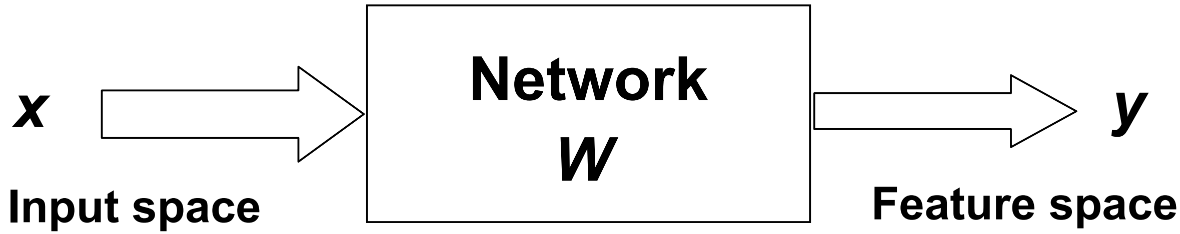
Discriminative Feature Extraction and Dimension Reduction

Berlin Chen, 2002

Introduction

- Goal: discovering significant patterns or features from the input data
  - Salient feature selection or dimensionality reduction



  - Compute an input-output mapping based on some desirable properties

Introduction

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Heteroscedastic Discriminant Analysis (HDA)

Introduction

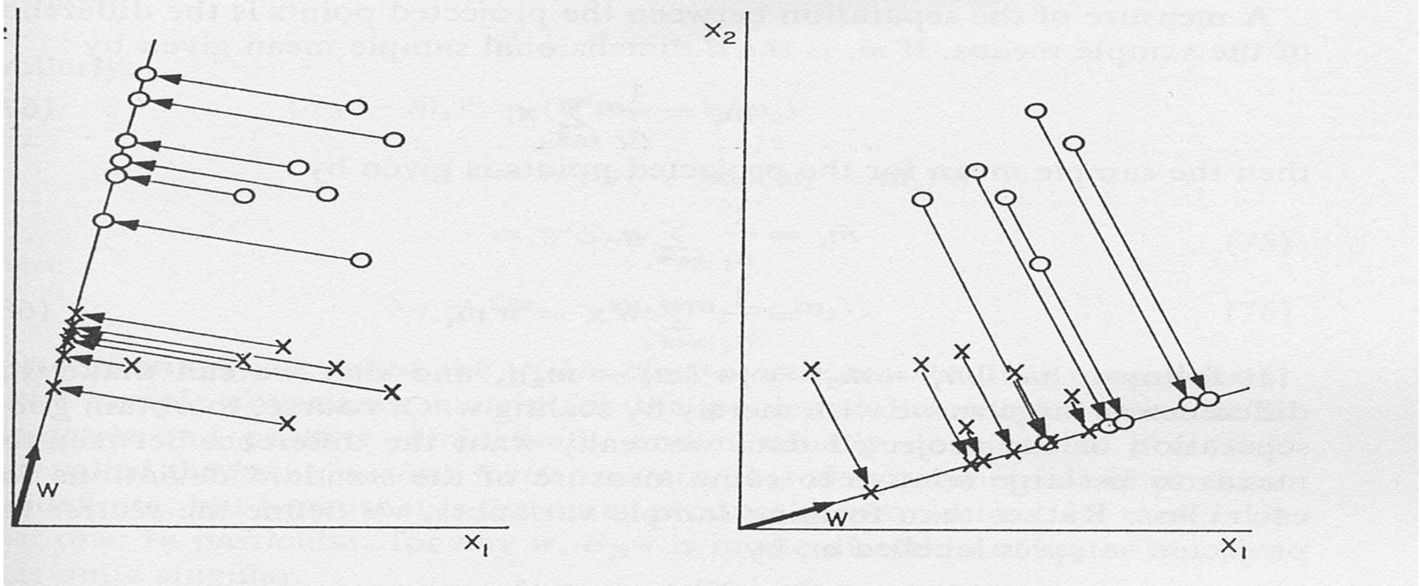


FIGURE 4.6. Projection of samples onto a line.

- Formulation
  - Model-free (nonparametric)
    - With/without prior information
  - Model-dependent (parametric)

Principle Component Analysis (PCA)

Pearson, 1901

- Known as Karhunen-Loẻve Transform (1947, 1963)
  - Or Hotelling Transform (1933)
- A standard technique commonly used for data reduction in statistical pattern recognition and signal processing
- A transform by which the data set can be represented by reduced number of effective features and still retain the most intrinsic information content
  - A small set of features to be found to represent the data samples accurately
- Also called “Subspace Decomposition”

Principle Component Analysis (PCA)

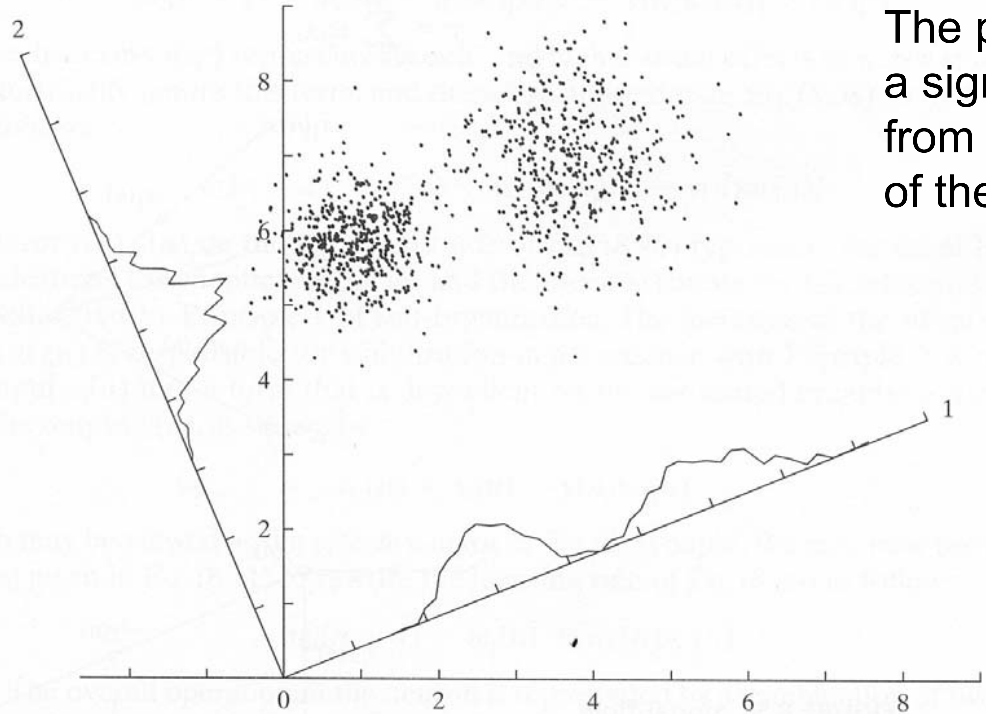


FIGURE 8.4 A cloud of data points is shown in two dimensions, and the density plots formed by projecting this cloud onto each of two axes, 1 and 2, are indicated. The projection onto axis 1 has maximum variance, and clearly shows the bimodal, or clustered character of the data.

