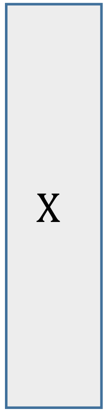
Lecture 4:

Linear Methods of Machine Learning:

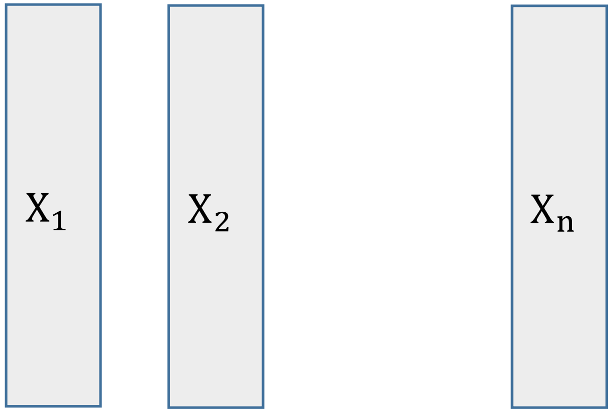
3) Projection Pursuit

Notations

$X = \begin{pmatrix} x_1 \\ \cdots \\ x_p \end{pmatrix} \in R^p$ - p-dimensional vector with components $x_1, x_2, \ldots, x_p$



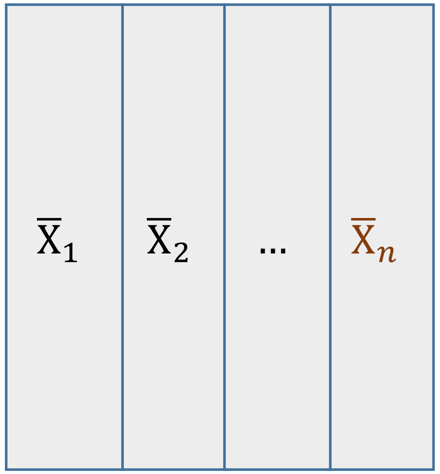
$\{X_1, X_2, \ldots, X_n\}$ – dataset, $X_i = \begin{pmatrix} x_{i1} \\ \cdots \\ x_{ip} \end{pmatrix}$, $i = 1, 2, \ldots, n$



Mean vector $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Centering dataset $\overline{X}_1 = X_1 - \overline{X}$, $\overline{X}_2 = X_2 - \overline{X}$, $\ldots$, $\overline{X}_n = X_n - \overline{X}$

described by $p \times n$ centering data matrix $\overline{\mathbf{X}} = (\overline{X}_1\ \overline{X}_2\ \ldots\ \overline{X}_n)$



Principal component analysis:

- $X \in \mathrm{R}^p$ - random vector
- $b \in \mathrm{R}^p$, $b^T \times b = 1$ - a direction in which projection $b^T \times X$ has maximal variance

$$I_{PCA(1)}(b) = Var(b^T \times X) = b^T \times cov(X) \times b = b^T \times \Sigma \times b$$

where $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}) \times (X_i - \overline{X})^T$ – sample covariance matrix - estimator of $Cov(X)$

- b_1 - found direction
- Look for another direction b, $b^T \times b = 1$, which is orthogonal to direction b_1 and in which projection $b^T \times X$ has maximal variance $I_{PCA(1)}(b) = Var(b^T \times X)$
- b_2 - found direction

After q steps: we found orthogonal matrix $B = (b_1 \ldots b_q)$ - $p \times q$ matrix, $B^T \times B = I_q$, which maximizes

$$I_{PCA(q)}(B) = Tr(B^T \times cov(X) \times B)$$

Independent component analysis based on curtosis:

$b \in R^p$, $b^T \times b = 1$ - a direction in which projection $b^T \times X$ has maximal measure of non-Gaussianity

described by Kurtotis $\mathbf{Kurt}(b^T \times X) = \mathbf{M}(b \times \mathbf{X})^4 - 3(\mathbf{M}(b \times \mathbf{X})^2)^2$

$$I_{ICA\text{-}1}(b) = \left|\mathrm{Kurt}(b^T \times X)\right| = \left|\frac{1}{T}\sum_{t=1}^{T}(b \times X_t)^4 - 3 \times \left(\frac{1}{T}\sum_{t=1}^{T}(b \times X_t)^2\right)^2\right|$$

