Methods for Intelligent Systems

Lecture Notes on Feature Projection

2009 - 2010

Simone Tognetti
tognetti@elet.polimi.it

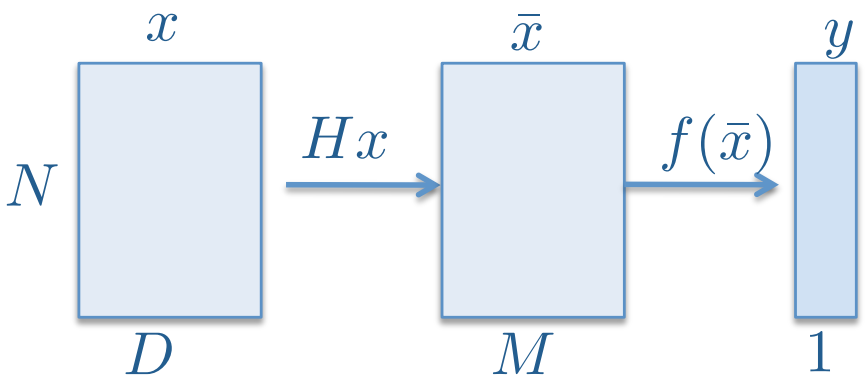
Department of Electronics and Information
Politecnico di Milano

Feature Projection

- Goal: Reduce the number of features D to improve classification accuracy
- Use a linear combination of original features

$$H : R^D \rightarrow R^M$$
$$\bar{x} = Hx$$

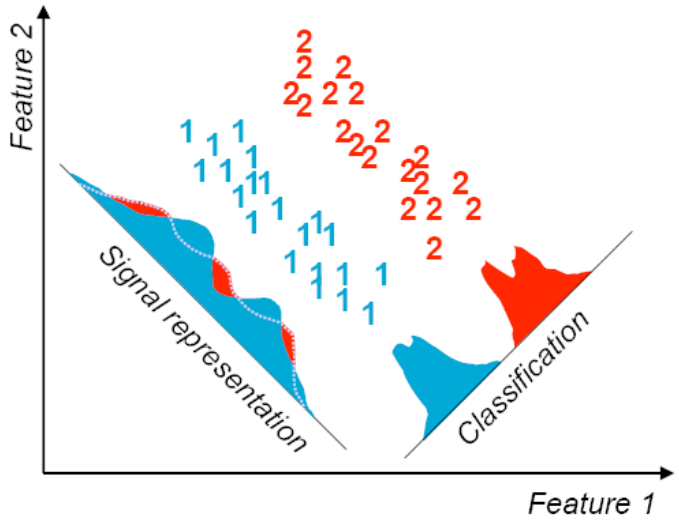
$$\bar{x} = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1N} \\ h_{21} & h_{22} & \dots & h_{2N} \\ \dots & & & \\ h_{M1} & h_{M2} & \dots & h_{MN} \end{bmatrix} x$$



- Projects data into different space in which classes might be best separated
- Resulting features could lose their meaning

Signal Classification vs Signal Representation

- Signal Representation (PCA)
  - Information associated with the data distribution (i.e. mean and variance)
  - No relationship with the classification problem
  - Data should have similar variances
- Signal Classification (LDA)
  - Information associated to discrimination capabilities (inter-class distance)
  - Tendency to over-fit the training data with poor generalization abilities.

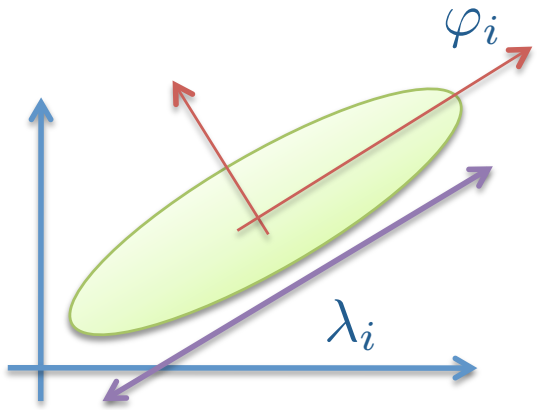


Principal Component Analysis (PCA)