The patterns show a significant difference from each other in one of the transformed axes

Principle Component Analysis (PCA)

- Suppose ***x*** is an *n*-dimensional zero mean random vector, $E_{\boldsymbol{x}}\{\boldsymbol{x}\} = \boldsymbol{0}$
  - If ***x*** is not zero mean, we can subtract the mean before processing the following analysis
  - ***x*** can be represented without error by the summation of *n* linearly independent vectors

$$\boldsymbol{x} = \sum_{i=i}^{n} \underbrace{y_i}\boldsymbol{\varphi}_i = \boldsymbol{\Phi y} \qquad \text{where} \quad \boldsymbol{y} = [y_1 \; . \; y_i \; . \; y_n]^T$$

$$\boldsymbol{\Phi} = \underbrace{[\boldsymbol{\varphi}_1 \; . \; \boldsymbol{\varphi}_i \; . \; \boldsymbol{\varphi}_n]}$$

The *i*-th component in the feature (mapped) space

The basis vectors

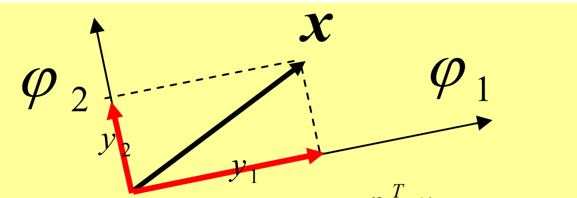
Principle Component Analysis (PCA)

– Further assume the column (basis) vectors of the matrix $\boldsymbol{\Phi}$ form an orthonormal set

$$\boldsymbol{\varphi}_i^T \boldsymbol{\varphi}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

- Such that y_i is equal to the projection of $\boldsymbol{x}$ on $\boldsymbol{\varphi}_i$

$$\forall_i \quad y_i = \boldsymbol{x}^T \varphi_i = \varphi_i^T \boldsymbol{x}$$



$$y_1 = \|\boldsymbol{x}\| \cos \theta_1 = \|\boldsymbol{x}\| \frac{\varphi_1^T \boldsymbol{x}}{\|\boldsymbol{x}\| \|\boldsymbol{\varphi}_1\|} = \varphi_1^T \boldsymbol{x}$$

, where $\|\boldsymbol{\varphi}_1\| = 1$

- y_i also has the following properties
  – Its mean is zero, too

$$E\{y_i\} = E\{\varphi_i^T \boldsymbol{x}\} = \varphi_i^T E\{\boldsymbol{x}\} = \varphi_i^T \boldsymbol{0} = 0$$

  – Its variance is

$$\sigma_i^2 = E\{y_i^2\} = E\{\varphi_i^T \boldsymbol{x}\boldsymbol{x}^T \varphi_i\} = \varphi_i^T E\{\boldsymbol{x}\boldsymbol{x}^T\}\varphi_i$$

$$= \varphi_i^T \boldsymbol{R} \varphi_i \qquad [\boldsymbol{R} \text{ is the (auto-)correlation matrix of } \boldsymbol{x}]$$

$$\boldsymbol{R} = E\{\boldsymbol{x}\boldsymbol{x}^T\} = \frac{1}{N}\sum_i \boldsymbol{x}_i \boldsymbol{x}_i^T$$

Principle Component Analysis (PCA)

– Further assume the column (basis) vectors of the matrix $\boldsymbol{\Phi}$ form an orthonormal set

- y_i also has the following properties
  - Its mean is zero, too

$$E\{y_i\} = E\{\varphi_i^T \boldsymbol{x}\} = \varphi_i^T E\{\boldsymbol{x}\} = \varphi_i^T \boldsymbol{0} = 0$$

  - Its variance is

$$\sigma_i^2 = E\{y_i^2\} = E\{\varphi_i^T \boldsymbol{x}\boldsymbol{x}^T \varphi_i\} = \varphi_i^T E\{\boldsymbol{x}\boldsymbol{x}^T\}\varphi_i$$

$$= \varphi_i^T \boldsymbol{R} \varphi_i \qquad [\boldsymbol{R} \text{ is the (auto-)correlation matrix of } \boldsymbol{x}]$$

$$\boldsymbol{R} = E\{\boldsymbol{x}\boldsymbol{x}^T\} = \frac{1}{N}\sum_i \boldsymbol{x}_i \boldsymbol{x}_i^T$$

- The correlation between two projections y_i and y_j is

$$E\{y_i y_j\} = E\{(\boldsymbol{\varphi}_i^T \boldsymbol{x})(\boldsymbol{\varphi}_j^T \boldsymbol{x})^T\} = E\{\boldsymbol{\varphi}_i^T \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{\varphi}_j\}$$

$$= \boldsymbol{\varphi}_i^T E\{\boldsymbol{x}\boldsymbol{x}^T\}\boldsymbol{\varphi}_j = \boldsymbol{\varphi}_i^T \boldsymbol{R} \boldsymbol{\varphi}_j$$

Principle Component Analysis (PCA)

- Minimum Mean-Squared Error Criterion
  - We want to choose only *m* of $\boldsymbol{\varphi}_i\text{'s}$ that we still can approximate $\boldsymbol{x}$ well in **mean-squared error criterion**

$$\boldsymbol{x} = \sum_{i=1}^{n} y_i \boldsymbol{\varphi}_i = \sum_{i=1}^{m} y_i \boldsymbol{\varphi}_i + \sum_{j=m+1}^{n} y_j \boldsymbol{\varphi}_j$$

$$\hat{\boldsymbol{x}}(m) = \sum_{i=1}^{m} y_i \boldsymbol{\varphi}_i$$

$$\bar{\varepsilon}(m) = E\left\{\left\|\hat{\boldsymbol{x}}(m) - \boldsymbol{x}\right\|^2\right\} = E\left\{\left(\sum_{j=m+1}^{n} y_j \boldsymbol{\varphi}_j^T\right)\left(\sum_{k=m+1}^{n} y_k \boldsymbol{\varphi}_k\right)\right\}$$

$$= E\left\{\sum_{j=m+1}^{n}\sum_{k=m+1}^{n} y_j y_k \boldsymbol{\varphi}_j^T \boldsymbol{\varphi}_k\right\}$$

$$= \sum_{j=m+1}^{n} E\left\{y_j^2\right\}$$

$$\because \boldsymbol{\varphi}_j^T \boldsymbol{\varphi}_k = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}$$

$$= \sum_{j=m+1}^{n} \sigma_j^2 = \sum_{j=m+1}^{n} \boldsymbol{\varphi}_j^T \boldsymbol{R} \boldsymbol{\varphi}_j$$

$$E\left\{y_j\right\} = 0$$

$$\sigma_j^2 = E\left\{y_j^2\right\} - \left(E\left\{y_j\right\}\right)^2 = E\left\{y_j^2\right\}$$

We should discard the bases where the projections have lower variances

Principle Component Analysis (PCA)

- Minimum Mean-Squared Error Criterion
  - If the orthonormal (basis) set φ_i's is selected to be the eigenvectors of the correlation matrix $\boldsymbol{R}$, associated with eigenvalues λ_i's
    - They will have the property that:

$$\boldsymbol{R}\boldsymbol{\varphi}_j = \lambda_j \boldsymbol{\varphi}_j$$

$\boldsymbol{R}$ is real and symmetric, therefore its eigenvectors form a orthonormal set

  - Such that the mean-squared error mentioned above will be

$$\bar{\varepsilon}(m) = \sum_{j=m+1}^{n} \sigma_j^2$$

$$= \sum_{j=m+1}^{n} \boldsymbol{\varphi}_j^T \boldsymbol{R} \boldsymbol{\varphi}_j = \sum_{j=m+1}^{n} \boldsymbol{\varphi}_j^T \lambda_j \boldsymbol{\varphi}_j = \sum_{j=m+1}^{n} \lambda_j$$

Principle Component Analysis (PCA)

- Minimum Mean-Squared Error Criterion
  - If the eigenvectors are retained associated with the *m* largest eigenvalues, the mean-squared error will be

$$\bar{\varepsilon}_{eigen}(m) = \sum_{j=m+1}^{n} \lambda_j \quad \left(\text{where } \lambda_1 \geq ... \geq \lambda_m \geq ... \geq \lambda_n\right)$$

