Information Visualization

High-Dimensional Data

Dr. David Koop
Schedule

• Today: High-Dimensional Data Lecture
• Tuesday, Oct. 26: No Class
• Thursday, Oct. 28: High-Dimensional Data Critique Due
What techniques might you use for high-dimensional data visualization?
High-Dimensional Data Visualization Techniques

- Scatterplot Matrix (SPLOM)
- Parallel Coordinates Plot (PCP)
- Heatmap
- Interactive Elements:
  - Brushing (Linked Highlighting)
  - Tooltips
Scatterplot Matrix

• Each pair of quantitative attributes has its own plot
Parallel Coordinate Plots

- Use multiple parallel axes, one for each dimension
- Each data item is encoded as a line mark
- Positive and negative correlation can be seen in these plots...
- …but ordering becomes important

[Munzner (ill. Maguire), 2014]
Brushing
Our High-Dimensional Data Focus

• Projection Understanding
• Tours
Dimensionality Reduction

• Attribute Aggregation: Use fewer attributes (dimensions) to represent items
• Combine attributes in a way that is more instructive than examining each individual attribute
• Example: Understanding the language in a collection of books
  - Count the occurrence of each non-common word in each book
  - Huge set of features (attributes), want to represent each with an aggregate feature (e.g. high use of "cowboy", lower use of "city") that allows clustering (e.g. "western")
  - Don't want to have to manually determine such rules
• Techniques: Principle Component Analysis, Multidimensional Scaling family of techniques
Principle Component Analysis (PCA)

[Image of PCA diagram showing original data space and component space with PC 1 and PC 2 axes]

[Text: Principle Component Analysis (PCA)]

[M. Scholz, CC-BY-SA 2.0]
Principal Component Analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. It's often used to make data easy to explore and visualize.

First, consider a dataset in only two dimensions, like (height, weight). This dataset can be plotted as points in a plane. But if we want to tease out variation, PCA finds a new coordinate system in which every point has a new (x, y) value. The axes don't actually mean anything physical; they're combinations of height and weight called “principal components” that are chosen to give one axes lots of variation.

Drag the points around in the following visualization to see PC coordinate system adjusts.

17 dimensions to 2

Here's the plot of the data along the first principal component. Already we can see something is different about Northern Ireland.

Now, see the first and second principal components, we see Northern Ireland a major outlier. Once we go back and look at the data in the table, this makes sense: the Northern Irish eat way more grams of fresh potatoes and way fewer of fresh fruits, cheese, fish and confectionary.

For more explanations, visit the Explained Visually project homepage.

Non-linear Dimensionality Reduction

\[ \Phi_{gen} : \mathcal{Z} \rightarrow \mathcal{X} \]

\[ \Phi_{extr} : \mathcal{X} \rightarrow \mathcal{Z} \]

original data space \( \mathcal{X} \)

component space \( \mathcal{Z} \)
Dimensionality Reduction in Visualization

5.2.2 Layout Quality

Fig. 8 shows the visual quality, normalized stress, and timing of Glimmer, Hybrid, and PivotMDS layouts on four data sets with known structure. In the case of grid, the correct shape is known. In the other three cases, the correct partitions of the points into clusters are available with these benchmark data sets, so the extent to which the color coding matches the spatial grouping created by an algorithm is a measure of its accuracy.

Qualitatively, with cancer, the Glimmer and PivotMDS algorithms indicate these two color-coded groups clearly with spatial position. Quantitatively, the stress of Glimmer is an order of magnitude lower than that of PivotMDS. Hybrid does separate the two groups but produces misleading subclusters in the orange group.

With shuttle_big, Hybrid produces a readable layout separating the red cluster from the other two but is slower by several hundred percent. Glimmer and PivotMDS both produce useful and qualitatively comparable layouts separating the clusters. The PivotMDS layout is twice as fast but has noticeable occlusion and much higher stress than the Glimmer layout.

The 10,000-point grid is accurately embedded by Glimmer and PivotMDS in comparable times. Hybrid is again slower but nevertheless terminated too soon, suffering from very noticeable qualitative distortion and with a much higher quantitative stress metric compared to that of the other layouts.

