Data Visualization

Michael Grossberg

Multi and High Dimensional Data
Review: Data Types

- **Data**
  - **Tabular**
    - **Ordered**
      - Categorical
        - Apples
        - Oranges
        - Bananas
      - Ordinal
        - Small
        - Medium
        - Large
      - Quantitative
        - 10 inches
        - 13 inches
        - 18.5 inches
    - Relational
    - Spatial
  - Trees
  - Networks
  - Intrinsic position
Multi-Dimensional Data

- Tabular data, containing
  - rows (records)
  - columns (dimensions)
- rows >> columns
- identifiers introduce semantics
- Independent & dependent variables
  - dependent are f(independent)
Muli-Dimensional Type

categorical

ordered

tabular

quantitative

10 inches

13 inches

18.5 inches

trees

networks

intrinsic position

apples

oranges

bananas

small

medium

large

relational

spatial
Limits to Displaying Dimensions

Limit? 4D? 5D? … 10D!??
High-Dimensional Data Visualization

❖ How many dimensions?
  ❖ ~50 – tractable with “just” vis
  ❖ ~1000 – need analytical methods
❖ How many records?
  ❖ ~ 1000 – “just” vis is fine
  ❖ >> 10,000 – need analytical methods
❖ Homogeneity
  ❖ Same data type?
  ❖ Same scales?
Analytic Component

- Scatterplot Matrices
- Parallel Coordinates
- Pixel-based visualizations/heat maps
- Multi-dimensional Scaling
- LDA

Range:

no / little analytics — strong analytics component
Geometric Methods
Parallel Coordinates

- Each axis represents dimension
- Lines connecting axis represent records
- Suitable for
  - all tabular data types
  - heterogeneous data
D3 Parallel Coordinates

http://bl.ocks.org/jasondavies/1341281
Parallel Coordinates

- Shows primarily relationships between adjacent axis
- Limited scalability (~50 dimensions, ~1-5k records)
- Transparency of lines
- Interaction is crucial
- Axis reordering
- Brushing
- Filtering

Algorithmic approaches:
- Choosing dimensions
- Choosing order
- Clustering & aggregating records
500-Axis Parallel Coordinate
Ambiguities
Star Plot/Radar Plot

- Similar to parallel coordinates
- Radiate from a common origin

Coekin 1969
D3: Star Plot

http://bl.ocks.org/kevinschaul/8213691
Trellis Plots

Multiple Plots (often with shared axis)
Small Multiples

- Like Trellis
- More plots
- Axis may not be important
- Each plot point on axis
- Use multiple views to show different partitions of a dataset

public support for vouchers, Andrew Gelman 2009

Multiple Line Charts

http://square.github.io/cubism/
Combining Various Charts
Scatterplot Matrices (SPLOM)

- Matrix of size $d \times d$
- Each row/column is one dimension
- Each cell plots a scatterplot of two dimensions

Scatterplot Matrices

- Limited scalability (~20 dimensions, ~500-1k records)
- Brushing is important
- Often combined with “Focus Scatterplot” as F+C technique

- Algorithmic approaches:
  - Clustering & aggregating records
  - Choosing dimensions
  - Choosing order
SPLOM Aggregation

Datavore: http://vis.stanford.edu/projects/datavore/splom/
SPLOM F+C, Navigation

Rolling the Dice
Multidimensional Visual Exploration using Scatterplot Matrix Navigation

Niklas Elmqvist
Pierre Dragicevic
Jean-Daniel Fekete
INRIA

http://youtu.be/E1birsp9iYk

Elmqvist
Combining PCs & Sploms
Connected Charts

C. Viau, M. J. McGuffin 2012
http://profs.etsmtl.ca/mmcguffin/research/
Data Reduction
Reducing Rows

Sampling  Filtering  Clustering

http://www.jasondavies.com/parallel-sets/

Later
Sampling

- Show (random) subset
- Efficient for large dataset
- For display purposes
- Outlier-preserving approaches

Ellis & Dix, 2006
Filtering

- Criteria to remove data
  - minimum variability
  - Range for dimension
  - consistency in replicates
Filter Example

http://square.github.io/crossfilter/
Clustering

http://www.jasondavies.com/parallel-sets/

Later
Dimension Reduction
Simple Random Projection

Experiments with Random Projection,
Sanjoy Dasgupta, 2000
Grand Tour

Travel a Dense Path on 2-plane projections in N-Space

Look at all the scatterplots (Movie)
Interesting Projections

- Search for 2D scatter plots that maximize/minimize some attribute
- Clumpiness
- Variance
- Non-gaussianness (entropy)
Scagnostics

Look for projections which look interesting

TN Dang, L Wilkinson - 2014
Searching for Projections
Principal Component Analysis (PCA)
1-D mean, stdev

Mean

$$\mu = E[x] = \frac{1}{n} \sum_{i=0}^{n} x_i$$

Variance = (Standard Deviation)$^2$

$$\sigma^2 = E[(x - \mu)^2] = \frac{1}{n} \sum_{i=0}^{n} (x_i - \mu)^2$$
Normal (1-D) distribution

\[ p(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x - \mu)^2}{2\sigma^2}} \]
N-D mean, (co-)variance

