Dimensionality reduction for information visualization

Jarkko Venna

Helsinki University of Technology / Numos Oy

Structure of the presentation

- Information visualization.
- Methods of dimensionality reduction
- Assessing the quality of visualizations
- New methods

Information visualization

The use of computer-supported, interactive, visual representations of abstract data to amplify cognition

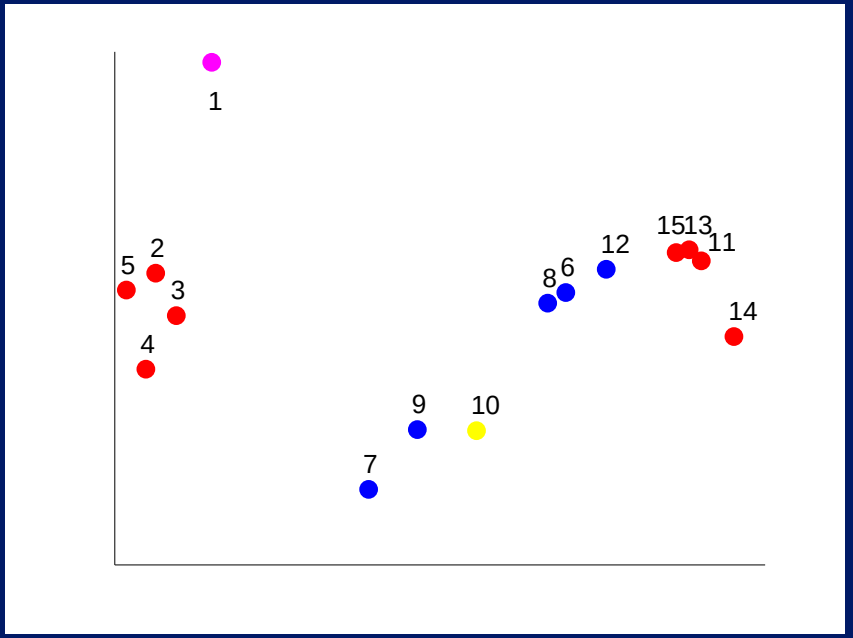
Card 1999

why visualize

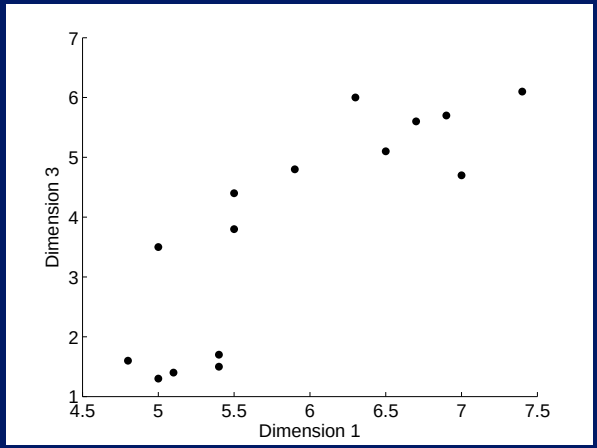
INPUT SPACE

Point	Var 1	Var 2	Var 3	Var 4
1	5.1	3.5	1.4	2.3
2	5.4	3.7	1.5	0.2
3	5.4	3.4	1.7	0.2
4	4.8	3.1	1.6	0.2
5	5.0	3.5	1.3	0.3
6	7.0	3.2	4.7	1.4
7	5.0	2.0	3.5	1.0
8	5.9	3.2	4.8	1.8
9	5.5	2.4	3.8	1.1
10	5.5	2.6	4.4	1.2
11	6.3	3.3	6.0	2.5
12	6.5	3.2	5.1	2.0
13	6.9	3.2	5.7	2.3
14	7.4	2.8	6.1	1.9
15	6.7	3.1	5.6	2.4

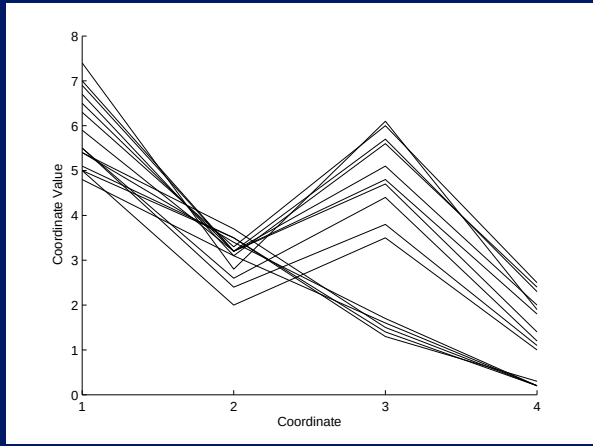
OUTPUT SPACE



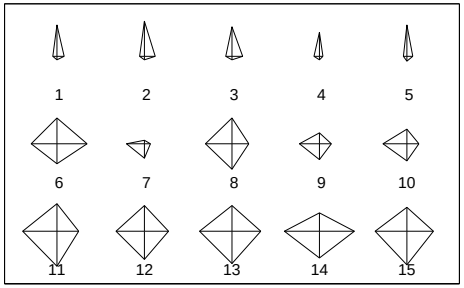
Classical methods for multivariate visualization


