Outline for Class 11: High Dimensionality

Objectives
By the end of the class, you will be able to…
• Describe general classes of techniques for visualizing data with many dimensions.
• Compare these classes of techniques critically, examining their strengths and weaknesses.
• Choose a multidimensional visualization technique for a particular problem and justify the choice.
• Explain the importance of ordering and filtering for multidimensional data and how this might be achieved.

Agenda for Class
• Review major classes of visualization techniques for multidimensional data.
• Discuss the research paper.

Reading Assignment
Ware: pp. 381-383 (multidimensional scaling), 348 (parallel coordinates), 182-185 (glyphs).


Notes:
1. Major techniques to display multidimensional data.
There are many different techniques, but they tend to be variations on several major themes. We have seen many of these before:

Parallel coordinates: Each dimension is plotted on a vertical axis, and these axes are lined up in parallel. Data items are then plotted as polylines connecting the item’s value on each dimension.
• Positive correlations between dimensions can be seen by a lack of line crossings. (A -> B in the image below)
• Negative correlations show lines meeting at a central point.
Glyphs: Different dimensions are mapped to different visual properties of objects drawn on the screen (e.g., size, shape, colour, angle, etc.)

One interesting type of glyph is the Chernoff Face, where different properties of the data are mapped to different facial features.

- Problems:
  - Faces implicitly give positive or negative meanings to the values, which may not be appropriate.
  - Some facial features may have more visual weight than others

[http://kspark.kaist.ac.kr/Human%20Engineering.files/Chernoff/Chernoff%20Faces.htm]

Star or whisker glyphs are a special type of glyph in which each dimension is represented as an axis directed radially outward from a central point. The length of the line represents that item’s value for that dimension.

Dimensionality reduction techniques (multidimensional scaling or MDS, principle component analysis, self-organizing maps, etc.): Create a lower dimensional layout of the data set (typically in 2D plane but could be 1D, 3D, etc.). In this lower dimensional layout, points that are close together are similar and points that are far apart are not
similar. Actual X and Y directions on the plane may have no clear meaning (MDS), or a non-intuitive meaning (PCA).

Example: Hand positions. In this case the 4096 D dataset is mapped to 2D, and the 2 dimensions do actually map to real dimensions (wrist rotation and finger extension) but this would not be the case with all dimensionality reduction techniques.

Why is dimensionality reduction useful?
- Can see high level clusters & outliers
- Assume true dimensionality is much lower than measure dimensionality
  - E.g. fisheries: want spawn rates. Measure water color, water temperature, catch rate…

How many dimensions should we reduce to?
- Often we see people reduce to 2 or 3 dimensions because this is easily displayed on a computer. But this may hide interesting structures in the data if the true dimensionality is greater than 2-3 dimensions.

**Small multiples**: Repeated charts or other elements in a line or grid. See Tufte Chap. 4. Variations on this idea include
- Scatterplot matrices (where each low level item is a scatterplot),
- Dimensional stacking (dimensions are embedded in a 2D table form hierarchically)
- Worlds-within-worlds (3D hierarchy rather than 2D)
- Table lens (where essentially graphs for each dimension are stacked next to each other)

**Pixel oriented techniques**: Where each pixel represents some data value. Example is the VisDB system by Keim & Kriegel:
• In the VisDB example, colouring represents the relevance of an item to a query. Alternately, colouring could represent data values directly.

2. Compare the scatterplot matrices, star glyphs, and parallel coordinates in the Yang et al. paper. What types of patterns and trends are easiest to see with each one?

Star glyphs:
• Compare how different individual items are overall, across all dimensions or a filtered subset. Then identify which dimensions they are similar on and which are different.
• Can get some sense of overall variability of the items.

Scatterplot matrices:
• Quickly scan many pairs of dimensions to find correlations

Parallel coordinates:
• Range of data on each dimension.
• High dimensional clusters and outliers (e.g. outlier lines at bottom of figure 2d).
• Level of correlation / similarity between dimensions (but only dimensions that are side-by-side). Note: this is also aided by the dimension spacing.

3. Why is automatic dimension ordering, spacing, & filtering important for multidimensional data?

There are several reasons why these are important:
• Order of dimensions strongly affects what patterns can be seen. E.g. in parallel coordinates you can only see correlations between adjacent dimensions. In many cases you will just see noise or miss important patterns if the ordering is not useful.
• Ordering and spacing also affects how we interpret displays E.g. Items placed close together are presumed to be similar (proximity grouping principle from the Gestalt laws). We want to use this intentionally to our advantage to show real similarities between dimensions (and avoid unintentionally suggesting patterns that are not real).
• When the dimensionality becomes large, it may be difficult to view all dimensions at once so filtering is necessary.
These actions can be done manually, but this becomes very cumbersome when there are many dimensions:

- There may be too many dimensions to try all combinations, leaving the user to guess which ones to try.
- Interaction of moving and filtering may become cumbersome itself.

4. **Advantages / disadvantages of the DOFSA technique**

**+**

- Can be applied to multiple visualization techniques.
- Filtering, spacing, and ordering together allow you to see some patterns where previously few if any patterns were visible. In other words, the hierarchical clustering is able to identify some structure in real datasets.
- Both automated and manual reordering/spacing/filtering are possible. Automatic provides a good starting point, which can then be tweaked manually.
- The automatic spacing in the parallel coordinates gives more room to dimensions that are different than those that are similar. These may be the more interesting ones to look at, so this acts somewhat like focus + context.

**-**

- The technique is only as good as the similarity measure used
- When interactively controlling spacing, users need to remember whether they have applied distortion (because the distances between items on the screen no longer accurately represent their high-dimensional distances).
- Makes assumptions that might not always be true (e.g., that users want to see dimensions that are different from each other). Fortunately there is manual control as well for when this happens.

5. **How could the DOFSA paper be improved?**

Because all the figures show filtering after ordering & spacing, it is hard to understand the benefits of each. It would be useful to see images such as filtering without ordering for comparison sake.

It was nice that the technique was shown with real data, but perhaps it would also have been useful to look at contrived data to examine whether all clusters are created as expected, and how different sizes of clusters / numbers of dimensions etc affect the result.

**Key Concepts for Today:**

1. There are several different approaches to visualizing multidimensional data, each with its own benefits and drawbacks.
2. Dimension ordering and filtering is important for many different multidimensional visualization techniques. One approach to achieve this automatically is by clustering dimensions hierarchically and using this hierarchy in the ordering.