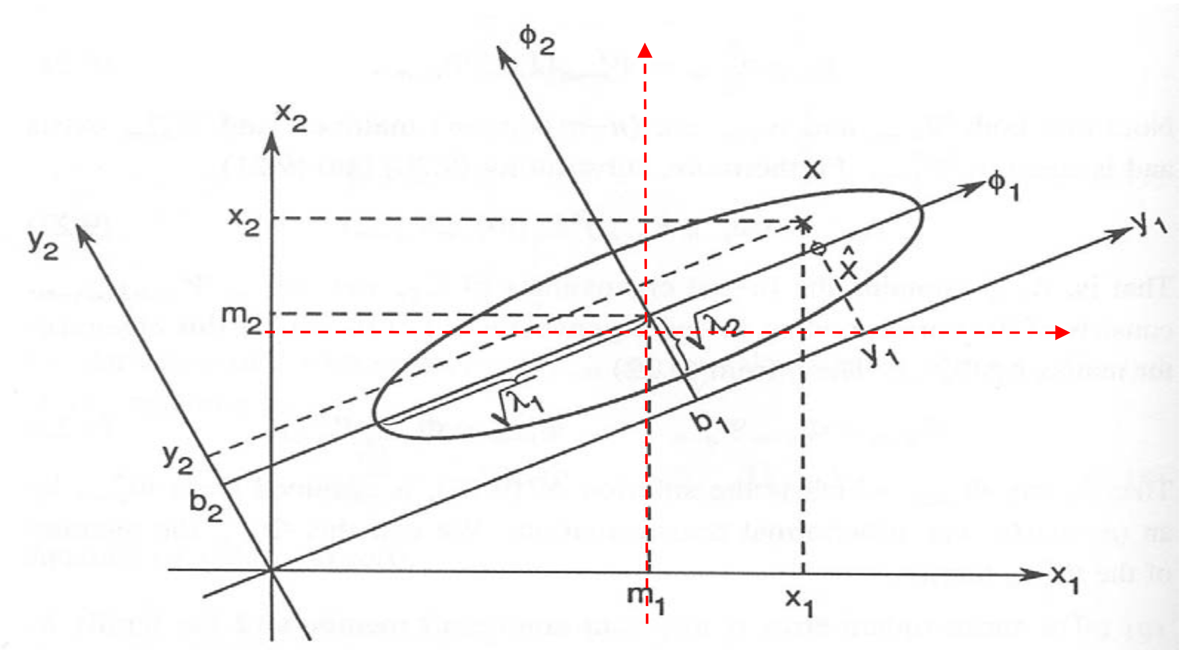
  - Any two projections y_i and y_j will be mutually uncorrelated

$$E\{y_i y_j\} = E\left\{\left(\boldsymbol{\varphi}_i^T \boldsymbol{x}\right)\left(\boldsymbol{\varphi}_j^T \boldsymbol{x}\right)^T\right\} = E\left\{\boldsymbol{\varphi}_i^T \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{\varphi}_j\right\}$$
$$= \boldsymbol{\varphi}_i^T E\left\{\boldsymbol{x}\boldsymbol{x}^T\right\}\boldsymbol{\varphi}_j = \boldsymbol{\varphi}_i^T \boldsymbol{R}\boldsymbol{\varphi}_j = \lambda_{\mathrm{j}}\boldsymbol{\varphi}_i^T \boldsymbol{\varphi}_j = 0$$

    - Good news for most statistical modeling
      - Gaussians and diagonal matrices

Principle Component Analysis (PCA)

- An two-dimensional example of Principle Component Analysis



Principle Component Analysis (PCA)

- Minimum Mean-Squared Error Criterion
  - It can be proved that $\bar{\varepsilon}_{eigen}(m)$ is the optimal solution under the mean-squared error criterion

To be minimized

constraints

$$\frac{\partial \varphi^T \boldsymbol{R} \varphi}{\partial \varphi} = 2\boldsymbol{R}\varphi$$

$$\text{Define: } J = \sum_{j=m+1}^{n} \boldsymbol{\varphi}_j^T \boldsymbol{R} \boldsymbol{\varphi}_j - \sum_{j=m+1}^{n} \sum_{k=m+1}^{n} \mu_{jk} \left(\boldsymbol{\varphi}_j^T \boldsymbol{\varphi}_k - \delta_{jk} \right)$$

Take derivation

$$\Rightarrow \forall_{m+1 \le j \le n} \frac{\partial J}{\partial \boldsymbol{\varphi}_j} = 2\boldsymbol{R}\boldsymbol{\varphi}_j - 2\sum_{k=m+1}^{n} \mu_{jk} \boldsymbol{\varphi}_k = 0 \quad \left(\text{where } \boldsymbol{\mu}_j^T = \left[\mu_{j\,m+1} \ldots \mu_{jn} \right] \right)$$

$$\Rightarrow \forall_{m+1 \le j \le n} \quad \boldsymbol{R}\boldsymbol{\varphi}_j = \boldsymbol{\Phi}_{n-m} \boldsymbol{\mu}_j \quad \left(\text{where } \boldsymbol{\Phi}_{n-m} = \left[\boldsymbol{\varphi}_{m+1} \ldots \boldsymbol{\varphi}_n \right] \right)$$

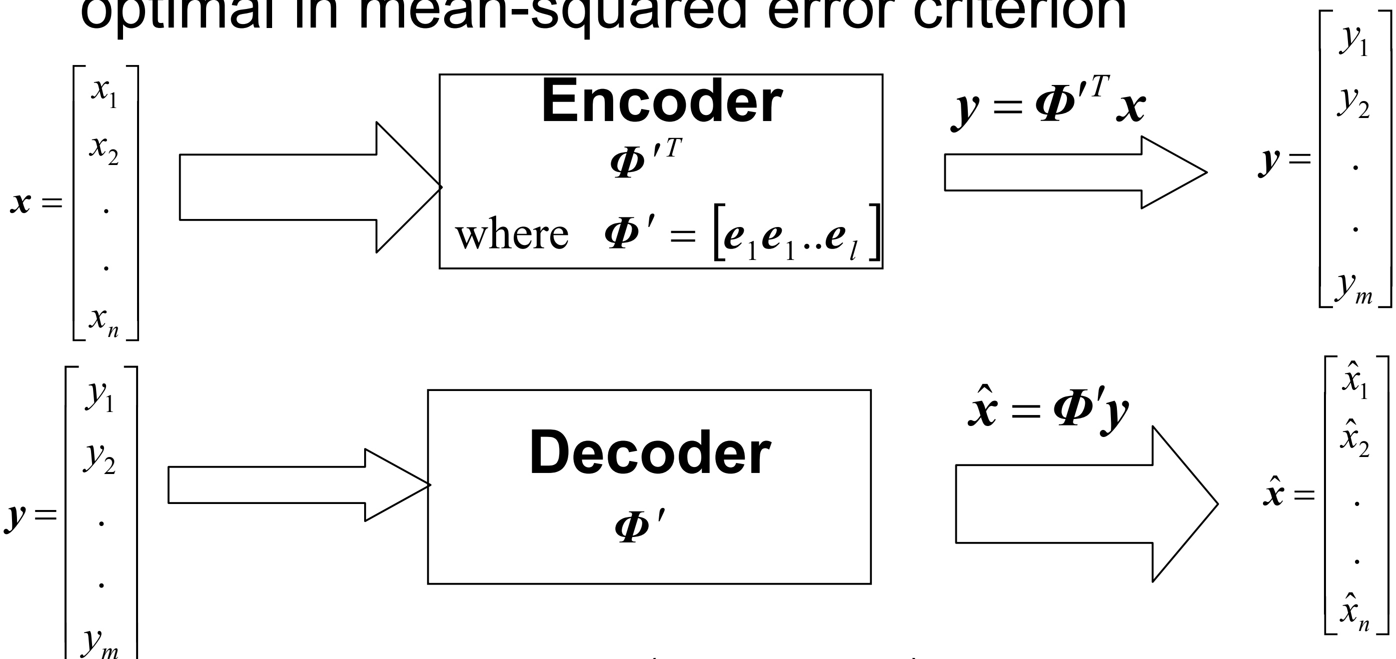
$$\Rightarrow \boldsymbol{R} \left[\boldsymbol{\varphi}_{m+1} \ldots \boldsymbol{\varphi}_n \right] = \boldsymbol{\Phi}_{n-m} \left[\boldsymbol{\mu}_{m+1} \ldots \boldsymbol{\mu}_n \right]$$

