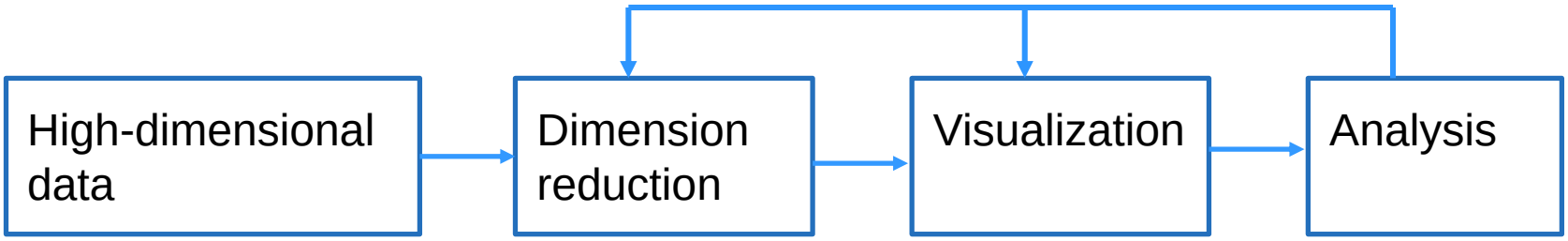
Dimension Reduction



Outline

- Feature selection
- Criteria for DR technique
- Linear Projection Techniques
  - Principle Component Analysis (PCA)
  - Factor Analysis
  - ICA, LDA, CCA
- Non-linear projection techniques
  - Multi-dimensional scaling
  - t-SNE (Stochastic Neighbor Embedding)
- Assisted and Guided Dimension Reduction

Introduction

High dimensional data are

- very hard to analyze with traditional knowledge discovery techniques, such as clustering, and
- very hard to visualize.

Dimension reduction (DR) techniques involve algorithms to either

- Remove irrelevant dimensions (*feature selection*) or
- Transform data in another space where less dimensions are needed to represent the data (*feature extraction/transform*)
  - Data are projected from a HD to a lower-dimensional space with techniques that aim to preserve the original distances.
  - Principal component analysis and Multidimensional Scaling

DR involves a *loss of information*. The goal of DR is to preserve important structures, such as clusters, outliers, correlations and manifolds (a 1D manifold is a line or circle; a 2D manifold a plane, torus or sphere)

Introduction

Dimension reduction aims at generating a LD visual representation that

- enables an overview of the data and their relations,
- supports navigation and browsing when the user supplies a query (with values and ranges for attributes) leading to a 2D point region.

DR builds maps in which distances between points represent similarities in HD data.

DR is often build on objective functions that measure the discrepancy between similarity and distance.

DR benefits from a tight coupling of algorithmic components and visualizations.

Introduction

Relation to subspace clustering:

- A dimension that does not contribute to a clusterable subspace probably is not important.
- A dimension that contributes to clusters where users are confident in representing true phenomena likely should be preserved.

Data has a dimensionality given by the observations and it has a lower *intrinsic dimensionality* that captures most of the variance.

Dimension reduction is about finding the intrinsic dimensionality.

Introduction

- Scores are used, e.g. in credit score assessment (depending on age, income, property) and medicine
- Scores are 1D values based on a weighed combination of several other values, where weighting adjusts e.g. for the different ranges.
- A score does not completely replace all the values it is made of; it involves some approximation.
- It is designed to minimize such errors.
- Scores are a result of dimension reduction and may involve a series of automatic and interactive steps.

Introduction

Scores in medicine:

- Framingham score to compute the risk for a cardiovascular event based on age, blood pressure, smoking behavior, total and HDL cholesterol (6 measures, Link)
- Child–Pugh score to assess the prognosis of chronic liver disease integrates 5 measures

Measure	1 point	2 points	3 points
Total bilirubin, µmol/L (mg/dL)	<34 (<2)	34–50 (2–3)	>50 (>3)
Serum albumin, g/dL	>3.5	2.8–3.5	<2.8
Prothrombin time, prolongation (s)	<4.0	4.0–6.0	> 6.0
Ascites	None	Mild (or suppressed with medication)	Moderate to Severe (or refractory)
Hepatic encephalopathy	None	Grade I–II	Grade III–IV

Points	Class	One year survival	Two year survival
5–6	A	100%	85%
7–9	B	81%	57%
10–15	C	45%	35%

Child-Pugh-Score, From: Link

Feature Selection

- Automatic feature selection is a part of machine learning.
- In the following, we describe
  - One feature selection method and then focus on visual analytics approaches that involve the user, enables her to steer the process based on visualizations of distributions, quality and interestingness measures.

Feature Selection

Correlation-based feature selection (Hall, 1999).

- Start with an empty set of features {S}
- Select a feature f_1 that provides a high *information gain* (a feature with high entropy) and add it $\{S\} \cup f_1$
- Add further features that are not strongly correlated to any feature $f_i \in \{S\}$ and compute the merit value M_S
- Terminate when the merit value M_S decreases.

$M_S \sim 1/ r_{ff}$ (feature-feature dependency, correlation)

$M_S \sim r_{cf}$ the feature-class dependency (see Hielscher, 2014 for a detailed description)

Dimension Reduction

- All DR techniques *minimize* some kind of *error* guided by some constraints, e.g. the distances between all pairs of points in low-D space should have the same relation like in HD space
- Linear techniques generate a new set of dimensions that is a linear combination of the original dimensions.
  - Suitable for data that follows a normal distribution
- Non-linear techniques are suitable for skewed or multimodal (e.g. bimodal) distributions.
- Advanced techniques can incorporate a priori knowledge, e.g. the results of clustering to create a layout with good separability.
- Major techniques (PCA, SNE, MDS) are available in R, Matlab and other libraries.

Dimension Reduction

Preprocessing (Dimension Space):

- Remove dimensions with low variance

  Visual aid: Sorted histogramm of variance in all dimensions

- Replace strongly correlating dimensions with a representative dimension, e.g. the average (after normalization)

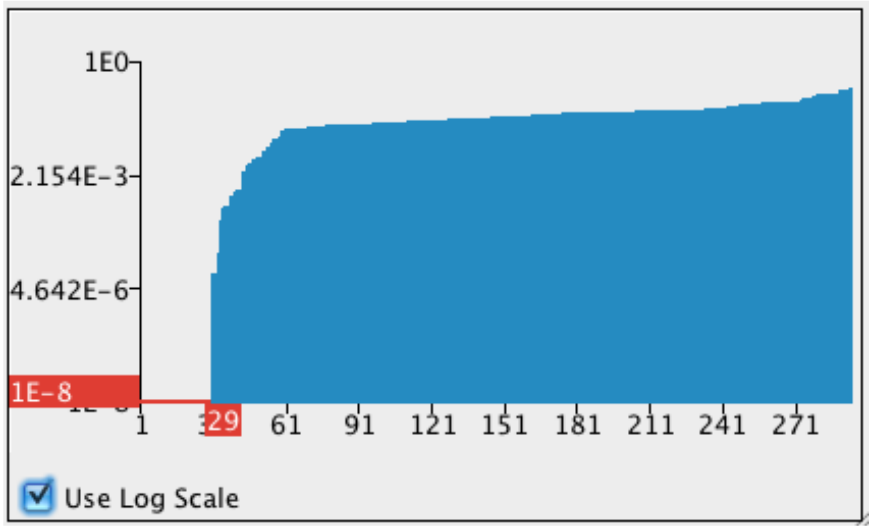
  Visual aid: Color-coded correlation matrix

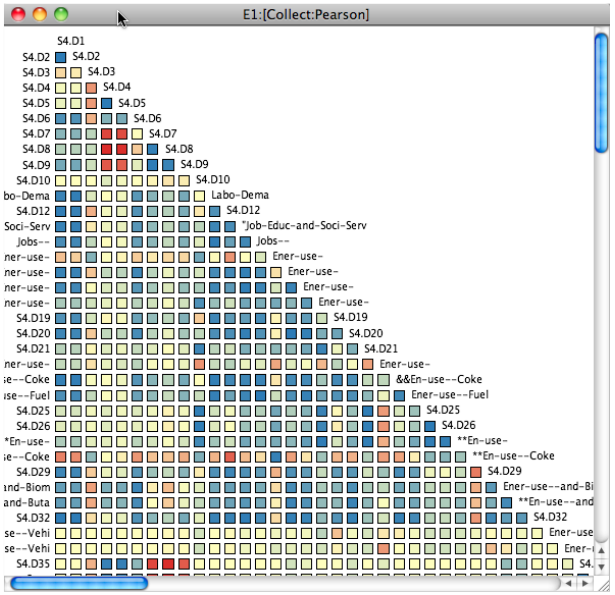
Preprocessing (Item Space)

- Outlier removal. Outliers severely influence the variance in involved dimensions and the correlation. Consider different algorithms.

Dimension Reduction

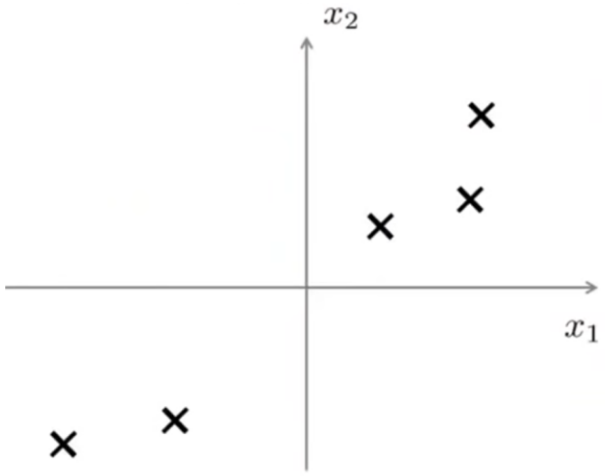
Preprocessing

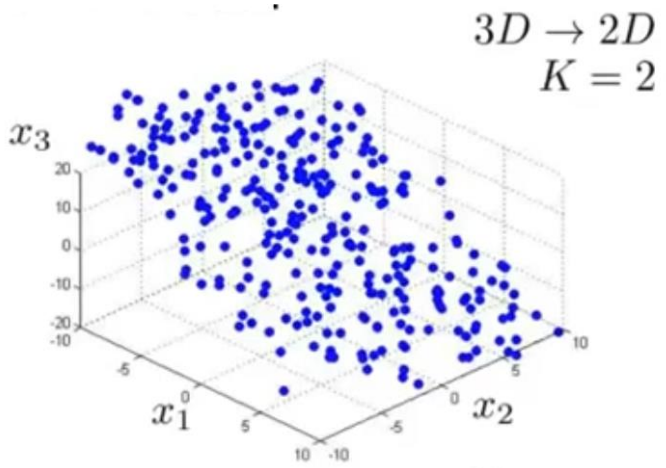


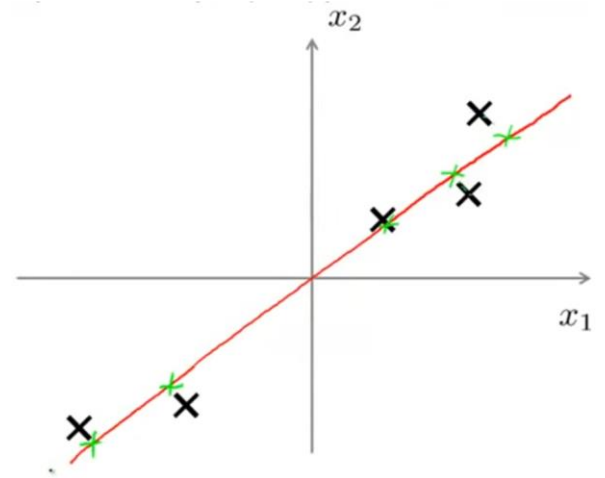


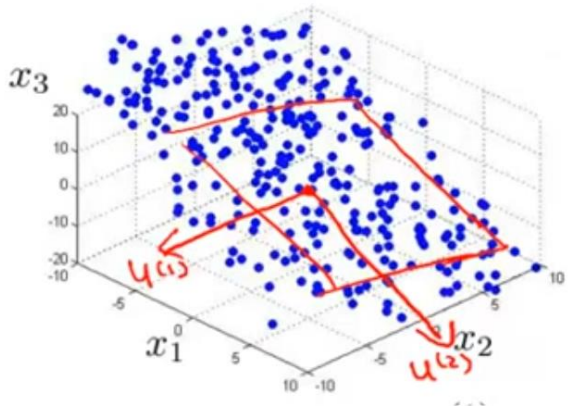
Variance of dimensions in log scale (left) and correlation matrix (right) as visual aids for parameter adjustment, e.g. thresholds for variance and correlation (From: Ingram, 2010).

