Advanced Data Analysis: Locality Preserving Projection

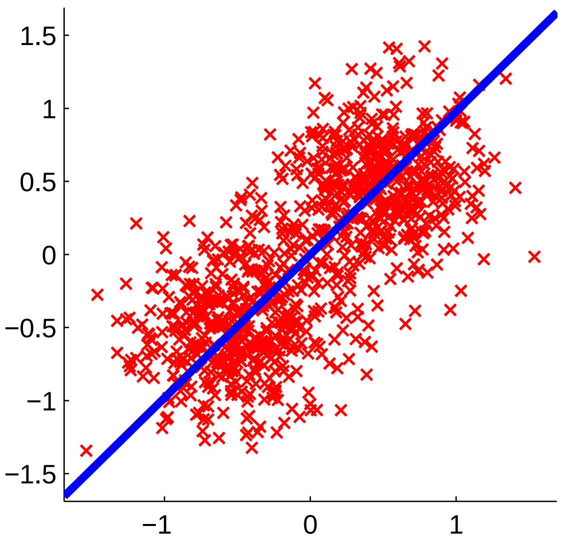
Masashi Sugiyama (Computer Science)

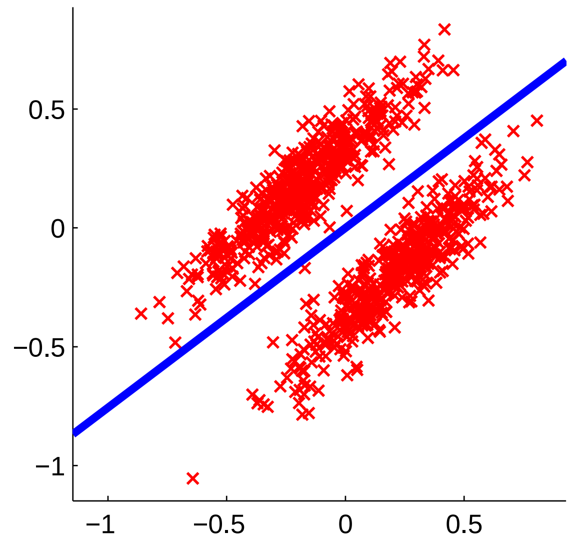
W8E-505, sugi@cs.titech.ac.jp

http://sugiyama-www.cs.titech.ac.jp/~sugi

Locality Preserving Projection (LPP)

- PCA finds a subspace which well describes the data.
- However, PCA can miss some interesting structures such as clusters.
- Another idea: Find a subspace which well preserves “local structures” in the data.





Similarity Matrix

- Similarity matrix $\boldsymbol{W}$: the "similar" $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are, the larger $\boldsymbol{W}_{i,j}$ is.

- Assumptions on $\boldsymbol{W}$:
  - Symmetric: $\boldsymbol{W}_{i,j} = \boldsymbol{W}_{j,i}$
  - Normalized: $0 \leq \boldsymbol{W}_{i,j} \leq 1$
  - Positive: $\boldsymbol{u}^\top \boldsymbol{W} \boldsymbol{u} \geq 0, \ \forall \boldsymbol{u} \neq 0$

- $\boldsymbol{W}$ is also called the affinity matrix.

Examples of Similarity Matrix

- Distance-based:

$$\boldsymbol{W}_{i,j} = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/\gamma^2) \qquad \gamma > 0$$

- Nearest-neighbor-based:

  $\boldsymbol{W}_{i,j} = 1$ if $\boldsymbol{x}_i$ is a k-nearest neighbor of $\boldsymbol{x}_j$ or $\boldsymbol{x}_j$ is a k-nearest neighbor of $\boldsymbol{x}_i$. Otherwise $\boldsymbol{W}_{i,j} = 0$.

