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

We want to reduce dimensionality while keeping intrinsic information.
We focus on linear dimensionality reduction:

- High-dimensional samples: \( \{ \mathbf{x}_i \}_{i=1}^{n} \), \( \mathbf{x}_i \in \mathbb{R}^d \)
- Embedding matrix: \( \mathbf{T} \)
- Embedded samples: \( \{ \mathbf{z}_i \}_{i=1}^{n} \), \( \mathbf{z}_i \in \mathbb{R}^r \)

Goal: Find appropriate embedding matrix \( \mathbf{T} \)
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
  \[ \{ x_i \}_{i=1}^{n} \quad x_i \in \mathbb{R}^d \]

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:**
  \[ S^{(t)} = \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^\top \]
  \[ \mu = \frac{1}{n} \sum_{i=1}^{n} x_i \]

- **PCA criterion:** maximize scatter after embedding
  \[ \max_T \left[ \text{tr}(T^\top S^{(t)} T (T^\top T)^{-1}) \right] \]
  normalization

- **Solution:** major eigenvectors of \( S^{(t)} \)
  \[ T_{PCA} = (\varphi_1 | \varphi_2 | \cdots | \varphi_r) \]
  \[ S^{(t)} \varphi = \lambda \varphi \]
  \[ \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \]
Examples of PCA

\[ \mathbb{R}^2 \iff \mathbb{R}^1 \]

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

\[ A_{i,j} = \exp \left( - \frac{||x_i - x_j||^2}{\sigma^2} \right) \]

Choice of affinity is arbitrary.
Local Scaling Heuristic

Zelnik-Manor & Perona (NIPS2005)

Local scaling based affinity matrix:

\[ A_{i,j} = \exp \left( -\frac{\|x_i - x_j\|^2}{\gamma_i \gamma_j} \right) \]

\[ \gamma_i : \text{scaling around the sample } x_i \]

\[ \gamma_i = \|x_i - x_i^{(k)}\| \]

\[ x_i^{(k)} : \text{k-th nearest neighbor sample of } x_i \]

A heuristic choice is \( k = 7 \).

NOTE: We may cross-validate \( k \) in supervised cases if necessary.
Locality Preserving Projection (LPP)

- Locality matrix:

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

- LPP criterion: put samples with large affinity close

\[ \min_T \left[ \text{tr}(T^\top S^{(l)} T (T^\top T)^{-1}) \right] \]

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

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

\[ S^{(l)} \varphi = \lambda \varphi \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \]
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
    \[ \{(x_i, y_i)\}_{i=1}^{n} \]
    \[ y_i \in \{1, 2, \ldots, c\} \]

- Put samples in the same class close
- Put samples in different classes apart
Fisher Discriminant Analysis (FDA)

- **Within-class scatter matrix:**
  \[ S^{(w)} = \sum_{m=1}^{c} \sum_{i:y_i=m} (\mathbf{x}_i - \mu_m)(\mathbf{x}_i - \mu_m)^\top \]
  \[ \mu_m = \frac{1}{n_m} \sum_{i:y_i=m} \mathbf{x}_i \]
  \[ n_m : \# \text{ of samples in class } m \]

- **Between-class scatter matrix:**
  \[ S^{(b)} = \sum_{m=1}^{c} n_m(\mu_m - \mu)(\mu_m - \mu)^\top \]
  \[ \mu = \frac{1}{n} \sum_{i} \mathbf{x}_i \]
  \[ n : \text{Total # of samples} \]
Fisher Discriminant Analysis (FDA)

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

\[
\max_T \left[ \text{tr}(T^\top S^{(b)} T (T^\top S^{(w)} T)^{-1}) \right]
\]

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

\[
T_{FDA} = (\varphi_1 | \varphi_2 | \cdots | \varphi_r)
\]

\[
S^{(b)} \varphi = \lambda S^{(w)} \varphi \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d
\]
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 \( \text{rank}(S^{(b)}) = c - 1 \), at most \( c - 1 \) features can be extracted.

