ICML2006, Pittsburgh, USA June 25-29, 2006

Local Fisher Discriminant Analysis for Supervised Dimensionality Reduction

Masashi Sugiyama

Tokyo Institute of Technology, Japan

Dimensionality Reduction

- High dimensional data is not easy to handle:

  Need to reduce dimensionality

- We focus on
  - Linear dimensionality reduction:

$$\boldsymbol{T}^\top \in \mathbb{R}^{R \times D}$$

$$\boldsymbol{z} \in \mathbb{R}^R \qquad \boldsymbol{z} = \boldsymbol{T}^\top \boldsymbol{x} \qquad \boldsymbol{x} \in \mathbb{R}^D$$

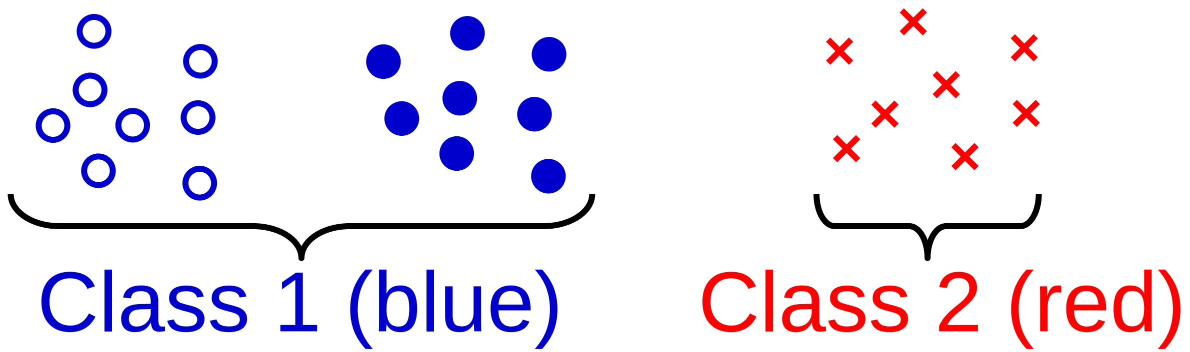
$$R \ll D$$

  - Supervised dimensionality reduction:

$$(\boldsymbol{x}, y) \qquad y \in \{1, 2, \dots, C\}$$

Within-Class Multimodality

One of the classes has several modes

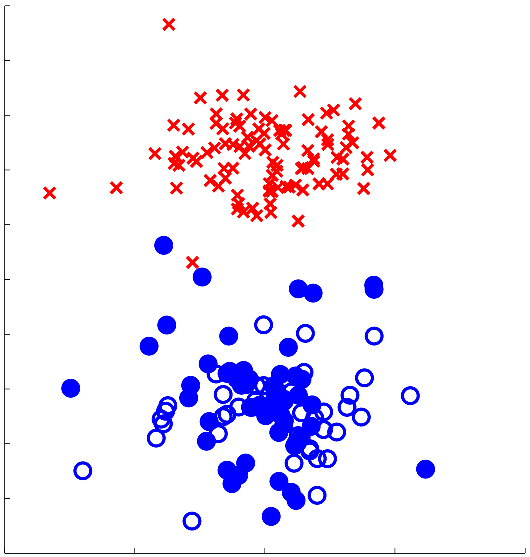


- Medical checkup:
  hormone imbalance (high/low) vs. normal
- Digit recognition:
  even (0,2,4,6,8) vs. odd (1,3,5,7,9)
- Multi-class classification:
  one vs. rest

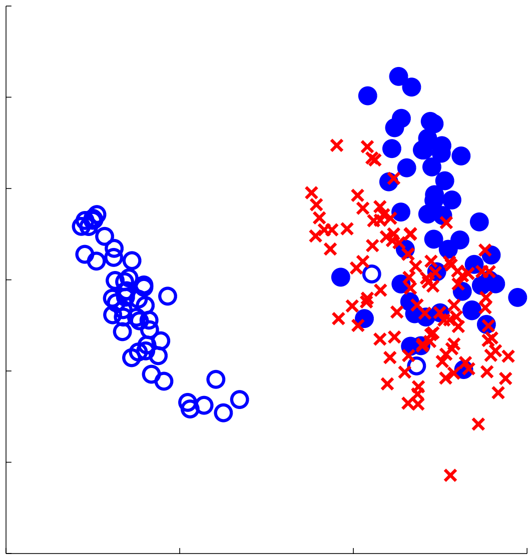
Goal of This Research

- We want to embed multimodal data so that
  - Between-class separability is maximized
  - Within-class multimodality is preserved

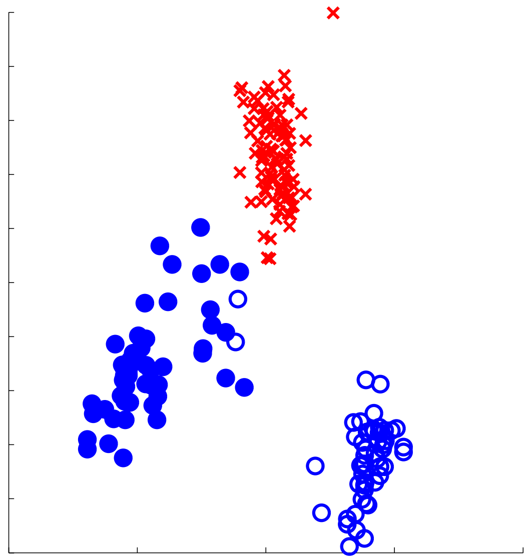
Separable but within-class multimodality lost