Dimension Reduction









Points in 2D/3D are projected to lower-dimensional spaces with minimum projection error

Principal Component Analysis

- Based on the original dimensions $x_1, \ldots, x_p$ PCA generates a new coordinate system with orthogonal dimensions (Hotelling, 1933)
- New dimensions (Principle Components, PC) are *linear combinations* of the original dimensions and are *sorted according to variance*
- Each new dimension carries a *loading* that characterizes how much variability of the data is explained.
- Starting from the dimension with the highest loading, take the first *n* dimensions until their cumulative loadings exceed a threshold, e.g. 95% (data compression)
- The projection error (in a least square sense) is minimized for any selection of *n*

Principal Component Analysis

Data Visualization

Country	GDP (trillions of US$)	Per capita GDP (thousands of intl. $)	Human Develop-ment Index	Life expectancy	Poverty Index (Gini as percentage)	Mean household income (thousands of US$)	...
Canada	1.577	39.17	0.908	80.7	32.6	67.293	...
China	5.878	7.54	0.687	73	46.9	10.22	...
India	1.632	3.41	0.547	64.7	36.8	0.735	...
Russia	1.48	19.84	0.755	65.5	39.9	0.72	...
Singapore	0.223	56.69	0.866	80	42.5	67.1	...
USA	14.527	46.86	0.91	78.3	40.8	84.3	...
...	...	...	...	...	...	...	

[resources from en.wikipedia.org]

Andrew Ng

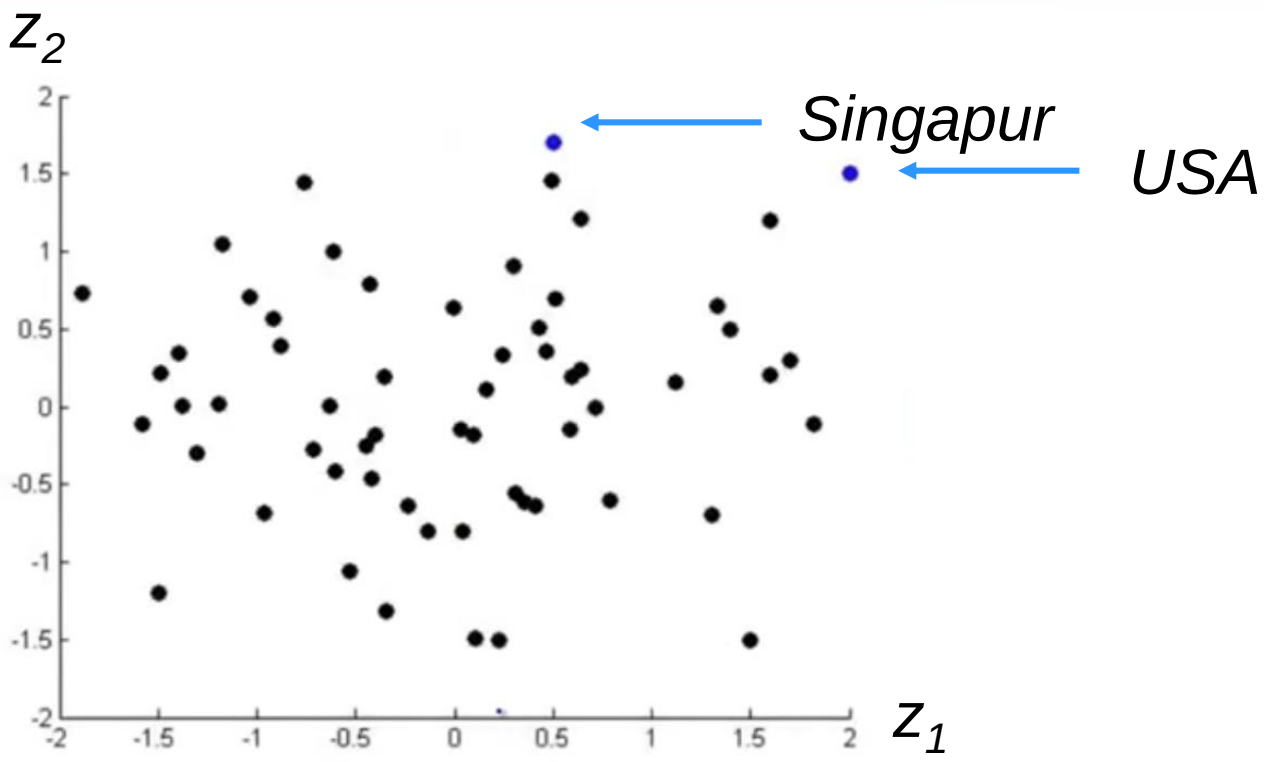
Data Visualization

Country	z_1	z_2
Canada	1.6	1.2
China	1.7	0.3
India	1.6	0.2
Russia	1.4	0.5
Singapore	0.5	1.7
USA	2	1.5
...	...	...

For each country 50 variables are available that characterize well-being and state of development.

PCA enables to reduce the data to 2 dimensions that explain most of the variance.

Principal Component Analysis



Each country now is a point in this new z_1, z_2 coordinate system. What does z_1 and z_2 denote?

z_1 is basically a measure for the *size of the country* and z_2 a measure for the *per person productivity and well-being*.

Principal Component Analysis

Approach:

- Standardize the data
- Determine the covariance matrix $Cov = 1/n\ XX^T$
- Apply an Eigenvalue analysis

$Cov = U\ \lambda U^T$

λ is a diagonal matrix (all elements are zero except the diagonal) with the Eigenvalues $\lambda_1 \leq \ldots \leq \lambda_p$

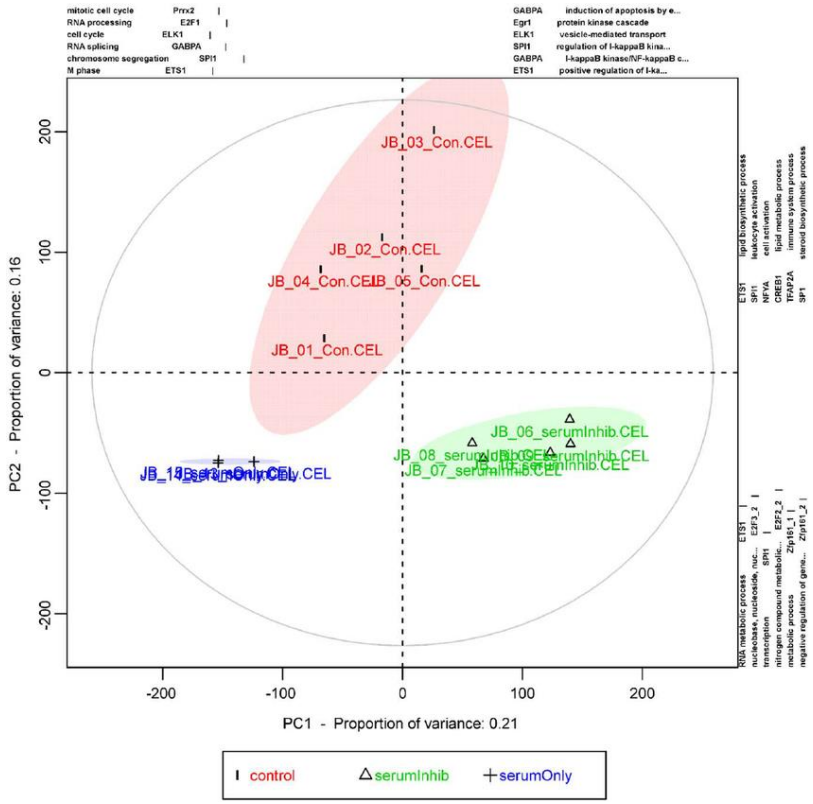
U is a orthogonal matrix containing the Eigenvectors (sorted according to the Eigenvalues).

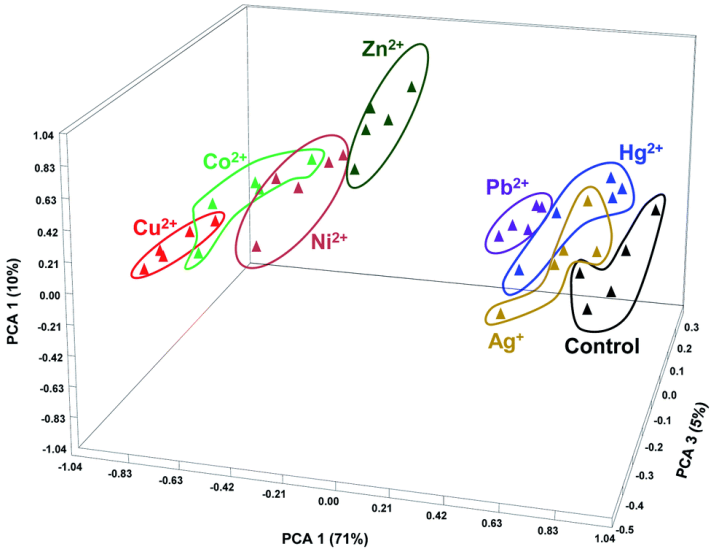
For dimension reduction from R^n to R^k, choose *k* such that the cumulative error is below a threshold (1%, 5%, 10%).

For visualization purposes, *k* is often 2 or 3 and the results are shown in a scatterplot.

Principal Component Analysis

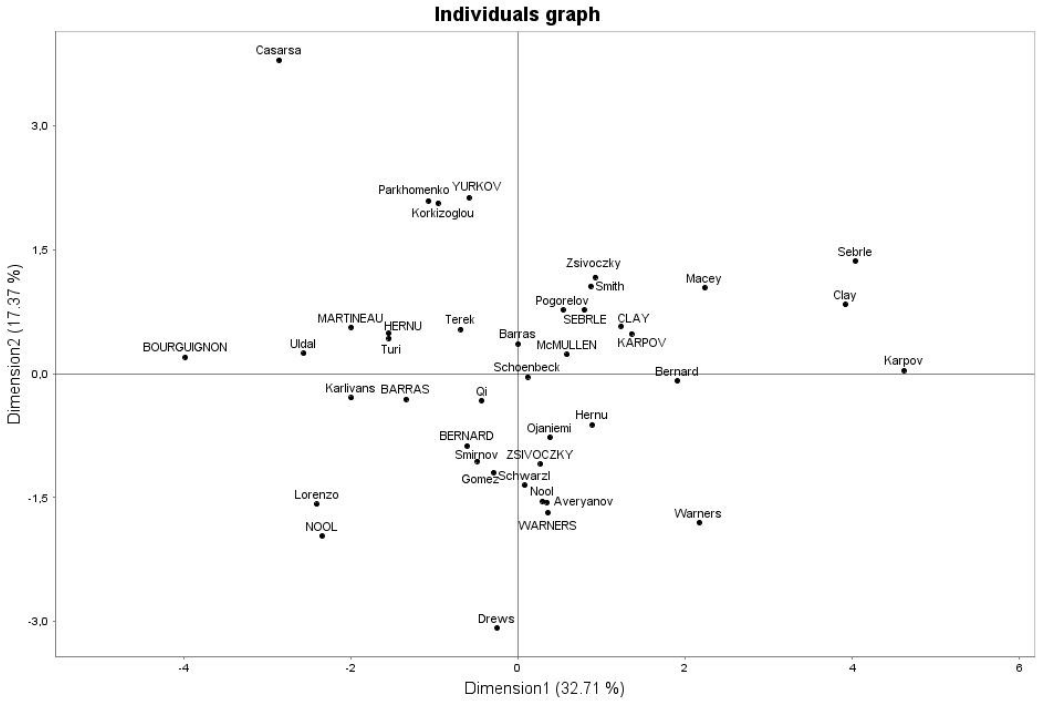
Score plots indicate distribution in the direction of the 2/3 largest components

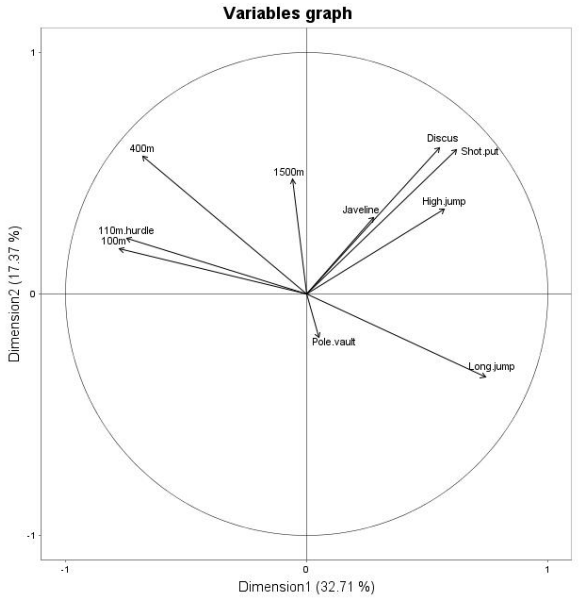




(From: Hansen: 2012)