The Glimmer layout of the docs data set is qualitatively better than the other three. It shows several spatially distinguishable clusters, color coded by blue, red, orange, and green. The green cluster is split into three parts. It took approximately 2 seconds with normalized stress of 0.157. Hybrid suffers from cluster occlusion. The stress is nearly twice as high as that of Glimmer, and the spatial embedding does not clearly separate any of the given clusters. PivotMDS is very fast but almost completely fails to show
Tasks in Understanding High-Dim. Data

**Task 1**

- **In**: HD data
- **Out**: 2D data

**What?**
- In High-dimensional data
- Out 2D data

**Why?**
- Produce
- Derive

**Task 2**

- **In**: 2D data
- **Out**: Scatterplot Clusters & points

**What?**
- In 2D data
- Out Scatterplot Clusters & points

**Why?**
- Discover
- Explore
- Identify

**How?**
- Encode
- Navigate
- Select

**Task 3**

- **In**: Scatterplot Clusters & points
- **Out**: Labels for clusters

**What?**
- In Scatterplot Clusters & points
- Out Labels for clusters

**Why?**
- Produce
- Annotate

---

[Munzner (ill. Maguire), 2014]
Probing Projections: Interaction Techniques for Interpreting Arrangements and Errors of Dimensionality Reductions

Julian Stahnke, Marian Dörk, Boris Müller, and Andreas Thom

Abstract
—We introduce a set of integrated interaction techniques to interpret and interrogate dimensionality-reduced data. Projection techniques generally aim to make a high-dimensional information space visible in form of a planar layout. However, the meaning of the resulting data projections can be hard to grasp. It is seldom clear why elements are placed far apart or close together and the inevitable approximation errors of any projection technique are not exposed to the viewer. Previous research on dimensionality reduction focuses on the efficient generation of data projections, interactive customisation of the model, and comparison of different projection techniques. There has been only little research on how the visualization resulting from data projection is interacted with. We contribute the concept of probing as an integrated approach to interpreting the meaning and quality of visualizations and propose a set of interactive methods to examine dimensionality-reduced data as well as the projection itself. The methods let viewers see approximation errors, question the positioning of elements, compare them to each other, and visualize the influence of data dimensions on the projection space. We created a web-based system implementing these methods, and report on findings from an evaluation with data analysts using the prototype to examine multidimensional datasets.

Index Terms
—Information visualization, interactivity, dimensionality reduction, multidimensional scaling.

1 INTRODUCTION
A primary goal of information visualization is to find patterns and relationships in multivariate datasets. Many visualization techniques have been developed towards this goal such as multiple coordinated views [2], parallel coordinates [14], scatterplot matrices [28], and dimensionality reductions such as multidimensional scaling (MDS) [5]. Dimensionality reductions are a particular class of techniques that synthesise high-dimensional data spaces onto projection spaces with much fewer dimensions, typically the two dimensions of the plane. While most visualization techniques juxtapose the different data dimensions as matrices or columns, dimensionality reductions integrate them into a planar canvas. The projection results in a so-called spatialisation (i.e., embedding) of data elements that approximately represents similarity as proximity and in turn dissimilarity as distance. Considering that the human perceptional system comprises a well-developed capacity for spatial reasoning, the assumption is that spatialisation would be a more natural way [31] to analyse high-dimensional datasets since groupings, separations, and other patterns among data elements become immediately discernible.

However, there are two major caveats linked with dimensionality reduction: first, it can be challenging to interpret the positions of projected elements, and second, the errors that occur with any projection technique are not exposed to the viewer. Previous research on dimensionality reduction focuses on the efficient generation of data projections, interactive customisation of the model, and comparison of different projection techniques. There has been only little research on how the visualization resulting from data projection is interacted with. We contribute the concept of probing as an integrated approach to interpreting the meaning and quality of visualizations and propose a set of interactive methods to examine dimensionality-reduced data as well as the projection itself. The methods let viewers see approximation errors, question the positioning of elements, compare them to each other, and visualize the influence of data dimensions on the projection space. We created a web-based system implementing these methods, and report on findings from an evaluation with data analysts using the prototype to examine multidimensional datasets.