Within-class multimodality preserved but non-separable



Separable and within-class multimodality preserved



Fisher Discriminant Analysis (FDA)

Fisher (1936)

- Within-class scatter matrix:

$$S^{(w)} = \sum_{c=1}^{C} \sum_{i:y_i=c} (x_i - \mu_c)(x_i - \mu_c)^\top$$

- Between-class scatter matrix:

$$S^{(b)} = \sum_{c=1}^{C} n_c(\mu_c - \mu)(\mu_c - \mu)^\top$$

- FDA criterion:

$$\max_{T} \left[\mathrm{tr}\left((T^\top S^{(w)} T)^{-1} T^\top S^{(b)} T \right) \right]$$

  - Within-class scatter is made small
  - Between-class scatter is made large

Interpretation of FDA

- Pairwise expressions:

n_c :Number of samples in class c
n :Total number of samples

$$\boldsymbol{S}^{(w)} = \frac{1}{2}\sum_{i,j=1}^{n} \boldsymbol{A}^{(w)}_{i,j}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top$$

$$\boldsymbol{A}^{(w)}_{i,j} = \begin{cases} 1/n_c & (y_i = y_j = c) \\ 0 & (y_i \neq y_j) \end{cases}$$

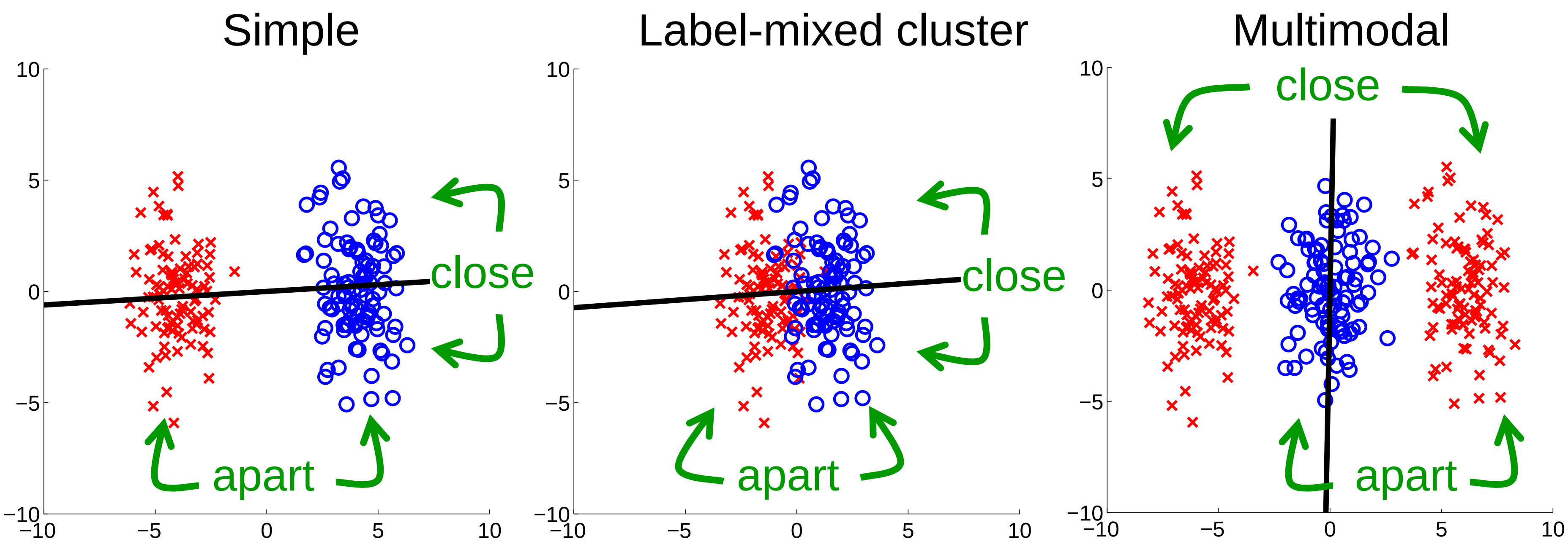
$$\boldsymbol{S}^{(b)} = \frac{1}{2}\sum_{i,j=1}^{n} \boldsymbol{A}^{(b)}_{i,j}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top$$

$$\boldsymbol{A}^{(b)}_{i,j} = \begin{cases} 1/n - 1/n_c & (y_i = y_j = c) \\ 1/n & (y_i \neq y_j) \end{cases}$$

- Samples in the same class are made close
- Samples in different classes are made apart

Examples of FDA

$\mathbb{R}^2 \Longrightarrow \mathbb{R}^1$



FDA does not take
within-class multimodality into account

NOTE: FDA can extract only *C-1* features since

$$\mathrm{rank}(\boldsymbol{S}^{(b)}) = C - 1$$

C :Number of classes

Locality Preserving Projection (LPP)

He & Niyogi (NIPS2003)

- Locality matrix:

$$\boldsymbol{S}^{(l)} = \frac{1}{2}\sum_{i,j=1}^{n} \boldsymbol{A}_{i,j}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top$$

- Affinity matrix:

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

- LPP criterion:

