and *Wine* (bottom): For each successor projection, the dissimilarity d is labeled with respect to the amount of all predecessor projections of the same sequence, which is graphically summarized at the end of each sequence.

Intra vs. Inter Set Differences: Be d_{Intra} the maximal intra set difference and be d_{Inter} the maximal inter set difference, then we get the pairs (d_{Intra}, d_{Inter}) for *Iris* as $(12.5, 0)$, for *Wine* as $(45, 2.3)$, for *Wdbc* as $(338, 12.5)$, and for *Cars* as $(2079, 243)$. It follows that the observed inter set differences that are caused by a coarser accuracy of ρ are rather small and negligible compared to the dominant intra set differences. Finally our experiments shows that a convenient choice of convergence parameter is $\rho = 0.1$.

6.5 Influence of Initial Projection

In this section, we investigate the influence of the initial projection $\mathbf{A}_0$. For this, we conduct the gradient ascent with noisy versions of the initial projection $\mathbf{A}_0 = \mathbf{A}^\pi$: be $\mathbf{R}_l$ a $2 \times n$ matrix with randomly chosen column vectors that have a p2-norm of l each, then a noisy version $\mathbf{A}_l^\pi$ is given by $\mathbf{A}_l^\pi = \mathbf{A}^\pi + \mathbf{R}_l$. The larger l the noisier $\mathbf{A}^\pi$ becomes. We did 100 runs of the ascent for each value $l = 0.2, 0.4, \ldots, 1.2$ and being started with $\mathbf{A}_l^\pi$.