\( C : \# \text{ of classes} \)
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)
**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

\[ S^{(w)}(w) = \frac{1}{2} \sum_{i,j=1}^{n} W_{i,j}^{(w)} (x_i - x_j)(x_i - x_j)^\top \]

\[ W_{i,j}^{(w)} = \begin{cases} 
1/n_{y_i} & (y_i = y_j) \\
0 & (y_i \neq y_j) 
\end{cases} \]

\[ S^{(b)} = \frac{1}{2} \sum_{i,j=1}^{n} W_{i,j}^{(b)} (x_i - x_j)(x_i - x_j)^\top \]

\[ W_{i,j}^{(b)} = \begin{cases} 
1/n - 1/n_{y_i} & (y_i = y_j) \\
1/n & (y_i \neq y_j) 
\end{cases} \]

\[ \max_T \left[ \text{tr}(T^\top S^{(b)} T (T^\top S^{(w)} T)^{-1}) \right] \]

Put samples in the same class close
Put samples in different classes apart
Local FDA (LFDA)

- **Local** within-class scatter matrix:
  \[
  S^{(lw)} = \frac{1}{2} \sum_{i,j=1}^{n} W^{(lw)}_{i,j} (x_i - x_j)(x_i - x_j)^\top
  \]
  \[
  W^{(lw)}_{i,j} = \begin{cases} 
  A_{i,j}/n_{y_i} & (y_i = y_j) \\
  0 & (y_i \neq y_j)
  \end{cases}
  \]

- **Local** between-class scatter matrix:
  \[
  S^{(lb)} = \frac{1}{2} \sum_{i,j=1}^{n} W^{(lb)}_{i,j} (x_i - x_j)(x_i - x_j)^\top
  \]
  \[
  W^{(lb)}_{i,j} = \begin{cases} 
  A_{i,j}(1/n - 1/n_{y_i}) & (y_i = y_j) \\
  1/n & (y_i \neq y_j)
  \end{cases}
  \]

- When \( A_{i,j} = 1 \), \( S^{(lw)} = S^{(l)} \) and \( S^{(lb)} = S^{(b)} \).
Local FDA (LFDA)

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

\[
\max_T \left[ \text{tr}(T^\top S^{(lb)} T (T^\top S^{(lw)} T)^{-1}) \right]
\]

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

\[
S^{(lb)} \varphi = \lambda S^{(lw)} \varphi
\]

\[
T_{LFDA} = (\sqrt{\lambda_1} \varphi_1 | \sqrt{\lambda_2} \varphi_2 | \cdots | \sqrt{\lambda_r} \varphi_r)
\]

\[\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r\]
Examples of LFDA

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

$C : \# \text{ of classes}$
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

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

- 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-PAMI 1996)
- **NCA**: Neighborhood component analysis (Goldberger et al. NIPS 2004)
- **MCML**: Maximally collapsing metric learning (Globerson & Roweis, NIPS 2005)
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: \( \{(x_i, y_i)\}_{i=1}^{n'} \)
  - Large number of unlabeled samples: \( \{x_i\}_{i=n'+1}^{n} \)

- 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:
    \[ S^{(lb)} \varphi = \lambda S^{(lw)} \varphi \]
  - PCA:
    \[ S^{(t)} \varphi = \lambda \varphi \]
- **SELF criterion:** weighted sum of LFDA & PCA
  \[ S^{(rlb)} \varphi = \lambda S^{(rlw)} \varphi \]
  - Regularized local between-class scatter matrix:
    \[ S^{(rlb)} = (1 - \beta)S^{(lb)} + \beta S^{(t)} \quad 0 \leq \beta \leq 1 \]
  - Regularized local within-class scatter matrix:
    \[ S^{(rlw)} = (1 - \beta)S^{(lw)} + \beta I \]
Visualization of Olivetti Face Images

- With/without glasses

SELF

PCA: label mixed

LFDA: overfit

\( \beta = 0.5 \)

\( \beta = 0 \)

\( \beta = 1 \)
LFDA and PCA are complementary.

**SELF** ($\beta = 0.5$) combines LFDA & PCA effectively.

Optimizing $\beta$ by cross-validation further improves the performance.

---

Data taken from semi-supervised learning book *(Chapelle et al., 2006)*

Red: significantly better by 5% t-test

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