Independent component analysis based on Shannon negative negentropy:

$b \in R^p$, $b^T \times b = 1$ - a direction in which projection $b^T \times X$ has maximal measure of non-Gaussianity

described by negentropy →

$$I_{ICA\text{-}2}(b) = \frac{1}{T}\sum_{t=1}^{T}\left\{\log_2\left(p\left(b^T \times X_t\right)\right)\right\}$$

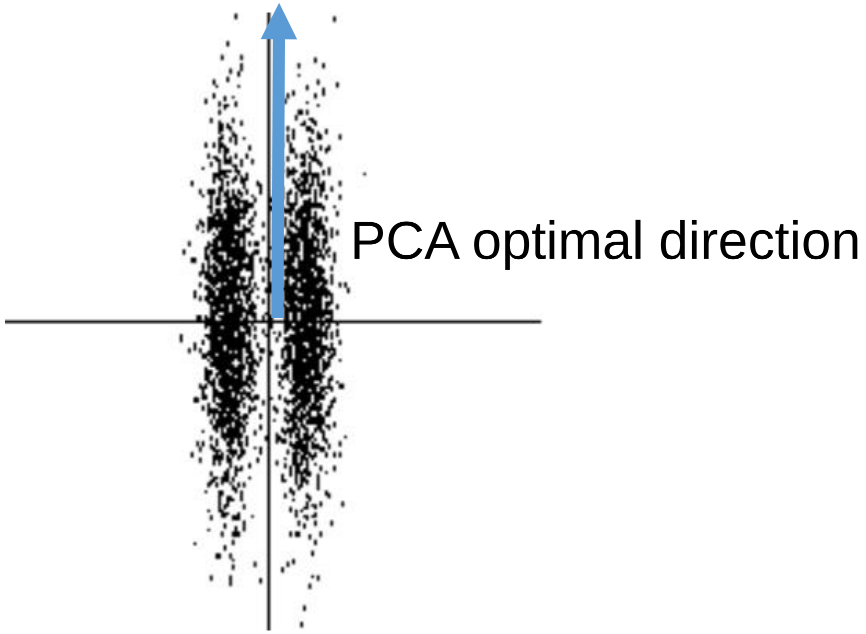
or by some negentropy approximation →

$$I_{ICA\text{-}3}(b) = \frac{1}{12}\left(\frac{1}{T}\sum_{t=1}^{T}\left(b^T \times X_t\right)^3\right)^2 + \frac{1}{48}\left(\frac{1}{T}\sum_{t=1}^{T}\left(b^T \times X_t\right)^4 - 3 \times \left(\frac{1}{T}\sum_{t=1}^{T}\left(b^T \times X_t\right)^2\right)^2\right)^2$$

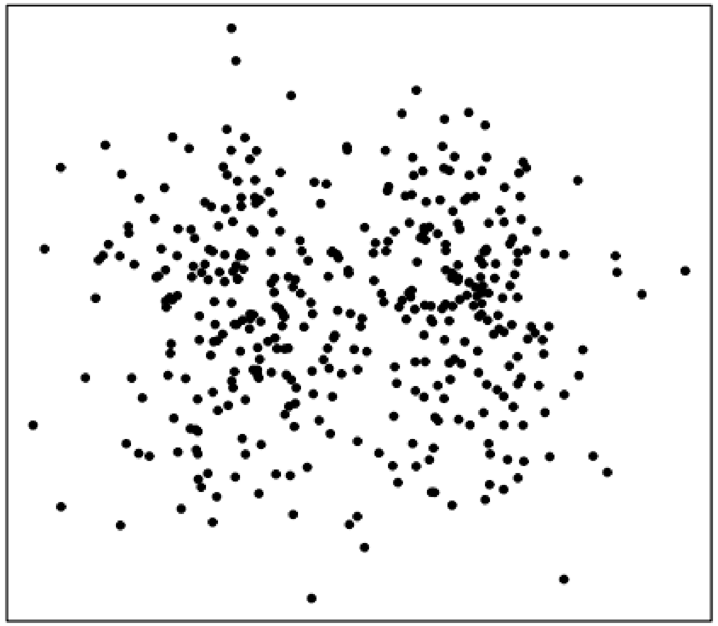
$$I_{ICA\text{-}4}(b) = \frac{1}{T}\sum_{t=1}^{T} h\left(b^T \times X_t\right), \qquad h(y) = \frac{1}{\alpha_1}\log\cosh\left(\alpha_1 \times y\right)$$