Dimension 3
Dimension 1
7
6
5
4
3
2
1
4.5
5
5.5
6
6.5
7
7.5

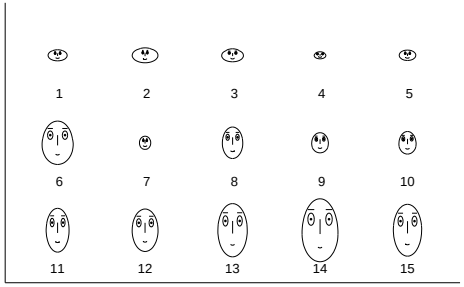
a


Coordinate Value
Coordinate
8
7
6
5
4
3
2
1
0
1
2
3
4

b


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

c

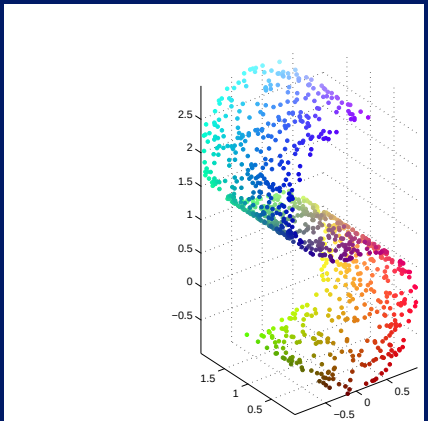

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

d

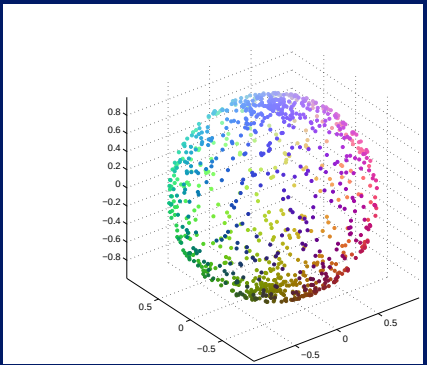
Dimensionality reduction methods

- Traditional approaches
  - Linear
  - Nonlinear distance preserving mappings
- Manifold learning methods
- Other methods

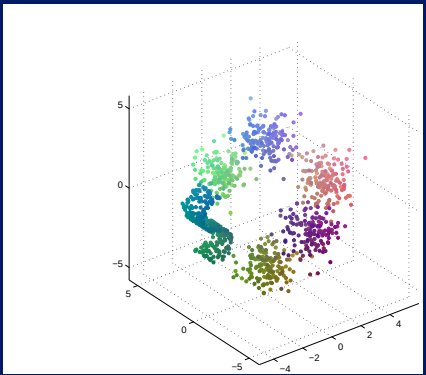
S-shaped manifold



Sphere



Clusters



Traditional approaches: Linear

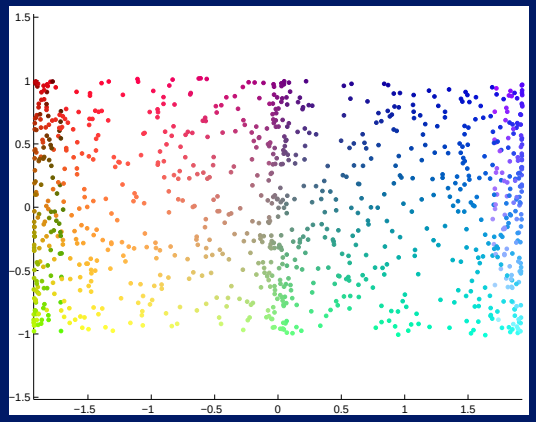
- Projection pursuit
- Principal component analysis (PCA)/
  linear Multidimensional Scaling
- The Grand Tour

Principal Component Analysis (PCA) / linear Multidimensional Scaling (MDS)

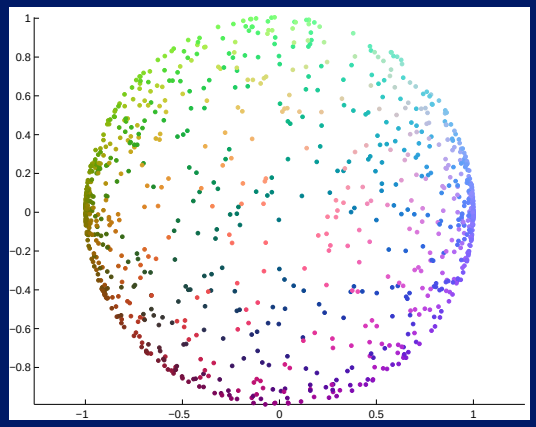
- The goal of PCA is to find linear components having maximal variance.
- Projection of the original data to the PCA subspace equals the configuration of points found by linear MDS (Classical scaling) that is calculated from the Euclidean distance matrix of the data.

