Multidimensional Scaling and Sammon's Mapping

Prof. Asim Tewari
IIT Bombay

Feature projection

- Principal component analysis (PCA)
- Non-negative matrix factorization (NMF)
- Kernel PCA
- Graph-based kernel PCA
- Linear discriminant analysis (LDA)
- Generalized discriminant analysis (GDA)
- Autoencoder

Principal Component Analysis

- PCA performs a linear transform of a data set X, which consists of a translation and a rotation.

$$y_k = (x_k - \bar{x}) \cdot E$$

where E is a rotation matrix that has to be determined from X. The corresponding inverse transform is

$$x_k = y_k \cdot E^T + \bar{x}$$

Principal Component Analysis

$$y_k = (x_k - \bar{x}) \cdot E$$

To determine the rotation matrix E, the variance of Y is maximized. The variance of Y can be written as

$$v_y = \frac{1}{n-1} \sum_{k=1}^{n} y_k^T y_k$$

$$= \frac{1}{n-1} \sum_{k=1}^{n} \left((x_k - \bar{x}) \cdot E\right)^T \cdot \left((x_k - \bar{x}) \cdot E\right)$$

Principal Component Analysis

$$
\begin{aligned}
v_y &= \frac{1}{n-1}\sum_{k=1}^{n} y_k^T y_k \\
&= \frac{1}{n-1}\sum_{k=1}^{n} ((x_k-\bar{x})\cdot E)^T \cdot ((x_k-\bar{x})\cdot E) \\
&= \frac{1}{n-1}\sum_{k=1}^{n} E^T \cdot (x_k-\bar{x})^T \cdot (x_k-\bar{x}) \cdot E \\
&= E^T \left(\frac{1}{n-1}\sum_{k=1}^{n} (x_k-\bar{x})^T \cdot (x_k-\bar{x}) \right) \cdot E \\
&= E^T \cdot C \cdot E
\end{aligned}
$$

where C is the covariance matrix of X.

Principal Component Analysis

- The elements of the covariance matrix are:

$$c_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k^{(i)} - \bar{x}^{(i)})(x_k^{(j)} - \bar{x}^{(j)}), \quad i,j = 1,\ldots,p$$

- The transformation matrix E should only represent a rotation, not a dilation, so we require:

$$E^T \cdot E = 1$$

Principal Component Analysis

- This leads to a constrained optimization problem that can be solved using Lagrange optimization.
- The variance $E^T \cdot C \cdot E$ can be maximized under the constraint $E^T \cdot E = 1$ using the Lagrange function

$$L = E^T C E - \lambda (E^T E - 1)$$

Principal Component Analysis

$$L = E^T C E - \lambda(E^T E - 1)$$

- The necessary condition for optima of L is

$$\frac{\partial L}{\partial E} = 0$$

$$\Leftrightarrow \quad CE = \lambda E$$

- This equation defines an eigenproblem which can be solved using linear algebra of homogeneous equation system

$$(C - \lambda I) \cdot E = 0$$

Principal Component Analysis

- The rotation matrix E is the concatenation of the eigenvectors of C.

$$E = (v_1, \ldots, v_p), \quad (v_1, \ldots, v_p, \lambda_1, \ldots, \lambda_p) = \operatorname{eig} C$$

- The variances in Y correspond to the eigenvalues λ1,…, λp of C because

$$CE = \lambda E \quad \Leftrightarrow \quad \lambda = E^T CE = v_y$$

Principal Component Analysis

$$X = \{(1,1),(2,1),(2,2),(3,2)\}$$

$$\bar{x} = \frac{1}{2} \cdot (4,3)$$

$$C = \frac{1}{3} \cdot \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

$$\lambda_1 = 0.8727$$

$$\lambda_2 = 0.1273$$

$$v_1 = \begin{pmatrix} -0.85065 \\ -0.52573 \end{pmatrix}$$

$$v_2 = \begin{pmatrix} 0.52573 \\ -0.85065 \end{pmatrix}$$

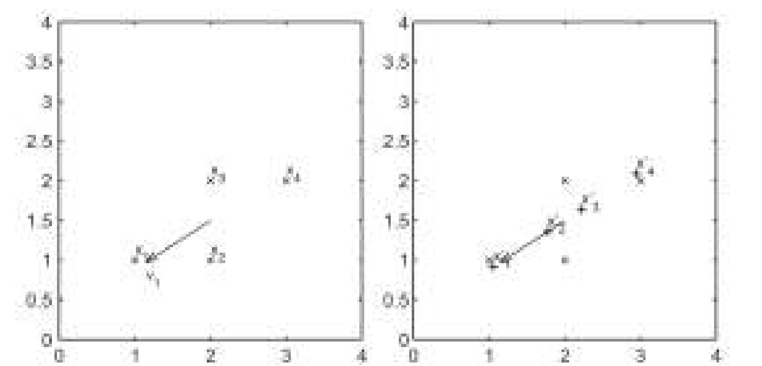
$$E = v_1 = \begin{pmatrix} -0.85065 \\ -0.52573 \end{pmatrix}$$