N-D Mean

\[ \mu = \mathbb{E}[\mathbf{x}] = \frac{1}{n} \sum_{i=0}^{n} \mathbf{x}_i \]

\[ \mu = \{ \mu_k \}, \quad \mu_k = \frac{1}{n} \sum_{i=0}^{n} x_{i,k} \]

N-D Covariance

\[ \Sigma_{j,k} = \frac{1}{n} \sum_{i=0}^{n} (x_{i,j} - \mu_j)(x_{i,k} - \mu_k) \]
Matrix Versions

\[ X = \begin{bmatrix}
    x_{1,1} & \cdots & x_{n,1} \\
    \vdots & \ddots & \vdots \\
    x_{1,m} & \cdots & x_{n,m}
\end{bmatrix} \]

N-D Covariance

\[ \Sigma = \tilde{X}^T \tilde{X} \]
N-D Normal Distribution

\[ f_{x}(x_1, \ldots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right), \]
Diagonalization

Symmetric matrix

Diagonalization (SVD)

Eigenvalues
Variance

First principal component

Eigenvectors
Eigenvectors = Principal Components

First principal component = direction of greatest variance
Least Squares
Data Oriented Coordinates

http://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-eigenvalues-and-dimension-reduction/
PCA in Vis Reduction

PCA from Harvard Cs171 student project

http://mu-8.com

Mercer and Pandian
Many other Linear Component Reductions

Linear Discriminant Analysis (LDA)

Separate classes

Latent Semantic Analysis (LSA)

Also

Factor Analysis

Dictionary Learning

Non-negative Matrix Factorization
Projection Pursuit

Independent Component Analysis (ICA)

Make components independent

Maximize Entropy/Non-gausianness
Non-Linear Dimension Reduction

Multi-Dimensional Scaling (MDS)

- Preserve distances between observations
- Kernel PCA

Manifold Learning

- Non-Linear map to lower dimensional space
Pixel Based Methods
Pixel Based Methods

Designing Pixel-Oriented Visualization Techniques: Theory and Applications
Daniel A. Keim, 2000
3D Pitfall: Occlusion & Perspective

Which one is the tallest bar?
What is the pattern in the data?

Gehlenborg and Wong, Nature Methods, 2012
3D Pitfall: Occlusion & Perspective

Which one is the tallest bar?
What is the pattern in the data?

Gehlenborg and Wong, Nature Methods, 2012
Pixel Based Displays

- Each cell is a “pixel”, value encoded in color / value
- Meaning derived from ordering
- If no ordering inherent, clustering is used
- Scalable – 1 px per item
- Good for homogeneous data
  - same scale & type
Bad Color Mapping

Normal Vision

Deuteranope Vision (“Red-Green Blindness”)
Good Color Mapping

Normal Vision

Deuteranope Vision ("Red-Green Blindness")
Color is relative!
Machine Learning

- Supervised Learning
  - A.K.A. Classification
  - Known labels (subset of rows)
  - Algorithms: label unlabeled rows

- Unsupervised Learning
  - A.K.A Clustering
  - Algorithm: label based on similarity

- Semi-Supervised Learning
  - Do both
Clustering

Partition

Hierarchical

Bi-Partite

Fuzzy
Machine Learning

- Supervised Learning
  - A.K.A. Classification
  - Known labels (subset of rows)
  - Algorithms: label unlabeled rows

- Unsupervised Learning
  - A.K.A Clustering
  - Algorithm: label based on similarity

- Semi-Supervised Learning
  - Do both
Partition Clustering

Each Point in a Unique Class
Centroid Based Clustering

Example Algorithm: K-Means Clustering

Partition
Distribution Based Clustering

Expectation Maximization (EM)

Partition
Density Based Clustering

DBSCAN

Group areas with Certain Density
OPTICS (more general)

Mean Shift

Estimated number of clusters: 3

Partition
Simple Graph Cutting Methods

(1) Similarity Score
(2) Pick Edge Threshold
(3) Cut (only connect stronger edges)
(4) Compute Connected components
Graph Cut Based Spectral Clustering

Partition
Hierarchical Clustering

Example: Ward Clustering
Comparison

Different Clustering Based on Different Assumptions
Bipartite Clustering (Biclustering)

Find Blocks in Data Matrices
Fuzzy Clustering

Membership in Cluster is Floating Point
Clustering Applications

- Clusters can be used to
  - order (pixel based techniques)
  - brush (geometric techniques)
  - aggregate

- Aggregation
  - cluster more homogeneous than whole dataset
  - statistical measures, distributions, etc. more meaningful

http://www.jasondavies.com/parallel-sets/
Clustered Heat Map
F+C Approach, with Dendrograms
Cluster Comparison

Caleydo Matchmaker

Lex, Streit, Partl, Kashofer, Schmalstieg 2010

https://www.youtube.com/watch?v=vi-G3LqHFZA
Aggregation

VisBricks

Heterogeneous Data
Heterogeneous Variables

Caleydo StratomeX

- Header block shows a summary, color indicates data type.
- Band = Intersection of two sets of patients.
- Blocks = Set of patients.
- Row = Gene.
- Column = Patient.
- Patients stratified by copy number.
- Patients stratified by clustering.
- Dependent column showing pathway.
- Dependent column showing survival.
- Dependent column showing clinical variable.
- Query wizard.

https://www.youtube.com/watch?v=s2ZofJ2GVHU
Glyphs
Chernoff Faces