$$\Rightarrow \boldsymbol{R}\boldsymbol{\Phi}_{n-m} = \boldsymbol{\Phi}_{n-m} \boldsymbol{U}_{n-m} \quad \left(\text{where } \boldsymbol{U}_{n-m} = \left[\boldsymbol{\mu}_{m+1} \ldots \boldsymbol{\mu}_n \right] \right)$$

Have a particular solution if $\boldsymbol{U}_{n-m}$ is a diagonal matrix and its diagonal elements is the eigenvalues $\lambda_{m+1} \ldots \lambda_n$ of $\boldsymbol{R}$ and $\boldsymbol{\varphi}_{m+1} \ldots \boldsymbol{\varphi}_n$ is their corresponding eigenvectors

Principle Component Analysis (PCA)

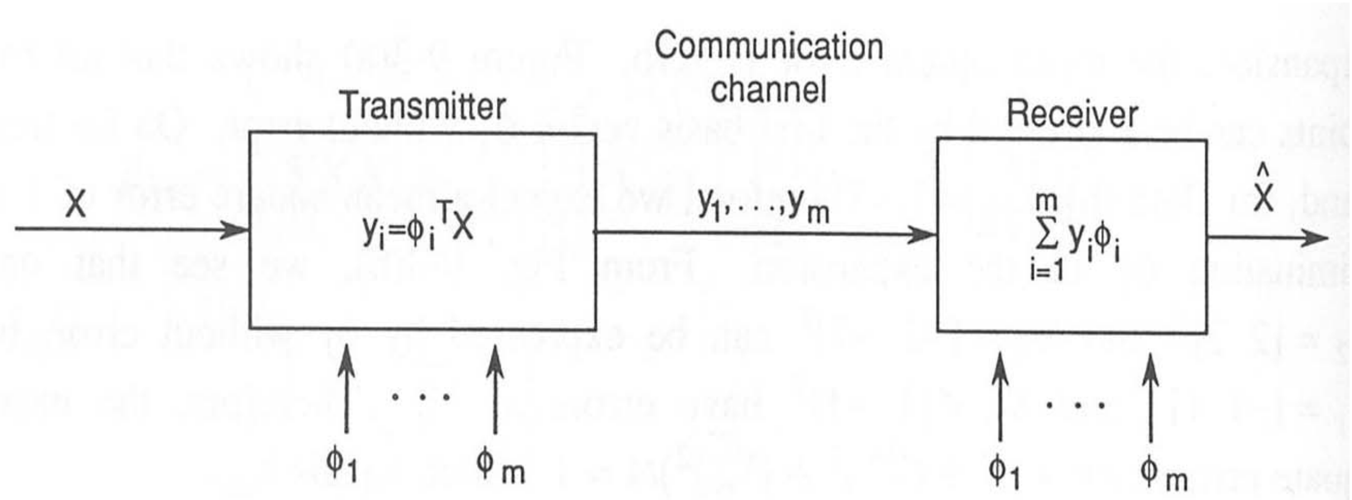
- Given an input vector ***x*** with dimensional *m*
  - Try to construct a linear transform Φ' (Φ' is an nxm matrix $m<n$) such that the truncation result, $\Phi'^T\mathbf{x}$, is optimal in mean-squared error criterion



$$\text{minimize } E_x\left(\left(\hat{x}\text{-}x\right)^T\left(\hat{x}\text{-}x\right)\right)$$

Principle Component Analysis (PCA)

- Data compression in communication



  - PCA is an optimal transform for signal representation and dimensional reduction, but not necessary for classification tasks, such as speech recognition
  - PCA needs no prior information (e.g. class distributions) of the sample patterns

Principle Component Analysis (PCA)

Hebbian-based Maximum Eigenfilter

$$y(n) = \sum_{i=1}^{m} w_i(n) x_i(n)$$

$$w_i(n+1) = \frac{w_i(n) + \eta y(n) x_i(n)}{\left(\sum_{j=1}^{m}\left[w_j(n) + \eta y(n) x_j(n)\right]^2\right)^{1/2}}$$

$$w_i(n+1) \approx w_i(n) + \eta y(n)\underbrace{\left[x_i(n) - y(n) w_i(n)\right]}_{x_i'(n)}$$

It had been proved that

$$\lim_{n \to \infty} \boldsymbol{w}(n) \to \boldsymbol{\varphi}_1 \text{ (the first principal component)}$$

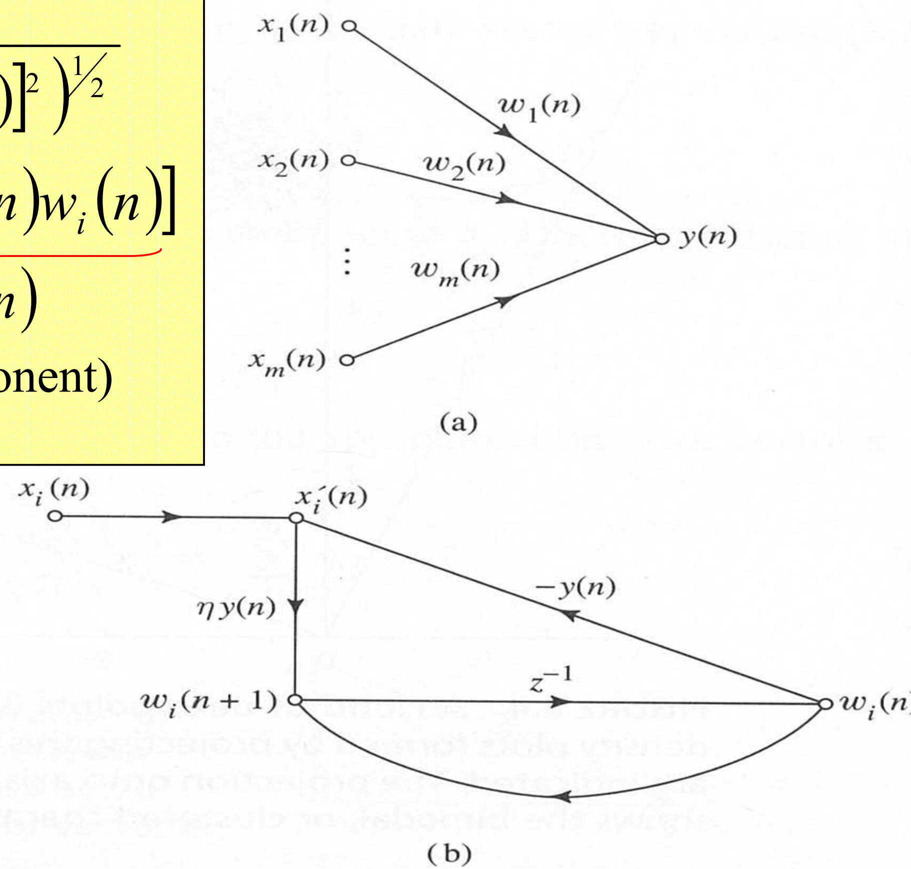
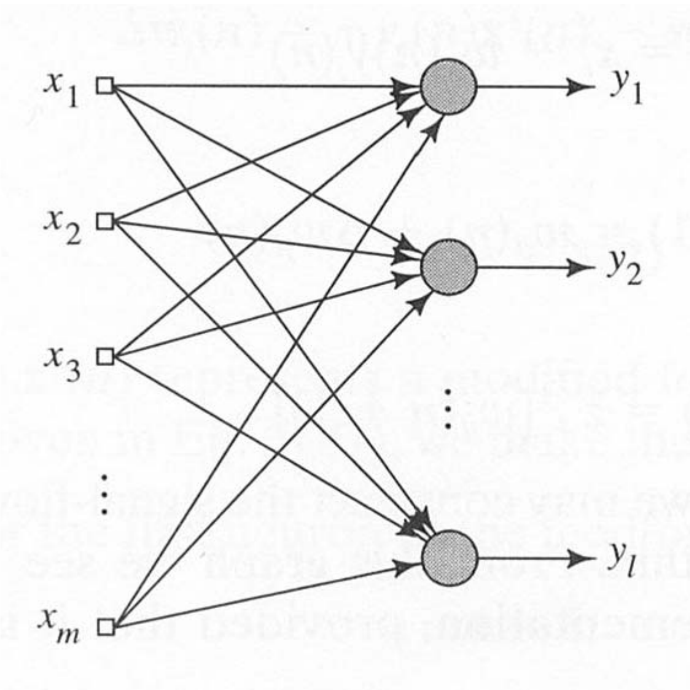


