8. KL-Transform and Linear Discriminant Analysis (LDA)

Linear discriminant analysis

From Wikipedia, the free encyclopedia.

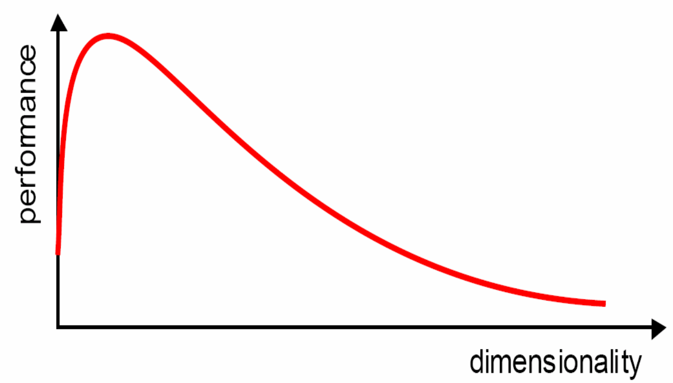
Linear discriminant analysis (LDA), is sometimes known as Fisher's linear discriminant, after its inventor, Ronald A. Fisher, who published it in *The Use of Multiple Measures in Taxonomic Problems* (1936). It is typically used as a feature extraction step before classification

Dimensionality Reduction

- Curse of dimensionality:
  - higher the dimension of the feature vectors

$\mapsto$ data sparsity

$\mapsto$undertrained classifier



Try to reduce dimension of feature vectors without loss of information

Established methods

- Two methods are established
  - Karhunen-Loeve transformation (KL transformation):
    - tries to describe the data as good as possible in a lower dimensional space
    - Based on principle component analysis (PCA)
  - Linear discriminant analysis (LDA):
    - try to optimize class separability
    - Also known as Fisher´s discriminant analysis

Idea of KL transform

Let $\vec{\phi}_i$ with i = 1...D be an orthonormal basis

Approximate the feature vector $\tilde{\vec{x}} = \sum_{i=1}^{d} y_i \vec{\phi}_i$

Expand the original feature vector $\vec{x} = \sum_{i=1}^{D} y_i \vec{\phi}_i$

with d < D

How do you determine the optimal basis?

Idea of KL transform

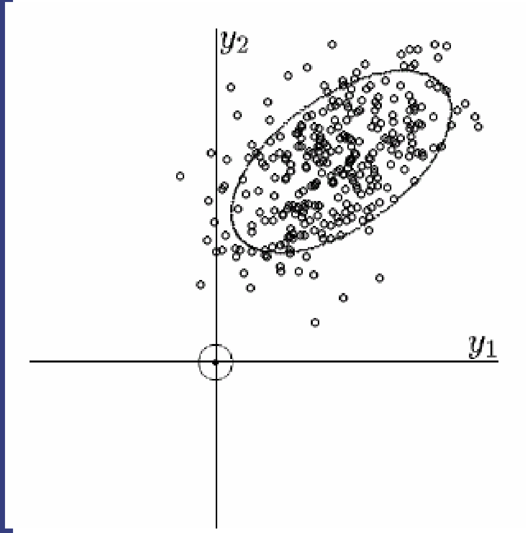
Minimize approximation error $\varepsilon_{\mathrm{d}} = \sum_{j=1}^{N} (\vec{x}_j - \tilde{\vec{x}}_j)^2$