Probing Projections

[Manuscript received 31 Mar. 2015; accepted 1 Aug. 2015; date of publication xx Aug. 2015; date of current version 25 Oct. 2015.]
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.
Probing Projection Goals

- Examining the Projection
- Exploring the Data
- Design Goals:
  - Show and correct approximation errors
  - Allow for multi-level comparisons
  - Spatial orientation
  - Consistent design
- Allow grouping of samples
  - Selections
  - Classes
  - Clusters

[J. Stahnke et al., 2015]
Tooltips with statistics

Elements can be analysed by viewing their values and comparing them to other elements to account for clustering of the data, and the distribution of values in matter in this regard: the spatial distribution of values in the projection when analysing a dimensionality-reduced projection. Two things it is important to be able to quickly reference original dimensions.

4.4 Analysing dimensions

It is possible to try to answer this question: what values would a fictive sample have to have to be projected to a certain spot? Or, phrased differently: what are the interpolated values for the projection space? We used a heatmap to project the contributions of the most deviating dimensions per group. Furthermore, hovering over a grouping’s thumbnail displays small density plots for each of them. The methods for comparing samples in which the sample deviates most. An individual sample can also be compared to other samples by selecting it and hovering over other elements to see how elements compare to other elements.

Elements can be analysed by viewing their values and comparing them to the dataset. A tooltip comparison is displayed as well. Because there is no single value for the dimensions, the means are used instead. The graphical representation also takes into account the deviation, and with increasing saturation for higher deviations. If there are too many deviations, they are displayed in text form for accuracy, as well as displayed as bar charts, with density plots of the samples, density plots for them are shown in the list of dimensions, as well as a text-based preview of the most deviating dimensions per group.

Analysing a single sample is done by hovering the mouse pointer over a dot on the projection. The values for the corresponding sample are indicated in the list of dimensions. Additionally, a tooltip appears, showing the sample’s absolute values, standard deviations, and graphical representations for each dimension.

Correct distances can be visualised by using a heatmap in the projection on the left. Brightness is used to avoid confusion with the group colours. This allows to visually assess the value distribution for a given dimension, with higher values being darker. They are displayed in text form for accuracy, as well as displayed as bar charts, with density plots of the samples, density plots for them are shown in the list of dimensions, as well as a text-based preview of the most deviating dimensions per group.

4.3 Comparing elements

Comparing the selection to the dataset. A tooltip will appear and visualise the differences. In many cases the user is interested in the positioning of the dots. (For some techniques, such as PCA, the positional contribution of each original dimension can be mapped to the projected dimensions. It would then be possible to display this as a biplot, creating meaningful axes.)

Clusters can also be saved and named as selections. One solution is to use a glyph plot, with the dots themselves being coded by colour, with the names or the number of samples, is displayed below the thumbnail. Where the user being able to switch between displaying classes, selections, or user-defined clusters using the respective eye icon.

Analysing groups works similarly. When selecting a group of elements, groupings can also be compared to each other, displaying dynamic heatmaps in the projection and density plots in the sidebar. Figure 2 shows how selected elements can be displayed as thumbnails in the list of dimensions, and on the projection itself.

Correct distances can be visualised by using a heatmap in the projection on the left. Brightness is used to avoid confusion with the group colours. This allows to visually assess the value distribution for a given dimension, with higher values being darker. They are displayed in text form for accuracy, as well as displayed as bar charts, with density plots of the samples, density plots for them are shown in the list of dimensions, as well as a text-based preview of the most deviating dimensions per group.

Figure 3. After selecting one group of samples, hovering over another element will display a tooltip that compares these groups (here selections).

Figure 4. Hovering over a dimension in the sidebar displays its distribution as a heatmap in the projection on the left. Brightness is used to avoid confusion with the group colours. This allows to visually assess the value distribution for a given dimension, with higher values being darker. They are displayed in text form for accuracy, as well as displayed as bar charts, with density plots of the samples, density plots for them are shown in the list of dimensions, as well as a text-based preview of the most deviating dimensions per group.