FIGURE 8.5 Signal-flow graph representation of maximum eigenfilter. (a) Graph of Eq. (8.36). (b) Graph of Eqs. (8.41) and (8.42).

Principle Component Analysis (PCA)
Hebbian-based Principal Analysis

- The Hebbian-based maximum eigenfilter can be expanded into a single layer feedforward network for principal component analysis (sanger, 1989)



$$y_j(n) = \sum_{i=1}^{m} w_{ji}(n) x_i(n), \quad j = 1, \ldots, J$$

$$\Delta w_{ji}(n) = \eta y_j(n) \Big[\underbrace{x_i(n) - \sum_{k=1}^{j} w_{ki}(n) y_k(n)}_{x_i'(n)} \Big]$$

$$w_{ji}(n+1) = w_{ji}(n) + \Delta w_{ji}(n)$$

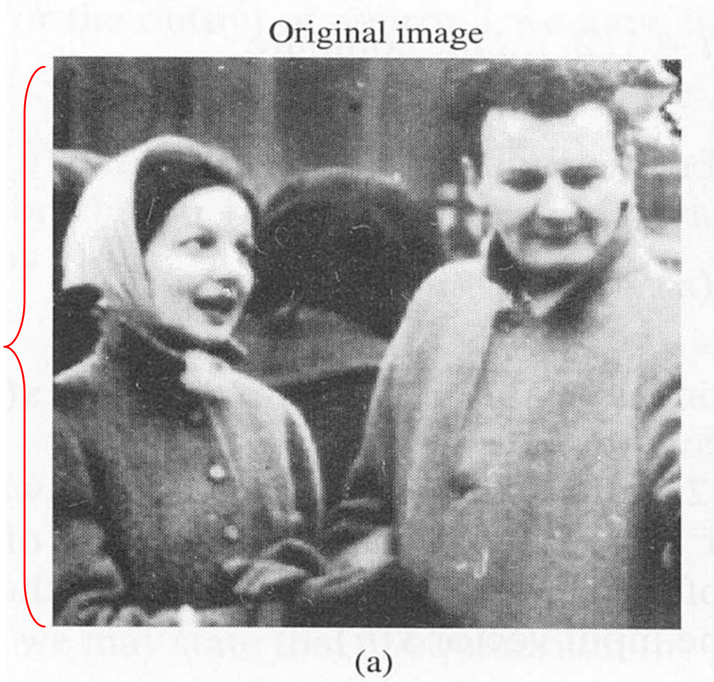
It had been proved that

$$\lim_{n \to \infty} \Delta \boldsymbol{w}_j(n) \to \boldsymbol{0}$$

$$\lim_{n \to \infty} \boldsymbol{w}_j(n) \to \boldsymbol{\varphi}_j \ \text{(the } j\text{-th principal component)}$$

Principle Component Analysis (PCA)
Hebbian-based Principal Analysis

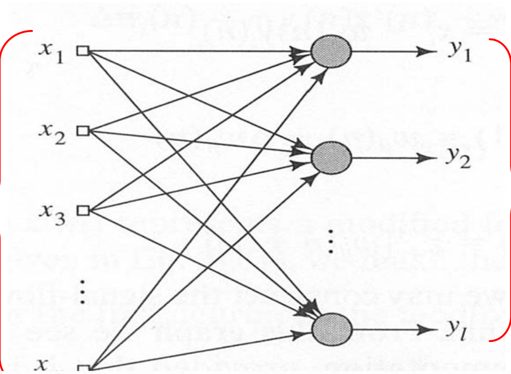
- Example: Image Coding



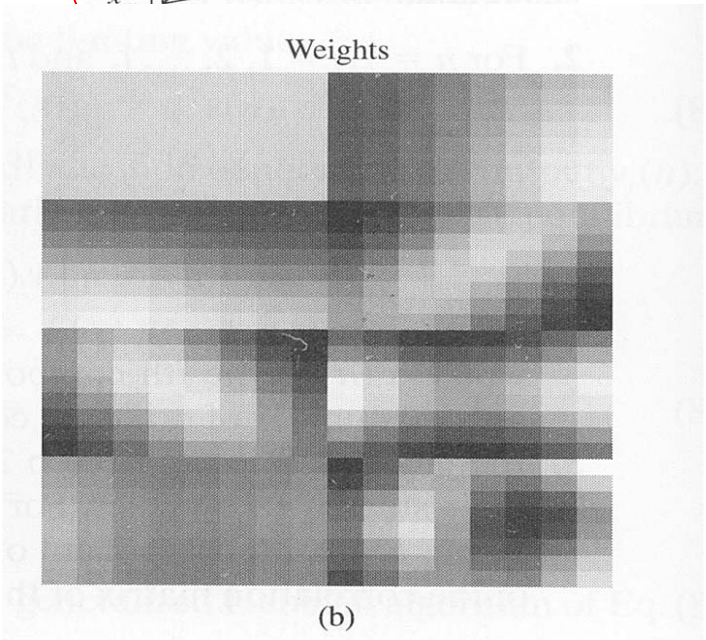
256

256

8x8
Non-overlapping
image block



8



Principle Component Analysis (PCA)
Hebbian-based Principal Analysis

- Example: Image Coding

Using first 8 components   15 to 1 compression

(c)   (d)

FIGURE 8.9 (a) An image of parents used in the image coding experiment. (b) 8 × 8 masks representing the synaptic weights learned by the GHA. (c) Reconstructed image of parents obtained using the dominant 8 principal components without quantization. (d) Reconstructed image of parents with 15 to 1 compression ratio using quantization.