- Combination of these two is also possible.

$$\boldsymbol{W}_{i,j} = \begin{cases} \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/\gamma^2) \\ 0 \end{cases}$$

LPP Criterion

- Idea: embed two close points as close, i.e., mininize

(A) $$\sum_{i,j=1}^{n} \|\boldsymbol{B}\boldsymbol{x}_i - \boldsymbol{B}\boldsymbol{x}_j\|^2 \boldsymbol{W}_{i,j} \ (\geq 0)$$

- (A) is expressed as $2\mathrm{tr}(\boldsymbol{B}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^\top \boldsymbol{B}^\top)$

(Homework!)

$$\boldsymbol{X} = (\boldsymbol{x}_1 | \boldsymbol{x}_2 | \cdots | \boldsymbol{x}_n)$$

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$$

$$\boldsymbol{D} = \mathrm{diag}(\textstyle\sum_{j=1}^{n} \boldsymbol{W}_{i,j})$$

- Since $\boldsymbol{B} = \boldsymbol{O}$ gives a meaningless solution, we impose

$$\boldsymbol{B}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^\top \boldsymbol{B}^\top = \boldsymbol{I}_m$$

LPP: Summary

- **LPP criterion**:

$$B_{LPP} = \operatorname*{argmin}_{B \in \mathbb{R}^{m \times d}} \operatorname{tr}(BXLX^\top B^\top)$$

$$\text{subject to } BXDX^\top B^\top = I_m$$

- **Solution** (see previous homework):

$$B_{LPP} = (\psi_d | \psi_{d-1} | \cdots | \psi_{d-m+1})^\top$$

  - $\{\lambda_i, \psi_i\}_{i=1}^m$:Sorted generalized eigenvalues and normalized eigenvectors of $XLX^\top \psi = \lambda XDX^\top \psi$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \qquad \langle XDX^\top \psi_i, \psi_j \rangle = \delta_{i,j}$$

- **LPP embedding of a sample** x :

$$z = B_{LPP}x$$

Generalized Eigenvalue Problem

$$\boldsymbol{A}\boldsymbol{\psi} = \lambda \boldsymbol{C}\boldsymbol{\psi}$$ (B)

- $\boldsymbol{C}$:positive symmetric matrix
- Then there exists a positive symmetric matrix $\boldsymbol{C}^{\frac{1}{2}}$ such that $(\boldsymbol{C}^{\frac{1}{2}})^2 = \boldsymbol{C}$.
  - Eigenvalue decomposition of $\boldsymbol{C}$:

$$\boldsymbol{C} = \sum_i \gamma_i \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^\top \qquad \gamma_i > 0$$

$$\boldsymbol{C}^{\frac{1}{2}} = \sum_i \sqrt{\gamma_i} \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^\top$$

Generalized Eigenvalue Problem

$$\boldsymbol{A}\boldsymbol{\psi} = \lambda \boldsymbol{C}\boldsymbol{\psi} \quad \text{(B)}$$

- Let $\boldsymbol{\phi} = \boldsymbol{C}^{\frac{1}{2}}\boldsymbol{\psi}$. Then (B) yields

$$\boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{\phi} = \lambda\boldsymbol{\phi} \quad \text{(C)}$$

- (C) is an ordinary eigenvalue problem.
- Ordinary eigenvectors are orthogonal:

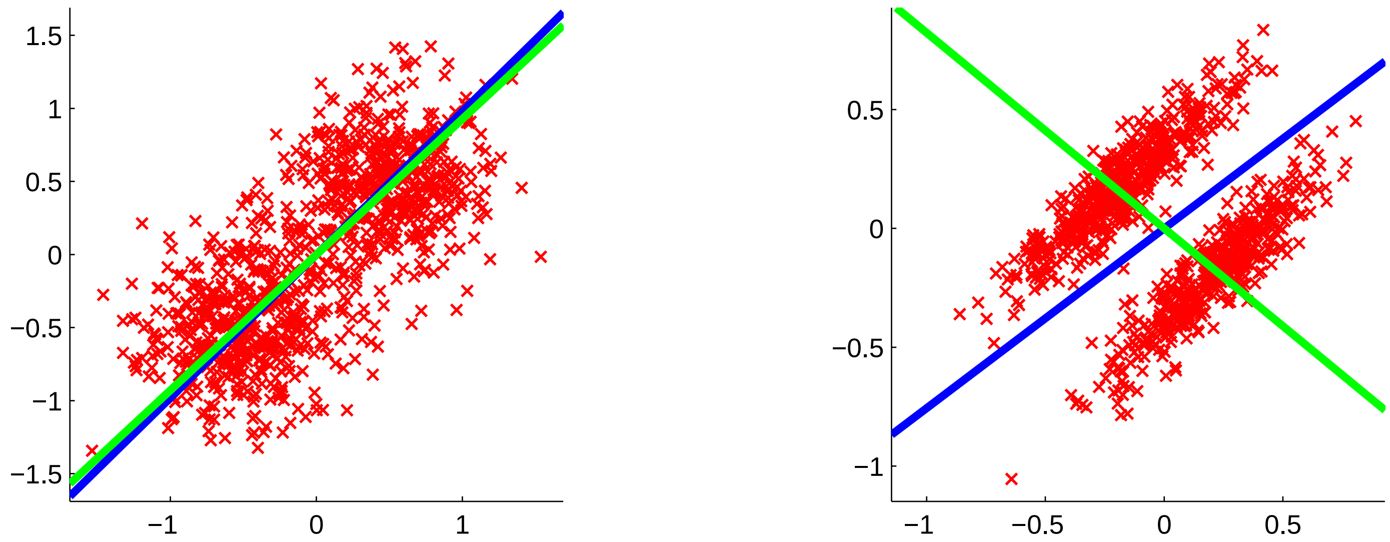
$$\langle \boldsymbol{\phi}_i, \boldsymbol{\phi}_j \rangle \propto \delta_{i,j} = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}$$

- Generalized eigenvectors are $\boldsymbol{C}$ -orthogonal:

$$\langle \boldsymbol{C}\boldsymbol{\psi}_i, \boldsymbol{\psi}_j \rangle \propto \delta_{i,j}$$

Examples

Blue: PCA Green: LPP

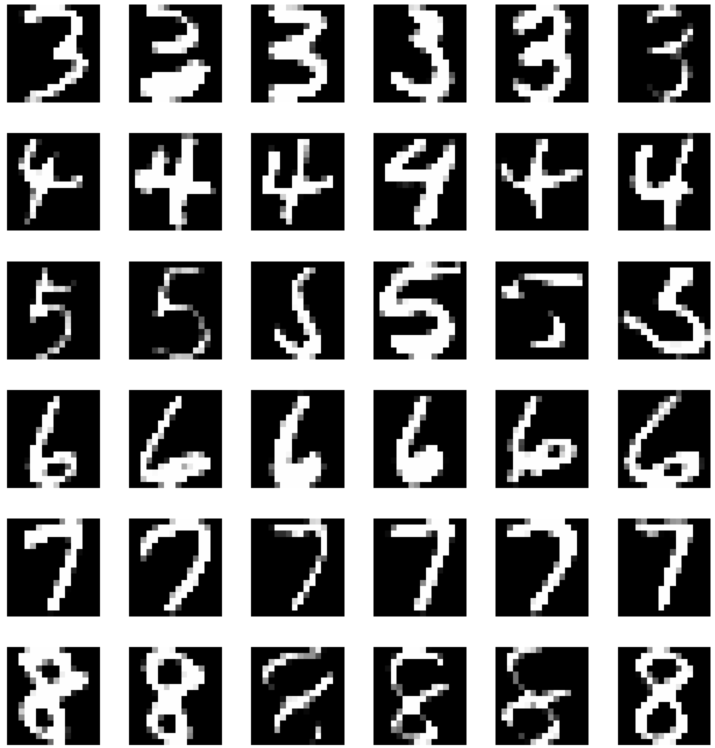


Note: Similarity matrix is defined by the nearest-neighbor-based method with 50 nearest neighbors.

- LPP can describe the data well, and also it preserves cluster structure.
- LPP is intuitive, easy to implement, analytic solution available, and fast.