j labels all the feature vectors $\vec{x}_j$

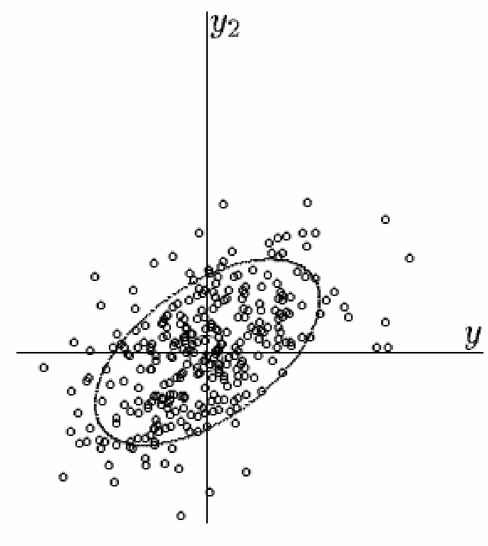
available in the training data

Effect of KL-transform

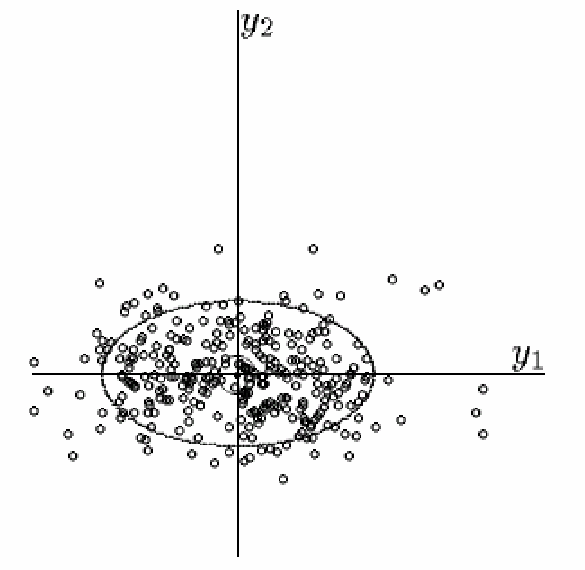
Original data



Center your data



KL-Transform



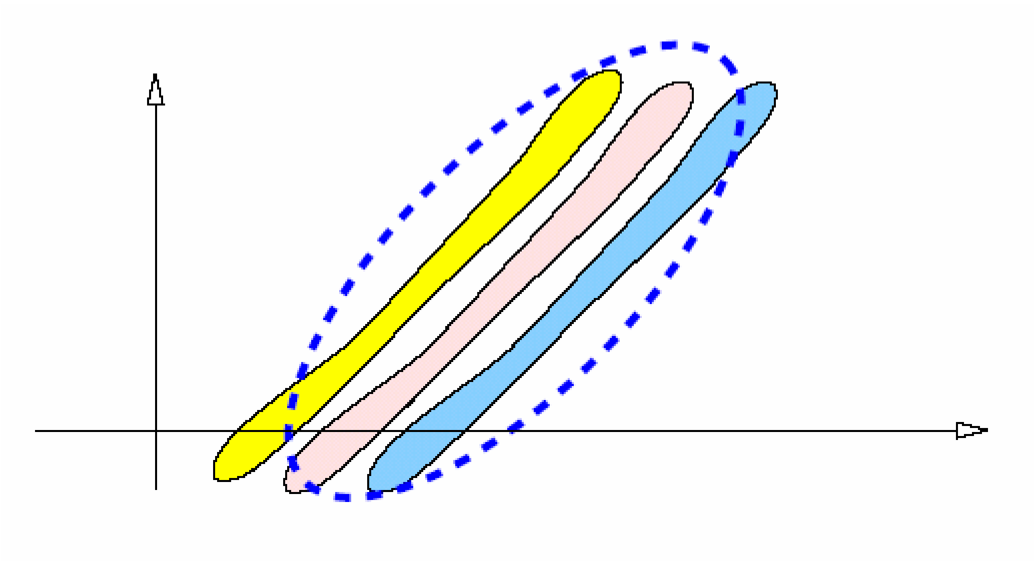