Projection Pursuit **(Friedman and Tukey, 1974; Huber, 1985)**

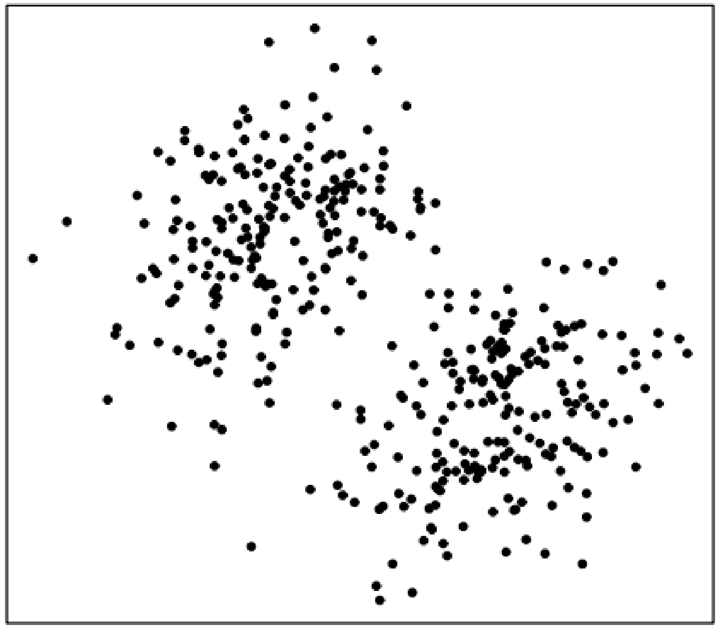
- data analysis technique that finds **interesting** low-dimensional linear orthogonal projections of a high-dimensional data
- for detecting unanticipated ‘structure’ – clusters, outliers, skewness, etc.



Does not allows to detect the clusters



Projection onto plane spanned two largest principal components



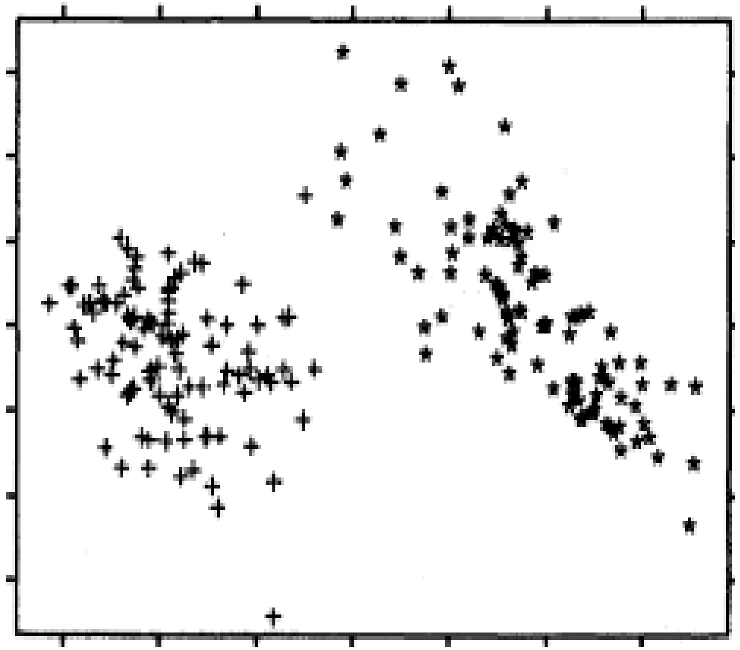
Projection onto plane spanned another **interesting** directions