- Hp: Data follows a multi-dimensional Gaussian distribution
- Goal: Find the principal component of the distribution that account for the maximum variance of data
- Covariance matrix

$$\Sigma_x = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2N} \\ \dots & & & \\ \sigma_{N1} & \sigma_{N2} & \dots & \sigma_N^2 \end{bmatrix}$$

$$\sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]$$



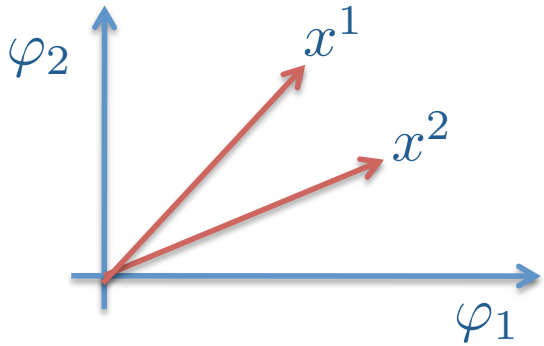
- Decomposition of Σ_x $\quad\quad \Sigma_x \varphi - \lambda\varphi = 0$
  - Eigen vectors φ are principal components of distribution
  - Eigen values λ are the variance of data along principal components

PCA and vectors

- Samples represented in a 2-D vector format

$$x = x_1\varphi_1 + x_2\varphi_2$$

φ_1, φ_2 are axis vectors



- Samples in a D-dimensional space

$$x = \sum_{i=1}^{D} x_i\varphi_i$$

- Suppose we want to represent the data in M-dimensional (M<D) space
  - Replace the features with a constant value

$$\hat{x}(M) = \sum_{i=1}^{M} x_i\varphi_i + \sum_{i=M+1}^{D} b_i\varphi_i$$

- We have an error

$$\Delta x(M) = x - \hat{x}(M) = \sum_{i=M+1}^{D}(x_i - b_i)\varphi_i$$

- Mean squared error

$$\epsilon^2(M) = E[\Delta x(M)^2] = \sum_{i=M+1}^{D} E[(x_i - b_i)^2]$$

PCA and mean squared error

- Find b_i that minimize $\epsilon^2(M)$

$$\frac{\partial}{\partial b_i} E[(x_i - b_i)^2] = -2(E[x_i] - b_i) = 0 \quad b_i = E[x_i]$$

$$\epsilon^2(M) = \sum_{i=M+1}^{D} E[(x_i - E[x_i])^2]$$

$$= \sum_{i=M+1}^{D} E[(x\varphi_i - E[x\varphi_i])^T (x\varphi_i - E[x\varphi_i])]$$

$$= \sum_{i=M+1}^{D} \varphi^T E[(x - E[x])(x - E[x])^T]\varphi = \sum_{i=M+1}^{D} \varphi^T \Sigma_x \varphi$$

- Adding the orthonormality constraint

$$\epsilon^2(M) = \sum_{i=M+1}^{D} \varphi^T \Sigma_x \varphi + \sum_{i=M+1}^{D} \lambda_i (1 - \varphi_i^T \varphi)$$

- Find φi that minimize $\epsilon^2(M)$

$$\frac{\partial}{\partial \varphi_i} \epsilon^2(M) = 2(\Sigma_x \varphi_i - \lambda_i \varphi_i = 0) \quad \Sigma_x \varphi_i = \lambda_i \varphi_i$$

PCA Step by step

1. Given the dataset x
2. Compute the covariance matrix Σ_x
3. Solve the characteristic equation

$$\Sigma_x \varphi - \lambda \varphi = 0$$

4. Chose the first M eigenvectors corresponding to the largest eigenvalues
5. Project the data

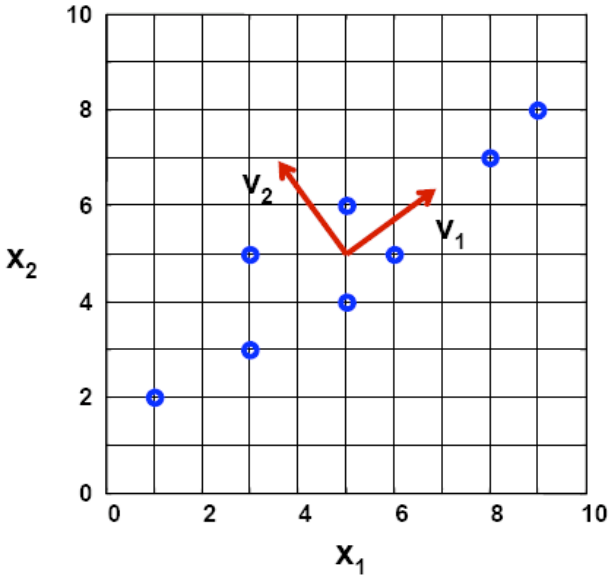
$$H = [\varphi_i | \varphi_2 | \dots | \varphi_M] \qquad \bar{x} = [\varphi_i | \varphi_2 | \dots | \varphi_M] x$$

