Linear Discriminant Analysis (LDA)

Dimensionality Reduction & Class Discrimination

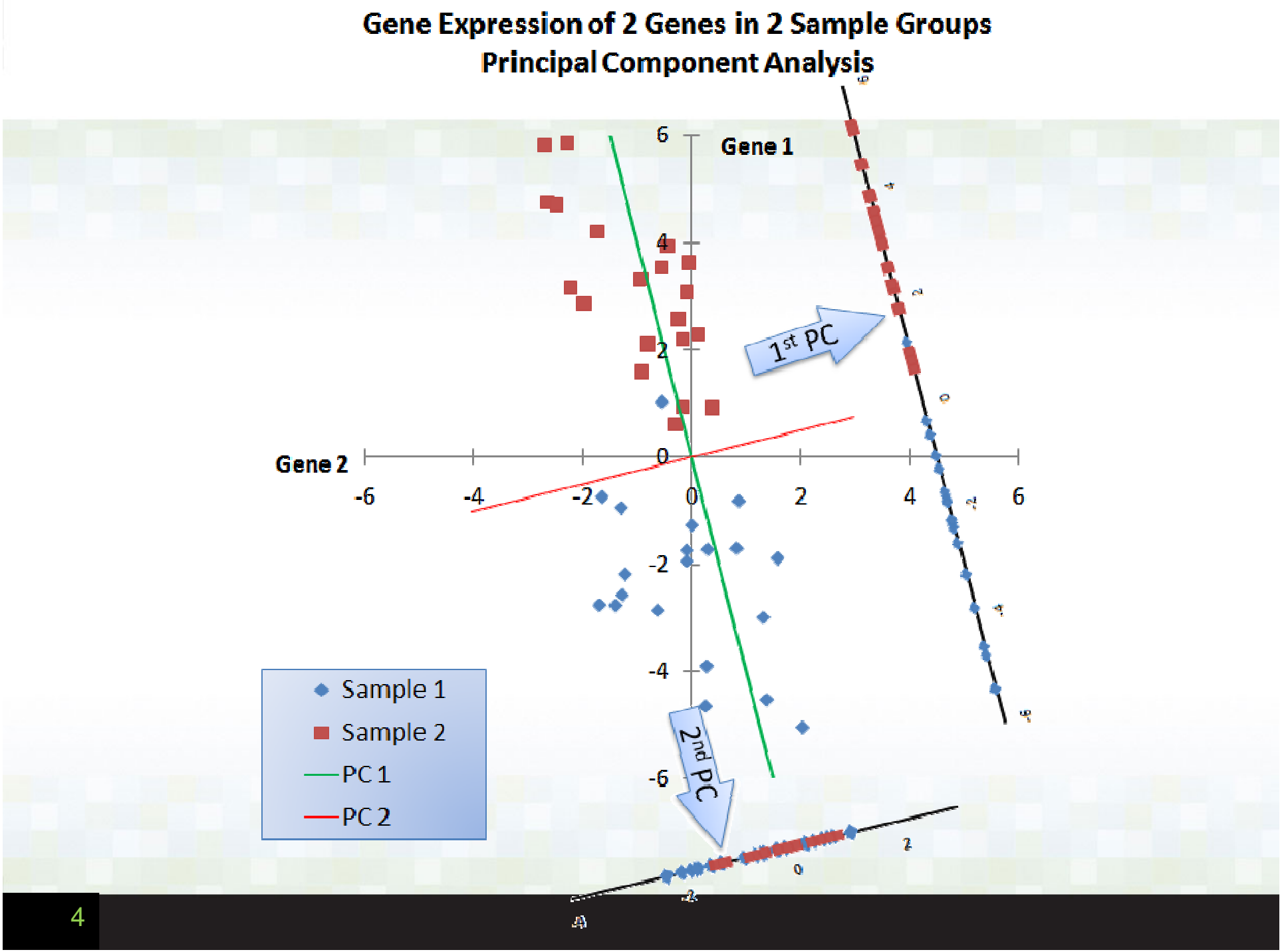
CS 6316 – Machine Learning
Fall 2017

OUTLINE

- Main Limitation of PCA
- Limitations of Logistic Regression
- Linear Discriminant Analysis (LDA)
- LDA Objective
- LDA Method
- Some Limitations of LDA
- PCA vs LDA
- LDA Example