Principle Component Analysis (PCA)
Adaptive Principal Components Extraction

- Both feedward and lateral connections are used

$$y_j(n) = \boldsymbol{w}_j^T(n)\boldsymbol{x}(n) + \boldsymbol{a}_j^T \boldsymbol{y}_{j-1}(n)$$

$$\boldsymbol{w}_j(n+1) = \boldsymbol{w}_j(n) + \eta\left[y_j(n)\boldsymbol{x}(n) - y_j^2(n)\boldsymbol{w}_j(n)\right]$$

$$\boldsymbol{a}_j(n+1) = \boldsymbol{a}_j(n) - \eta\left[y_j(n)\boldsymbol{y}_{j-1}(n) + y_j^2(n)\boldsymbol{a}_j(n)\right]$$

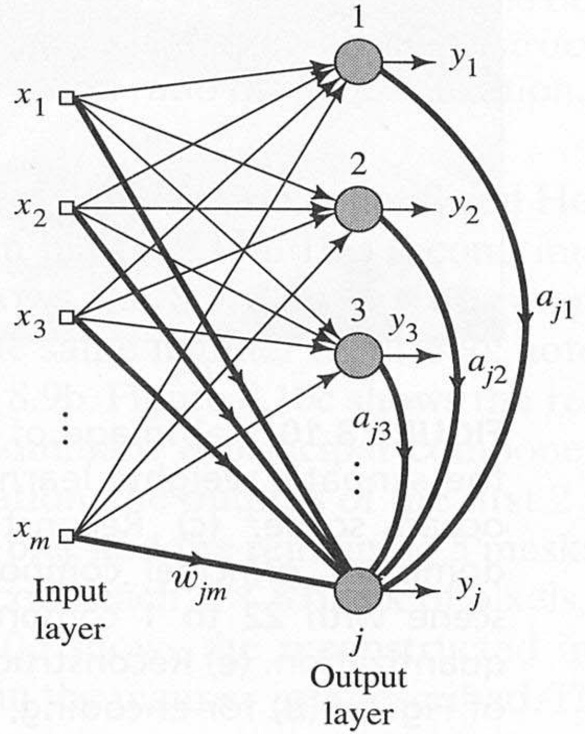


FIGURE 8.11 Network with feedforward and lateral connections for deriving the APEX algorithm.

Principle Component Analysis (PCA)
Eigenface and Eigenvoice

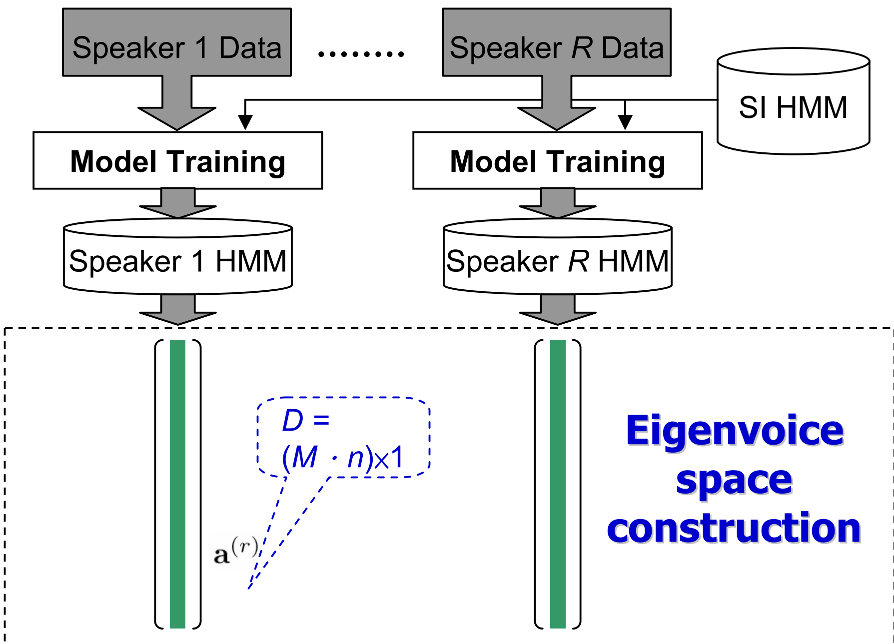
- Eigenface in face recognition (1990)
  - Consider an individual image to be a linear combination of a small number of face components or "eigenface" derived from a set of reference images
  - Steps
    - Convert each of the *L* reference images into a vector of floating point numbers representing light intensity in each pixel
    - Calculate the coverance/correlation matrix between these reference vectors
    - Apply Principal component Analysis (PCA) find the eigenvectors of the matrix: the eigenfaces
    - Besides, the vector obtained by averaging all images are called "eigenface 0". The other eigenface from "eigenface 1" onwards model the variations from this average face

Principle Component Analysis (PCA)
Eigenface and Eigenvoice

- Eigenface in face recognition (1990)
  - Steps
    - Then the faces are then represented as eigenvoice 0 plus a linear combination of the remain K ($K \leq L$) eigenfaces
  - The Eigenface approach persists the minimum mean-squared error criterion
  - Incidentally, the eigenfaces are not themselves usually plausible faces, only directions of variations between faces

Principle Component Analysis (PCA)
Eigenface and Eigenvoice

- Eigenvoice in speaker adaptation (PSTL, 2000)
  - Steps
    - Concatenating the regarded parameters for each speaker *r* to form a huge vector $\mathbf{a}^{(r)}$ (a supervectors)
    - SD model mean parameters (μ)



Let each new speaker S be represented by a point P in K-space

$$P = e(0) + w(1) * e1 + \cdots + w(K) * e(K).$$

Principal Component Analysis

Principle Component Analysis (PCA)
Eigenface and Eigenvoice

- Eigenvoice in speaker adaptation

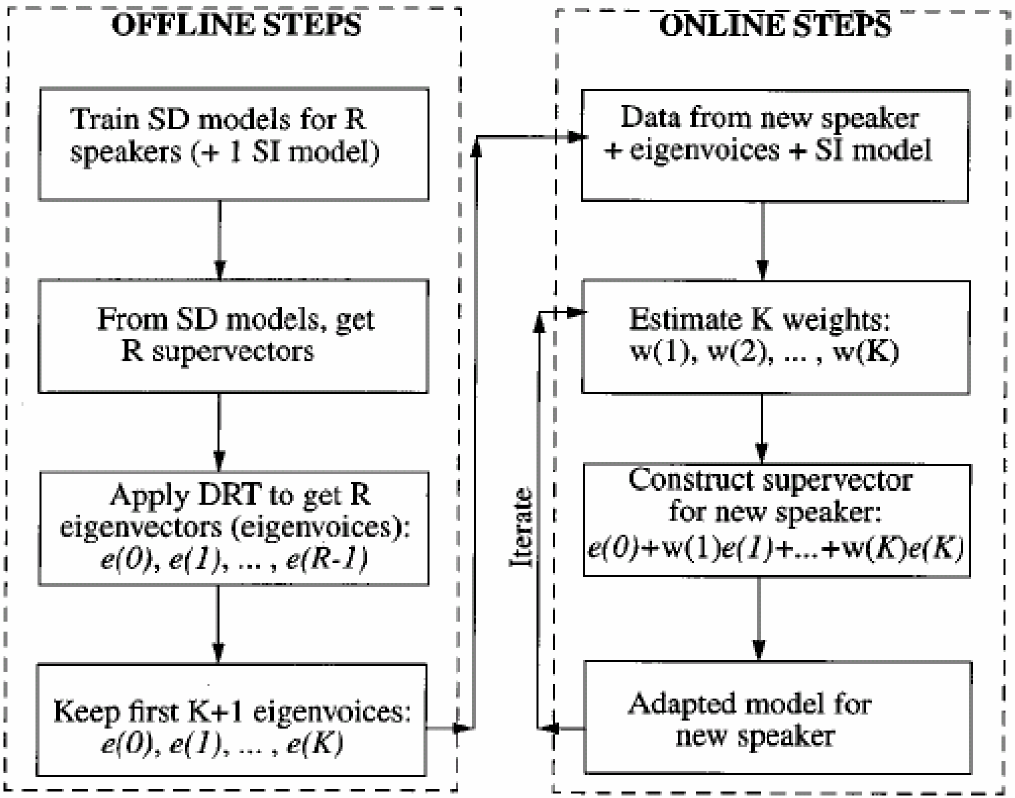
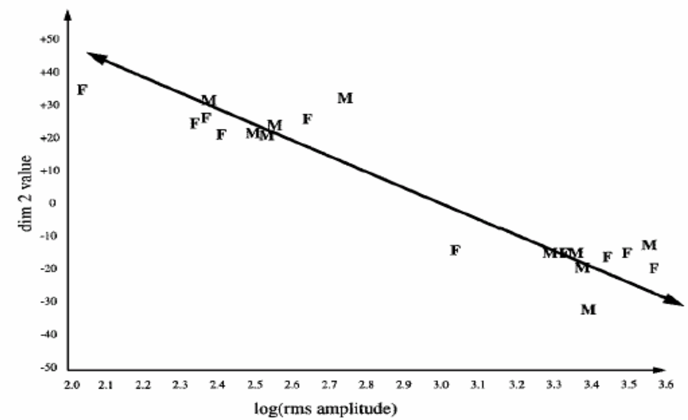