PCA Example

- Dataset $x = \{(1,2),(3,3),(3,5),(5,4),(5,6),(6,5),(8,7),(9,8)\}$
- Covariance matrix

$$\Sigma_x = \begin{bmatrix} 7.1429 & 4.8571 \\ 4.8571 & 4.0000 \end{bmatrix}$$

$$\mu = (5,5)$$



- Decomposition of covariance matrix

$$\Sigma_x \varphi - \lambda \varphi = 0$$

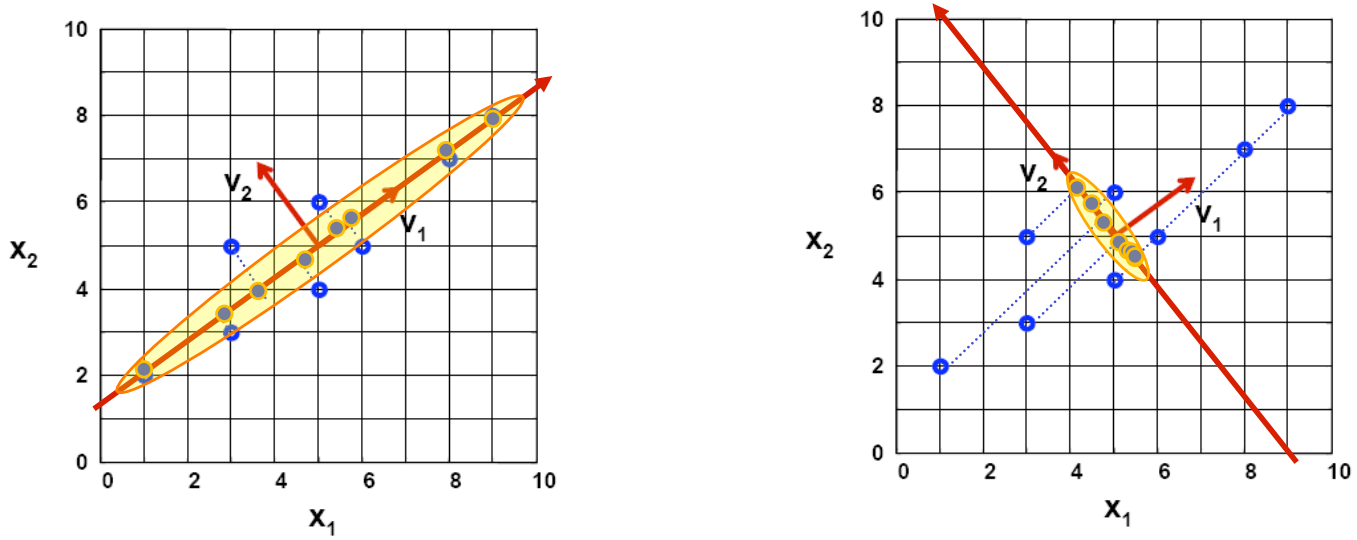
$$\begin{vmatrix} 7.1429 - \lambda_1 & 4.8571 \\ 4.8571 & 4.0000 - \lambda_2 \end{vmatrix} = 0$$

$$H = [\varphi_1 | \varphi_2] = \begin{bmatrix} -0.8086 & -0.5883 \\ -0.5883 & 0.8086 \end{bmatrix}$$

$$\lambda_1 = 10.6764, \lambda_2 = 0.4664$$

PCA Example (2)

- Projection into principal components



- Ration of component variance and total variance $\frac{\lambda_i}{\sum_{i=1}^{D} \lambda_i}$