Figure 5. An individual sample can also be compared to other samples by selecting it and hovering over other elements to see how elements compare to other elements.

Figure 6. Elements can be analysed by viewing their values and comparing them to other elements to account for clustering of the data, and the distribution of values in matter in this regard: the spatial distribution of values in the projection when analysing a dimensionality-reduced projection. Two things it is important to be able to quickly reference original dimensions.
Comparing Two Groups

![South America vs Northern Europe comparison]

**South America** 3 samples
**Northern Europe** 9 samples

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Mean</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educational attainment</td>
<td>50</td>
<td>77</td>
</tr>
<tr>
<td>Employees working very...</td>
<td>18</td>
<td>6.2</td>
</tr>
<tr>
<td>Life expectancy</td>
<td>75</td>
<td>81</td>
</tr>
<tr>
<td>Life satisfaction</td>
<td>7.1</td>
<td>7.4</td>
</tr>
<tr>
<td>Self-reported health</td>
<td>65</td>
<td>77</td>
</tr>
<tr>
<td>Student skills</td>
<td>420</td>
<td>500</td>
</tr>
<tr>
<td>Time devoted to leisure...</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>Years in education</td>
<td>16</td>
<td>19</td>
</tr>
</tbody>
</table>

[J. Stahnke et al., 2015]
Heatmap from Dimension Hover

D. Koop, CSCI 628, Fall 2021

[Heatmap image]

Heatmap from Dimension Hover

[J. Stahnke et al., 2015]
Fig. 5. Comparing an individual sample with a class from the well-known Iris flower data set. In addition to the distribution of dimensions of the class and the value of the sample, the visualization also indicates sample-centric distortions using grey and white halos.

The grid is constructed by dividing up the projection space into a number of cells, with their size based on the mean of distances for the closest three points from each point. A minimum value is used to prevent the grid from getting too small to be useful, and calculations taking too long.

The values for each grid cell are calculated by averaging the values of the points in the cell, or, if there are none, the three closest points for the cell, weighting them according to their distances from the cell. This ensures smooth transitions over large gaps in the projection space, while being responsive to abrupt changes at the same time.

4.4.2 Density plots

While the heatmaps show how the values are spatially distributed in the projection space, kernel density plots in the list of dimensions show their value distributions. In the prototype, currently the plots are generated roughly equivalent to R’s bw.nrd0 function which uses Silverman’s ‘rule of thumb’ [24, p. 48]. Percentiles are indicated on the density plots to support the visual assessment. Used together with brushing and linking, it is possible to assess how a sample, or a group of samples, relates to the whole dataset.

Markers or sub-plots for selected elements are shown on the density plots in the list of dimensions, providing dynamic highlights of samples being examined (see Figure 5, lower right). Additional markers display the dimensional values mouse position in the projection space, based on the calculations done for the heatmap, making it possible to gradually track value progressions for multiple dimensions.

4.5 Examining projection errors

Besides exploring the distribution of samples and dimensions, the visualization environment allows for the integrated examination of projection errors by providing per-sample halos, distance corrections, and a dendrogram.

4.5.1 Error halos

The cumulative distance error for each point is displayed as a halo around the dot, with the radius corresponding to the relative amount of error and the value indicating the direction of the error (see Figure 6). This is intended to help the user visually understand the quality of the projection and find potentially unreliable spots. Hovering over a dot shows the errors in relation to the hovered point, to check if the distances between certain points are correct or just projection artefacts, and learn which points should be closer together or further apart.

Halos were chosen because their visual properties are a good match for the properties of the error they represent. The error does not belong to the sample, but to the projection, and as such should be displayed by the projection. A halo is clearly connected to the dot, but also part of the projection. Size was chosen as a very intuitive metric to display the amount of error, with points with a large error standing out easily. The brightness of the halo displays the direction of the error. If the other points are farther away in the projection than in the high-dimensional space, the halo is white; if they are too close in the projection, the halo is dark. White was chosen for points that should be nearer because it stands out more, and while using the prototype ourselves, we often ended up looking for ‘similar’ samples.

