EEL-5840 / EEL-4930 Elements of Machine Intelligence

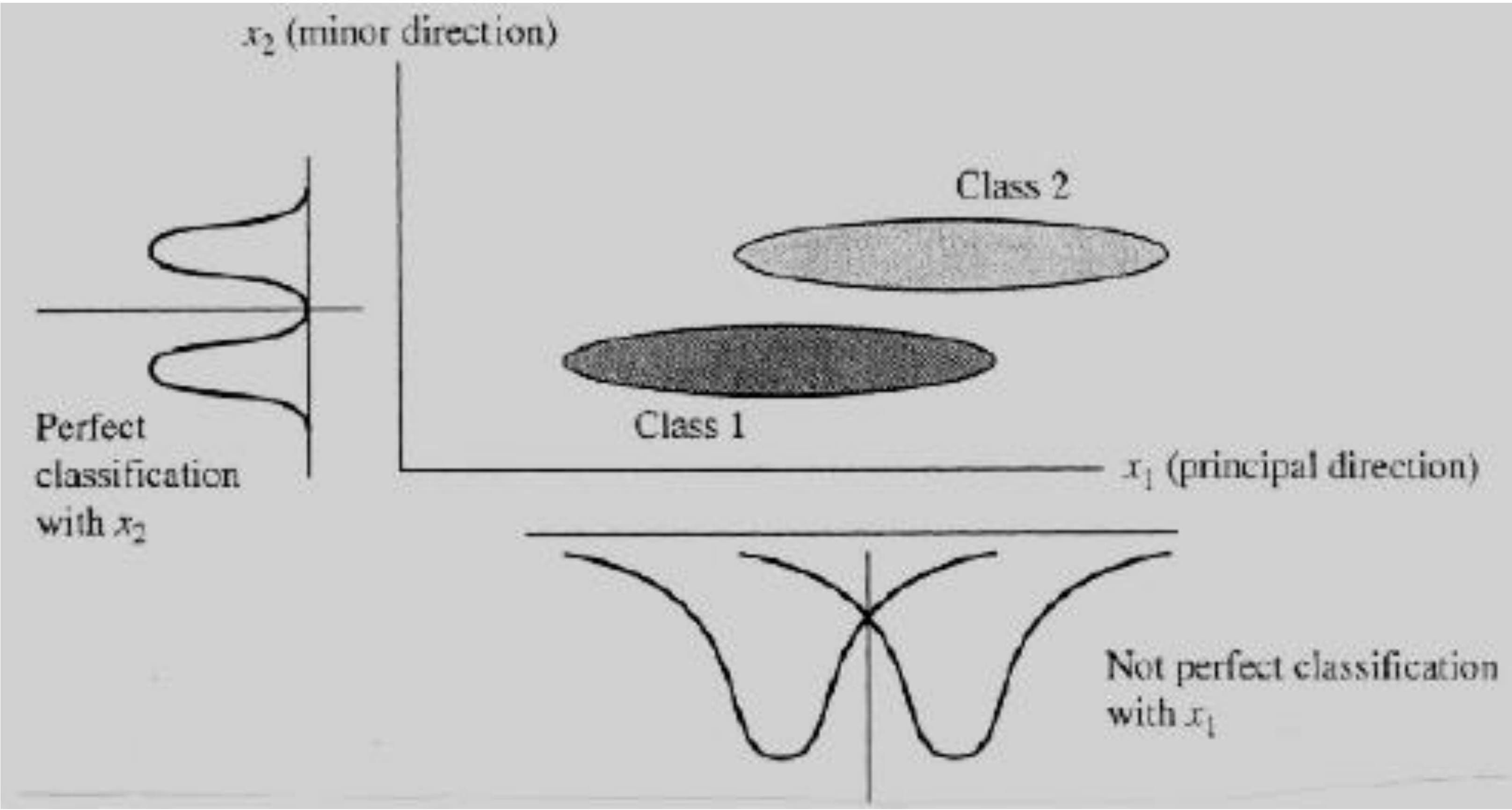
Linear Discriminant Analysis (LDA)

Principal Component Analysis (PCA)

- PCA and classification
  - PCA is **not** always an optimal dimensionality-reduction procedure for classification purposes.
- Multiple classes and PCA
  - Suppose there are *C* classes in the training data.
  - PCA is based on the sample covariance which characterizes the scatter of the entire data set, *irrespective of class-membership*.
  - The projection axes chosen by PCA might not provide good discrimination power.