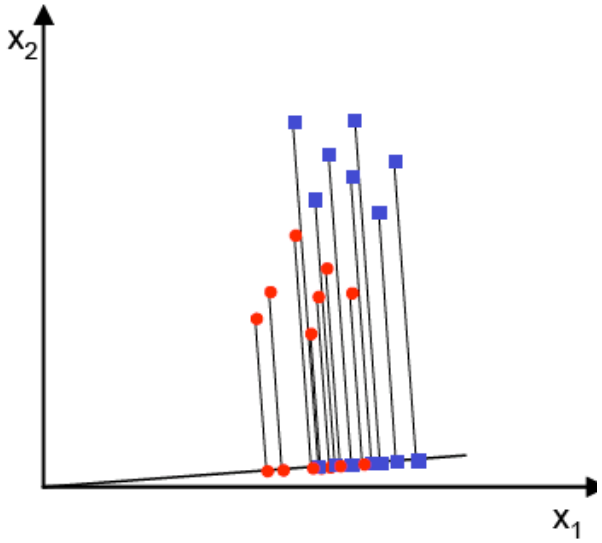
$$\frac{\lambda_1}{\sum_{i=1}^{D} \lambda_i} = 0.9581 \qquad \frac{\lambda_2}{\sum_{i=1}^{D} \lambda_i} = 0.0419$$

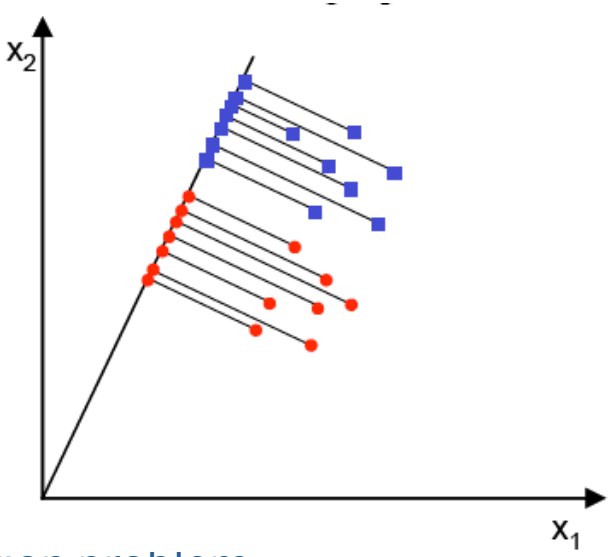
PCA Notes

- PCA aims to decompose the covariance matrix
- $\sum_x$ is estimated under the assumption of a Gaussian Distributions
  - If data are not normally distributed, PCA de-correlates features
- PCA does not use class labels to project data
- Projection depends only on the data structure
- By stretching one of the principal component, the distribution of data does not change
- There is no guarantee that principal components best separate classes

Fisher's Linear Discriminant Analysis (LDA)

- LDA is a linear projection
- the projections maximizes **intra-class** separability (among different classes) and minimizes **inter-class** separability (in the same class**).**





- The projection is find by solving an optimization problem
  - Which measure should be minimized?

Fisher's Linear Discriminant Analysis (LDA)

- A possible measure of separability: sample mean
  - Sample mean for class ω_i

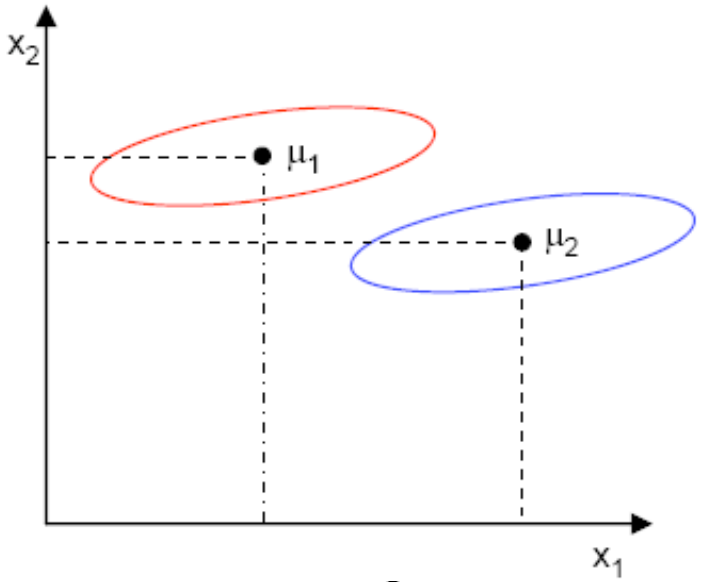
$$\mu_i = \frac{1}{N_i} \sum_{x \in \omega_i} x$$

  - Projected mean

$$\tilde{\mu}_i = \frac{1}{N_i} \sum_{x \in \omega_i} W^T x = W^T \mu_i$$

  - Measure of separation

$$J(W) = |\tilde{\mu}_1 - \tilde{\mu}_2|$$



- the distance between the projected means is not enough: it does not take in account the standard deviation