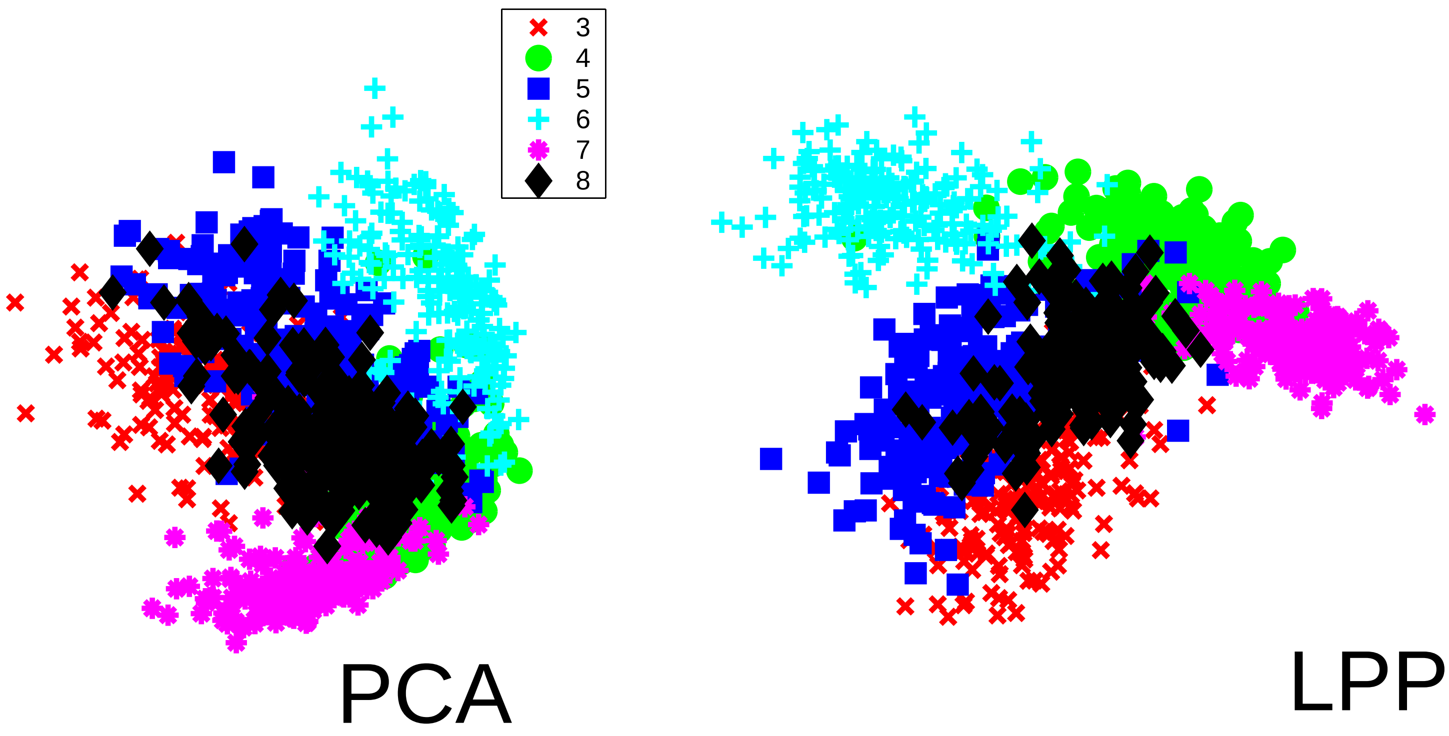
Examples (cont.)

- Embedding hand-written numerals from 3 to 8.
- Each image consists of 16x16 pixels.



Examples (cont.)

- LPP finds slightly clearer clusters than PCA?