Linear Discriminant Analysis

Photo by Jamie McCaffrey ~ https://www.flickr.com/photos/15609463@N03/14898932531


Gene Expression of 2 Genes in 2 Sample Groups
Principal Component Analysis
Gene 1
Gene 2
1st PC
2nd PC
Sample 1
Sample 2
PC 1
PC 2
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6

Main Limitation of PCA

- The main limitation of PCA is that it does not consider class separability since it does not take into account the class label of the feature vector
  - PCA simply performs a coordinate rotation that aligns the transformed axes with the directions of maximum variance
  - *There is no guarantee that the directions of **maximum variance** will contain good features for **discrimination***

Limitations of Logistic Regression

- Logistic regression (LR) is a simple and powerful linear classification algorithm
- However, it has a number of limitations that suggest the need for alternative linear classification algorithms
- **Limitations of LR**:
  - Two-class problems: *LR is intended for binary classification problems*
  - Unstable with well separated classes
  - Unstable with few examples: *LR can become unstable when there are few examples from which to estimate the parameters*

Limitations of Logistic Regression

- Linear Discriminant Analysis (LDA) does address each of these points
- It is the go-to linear method for multi-class classification problems
- Even with binary-classification problems, it is a good idea to try both LR and LDA

LDA

Linear Discriminant Analysis (LDA) is most commonly used as dimensionality reduction technique in the pre-processing step for pattern-classification and machine learning applications. The goal is to project a dataset onto a lower-dimensional space with good class-separability in order avoid overfitting ("curse of dimensionality") and also reduce computational costs

Dimensionality reduction does not only help reducing computational costs for a given classification task, but it can also be helpful to avoid overfitting by minimizing the error in parameter estimation ("curse of dimensionality")

LDA

- The LDA approach is very similar to a Principal Components Analysis (PCA), but in addition to finding the component axes that maximize the variance of the data [this is PCA], it additionally is interested in the axes that maximizes the separation between multiple classes [LDA]

LDA: *Simplifying* Assumptions (About the Data)

1. The data is Gaussian
   - *Each variable is shaped like a bell curve when plotted*
2. Each attribute has the same variance
   - *That values of each variable vary around the mean by the same amount on average*

With these assumptions, the LDA model estimates the mean and variance from your data for each class

LDA: All about Mean and Variance

- The purpose of linear discriminant analysis (LDA) is to estimate the probability that a sample belongs to a specific class given the data sample itself. That is to estimate

  $\Pr(C = c_i | X = x)$, $where\ C = \{c_1, c_2, \ldots, c_m\}$

  is the set of class identifiers, X is the domain, and x is the specific sample

- Applying Bayes Theorem results in:

$$Pr(C = c_i | X = x) = \frac{Pr(X=x|C=c_i)Pr(C=c_i)}{\sum_{j=1}^{m} Pr(X=x|C=c_j)Pr(C=c_j)}$$

LDA: All about Mean and Variance

$$Pr(C = c_i|X = x) = \frac{Pr(X=x|C=c_i)Pr(C=c_i)}{\sum_{j=1}^{m} Pr(X=x|C=c_j)Pr(C=c_j)}$$

- The probability of a particular class $\Pr(C = c_i)$ can be estimated by the frequency of the class c_i in the training data
- Remember, LDA assumes that each class can be modeled as a multivariate Gaussian distribution with each class sharing a common covariance matrix

LDA Objective

- LDA seeks to reduce dimensionality while preserving as much of the class discriminatory information as possible