Principal Component Analysis

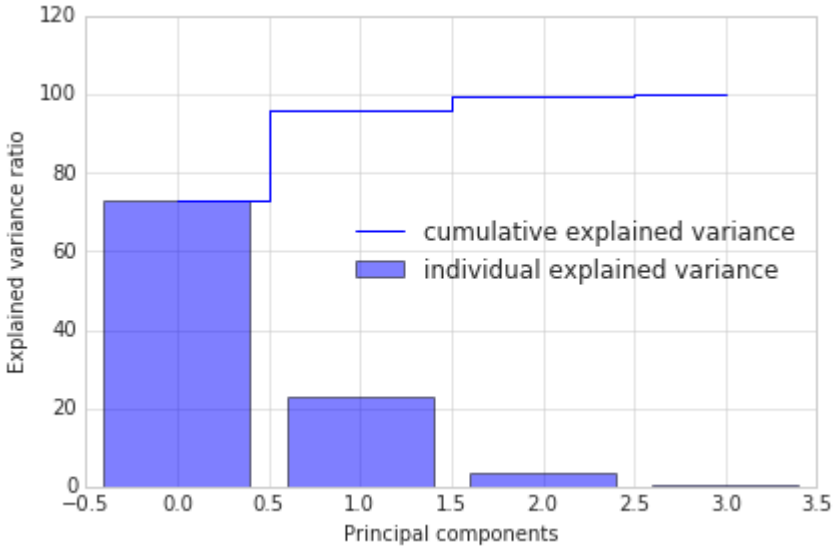


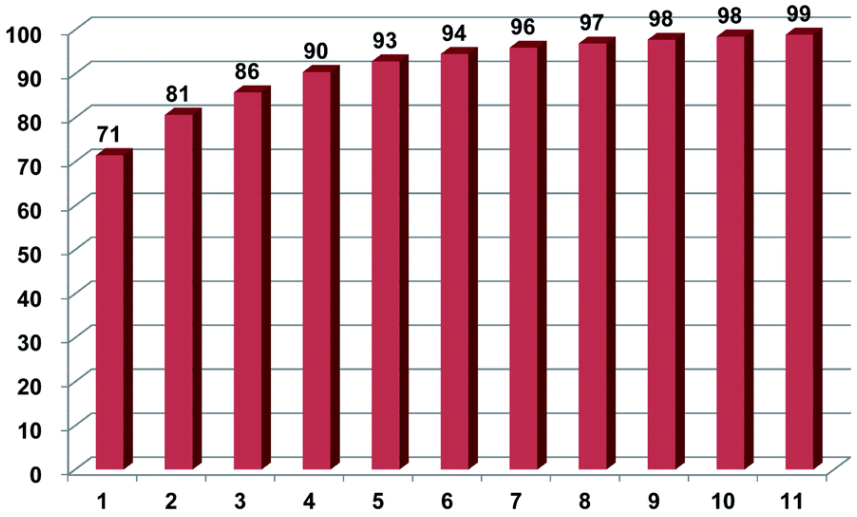


Left: Results of sports man in decathon are analyzed. The 1st PC basically discriminates the best from other sportsman. The 2nd PC discriminates those better in fast running and jumping from the stronger athletes (shot put and discus).

Right: The influence of the different disciplines on the result (From: Link)

Principal Component Analysis



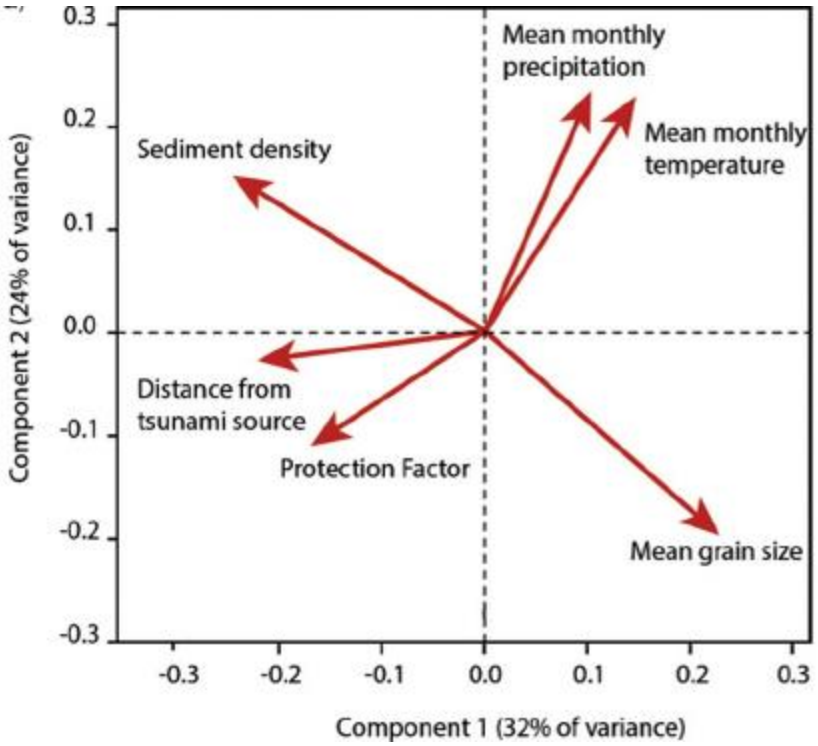


Scree plots: Individual and cumulative variance explained by the first *n* PCs (From: Link)

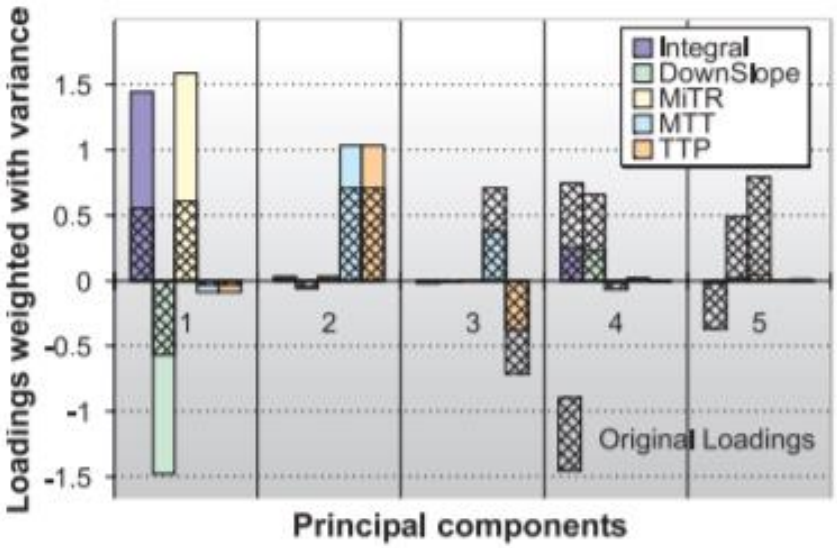
What is missing?

The influence of the original variables on the principal components. What is PC1 related to the original data?

Principal Component Analysis

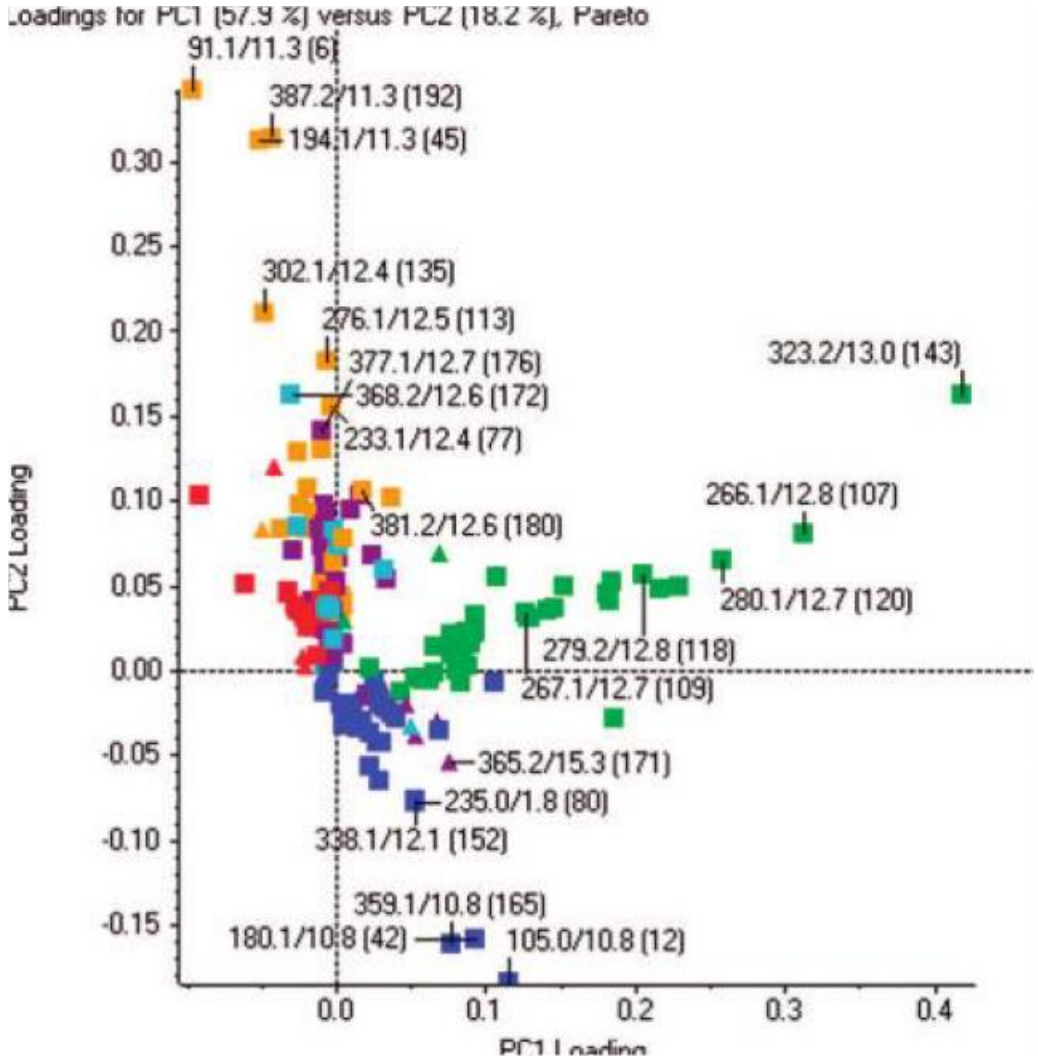


The classical „loading plot“ only indicates how the variables contribute to PC1 and PC2 (From: Kain, 2015).



The original variables (derived from time-dependent medical data are „Integral“, „Downslope“, …). Integral, downslope and MiTR have a strong influence on PC1, MTT and TTP on PC2. The hatched scaling shows the extent after dividing for the variance in this dimension (From: Oeltze, 2007).

Principal Component Analysis



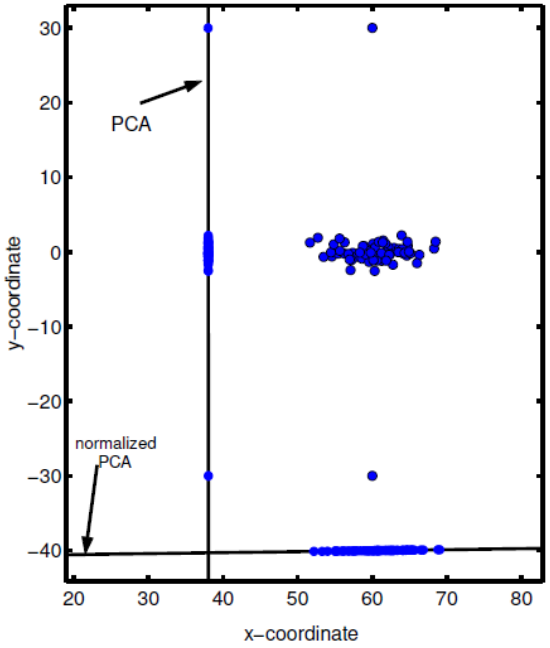
In real-world examples with many dimensions, the loadings plot may get messy.

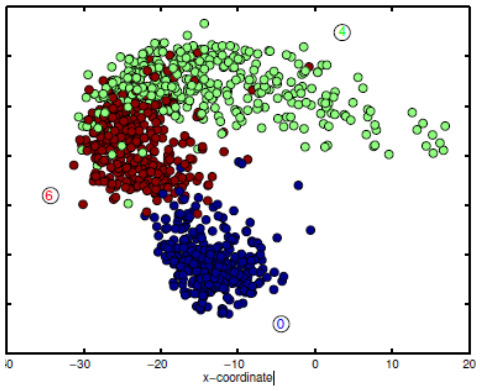
Example from analytical chemistry. Colors indicate classes of measurements, e.g. from mass spectroscopy (From: Ivosev, 2008).