Principal Component Analysis (PCA)

- PCA and classification (cont’d)



Linear Discriminant Analysis (LDA)

- What is the goal of LDA?
  - Perform dimensionality reduction "**while preserving as much of the class discriminatory information as possible**".
  - Seeks to find directions along which the classes are best separated.
  - Takes into consideration the scatter *within-classes* but also the scatter *between-classes*.
  - More capable of distinguishing image variation due to identity from variation due to other sources such as illumination and expression.

Linear Discriminant Analysis (LDA)

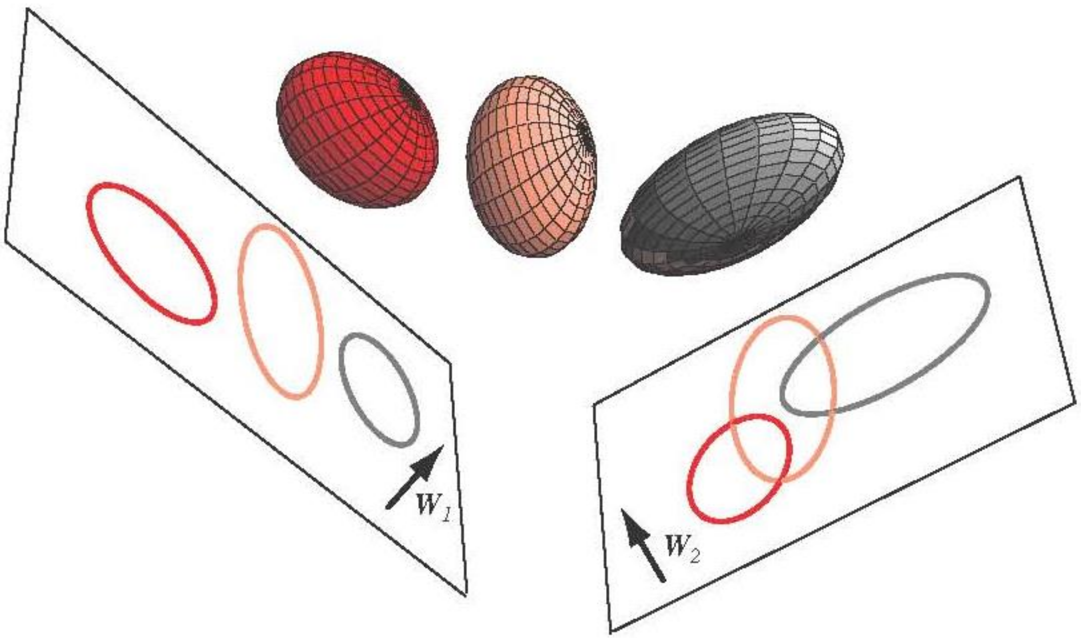


FIGURE 3.6. Three three-dimensional distributions are projected onto two-dimensional subspaces, described by a normal vectors $\mathbf{W}_1$ and $\mathbf{W}_2$. Informally, multiple discriminant methods seek the optimum such subspace, that is, the one with the greatest separation of the projected distributions for a given total within-scatter matrix, here as associated with $\mathbf{W}_1$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Linear Discriminant Analysis (LDA)

- Notation

- Suppose there are C classes
- Let $\boldsymbol{\mu}_i$ be the mean vector of class i, $i = 1, 2, .., C$
- Let M_i be the number of samples within class i, $i = 1, 2, .., C$,
- Let $M = \sum_{i=0}^{C} M_i$ be the total number of samples. and

<u>Within-class scatter matrix:</u>

$$S_w = \sum_{i=1}^{C} \sum_{j=1}^{M_i} (x_j - \boldsymbol{\mu}_i)(x_j - \boldsymbol{\mu}_i)^T$$

<u>Between-class scatter matrix:</u>

(S_b has at most rank C-1)

$$S_b = \sum_{i=1}^{C} (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$$

$$\boldsymbol{\mu} = 1/C \sum_{i=1}^{C} \boldsymbol{\mu}_i \quad \text{(mean of entire data set)}$$

