PAKDD2008 May 20-23, 2008

Semi-Supervised Local Fisher Discriminant Analysis for Dimensionality Reduction

Masashi Sugiyama (Tokyo Tech.)
Tsuyoshi Ide (IBM)
Shinichi Nakajima (Nikon)
Jun Sese (Ochanomizu Univ.)

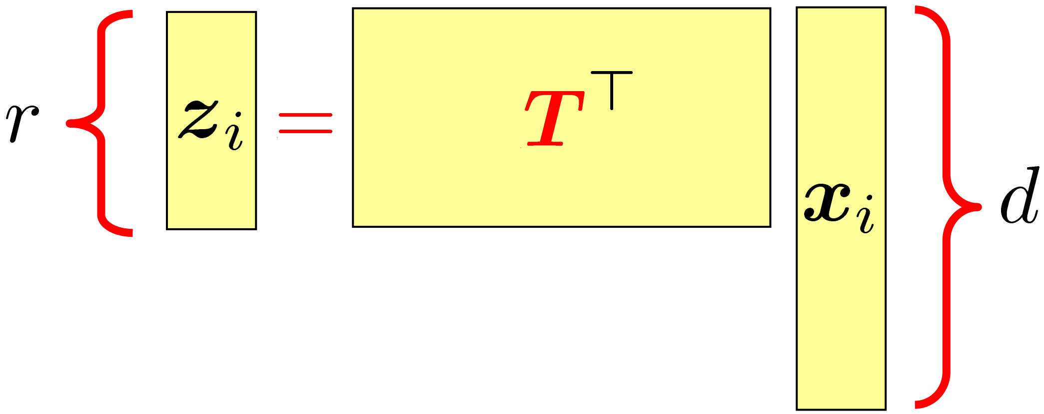
Dimensionality Reduction

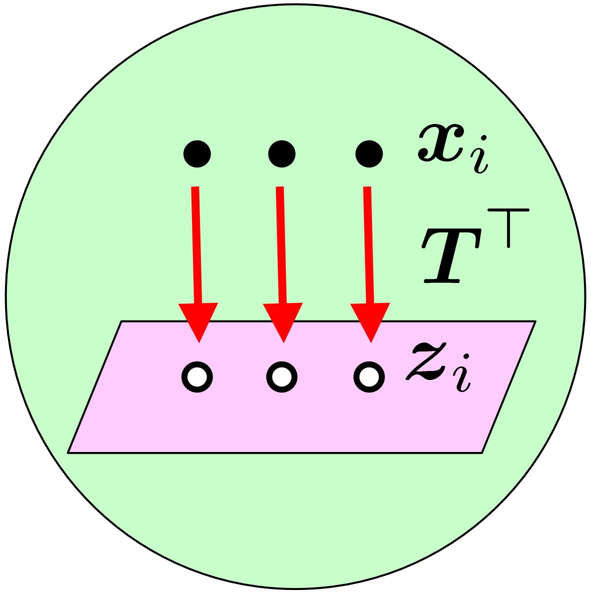
- Curse of dimensionality: High-dimensional data is hard to deal with

→ We want to reduce dimensionality while keeping intrinsic information

Linear Dimensionality Reduction

- We focus on linear dimensionality reduction:
  - High-dimensional samples: $\{\boldsymbol{x}_i\}_{i=1}^n$ $\boldsymbol{x}_i \in \mathbb{R}^d$
  - Embedding matrix: $\boldsymbol{T}$
  - Embedded samples: $\{\boldsymbol{z}_i\}_{i=1}^n$ $\boldsymbol{z}_i \in \mathbb{R}^r$





- Goal: Find appropriate embedding matrix $\boldsymbol{T}$

Organization

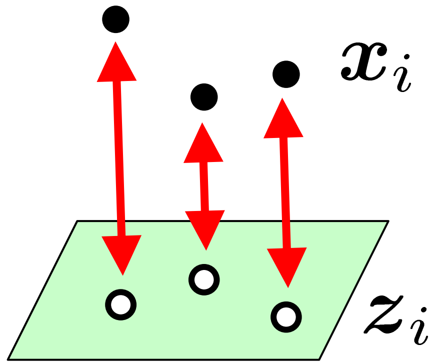
1. Linear dimensionality reduction
2. Unsupervised methods:
   - Principal component analysis (PCA)
   - Locality preserving projection (LPP)
3. Supervised methods:
   - Fisher discriminant analysis (FDA)
   - Local Fisher discriminant analysis (LFDA)
4. Semi-supervised method:
   - Semi-supervised LFDA (SELF)
5. Conclusions

Principal Component Analysis (PCA)

Unsupervised learning:

- Unlabeled samples

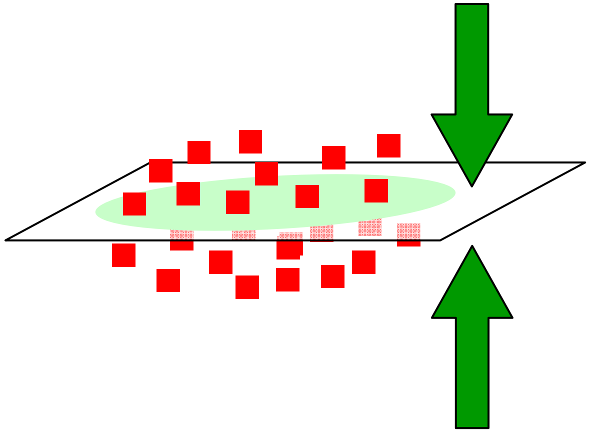
$\{\boldsymbol{x}_i\}_{i=1}^{n}$ $\boldsymbol{x}_i \in \mathbb{R}^d$



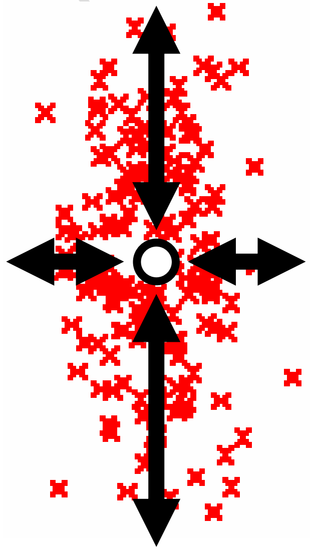
Basic idea of PCA:

- Find the embedding subspace that gives the best approximation to the original samples

- Equivalent to finding the embedding subspace with the largest variance



Principal Component Analysis (PCA)



- Total scatter matrix:

$$\boldsymbol{S}^{(t)} = \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \qquad \boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$$