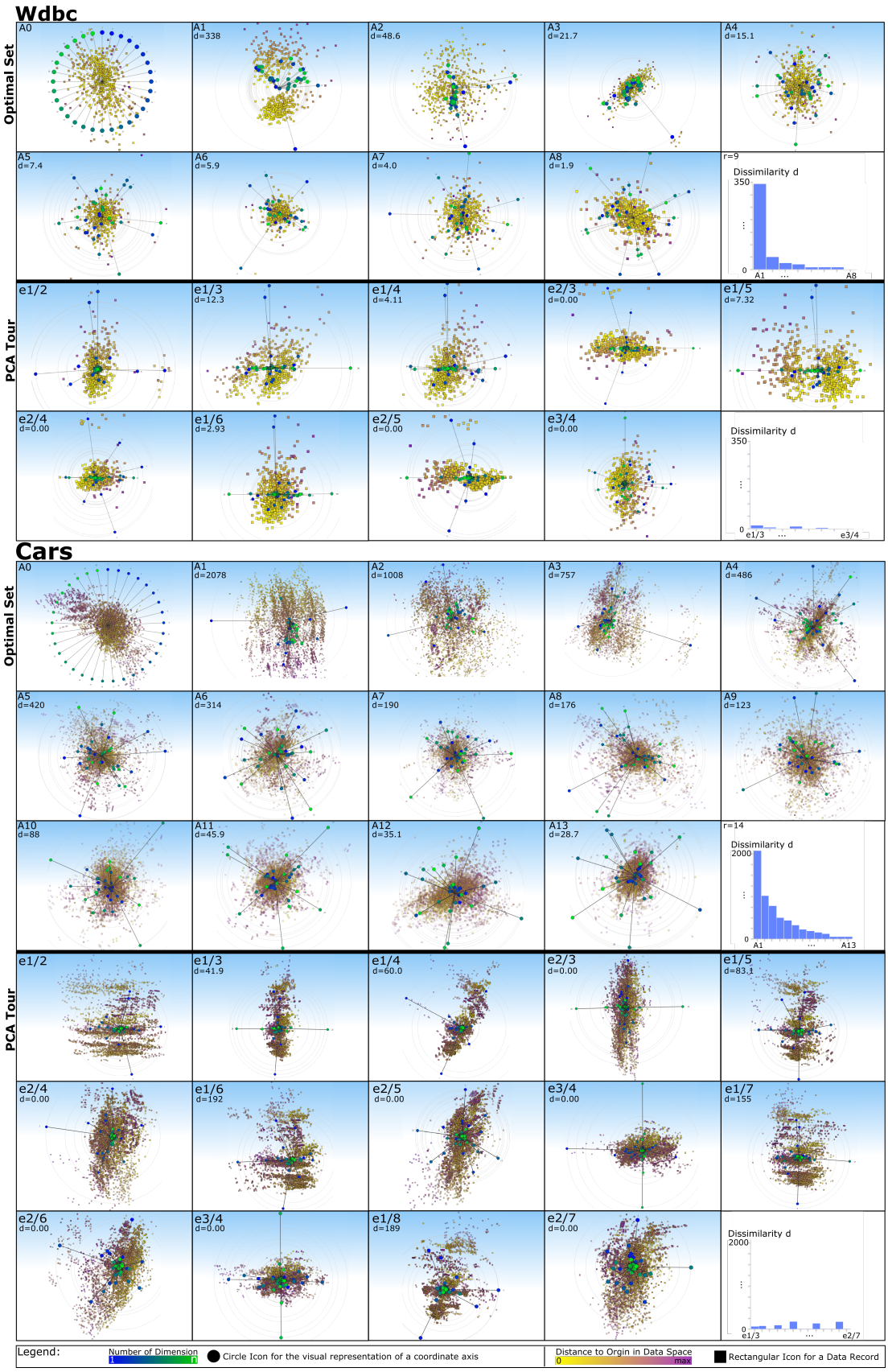

Fig. 5. Optimal set of projections compared to a subset of projections of the PCA tour for the test data *Wdbc* (top) and *Cars* (bottom): For each successor projection, the dissimilarity d is labeled with respect to the amount of all predecessor projections of the same sequence, which is graphically summarized at the end of each sequence.