Fig. 1. Block diagram for eigenvoice speaker adaptation

Principle Component Analysis (PCA)
Eigenface and Eigenvoice

- Eigenvoice in speaker adaptation
  - Dimension 1 (eigenvoice 1):
    - Correlate with pitch or sex
  - Dimension 2 (eigenvoice 2):
    - Correlate with amplitude
  - Dimension 3 (eigenvoice 3):
    - Correlate with second-formant movement



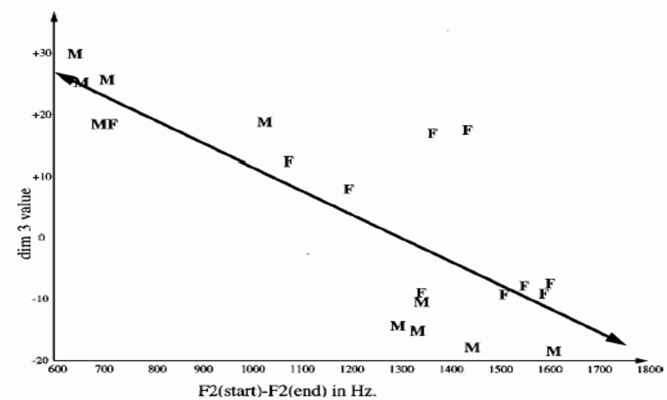


Fig. 4. Dimension 3 versus F2(start)–F2(end) for "U," extreme M and F in each speaker set

Linear Discriminant Analysis

- Given a set of sample vectors with labeled (class) information, try to find a linear transform ***W*** such that the ratio of **average between-class variation** over **average within-class variation** is maximal

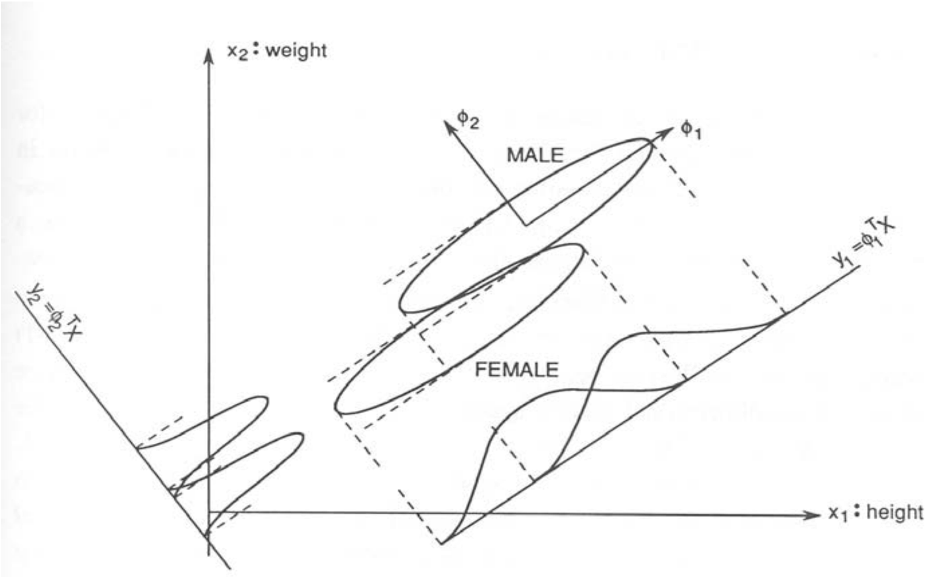


Fig. 10-1 An example of feature extraction for classification.

Linear Discriminant Analysis (LDA)

- Suppose there are *N* sample vectors $\boldsymbol{x}_i$ with dimensionality *n*, each of them is belongs to one of the *J* classes $g(\boldsymbol{x}_i) = j, \quad j \in \{1,2,\ldots, J\}, g(\cdot)$ is class index
  - The sample mean is: $\overline{\boldsymbol{x}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i$
  - The class sample means are: $\overline{\boldsymbol{x}}_j = \frac{1}{N_j} \sum_{g(\boldsymbol{x}_i)=j} \boldsymbol{x}_i$
  - The class sample covariances are: $\boldsymbol{\Sigma}_j = \frac{1}{N_j} \sum_{g(\boldsymbol{x}_i)=j} \left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_j\right)\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_j\right)^T$
  - The **average within-class variation** before transform

$$\boldsymbol{S}_w = \frac{1}{N} \sum_j N_j \boldsymbol{\Sigma}_j$$

  - The **average between-class variation** before transform

$$\boldsymbol{S}_b = \frac{1}{N} \sum_j N_j \left(\overline{\boldsymbol{x}}_j - \overline{\boldsymbol{x}}\right)\left(\overline{\boldsymbol{x}}_j - \overline{\boldsymbol{x}}\right)^T$$

Linear Discriminant Analysis (LDA)

- If the transform $\boldsymbol{W} = [\boldsymbol{w}_1 \boldsymbol{w}_2 \dots \boldsymbol{w}_m]$ is applied
  - The sample vectors will be $\boldsymbol{y}_i = \boldsymbol{W}^T \boldsymbol{x}_i$
  - The sample mean will be $\overline{\boldsymbol{y}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{W}^T \boldsymbol{x}_i = \boldsymbol{W}^T \left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \right) = \boldsymbol{W}^T \overline{\boldsymbol{x}}$
  - The class sample means will be $\overline{\boldsymbol{y}}_j = \frac{1}{N_j} \sum_{g(\boldsymbol{x}_i)=j} \boldsymbol{W}^T \boldsymbol{x}_i = \boldsymbol{W}^T \overline{\boldsymbol{x}}_j$
  - The **average within-class variation** will be

$$\widetilde{\boldsymbol{S}}_w = \frac{1}{N}\sum_j N_j \left\{ \frac{1}{N_j} \cdot \sum_{g(\boldsymbol{x}_i)=j} \left(\boldsymbol{W}^T \boldsymbol{x}_i - \frac{1}{N_j} \sum_{g(\boldsymbol{x}_i)=j} \left(\boldsymbol{W}^T \boldsymbol{x}_i \right) \right) \left(\boldsymbol{W}^T \boldsymbol{x}_i - \frac{1}{N_j} \sum_{g(\boldsymbol{x}_i)=j} \left(\boldsymbol{W}^T \boldsymbol{x}_i \right) \right)^T \right\}$$

$$= \boldsymbol{W}^T \left\{ \frac{1}{N} \sum_j N_j \boldsymbol{\Sigma}_j \right\} \boldsymbol{W}$$

$$= \boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W}$$

Linear Discriminant Analysis (LDA)