- PCA criterion: maximize scatter after embedding

$$\max_{\boldsymbol{T}} \left[\mathrm{tr}\left(\boldsymbol{T}^\top \boldsymbol{S}^{(t)} \boldsymbol{T} \underbrace{(\boldsymbol{T}^\top \boldsymbol{T})^{-1}}_{\text{normalization}}\right) \right]$$

normalization

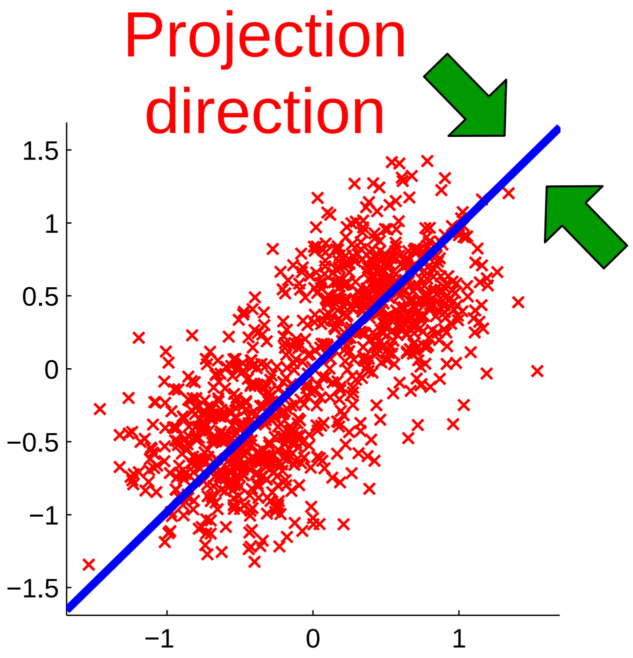
- Solution: major eigenvectors of $\boldsymbol{S}^{(t)}$

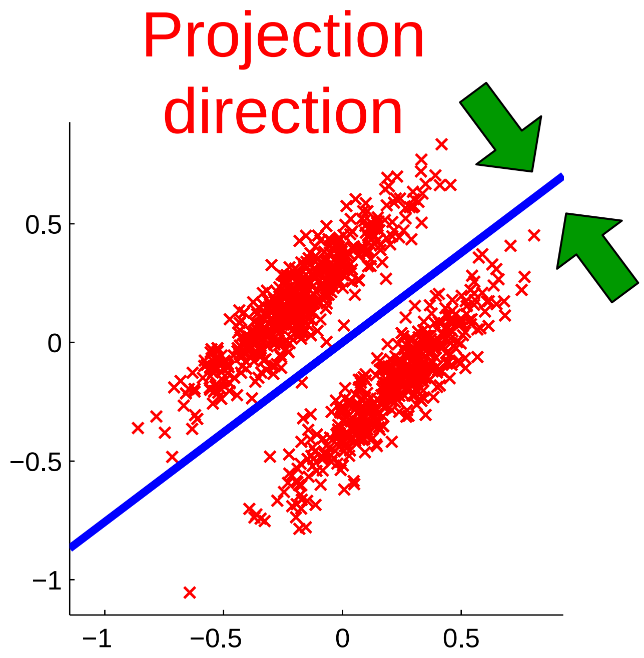
$$\boldsymbol{T}_{PCA} = (\boldsymbol{\varphi}_1 | \boldsymbol{\varphi}_2 | \cdots | \boldsymbol{\varphi}_r)$$

$$\boldsymbol{S}^{(t)} \boldsymbol{\varphi} = \lambda \boldsymbol{\varphi} \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

Examples of PCA

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





- Global structure is well preserved.
- But, local structure such as clusters is not necessarily preserved.

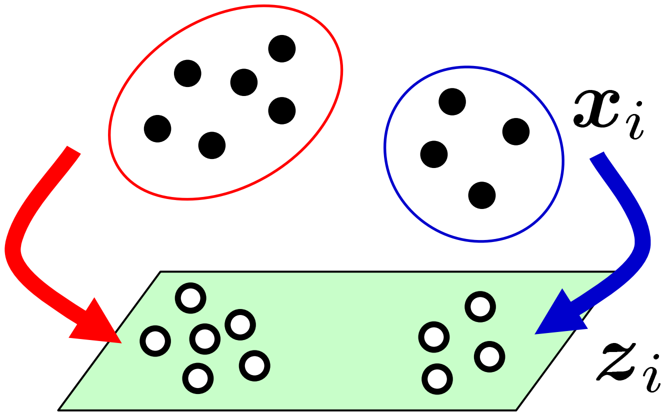
Organization

1. Linear dimensionality reduction
2. Unsupervised methods:
   - Principal component analysis (PCA)
   - Locality preserving projection (LPP)
3. Supervised methods:
   - Fisher discriminant analysis (FDA)
   - Local Fisher discriminant analysis (LFDA)
4. Semi-supervised method:
   - Semi-supervised LFDA (SELF)
5. Conclusions

Locality Preserving Projection (LPP)

He & Niyogi (NIPS2003)

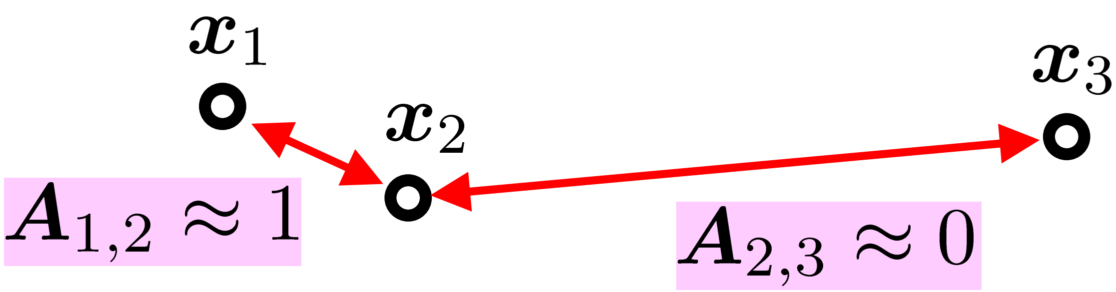
- Basic idea: Embed similar samples close

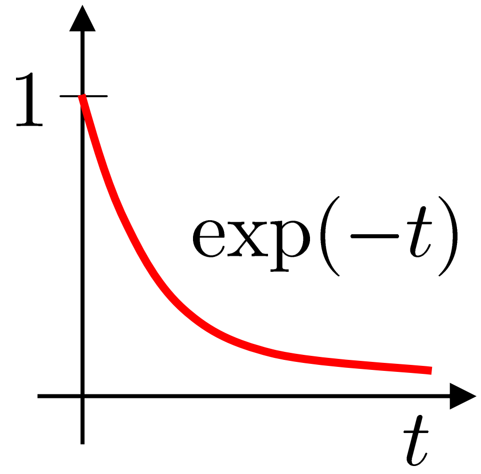


⇒ Local structure tends to be preserved.

Affinity Matrix

- Nearby samples have large affinity
- Far-apart samples have small affinity





- Example:

$$\boldsymbol{A}_{i,j} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\sigma^2}\right)$$

- Choice of affinity is arbitrary.

Local Scaling Heuristic

Zelnik-Manor & Perona (NIPS2005)

- Local scaling based affinity matrix:

$$A_{i,j} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\gamma_i \gamma_j} \right)$$

- 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 $\boldsymbol{x}_i$

- A heuristic choice is $k = 7$.