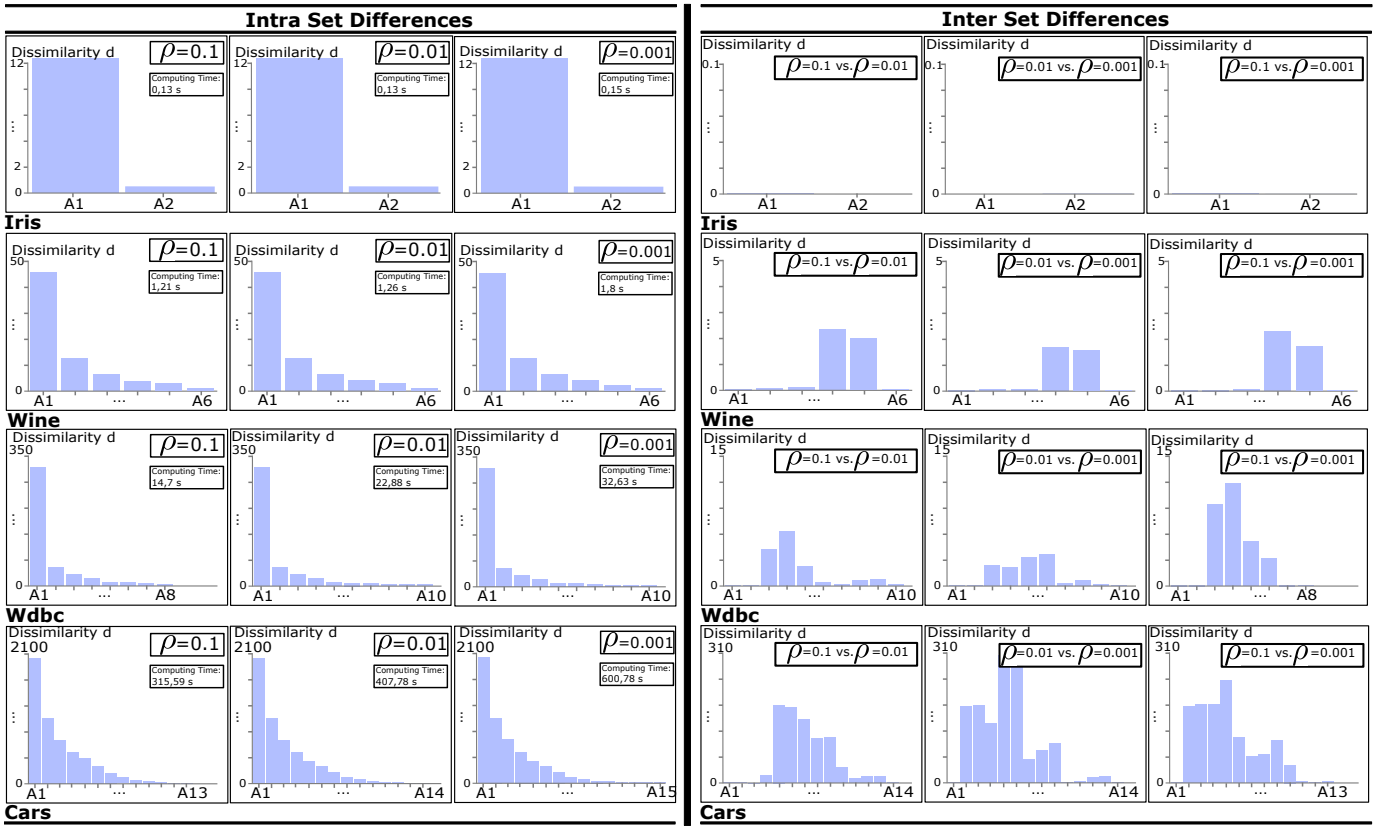

Fig. 6. Influence of the convergence parameter ρ for the parametrization $\rho = 0.1$, $\rho = 0.01$, and $\rho = 0.001$ on the behavior of the dissimilarity d corresponding to the resulting optimal set of projections: (left) Intra Set and (right) Inter Set Differences for the test data.

Be $d(\mathbf{A}_i, l)$ the dissimilarity to the subset of predecessor projections $\{\mathbf{A}_l^\pi, \mathbf{A}_1, \ldots, \mathbf{A}_{i-1}\}$ for such a set of projections that results if the ascent starts with $\mathbf{A}_l^\pi$. We define the inter set dissimilarity $d(\mathbf{A}_i, l, 0) = |d(\mathbf{A}_i, l) - d(\mathbf{A}_i, 0)|$ of projection $\mathbf{A}_i$ to the projections with the same index i that result by starting the ascent at $\mathbf{A}^\pi$. For a projection $\mathbf{A}_i$ over all runs $j = 1, \ldots, 100$ with the same l, we stored its mean inter set dissimilarity $d_\mu(\mathbf{A}_i, l) = \frac{1}{100} \sum_{j=1}^{100} d_j(\mathbf{A}_i, l, 0)$ as well as the maximum/minimum dissimilarity $d_{\min}(\mathbf{A}_i, l, 0)$/ $d_{\max}(\mathbf{A}_i, l, 0)$.

Figure 7 illustrates the results. It can be seen that the mean inter set dissimilarity $d_\mu(\mathbf{A}_i, l)$ is stable: an unstable behavior could be recognized by an exponential growth in $d_\mu(\mathbf{A}_i, l)$. However, this is not what we observe. Instead, we observe a converging behavior against similar projections even though the initial projection becomes diffuse. Figure 8 (a) illustrates this ability in detail for randomly chosen runs of the *Wine* dataset for different l: the first five projections of the optimal set can be seen each. Especially the different projections $\mathbf{A}_1$ stably show the main pattern in that data, which is shaped like a rotated version of the letter 'u', independently to the start projection $\mathbf{A}_0$.

We are also interested in the behavior if the initial projection is randomly chosen, for instance by a user-based interaction. Regarding this, Figure 8 (b) row-wise illustrates the first five projections of the optimal set with respect to three randomly chosen inital projections $\mathbf{A}_0$ of the *Wine* data. Interestingly, prominent data patterns, such as the 'u' pattern in $\mathbf{A}_1$, are still visited in each case. In fact, our experiments show that relevant data patterns are found independently of the chosen start projection $\mathbf{A}_0$ (please find further examples in the supplemental material).

7 Discussion and Limitations

Improvements, advantages, and limitations of our approach will be discussed in this section.