PCA projections of the toy data sets

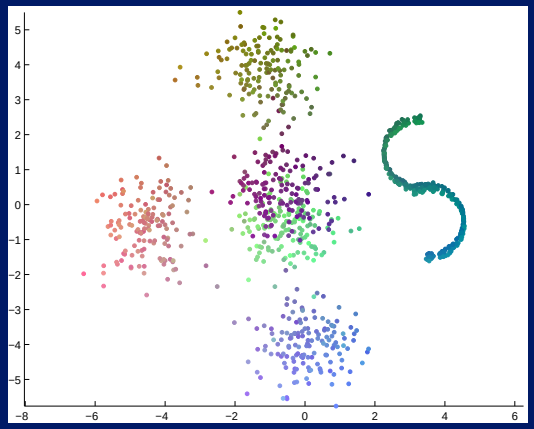
S-shaped manifold



Sphere



Clusters



Traditional approaches: Distance Preserving Mappings

- Traditional Multidimensional Scaling (MDS)
- Isomap
- Curvilinear Component Analysis (CCA)

Traditional MDS

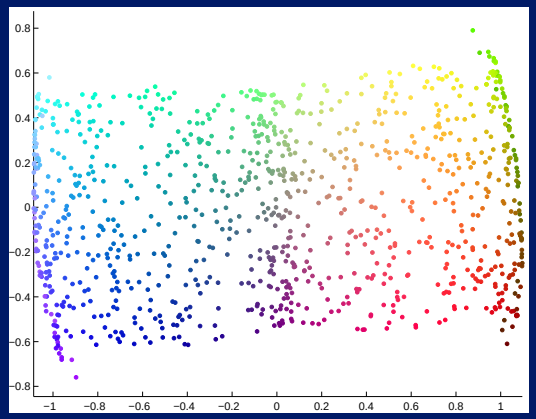
- There are several different variants of Multidimensional Scaling (MDS)
- The goal: to find a configuration of points that preserves the pairwise distance matrix of the data.
- The simplest nonlinear Multidimensional Scaling method is metric MDS. Its cost function, *raw stress*, is

$$E = \sum_{ij} (d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{y}_i, \mathbf{y}_j))^2, \tag{1}$$

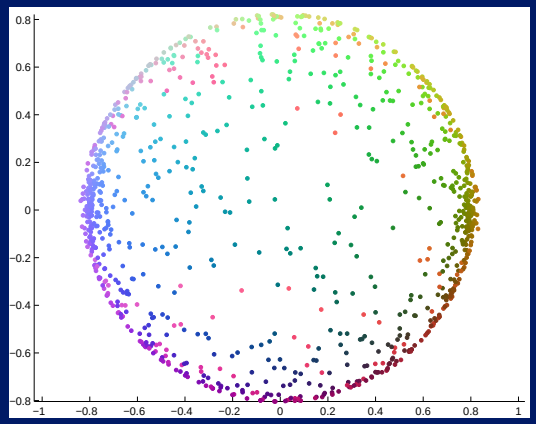
- Other variants include Sammon's mapping and nonlinear MDS.