NOTE: We may cross-validate k in supervised cases if necessary

Locality Preserving Projection (LPP)

- Locality matrix:

$A_{i,j}$:Affinity matrix

$$S^{(l)} = \frac{1}{2n} \sum_{i,j=1}^{n} A_{i,j}(x_i - x_j)(x_i - x_j)^\top$$

- LPP criterion: put samples with large affinity close

$$\min_{T} \left[\mathrm{tr}(T^\top S^{(l)} T \underbrace{(T^\top T)^{-1}}_{\text{Normalization}}) \right]$$

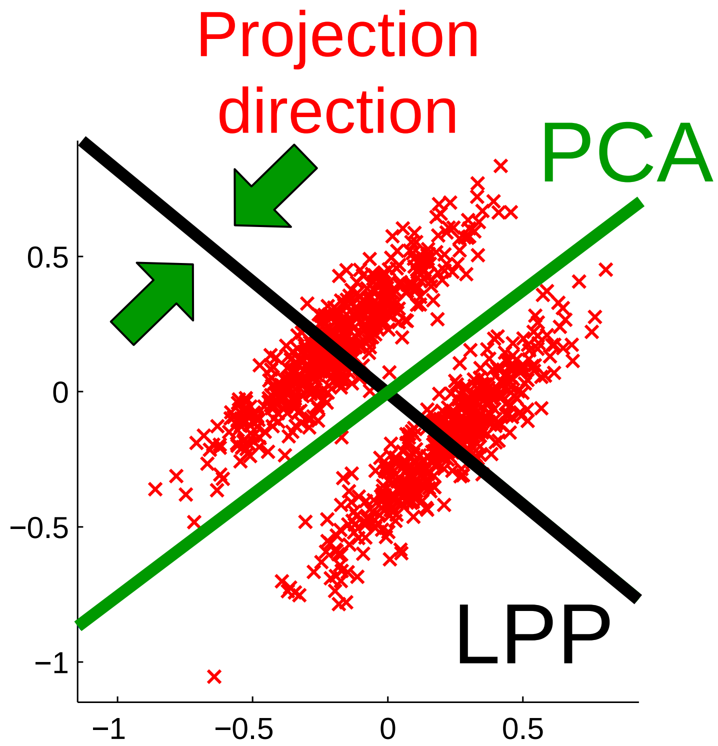
Normalization

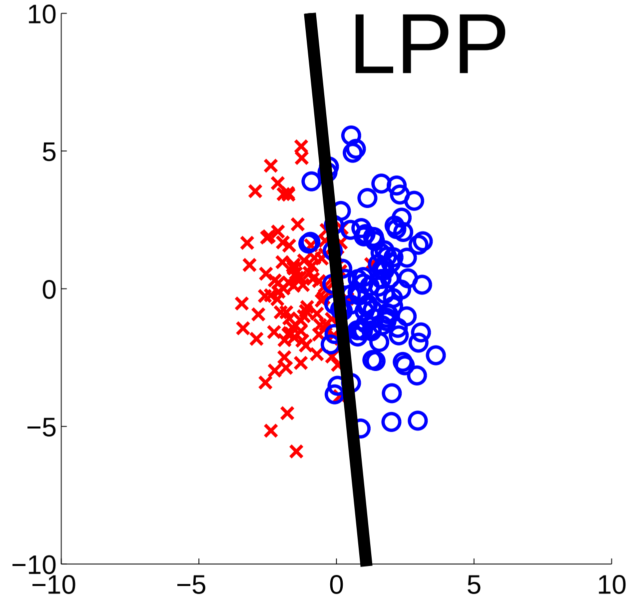
- Solution: minor eigenvectors of $S^{(l)}$

$$T_{LPP} = (\varphi_d | \varphi_{d-1} | \cdots | \varphi_{d-r+1})$$

$$S^{(l)}\varphi = \lambda\varphi \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

Examples of LPP





- Cluster structure tends to be preserved.
- Class-separability is not taken into account due to unsupervised nature.

Organization

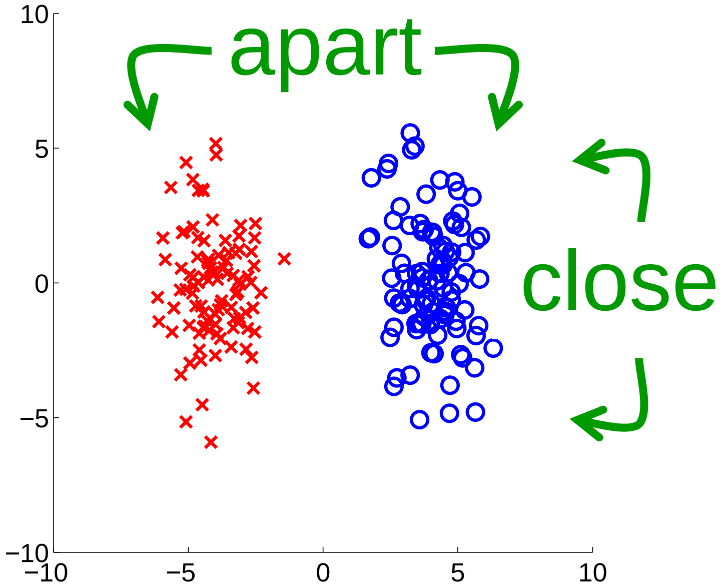
1. Linear dimensionality reduction
2. Unsupervised methods:
   - Principal component analysis (PCA)
   - Locality preserving projection (LPP)
3. Supervised methods:
   - Fisher discriminant analysis (FDA)
   - Local Fisher discriminant analysis (LFDA)
4. Semi-supervised method:
   - Semi-supervised LFDA (SELF)
5. Conclusions

Supervised Dimensionality Reduction

- Supervised learning:
  - Labeled samples

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \qquad y_i \in \{1, 2, \ldots, c\}$$

- Put samples in the same class close
- Put samples in different classes apart



Fisher Discriminant Analysis (FDA)

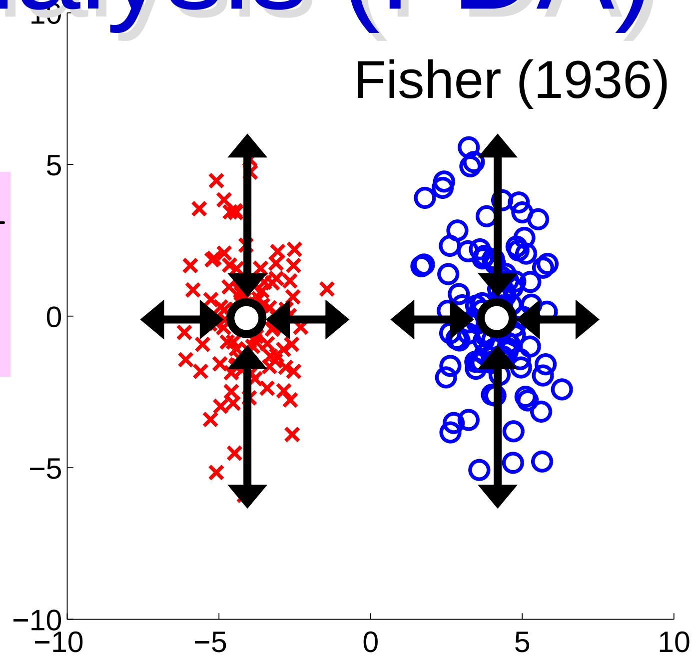
Fisher (1936)