The projected data

$$Y = \{1.1135, 0.2629, -0.2629, -1.1135\}$$

Inverse PCA yields

$$X' = \{\ (1.0528, 0.91459),\ (1.7764, 1.3618),\ (2.2236, 1.6382),\ (2.9472, 2.0854)\ \} \neq X$$



Multidimensional Scaling

- Multidimensional scaling (MDS) is a linear mapping based on matrix decomposition. Given the data matrix $X \in \mathbb{R}^{n \times p}$, eigen decomposition of the product matrix XX^T yields

$$XX^T = Q\Lambda Q^T = (Q\sqrt{\Lambda}^T) \cdot (\sqrt{\Lambda} Q^T) = (Q\sqrt{\Lambda}^T) \cdot (Q\sqrt{\Lambda}^T)^T$$

Multidimensional Scaling

$$XX^T = Q\Lambda Q^T = (Q\sqrt{\Lambda}^T)\cdot(\sqrt{\Lambda}Q^T) = (Q\sqrt{\Lambda}^T)\cdot(Q\sqrt{\Lambda}^T)^T$$

Using this eigendecomposition, an estimate for *X* is

$$Y = Q\sqrt{\Lambda}^T$$

MDS of a feature data set *X* yields the same results as PCA.

However, MDS also produces an approximate feature space representation *Y* for Euclidean distance matrix D.

Multidimensional Scaling

Let us transform $\tilde{\mathbb{X}}$ to $\mathbb{X}$ in a coordinate system with an arbitrarily chosen origin $\tilde{x}_a, a \in \{1, \ldots, n\}$

Where $\tilde{X} \in \mathbb{R}^{n \times p}$ and $X \in \mathbb{R}^{n \times q}$, $q < p$

so

$$x_k = \tilde{x}_k - \tilde{x}_a$$

$k = 1, \ldots, n,$ and

$$\tilde{x}_i - \tilde{x}_j = x_i - x_j$$

$i, j = 1, \ldots, n.$

Multidimensional Scaling

$$\tilde{X} \rightarrow X$$

$k = 1, \ldots, n$, and $\quad x_k = \tilde{x}_k - \tilde{x}_n$

$i, j = 1, \ldots, n. \quad \tilde{x}_i - \tilde{x}_j = x_i - x_j$

Taking the scalar product of each side with itself yields

$$(\tilde{x}_i - \tilde{x}_j)(\tilde{x}_i - \tilde{x}_j)^T = (x_i - x_j)(x_i - x_j)^T$$

$$\Rightarrow d_{ij}^2 = x_i x_i^T - 2x_i x_j^T + x_j x_j^T = d_{in}^2 - 2x_i x_j^T + d_{jn}^2$$

$$\Rightarrow x_i x_j^T = (d_{in}^2 + d_{jn}^2 - d_{ij}^2)/2$$

so the product matrix XX^T can be computed from the Euclidean distance matrix D.

Multidimensional Scaling

$$\tilde{X} \to X$$

$k = 1, \ldots, n$, and $\quad x_k = \tilde{x}_k - \tilde{x}_n$

$i, j = 1, \ldots, n. \quad \tilde{x}_i - \tilde{x}_j = x_i - x_j$

Taking the scalar product of each side with itself yields

$$(\tilde{x}_i - \tilde{x}_j)(\tilde{x}_i - \tilde{x}_j)^T = (x_i - x_j)(x_i - x_j)^T$$

$$\Rightarrow d_{ij}^2 = x_i x_i^T - 2 x_i x_j^T + x_j x_j^T = d_{in}^2 - 2 x_i x_j^T + d_{jn}^2$$

$$\Rightarrow x_i x_j^T = (d_{in}^2 + d_{jn}^2 - d_{ij}^2)/2$$

so the product matrix XX^T can be computed from the Euclidean distance matrix D.

Multidimensional Scaling

$$\tilde{X} = \{(1,1), (2,1), (2,2), (3,2)\}$$

$$\bar{x} = \frac{1}{2} \cdot (4,3)$$

Subtracting the mean yields

$$X = \left\{(-1, -\frac{1}{2}), (0, -\frac{1}{2}), (0, \frac{1}{2}), (1, \frac{1}{2})\right\}$$

and the product matrix

$$XX^T = \frac{1}{4}\begin{pmatrix} 5 & 1 & -1 & -5 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -5 & -1 & 1 & 5 \end{pmatrix}$$

The largest eigenvalue and the corresponding eigenvector of XX^T are

$$\lambda_1 \approx 2.618 \qquad v_1 \approx \begin{pmatrix} -0.6882 \\ -0.1625 \\ 0.1625 \\ 0.6882 \end{pmatrix}$$