Metric MDS projections of the toy data sets

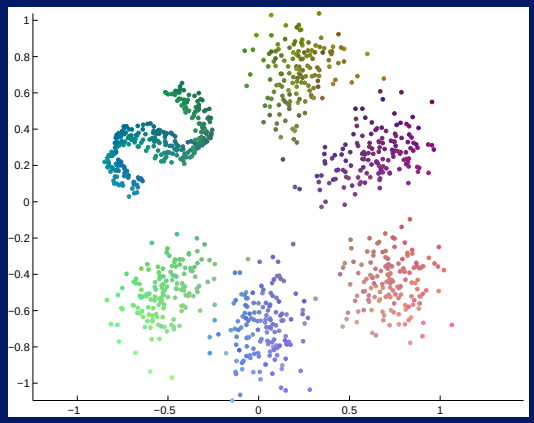
S-shaped manifold



Sphere



Clusters

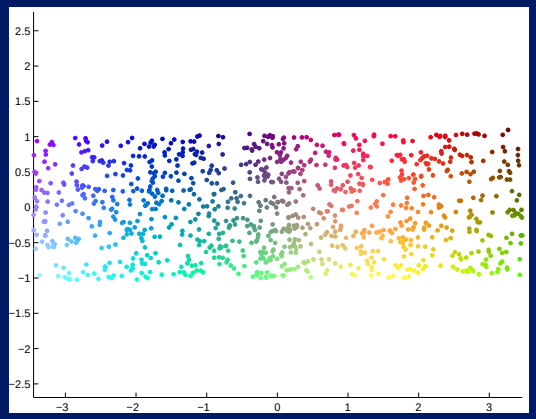


Isomap

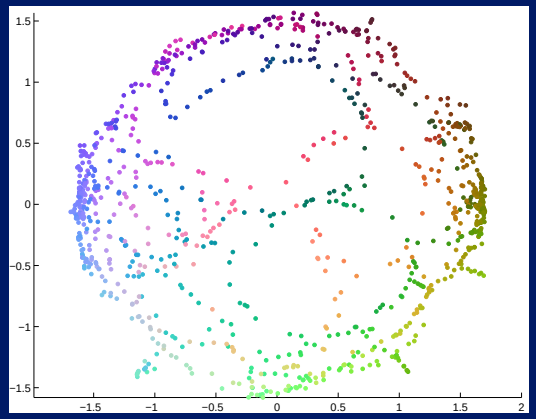
- Originally presented as a manifold learning method
- Form the k-nearest-neighbor graph. Each edge has a weight that is the Euclidean distance between the points it connects.
- Calculate the shortest path distances between points on the graph.
- Find the configuration of points by using linear MDS on the shortest path distance matrix.

Isomap projections of the toy data sets

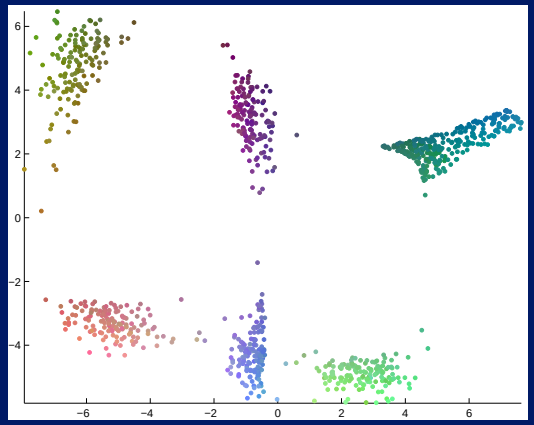
S-shaped manifold



Sphere



Clusters



Curvilinear Component Analysis (CCA)

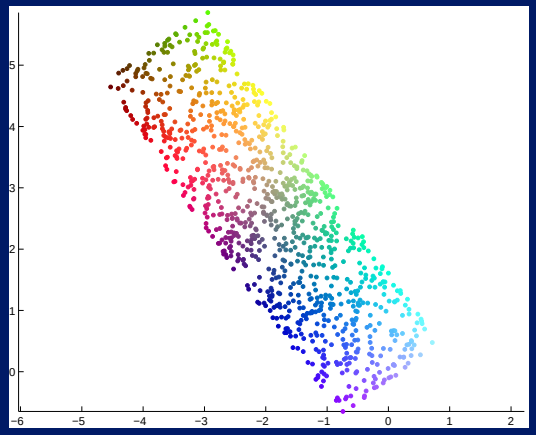
- Concentrates on preserving only the distances between points that are proximate in the *output space* instead of all pairwise distances.
- The cost function

$$E = \frac{1}{2} \sum_i \sum_{j \neq i} (d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{y}_i, \mathbf{y}_j))^2 F(d(\mathbf{y}_i, \mathbf{y}_j), \sigma) \ , \qquad (2)$$

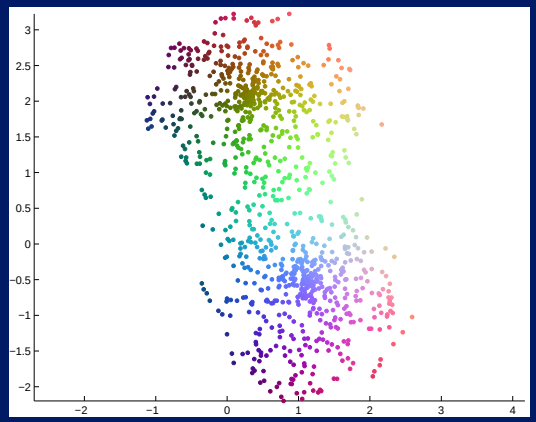
- The width of the area of influence around each data point σ is slowly reduced to zero during the optimization.

CCA projections of the toy data sets

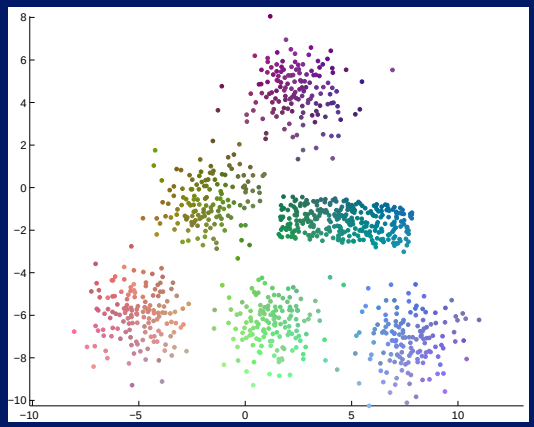
S-shaped manifold



Sphere



Clusters



Manifold Learning

The goal is to find and unfold the nonlinear manifold assumed to lie in the high dimensional data space

