High Dimensionality

Lecture 10 CPSC 533C, Fall 2004

20 Oct 2004

Reading

Hyperdimensional Data Analysis Using Parallel Coordinates
Edward J. Wegman. Journal of the American Statistical Association,
Vol. 85, No. 411. (Sep., 1990), pp. 664-675.

Fast Multidimensional Scaling through Sampling, Springs and Interpolation
Alistair Morrison, Greg Ross, Matthew Chalmers,
Information Visualization 2(1) March 2003, pp. 68-77.

Cluster Stability and the Use of Noise in Interpretation of Clustering
George S. Davidson, Brian N. Wylie, Kevin W. Boyack, Proc InfoVis 2001.

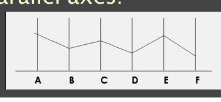
Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration Of High Dimensional Datasets
Jing Yang, Wei Peng, Matthew O. Ward and Elke A. Rundensteiner. Proc. InfoVis 2003.

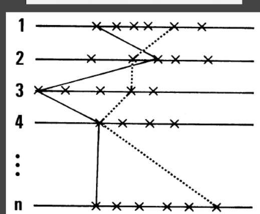
Optional:
- Visualizing the non-visual: spatial analysis and interaction with information from text documents. James A. Wise et al, Proc. InfoVis 1995
- Hierarchical Parallel Coordinates for Visualizing Large Multivariate Data Sets Ying-Huey Fua, Matthew O. Ward, and Elke A. Rundensteiner, IEEE Visualization '99.
- Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry. Alfred Inselberg and Bernard Dimsdale, IEEE Visualization '90.

2

Parallel Coordinates

only 2 orthogonal axes in the plane
instead, use parallel axes!





3

[Hyperdimensional Data Analysis Using Parallel Coordinates. Edward J. Wegman. Journal of the American Statistical Association, Vol. 85, No. 411. (Sep., 1990), pp. 664-675.]

PC: Correllation

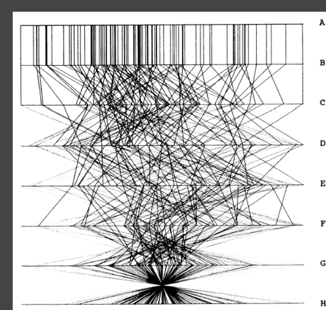


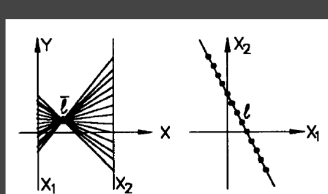
Figure 3. Parallel Coordinate Plot of Six-Dimensional Data Illustrating Correlations of ρ = 1, .8, .2, 0, −.2, −.8, and −1.

[Hyperdimensional Data Analysis Using Parallel Coordinates. Edward J. Wegman. Journal of the American Statistical Association, Vol. 85, No. 411. (Sep., 1990), pp. 664-675.]

4

PC: Duality

rotate-translate
point-line
- pencil: set of lines coincident at one point
- not critical to understand projective plane details!



[Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry. Alfred Inselberg and Bernard Dimsdale, IEEE Visualization '90.]

5

PC: Axis Ordering

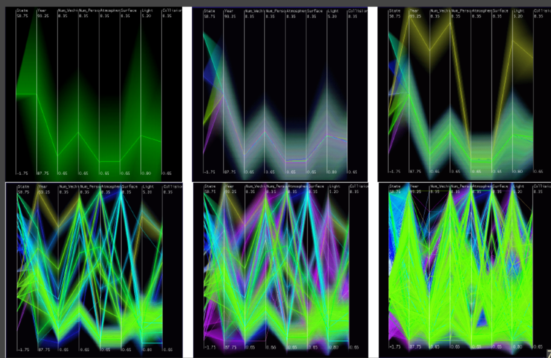
geometric interpretations
- hyperplane, hypersphere
- points do have intrinsic order

infovis
- no intrinsic order, what to do?
- indeterminate/arbitrary order
  - weakness of many techniques
  - downside: human-powered search
  - upside: powerful interaction technique
- most implementations
  - user can interactively swap axes

Automated Multidimensional Detective
- [Inselberg 99]
- machine learning approach

6

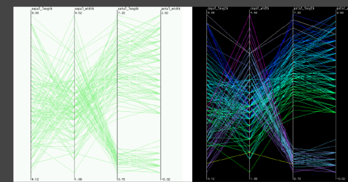
Hierarchical Parallel Coords: LOD



[Hierarchical Parallel Coordinates for Visualizing Large Multivariate Data Sets. Ying-Huey Fua, Matthew O. Ward, and Elke A. Rundensteiner, IEEE Visualization '99.]

