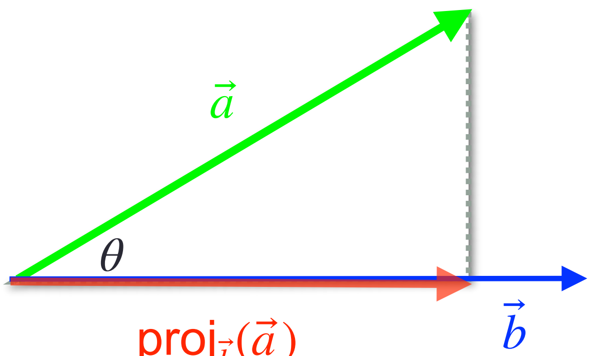
PRINCIPAL COMPONENT ANALYSIS (PCA)

Recap:
Projection of a vector onto another vector



$\text{proj}_{\vec{b}}(\vec{a})$
with length a_1

Right triangle:

$$\cos(\theta) = \frac{a_1}{\|\vec{a}\|} \qquad \Rightarrow \quad a1 = \|\vec{a}\| \cos\theta$$

Dot product of $\vec{a}$ and $\vec{b}$:

$$\vec{a}\cdot\vec{b} = \|\vec{a}\|\,\|\vec{b}\| \cos\theta \qquad \Rightarrow \quad \|\vec{a}\| \cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{b}\|}$$

(Length) $\Rightarrow \quad a_1 = \frac{\vec{a}\cdot\vec{b}}{\|\vec{b}\|} = \frac{a^T b}{\|\vec{b}\|}$

(Vector) $\Rightarrow \quad \text{proj}_{\vec{b}}(\vec{a}) = a_1 \frac{\vec{b}}{\|\vec{b}\|}$

Principal Component Analysis (PCA)

- **Motivation:** Many data sets have the property that the data points all lie close to a manifold of much lower dimensionality than that of the original space.

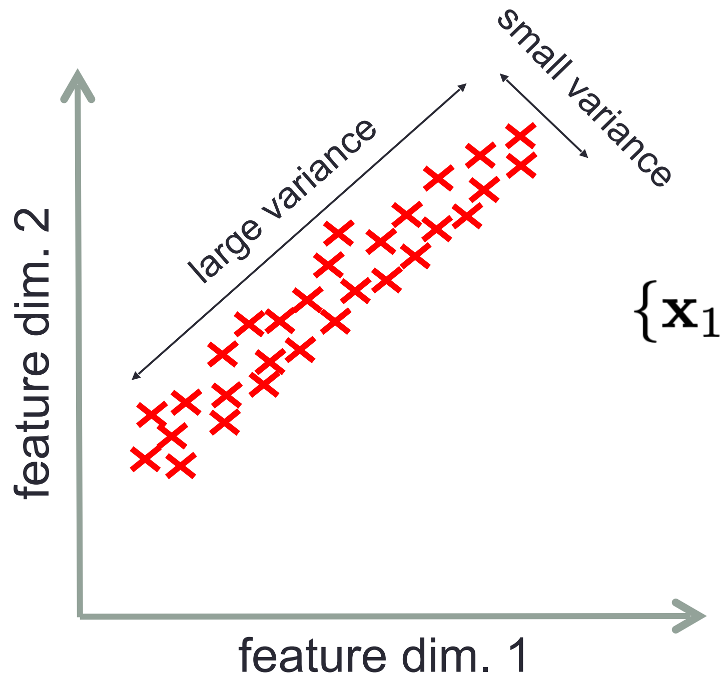
- **Used for:** dimensionality reduction, lossy data compression, feature extraction, data visualization;
to achieve easier class-separability in the case of classification problems

- **Unsupervised** method

Principal Component Analysis (PCA)

- Re-expressing the dataset to extract relevant information embedded in the variance of data samples, to reduce the redundancy.