Within-class scatter matrix:

$$\boldsymbol{S}^{(w)} = \sum_{m=1}^{c} \sum_{i:y_i=m} (\boldsymbol{x}_i - \boldsymbol{\mu}_m)(\boldsymbol{x}_i - \boldsymbol{\mu}_m)^\top$$

$$\boldsymbol{\mu}_m = \frac{1}{n_m} \sum_{i:y_i=m} \boldsymbol{x}_i$$

n_m : # of samples in class m

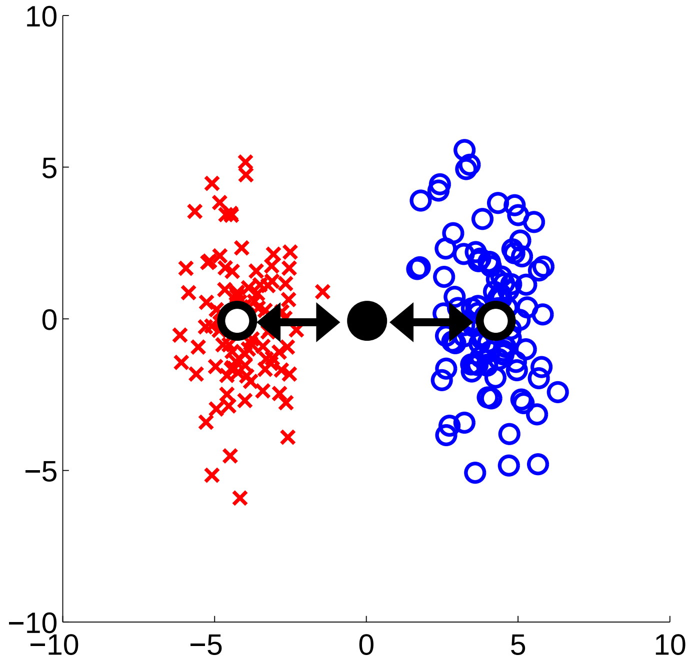


Between-class scatter matrix:

$$\boldsymbol{S}^{(b)} = \sum_{m=1}^{c} n_m (\boldsymbol{\mu}_m - \boldsymbol{\mu})(\boldsymbol{\mu}_m - \boldsymbol{\mu})^\top$$

$$\boldsymbol{\mu} = \frac{1}{n} \sum_{i}^{n} \boldsymbol{x}_i$$

n : Total # of samples



Fisher Discriminant Analysis (FDA)

- FDA criterion:
  - Increase between-class scatter
  - Reduce within-class scatter

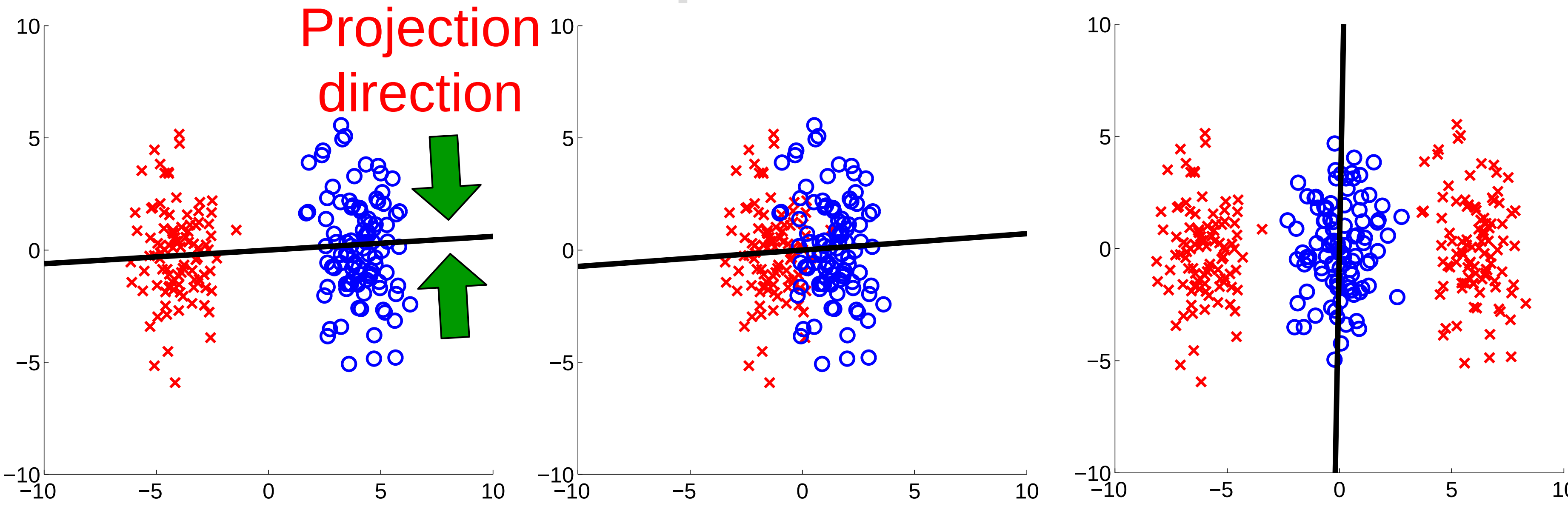
$$\max_{\boldsymbol{T}} \left[\mathrm{tr}(\boldsymbol{T}^\top \boldsymbol{S}^{(b)} \boldsymbol{T} (\boldsymbol{T}^\top \boldsymbol{S}^{(w)} \boldsymbol{T})^{-1}) \right]$$

- Solution: major eigenvectors of between/within-class scatter matrices

$$\boldsymbol{T}_{FDA} = (\boldsymbol{\varphi}_1 | \boldsymbol{\varphi}_2 | \cdots | \boldsymbol{\varphi}_r)$$

$$\boldsymbol{S}^{(b)} \boldsymbol{\varphi} = \lambda \boldsymbol{S}^{(w)} \boldsymbol{\varphi} \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

