High Dimensional Data

Daniel Archambault

Previously in CSCM27...

- What do coordinated multiple views provide?
- What is the relationship to dynamic data?
- When should we apply animation?
- When should we apply small multiples?

Previously in CSCM27... (2)

- Get your n-dimensional glasses on...
- We are looking at visualising high dimensional data

Dimensionality Reduction and High Dimensional Data

Thanks

- Huge thanks to Tamara Munzner (my PhD adviser)
- Many of the figures are from her work
- This lecture is based off of her lectures

High Dimensional Data

- Three or more dimensions
- We deal primarily with continuous dimensions
- Cannot visualise directly
- A number of ways to look at this data

Overview

- Many methods to visualise high dimensional data

1. Scatter plot matrices (SPLOMs)
2. Parallel Coordinates
3. Glyphs and Glyph Matrix
4. Principal Component Analysis
5. Multidimensional Scaling

- Other methods exist, but we focus on these

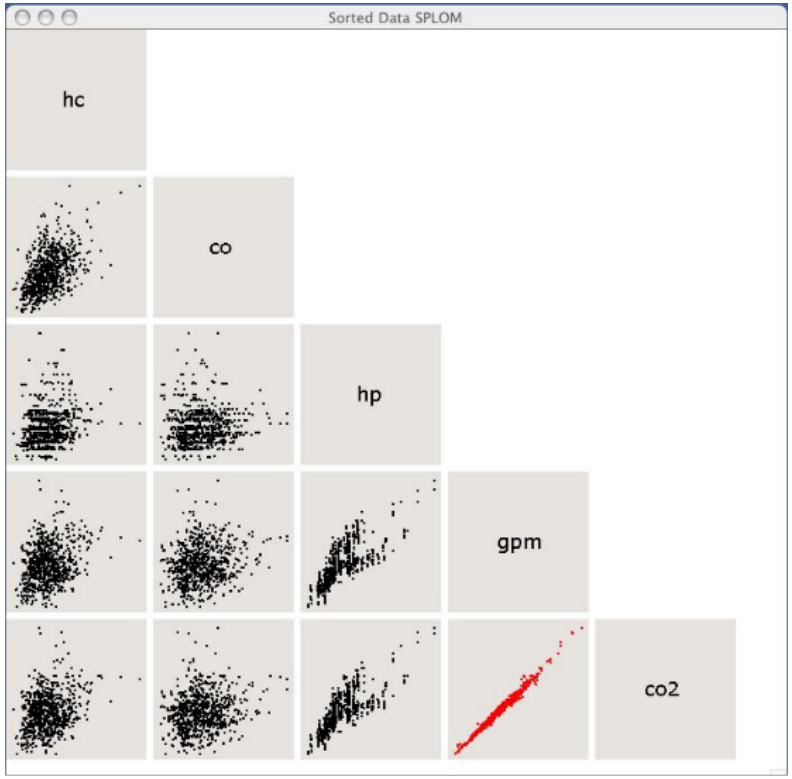
Axis Aligned Visualisations

- Positives
  - The x and y dimensions have a clear meaning
  - Positions in space have a clear meaning
- Negatives
  - If feature is not axis aligned, may be harder to see...

Use Coordinated Multiple Views!

- Coordinated multiple views can be used
- In this case, it is a scatterplot matrix (SPLOM)
- Good news is that all pairs of dimensions are visible
- Axis is interpretable and so is position
- Bad news, takes up a lot of screen space
- Off axis features may be difficult to see...

Scatter Plot Matrix (SPLOMS)