Fisher's Linear Discriminant Analysis (LDA)

- Within-class scatter matrix

$$S_\omega = \sum_{\omega_i \in Y} \sum_{x \in \omega_i} (x - \mu_i)(x - \mu_i)^T$$

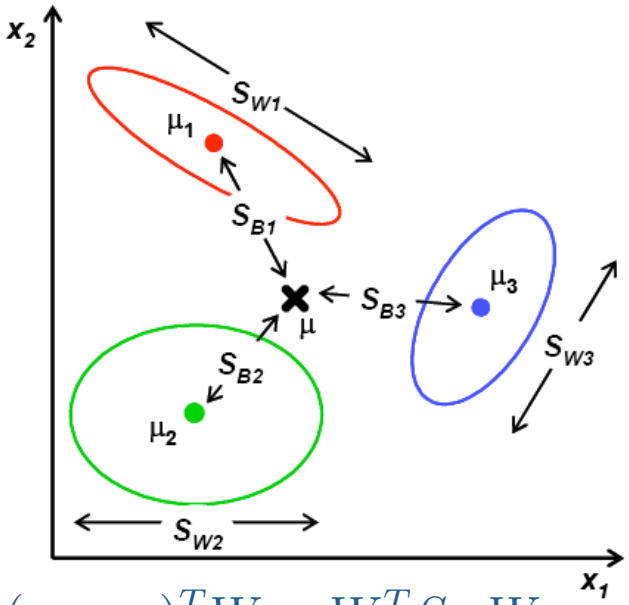
- Between-class scatter matrix

$$S_b = \sum_{\omega_i \in \mathcal{Y}} N_i (\mu - \mu_i)(\mu - \mu_i)^T$$

- Measure of separation

$$J(W) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

  - Projected means are well separated
  - Projected variance are small



$$\tilde{s}_i^2 = \sum_{x \in \omega_i} (W^T x - X^T \mu_i) = \sum_{x \in \omega_i} W^T (x - \mu_i)(x - \mu_i)^T W = W^T S_{\omega_i} W$$

$$\tilde{s}_1^2 + \tilde{s}_2^2 = W^T S_\omega W$$

$$(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (W^T \mu_1 - W^T \mu_2)^2 = W^T (\mu_1 - \mu_2)(\mu_1 - \mu_2) W = W^T S_b W$$

$$J(W) = \frac{W^T S_b W}{W^T S_w W}$$

Fisher's Linear Discriminant Analysis (LDA)

- Find W that maximizes J(W)

$$\frac{\partial}{\partial W} J(W) = \frac{\partial}{\partial W}\left[\frac{W^T S_b W}{W^T S_w W}\right] = 0$$

$$(W^T S_w W)\frac{\partial}{\partial W}(W^T S_b W) - (W^T S_b W)\frac{\partial}{\partial W}(W^T S_w W) = 04$$