- Locally Linear Embedding (LLE)
- Laplacian Eigenmap
- Charting
- Maximum Variance Unfolding (MVU)

Locally Linear Embedding (LLE)

- The geometry of the data can be captured by calculating the linear coefficients that reconstruct each data point from its k nearest neighbors.
- The optimal reconstruction weights are found by minimizing:

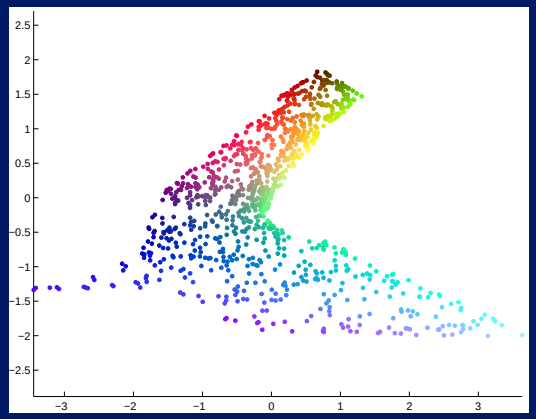
$$E(\mathbf{W}) = \sum_i |\mathbf{x}_i - \sum_j W_{ij}\mathbf{x}_j|^2. \tag{3}$$

- For visualization the configuration of points is found by minimizing

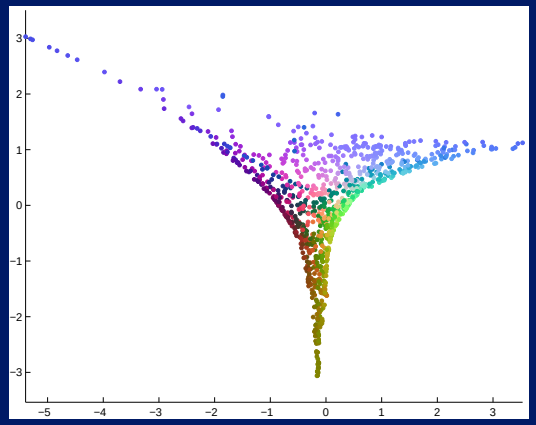
$$E(\mathbf{Y}) = \sum_i |\mathbf{y}_i - \sum_j \mathbf{W}_{ij}\mathbf{y}_j|^2 \ , \tag{4}$$

LLE projections of the toy data sets

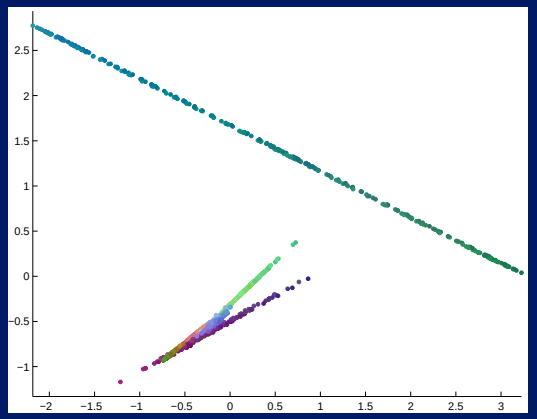
S-shaped manifold



Sphere



Clusters



Other Approaches

- The Self-Organizing Map (SOM)
- Stochastic Neighbor Embedding

Stochastic Neighbor Embedding (SNE)

- SNE tries to preserve the probability of points being a neighbors
- The probability p_{ij} of the point i being a neighbor of point j in the input space

$$p_{ij} = \frac{\exp\left(-d(\mathbf{x}_i, \mathbf{x}_j)/\sigma_i^{(i)}\right)}{\sum_{k \neq i} \exp\left(-d(\mathbf{x}_i, \mathbf{x}_k)/\sigma_i^{(i)}\right)}, \tag{5}$$

- The probability of the point i being a neighbor of point j in the output space

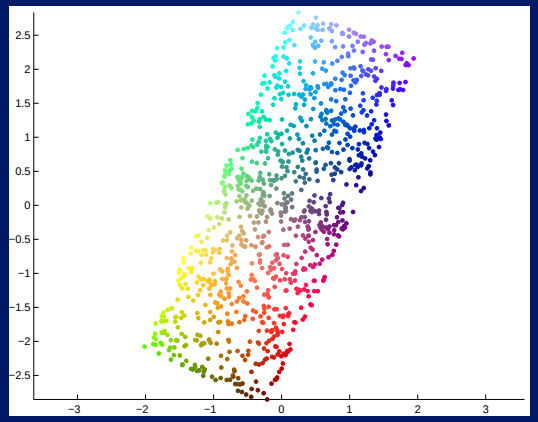
$$q_{ij} = \frac{\exp\left(-\|\mathbf{y}_i - \mathbf{y}_j\|^2/\sigma_i^{(o)}\right)}{\sum_{k \neq i} \exp\left(-\|\mathbf{y}_i - \mathbf{y}_k\|^2/\sigma_i^{(o)}\right)}. \tag{6}$$