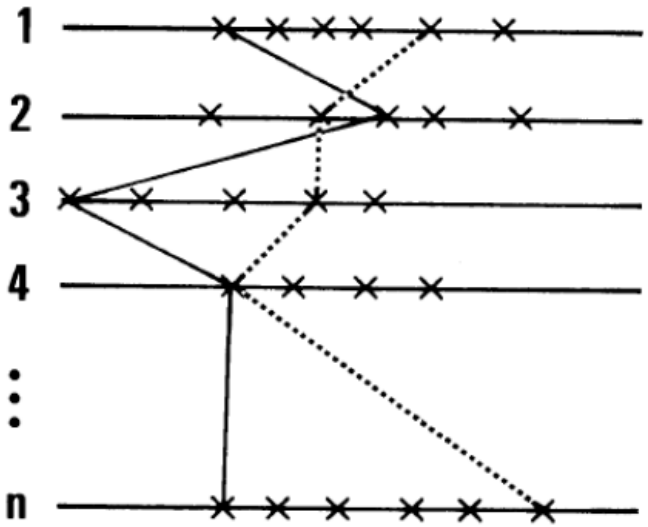
L. Wilkinson, A. Anand and R. Grossman, "High-Dimensional Visual Analytics: Interactive Exploration Guided by Pairwise Views of Point Distributions," in IEEE Transactions on Visualization and Computer Graphics, 12:(6)1363-1372, 2006.

- Screenspace requirements can be reduced by computing

Parallel Coordinates

- Parallel coordinates are far more compact than SPLOMS
- Good for correlations anti-correlations between dimensions
- Good for point clusters in a few dimensions
- Problem: how to pick a good order for the dimensions
- Off axis features may be a bit difficult...

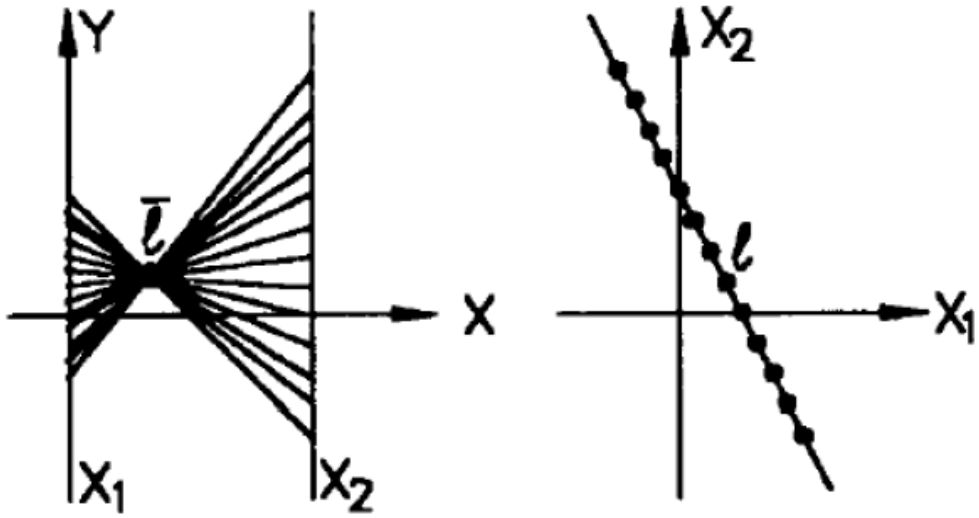
Parallel Coordinates



Hyperdimensional Data Analysis Using Parallel Coordinates. Edward J. Wegman. Journal of the American Statistical Association, 85(411), Sep 1990, p 664-675.