- Assume we have a set of D-dimensional samples

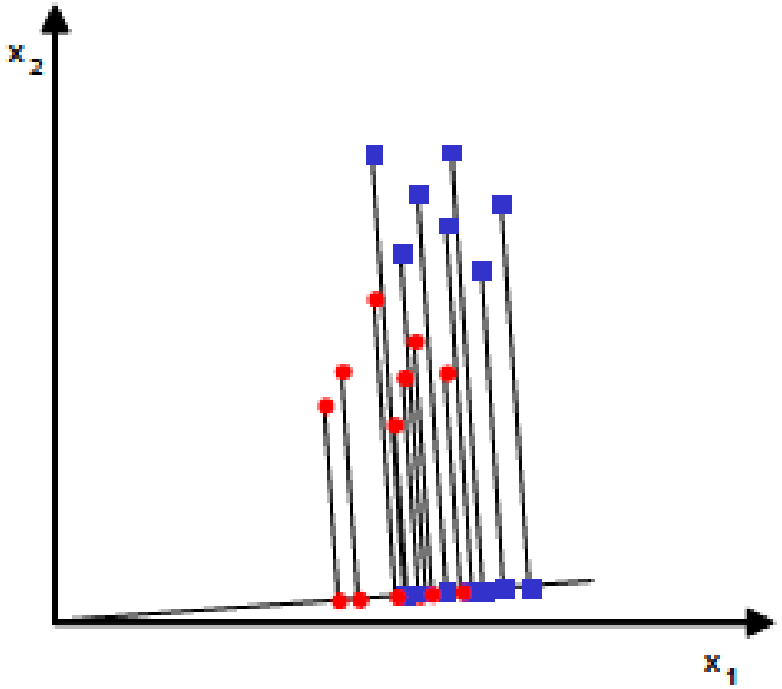
$$\{x^1, x^2, \ldots x^N\}$$

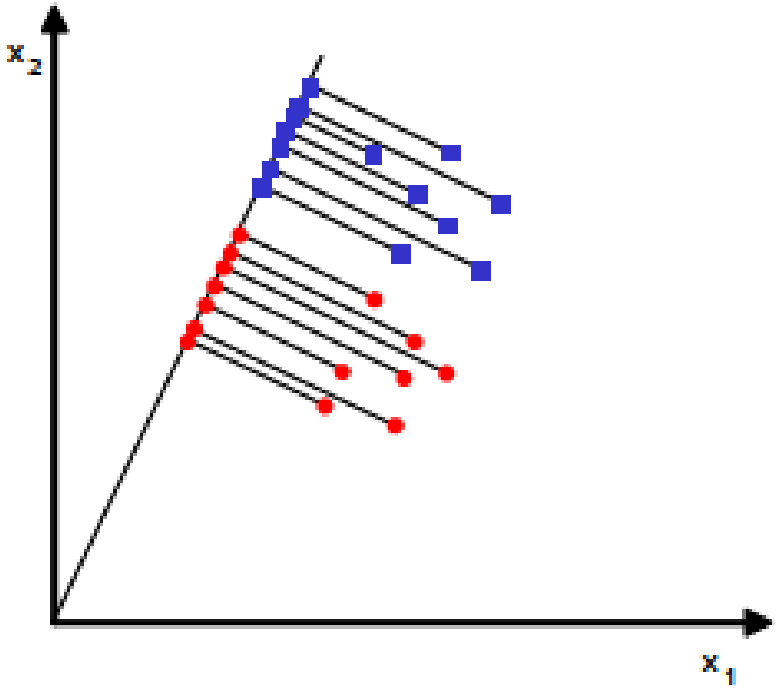
N_1 of which belong to class ω_1, and N_2 to class ω_2

- Wish to obtain a scalar y by projecting the samples x onto a line $y = w^T x$

LDA Objective

- Of all the possible lines we would like to select the one that maximizes the separability of the scalars