So the 1D MSD projection is

$$Y \approx \begin{pmatrix} -0.6882 \\ -0.1625 \\ 0.1625 \\ 0.6882 \end{pmatrix} \sqrt{2.618} \approx \begin{pmatrix} -1.1135 \\ -0.2629 \\ 0.2629 \\ 1.1135 \end{pmatrix}$$

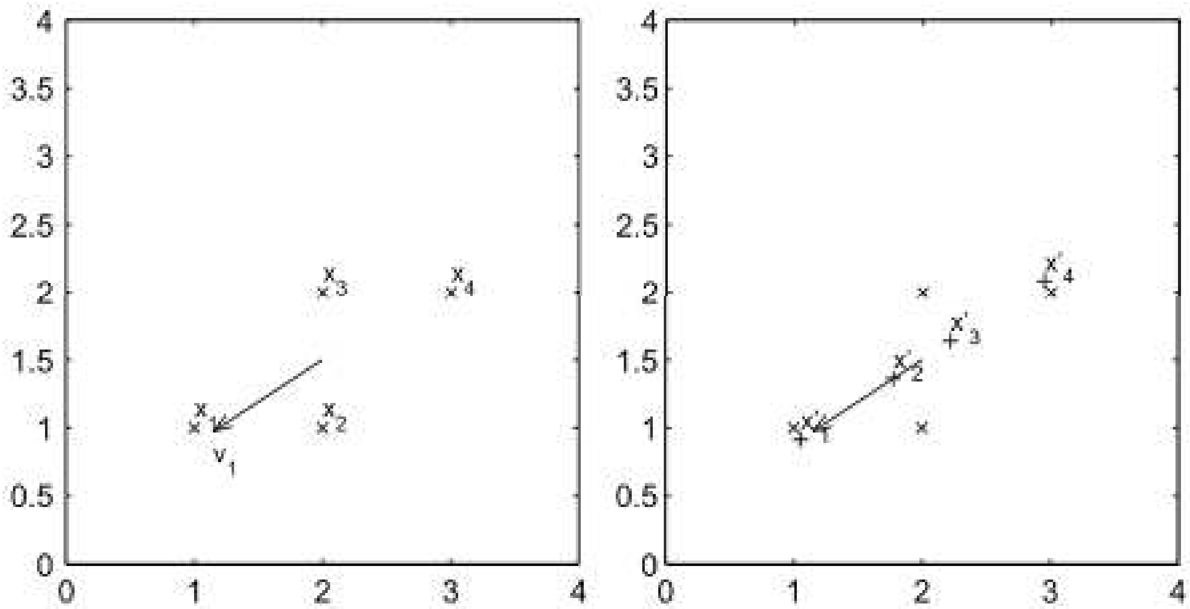
Multidimensional Scaling

$$D^x = \begin{pmatrix} 0 & 1 & \sqrt{2} & \sqrt{5} \\ 1 & 0 & 1 & \sqrt{2} \\ \sqrt{2} & 1 & 0 & 1 \\ \sqrt{5} & \sqrt{2} & 1 & 0 \end{pmatrix} \approx \begin{pmatrix} 0 & 1 & 1.4142 & 2.2361 \\ 1 & 0 & 1 & 1.4142 \\ 1.4142 & 1 & 0 & 1 \\ 2.2361 & 1.4142 & 1 & 0 \end{pmatrix}$$

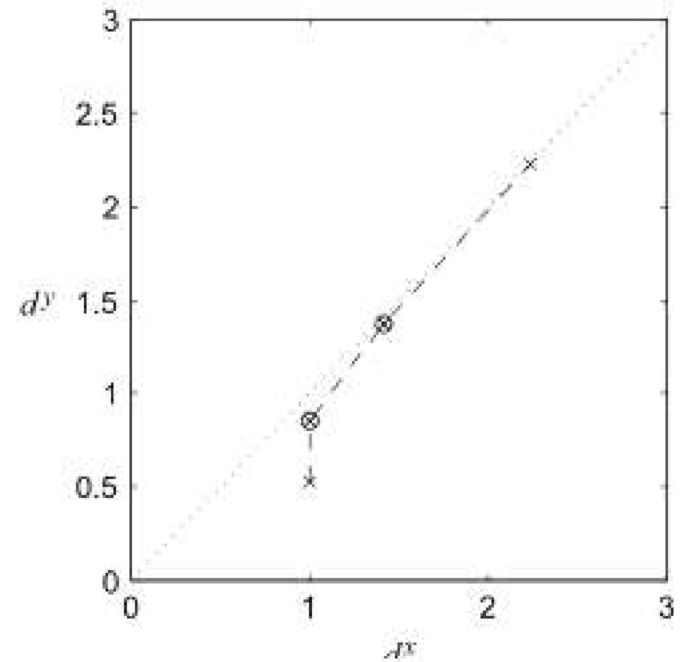
$$D^y \approx \begin{pmatrix} 0 & 0.8507 & 1.3764 & 2.2270 \\ 0.8507 & 0 & 0.5257 & 1.3764 \\ 1.3764 & 0.5257 & 0 & 0.8507 \\ 2.2270 & 1.3764 & 0.8507 & 0 \end{pmatrix}$$