Size and brightness were chosen over colour or shape, as using colour would have clashed with the colouring of the dots, and different shapes were not as glanceable as changes in brightness.

4.5.2 Distance correction

After examining the approximation errors, the viewer might decide that the errors of a certain point warrant more attention. They can then visualize them by selecting to view the high-dimensional distances between the selected point and all others. This removes all projection errors when it comes to distance, but for the selected sample only. The new, corrected positions for each point are calculated by taking the vector between the selected and the other point and multiplying it by the distance error ratio between the two. The angle between them is kept as is. As a result, the other point moves directly towards or away from the specified point, correcting the distance.

Showing Error via Sample-centric Halos

[J. Stahnke et al., 2015]
White: higher levels of similarity
Gray: lower levels of similarity
User Study & Results

• Types of Questions:
  - How would you try to characterize the type X?
  - In what way are X and Y different in their properties?
  - Are the projections of X and Y correct or do they deviate? How do you interpret this?
  - Can you discover which parts of the cluster combinations are A, B, and C?

• Discussion:
  - Learnability: need more effective mechanisms for grasping the concepts behind dimensionality reduction
  - Manipulation: What happens with results?
  - Large data: What about text corpora?

[J. Stahnke et al., 2015]
Different Projections

Fig. 1. Projections are like shadows. They are useful for obtaining a sense of the shape of a data set, when many projections are viewed. What may look like a horse in one projection may be revealed as carefully oriented pair of hands by another projection.

\[ X = \begin{bmatrix} 2 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 4 & 0 \\ & & & & & & & & & & & \end{bmatrix} \text{ and } A_1 = \begin{bmatrix} 2 & 4 & 1 & 0 & 0 & 0 & 1 & 5 & 0 & 1 & 5 & 0 \\ & & & & & & & & & & \end{bmatrix} \]

Then \[ XA_1 = \begin{bmatrix} 2 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 4 & 0 \\ & & & & & & & & & & & \end{bmatrix} \]

is the first two columns of the data matrix. If instead \[ A_2 = \begin{bmatrix} 2 & 4 & 0 & .71 & 0 & .42 & 0 & .71 & 0 & .42 & .84 & 3 & 5 \\ & & & & & & & & & & & \end{bmatrix} \]

then \[ XA_2 = \begin{bmatrix} 2 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 4 & 0 \\ & & & & & & & & & & & \end{bmatrix} \]

a combination of all three variables.

[D. Cook et al., 2008]
What Structure is Interesting?

When we use tours, what are we looking for in data? Anything that is not normally distributed, a little unusual or unexpected. For example, clusters of points, outliers, linear dependencies, non-linear relationships, and low-dimensional sub-structures. All of these can be present in multivariate data but hidden from the viewer who only chooses a few static projections. Figures 4 and 5 show some examples.

In Figure 4 a scatterplot matrix of all pairwise plots is shown at left, and a tour projection is shown at right. The pairwise plots show some linear association between three variables, particularly between the variables TEMP and PRESS, and TEMP and CODE. However, viewing the data in a tour reveals that the three variables are really perfectly related, with perhaps a slight nonlinear association. The projection of the data revealing the perfect relationship is:

\[
A_1 = \begin{bmatrix}
0.720 \\
0.470 \\
0.720
\end{bmatrix},
\]

\[
A_2 = \begin{bmatrix}
0.668 \\
0.671 \\
0.671
\end{bmatrix}.
\]

In Figure 5 the pairwise scatterplots (at left) suggest there is some clustering of the data points in this six variable data set. The tour projection (right) reveals three well-separated clusters. The projection revealing the clusters is:

\[
A_3 = \begin{bmatrix}
0.191 \\
0.573 \\
0.191
\end{bmatrix}.
\]
Tours help explore projections

[Non-linear association]

[D. Cook et al., 2008]
Going beyond 2D and 3D to visualise higher dimensions, for ordination, clustering & other models

D. Cook
Toward Comparing DNNs with UMAP Tour

M. Li and C. Scheidegger