Why discarding affine transformations?: In Information Visualization, depending on the application and the dataset, different goals of a visual analysis are possible. Among them there are universal goals that are relevant to all applications: the segmentation of the data points into meaningful clusters. While there is a large amount of cluster definitions for an automatic clustering, *visual clustering* [29] has been established as an interesting alternative, i.e., an interactive process where the user manually marks the clusters in appropriate projections. This has the advantage that no prior knowledge about shape or properties of the clusters are necessary. We argue that for a visual clustering, affine transformations of the projections are of less relevance: regions that are clearly visually distinguishable remain distinguishable after an affine transformation. Please note that the human perception system is not rotationally invariant, and thus the setup for the presentation of projections influences the users' capability to recognize important structures. Nevertheless, it is paramount to have such a set of projections available that mutually bear the most structural information regarding the data, which is the focus of this work. Then, to ask for a well designed presentation in order to show the set of projections to the user is not within this paper's focus.

Relation to quality metrics: Quality metrics measure the quality of a single projection. In contrast, our approach measures the quality of a projection relative to a number of already present projections. In this sense, our approach is orthogonal to existing quality metrics. In fact, they can be used as starting point of our approach.

Dependence on the starting projection: Our approach is parametrized by a start projection $\mathbf{A}_0$ (cf. Sec. 6.5). Even though the choice of this projection influences the set of optimal projections, the relevant data patterns, or variations, stably remain visible. Hence, the final result regarding a visual search is less dependent on $\mathbf{A}_0$.

Completeness of sets of projections: Given a dataset with $\text{rank}(\overline{\mathbf{D}}) = n+1$, the space of all Star Coordinates under discarding of affine transformations is completely described by $n_b = \text{abs}(\frac{n}{2})$ linear independent projections: we can find n_b linear independent projections $\mathbf{A}_1, ..., \mathbf{A}_{n_b}$ with $d(\mathbf{A}_j, \mathbf{A}_1, ..., \mathbf{A}_{j-1}, \mathbf{A}_{j+1}, ..., \mathbf{A}_{n_b}) > 0$ for $j = 1, ..., n_b$, and $d(\mathbf{B}, \mathbf{A}_1, ..., \mathbf{A}_{n_b}) = 0$ for any projection matrix $\mathbf{B}$. The value of $n/2$ intended projections can be explained as follows: the space of all projections is $2n$-dimensional, since a projection matrix $\mathbf{A}_i$ consist of $2n$ independent entries. By discarding affine transformations, each