The quality of this mapping can be visualized by a so-called Shepard diagram, a scatter plot of the distances d^y_{ij} in the projection versus the distances d^x_{ij} of the original data

Multidimensional Scaling

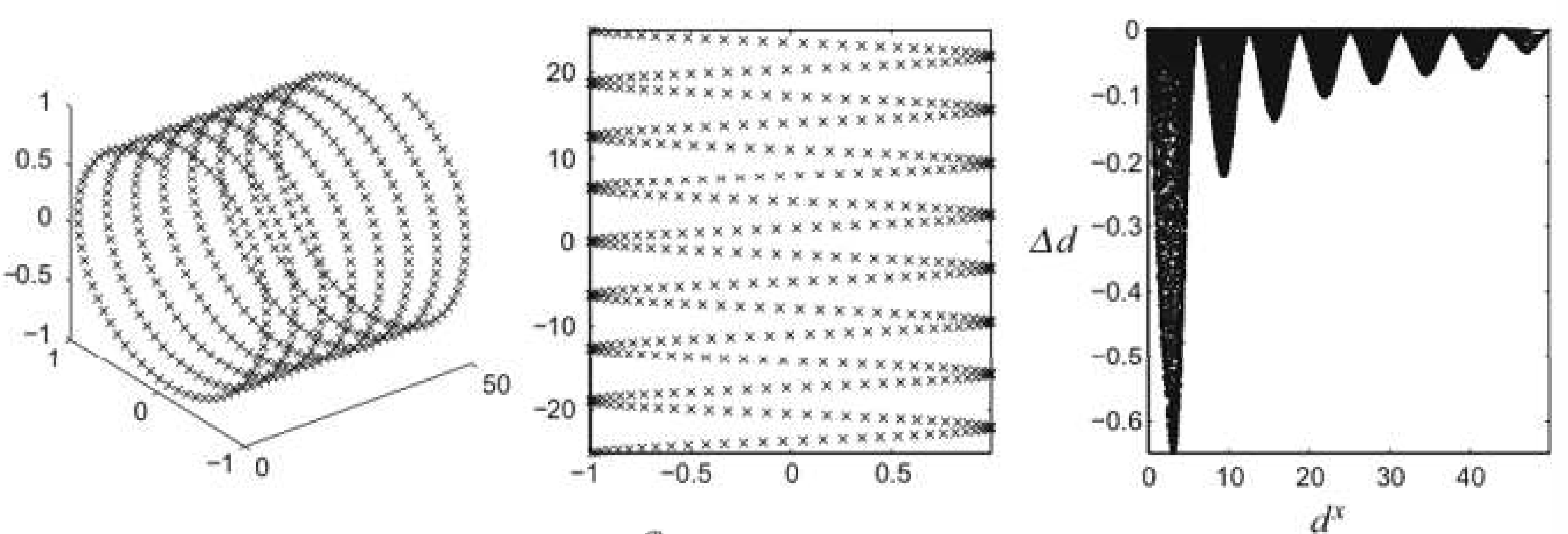


Principal component analysis (four points data set)



Shepard diagram for PCA/MDS projection (four points data set)