- If the transform $\boldsymbol{W} = [\boldsymbol{w}_1 \boldsymbol{w}_2 \ldots. \boldsymbol{w}_m]$ is applied
  - The **average between-class variation** will be

    $$\widetilde{\boldsymbol{S}}_b = \boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W}$$

  - Try to find optimal $\boldsymbol{W}$ such that the following criterion function is maximized

    $$J(\boldsymbol{W}) = \frac{\left|\widetilde{\boldsymbol{S}}_b\right|}{\left|\widetilde{\boldsymbol{S}}_w\right|} = \frac{\left|\boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W}\right|}{\left|\boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W}\right|}$$

    - A close form solution: the column vectors of an optimal matrix are the generalized eigenvectors corresponding to the largest eigenvalues in $\boldsymbol{W}$

      $$\boldsymbol{S}_b \boldsymbol{w}_i = \lambda_i \boldsymbol{S}_w \boldsymbol{w}_i$$

    - That is, $\boldsymbol{w}_i$'s are the eigenvectors corresponding to the largest eigenvalues of

      $$\boldsymbol{S}_w^{-1} \boldsymbol{S}_b \boldsymbol{w}_i = \lambda_i \boldsymbol{w}_i$$

Linear Discriminant Analysis (LDA)

- Proof: $\because \hat{\boldsymbol{W}} = \arg\max_{\hat{\boldsymbol{W}}} J(\boldsymbol{W}) = \arg\max_{\hat{\boldsymbol{W}}} \frac{\left|\tilde{\boldsymbol{S}}_b\right|}{\left|\tilde{\boldsymbol{S}}_w\right|} = \arg\max_{\hat{\boldsymbol{W}}} \frac{\left|\boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W}\right|}{\left|\boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W}\right|}$

Or, for each column vector $\boldsymbol{w}_i$ of $\boldsymbol{W}$, we want to find that :

The qradtic form has optimal solution : $\lambda_i = \frac{\boldsymbol{w}_i^T \boldsymbol{S}_b \boldsymbol{w}_i}{\boldsymbol{w}_i^T \boldsymbol{S}_w \boldsymbol{w}_i}$

$$\Rightarrow \frac{\partial \lambda_i}{\partial \boldsymbol{w}_i} = \frac{2\boldsymbol{S}_b \boldsymbol{w}_i \left(\boldsymbol{w}_i^T \boldsymbol{S}_w \boldsymbol{w}_i\right) - 2\boldsymbol{S}_w \boldsymbol{w}_i \left(\boldsymbol{w}_i^T \boldsymbol{S}_b \boldsymbol{w}_i\right)}{\left(\boldsymbol{w}_i^T \boldsymbol{S}_w \boldsymbol{w}_i\right)^2} = 0$$

$$\Rightarrow \frac{\boldsymbol{S}_b \boldsymbol{w}_i \left(\boldsymbol{w}_i^T \boldsymbol{S}_w \boldsymbol{w}_i\right)}{\left(\boldsymbol{w}_i^T \boldsymbol{S}_w \boldsymbol{w}_i\right)^2} - \frac{\boldsymbol{S}_w \boldsymbol{w}_i \left(\boldsymbol{w}_i^T \boldsymbol{S}_b \boldsymbol{w}_i\right)}{\left(\boldsymbol{w}_i^T \boldsymbol{S}_w \boldsymbol{w}_i\right)^2} = 0$$

$$\frac{\boldsymbol{S}_b \boldsymbol{w}_i}{\boldsymbol{w}_i^T \boldsymbol{S}_w \boldsymbol{w}_i} - \frac{\boldsymbol{S}_w \boldsymbol{w}_i}{\boldsymbol{w}_i^T \boldsymbol{S}_w \boldsymbol{w}_i} \lambda_i = 0 \quad \left(\because \lambda_i = \frac{\boldsymbol{w}_i^T \boldsymbol{S}_b \boldsymbol{w}_i}{\boldsymbol{w}_i^T \boldsymbol{S}_w \boldsymbol{w}_i} \right)$$

$$\Rightarrow \boldsymbol{S}_b \boldsymbol{w}_i - \lambda_i \boldsymbol{S}_w \boldsymbol{w}_i = 0 \Rightarrow \boldsymbol{S}_b \boldsymbol{w}_i = \lambda_i \boldsymbol{S}_w \boldsymbol{w}_i$$

$$\Rightarrow \boldsymbol{S}_w^{-1} \boldsymbol{S}_b \boldsymbol{w}_i = \lambda_i \boldsymbol{w}_i$$

Heteroscedastic Discriminant Analysis (HDA)

IBM, 2000

- Heteroscedastic : A set of statistical distributions having different variances
- LDA does not consider individual class covariances and may therefore generate suboptimal results
  - Modified the LDA objective function

$$H(\boldsymbol{W}) = \prod_{j=1}^{J} \left(\frac{\left| \boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W} \right|}{\left| \boldsymbol{W}^T \boldsymbol{\Sigma}_j \boldsymbol{W} \right|} \right)^{Nj} = \frac{\left| \boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W} \right|}{\prod_{j=1}^{J} \left| \boldsymbol{W}^T \boldsymbol{\Sigma}_j \boldsymbol{W} \right|^{Nj}}$$

  - Take the log and rearrange terms

$$\log H(\boldsymbol{W}) = -\left(\sum_{j=1}^{J} N_j \log \left| \boldsymbol{W}^T \boldsymbol{\Sigma}_j \boldsymbol{W} \right| \right) + N \log \left| \boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W} \right|$$

  - However the dimensions of the HDA projection can often be highly correlated
    - An other transform can be further composed into HDA

Heteroscedastic Discriminant Analysis (HDA)

- The difference in the projections obtained from LDA and HDA for 2-class case

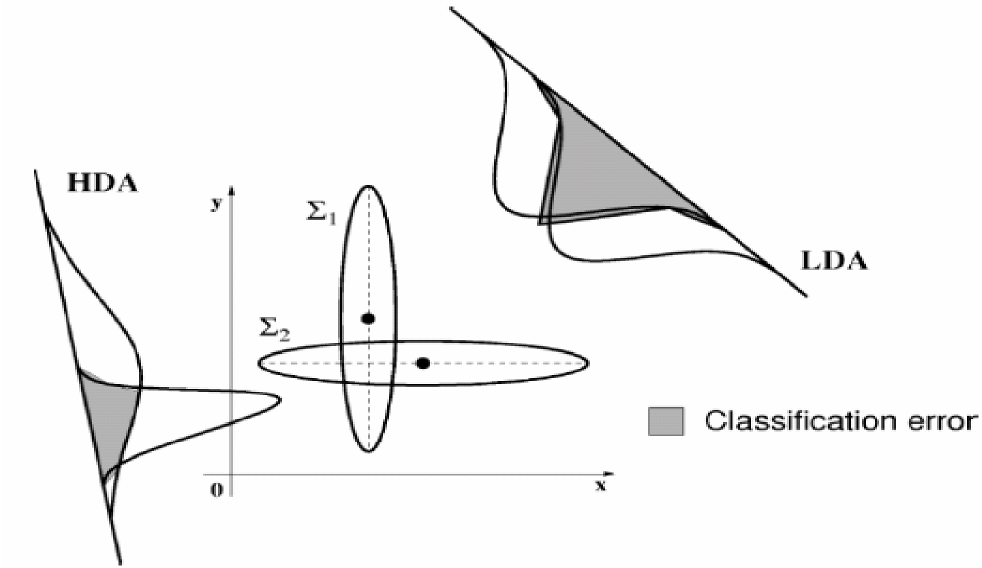


Fig. 1. Difference between LDA and HDA.

  – Clearly, the HDA provides a much lower classification error than LDA theoretically
    - However, most statistical modeling assume data samples are Gaussian and have **diagonal** covariance matrices