Linear Discriminant Analysis (LDA)

- Methodology

projection matrix

$$\mathbf{y} = U^T \mathbf{x}$$

  - LDA computes a transformation that maximizes the between-class scatter while minimizing the within-class scatter:

$$\max \frac{|U^T S_b U|}{|U^T S_w U|} = \max \frac{|\tilde{S}_b|}{|\tilde{S}_w|}$$

$\tilde{S}_b, \tilde{S}_w$: scatter matrices of the projected data $\mathbf{y}$

Linear Discriminant Analysis (LDA)

- Linear transformation implied by LDA
  - The LDA solution is given by the eigenvectors of the *generalized eigenvector problem*:

$$S_B u_k = \lambda_k S_W u_k$$

  - The linear transformation is given by a matrix U whose columns are the eigenvectors of the above problem (i.e., called *Fisherfaces*).

$$\begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_K \end{bmatrix} = \begin{bmatrix} u_1^T \\ u_2^T \\ \dots \\ u_K^T \end{bmatrix} (x - \boldsymbol{\mu}) = U^T (x - \boldsymbol{\mu})$$

  - **Important:** Since S_b has at most rank C-1, the max number of eigenvectors with non-zero eigenvalues is C-1 (i.e., max dimensionality of sub-space is C-1)

Linear Discriminant Analysis (LDA)

- Does S_w^{-1} always exist?
  - If S_w is non-singular, we can obtain a conventional eigenvalue problem by writing:

$$S_w^{-1} S_B u_k = \lambda_k u_k$$

  - In practice, S_w is often singular since the data are image vectors with large dimensionality while the size of the data set is much smaller ($M << N$)

Linear Discriminant Analysis (LDA)

- Does S_w^{-1} always exist? – cont.
  - To alleviate this problem, we can use PCA first:

    1) PCA is first applied to the data set to reduce its dimensionality.

$$\begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_N \end{bmatrix} --> PCA --> \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_K \end{bmatrix}$$

    2) LDA is then applied to find the most discriminative directions:

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_K \end{bmatrix} --> LDA --> \begin{bmatrix} z_1 \\ z_2 \\ \cdots \\ z_{C-1} \end{bmatrix}$$

Linear Discriminant Analysis (LDA)

- Some terminology
  - Most Expressive Features (MEF): the features (projections) obtained using PCA.
  - Most Discriminating Features (MDF): the features (projections) obtained using LDA.

- Numerical problems
  - When computing the eigenvalues/eigenvectors of $S_w^{-1}S_B u_k = \lambda_k u_k$ numerically, the computations can be unstable since $S_w^{-1}S_B$ is not always symmetric.

Linear Discriminant Analysis (LDA)

- Factors unrelated to classification
  - MEF vectors show the tendency of PCA to capture major variations in the training set such as lighting direction.
  - MDF vectors discount those factors unrelated to classification.

Figure 2. A sample of MEF and MDF vectors treated as images. The MEF vectors show the tendency of the principal components to capture major variations in the training set, such as lighting direction. The MDF vectors show the ability of the MDFs to discount those factors unrelated to classification. The training images used to produce these vectors are courtesy of the Weizmann Institute.

Linear Discriminant Analysis (LDA)

- Case Study: PCA versus LDA
  - A. Martinez, A. Kak, "PCA versus LDA", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228-233, 2001.

- Is LDA always better than PCA?
  - There has been a tendency in the computer vision community to prefer LDA over PCA.
  - This is mainly because LDA deals directly with discrimination between classes while PCA does not pay attention to the underlying class structure.
  - Main results of this study:

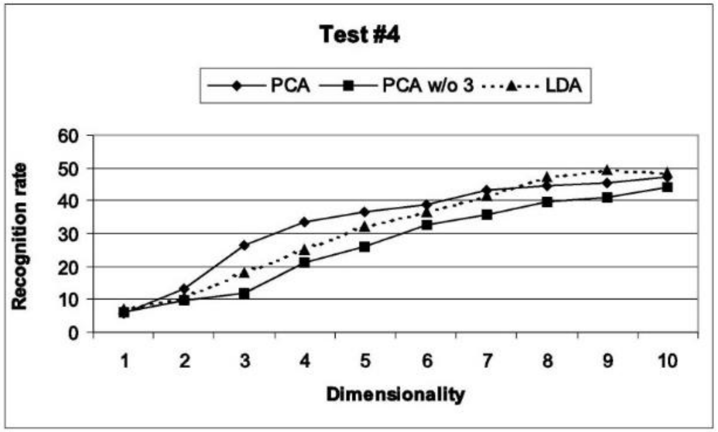
(1) When the training set is small, PCA can outperform LDA.