- Order dimensions left to right or top to bottom

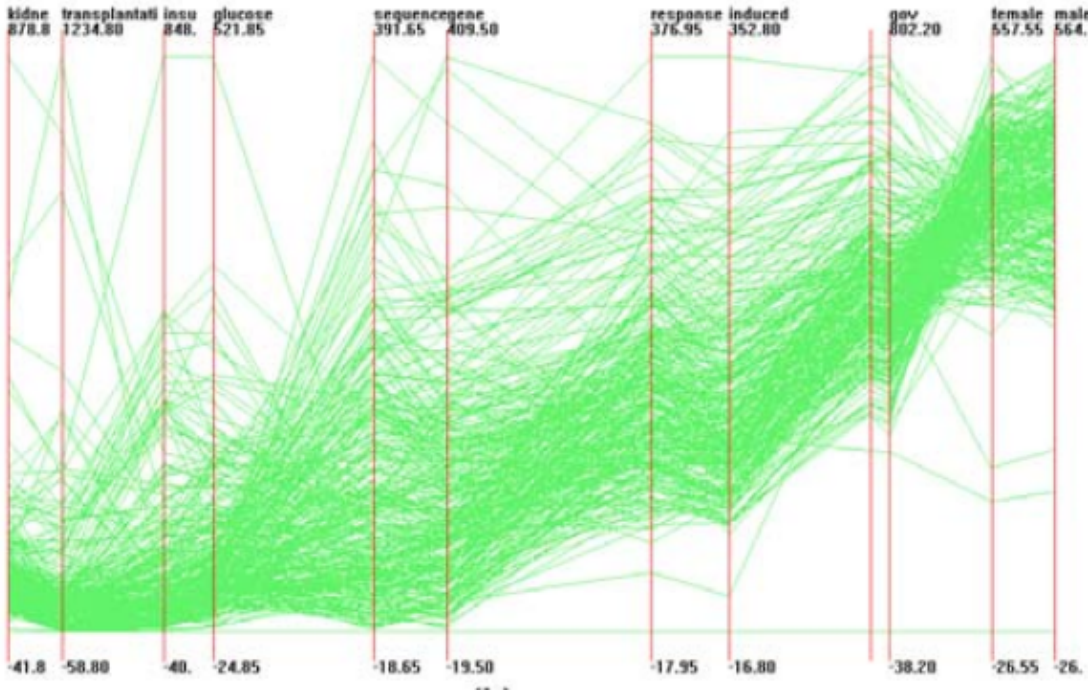
Why It Works



Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry. Alfred Inselberg and Bernard Dimsdale, IEEE Visualization 1990.

- Clusters of polylines are clusters of points!
- Correlations correspond to single point intersections

Parallel Coordinates

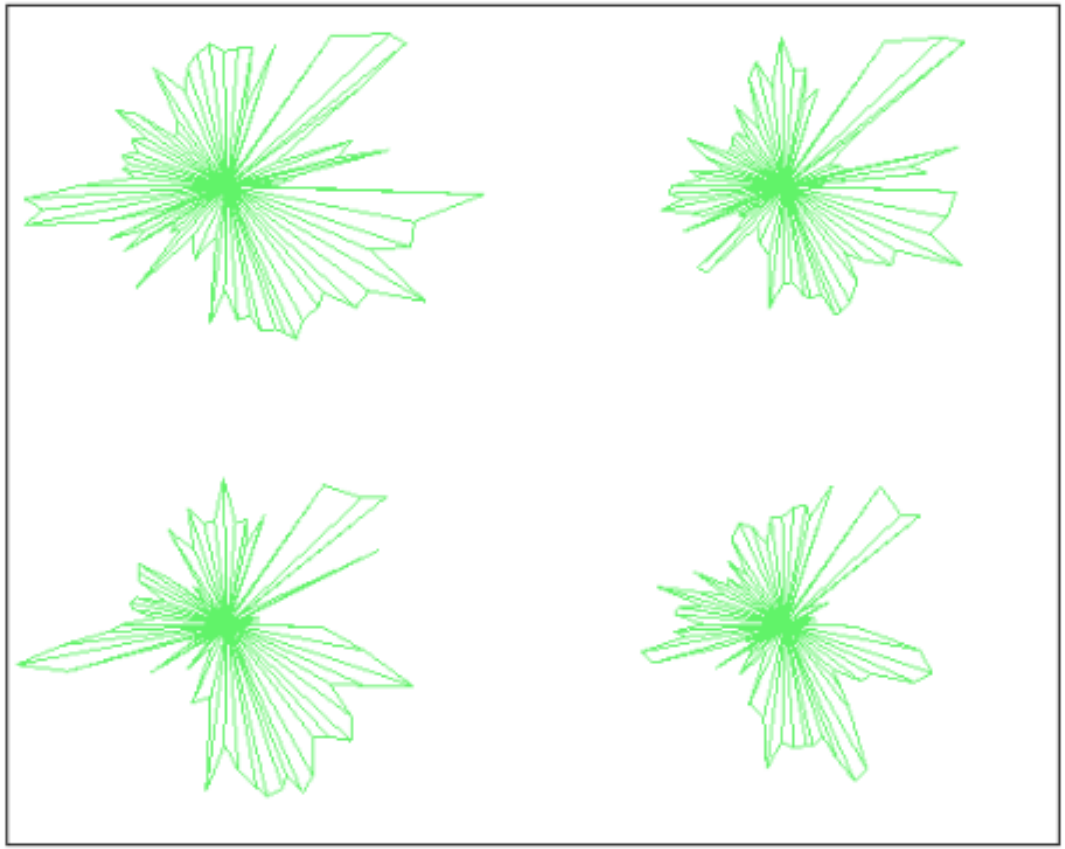


Jing Yang, Wei Peng, M. O. Ward and E. A. Rundensteiner, "Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets," IEEE Symposium on Information Visualization 2003 (IEEE Cat. No.03TH8714), Seattle, WA, USA, 2003, pp. 105-112.

