Locality Preserving Projection

- PCA: Find 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.
- This could be achieved by embedding the data so that two close points are kept close.

Similarity Matrix

- Define $\boldsymbol{W}$ so that the "similar" $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are, the larger $\boldsymbol{W}_{i,j}$ is.
- Assumptions on $\boldsymbol{W}$:
  - Symmetric: $\boldsymbol{W}_{i,j} = \boldsymbol{W}_{j,i}$
  - Positive entries: $\boldsymbol{W}_{i,j} \geq 0$
  - Invertible: $\exists \boldsymbol{W}^{-1}$
- $\boldsymbol{W}$ is often called the similarity matrix or affinity matrix

Examples of Similarity Matrix

- Distance-based:

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

- 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 two is also possible.

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

LPP Criterion

- Embedding two close points as close:

$$\min_{\boldsymbol{B}\in\mathbb{R}^{m\times d}} \left[\sum_{i,j=1}^{n} \|\boldsymbol{B}\boldsymbol{x}_i - \boldsymbol{B}\boldsymbol{x}_j\|^2 \boldsymbol{W}_{i,j}\right],$$

- Exercise:

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

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

$$\boldsymbol{X} = (\boldsymbol{x}_1|\boldsymbol{x}_2|\cdots|\boldsymbol{x}_n) \qquad \boldsymbol{D} = \mathrm{diag}(\textstyle\sum_{j=1}^{n} \boldsymbol{W}_{i,j})$$

Constraint

$$\min_{\boldsymbol{B}\in\mathbb{R}^{m\times d}} \left[\mathrm{tr}(\boldsymbol{B}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^\top\boldsymbol{B}^\top)\right]$$

- Since $\boldsymbol{B}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^\top\boldsymbol{B}^\top$ is positive semi-definite,

$$\min_{\boldsymbol{B}\in\mathbb{R}^{m\times d}} \left[\mathrm{tr}(\boldsymbol{B}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^\top\boldsymbol{B}^\top)\right] = 0$$

- A trivial solution exists: $\boldsymbol{B} = \boldsymbol{O}$
- To avoid this, we impose some constraint:

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

LPP Embedding

$$\boldsymbol{B}_{LPP} = \underset{\boldsymbol{B}\in\mathbb{R}^{m\times d}}{\operatorname{argmin}} \left[\operatorname{tr}(\boldsymbol{B}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^\top\boldsymbol{B}^\top)\right]$$
$$\text{subject to } \boldsymbol{B}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^\top\boldsymbol{B}^\top = \boldsymbol{I}_m$$

- Previous homework gives a solution:

$$\boldsymbol{B}_{LPP} = (\boldsymbol{\psi}_{d-m+1}|\boldsymbol{\psi}_{d-m+2}|\cdots|\boldsymbol{\psi}_d)^\top$$

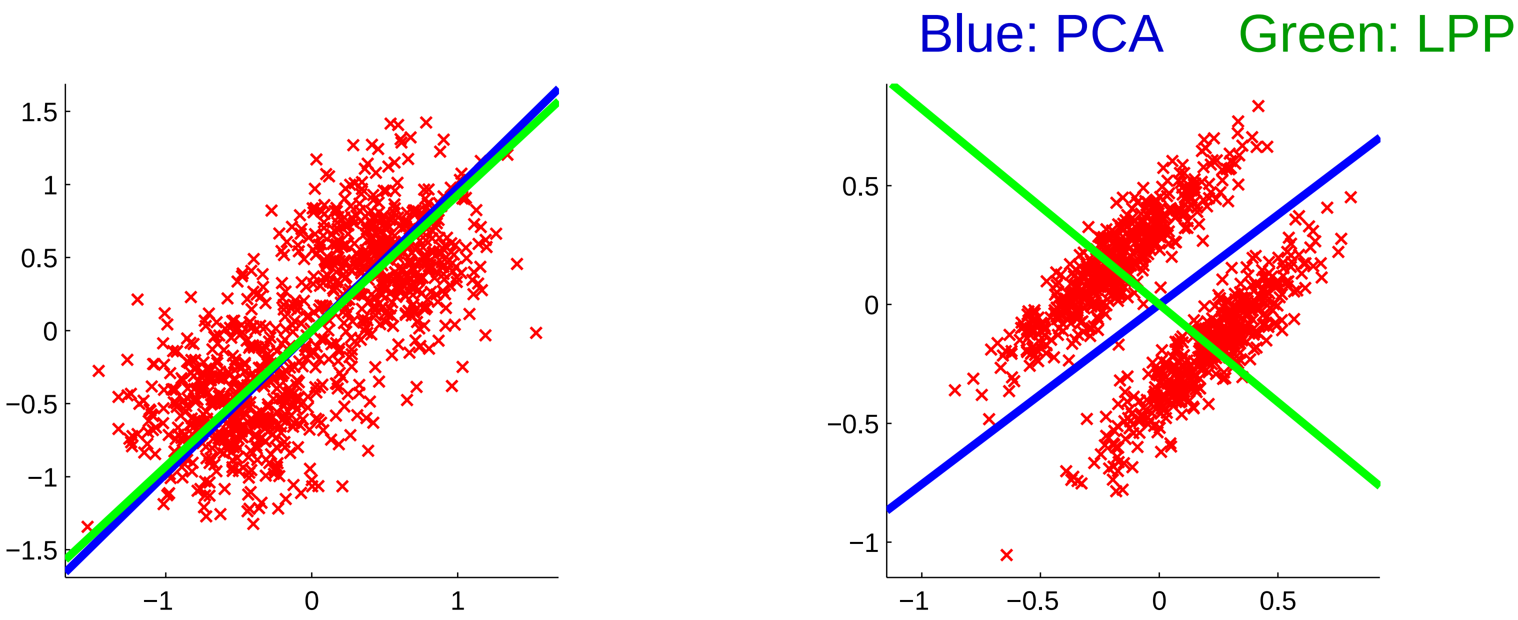
- $\{\lambda_i, \boldsymbol{\psi}_i\}_{i=1}^m$:Generalized eigenvalues and eigenvectors of $\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^\top\psi = \lambda\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^\top\psi$