Examples of FDA



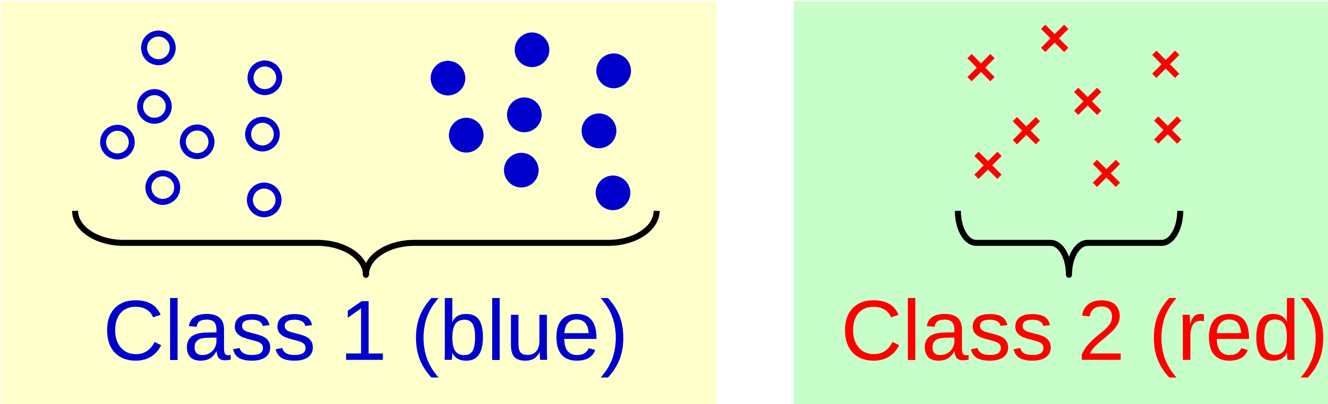
- Samples in different classes are separated from each other.
- But, FDA does not work well in the presence of within-class multi-modality.
- Since $\mathrm{rank}(\boldsymbol{S}^{(b)}) = c - 1$, at most $c - 1$ features can be extracted.

c : # of classes

Organization

1. Linear dimensionality reduction
2. Unsupervised methods:
   - Principal component analysis (PCA)
   - Locality preserving projection (LPP)
3. Supervised methods:
   - Fisher discriminant analysis (FDA)
   - Local Fisher discriminant analysis (LFDA)
4. Semi-supervised method:
   - Semi-supervised LFDA (SELF)
5. Conclusions

Within-class Multi-modality

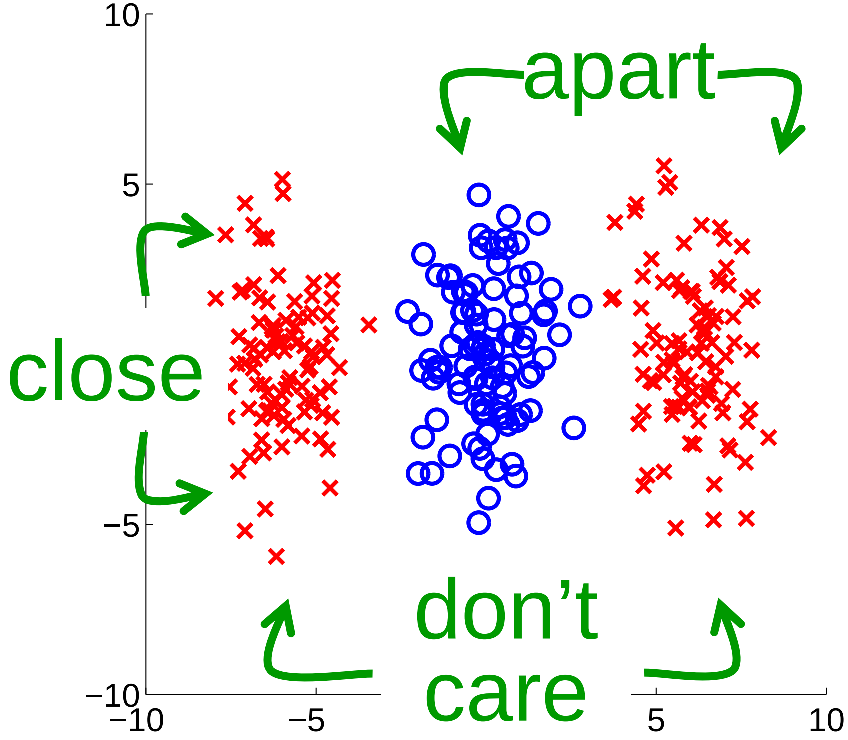


- **Medical diagnosis:**
  Hormone imbalance (too high/low) vs. normal
- **Digit recognition:**
  Even (0,2,4,6,8) vs. odd (1,3,5,7,9)
- **Multi-class classification:**
  one class vs. the others (i.e, one-versus-rest)

Local FDA (LFDA)

Sugiyama (JMLR2007)

- Basic idea:
  - Put nearby samples in the same class close
  - Don’t care far-apart samples in the same class
  - Put samples in different classes apart



LPP and FDA are combined!

Pairwise Expression of Scatter Matrices

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

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

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

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

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

Put samples in the same class close

Put samples in different classes apart

Local FDA (LFDA)

$\boldsymbol{A}_{i,j}$:Affinity matrix

- Local within-class scatter matrix:

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

$$\boldsymbol{W}_{i,j}^{(lw)} = \begin{cases} \boldsymbol{A}_{i,j}/n_{y_i} & (y_i = y_j) \\ 0 & (y_i \neq y_j) \end{cases}$$

- Local between-class scatter matrix:

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

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

- When $\boldsymbol{A}_{i,j} = 1$, $\boldsymbol{S}^{(lw)} = \boldsymbol{S}^{(l)}$ and $\boldsymbol{S}^{(lb)} = \boldsymbol{S}^{(b)}$.

Local FDA (LFDA)

- LFDA criterion:
  - Increase local between-class scatter
  - Reduce local within-class scatter

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

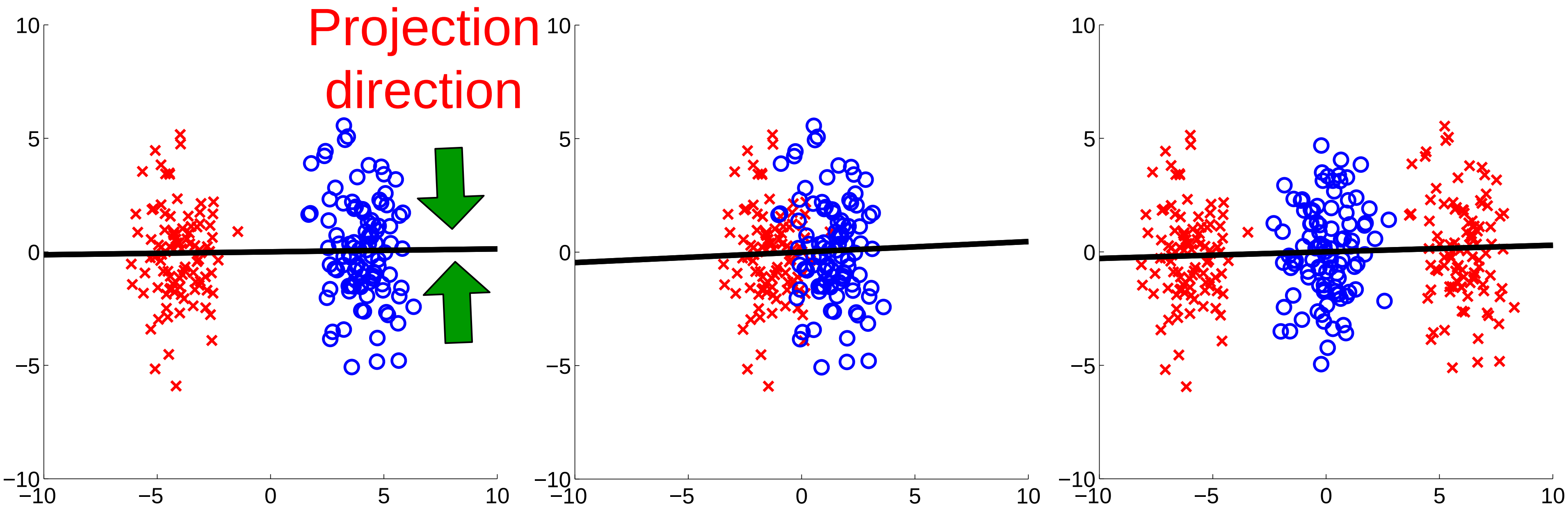
- Solution: major eigenvectors of local between/within-class scatter matrices