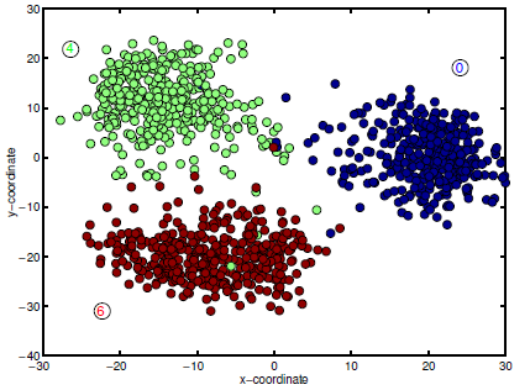
Principal Component Analysis

Outlier sensitivity and cluster preservation

- Classic PCA is highly sensitive to outliers, whereas normalized PCA is more robust.
- Classic PCA does not preserve clusters well, whereas supervised PCA improves separability.

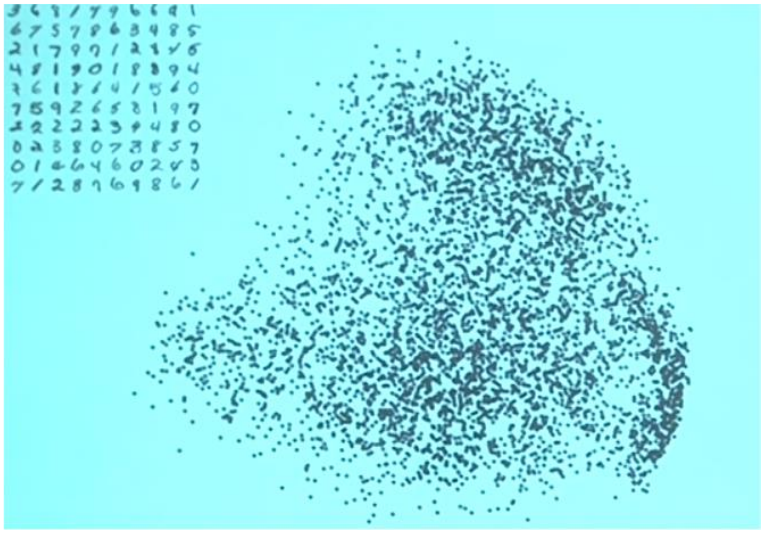


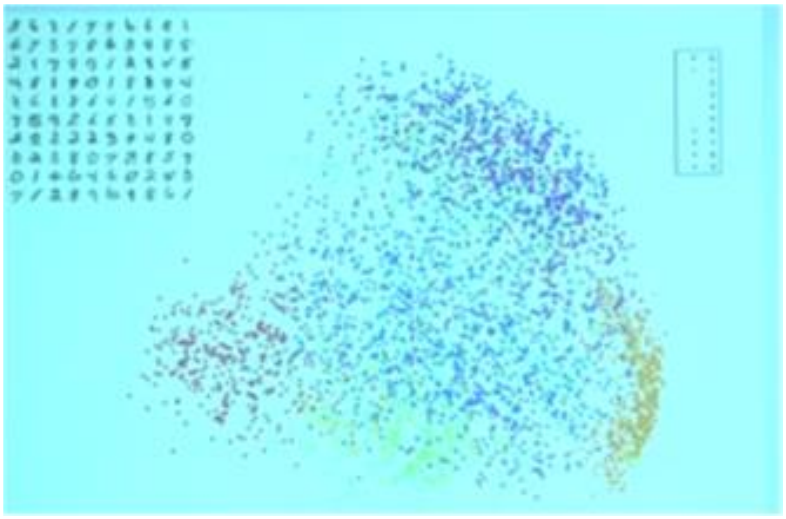




From: Koren, 2004

Principal Component Analysis





Application to the digit dataset. Many handwritten numbers (0, …, 9) are analyzed pixelwise. High overlap means high similarity. Left: Without further help, structure hardly recognizeable. Right: With color-coding, it seems that the digits are separated well. The largest variance (horizontal line) is between 0 and 1 (Courtesy of L. van der Maaten).

Principal Component Analysis

Discussion PCA: (based on Fodor, 2002)

- Linear projection method
- PCA assumes normally distributed data → clusters are not preserved.
- Dimensions need to be normalized (zero means) and scaled (divide by the range or σ) (*autoscaling*)
- Drawback of autoscaling: Noisy measurements are scaled up whereas large peaks in meaningful data get reduced → Pareto scaling as alternative (divide each value by the standard deviation)
- Strongly correlating dimensions hamper the result (should be removed upfront)
- Interpretability of the new dimensions is challenging; often domain scientists are not satisfied.

Principal Component Analysis

Alternative use of PCA:

- Determine a set of dimensions from the original data that explains most of the data:
  - Take the new dimension with the lowest loading and determine the original dimension *d* that contributes most to it. → remove *d*
  - Continue with the second lowest and further loadings until the number of dimensions is reduced sufficiently or a quality criterion is fulfilled.

Principal Component Analysis

Interactive PCA (Jeong, 2009)

Users may

- adjust weights for each dimension
- Manipulate points (e.g. by selecting a rectangular region and draggingit)

(YouTube)

Factor Analysis

Factor analysis

- linear dimension reduction technique based on 2nd order statistics, but no ranking of dimensions (as in PCA).
- **Idea**: Identify common causes (latent factors) of measured data by analyzing correlations.

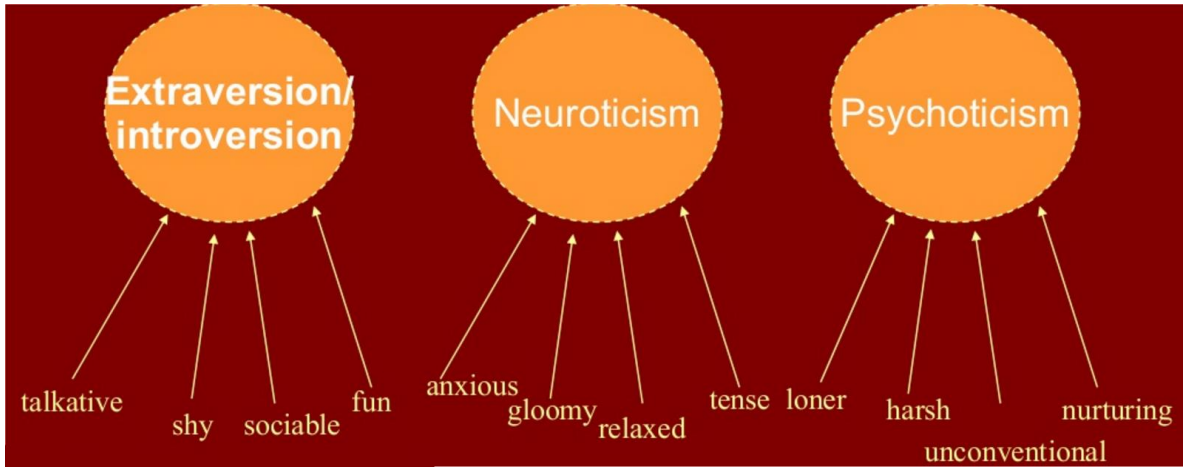
  Variance of the data is explained by smaller number of factors/clusters
  consisting of different variables
- **Origin**: Intelligence tests, i.e. search for verbal intelligence, mathematical intelligence, … based on (many test results from many test persons), e.g. test result ~ 10 x verbal intelligence + 8 x mathematical intelligence. 10 and 8 are loading
- Further applications: Language recognition (decompose different voices)

Factor Analysis

Approach:

- Date is first normalized and centralized (0 mean)
- Pairwise correlations of items are evaluated w.r.t. to the scalar product of the items yielding an angle α (dot product of two vectors). Based on the analysis of α subspaces of highly correlating variables is determined.
- Observed variables are modelled as linear combinations of the potential factors, plus "error" terms.

Factor Analysis

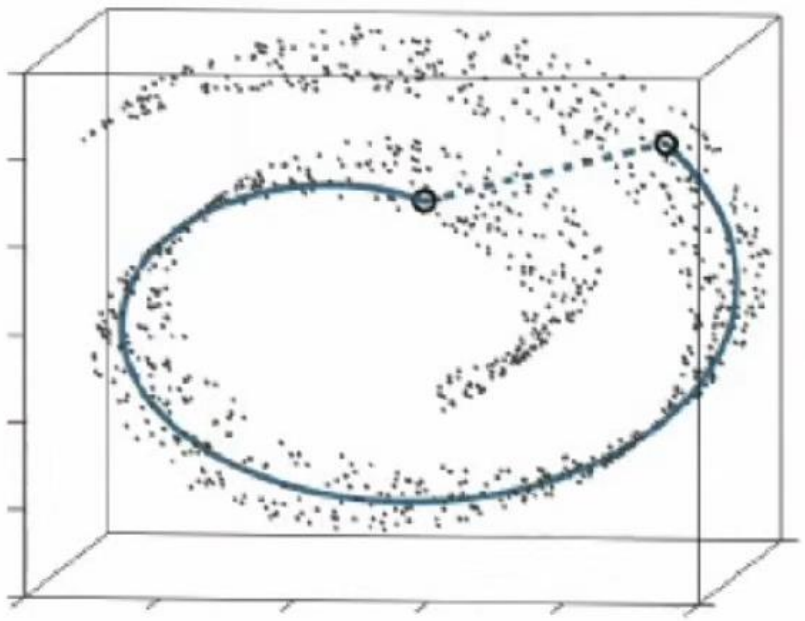


From: Link

- Discussion: Models can be quite complex, i.e. if variables contribute to different factors.
- Computed factors need interpretation

Non-Linear Projection Techniques

- Non-linear techniques are motivated by the fact that large distances are often not interesting or reliable.
- Instead small distances often on a manifold should be preserved.



An example of large distances in the Swiss roll dataset that needs not to be preserved (Courtesy of L. van der Maaten)

Multi-Dimensional Scaling

- Non-linear iterative optimization method where the distance of points in R^n is preserved optimally when transforming to R^k
- Introduced by Torgerson in 1952
- An optimization problem based on a stress function is solved with a non-linear optimization method (e.g. gradient descent, simulated annealing)
- Gradient descent: simpler, but more sensitive to local minima

$$\sum_{j=I\ldots N-1} \sum_{i=i+I..j} \left(\|(x_i, y_i) - (x_j, y_j)\| - d_{ij} \right)^2$$

The true distance of the points in HD space

The distance in LD space

See Video

Multi-Dimensional Scaling

MDS is a similarity-based projection in LD-space →

At the core of MDS is a **distance metric**.

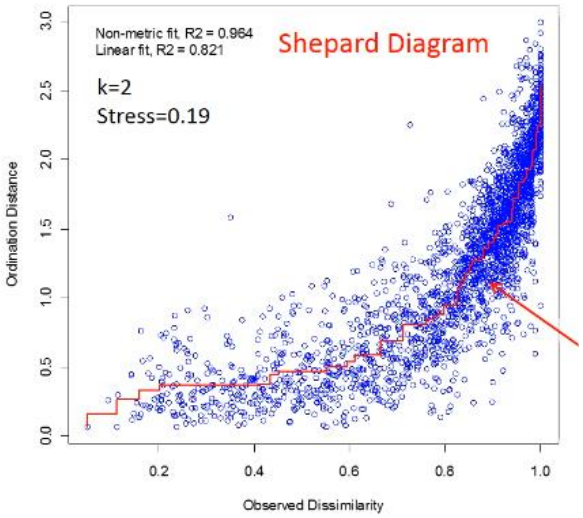
Typical choices are

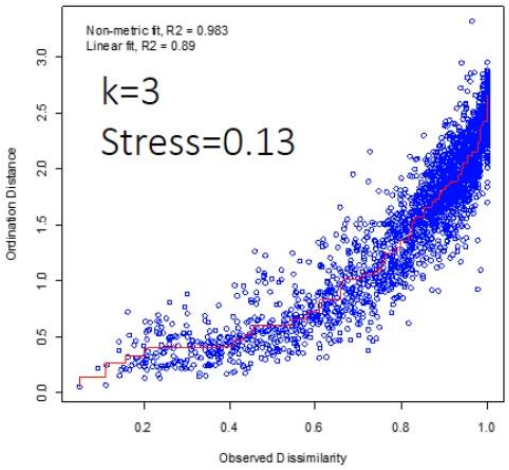
- Euclidean distance,
- city block distance or
- angle between feature vectors, e.g. in text analytics, where large set of documents are shown.