$$(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d) \qquad (\langle\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^\top\boldsymbol{\psi}_i, \boldsymbol{\psi}_j\rangle = \delta_{i,j})$$

- LPP Embedding of a sample $\boldsymbol{x}'$:

$$\boldsymbol{z}' = \boldsymbol{B}_{LPP}\boldsymbol{x}'$$

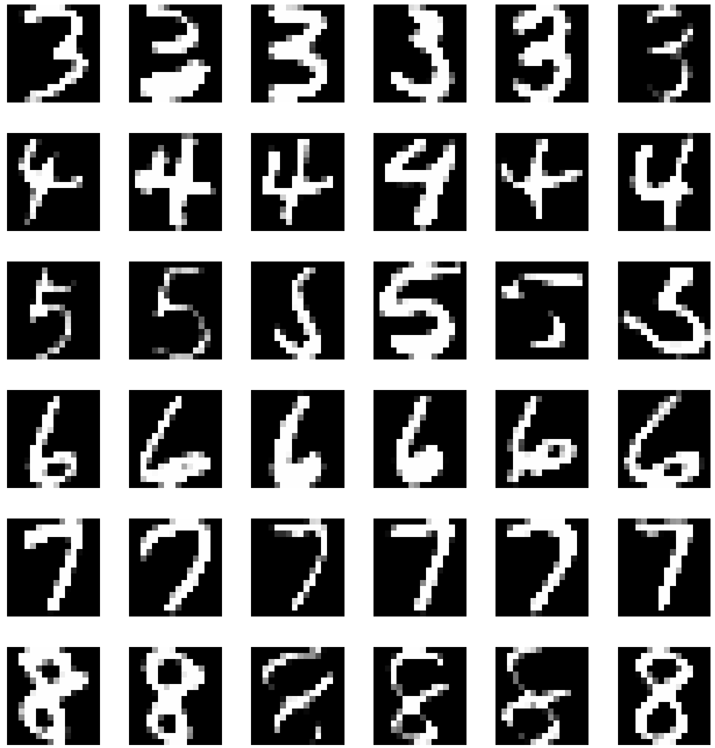
Examples



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

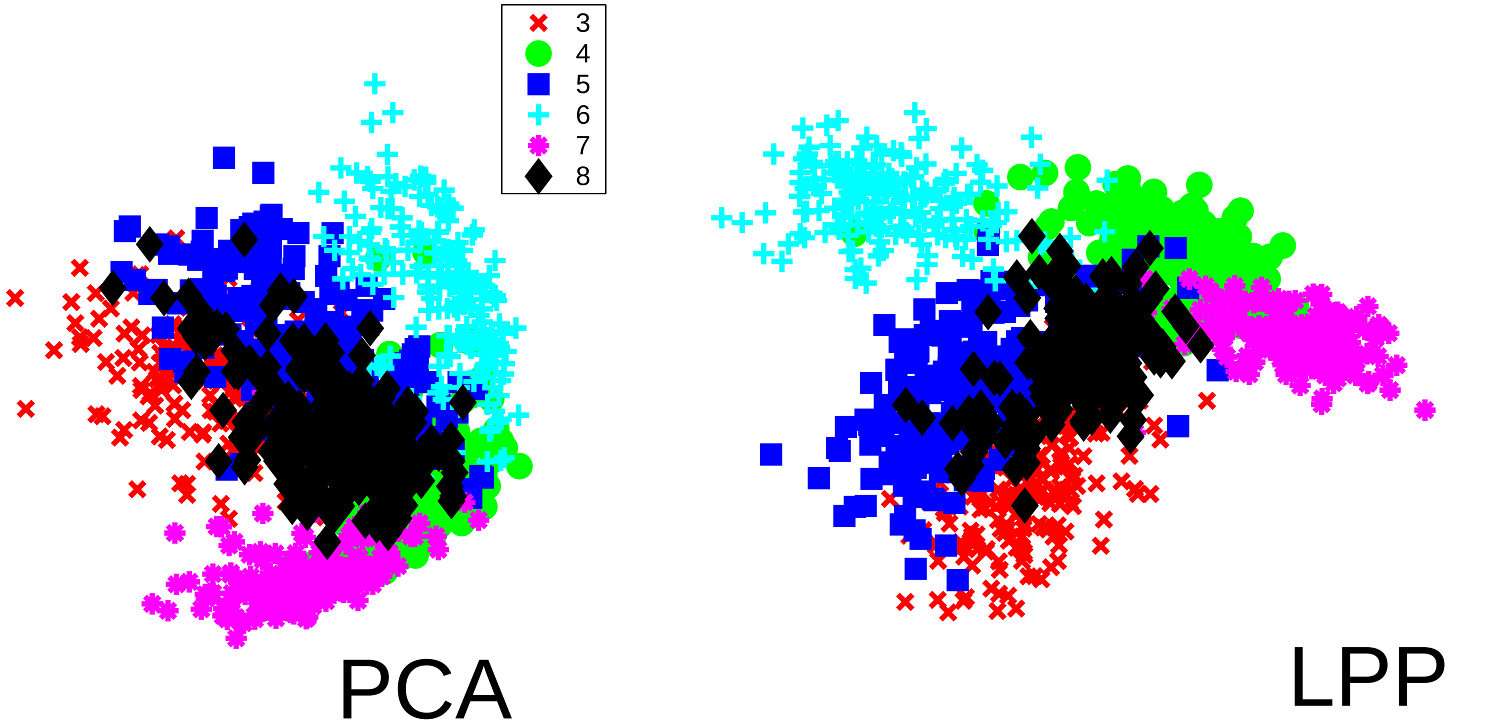
Examples (cont.)