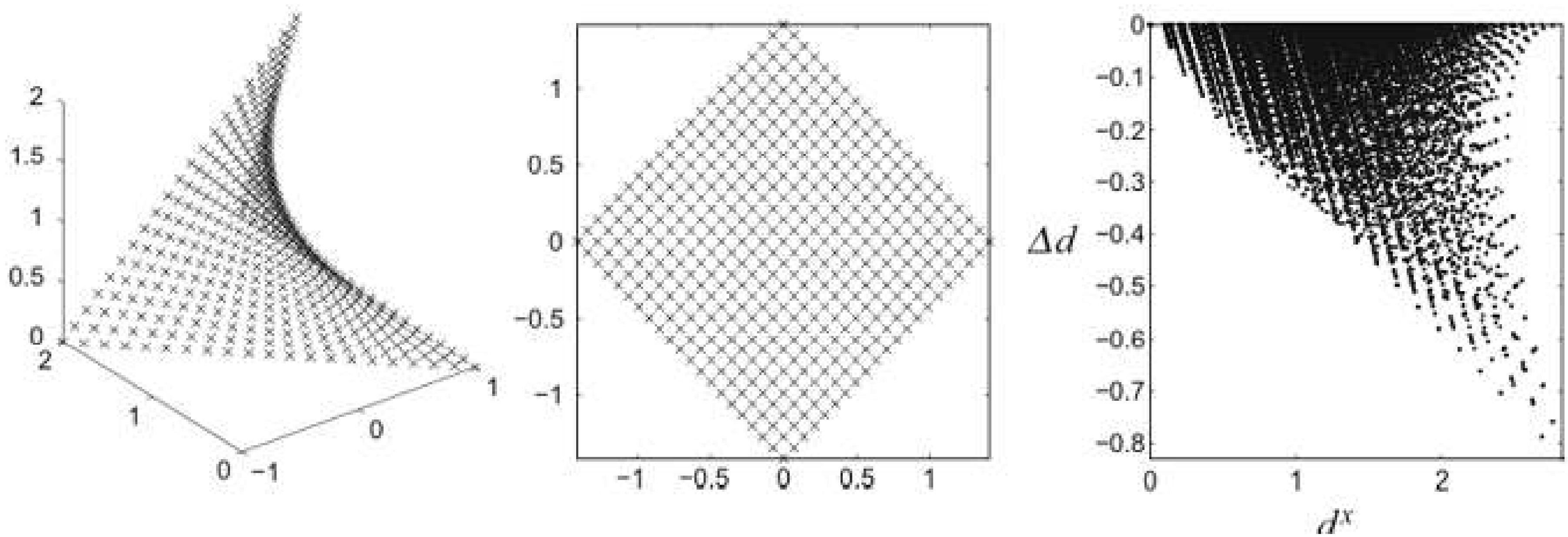
Multidimensional Scaling



$$X = \{(t, \sin t, \cos t)^T \mid t \in \{0, 0.1, 0.2, \ldots, 50\}\}$$

Helix data set, PCA/MDS projection, and projection errors

Multidimensional Scaling



$$\mathcal{X} = \{((s_1 - 1)\cdot(s_2 - 1), s_1, s_2)^T \mid s_1, s_2 \in \{0, 0.1, 0.2, \ldots, 2\}\}$$

Bent square data set, PCA/MDS projection, and projection errors

Sammon mapping

- The idea of Sammon mapping is to map a data set $X \subset \mathbb{R}^p$ to a data set $Y \subset \mathbb{R}^q$ so that distances between pairs of elements of *X* are similar to the corresponding distances between pairs of elements of *Y*.

$$d_{ij}^{x} \approx d_{ij}^{y}$$

Y is found by minimizing the error between D^x and D^y. Candidates for this error functional are

Sammon mapping

Y is found by minimizing the error between D^x and D^y. Candidates for this error functional are

$$E_1 = \frac{1}{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\left(d_{ij}^{x}\right)^2}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\left(d_{ij}^{y}-d_{ij}^{x}\right)^2$$

$$E_2 = \sum_{i=1}^{n}\sum_{j=i+1}^{n}\left(\frac{d_{ij}^{y}-d_{ij}^{x}}{d_{ij}^{x}}\right)^2$$

$$E_3 = \frac{1}{\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}^{x}}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\frac{\left(d_{ij}^{y}-d_{ij}^{x}\right)^2}{d_{ij}^{x}}$$

Sammon mapping

$$\frac{\partial d_{ij}^{y}}{\partial y_k} = \frac{\partial}{\partial y_k}\|y_i - y_j\| = \begin{cases} \frac{y_k - y_j}{d_{kj}^{y}} & \text{if } i = k \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial E_S}{\partial y_k} = \frac{2}{\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}^{x}} \sum_{\substack{j=1 \\ j\neq k}}^{n} \left(\frac{1}{d_{kj}^{x}} - \frac{1}{d_{kj}^{y}}\right)(y_k - y_j)$$

$$\frac{\partial^2 E_S}{\partial y_k^2} = \frac{2}{\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}^{x}} \sum_{\substack{j=1 \\ j\neq k}}^{n} \left(\frac{1}{d_{kj}^{x}} - \frac{1}{d_{kj}^{y}} - \frac{(y_k - y_j)^2}{(d_{kj}^{y})^3}\right)$$

Sammon mapping

$$X = \{(1,1), (2,1), (2,2), (3,2)\}$$

We initialize *Y = {1,2,3,4}*

$$D^y = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{pmatrix}$$

and the initial Sammon error

$$E_3 = \frac{1}{3 \cdot 1 + 2 \cdot \sqrt{2} + \sqrt{5}} \cdot \left(2 \cdot \frac{\left(2 - \sqrt{2}\right)^2}{\sqrt{2}} + \frac{\left(3 - \sqrt{5}\right)^2}{\sqrt{5}} \right) \approx 0.0925$$