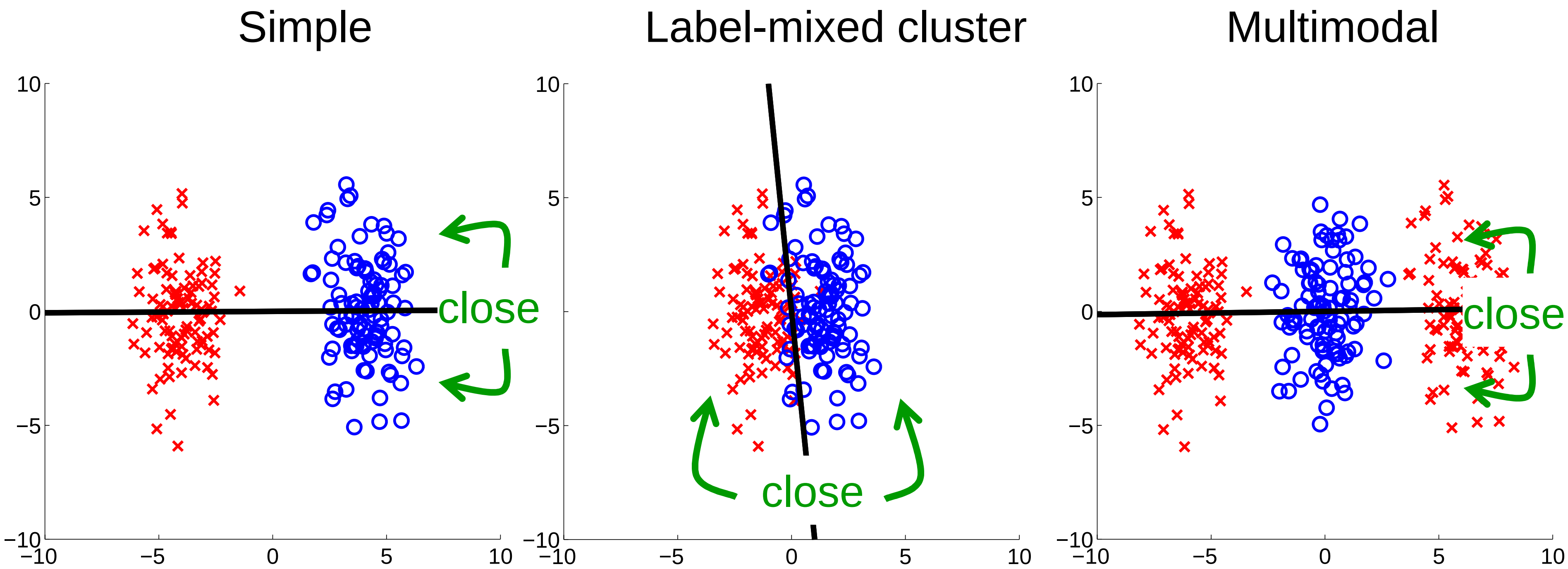
$$\min_{\boldsymbol{T}} \left[\mathrm{tr}(\boldsymbol{T}^\top \boldsymbol{S}^{(l)} \boldsymbol{T})\right]$$

$$\text{subject to } \boldsymbol{T}^\top \boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^\top \boldsymbol{T} = \boldsymbol{I}$$

  - Nearby samples in original space are made close
  - Constraint is to avoid $\boldsymbol{T} = \boldsymbol{O}$

Examples of LPP

$$\mathbb{R}^2 \Longrightarrow \mathbb{R}^1$$

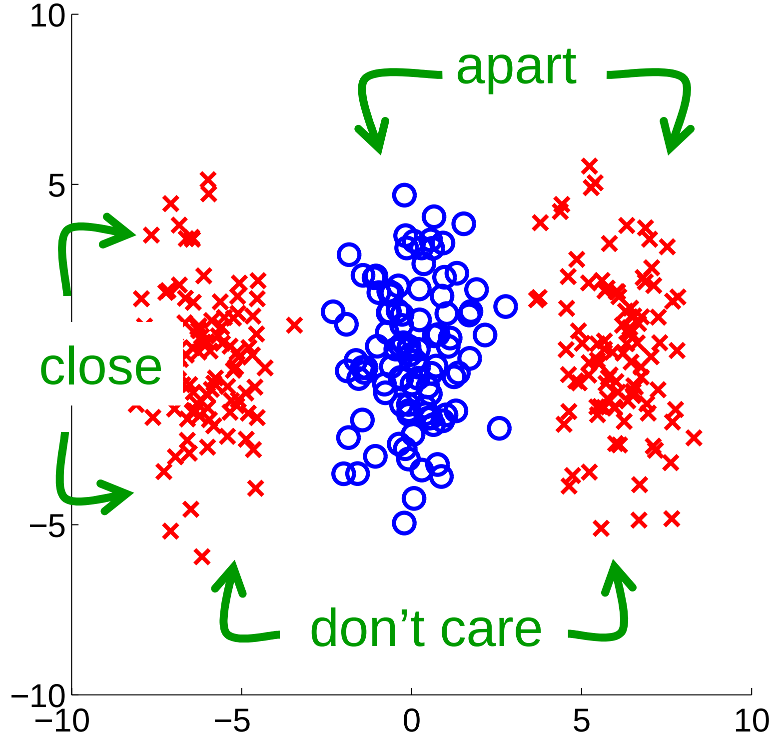


LPP does not take between-class separability into account (unsupervised)

Our Approach

We combine FDA and LPP

- Nearby samples in the same class are made close
- Far-apart samples in the same class are not made close
- Samples in different classes are made apart