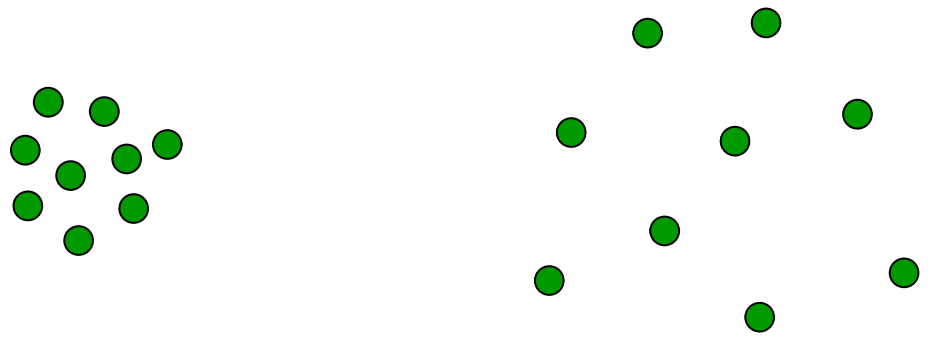
Drawbacks of LPP

- Obtained result depends on the similarity matrix $\boldsymbol{W}$.
- Appropriately constructing similarity matrix (e.g., k, γ) is not always easy.

Local Scaling of Samples

- Density of samples may be locally different.

Dense region Sparse region



- Using the same γ globally in the similarity matrix may not be appropriate.

$$\boldsymbol{W}_{i,j} = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/\gamma^2)$$

Local Scaling Heuristic

- γ_i : scaling around the sample $\boldsymbol{x}_i$

$$\gamma_i = \|\boldsymbol{x}_i - \boldsymbol{x}_i^{(k)}\|$$

$\boldsymbol{x}_i^{(k)}$: k-th nearest neighbor sample of $\boldsymbol{x}_i$

- Local scaling based similarity matrix:

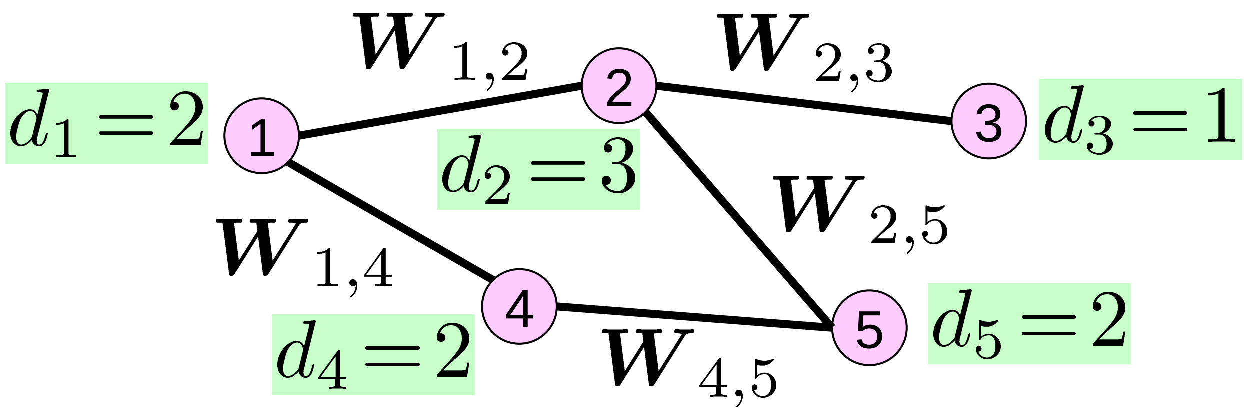
$$\boldsymbol{W}_{i,j} = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/(\gamma_i\gamma_j))$$

- A heuristic choice is $k = 7$.

L. Zelnik-Manor & P. Perona, Self-tuning spectral clustering, Advances in Neural Information Processing Systems 17, 1601-1608, MIT Press, 2005.

Graph Theory

- Graph: A set of vertices and edges
- Adjacency matrix $\boldsymbol{W}$: $\boldsymbol{W}_{i,j}$ is the number of edges from i -th to j -th vertices.
- Vertex degree d_i :Number of connected edges at i -th vertex.