Hierarchical Clustering

proximity-based coloring



[Hierarchical Parallel Coordinates for Visualizing Large Multivariate Data Sets. Ying-Huey Fua, Matthew O. Ward, and Elke A. Rundensteiner, IEEE Visualization '99.]

interaction lecture later:

- structure-based brushing
- extent scaling

Dimensionality Reduction

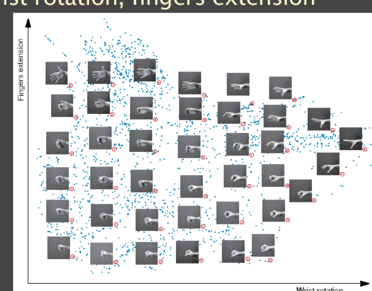
mapping multidimensional space into space of fewer dimensions

- typically 2D for infovis
- keep/explain as much variance as possible
- show underlying dataset structure
- multidimensional scaling (MDS)

minimize differences between interpoint distances in high and low dimensions

Dimensionality Reduction: Isomap

4096 D: pixels in image
2D: wrist rotation, fingers extension



[A Global Geometric Framework for Nonlinear Dimensionality Reduction. J. B. Tenenbaum, V. de Silva, and J. C. Langford. Science 290(5500), pp 2319--2323, Dec 22 2000]

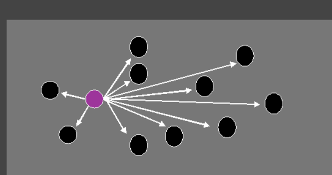
Naive Spring Model

repeat for all points

- compute spring force to all other points
- difference between high dim, low dim distance
- move to better location using computed forces

compute distances between all points

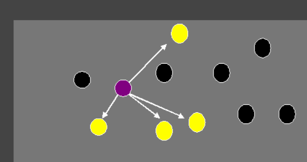
- O(n^2) iteration, O(n^3) algorithm



Faster Spring Model [Chalmers 96]

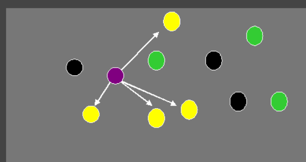
compare distances only with a few points

- maintain small local neighborhood set



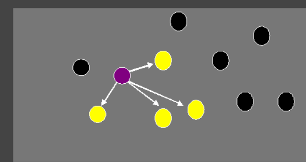
Faster Spring Model [Chalmers 96]

compare distances only with a few points
- maintain small local neighborhood set
- each time pick some randoms, swap in if closer



Faster Spring Model [Chalmers 96]

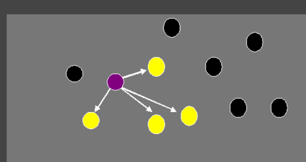
compare distances only with a few points
- maintain small local neighborhood set
- each time pick some randoms, swap in if closer



Faster Spring Model [Chalmers 96]

compare distances only with a few points
- maintain small local neighborhood set
- each time pick some randoms, swap in if closer

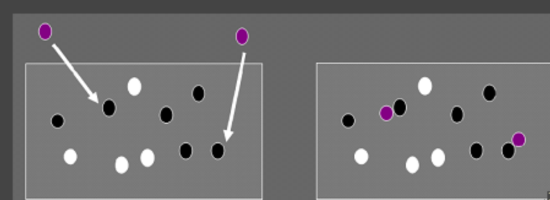
small constant: 6 locals, 3 randoms typical
- O(n) iteration, O(n2) algorithm



Parent Finding [Morrison 02, 03]

lay out a root(n) subset with [Chalmers 96]
for all remaining points
- find "parent": laid-out point closest in high D
- place point close to this parent

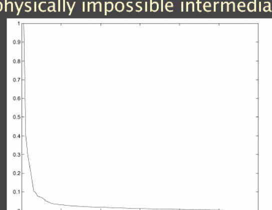
O(n^5/4) algorithm



True Dimensionality: Linear

how many dimensions is enough? > 2 or 3?
- knee in error curve

example: measured materials from graphics
linear PCA: 25
- can get physically impossible intermediate points

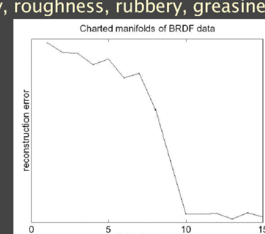


[A Data-Driven Reflectance Model, SIGGRAPH 2003, W Matusik, H. Pfister
M. Brand and L. McMillan, graphics.lcs.mit.edu/~wojciech/pubs/sig2003.pdf]

True Dimensionality: Nonlinear

nonlinear MDS: 10-15
- all intermediate points possible

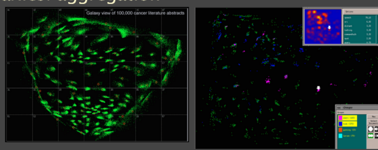
categorizable by people
- red, green, blue, specular, diffuse, glossy, metallic,
- plastic-y, roughness, rubbery, greasiness, dustiness...