Local Fisher Discriminent Analysis

$$\max_{\boldsymbol{T}} \left[\mathrm{tr}((\boldsymbol{T}^\top \widetilde{\boldsymbol{S}}^{(w)} \boldsymbol{T})^{-1} \boldsymbol{T}^\top \widetilde{\boldsymbol{S}}^{(b)} \boldsymbol{T}) \right]$$

- Local within-class scatter matrix:

$$\widetilde{\boldsymbol{S}}^{(w)} = \frac{1}{2} \sum_{i,j=1}^{n} \widetilde{\boldsymbol{A}}_{i,j}^{(w)} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top$$

$$\widetilde{\boldsymbol{A}}_{i,j}^{(w)} = \begin{cases} \boldsymbol{A}_{i,j}/n_c & (y_i = y_j = c) \\ 0 & (y_i \neq y_j) \end{cases}$$

- Local between-class scatter matrix:

$$\widetilde{\boldsymbol{S}}^{(b)} = \frac{1}{2} \sum_{i,j=1}^{n} \widetilde{\boldsymbol{A}}_{i,j}^{(b)} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top$$

$$\widetilde{\boldsymbol{A}}_{i,j}^{(b)} = \begin{cases} \boldsymbol{A}_{i,j}(1/n - 1/n_c) & (y_i = y_j = c) \\ 1/n & (y_i \neq y_j) \end{cases}$$

How to Obtain Solution

$$\boldsymbol{T}_{LFDA} = \underset{\boldsymbol{T}}{\mathrm{argmax}} \left[\mathrm{tr}((\boldsymbol{T}^\top \widetilde{\boldsymbol{S}}^{(w)} \boldsymbol{T})^{-1} \boldsymbol{T}^\top \widetilde{\boldsymbol{S}}^{(b)} \boldsymbol{T}) \right]$$

- Since LFDA has a similar form to FDA, solution can be obtained just by solving a generalized eigenvalue problem:

$$\widetilde{\boldsymbol{S}}^{(b)} \boldsymbol{\varphi} = \lambda \widetilde{\boldsymbol{S}}^{(w)} \boldsymbol{\varphi}$$

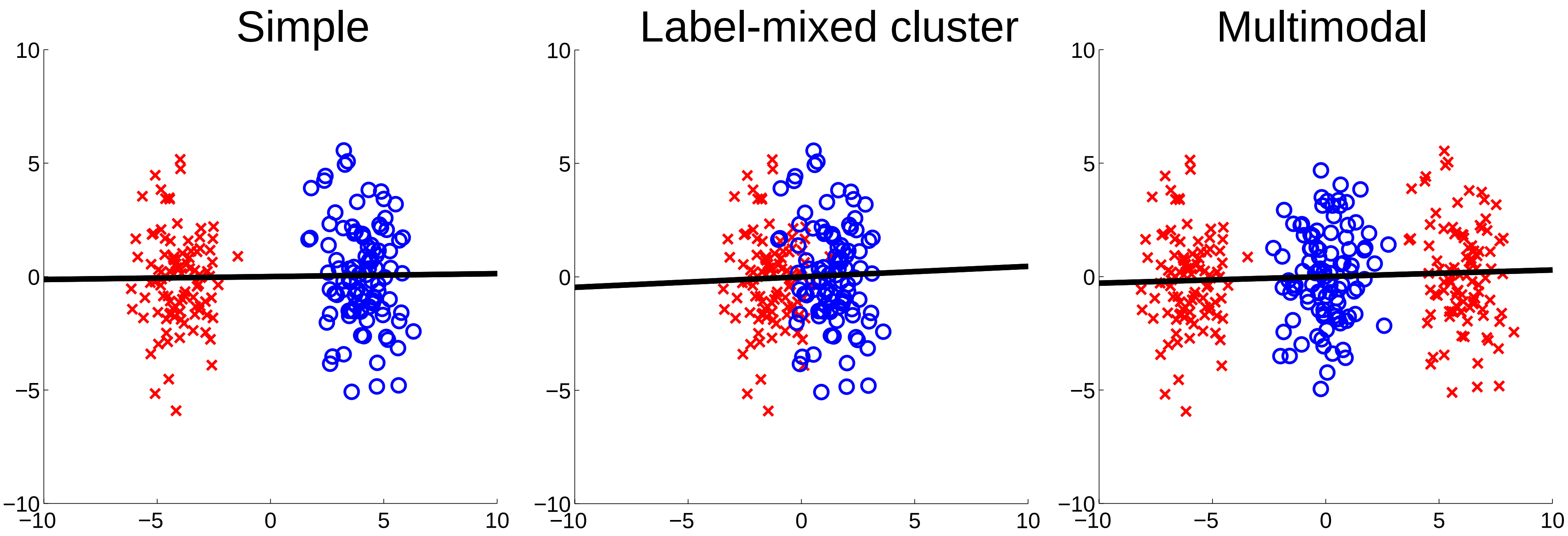
$$\boldsymbol{T}_{LFDA} = (\widetilde{\boldsymbol{\varphi}}_1 | \widetilde{\boldsymbol{\varphi}}_2 | \cdots | \widetilde{\boldsymbol{\varphi}}_R)$$

$$\widetilde{\boldsymbol{\varphi}}_1, \widetilde{\boldsymbol{\varphi}}_2, \ldots, \widetilde{\boldsymbol{\varphi}}_R$$

$$\widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \cdots \geq \widetilde{\lambda}_D$$

Examples of LFDA

$$\mathbb{R}^2 \Longrightarrow \mathbb{R}^1$$



LFDA works well for all three cases!