The Swiss Banknote data set: 100 genuine and 100 forged Swiss bank notes

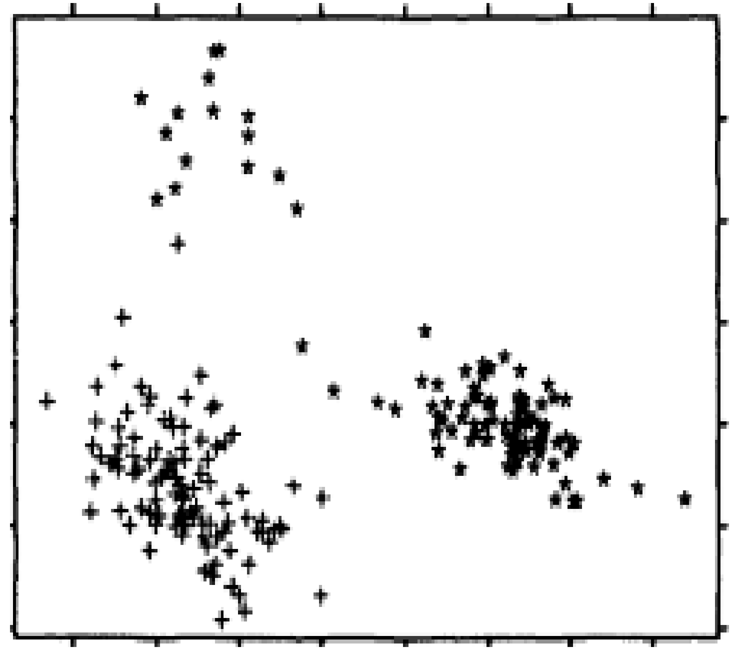
Each banknote – point in R^6 - is described by six variables:

- width of bank note;
- height on left side;
- height on right side;
- lower margin;
- upper margin;
- diagonal of inner box

Projection onto plane spanned two largest principal components



Projection onto plane spanned another directions allows to detect two distinct groups of forged notes



Projection Pursuit - data analysis technique that finds **interesting** projections of a high-dimensional data by optimizing a certain **objective function** called **projection index** $I_1(b)$ or $I_q(B)$

Projection index:

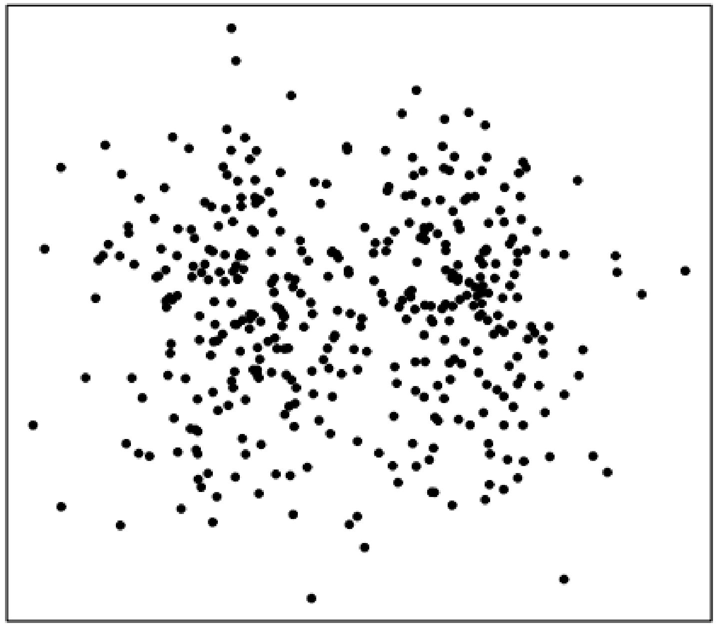
- how to choose
- how to optimize the chosen Index