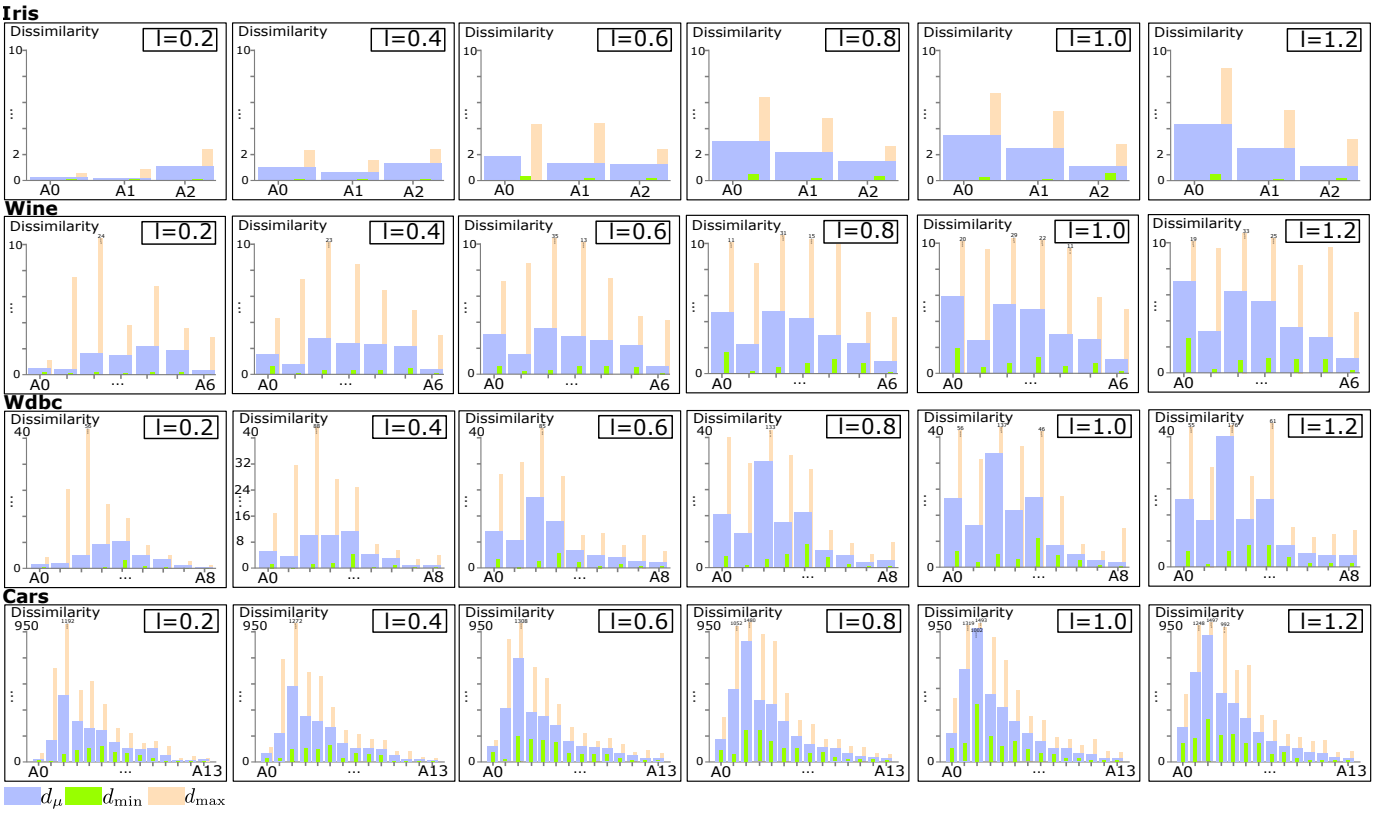

Fig. 7. Influence of varying the start projection $\mathbf{A}_0$ on the dissimilarity corresponding to the resulting optimal set of projections, which is statistically measured by the mean inter set dissimilarity d_μ and the maximum/minimum dissimilarity $d_{\min}/d_{max}$ for the test data.

projection matrix $\mathbf{A}_i$ loses 4 degrees of freedom. A matrix $\mathbf{A}_i$ with all its affine transformations forms a 4-dimensional subspace of the space of all transformation matrices. Hence, $n/2$ projection matrices with its affine transformations are enough to cover the transformation space.

Relation to MDS and PCA: MDS usually provides one as distance-preserving as possible projection, from nD to 2D space. Applying an affine transformation on a MDS $\mathbf{M}$, such as rotation and scaling, yield an equivalent or identical MDS configuration. Following (7), two MDS configurations $\mathbf{M}_i$ and $\mathbf{M}_j$ are identical, i.e. they convey the same information, if $d(\mathbf{M}_i,\mathbf{M}_j)=0$, meaning they can be mapped to each other by affine transformations. In addition, PCA provides a set of partly relevant projections. It contains relevant projections but also a number of them that lead to visual noise. See an example of this behavior for the *Wine* dataset in the supplemental material. The ratio of relevant projections is higher with our approach, meaning that our set is less repetitive, making a user-based visual search more feasible. This is reflected by the fact that the number of projections grows quadratically in n for PCA, for our optimal set it grows linearly in n. Moreover, PCA requires Gaussian-distributed data to perform optimally, otherwise relevant data patterns might be undetectable. Our approach does not have such a requirement. Lastly, PCA performs differently if different data normalization approaches are used. Since affine transformations do not affect the result of our approach, our optimal sets behave more stably regarding data normalization.

8 Conclusion

We provided a novel approach to measure the dissimilarity of multivariate projections disregarding affine transformations. It is based on the idea that a new projection in a tour should have a large dissimilarity to all projections that were already presented, in order to ensure the presentation of new data insights. Based on this measure, a small set of optimal projections is automatically selected by our approach. It makes a projection-based visual search more feasible for a user, since the number of projections is restricted to $n/2$. For the future, we are interested in the investigation of further measures that can be applied to a number of projections. For instance, to automatically detect a number of prominent projections which optimally describe the data.