$$(W^T S_w W)2S_b W - (W^T S_b W)2S_w W = 0$$

  - Dividing by $(W^T S_w W)$

$$\frac{W^T S_w W}{W^T S_w W} S_b W - \frac{W^T S_b W}{W^T S_w W} S_w W = 0$$

$$S_b W - J S_w W = 0$$

$$S_w^{-1} S_b W - J W = 0$$

  - Solving the generalized eigenvector problem we obtain W that maximize J

LDA step by step

1. Given the dataset x
2. Computes the class statistic

$$S_{\omega_i} = \sum_{x \in \omega_i} (x - \mu_i)(x - \mu_i)^T$$

3. Compute the within-class scatter matrix

$$S_\omega = \sum_{i=1..|Y|} S_{\omega_i}$$

4. Compute the between-class scatter matrix

$$S_b = \sum_{\omega_i \in \mathcal{Y}} N_i (\mu - \mu_i)(\mu - \mu_i)^T$$

5. Solve the generalized eigenvalue problem

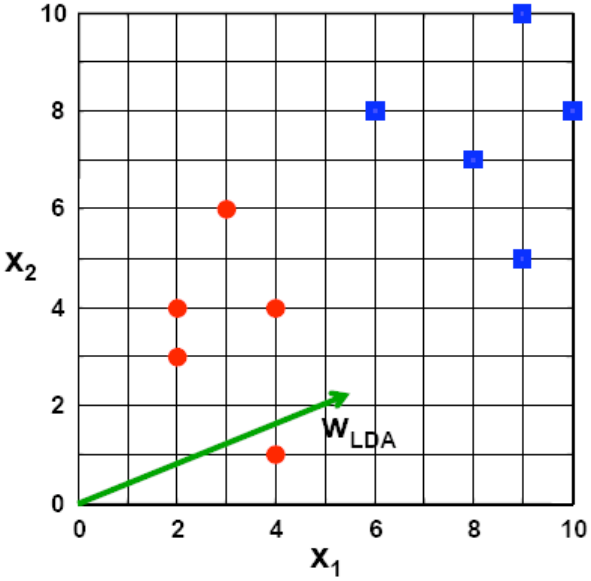
$$S_w^{-1} S_b W - JW = 0$$

LDA example

1. Dataset

$$x=\{(4,1), (2,4), (2,3), (3,6), (4,4) , (9,10), (6,8), (9,5), (8,7), (10,8)\}$$

$$y = \{1, 1, 1, 1, 1, 2, 2, 2, 2, 2\}$$



2. Compute class statistic

$$S_{\omega_i} = \sum_{x \in \omega_i} (x - \mu_i)(x - \mu_i)^T$$

   - Class 1

$$\mu_1 = [3, 3.6] \quad S_{\omega_1} = \begin{bmatrix} 4 & -2 \\ -2 & 13.2 \end{bmatrix}$$

   - Class 2

$$\mu_2 = [8.4, 7.6] \quad S_{\omega_2} = \begin{bmatrix} 9.2 & -0.2 \\ -0.2 & 13.2 \end{bmatrix}$$

3. Compute the within-class scatter matrix

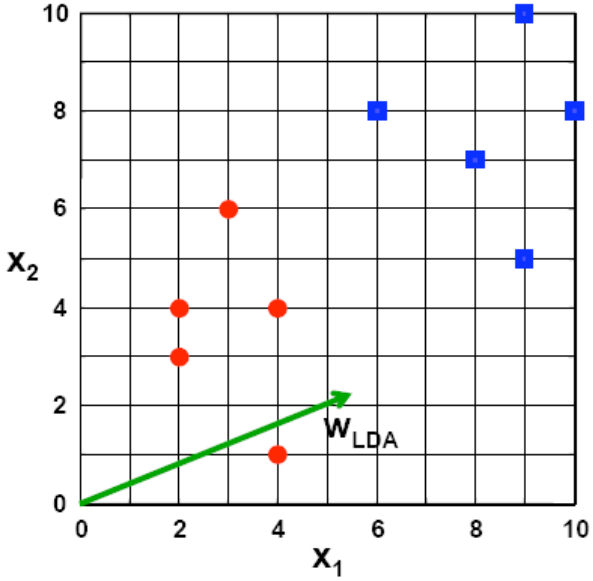
$$S_\omega = \begin{bmatrix} 13.2 & -0.2 \\ -0.2 & 26.4 \end{bmatrix}$$

LDA example (2)

4. Compute the between-class scatter matrix

$$S_b = \sum_{\omega_i \in \mathcal{Y}} N_i (\mu - \mu_i)(\mu - \mu_i)^T$$

$$S_b = \begin{bmatrix} 72.9 & 54 \\ 54 & 40 \end{bmatrix}$$



5. Solve the generalized eigen value problem

$$S_w^{-1} S_b W - JW = 0$$

$$|S_w^{-1} S_b - \lambda I| = 0$$

$$\begin{vmatrix} 5.9462 - \lambda_1 & 4.4046 \\ 2.5410 & 1.8822 - \lambda_2 \end{vmatrix} = 0$$

$$W = \begin{bmatrix} 0.9196 & -0.5952 \\ 0.3930 & 0.8036 \end{bmatrix} \qquad \lambda_1 = 7.8284 \; \lambda_2 = 0$$

Note on LDA

- Produces only C-1 projections
- Hp of unimodal distribution of data within each class.
  - If data are highly non linear the resulting projection is sub optimal
- Non-parametric Linear Discriminant Analysis remove the unimodal assumpion
  - S_b is computed with local information trough a K-NN
  - S_b result in a full rank matrix and projection is made over more than c-1 classes
- LDA fails when the information is contained in the variance rather than the mean
  - Encode the variance as a new feature!

Other dimensionality reduction techniques

- Kernel PCA
- Indipendent Component Analysis (ICA)
- Multilayer Perceptron
- Self organizing maps (SOMs)
- Sammon's map
- Support vectors machine (SVM)
  - Margin maximization