Glyph Representations

- Glyphs can be designed to represent many dimensions in 2D
- Dimensions do not need to be continuous
- Designs may be able to highlight off axis features
- Positive: more than pairs of dimensions at once
- However, glyph design is hard
- It may take time to learn glyphs in order to understand them

Glyph Representation of High Dimensional Data



Jing Yang, Wei Peng, M. O. Ward and E. A. Rundensteiner, "Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets," IEEE Symposium on Information Visualization 2003 (IEEE Cat. No.03TH8714), Seattle, WA, USA, 2003, pp. 105-112.

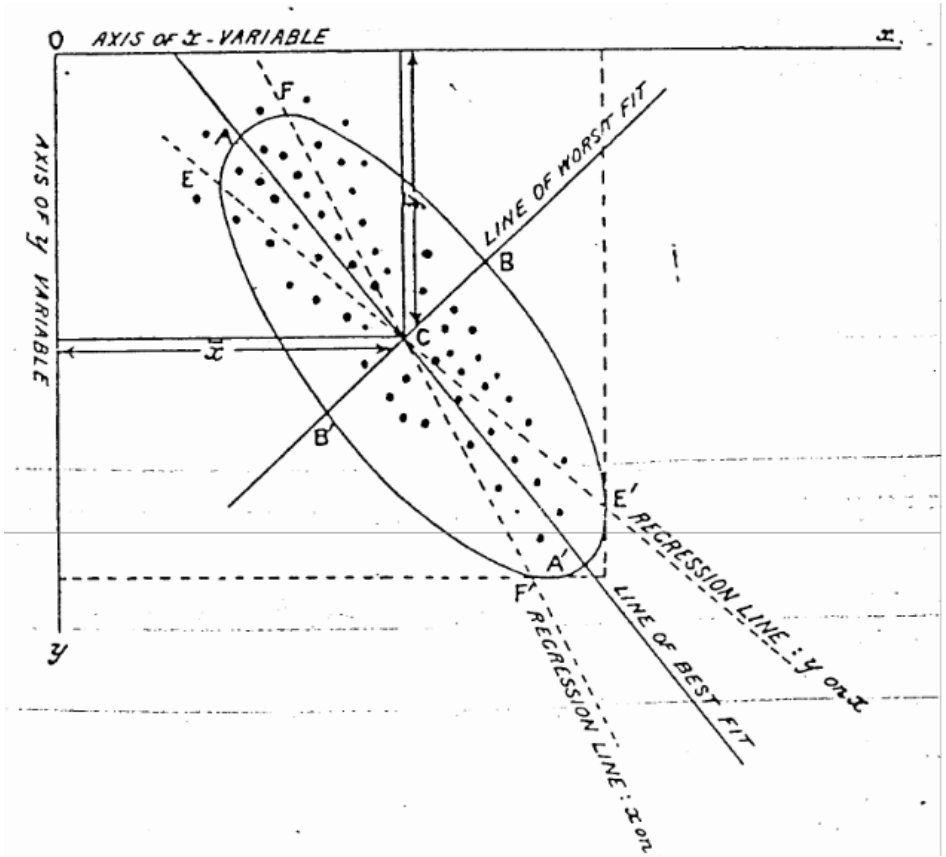
Many D to 2D

- Previous methods are usually axis aligned
- Advantage: x and y are interpretable
- Disadvantage: features are not always axis aligned
- Employ a projection or mapping down to 2D?

Principal Component Analysis (PCA)

- Given a point set, compute maximum directions of variance
- Take a photograph of the data in HD space
- Look at this photograph
- Essentially minimises distance to the projection plane
- Good news: off axis features can be captured
- Two vectors that define plane can be interpreted
  - zero or low means lower influence
- Problem: information loss through projection
- Orientation of plane in high dimension hard to interpret?