Appendix 1:

Proof that (4)-(7) is the solution of the minimization problem (3): Defining $\overline{\mathbf{Q}}_i = \begin{pmatrix} \mathbf{Q}_i & \mathbf{r}_i \\ 0\ 0 & 1 \end{pmatrix}$ for $i=1,...,r$, we can write (2) in homogeneous coordinates

$$\overline{\mathbf{E}} = \overline{\mathbf{B}} \cdot \overline{\mathbf{Data}} - \frac{1}{r} \sum_{i=1}^{r} \overline{\mathbf{Q}}_i \cdot \overline{\mathbf{A}}_i \cdot \overline{\mathbf{Data}}. \tag{14}$$

Note that $\mathbf{E}$ in (2) and $\overline{\mathbf{E}}$ in (14) are identical except for an additional zero row in $\overline{\mathbf{E}}$. Introducing the $3 \times (2r+1)$ matrix of all unknown affine transformation parameters

$$\overline{\mathbf{X}} = \begin{pmatrix} \mathbf{Q}_1 & \dots & \mathbf{Q}_r & \mathbf{r}_1 + \dots + \mathbf{r}_r \\ 0\ 0 & \dots & 0\ 0 & r \end{pmatrix}, \tag{15}$$

(14) can be written as

$$\overline{\mathbf{E}} = \overline{\mathbf{B}} \cdot \overline{\mathbf{Data}} - \frac{1}{r} \overline{\mathbf{X}} \cdot \overline{\mathbf{A}} \cdot \overline{\mathbf{Data}}. \tag{16}$$

Then the condition for $\overline{\mathbf{X}}$ to minimize $\|\overline{\mathbf{E}}\|_{Fr}^2$ is

$$\overline{\mathbf{B}} \cdot \overline{\mathbf{Data}} \cdot (\overline{\mathbf{A}} \cdot \overline{\mathbf{Data}})^T = \frac{1}{r} \overline{\mathbf{X}} \cdot \overline{\mathbf{A}} \cdot \overline{\mathbf{Data}} \cdot (\overline{\mathbf{A}} \cdot \overline{\mathbf{Data}})^T \tag{17}$$

which can be solved to

$$\overline{\mathbf{X}} = r\, \overline{\mathbf{B}} \cdot \overline{\mathbf{D}} \cdot \overline{\mathbf{A}}^T \cdot (\overline{\mathbf{A}} \cdot \overline{\mathbf{D}} \cdot \overline{\mathbf{A}}^T)^{-1}. \tag{18}$$

Inserting this into (16) gives (6) with (4) and (5).