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



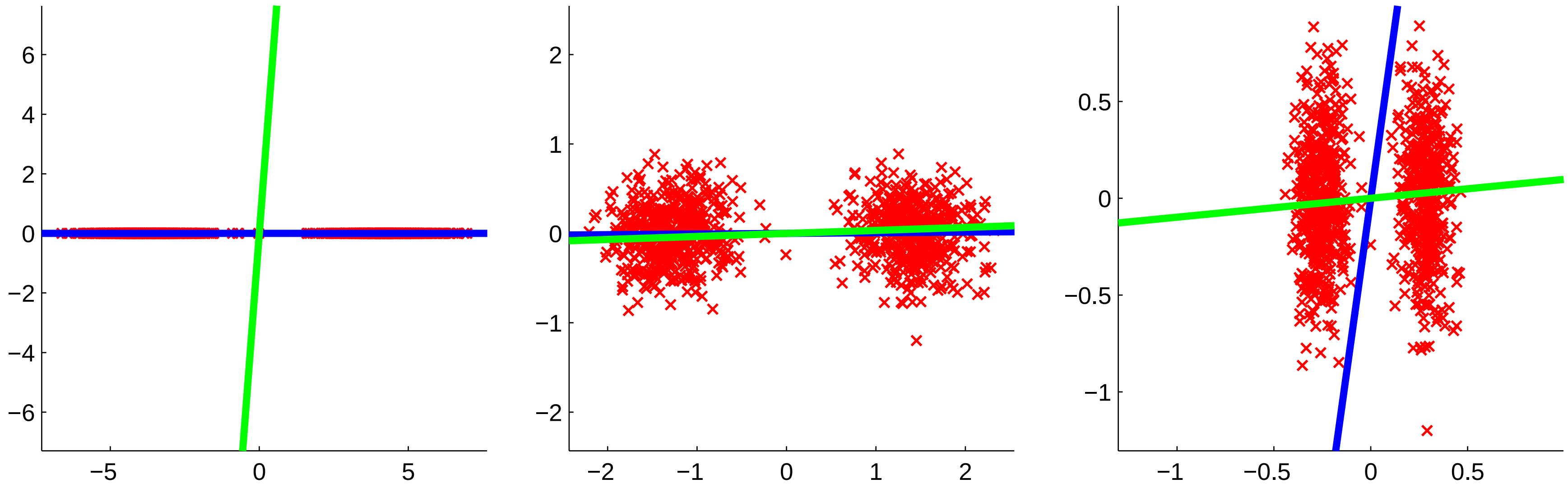
Examples (cont.)

- LPP can find slightly clear clusters than PCA.



Drawback

Blue: PCA Green: LPP



- Rescaling of the data affects the results of LPP (as well as PCA).
- “cm vs. kg” and “m vs. kg” can differ.

Homework

1. Data visualization: Embed your data sets into 2- or 3-dimensional subspace by PCA or LPP.
2. Data mining: Find something “interesting” from the visualized data.

Have a nice Golden Week!