Fundamental idea: *The useful information content in the data is represented by the variance of data samples.*

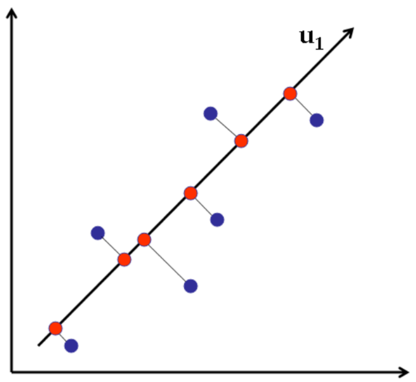


$$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$$

$$\mathbf{x}_n \in \mathbb{R}^2$$

Principal Component Analysis (PCA)

- We are looking for a **projection vector** that accounts for as much of the variability in the data as possible after projection.



$$y_n = \frac{\mathbf{u}_1^T \mathbf{x}_n}{\| \mathbf{u}_1 \|} \quad \text{where} \quad y_n \in \mathbb{R}^M$$

We will constrain the projection vectors to have unit norm:

$$\| \mathbf{u}_1 \| = \sqrt{\mathbf{u}_1^T \mathbf{u}_1} = 1$$

Recall: Statistical Properties of Projected Data

$$\mathbf{x}_n \in \mathbb{R}^D \qquad \boldsymbol{\mu}_x = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \qquad \boldsymbol{\Sigma}_x = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_x)(\mathbf{x}_n - \boldsymbol{\mu}_x)^T$$

$$y_n = \mathbf{u}_1^T\mathbf{x}_n$$

$$y_n \in \mathbb{R}^M$$

Recall: Statistical Properties of Projected Data

$$\mathbf{x}_n \in \mathbb{R}^D \qquad \boldsymbol{\mu}_x = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \qquad \boldsymbol{\Sigma}_x = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_x)(\mathbf{x}_n - \boldsymbol{\mu}_x)^T$$

$$y_n = \mathbf{u}_1^T\mathbf{x}_n$$

$$y_n \in \mathbb{R}^M$$

$$\mu_y = \frac{1}{N}\sum_{n=1}^{N}\mathbf{u}_1^T\mathbf{x}_n = \mathbf{u}_1^T\boldsymbol{\mu}_x,$$

$$\begin{aligned}\sigma_y^2 &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{u}_1^T\mathbf{x}_n - \mathbf{u}_1^T\boldsymbol{\mu}_x)^2 \\ &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{u}_1^T(\mathbf{x}_n - \boldsymbol{\mu}_x))^2 \\ &= \mathbf{u}_1^T\underbrace{\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_x)(\mathbf{x}_n - \boldsymbol{\mu}_x)^T}_{\boldsymbol{\Sigma}_x}\mathbf{u}_1\end{aligned}$$

variance of the transformed data → $\boxed{\sigma_y^2 = \mathbf{u}_1^T\boldsymbol{\Sigma}_x\mathbf{u}_1}$

We want to maximize this quantity as a result of the transformation

PCA for M=1

- **Given** data: $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ where $\mathbf{x}_n \in \mathbb{R}^D$.

- **Goal** is to find a linear transformation of data points onto an M-dimensional subspace:

$$y_n = \mathbf{u}_1^T \mathbf{x}_n \quad \text{where} \quad y_n \in \mathbb{R}^M$$

$$\begin{aligned} \mathbf{u}_1 &= \arg\max_{\mathbf{u}_1} [\sigma_y^2] \\ &= \arg\max_{\mathbf{u}_1} [\mathbf{u}_1^T \boldsymbol{\Sigma}_x \mathbf{u}_1] \end{aligned}$$

under the constraint that $\boldsymbol{u}_1^T \boldsymbol{u}_1 = 1$ in order to avoid $\| \boldsymbol{u}_1 \| \to \infty$

Solving this equality constrained optimization problem with the method of Lagrange multipliers.