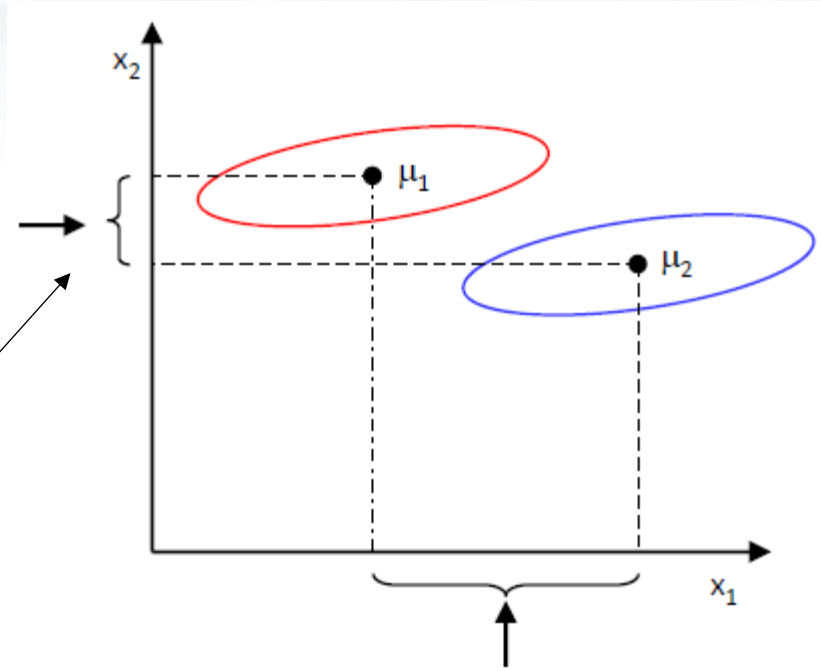


LDA Objective

- To find a good projection vector, we need to define a measure of separation:
  - E.g. mean vector of each class in x-space and y-space

- We could then choose the distance between the projected means as our objective function
- However, the distance between projected means is **not** a good measure since it does not account for the standard deviation within classes

Separation of Means



- Y-axis; this axis yields better class separability
- X-axis; this axis has a larger distance between means

LDA Solution

- Fisher suggested ("Fisher's LDA") maximizing the difference between the means, normalized by a measure of the within-class scatter

- So this involves calculating a
  - within-class scatter matrix
  - between-class scatter matrix

e.g. for Iris data set these scatter matrices will be 4x4-dimensional (4 = number of dimensions = number of attributes)

d-dimensional mean vectors

- As an example for the iris data set, the computation of the mean vectors m_i, (i=1, 2, 3) of the 3 different flower classes:

$$m_i = \begin{bmatrix} \mu_{\omega_i}(sepal\ length) \\ \mu_{\omega_i}(sepal\ width) \\ \mu_{\omega_i}(petal\ length) \\ \mu_{\omega_i}(petal\ width) \end{bmatrix}, with\ i = 1, 2, 3$$

where ω_i=class i

Within-class Scatter Matrix

$$S_W = \sum_{i=1}^{c} S_i$$

where c is the number of classes, and

$$S_i = \sum_{x \in D_i}^{n} (x - m_i)(x - m_i)^T$$

(scatter matrix for every class)

where D is the dimension (attributes) and

$$m_i = \frac{1}{n_i} \sum_{x \in D_i}^{n} x_k$$

Within-class Scatter Matrix

- Alternatively, we could also compute the class-covariance matrices
- Since in this example (iris) all classes have the same sample size we can drop the N-1 term (N=sample size of respective class), the equation becomes:

$$S_W = \sum_{i=1}^{c} (N_i - 1)\Sigma_i$$

where Σ represents the covariance matrix

Between-class scatter

$$S_B = \sum_{i=1}^{c} N_i (m_i - m)(m_i - m)^T$$

where m is the overall mean, and m_i and N_i are the sample mean and sizes of the respective classes (c)

Solving the generalized eigenvalue problem for the matrix:

$S_W^{-1} S_B$ to obtain the linear discriminants

Eigenvectors: will form the new axes of the new feature subspace, and

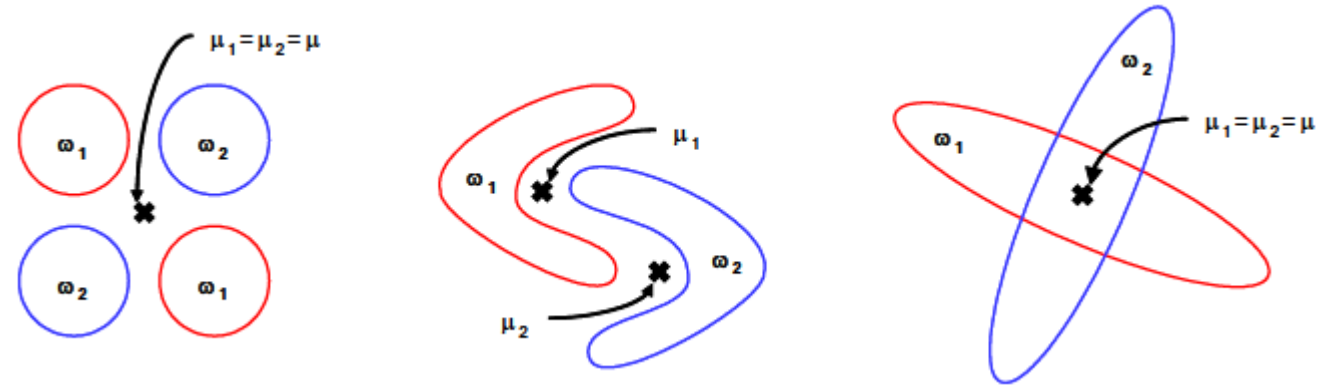
Eigenvalues: how "informative" new "axes" are

LDA Method (Summary)

1. Compute the d-dimensional mean vectors for the different classes from the dataset.
2. Compute the scatter matrices (in-between-class and within-class scatter matrix).
3. Compute the eigenvectors ($\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d$) and corresponding eigenvalues ($\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_d$) for the scatter matrices.
4. Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a $d \times k$ dimensional matrix $\boldsymbol{W}$ (where every column represents an eigenvector).
5. Use this $d \times k$ eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the matrix multiplication: $\boldsymbol{Y} = \boldsymbol{X} \times \boldsymbol{W}$ (where $\boldsymbol{X}$ is a $n \times d$-dimensional matrix representing the n samples, and $\boldsymbol{y}$ are the transformed $n \times k$-dimensional samples in the new subspace).

Some Limitations of LDA

- LDA is a parametric method (it assumes unimodal Gaussian likelihoods)
  - If the distributions are significantly non-Gaussian, the LDA projections may not preserve complex structure in the data needed for classification

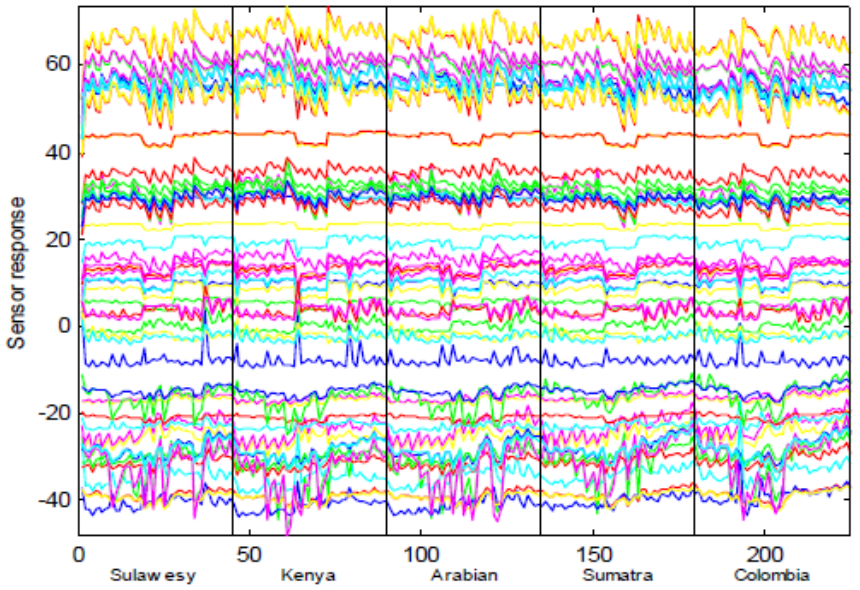


- LDA will also fail if discriminatory info is not in the mean but in the variance of the data