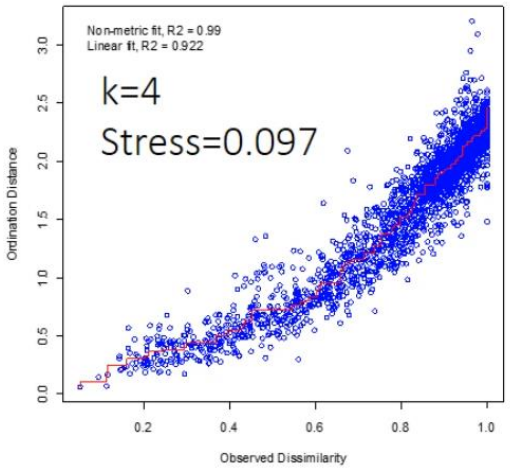
$$\theta = \arccos\left(\frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{x}||\mathbf{y}|}\right) \qquad \mathbf{x}\cdot\mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

Users may select a metric, fix certain points as constraints, employ cluster labels (in the distance metric)

Multi-Dimensional Scaling







MDS introduces an error indicated by the stress value. The Shepard diagram illustrates the error.

With more dimensions, stress reduces and accuracy increases.

Interpretation of stress values: >0,2 poor, 0.1-0.2 fair, 0.05-0,1 good, <0.05 excellent (From: Link).

Multi-Dimensional Scaling

Discussion:

- The stress value indicates (roughly) the (global) quality of the reduction.
- Depending on the analysts‘ question, a layout with slightly higher stress value may be better, e.g. for preserving clusters

Applications:

- MDS is widely used to show the results of subspace clustering (recall A. Tatu‘s work)
- Text analytics (visualization of textual documents with various tags and properties such as creation date, word count, authors, …)

Multi-Dimensional Scaling

MDS requires expensive HD distance and matrix computations.

Remarks on performance:

- Original algorithm has an $O(n^3)$ complexity which is not feasible for large datasets (Torgerson, 1952)
- $O(n^2)$ complexity algorithm that is widely used (Chalmers, 1996)
- Even faster stochastic algorithm (Morrison, 2002)

  Instead of comparing each point with all others, only small sets of points are considered (the top N nearest neighbors v_{max} and a random sample of size s_{max}).

Reduces $N(N-1)$ force computations to $N\,(v_{max} + s_{max})$.

Despite these improvements for really large datasets, MDS is still too slow to support a trial-and-error approach.

Multi-Dimensional Scaling

- Modern algorithms, force-directed layout
  - Forces are defined such that similar objects are pulled together and dissimilar objects push each other away (see video for explanation)
- At some point, the stress function does not decrease any further. This may be a local optimum.

Comparison to PCA:

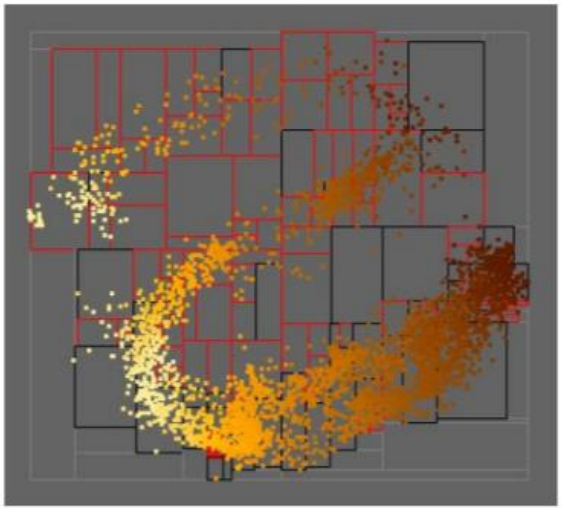
- The axis do not mean anything.

Parameters

  - Termination criteria (certain value of the stress function or certain gradient, i.e. the solution does not improve anymore)
  - Parameters for influencing acceleration (s_{max}, v_{max}). Typical values: $s_{max} = 10$, $v_{max} = 5$

Multi-Dimensional Scaling

Solution: Progressive MDS for a subset of the points, may be terminated or refined at selected positions (Williams, Munzner, 2010)



Based on a hierarchical data structure, first only a few thousand points are projected with the current parameters.

The user may refine locally or wait for the full transformation to be computed (150 seconds for this S-shaped dataset with 50.000 points).

Multi-Dimensional Scaling

Progressive steerable MDS

- Users can
  - immediately start to explore the dataset based on a first overview.
  - drill down and select local areas of particular interest that will be filled very fast.

A global method that is hard to control was transformed in a local method that can be steered.

Projection Pursuit

- Linear dimension reduction algorithm (Friedman, Tukey, 1974)
- Computationally very efficient
- Each dimension is associated with an index that measures how well this dimension separates the data
- Resulting projections (mostly 2D) reveal clusters well
- Projection pursuit may be applied to individual clusters.
- Combination of clustering and projection as an effective tool for exploratory data analysis

Projection Pursuit

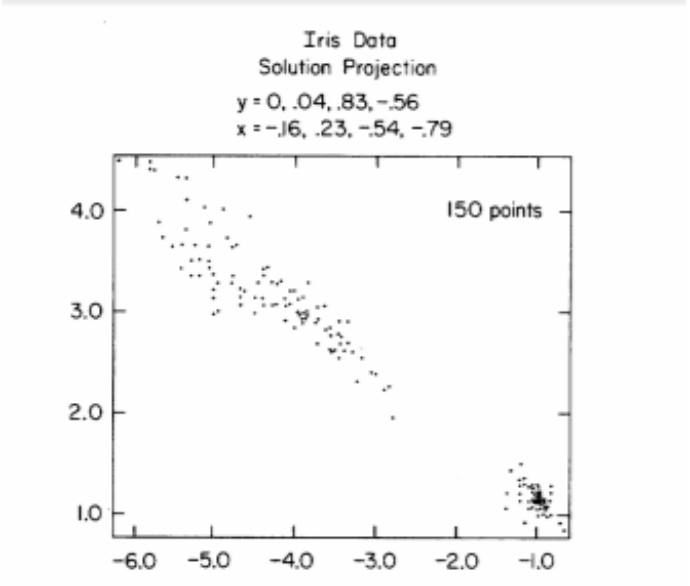
Approach:

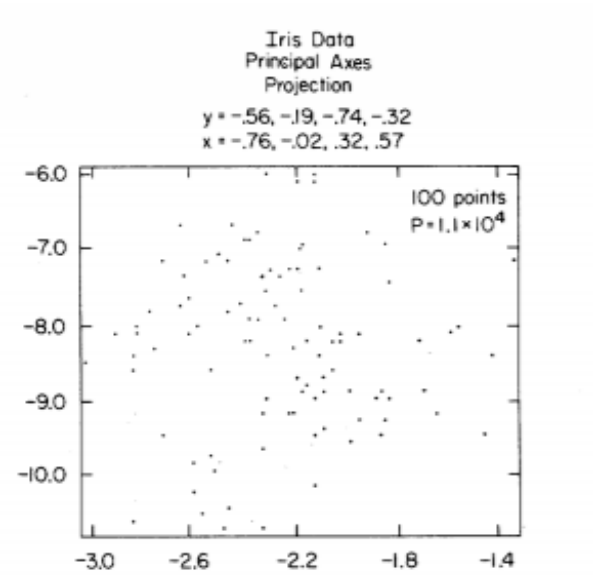
- The authors observed researchers exploring multivariate data, e.g. by rotating 3D views and analyzing which projections were carefully inspected: projections
  - where the data are largely spread (occupying large screen space) and
  - where regions with high local density are recognizeable.
- Projection index is defined as a combined measure (product) of *spread* and *local density*
- *Spread* is a modified (trimmed) version of std. dev.
- *Local density* is defined as the average distance of point pairs within a cutoff radius R
- Preprocessing: Usually, all dimensions should have similar variance.

Projection Pursuit

- The algorithm starts with an initial set of directions, e.g. the principal components, original directions, random choice, ...
- Iteratively, the directions are optimized w.r.t. the projection index until convergence, e.g. improvement <1%
- A local optimum is computed.
- Different starting points and projections are recommended.
- 20-30 iterations are usually sufficient.
- Result is rather robust against the choice of the cutoff radius.

Projection Pursuit





Projection Pursuit analysis applied to the four-dimensional Iris dataset (left). The cluster on the upper left (100 objects) is displayed with 2 principal component directions (From: Friedman, 1974).

Stochastic Neighborhood Embedding

- Major goal: preserve both global and local structure of the data when mapping from HD to 2 or 3 data.
- Global structure: primarily clusters at different scales
- Local structure: distances and neighbors
- Basic technique: Stochastic Neighborhood Embedding (Hinton, Roweis, 2002)
  - Convert HD Euclidean distances between points in conditional probabilities that represent similarities
  - Similarity of x_i and x_j is the cond. Probability $p_{j|i}$ that x_i picks x_j as neighbor.
  - Probability is computed according to a Gaussian centered at x_i.

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/2\sigma_i^2)}$$

Stochastic Neighborhood Embedding

In the same way, for the map points y_i y_j the similarity is computed

$$q_{j|i} = \frac{exp(-||y_i - y_j||^2)}{\sum_{k \neq i} exp(-||y_i - y_k||^2)}$$

The points are iteratively moved until the two distributions are as close as possible.

Difference between distributions is measured with the Kullback Leibler divergence.

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} log \frac{p_{j|i}}{q_{j|i}}$$

Stochastic Neighborhood Embedding

Some notes on the KL divergence:

- KL is not symmetric ($KL(P_i|Q_i) \neq KL(Q_i|P_i)$) (see prev. Eq.)
- KL leads to **high costs** if close points (x_i, x_j) are mapped to distant points (y_i, y_j)

  but to low costs if distant points (x_i, x_j) are mapped to close points (y_i, y_j) → local similarity is preserved well

Essential parameter is the variance of the Gaussian σ:

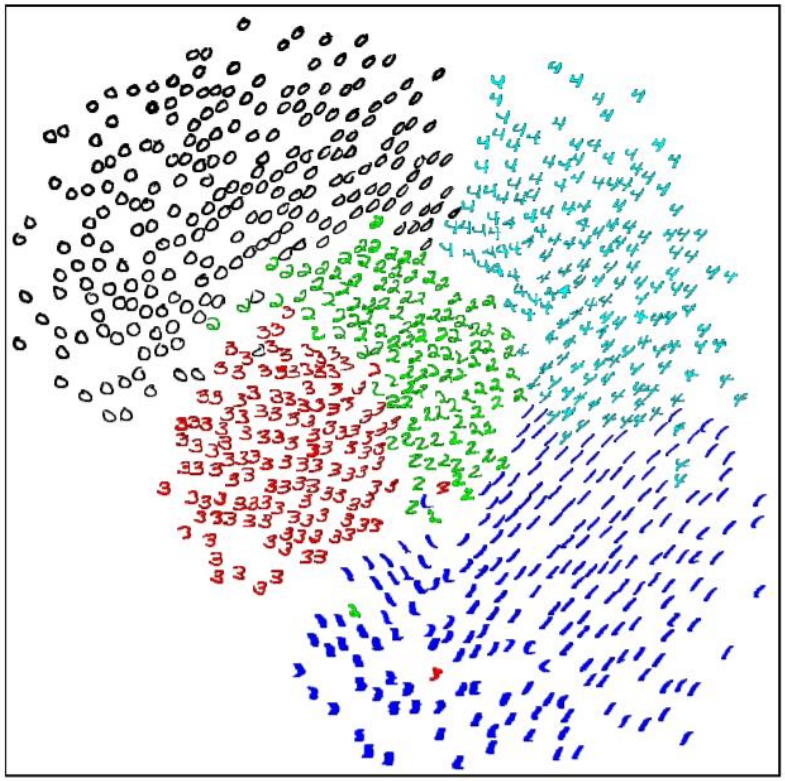
- Choice of σ is locally adapted to the density: Low σ in high-density areas and high σ in sparse regions.
- Value is computed based on Shannon entropy.
- Results are stable against small changes of σ.