This is the first (leftmost) value of the Sammon error function shown in Fig

Sammon mapping

For this initialization, the error gradients are

$$\frac{\partial E_3}{\partial y_1} = \frac{2}{3 \cdot 1 + 2 \cdot \sqrt{2} + \sqrt{5}} \cdot \left(-\frac{2-\sqrt{2}}{\sqrt{2}} - \frac{3-\sqrt{5}}{\sqrt{5}} \right) \approx -0.1875$$

$$\frac{\partial E_3}{\partial y_2} = \frac{2}{3 \cdot 1 + 2 \cdot \sqrt{2} + \sqrt{5}} \cdot \left(-\frac{2-\sqrt{2}}{\sqrt{2}} \right) \approx -0.1027$$

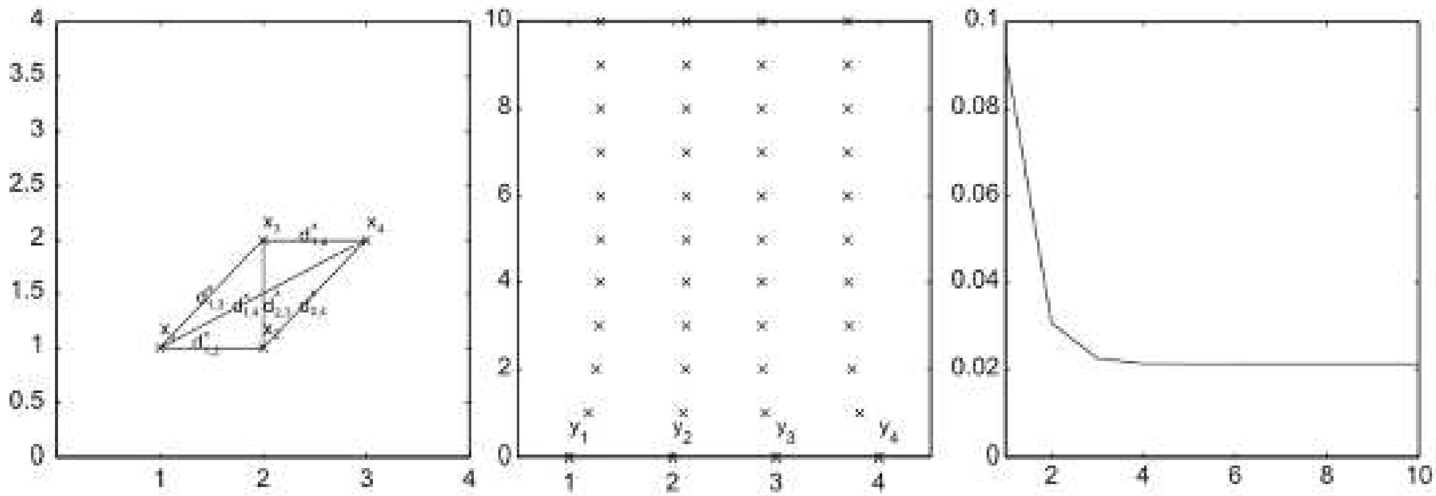
$$\frac{\partial E_3}{\partial y_3} = \frac{2}{3 \cdot 1 + 2 \cdot \sqrt{2} + \sqrt{5}} \cdot \left(\frac{2-\sqrt{2}}{\sqrt{2}} \right) \approx 0.1027$$

$$\frac{\partial E_3}{\partial y_4} = \frac{2}{3 \cdot 1 + 2 \cdot \sqrt{2} + \sqrt{5}} \cdot \left(\frac{3-\sqrt{5}}{\sqrt{5}} + \frac{2-\sqrt{2}}{\sqrt{2}} \right) \approx 0.1875$$