PCA vs. LDA

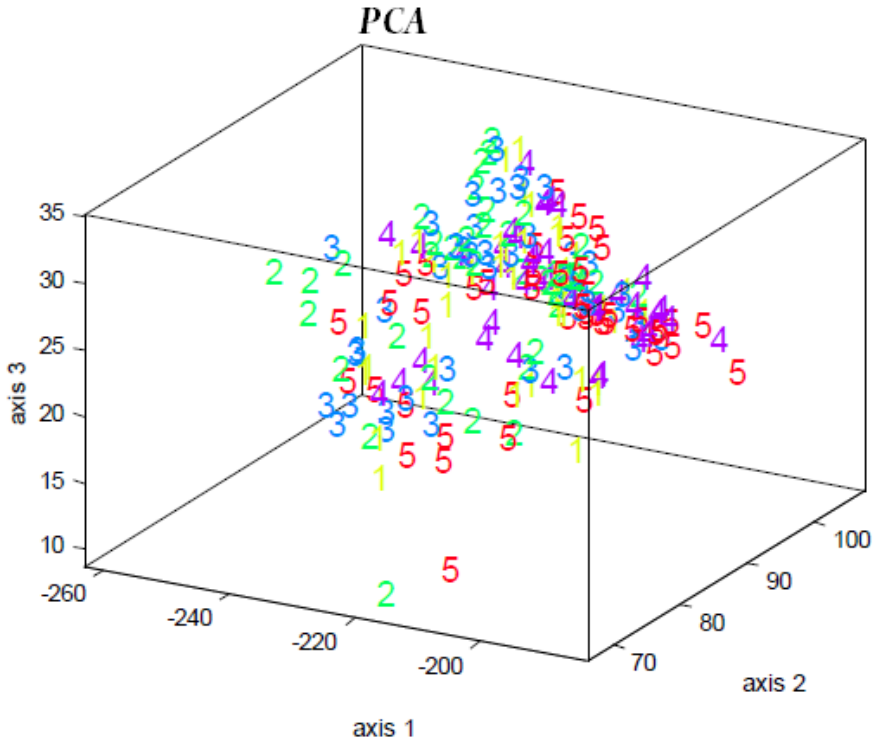
This example illustrates the performance of PCA and LDA on an odor recognition problem

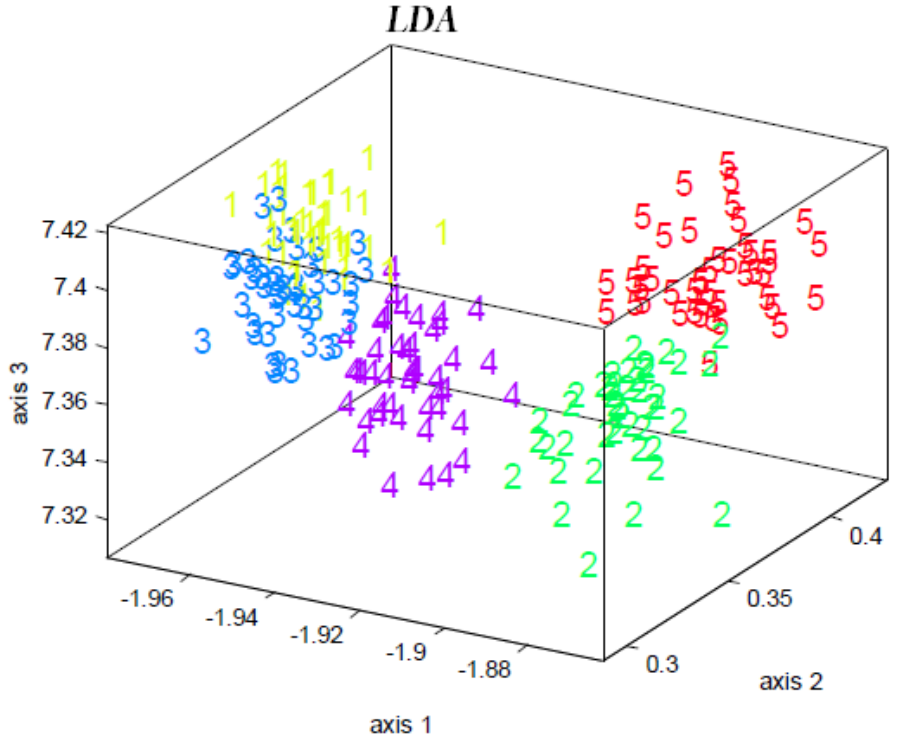
- Five types of coffee beans were presented to an array of gas sensors
- For each coffee type, 45 "sniffs" were performed and the response of the gas sensor array was processed in order to obtain a 60-dimensional feature vector