- The cost function

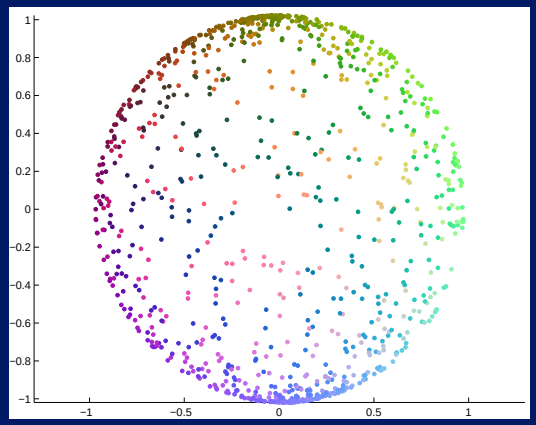
$$E = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{7}$$

SNE projections of the toy data sets

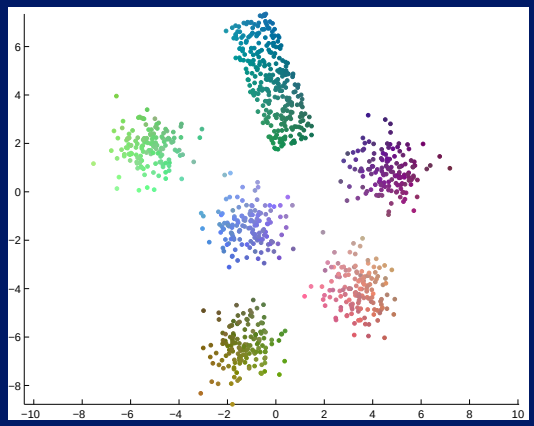
S-shaped manifold



Sphere



Clusters



Assessing the quality of visualizations

- The final truth can only be found by usability studies
- Quality measures can be used to verify that the relevant features in the data are represented accurately
- Commonly used methods to asses the quality of visualizations
  - Qualitative assessment by looking at the visualization
  - Distance preservation measures
  - Classification rate

A new visualization task: Visual Neighbor Retrieval

The task of the user is to identify the neighbors of a data point by looking at the visualization.

The task of the visualization system is to produce a single image that allows the neighbors of data points to be selected as well as possible without prior knowledge of which data points neighbors are studied.

Precision and Recall

- precision and recall are used to measures the quality of information retrieval systems
- $$\text{precision} = \frac{N_{TP}}{k} = 1 - \frac{N_{FP}}{k}, \tag{8}$$

  where N_{TP} is the number of the true positives, N_{FP} is the number of the false positives and k is the number of the retrieved items.
- $$\text{recall} = \frac{N_{TP}}{r} = 1 - \frac{N_{MISS}}{r}, \tag{9}$$

  where N_{MISS} is the number of the misses, the relevant objects not retrieved, and r is the total number of the relevant objects.

Precision and Recall in a visualization

- The precision and recall are calculated separately for each data points neighborhood
- r=number of nearest neighbors in the data sets
- k=number of neighbors studied in the visualization
- The precision and recall measures are averaged to get the overall measure for the visualization.

Trustworthiness of a visualization

$$M_{trust}(k) = 1 - A(k) \sum_{i=1}^{N} \sum_{j \in U_k(i)} (r(i,j) - k), \qquad (10)$$

- $A(k)$ scales the measure to be between zero and one:
- $U_k(i)$ is the set of points that are in the neighborhood of the data point i in the output space but not in the input space.
- The errors are quantified by ranks ($r(i,j)$) instead of just counted as in precision

Continuity of a visualization

$$M_{cont}(k) = 1 - A(k) \sum_{i=1}^{N} \sum_{j \in V_k(i)} (\hat{r}(i,j) - k) \; . \tag{11}$$

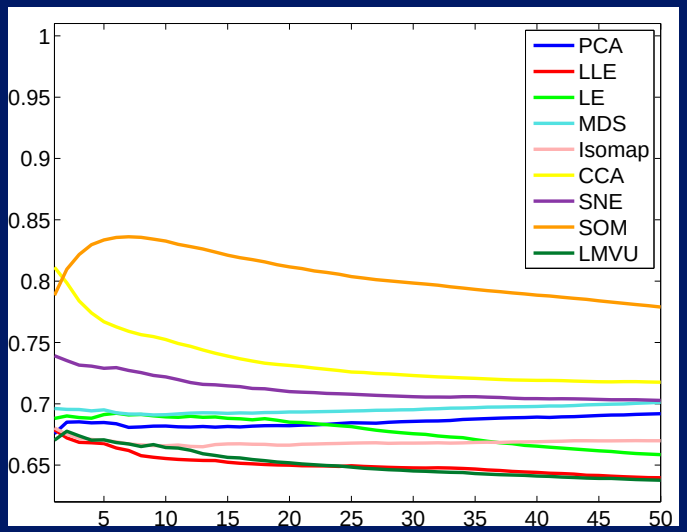
- $A(k)$ scales the measure to be between zero and one:
- $V_k(i)$ is the set of points that are in the neighborhood of the data point i in the input space but not in the output space.
- The errors are quantified by ranks ($\hat{r}(i,j)$) instead of just counted as in recall