With step length α *= 1, the estimate of Y = (1.1875, 2.1027, 2.8973, 3.8125)*

Sammon mapping

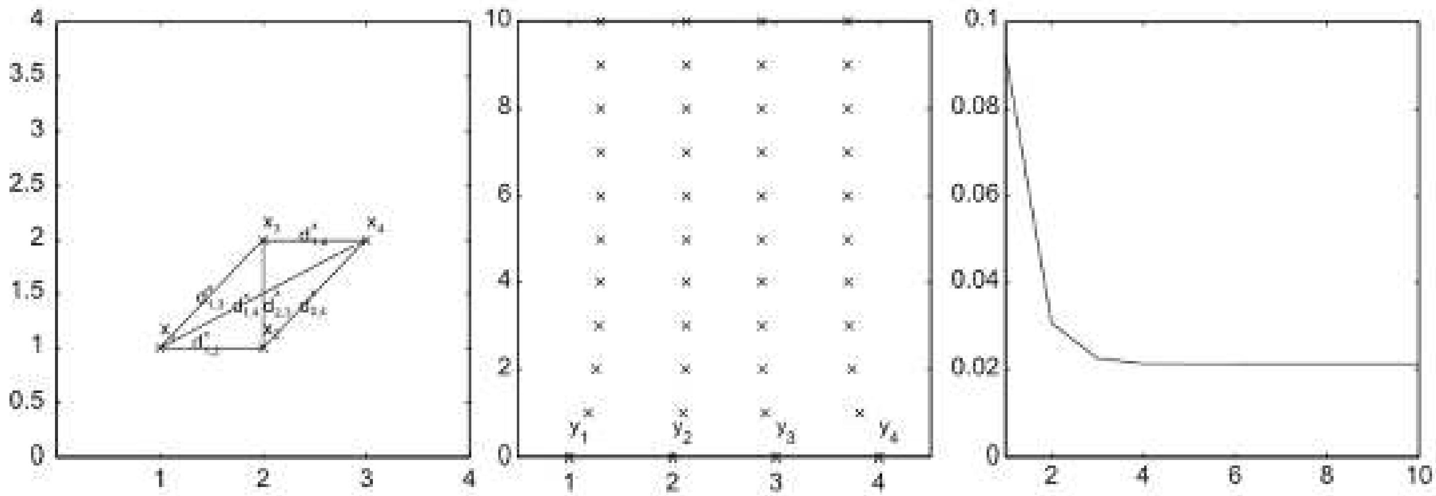
$$X = \{(1,1), (2,1), (2,2), (3,2)\}$$



Four points data set: pairwise distances, Sammon mapping after 0,…. 10 iterations, Sammon error function

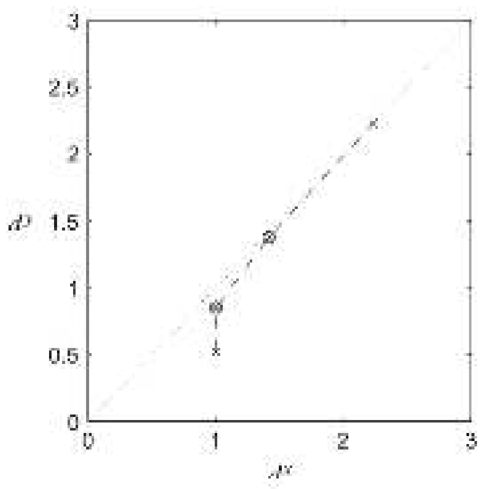
Sammon mapping

$$X = \{(1,1), (2,1), (2,2), (3,2)\}$$

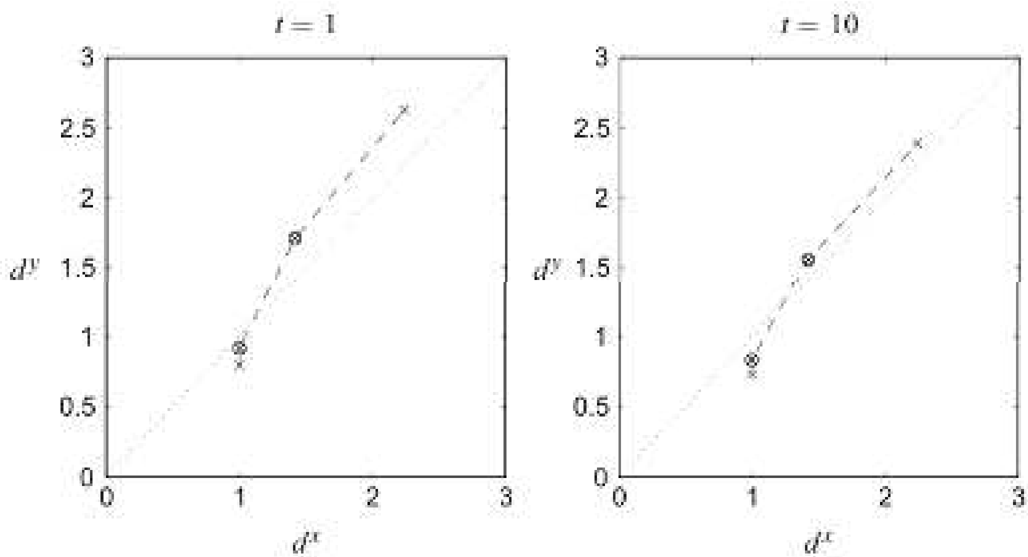


Four points data set: pairwise distances, Sammon mapping after 0,…. 10 iterations, Sammon error function