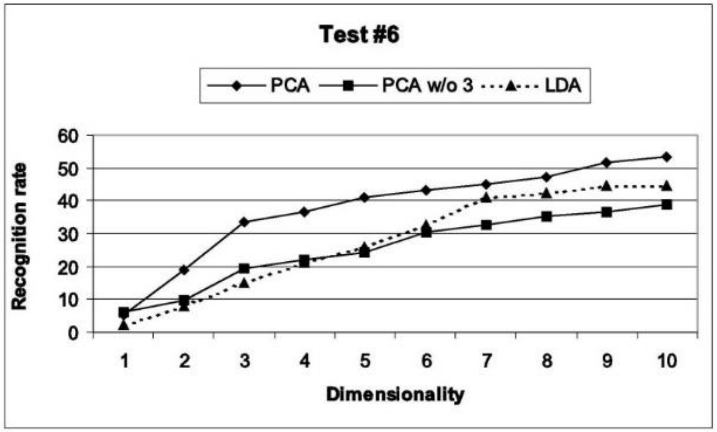
(2) When the number of samples is large and representative for each class, LDA outperforms PCA.

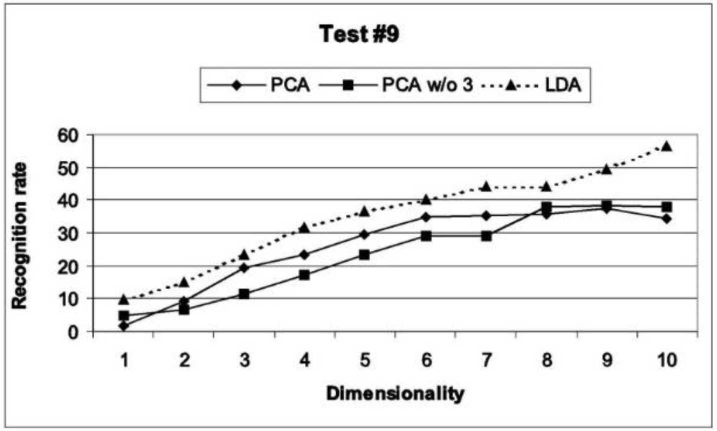
Linear Discriminant Analysis (LDA)

- Is LDA always better than PCA? – cont.