Note: Usually $\mathrm{rank}(\widetilde{\boldsymbol{S}}^{(b)}) \gg C$ so LFDA can extract more than C features (cf. FDA)

Neighborhood Component Analysis (NCA)

Goldberger, Roweis, Hinton & Salakhutdinov (NIPS2004)

- Minimize leave-one-out error of a stochastic k-nearest neighbor classifier
- Obtained embedding is separable
- NCA involves non-convex optimization

  → There are local optima
- No analytic solution available

  → Slow iterative algorithm
- LFDA has analytic form of global solution

Maximally Collapsing Metric Learning (MCML)

Globerson & Roweis (NIPS2005)

- Idea is similar to FDA
  - Samples in the same class are close (“one point”)
  - Samples in different classes are apart
- MCML involves non-convex optimization
- There exists a nice convex approximation

  → Non-global solution
- No analytic solution available

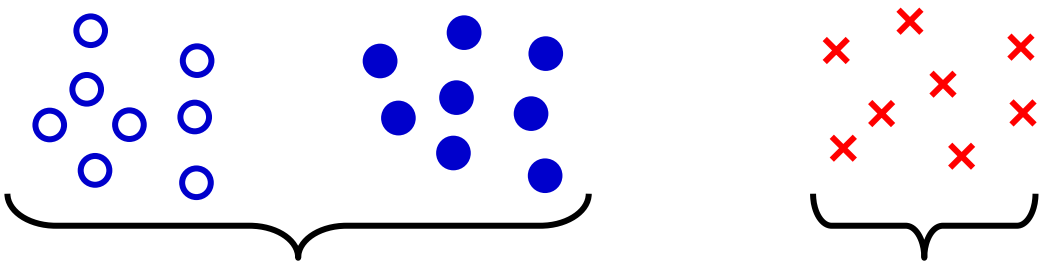
  → Slow iterative algorithm

Simulations

- Visualization of UCI data sets:
  - Letter recognition (D=16)
  - Segment (D=18)
  - Thyroid disease (D=5)
  - Iris (D=4)

$\mathbb{R}^D \Longrightarrow \mathbb{R}^2$

- Extract 3 classes from original data
- Merge 2 classes



Class 1 (blue) Class 2 (red)

Summary of Simulation Results

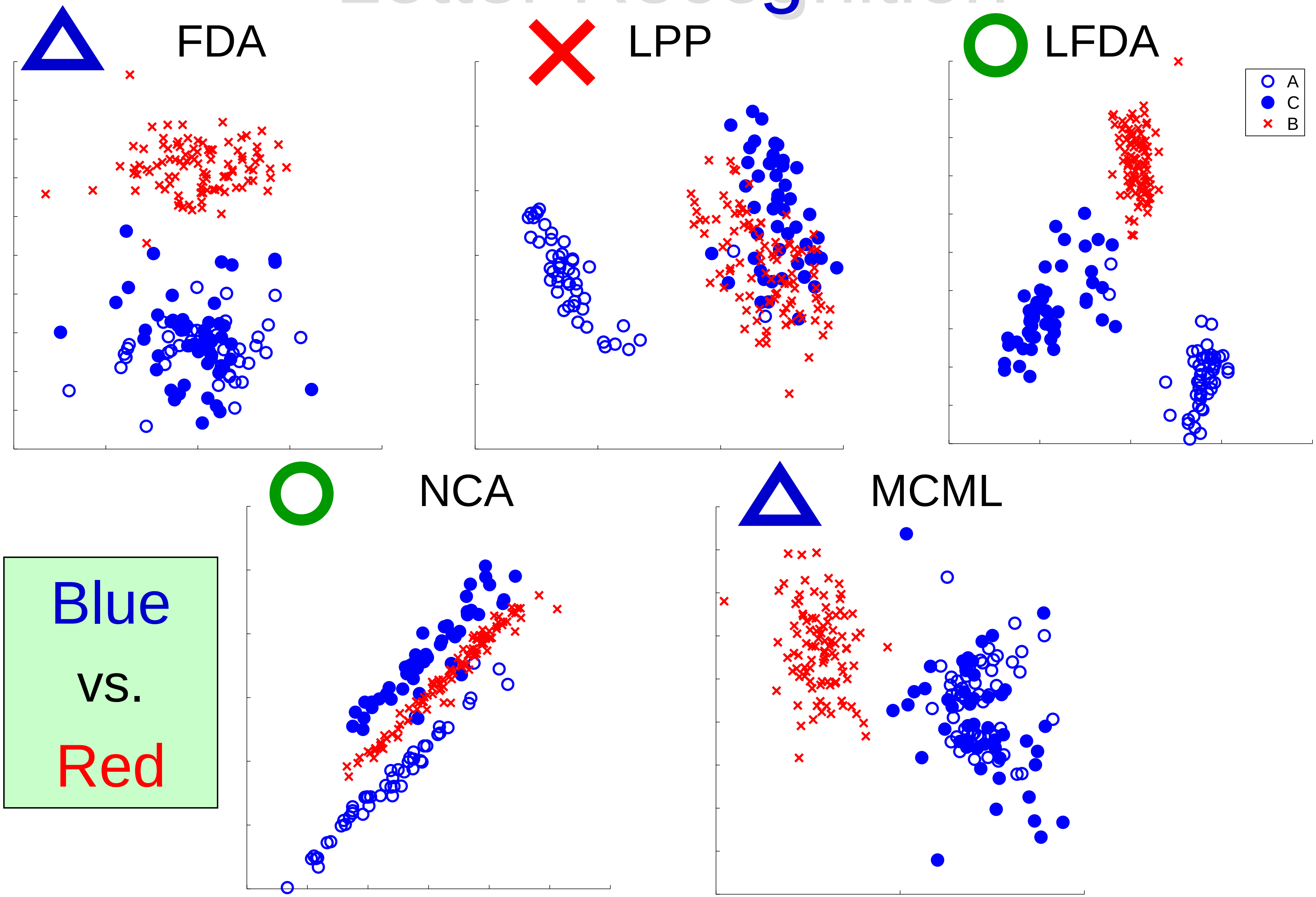
	Lett	Segm	Thyr	Iris	Comments
FDA	△	△	△	×	No multi-modal
LPP	×	×	○	○	No label-separability
LFDA	○	○	○	○	
NCA	○	×	○	○	Slow, local optima
MCML	△	○	○	○	Slow, no multi-modal