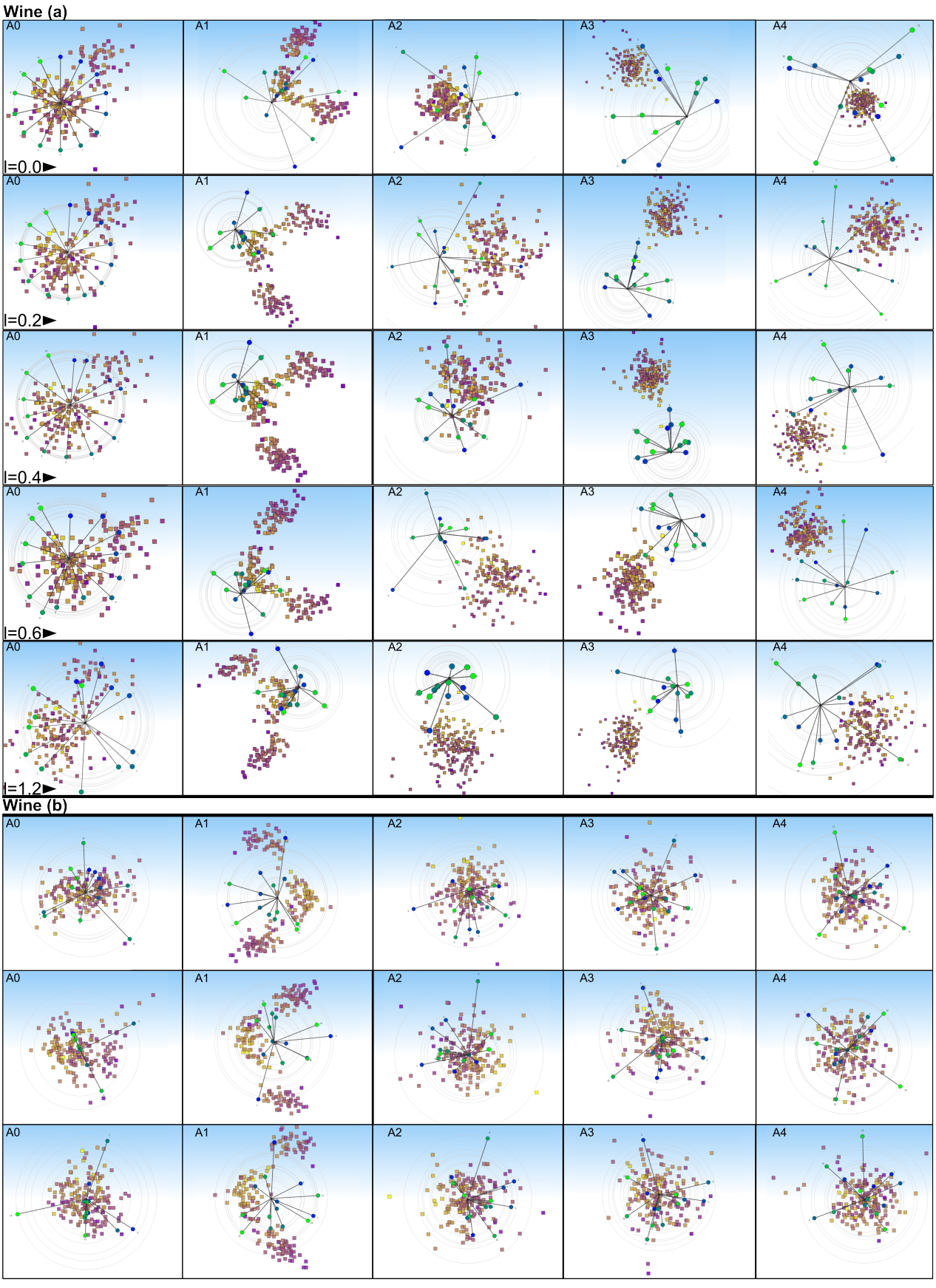

Fig. 8. Case study: (a) Influence of noise to the initial radial configuration and the resulting set of optimal projections for the wine data.(b) Random walk for the *Wine* dataset: the inital projections $\mathbf{A}_0$ are randomly chosen.

REFERENCES

[1] G. Albuquerque, M. Eisemann, D. J. Lehmann, H. Theisel, and M. Magnor. Improving the Visual Analysis of High-dimensional Datasets Using Quality Measures. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (IEEE VAST)*, pages 19–26, 2010.

[2] G. Albuquerque, M. Eisemann, and M. A. Magnor. Perception-based Visual Quality Measures. *In Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (IEEE VAST)*, pages 13–20, 2011.

[3] D. Asimov. The Grand Tour: a Tool for Viewing Multidimensional Data. *Journal on Scientific and Statistical Computing*, 6(1):128–143, 1985.

[4] A. Asuncion and D. Newman. UCI Machine Learning Repository, 2007.

[5] E. Bertini. Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization. *IEEE Transactions on Visualization and Computer Graphics*, 17:2203–2212, 2011.

[6] D. Cook, A. Buja, J. Cabreta, and C. Hurley. Grand Tour and Projection Pursuit. *Journal of Computational and Statistical Computing*, 4(3):155–172, 1995.

[7] K. M. Daniels, G. G. Grinstein, A. Russell, and M. Glidden. Properties of normalized radial visualizations. *IEEE Information Visualization*, 11(4):273–300, 2012.