Stochastic Neighborhood Embedding

Optimization:

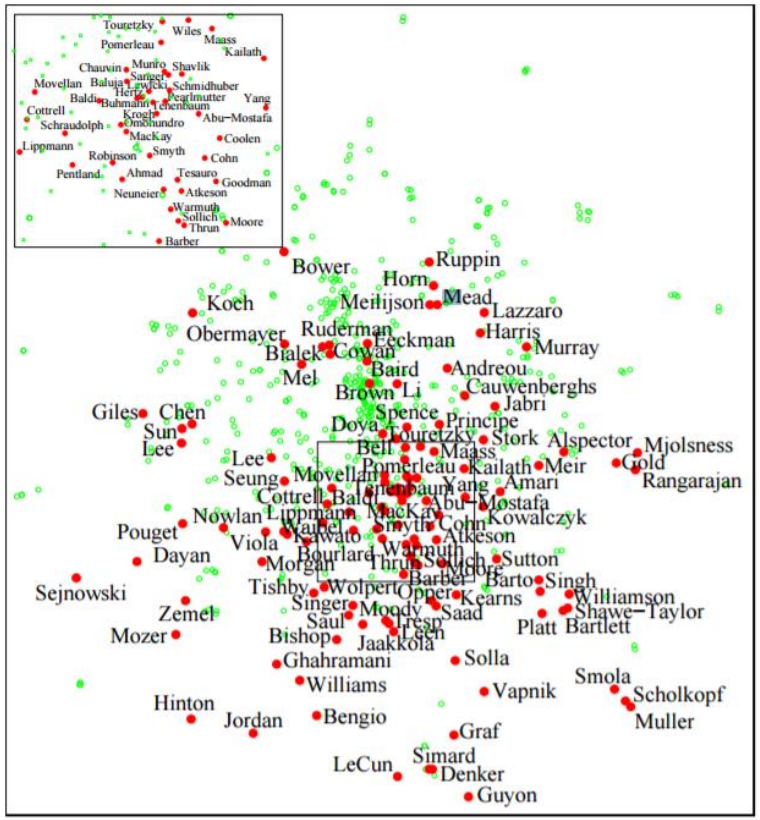
- Naive implementation often leads to poor local minima.
- **Refinement**: Add Gaussian noise to the points and reduce the amount of noise in every step.
- Much better results are possible but two parameters need to be adjusted: the initial amount of noise and the decay.
- Authors suggest to rerun the algorithm several times to determine good values.

Stochastic Neighborhood Embedding



Result of running the SNE algorithm on 3000 256-dimensional grayscale images of handwritten digits represented as vectors xi The classes are separated even SNE had no information about class labels (From Hinton, Roweis, 2002)

Stochastic Neighborhood Embedding



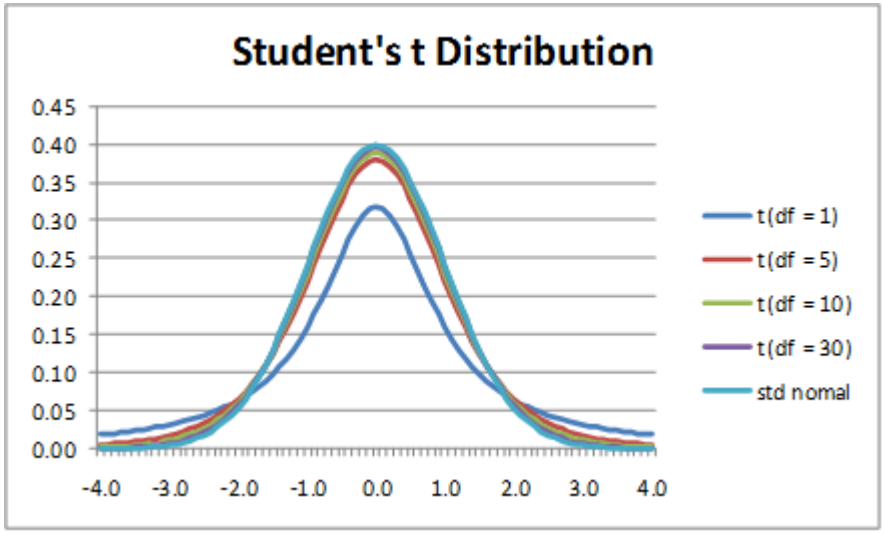
Crowding problem: too many points (authors, co-author info) are mapped centrally (Courtesy L. van der Maaten).

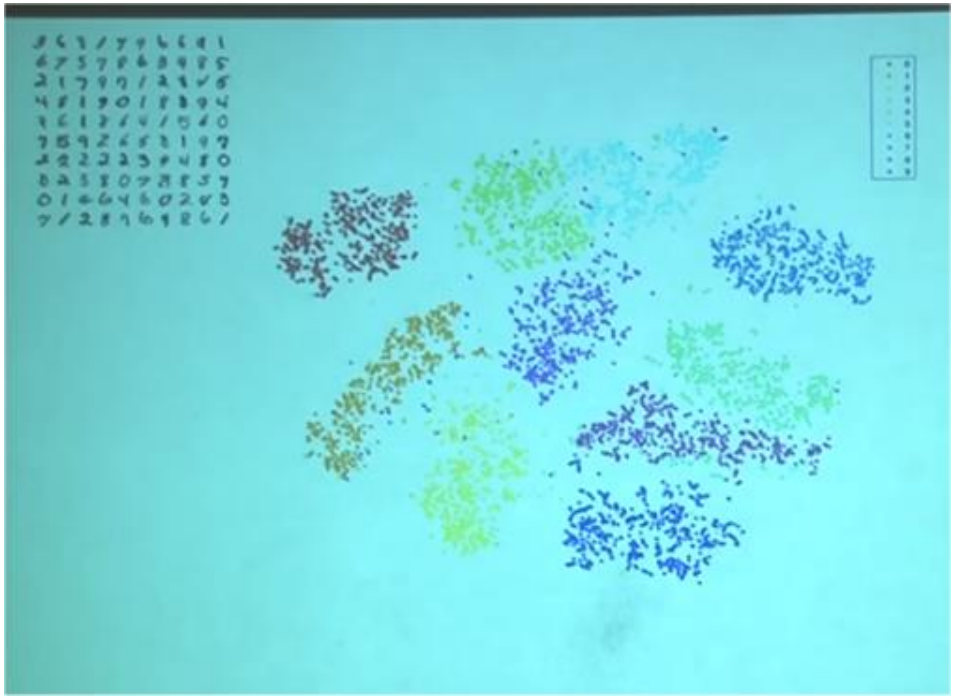
t-SNE

- t-SNE enables a different measure of the distribution of the map points, namely the student t-distribution with one DOF (van der Maaten, 2008)

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$$

- The t-distribution has a wider tail and thus leads to a reduction of the crowding problem





With t-SNE, the hand-written digits are nicely clustered

(Courtesy L. van der Maaten).

Grand Tour

Empirical Guidance

Assessment of Projection Quality

- Projection involves an approximation error.
- A global metric, such as the stress value, does not convey strong local variations of this error.
- Attempts were made to visualize the distortion related to a projection, e.g. by computing a voronoi diagramm and measuring distortion per point leading to color coding of the voronoi cell.

Assisted Dimension Reduction

Instead of largely automatic processes with little user input, assisted techniques serve the user better.

We discuss four approaches:

- *Visual hierarchical DR* (Yang, 2002)
  A hierachy is created by means of dissimilarity and used to select dimensions
- *Interactive DR through user-defined quality metrics* (Johannson, 2009)
  QM relates to the preservation of correlation, clusters, …
- *Assisted search for multiple subspaces* (Krause, 2016)
- *Guidance* for dimensional reduction (Ingram, 2010)
  More comprehensive support, including filtering, feature transformation, …

Assisted Dimension Reduction

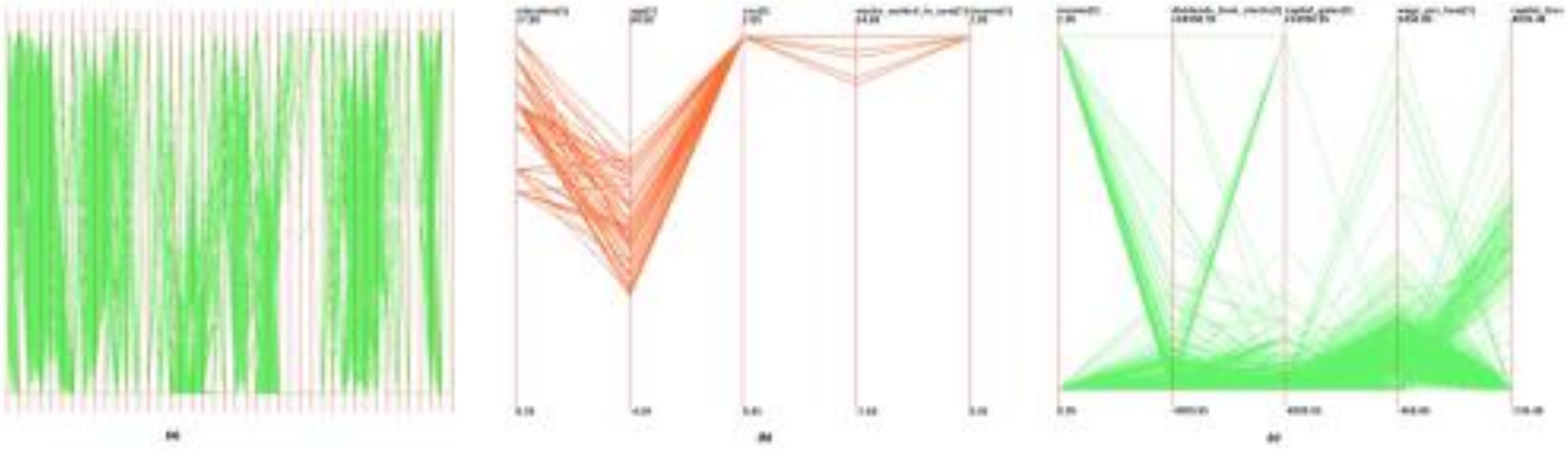
Visual hierarchical DR (Yang, 2002):

- Dimensions are analyzed w.r.t. similarity and are clustered.
- Clustering is performed with different similarity thresholds, leading to a hierarchy with low and high level clusters
- For each low level cluster, a representative dimension is computed automatically or selected by the user.
- Dimension hierarchy is navigated to select/deselect clusters and dimensions.
- Results are shown with PC plots, SP matrices

Assisted Dimension Reduction

- Similarity between dimensions: Dimensions are centralized (zero means), normalized (range -1, 1) and values are sorted for any pair of dimensions (leading to sequences i_1, i_2 , i_3, and j_1, j_2, j_3)
  - An alternative, computationally more expensive, would be a correlation coefficient.
- A pair of dimensions is similar if a certain portion α of corresponding value pairs (i_k, j_k) has an absolute distance below a threshold θ.
- With an increasing threshold, higher level clusters are obtained.

Assisted Dimension Reduction



The census dataset with 42 dimensions is too large for detailed inspection (left). Two subspaces of similar dimensions are analyzed (middle, right). Individual items are now visible (From: Yang, 2002).

Assisted Dimension Reduction

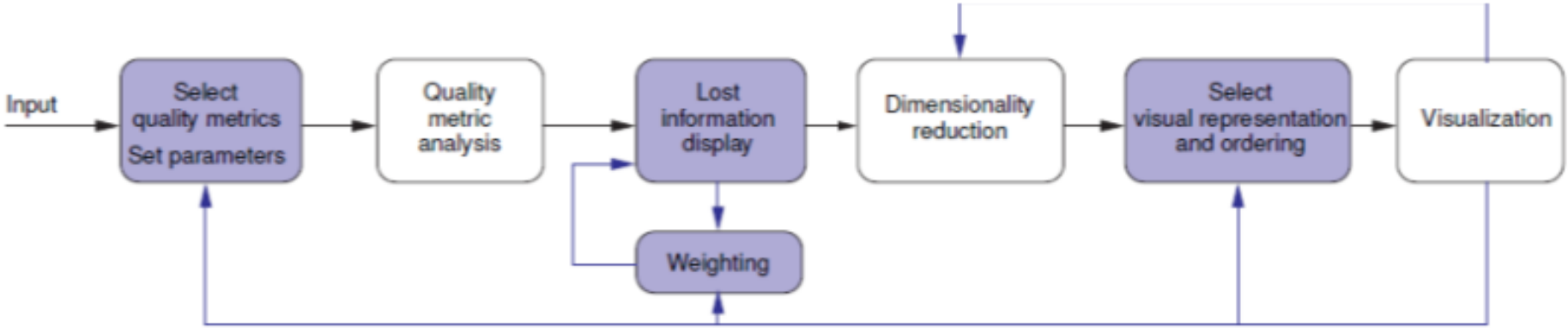
DR through user-defined quality metrics (Johannson, 2009)

Goal: From a set of dimensions {M} reduce to a subset {K} by removing |M| - |K| dimensions

- Neither new dimensions are created (as in PCA) nor projection techniques are applied (as in MDS)
- Lower dimensional result is explored with InfoVis. Techniques, such as Parallel coordinates plots and Scatter plot matrices

Strategy: Since there is no single best way to perform the reduction, consider distinct features that may be preserved and support the selection of the user with related quality metrics.

Assisted Dimension Reduction



Workflow of the dimensionality reduction (blue boxes are specified by the user) (From: Johansson, 2009)

Correlation preservation:

Pearson correlation is computed for each pair of variables $r(x_i,x_j)$.

If $|r(x_i,x_j)| < \theta$ summarize the pair to one variable

Assisted Dimension Reduction

Outlier detection (acc. Johansson, 2009).