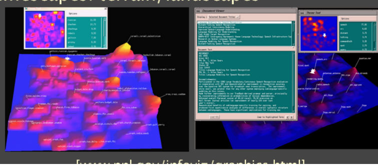
[A Data-Driven Reflectance Model, SIGGRAPH 2003, W Matusik, H. Pfister

Themescapes/Galaxies

MDS output: beyond just drawing points
- galaxies: aggregation



- themescapes: terrain/landscapes



[www.pnl.gov/infoviz/graphics.html]

Cluster Stability

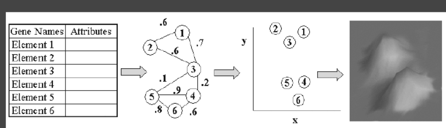
display
- also terrain metaphor

underlying computation
- energy minimization (springs) vs. MDS
- weighted edges

do same clusters form with different random start points?

"ordination"
- spatial layout of graph nodes

Approach



normalize within each column

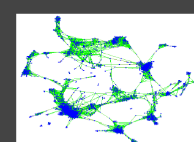
similarity metric
- discussion: Pearson's correllation coefficient

threshold value for marking as similar
- discussion: finding critical value

Graph Layout

criteria
- distance in layout matching graph-theoretic distance
  - vertices one hop away close
  - vertices many hops away far
- insensitive to random starting positions
  - major problem with previous work!
- tractable computation

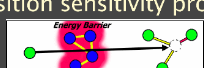
force-directed placement
- discussion: energy minimization
- others: gradient descent, etc
- discussion: termination criteria

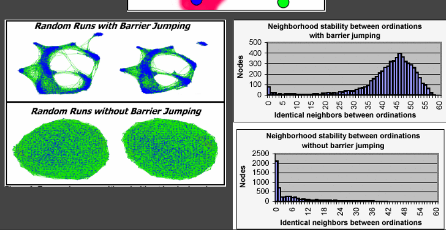


Barrier Jumping

same idea as simulated annealing
- but compute directly
- just ignore repulsion for fraction of vertices

solves start position sensitivity problem



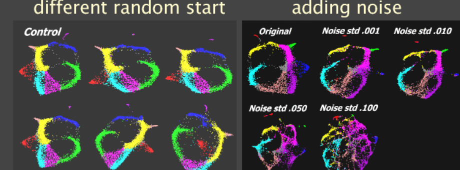


Results

efficiency
- naive approach: O(V^2)
- approximate density field: O(V)

good stability
- rotation/reflection can occur

different random start　　adding noise



Critique

real data
- suggest check against subsequent publication!

give criteria, then discuss why solution fits

visual + numerical results
- convincing images plus benchmark graphs

detailed discussion of alternatives at each stage

specific prescriptive advice in conclusion

Dimension Ordering

in NP, like most interesting infovis problems
- heuristic

divide and conquer
- iterative hierarchical clustering
- representative dimensions

choices
- similarity metrics
- importance metrics
  - variance
- ordering algorithms
  - optimal
  - random swap
  - simple depth-first traversal

Spacing, Filtering

same idea: automatic support

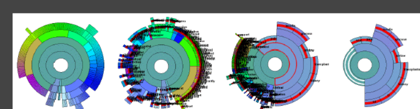
interaction

- manual intervention

- structure-based brushing
- focus+context, next week

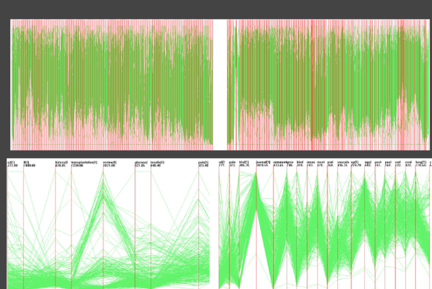
Results: InterRing

raw, order, distort, rollup (filter)



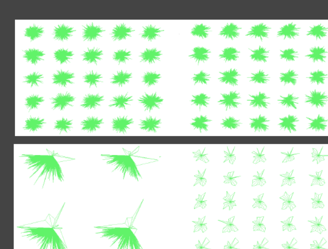
Results: Parallel Coordinates

raw, order/space, zoom, filter



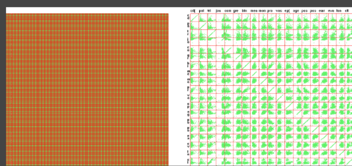
Results: Star Glyphs

raw, order/space, distort, filter



Results: Scatterplot Matrices

raw, filter



Critique

pro

approach on multiple techniques,
real data!

con

always show order then space then filter
- hard to tell which is effective
- show ordered vs. unordered after zoom/filter?

Software, Data Resources

www.cs.ubc.ca/~tmm/courses/cpsc533c-04-fall/resources.html