How to choose Projection index - depends on the data analysis task

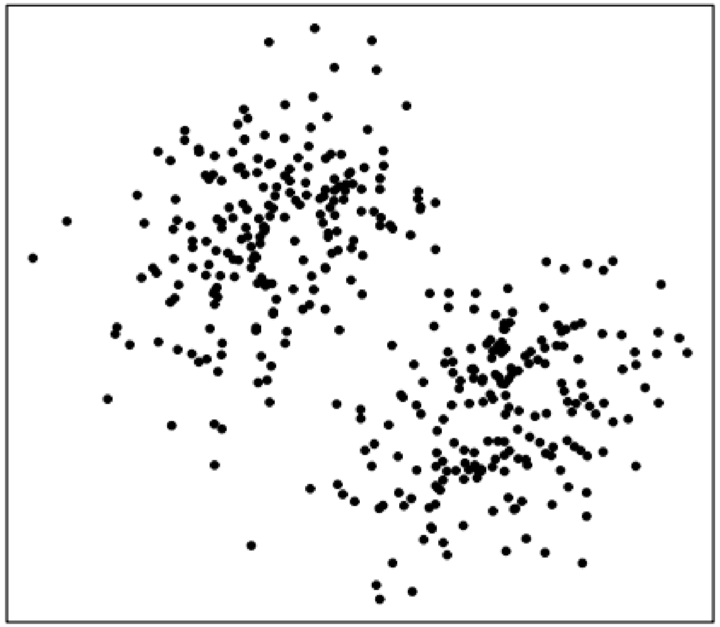
- examples: $I_{PCA(1)}$ $(I_{PCA(q)})$ – in PCA, $I_{ICA\text{-}1}$ $(I_{ICA\text{-}3}, I_{ICA\text{-}3})$ – in ICA
- another examples - Projection indexes for visualization, density estimation, regression, etc.

How to optimize the chosen Index - depends on index

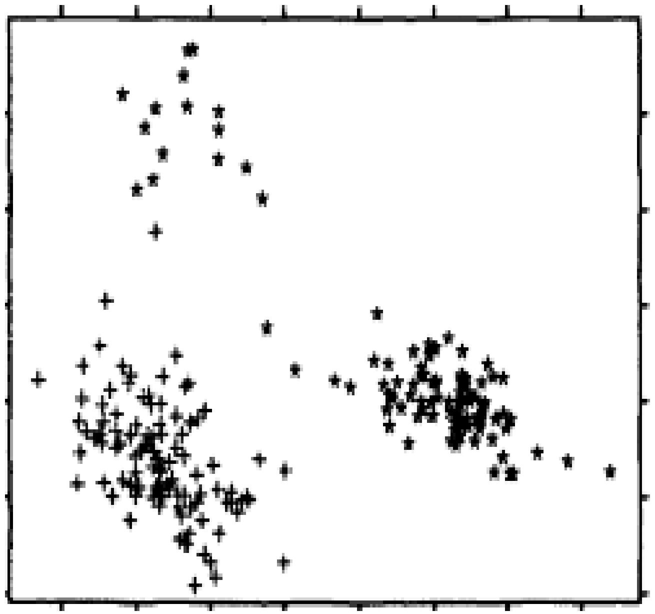
- sometimes – solution to the Eigenvalues problem (PCA, LDA, canonical correlations)
- in other cases – general optimization methods (gradient descend, ‘Newton’ descend, etc.)



Projection onto plane based on PCA-index



Projection onto plane based on ‘Holes’-index



Projection onto plane based on ‘Hermit’-index in ‘Swiss-banknote’ task

Projection index for supervised Linear Discriminant Analysis (LDA) (classification in general case)

Training p-dimensional 'labeled' datasets $\{X_{ij}\}$ from different classes:

- m classes indexed by j: $j = 1, 2, \ldots, m$
- $\mathbf{X}_j = \{X_{1j}, X_{2j}, \ldots, X_{n(j),j}\}$ – data from j^{th}-class indexed by $i = 1, 2, \ldots, n(j), j = 1, 2, \ldots, m$