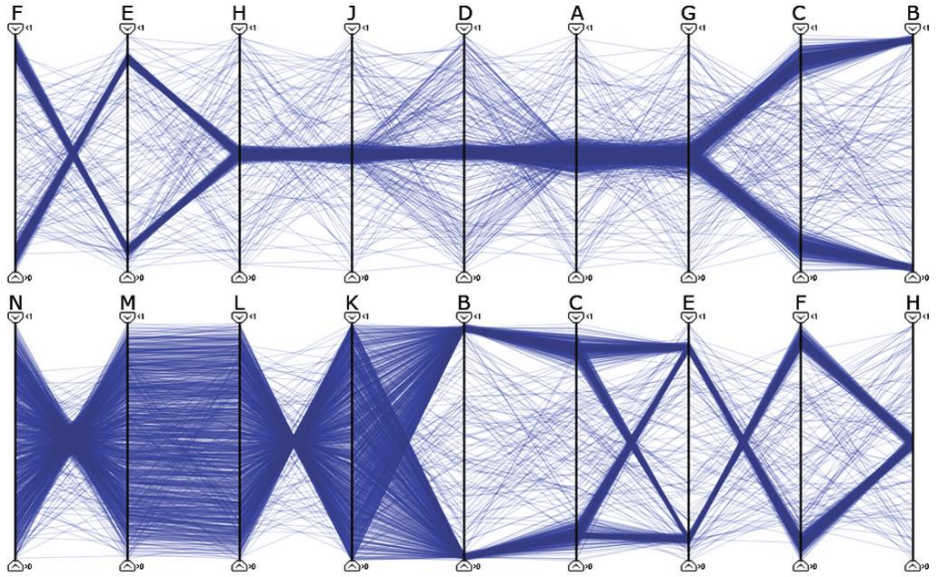
- Outliers are determined with a density-based approach (an outlier has very few neighbors in a certain distance) and a grid-based approach (an outlier is one of very few elements of a cell).
- Outliers are computed for all pairs of variables and for higher dimensions (3, 4, 5, up to a maximum).
- For each element, it is recorded for which variable combination it is an outlier.
- *Outlier preservation rate*: Portion of outliers in HD space that are preserved after reduction to LD space.

Assisted Dimension Reduction

Cluster Detection:

- A density-based clustering approach detects all subspace clusters up to a maximum number of dimensions.
- Clusters are assessed w.r.t. significance according to the cluster dominance factor (CD) (recall Kailing, 2004)
- Ideally, only dimensions are removed that do not contribute to a subspace cluster.
- Information loss relates to the portion of dimensions contributing to an SS cluster that are removed (weighting according to the CD factor).

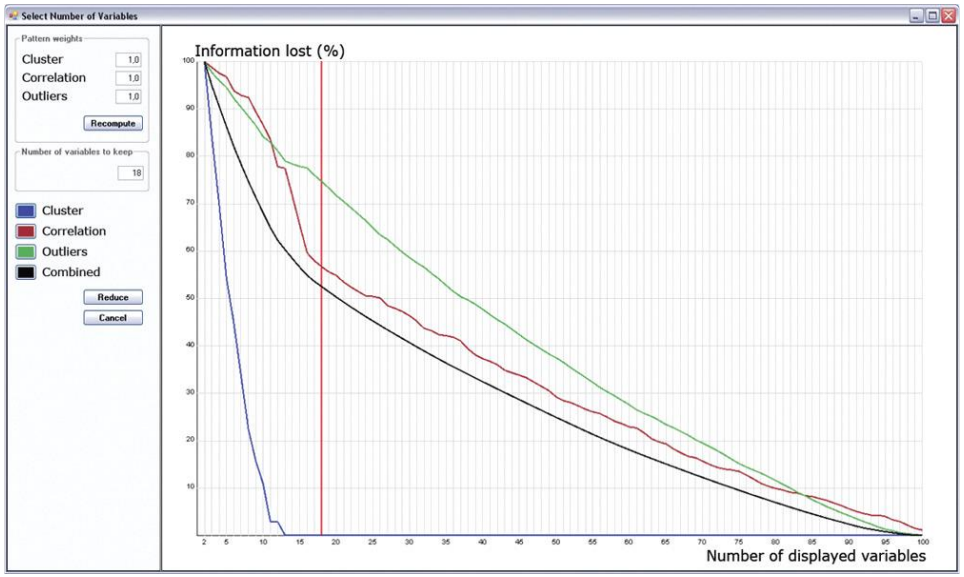
Assisted Dimension Reduction



A synthetic data set reduced to 9 variables using different quality metric weights and variable orders.

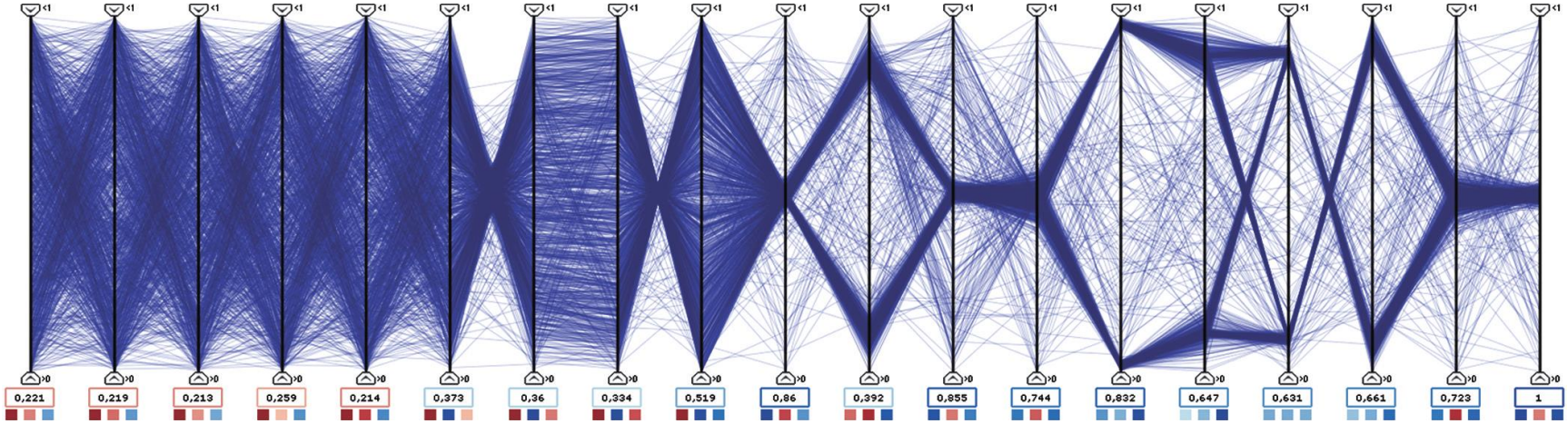
Top view: clustering is assigned a large weight and variables are ordered to enhance cluster structures.
Bottom view: a corresponding weighting and ordering is made for correlation structures (From: Johansson, 2009).

Assisted Dimension Reduction



A diagramme conveys the information loss involved in removing variables. The loss is categorized according to clusters, outliers, … and enables to select a target number that preserves the desired aspects well (From: Johansson, 2009).

Assisted Dimension Reduction



Visual aids below the dimensions: the five leftmost dimensions contain primarily noise (low quality value and red color of the left icons). Importance for clustering and outlier are represented by the two other icons (From: Johansson, 2009).

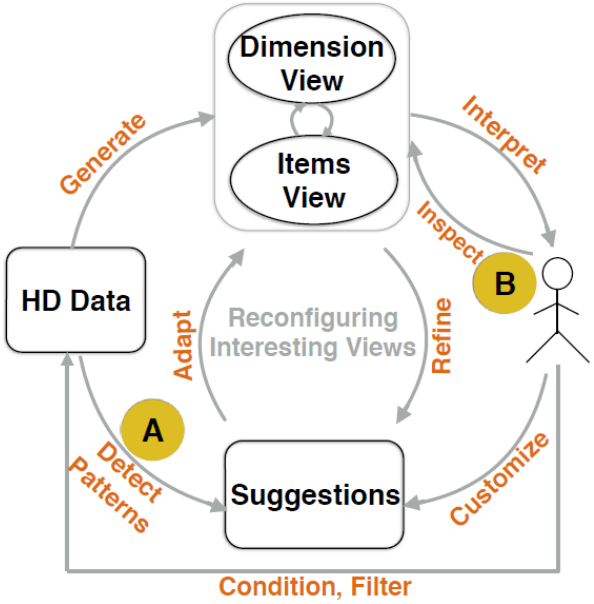
Assisted Dimension Reduction

So far, assistance was discussed for the reduction to *one* subspace.

- Often there are more (smaller) subspaces that are relevant depending on the analysts‘ questions (like in subspace clustering)
- Subspace clustering is focussed on one analytical aspect: finding dense areas
- For other visual patterns, e.g. outliers, (also negative) correlations other subspaces are relevant.

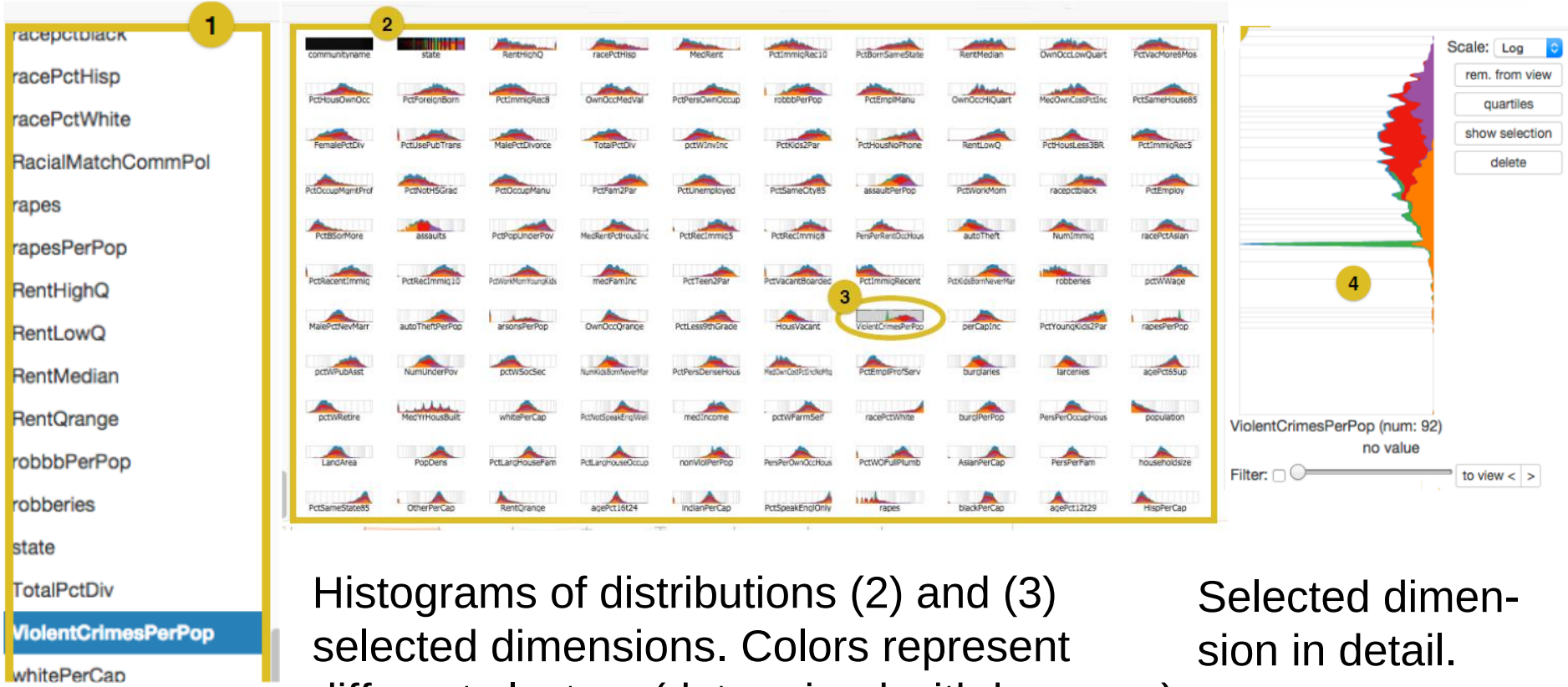
Goal: Assist the user to find interesting subspaces.
Consider also reduction in attribute space (Krause, 2016)

Assisted Dimension Reduction



- A combination of visual representations (views) is used to interpret subspaces.
- Filters enable reduction in item and dimension space.
- Reduction may be started by automatically detected patterns (A) or by user-defined subspaces.
- Subspaces are iteratively refined (From: Krause, 2016).