Comparison of visualization methods; S-data

Gene expression compendium

Trustworthiness

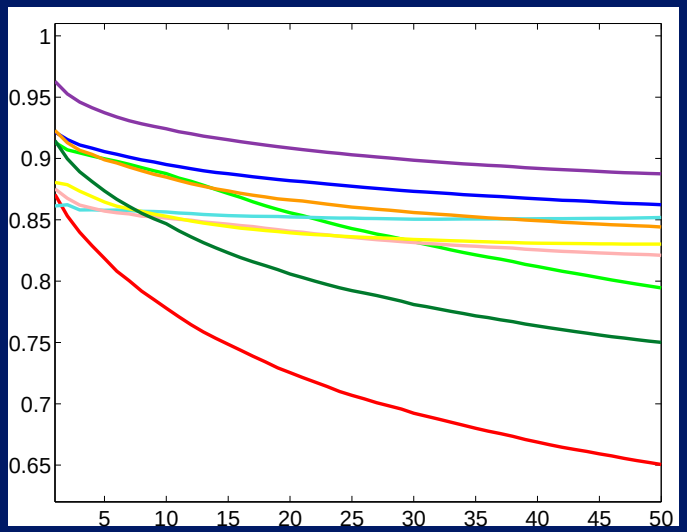
M_{trust}



k

Continuity

M_{cont}



k

New methods aimed for the visual neighbor retrieval task

- Neighbor Retrieval Visualizer (NeRV)
- Local MDS

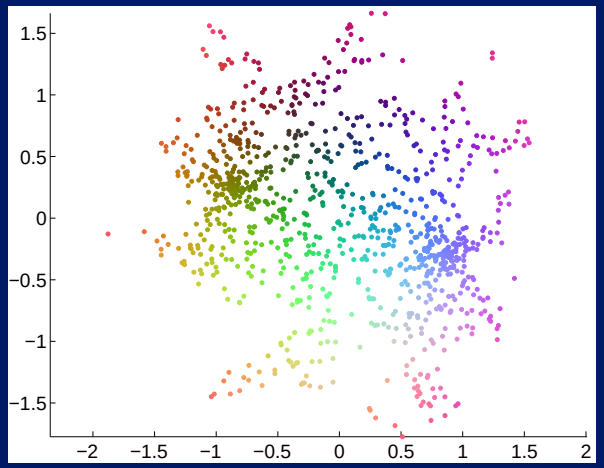
Neighbor Retrieval Visualizer (NeRV)

- It can be shown that SNE optimizes a kind of smoothed recall measure.
- Typically optimizing recall leads to low precision
- By reversing the Kullback-Liebler divergence in the cost function of SNE we get a method that optimizes smoothed precision
- NeRV cost function

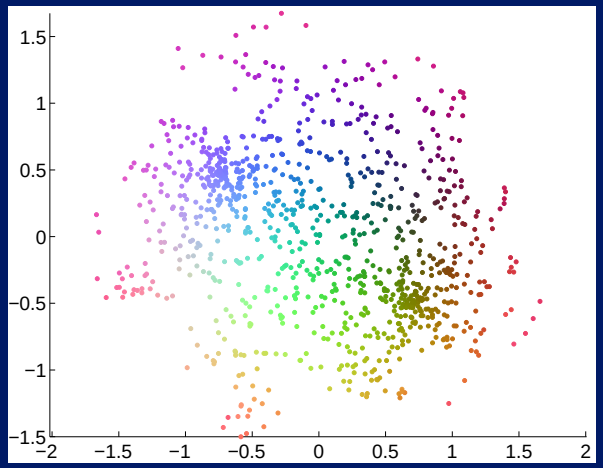
$$\begin{aligned} E_{\text{NeRV}} &= \lambda E_i[D_{KL}(p_i, q_i)] + (1-\lambda) E_i[D_{KL}(q_i, p_i)] \\ &= \lambda \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1-\lambda) \sum_i \sum_{j \neq i} q_{ij} \log \frac{q_{ij}}{p_{ij}}, \end{aligned} \quad (12)$$