Sammon mapping

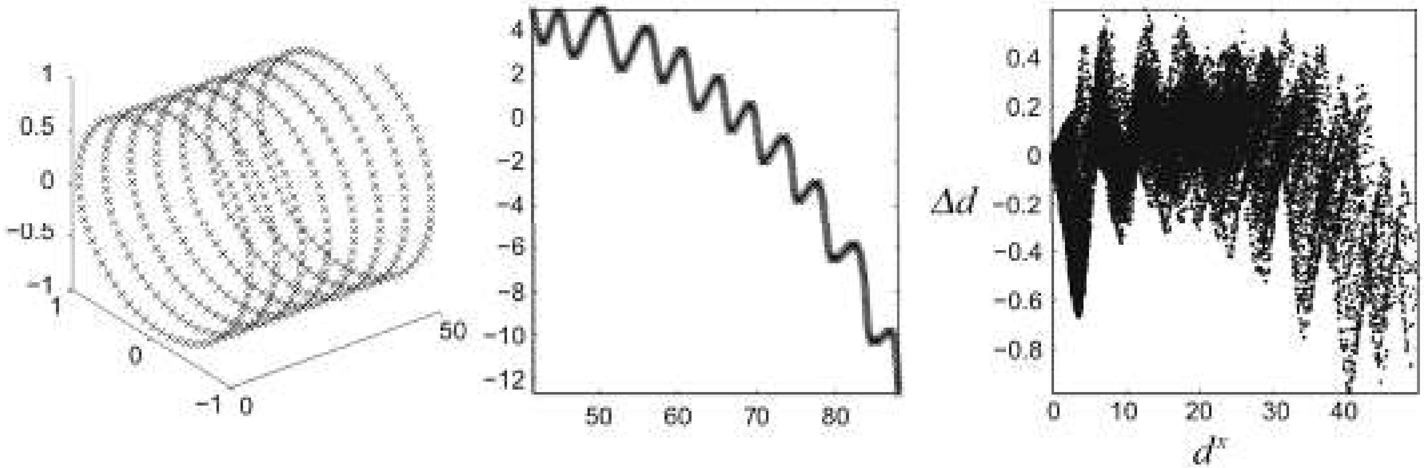


Shepard diagram for PCA/MDS projection (four points data set)



Shepard diagrams for Sammon projection (four points data set).

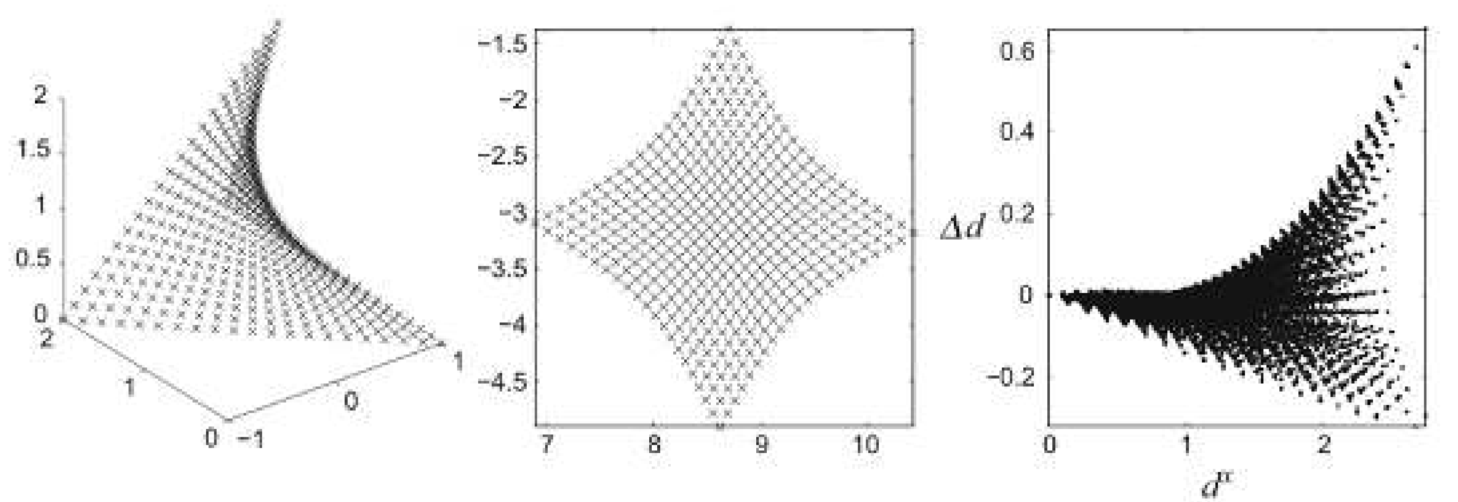
Sammon mapping



$$\mathbb{X} = \{(t, \sin t, \cos t)^T \mid t \in \{0, 0.1, 0.2, \ldots, 50\}\}$$

Helix data set, Sammon projection, and projection errors

Sammon mapping



$$X = \{((s_1 - 1)\cdot(s_2 - 1), s_1, s_2)^T \mid s_1, s_2 \in \{0, 0.1, 0.2, \ldots, 2\}\}$$

Bent square data set, Sammon projection, and projection errors