Let’s start simple

- d=1
- Decompose into mean and best direction

$$\tilde{\vec{x}}_j = \vec{\mu} + a_j \vec{e}$$

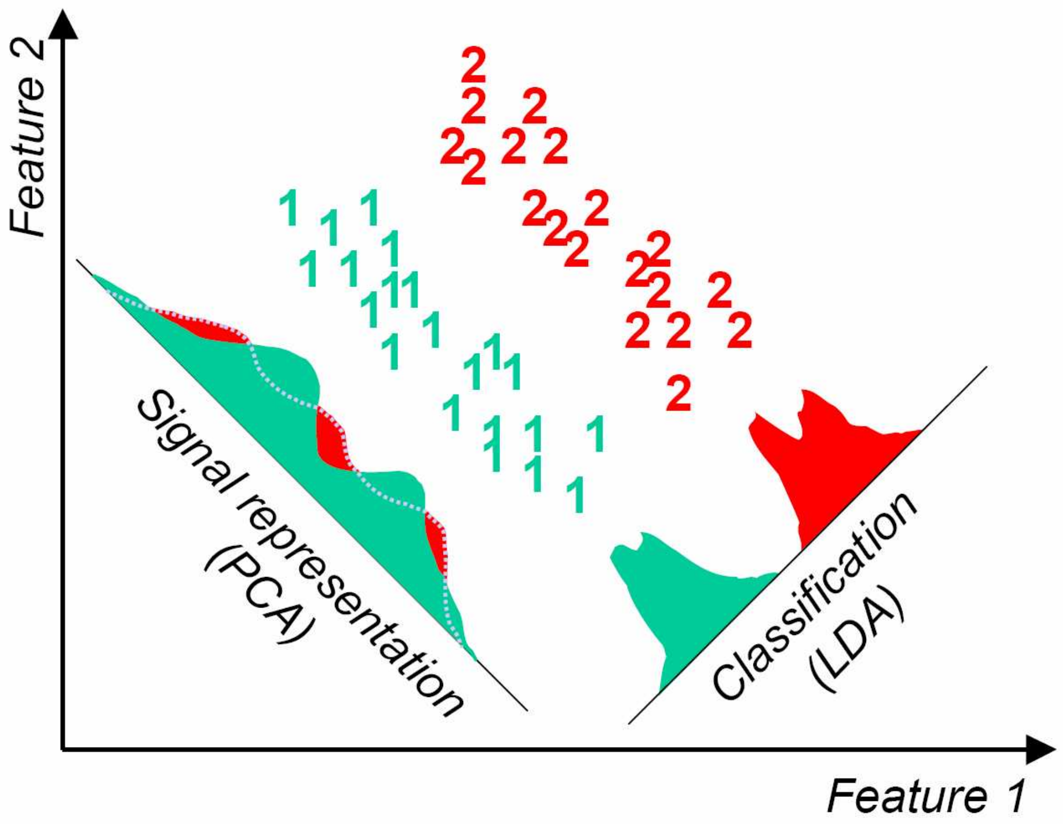
What does the KL transform to this data

Distribution of data for three classes:

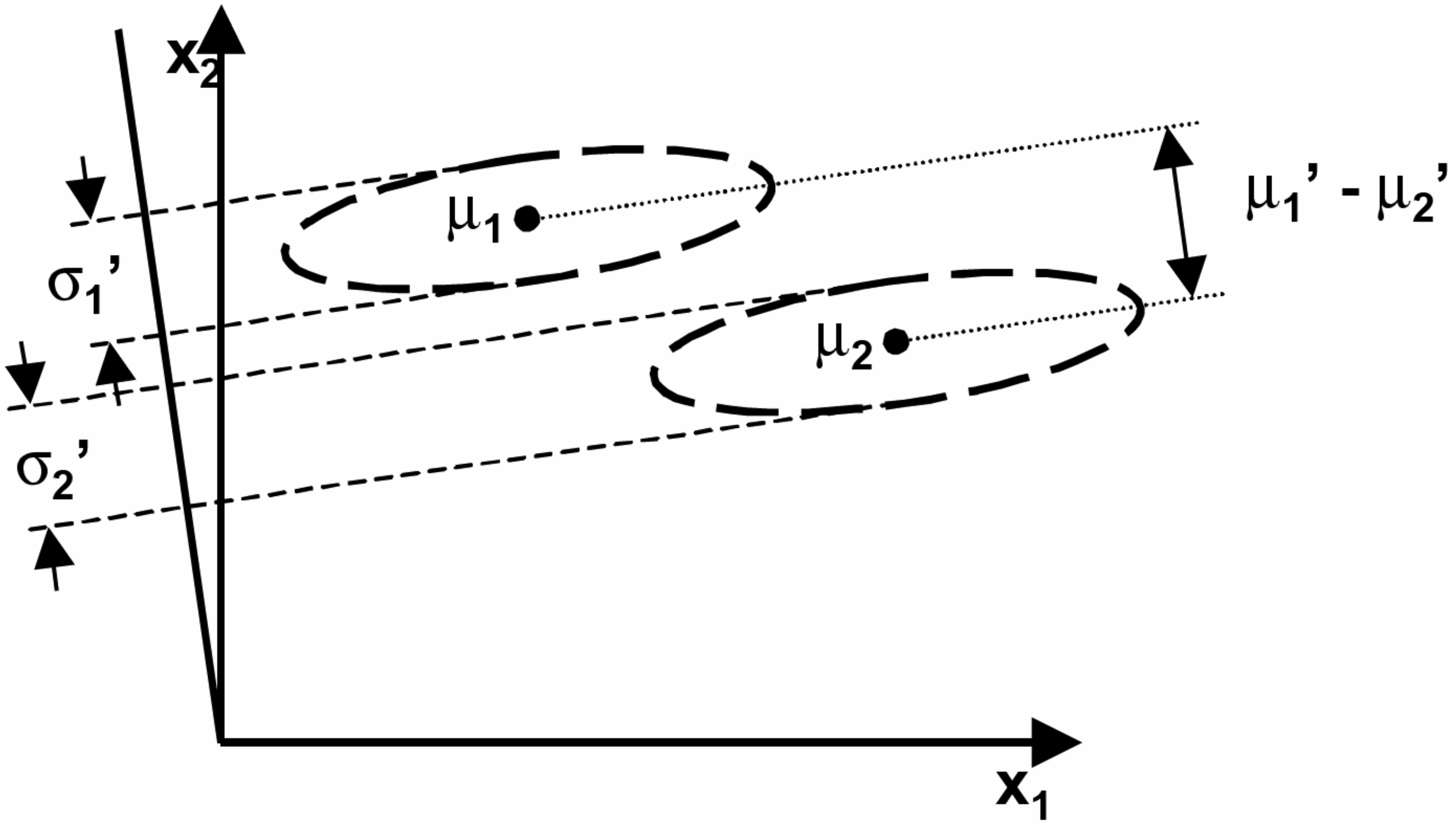


What will the KL-transform do?

Difference KL transform/LDA



Idea of LDA: try to maximize class separability



Definition: between-class-scatter-matrix

Between - class - scatter - matrix

$$S_b = \sum_{k=1}^{K} p_k (\vec{\mu}_k - \vec{\mu})(\vec{\mu}_k - \vec{\mu})^t$$

with :

K : number of classes

$$p_k = \frac{\mathrm{N_k}}{\sum_{l=1}^{K} N_l} \quad \text{(fraction of data belonging to class k)}$$

$$\vec{\mu}_k = \frac{1}{\mathrm{N_k}} \sum_{i=1}^{N_k} \vec{\mathrm{x}}_{\mathrm{i,k}} \quad \text{(mean vector of class k)}$$

$\vec{\mu}$: mean of all vectors

Definition: within-class-scatter-matrix

Within - class - scatter - matrix

$$S_w = \sum_{k=1}^{K} p_k \Sigma_k$$

with

$$\Sigma_k = \frac{1}{\mathrm{N_k}} \sum_{i=1}^{N_k} (\vec{\mathrm{x}}_{\mathrm{i,k}} - \vec{\mu}_k)(\vec{\mathrm{x}}_{\mathrm{i,k}} - \vec{\mu}_k)^{\mathrm{t}} \quad \text{(covariance matrix of class k)}$$

LDA

- Maximize class separability
- Keep variance of all classes roughly constant

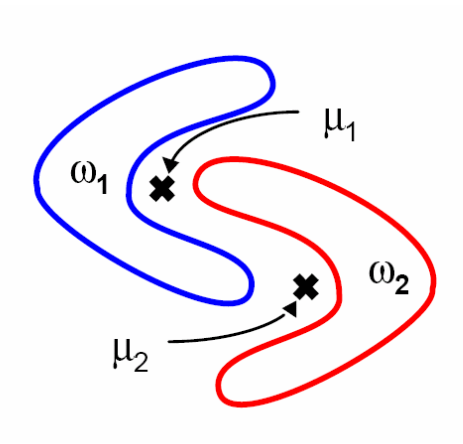
$\mapsto$ optimization problem with constraint