Principal Component Analysis (PCA)



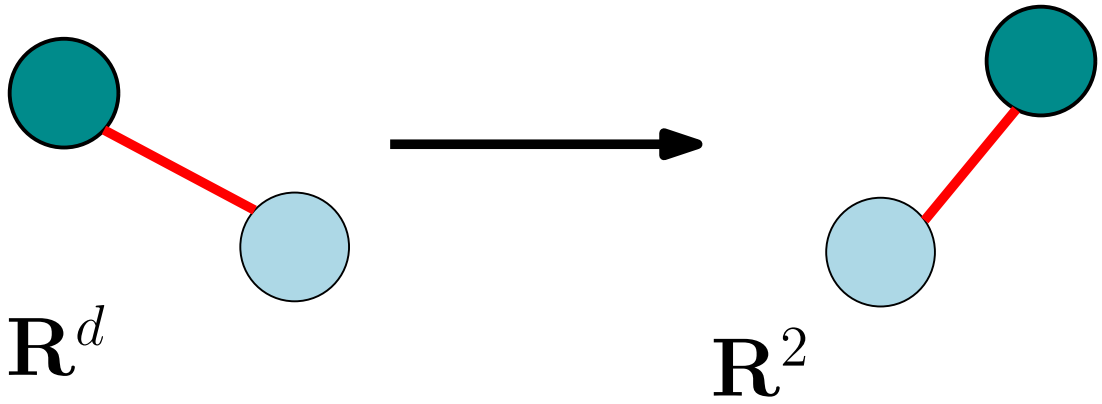
Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine. 2 (11): 559–572.

- compute using matrix algebra on the covarience matrix

Multidimensional Scaling

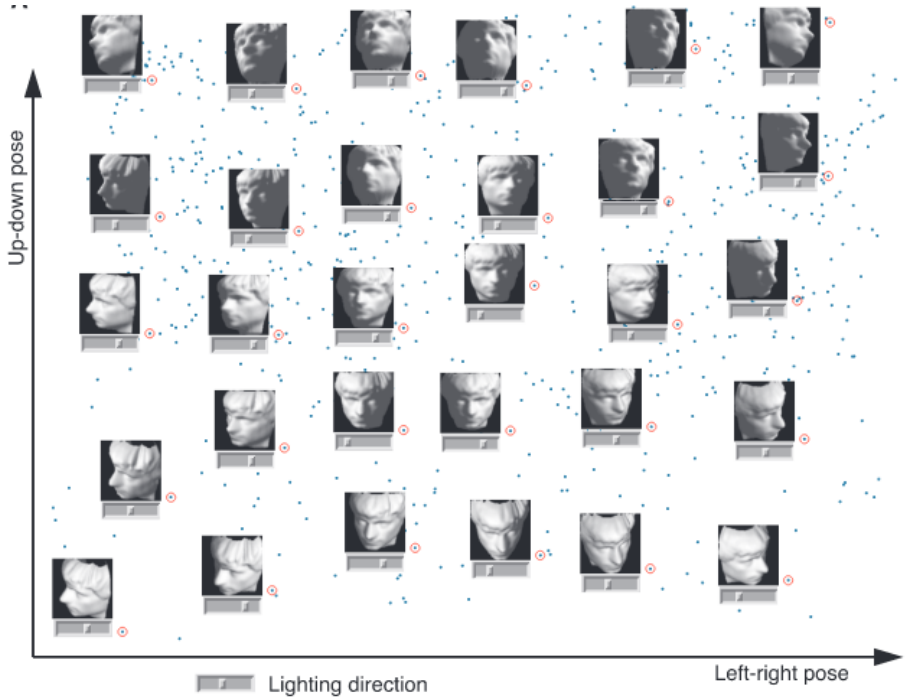
- Don't project anything, instead preserve distances
- Create a two dimensional embedding of points whereby:
  - LD distance $\propto$ HD distance
  - stuff that is close is close
  - stuff that is far is far
- Advantage: now nothing needs to be axis aligned
- Not a projection and more flexibility wrt information loss
- Problem: axis not interpretable anymore
- Distance matrix can be expensive to compute

Multidimensional Scaling (MDS)



- Mapping of high dimensional data to lower dimension (often 2D)
- Principle: distances in HD proportional to LD

Why Do MDS?