- $\lambda \in [0 \ldots 1]$ selects the trade off between precision and recall

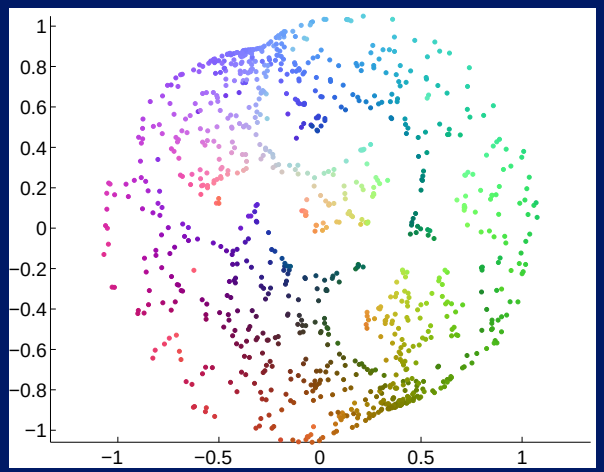
Example: NeRV projections of a sphere



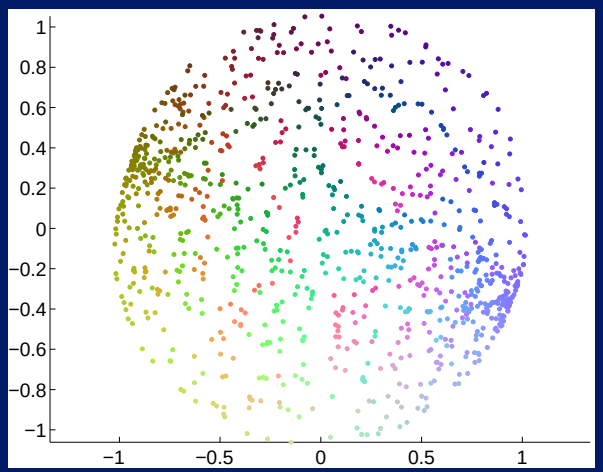
$\lambda = 0$



$\lambda = 0.1$



$\lambda = 0.9$



$\lambda = 1.0$

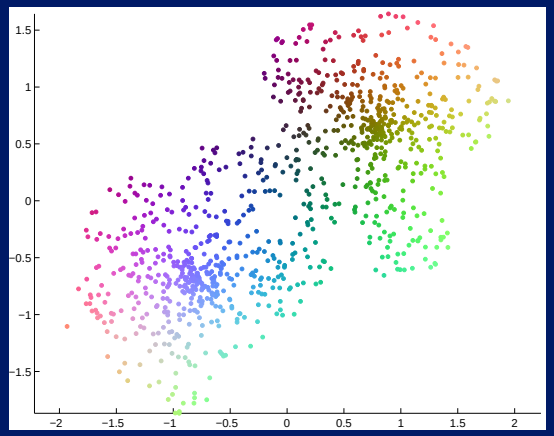
Local MDS

- CCA concentrates on preserving distances between close-by points in the *visualization*. This results in good trustworthiness.
- by adding a term to the cost function that concentrates on preserving distances between close-by points in the *original* space the formation of discontinuities in the mapping is discouraged.
- The cost function is

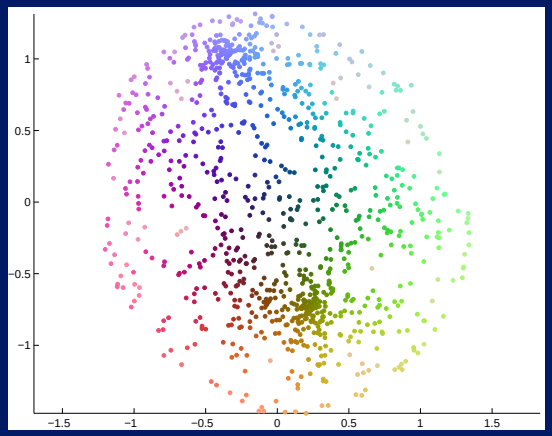
$$E = \frac{1}{2}\sum_i \sum_{j \neq i} [(1-\lambda)(d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{y}_i, \mathbf{y}_j))^2 F(d(\mathbf{y}_i, \mathbf{y}_j), \sigma_i) + \lambda(d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{y}_i, \mathbf{y}_j))^2 F(d(\mathbf{x}_i, \mathbf{x}_j), \sigma_i)] \; .$$

- $F(d(\bullet_i, \bullet_j), \sigma_i)$ is the area of influence around the data point i

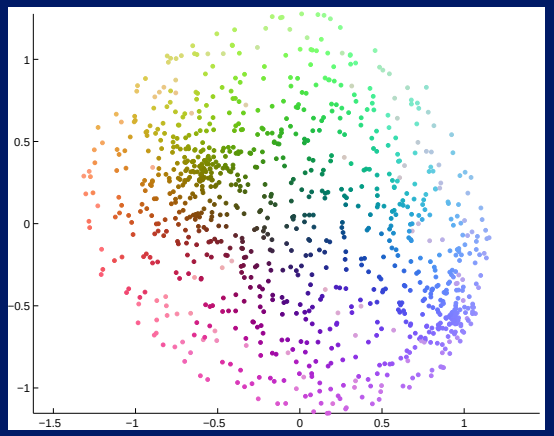
Example: local MDS projections of a sphere



$\lambda = 0$



$\lambda = 0.1$



$\lambda = 0.5$

Further information

http://www.cis.hut.fi/projects/mi