$$\boldsymbol{S}^{(lb)} \boldsymbol{\varphi} = \lambda \boldsymbol{S}^{(lw)} \boldsymbol{\varphi}$$

$$\boldsymbol{T}_{LFDA} = (\sqrt{\lambda_1}\boldsymbol{\varphi}_1 | \sqrt{\lambda_2}\boldsymbol{\varphi}_2 | \cdots | \sqrt{\lambda_r}\boldsymbol{\varphi}_r)$$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$$

Examples of LFDA



- Between-class separability is preserved.
- Within-class cluster structure is also preserved.
- Since $\mathrm{rank}(\boldsymbol{S}^{(lb)}) \gg c$ in general, no upper limit on the number of features to extract

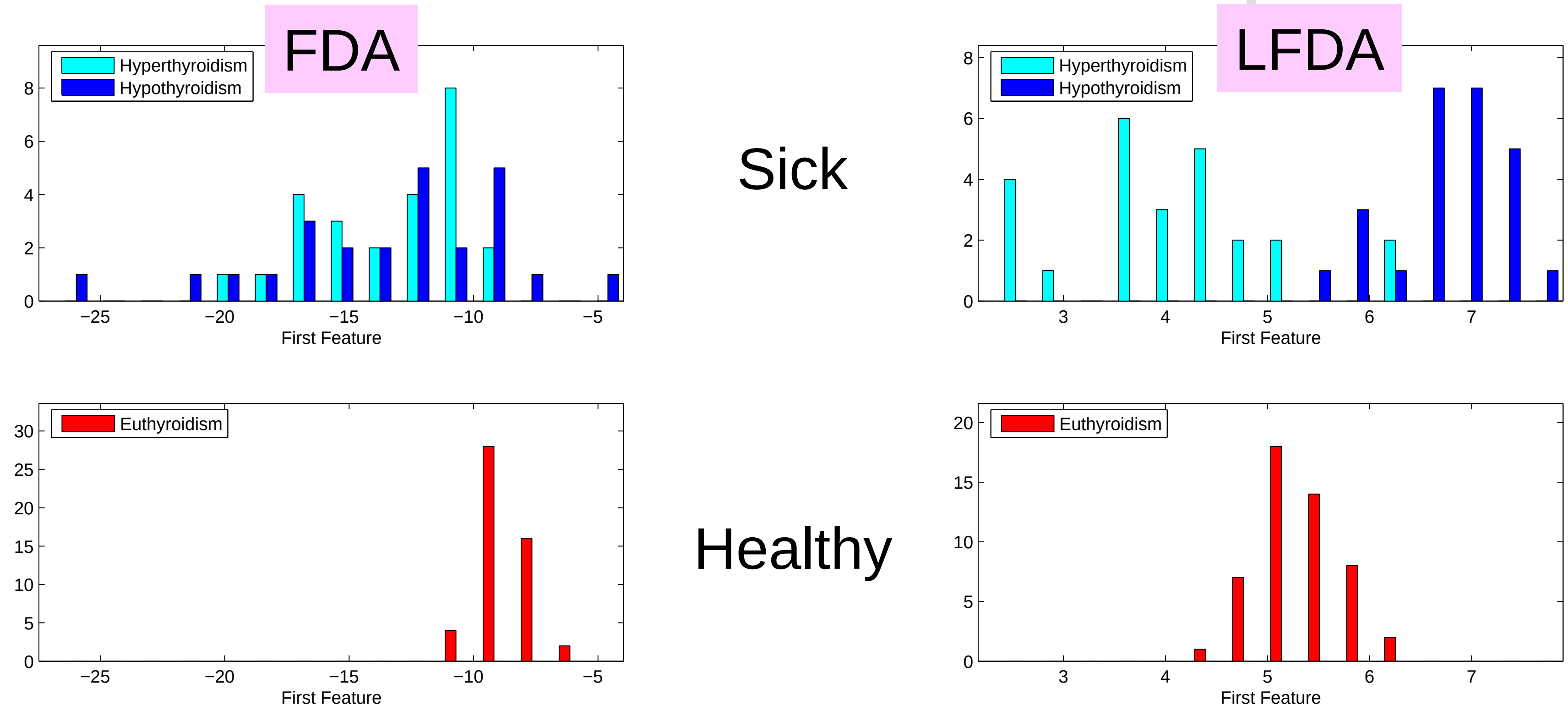
c : # of classes

Examples of LFDA (cont.)

- Analysis of thyroid disease data (5-dim):
  - T3-resin uptake test.
  - Total Serum thyroxin as measured by the isotopic displacement method.

  etc.
- Label: healthy or disease
- Two types of thyroid diseases:
  - Hyper-functioning: thyroid works too strongly
  - Hypo-functioning: thyroid works too weakly

Visualization in 1-dim Space



- Healthy/sick are nicely separated.
- Hyper-/hypo-functioning are mixed.

- Healthy/sick and hyper-/hypo-functioning are both nicely separated.
- LFDA feature has high (negative) correlation to thyroid's functioning level.

Classification Error by 1-NN

	LFDA	LDI	NCA	MCML	LPP	PCA
banana	13.7(0.8)	13.6(0.8)	14.3(2.0)	39.4(6.7)	13.6(0.8)	13.6(0.8)
b-cancer	34.7(4.3)	36.4(4.9)	34.9(5.0)	34.0(5.8)	33.5(5.4)	34.5(5.0)
diabetes	32.0(2.5)	30.8(1.9)	—	31.2(2.1)	31.5(2.5)	31.2(3.0)
f-solar	39.2(5.0)	39.3(4.8)	—	—	39.2(4.9)	39.1(5.1)
german	29.9(2.8)	30.7(2.4)	29.8(2.6)	31.3(2.4)	30.7(2.4)	30.2(2.4)
heart	21.9(3.7)	23.9(3.1)	23.0(4.3)	23.3(3.8)	23.3(3.8)	24.3(3.5)
image	3.2(0.8)	3.0(0.6)	—	4.7(0.8)	3.6(0.7)	3.4(0.5)
ringnorm	21.1(1.3)	17.5(1.0)	21.8(1.3)	22.0(1.2)	20.6(1.1)	21.6(1.4)
splice	16.9(0.9)	17.9(0.8)	—	17.3(0.9)	23.2(1.2)	22.6(1.3)
thyroid	4.6(2.6)	8.0(2.9)	4.5(2.2)	18.5(3.8)	4.2(2.9)	4.9(2.6)
titanic	33.1(11.9)	33.1(11.9)	33.0(11.9)	33.1(11.9)	33.0(11.9)	33.0(12.0)
twonorm	3.5(0.4)	4.1(0.6)	3.7(0.6)	3.5(0.4)	3.7(0.7)	3.6(0.6)
waveform	12.5(1.0)	20.7(2.5)	12.6(0.8)	17.9(1.5)	12.4(1.0)	12.7(1.2)
Comp. Time	1.00	1.11	97.23	70.61	1.04	0.91