Mean vector in j^{th}-class: $\overline{X}_j = \frac{1}{n(j)} \sum_{i=1}^{n(j)} X_{ij}$ $\qquad j = 1, 2, \ldots, m$

Total mean vector: $\overline{X} = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n} X_{ij}, \quad n = \sum_{j=1}^{m} n(j)$

Within-class scatter matrix: $\Sigma_{within} = \sum_{j=1}^{m} \sum_{i=1}^{n(j)} (X_{ij} - \overline{X}_j) \times (X_{ij} - \overline{X}_j)^T$

Between-class scatter matrix: $\Sigma_{between} = \sum_{j=1}^{m} n(j) \times (\overline{X}_j - \overline{X}) \times (\overline{X}_j - \overline{X})^T$

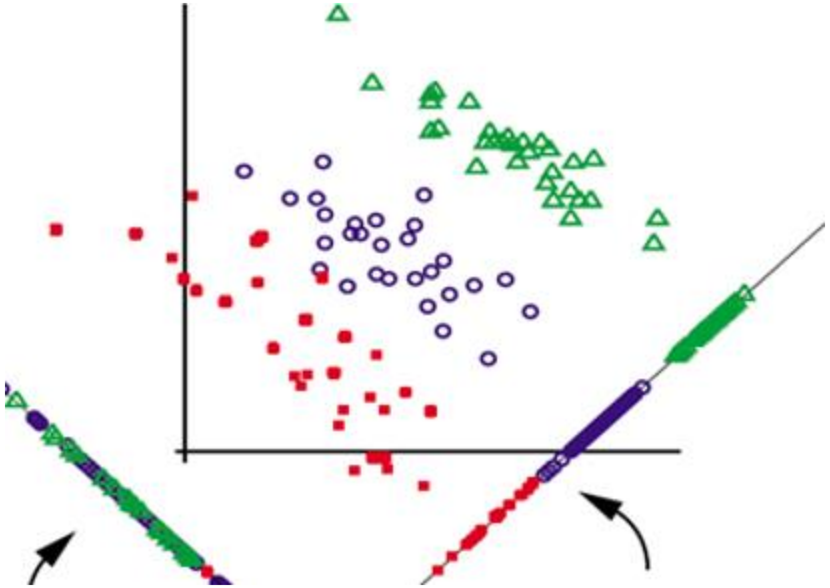
Interesting projections – the projections in which there are **the biggest difference between the observations from different classes** – in which the classes are clustered in the view.

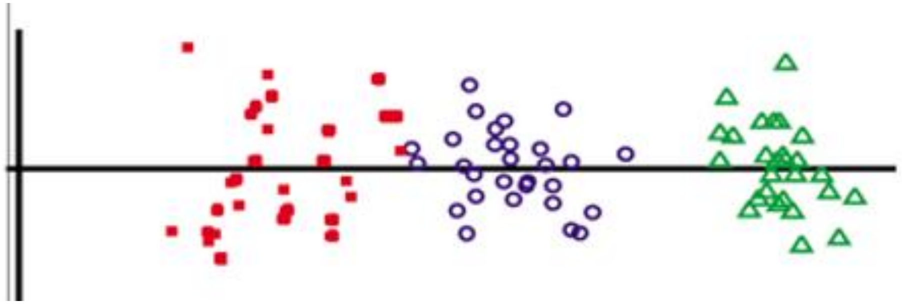
The projection Index reflects certain measure of between-class variation relative to within-class variation: in one-dimensional direction b

$$I(b) = (b^T \times \Sigma_{between} \times b)/(b^T \times \Sigma_{within} \times b) \rightarrow \max$$

The projection Index in multidimensional case: orthogonal matrix projection $p \times q$ matrix B