LDA is not always better when training set is small



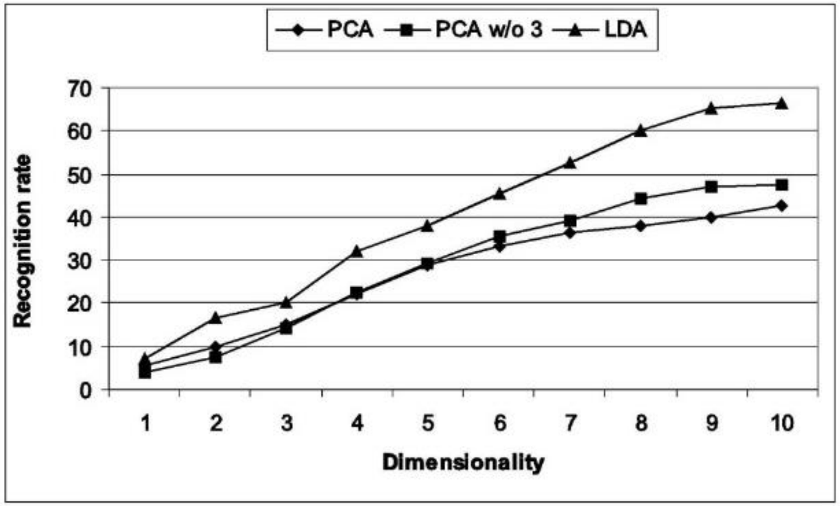


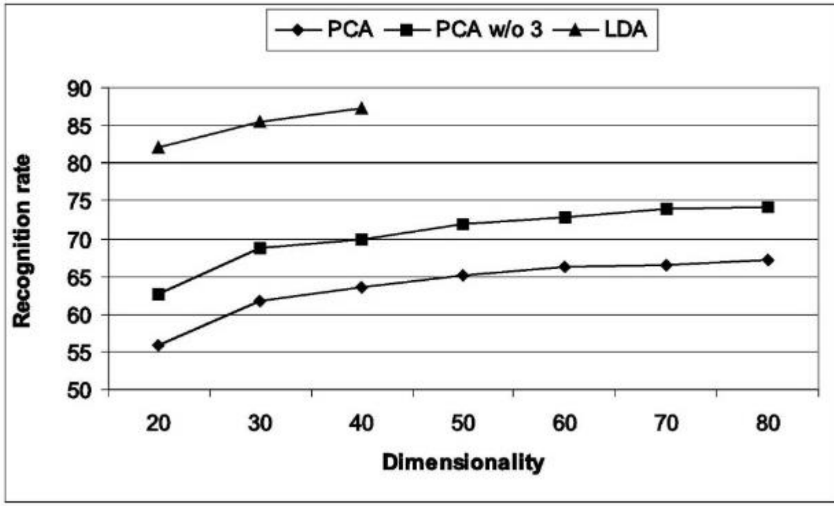


Linear Discriminant Analysis (LDA)

- Is LDA always better than PCA? – cont.

LDA outperforms PCA when training set is large





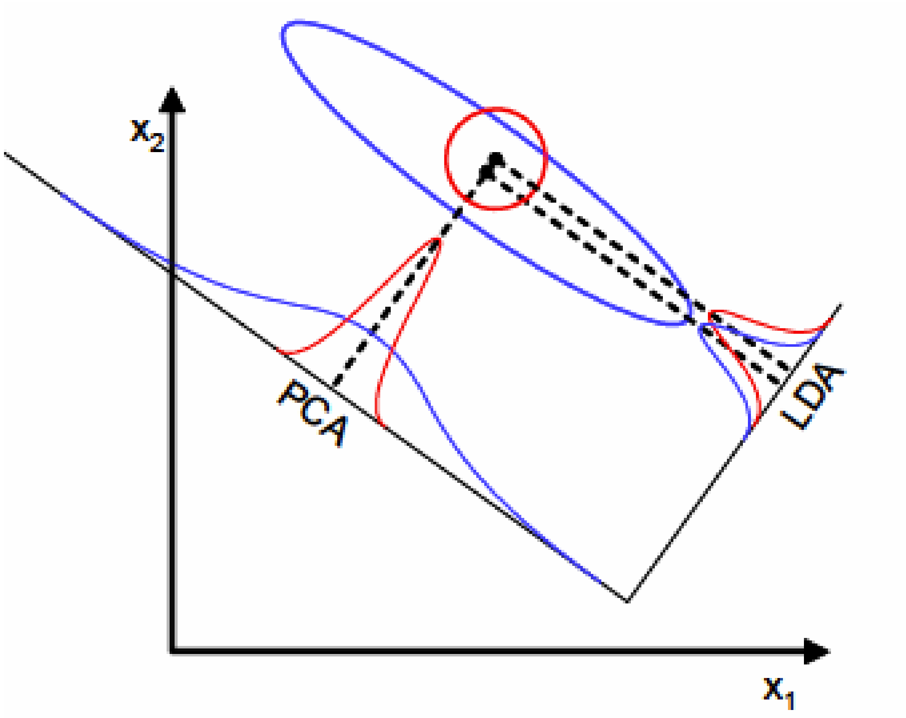
Limitations of LDA

- LDA produces at most C-1 feature projections

- LDA is a parametric method since it assumes unimodal Gaussians likelihoods
  - If class distributions are non-Gaussian, the LDA projects will not be able to preserve complex structure of data.

- LDA will fail when the discriminatory information is not in the mean but instead in the variance of the data.

Limitations of LDA (cont.)



Questions