- Mean and Std. of misclassification rate. Dim is chosen by cross-validation.
- Blue: Data with within-class multimodality, Red: Significantly better by 5% t-test
- LDI : Local discriminant information (Hastie & Tibshirani, IEEE-PAMI1996)
- NCA : Neighborhood component analysis (Goldberger et al. NIPS2004)
- MCML : Maximally collapsing metric learning (Globerson & Roweis, NIPS2005)

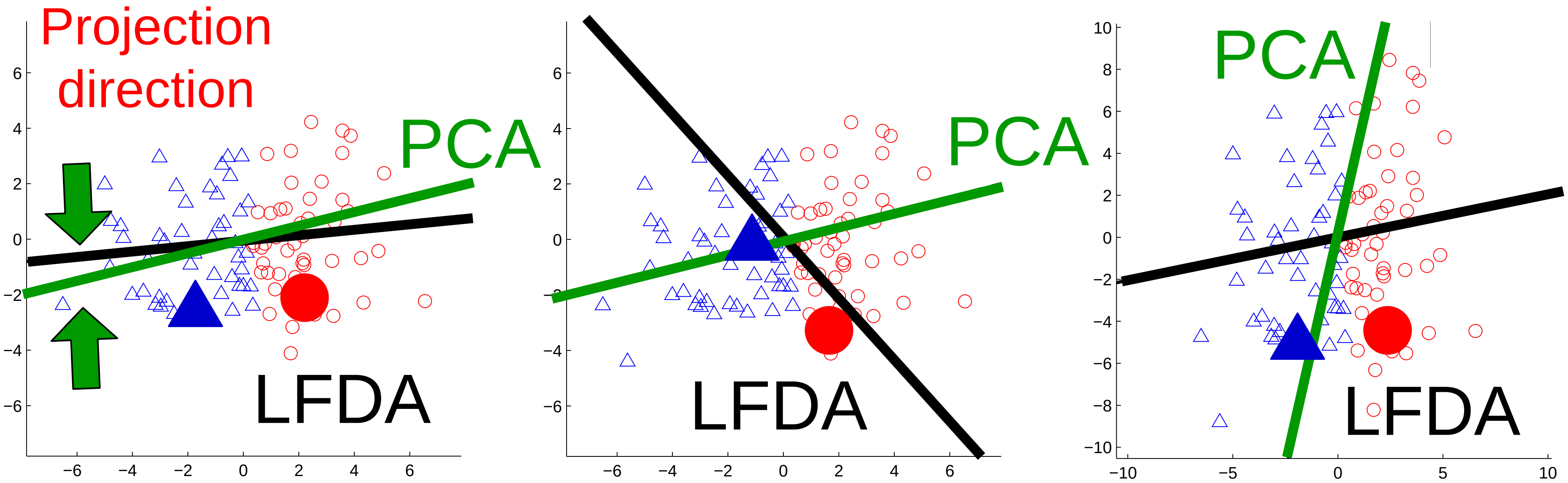
Organization

1. Linear dimensionality reduction
2. Unsupervised methods:
   - Principal component analysis (PCA)
   - Locality preserving projection (LPP)
3. Supervised methods:
   - Fisher discriminant analysis (FDA)
   - Local Fisher discriminant analysis (LFDA)
4. Semi-supervised method:
   - Semi-supervised LFDA (SELF)
5. Conclusions

Semi-supervised Dimensionality Reduction

- Semi-supervised learning:
  - Small number of labeled samples: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n'}$
  - Large number of unlabeled samples: $\{\boldsymbol{x}_i\}_{i=n'+1}^{n}$
- Supervised dimensionality reduction method tends to overfit labeled samples.
- We want to utilize unlabeled samples.

LFDA and PCA in Semi-supervised Setting



- LFDA tends to overfit.
- PCA does not use label information
- LFDA and PCA tend to be complementary.

Semi-supervised LFDA (SELF)

- Basic idea: Combine LFDA and PCA
- Key fact: Both involve similar eigenproblems.
  - LFDA: $\boldsymbol{S}^{(lb)}\boldsymbol{\varphi} = \lambda \boldsymbol{S}^{(lw)}\boldsymbol{\varphi}$
  - PCA: $\boldsymbol{S}^{(t)}\boldsymbol{\varphi} = \lambda \boldsymbol{\varphi}$
- SELF criteiron: weighted sum of LFDA & PCA

$$\boldsymbol{S}^{(rlb)}\boldsymbol{\varphi} = \lambda \boldsymbol{S}^{(rlw)}\boldsymbol{\varphi}$$

  - Regularized local between-class scatter matrix:

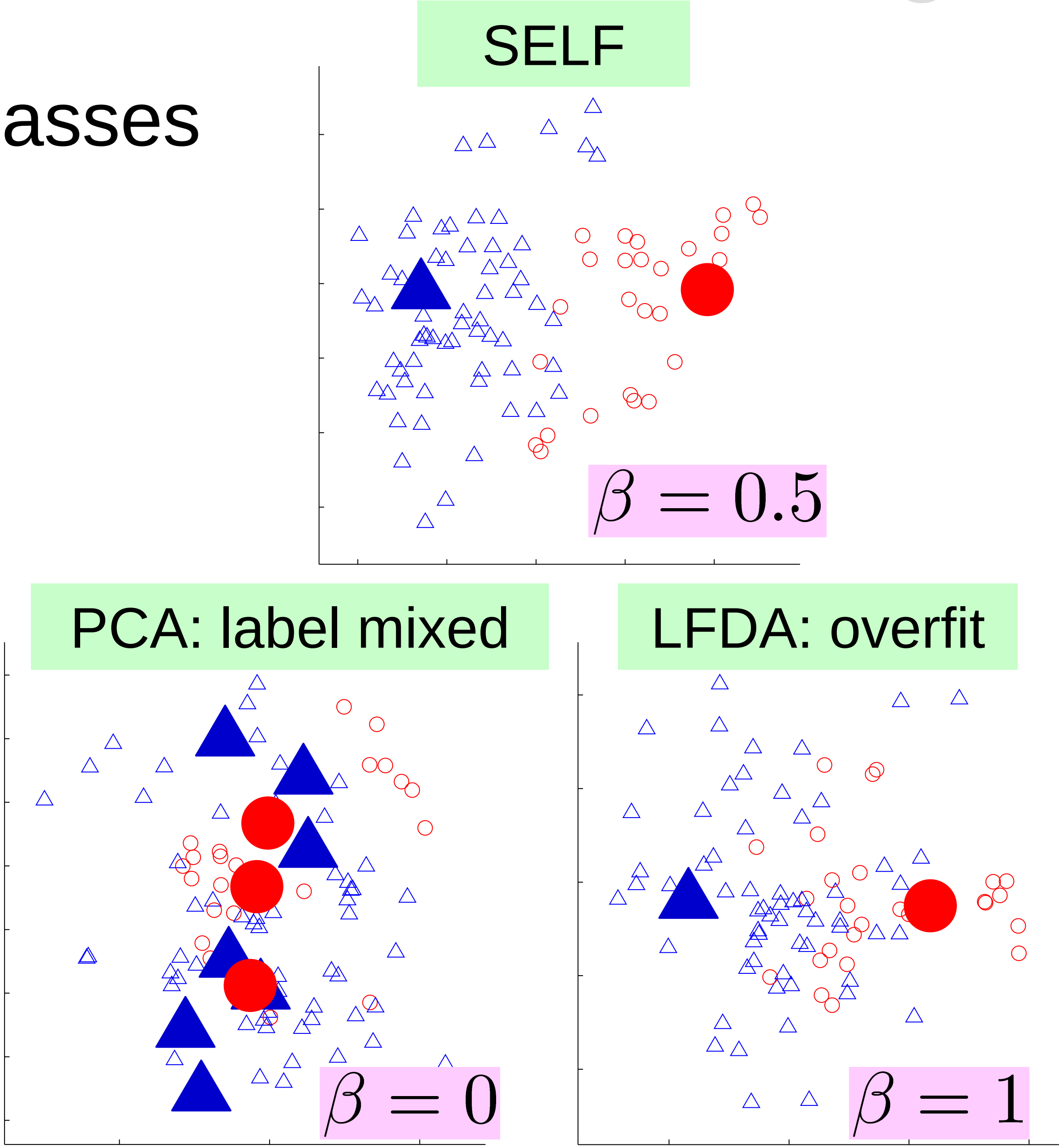
$$\boldsymbol{S}^{(rlb)} = (1-\beta)\boldsymbol{S}^{(lb)} + \beta \boldsymbol{S}^{(t)} \quad 0 \le \beta \le 1$$

  - Regularized local within-class scatter matrix:

$$\boldsymbol{S}^{(rlw)} = (1-\beta)\boldsymbol{S}^{(lw)} + \beta \boldsymbol{I}$$

Visualization of Olivetti Face Images

- With/without glasses



Classification Error

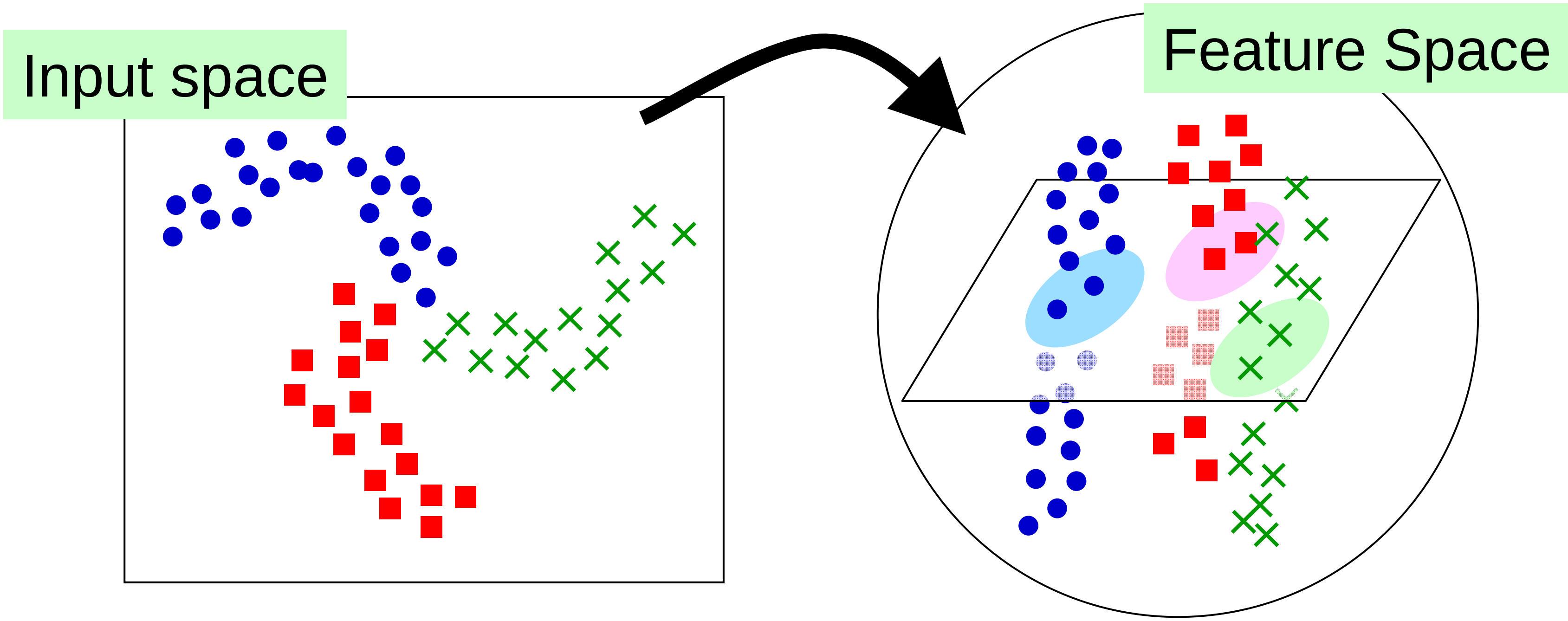
	LFDA	SELF ($\beta = 0.5$)	PCA	SELF (CV)
SSL1	14.9(1.8)	6.0(1.3)	6.2(1.1)	6.0(1.4)
SSL2	15.7(0.9)	9.6(1.1)	11.2(0.8)	10.3(2.4)
SSL3	21.1(3.9)	14.3(1.8)	15.5(1.0)	14.1(1.4)
SSL4	33.4(3.5)	36.6(2.4)	48.7(2.4)	33.4(3.7)
SSL5	27.5(2.3)	27.2(2.3)	31.0(1.9)	27.3(2.9)
SSL6	38.1(1.5)	35.4(2.4)	27.3(2.7)	27.0(2.7)
SSL7	29.4(2.4)	29.1(2.4)	29.3(1.6)	27.7(1.4)

- Data taken from semi-supervised learning book (Chapelle et al., 2006)
- Red: significantly better by 5% t-test

- LFDA and PCA are complementary.
- SELF($\beta = 0.5$) combines LFDA & PCA effectively.
- Optimizing β by cross-validation further improves the performance.

Non-linear Extension of SELF by Kernelization

- Standard kernel trick allows us to obtain a non-linear version of SELF.



Conclusions

- Semi-supervised LFDA (SELF) : Combination of LFDA and PCA
  - Between-class separability enhanced.
  - Within-class local structure preserved.
  - Global data structure preserved.
  - Closed-form solution exists.
  - Computationally fast and stable.
  - Non-linear extension of SELF by kernelization