○ Separable and multimodality preserved

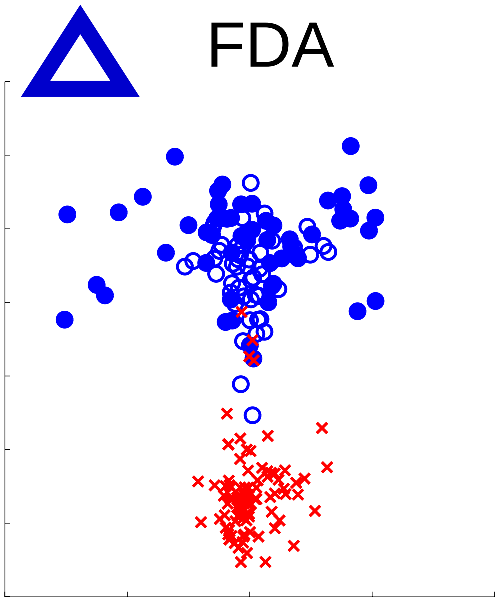
△ Separable but no multimodality

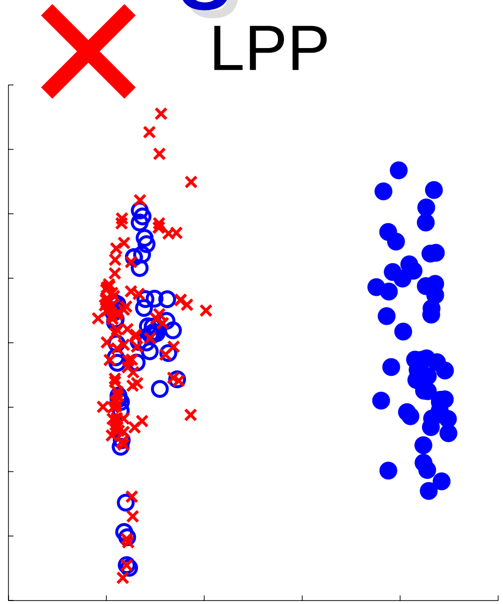
× Multimodality preserved but no separability