Spectral Graph Theory

- Spectral graph theory studies relationships between the properties of a graph and its adjacency matrix.
- Graph Laplacian $\boldsymbol{L}$:

$$\boldsymbol{L}_{i,j} = \begin{cases} d_i & (i = j) \\ -1 & (i \neq j \text{ and } \boldsymbol{W}_{i,j} > 0) \\ 0 & (\text{otherwise}) \end{cases}$$

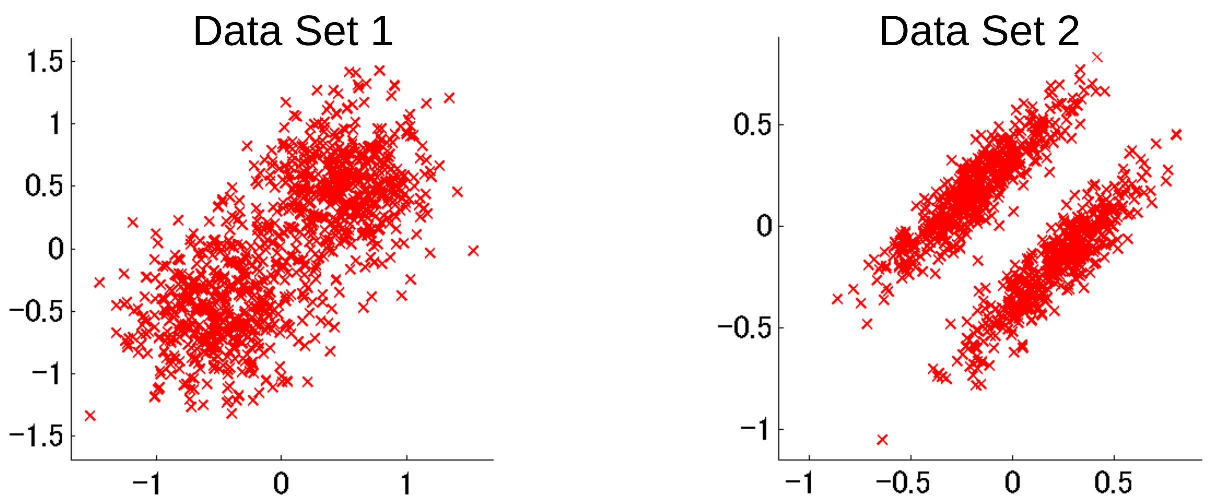
Relation to Spectral Graph Theory

- Suppose our similarity matrix $\boldsymbol{W}$ is defined by nearest neighbors.
- Consider the following graph:
  - Each vertex corresponds to each point $\boldsymbol{x}_i$
  - Edge exists if $\boldsymbol{W}_{i,j} > 0$
- $\boldsymbol{W}$ is the adjacency matrix.
- $\boldsymbol{D}$ is the diagonal matrix of vertex degrees.
- $\boldsymbol{L}$ is the graph Laplacian.

Homework

1. Implement LPP and reproduce the 2-dimensional examples shown in the class (data sets 1 and 2).

http://sugiyama-www.cs.titech.ac.jp/~sugi/data/DataAnalysis



Test LPP with your own (artificial or real) data and analyze the characteristics of LPP.

Homework (cont.)

2. Prove

$$\sum_{i,j=1}^{n} \|\boldsymbol{B}\boldsymbol{x}_i - \boldsymbol{B}\boldsymbol{x}_j\|^2 \boldsymbol{W}_{i,j} = 2\mathrm{tr}(\boldsymbol{B}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^\top \boldsymbol{B}^\top)$$

$$\boldsymbol{X} = (\boldsymbol{x}_1 | \boldsymbol{x}_2 | \cdots | \boldsymbol{x}_n)$$

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$$

$$\boldsymbol{D} = \mathrm{diag}(\textstyle\sum_{j=1}^{n} \boldsymbol{W}_{i,j})$$

Suggestion

- If you are interested in spectral graph theory, the following book would be interesting.
  - Chung, F. R. K., *Spectral Graph Theory*, American Mathematical Society, 1997.

- Read the following article for the next class:
  - M. Sugiyama: Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis, Journal of Machine Learning Research, 8(May), 1027-1061, 2007.

http://www.jmlr.org/papers/volume8/sugiyama07b/sugiyama07b.pdf

Advanced Data Analysis: Golden Week Special Homework

- A 4-dimensional data set (“data set X”) is available from

  http://sugiyama-www.cs.titech.ac.jp/~sugi/data/DataAnalysis/4d-x.txt

  Apply PCA and extract information.