Results

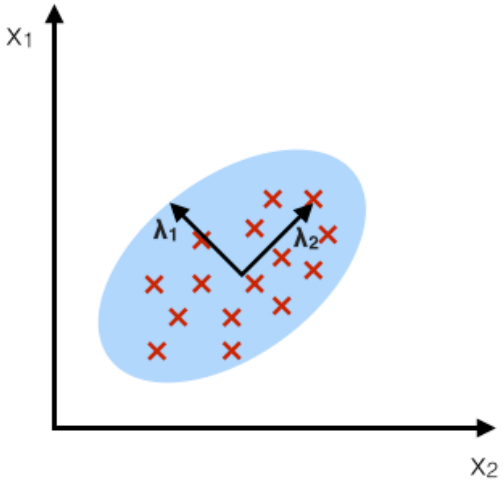
- From the 3D scatter plots it is clear that LDA outperforms PCA in terms of class discrimination
- This is one example where the discriminatory information is not aligned with the direction of maximum variance



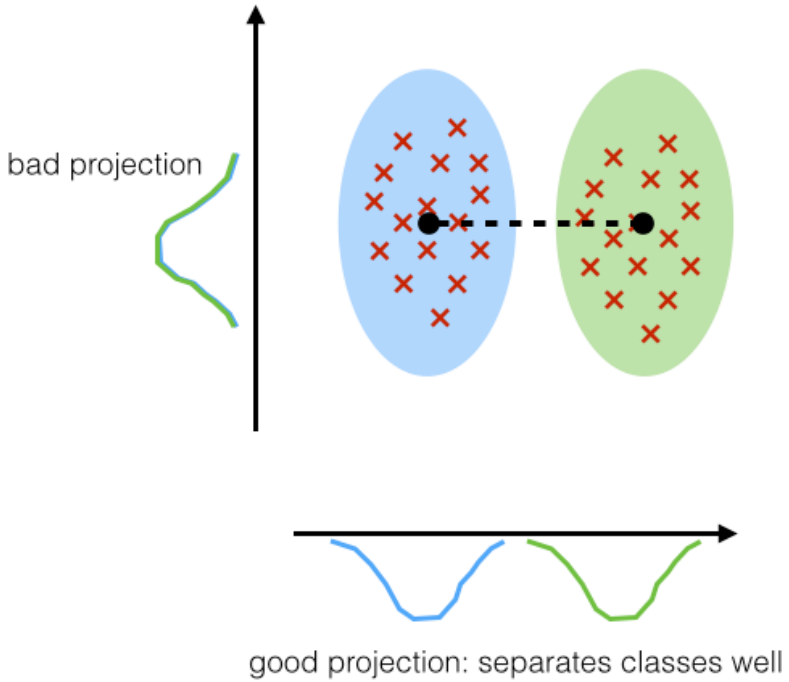


PCA vs. LDA

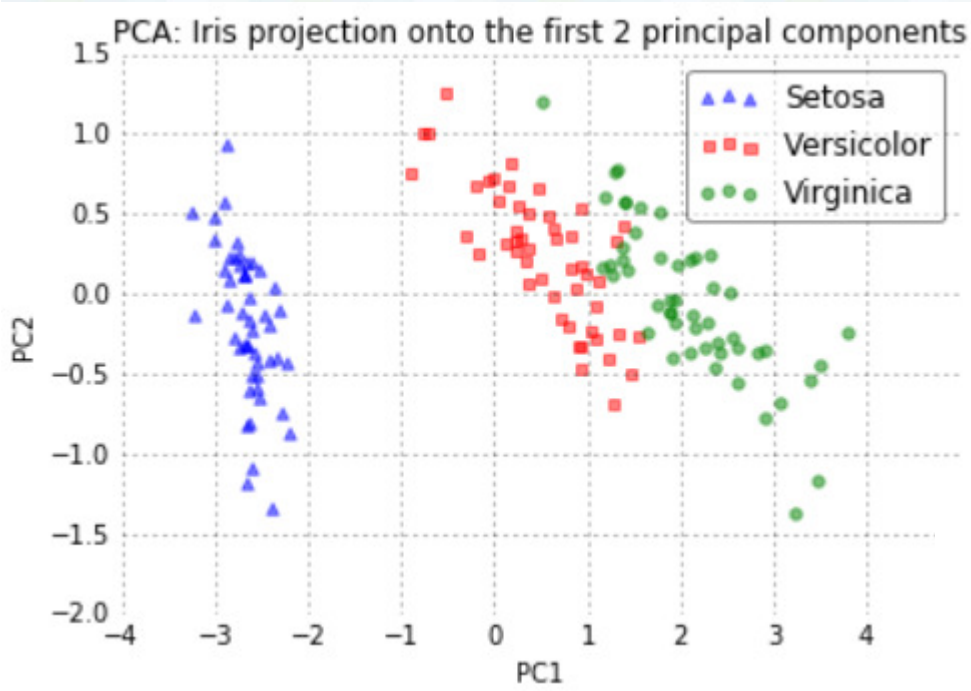
PCA:
component axes that maximize the variance