[8] L. Di Caro, V. Frias-Martinez, and E. Frias-Martinez. Analyzing the Role of Dimension Arrangement for Data Visualization in RadViz. In *Proceedings of the 14th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part II*, PAKDD'10, pages 125–132, Berlin, Heidelberg, 2010. Springer-Verlag.

[9] R. A. Fisher. The use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, pages 179 – 188, 1936.

[10] M. Forina. PARVUS - An Extendible Package for Data Exploration, Classification and Correlation., Institute of Pharmaceutical and Food Analysis and Technologies, Genoa, Italy.

[11] J. H. Friedman and J. W. Tukey. A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers*, 23:881–890, September 1974.

[12] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, and E. Stanley. DNA Visual and Analytic Data Mining. In *Proceedings of the 8th Conference on Visualization*, pages 437 – 441, Los Alamitos, CA, USA, 1997. IEEE Computer Society Press.

[13] S. Ingram, T. Munzner, and M. Olano. Glimmer: Multilevel MDS on the GPU. *IEEE Transactions on Visualization and Computer Graphics*, 15(2):249–261, 2009.

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

[15] E. Kandogan. Star Coordinates: A Multi-Dimensional Visualization Technique with Uniform Treatment of Dimensions. *In Proceedings of the IEEE Information Visualization Symposium*, pages 9–12, 2000.

[16] E. Kandogan. Visualizing Multi-Dimensional Clusters, Trends, and Outliers Using Star Coordinates. *In Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 107 – 116, 2001.

[17] D. J. Lehmann and H. Theisel. Orthographic Star Coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2615–2624, 2013.

[18] K. Nakai and M. Kanehisa. Expert System for Predicting Protein Localization Sites in Gram-Negative Bacteria. *Proteins: Structure, Function and Genetics*, 11(2):95–110, 1991.

[19] L. Nováková and O. Štepánková. Multidimensional Clusters in RadViz. In *Proceedings of the 6th WSEAS International Conference on Simulation, Modelling and Optimization*, pages 470–475, Stevens Point, Wisconsin, USA, 2006. World Scientific and Engineering Academy and Society (WSEAS).

[20] P. Pagliosa, F. V. Paulovich, R. Minghim, H. Levkowitz, and L. G. Nonato. Projection Inspector: Assessment and Synthesis of Multidimensional Projections. *Neurocomputing*, 150, Part B(0):599 – 610, 2015.

[21] F. V. Paulovich, C. T. Silva, and L. G. Nonato. Two-Phase Mapping for Projecting Massive Data Sets. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1281–1290, Nov. 2010.

[22] J. Schneidewind, M. Sips, and D. Keim. Pixnostics: Towards Measuring the Value of Visualization. *Symposium On Visual Analytics Science And Technology*, pages 199–206, 2006.

[23] L. Shao, M. Behrisch, T. Schreck, T. von Landesberger, M. Scherer, S. Bremm, and D. A. Keim. Guided Sketching for Visual Search and Exploration in Large Scatter Plot Spaces. *EuroVA International Workshop on Visual Analytics*, 2014.

[24] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting Good Views of High-dimensional Data using Class Consistency. *Computer Graphics Forum (Proc. EuroVis 2009)*, 28(3):831–838, 2009.

[25] W. Street, W. Wolberg, and O. Mangasarian. Nuclear Feature Extraction for Breast Tumor Diagnosis. *SPIE International Symposium on Electronic Imaging: Science and Technology*, 1993.

[26] A. Tatu, G. Albuquerque, M. Eisemann, P. Bak, H. Theisel, M. Magnor, and D. Keim. Automated Analytical Methods to Support Visual Exploration of High-Dimensional Data. *IEEE Transactions on Visualization and Computer Graphics*, 17:584–597, 2011.

[27] W. S. Torgerson. Multidimensional Scaling: I. Theory and Method. *Psychometrika*, 17:401–419, 1952.

[28] L. Wilkinson, A. Anand, and R. Grossman. Graph-Theoretic Scagnostics. *IEEE Symposium on Information Visualization*, pages 157–164, 2005.

[29] H. Zhou, X. Yuan, H. Qu, W. Cui, and B. Chen. Visual Clustering in Parallel Coordinates. *Computer Graphics Forum*, (27):1047 – 1054, 2008.