Assisted Dimension Reduction



Histograms of distributions (2) and (3) selected dimensions. Colors represent different clusters (determined with k means)

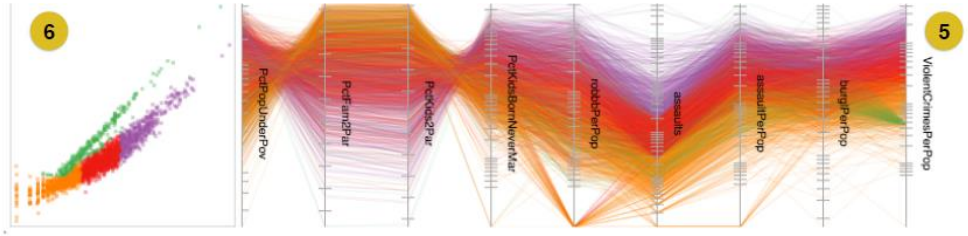
List of dimensions

Selected dimension in detail.

A filter can be applied to partition the values (4)

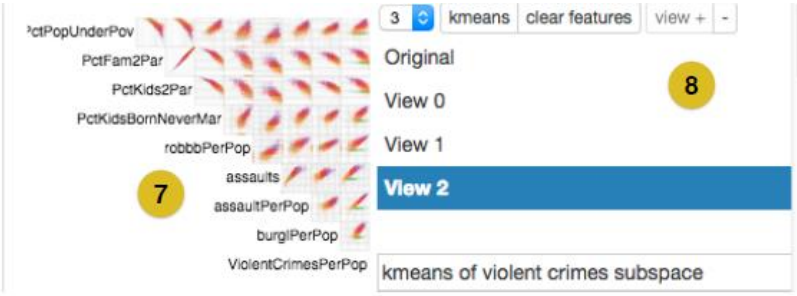
(From: Krause, 2016)

Assisted Dimension Reduction



Further parts of the „Seek a view“ UI:
(5) A parallel coordinates plot of the currently selected subspace.

(6) The first two principal components

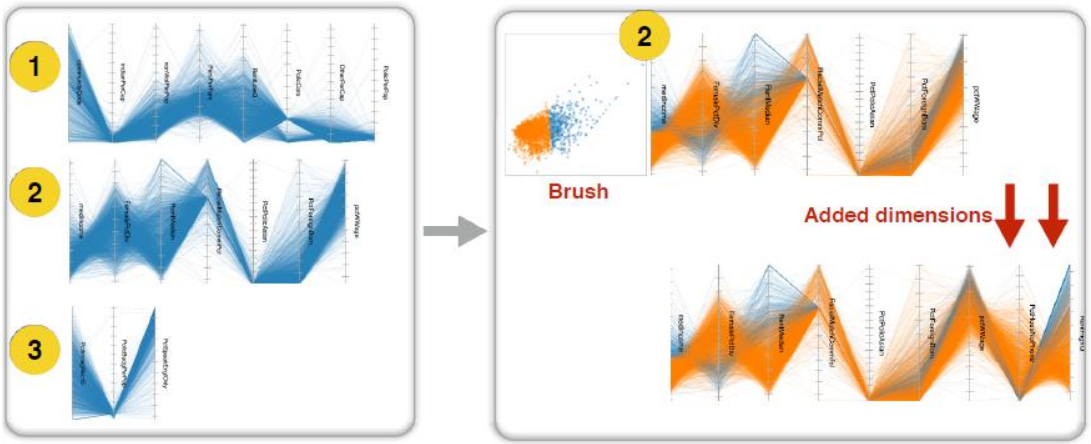


A scatterplot matrix (7).

The current view may be labeled and added.

(From: Krause, 2016)

Assisted Dimension Reduction



Subspaces are presented separately (left). A PCA visualizations is used for brushing and selecting dimensions to be added (right) (From: Krause, 2016).

Assisted Dimension Reduction

Evaluation of Seek a View (Krause, 2016):

- Employ a crime data set from the US FBI report (2000 items, 147 dimensions).
- **Analysts‘s question**: What affects number of violent crimes per population most?
- Subspace clustering reveals clusters with different amount of robberies.
- A clear relation to „number of homeless people“ occurs.
- Also the family situation, e.g. „portion of kids with two parents“, „portion of kids never married“ is related to crime rate.

Guidance

Dim. Reduction involves many steps:

- Feature transformation (log, sqrt, …)
- Normalization
- Filtering or subset selection (item space)
- Annotation of data, e.g. with clustering results
- Search for highly correlated variables (to merge),
- Search for variables with low variance (to remove),
- Selection of a technique,
- Adjustment of parameters,
- Visual analysis of results
- Further loops of this process or parts thereof.

Guidance

Why guidance is needed?

- Large number of DR methods,
- Complexity of the underlying mathematics,
- Complex interaction between choices by the user and properties of the data → a single set of techniques/ params often cannot be reused for other data.

DimStiller (Ingram, 2010):

- Suggests steps to reduce the data to their *intrinsic dimensions,* showing how parameters influence this estimation.
- Guidance based on predefined workflows that can be (re)used, adapted (e.g. w.r.t. parameters), and stored.

Guidance

DimStiller - Target users:

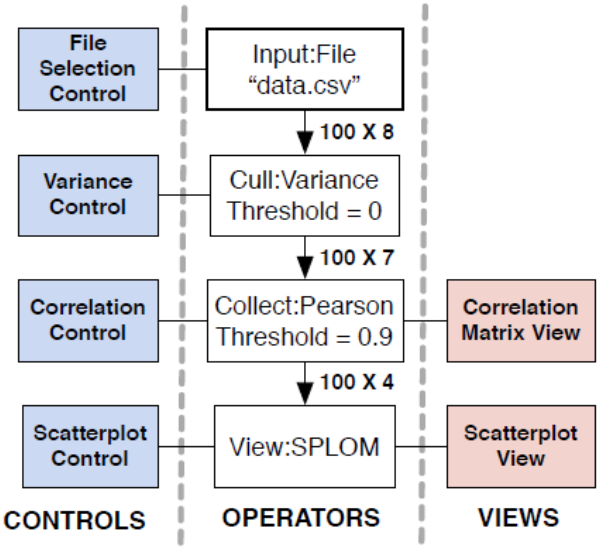
- Users with intermediate competence.
- No professionals of data analysis and no novices without any understanding of concepts/methods for data analysis.

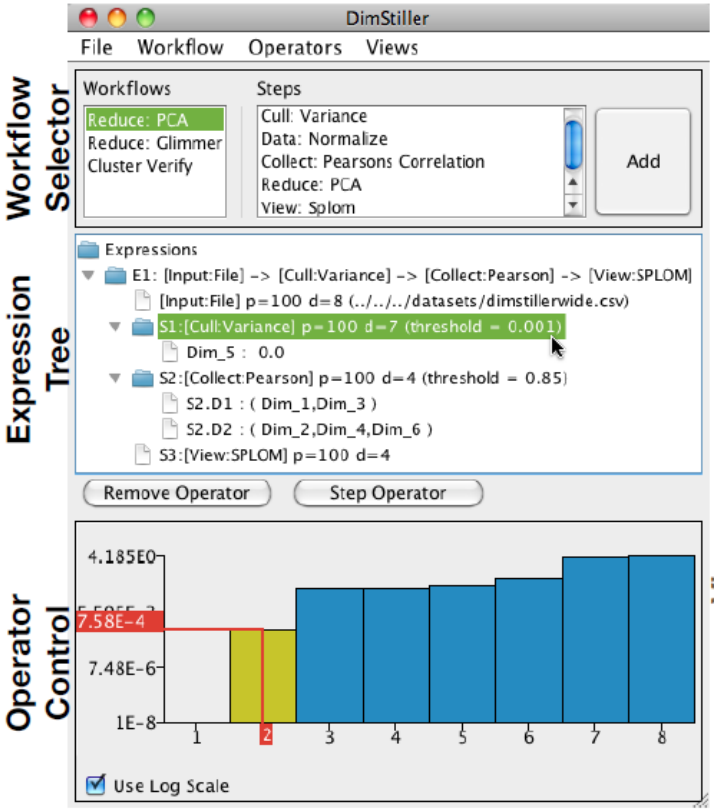
Tasks:

- Analysis of messy real-world datasets with parts of the data being very noisy, quality problems leading to outliers, many dimensions and items.

A set of workflows is provided as *optional support*.

Guidance





Dimstiller: Concept (left). Variance and correlation control reduce from 8 to 4 dimensions. User Interface (right) where the PCA workflow is selected (From: Ingram, 2010).

Summary

- Dimension reduction is a process of *data compression* and benefits from powerful algorithms and expressive visual aids.
- Often just one (2D/3D) view is generated; no matter what the intrinsic dimensions of the data is.
- Variance in dimensions and correlations guide DR algorithms.
- Often, DR creates synthetic dimensions (PCA, MDS) that may capture the structure, but are hard to interpret.
- Visual aids indicate the consequences of adding/removing dimensions or removing outliers.
- Instead of a *single* reduced set of dimensions, several subsets may help to reveal all interesting patterns in the data.
- Guidance is essential to support the whole process including preprocessing, method selection, refinement.

Essential Links

- Multidimensional Scaling
- Lecture from L. van der Maaten (t-SNE)

Outlook

Aim at a better understanding

- how DR is actually used in real analytic situations with large messy real world data.
- of the perceptual consequences of the DR method and subsequent visualization
  - Can users correctly detect all clusters? Can they sort clusters according to size and to distance? How fast are they?
  - Task-based experiments and eye tracking studies may help to get background.

References

M Chalmers. “A linear iteration time layout algorithm for visualising high-dimensional data”, *Proc. of IEEE Visualization'96*., pp. 127-131, 1996.

I. K. Fodor. A survey of dimension reduction techniques. *Technical Report UCRL-ID-148494*, Lawrence Livermore National Laboratory, 2002

M. Hall 1999, Correlation-based Feature Selection for Machine Learning

M Hansen, TA Gerds, OH Nielsen, JB Seidelin, JT Troelsen, J Olsen. “pcaGoPromoter-an R package for biological and regulatory interpretation of principal components in genome-wide gene expression data”, *PloS one* 7 (2), e32394, 2012

Tommy Hielscher, Myra Spiliopoulou, Henry Völzke, Jens-Peter Kühn: Using Participant Similarity for the Classification of Epidemiological Data on Hepatic Steatosis. *CBMS 2014*: 1-7

G Hinton, S Roweis. “Stochastic neighbor embedding”, *Advances in neural information processing systems* 15, 833-840

HOTELLING, H., 1933. Analysis of a Complex of Statistical Variables Into Principal Components, *Journal of Educational Psychology*, volume 24, pages 417-441

S Ingram, T Munzner, M Olano.„Glimmer: „Multilevel MDS on the GPU“, *IEEE Transactions on Visualization and Computer Graphics* 15 (2), 249-261, 2009

S Ingram, T Munzner, V Irvine, M Tory, S Bergner, T Möller. „DimStiller: Workflows for dimensional analysis and reduction“, *Proc. of IEEE Symposium on Visual Analytics Science and Technology (VAST),* 3-10, 2010.

G. Ivosev, L. Burton, R. Bonner. “Dimensionality Reduction and Visualization in Principal Component Analysis”, *Anal. Chem.* **,** *80,* 4933–4944

S Johansson, J Johansson. “Interactive dimensionality reduction through user-defined combinations of quality metrics”, *IEEE Transactions on Visualization and Computer Graphics* 15 (6), 993-1000, 2009

References (II)

C. Kain, C. Gomez, D. Hart, C. Chague-Goff. „Analysis of environmental controls on tsunami deposit texture“, *Marine Geology*, Volume 368(1): 1–14, 2015.

Koren Y, Carmel L. “Robust linear dimensionality reduction*”, IEEE Trans Vis Comput Graph.* 2004 Jul-Aug;10(4):459-70

J Krause, A Dasgupta, JD Fekete, E Bertini. „SeekAView: An Intelligent Dimensionality Reduction Strategy for Navigating High-Dimensional Data Spaces“, *Proc. of LDAV 2016-IEEE 6th Symposium on Large Data Analysis and Visualization*

A Morrison, G Ross, M Chalmers. “A hybrid layout algorithm for sub-quadratic multidimensional scaling”, *Proc. of IEEE Symposium on Information Visualization,* pp. 152-158 , 2002.

S Oeltze, H Doleisch, H Hauser, P Muigg, B Preim. “Interactive visual analysis of perfusion data”, *IEEE transactions on visualization and computer graphics,* Vol.13 (6), 1392-1399, 2007.

Torgerson, Warren S. "Multidimensional scaling: I. Theory and method." *Psychometrika* 17.4 (1952): 401-419.

L van der Maaten, G Hinton. “Visualizing data using t-SNE”, The Journal of Machine Learning Research 9 (2579-2605), 85

M Williams, T Munzner. „Steerable, progressive multidimensional scaling“, *Proc. of IEEE Symposium on Information Visualization*, pp. 57-64, 2004.

Jing Yang, Matthew O. Ward, Elke A. Rundensteiner, Shiping Huang: Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets. *Proc. of VisSym*, pp. 19-28, 2003.

References (III)