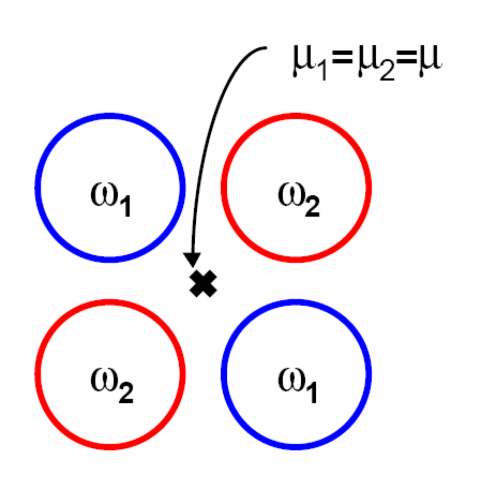
Solution

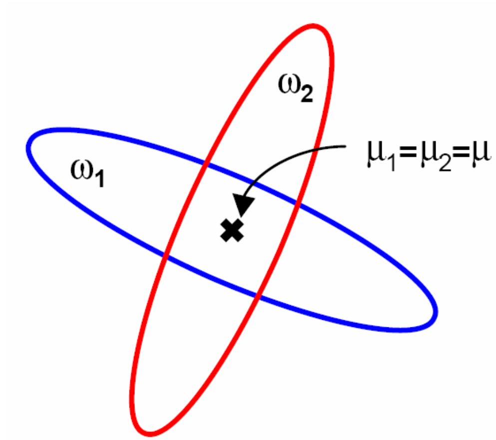
$$\boxed{S_b \vec{\phi}_i = \lambda_i S_w \vec{\phi}_i}$$

Limitations of LDA

LDA implicitly assumes
Gaussian distribution of data

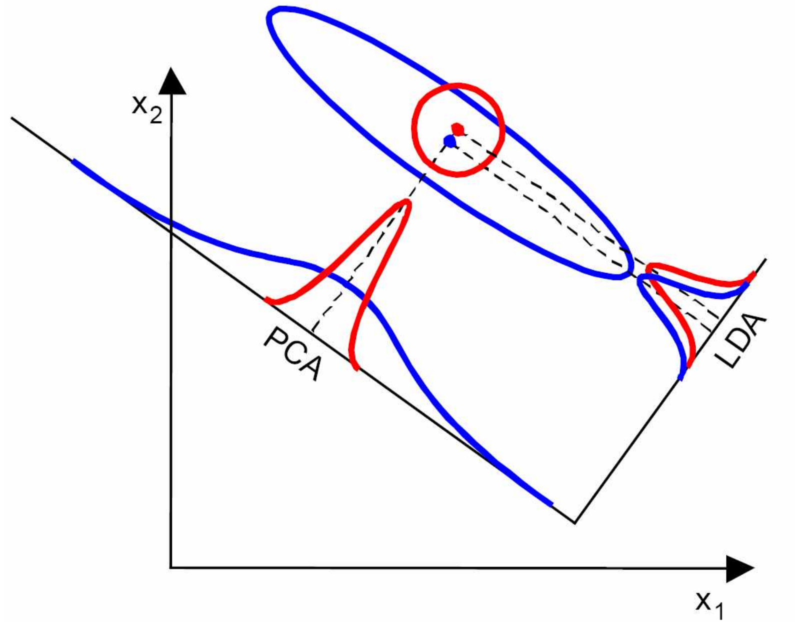






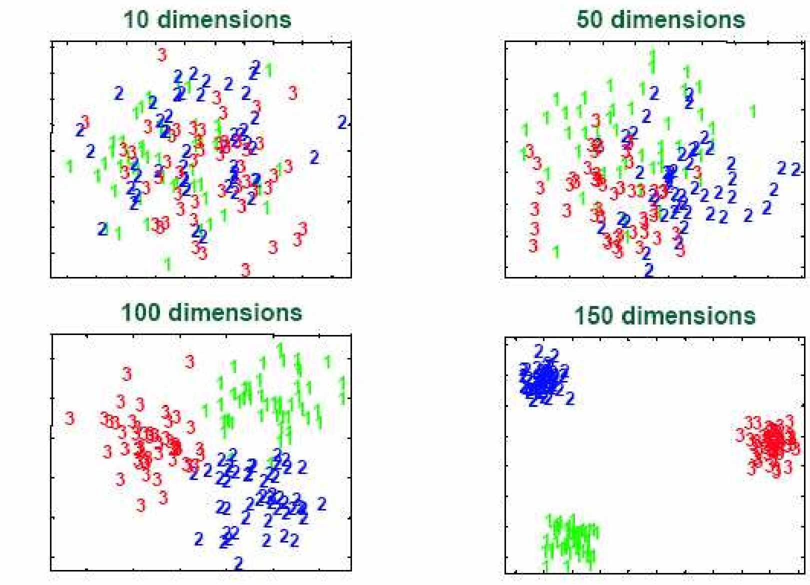
Limitations of LDA

LDA implicitly assumes that the mean is the discriminating factor, not variance



Limitations of LDA

LDA may overfit the data



Example:
-three multivariate Gaussian distributions with zero mean
-50 samples drawn from each Gaussian

Summary

- To increase performance of classifier
  - Use KL transform (PCA)
  - LDA
- LDA has limitations but improved versions exist