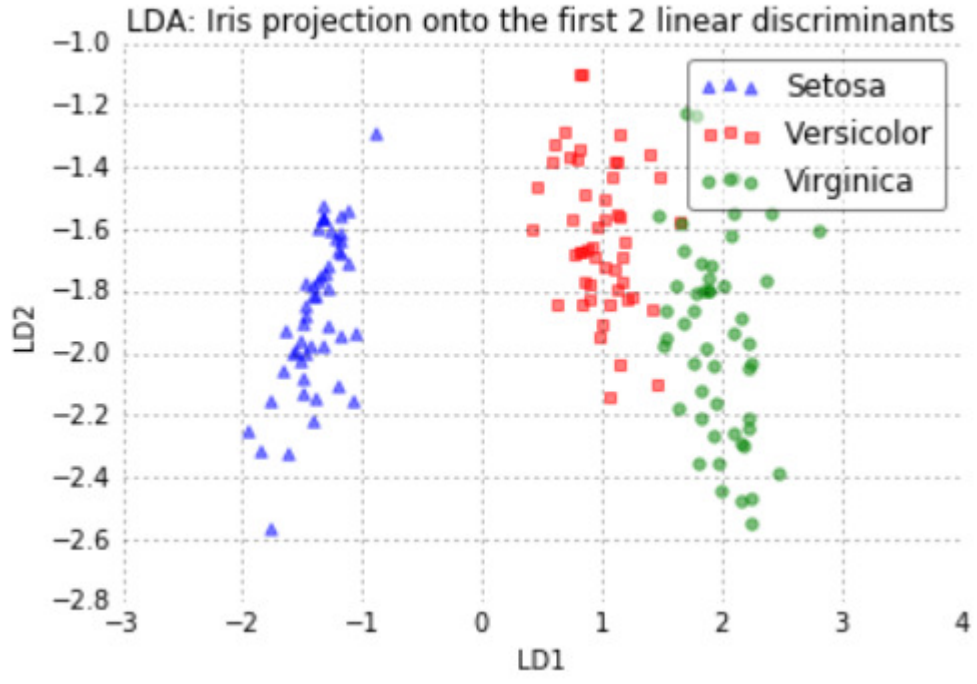


LDA:
maximizing the component axes for class-separation



PCA vs LDA





We can observe better class seperability

LDA Example

References:

~Linear Discriminants Analysis ~ R. Gutierrez-Osuna ~ Pattern Analysis ~ TAMU

~Linear Discriminant Analysis for Machine Learning ~ J.Brownlee

~Linear Discriminant Analysis ~ S.Raschka

~Fisher's Linear Discriminant Analysis (LDA) ~ H.Goel