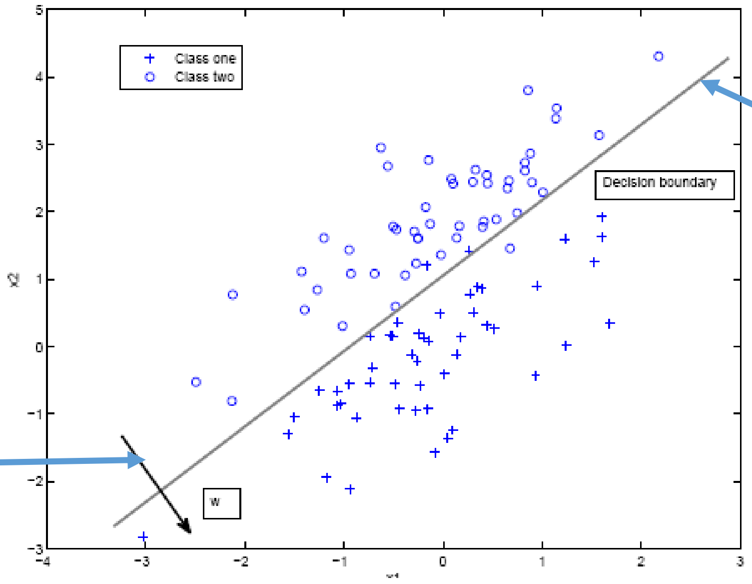
$$I(B) = 1 - \frac{\mathrm{Det}(B^T \times \Sigma_{within} \times B)}{\mathrm{Det}(B^T \times (\Sigma_{within} + \Sigma_{between}) \times B)}$$





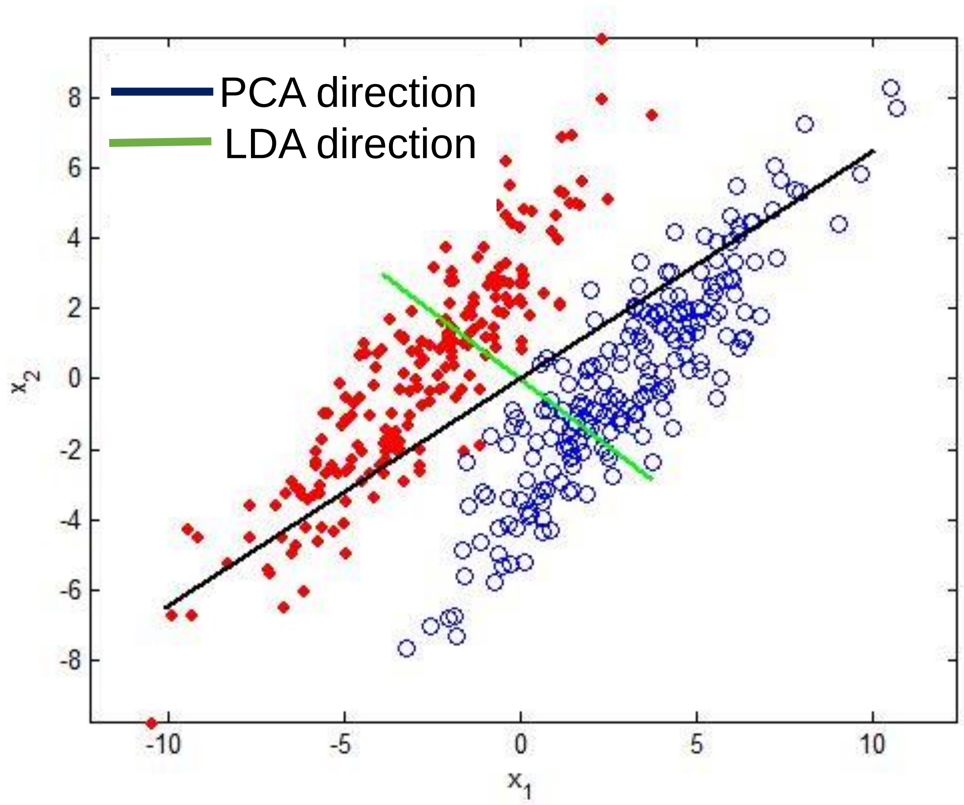
Different directions/projections

The best direction



The best separating space

The best projecting direction


PCA direction
LDA direction
8
6
4
2
0
-2
-4
-6
-8
-10
-5
0
5
10
x_1
x_2