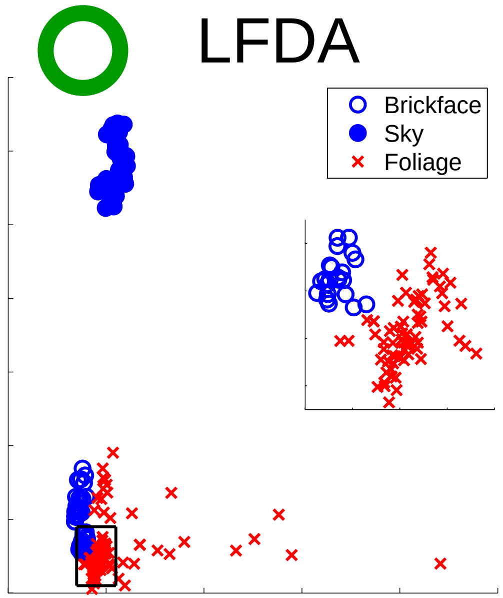
Letter Recognition



Segment



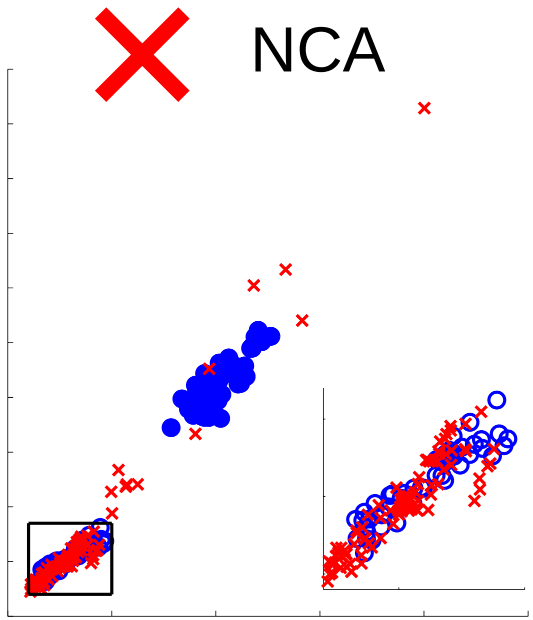


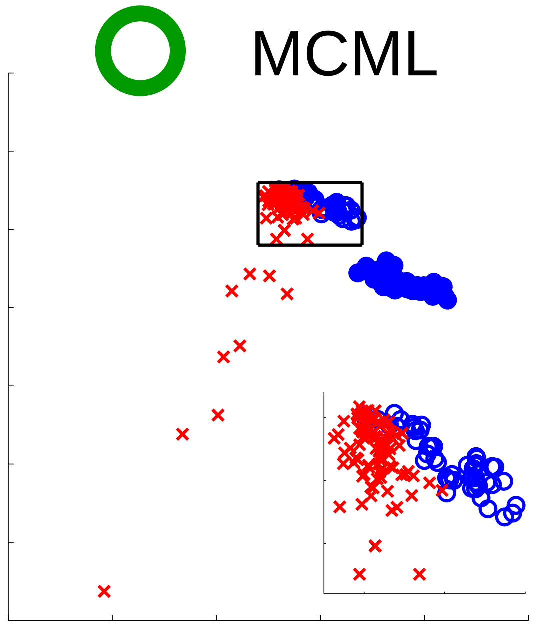


Blue

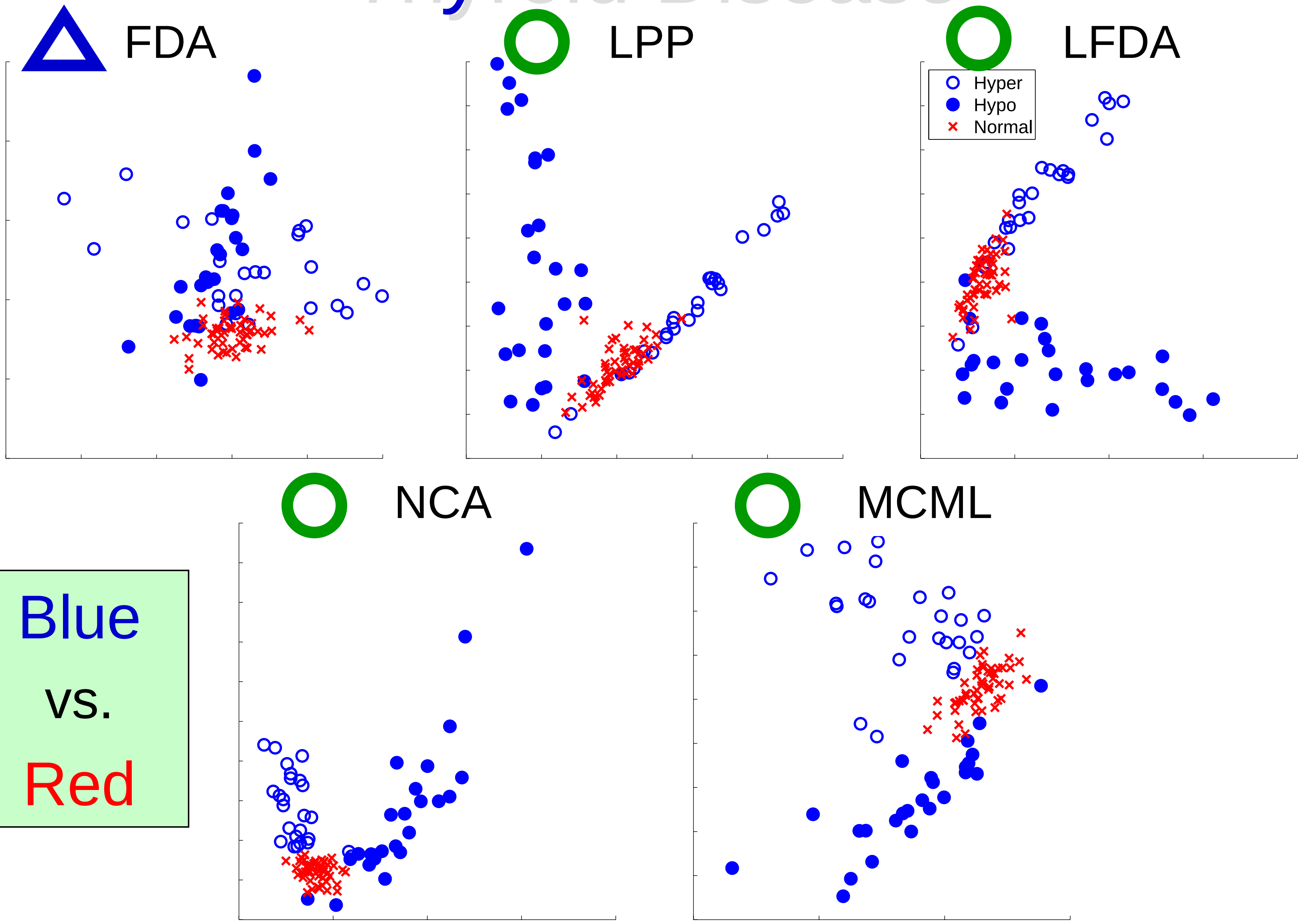
vs.