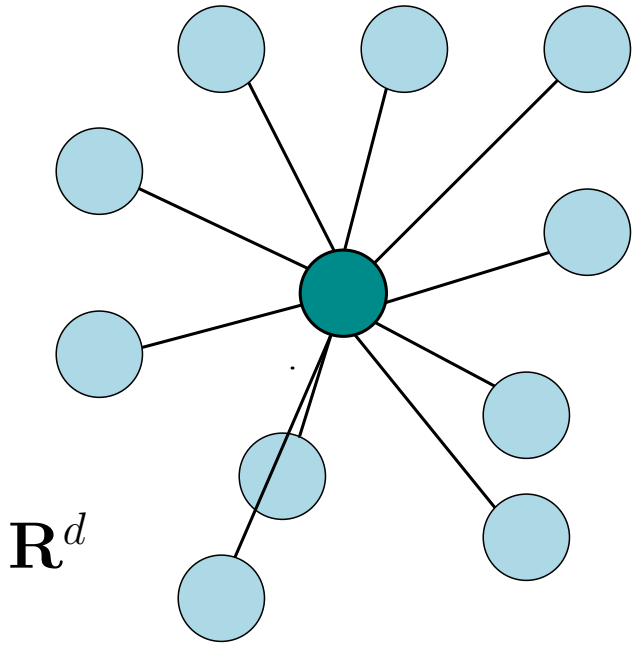
A Global Geometric Framework for Nonlinear Dimensionality Reduction. J. B. Tenenbaum, V. de Silva, and J. C. Langford. Science 290(5500):2319–2323, 2000

- Axis does not mean anything
  - Points that are near are similar
  - Points that are far different

How to Do MDS?

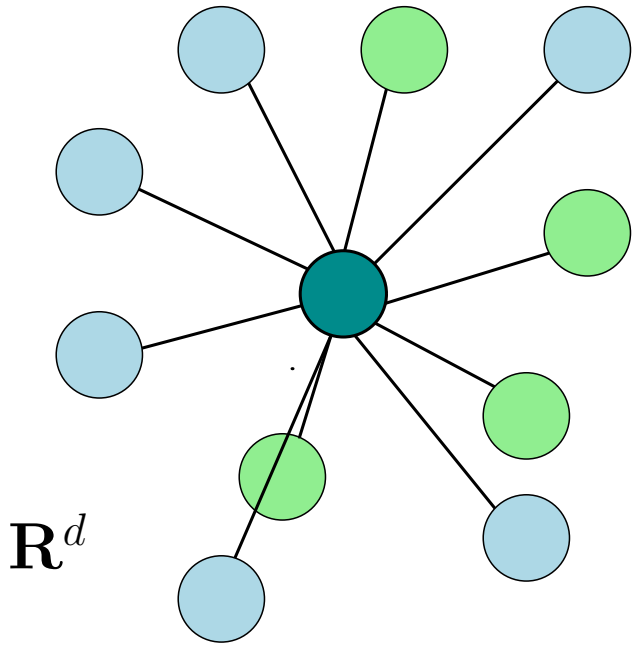
- Can be computed using matrix algebra on distance matrix
  - classical MDS
- However, proportionality is not very good
- Better approach use a spring algorithm
  - compute distance to all points in HD & LD
  - nudge point with force such that distances are more proportional
  - iterate until convergence

Compute All HD Distances



- Compute all pairwise distances in HD space
- Expensive as there are $O(n^2)$

Sample HD Distances Intelligently



- Select a local neighbourhood for each point
- Use as anchors to layout everything else

Implementations and Examples

- PCA example
  - `http://bl.ocks.org/hardbyte/40cd6622cffbe98055d3`
- MDS code available in Java
  - use to create a projection and then read in with D3
  - `http://algo.uni-konstanz.de/software/mdsj/`
- Visualisation of MDS result
  - `http://bl.ocks.org/dahtah/4482115`

Quiz Time...

- Name a dimensionality reduction technique
- Discuss advantages and disadvantages