Red





Thyroid Disease



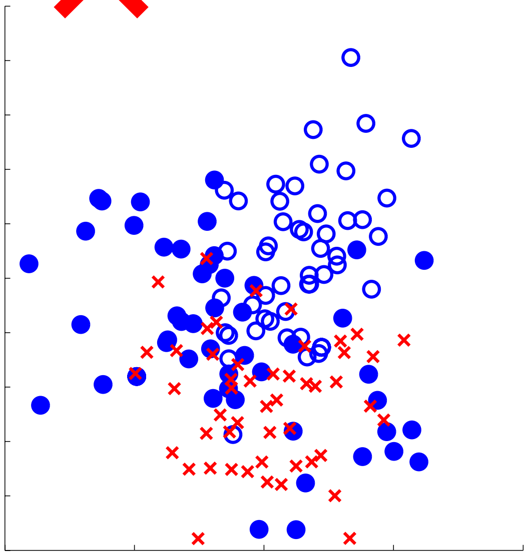
Blue

vs.

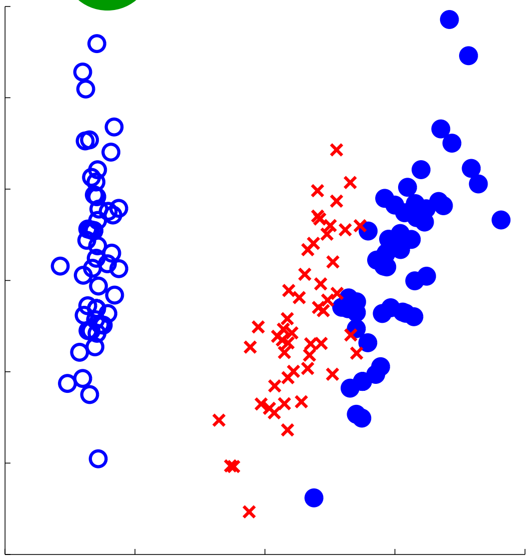
Red

Iris

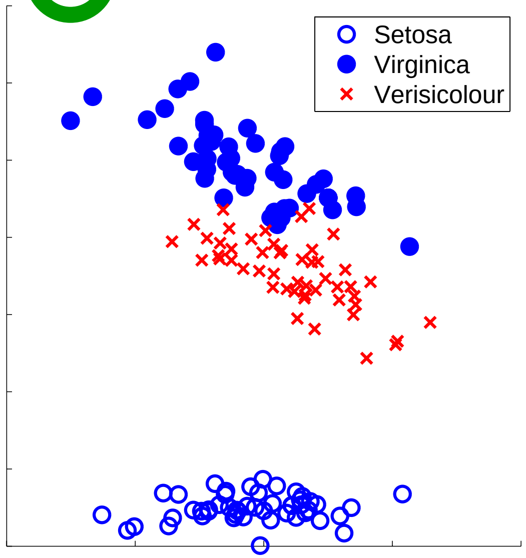
FDA



LPP



LFDA

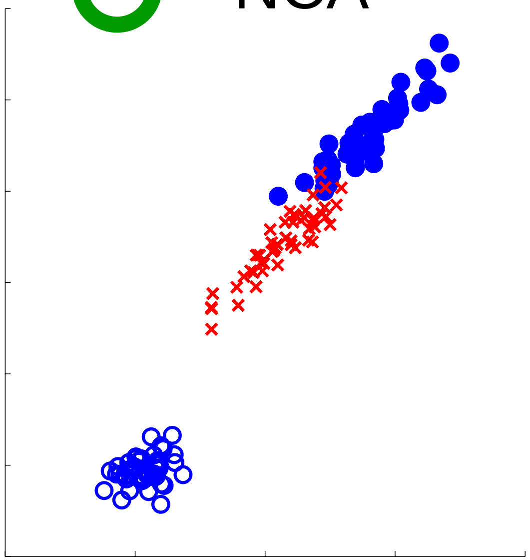


Blue

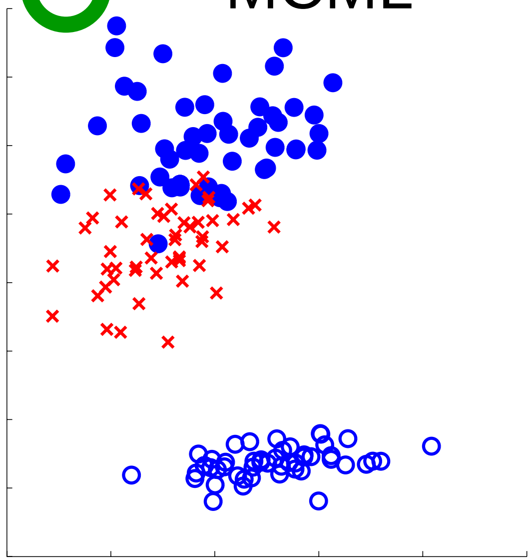
vs.

Red

NCA



MCML



Kernelization

- LFDA can be non-linearized by kernel trick

$$\langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

- FDA: Kernel FDA — Mika *et al*. (NNSP1999)
- LPP: Laplacian eigenmap — Belkin & Niyogi (NIPS2001)
- MCML: Kernel MCML — Globerson & Roweis (NIPS2005)
- NCA: not available yet?

Conclusions

- LFDA effectively combines FDA and LPP.
- LFDA is suitable for embedding multimodal data.
- Same as FDA, LFDA has analytic optimal solution thus computationally efficient.
- Same as LPP, LFDA needs to pre-specify affinity matrix.
- We used local scaling method for computing affinity, which does not include any tuning parameter.

Zelnik-Manor & Perona (NIPS2004)