PCA for M=1

$$\boldsymbol{u}_1 = \arg\max_{\boldsymbol{u}_1}[\boldsymbol{u}_1^T \boldsymbol{\Sigma}_x \boldsymbol{u}_1] \qquad \text{s.t.} \qquad \boldsymbol{u}_1^T \boldsymbol{u}_1 = 1$$

$$\mathcal{L}(\boldsymbol{u}_1, \lambda_1) = \boldsymbol{u}_1^T \boldsymbol{\Sigma}_x \boldsymbol{u}_1 + \lambda_1 (1 - \boldsymbol{u}_1^T \boldsymbol{u}_1)$$

$$\frac{\partial \mathcal{L}(\boldsymbol{u}_1, \lambda_1)}{\partial \boldsymbol{u}_1} = 2\boldsymbol{\Sigma}_x \boldsymbol{u}_1 - \lambda_1 2\boldsymbol{u}_1 \stackrel{!}{=} 0$$

$$\boxed{\boldsymbol{\Sigma}_x \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1}$$

One can obtain solutions to this by eigendecomposition.

This means that $\boldsymbol{u}_1$ *is an <u>eigenvector</u> of the covariance matrix with the <u>eigenvalue</u>* λ_1

PCA for M=1

$$\boldsymbol{u}_1 = \arg\max_{\boldsymbol{u}_1} [\boldsymbol{u}_1^T \boldsymbol{\Sigma}_x \boldsymbol{u}_1] \qquad \text{s.t.} \qquad \boldsymbol{u}_1^T \boldsymbol{u}_1 = 1$$

$$\mathcal{L}(\boldsymbol{u}_1, \lambda_1) = \boldsymbol{u}_1^T \boldsymbol{\Sigma}_x \boldsymbol{u}_1 + \lambda_1 (1 - \boldsymbol{u}_1^T \boldsymbol{u}_1)$$

$$\frac{\partial \mathcal{L}(\boldsymbol{u}_1, \lambda_1)}{\partial \boldsymbol{u}_1} = 2\boldsymbol{\Sigma}_x \boldsymbol{u}_1 - \lambda_1 2 \boldsymbol{u}_1 \stackrel{!}{=} 0$$

$$\boldsymbol{\Sigma}_x \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1$$

going further..

$$\boxed{\lambda_1 = \boldsymbol{u}_1^T \boldsymbol{\Sigma}_x \boldsymbol{u}_1 = \sigma_y^2}$$

λ_1 corresponds to the variance of the transformed data.

Solution: ***We can obtain largest variance by projecting onto the eigenvector $\boldsymbol{u}_1$ with the largest eigenvalue λ_1***

PCA for M>1

- We can perform the linear transformation to a higher dimension than one (i.e., M>1) using multiple vectors.

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_M \end{pmatrix}}_{\mathbf{y}_n} = \underbrace{\begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_M^T \end{pmatrix}}_{\mathbf{U}^T} \cdot \underbrace{\begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix}}_{\mathbf{x}_n}$$

$$\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_M]$$

$$\| \mathbf{u}_m \| = 1 \quad \forall m$$

$$\mathbf{y}_n \in \mathbb{R}^M$$

PCA for M>1

- Here we solve the eigenvalue/eigenvector problem for:

$$\boxed{\boldsymbol{\Sigma}_x}\mathbf{U} = \mathbf{U}\mathbf{L} \qquad \text{where} \qquad \mathbf{L} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_M \end{pmatrix}$$

eigenvalues of the decomposition

in order to obtain the projection vectors $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_M]$

Solution: *We will select the M eigenvectors of the <u>covariance matrix</u> which correspond to the M largest eigenvalues.*

PCA for M>1

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_M \end{pmatrix}}_{\mathbf{y}_n} = \underbrace{\begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_M^T \end{pmatrix}}_{\mathbf{U}^T} \cdot \underbrace{\begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix}}_{\mathbf{x}_n}$$

- We construct the transformation matrix $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_M]$ by selecting the M eigenvectors of $\mathbf{\Sigma}_x$ which correspond to the M largest eigenvalues.

- These eigenvectors are called *principal components.*
- Covariance matrix of the transformed data $\mathbf{y}_n$ becomes a diagonal matrix:

$$\mathbf{\Sigma}_y = \mathbf{L}$$

$$\mathbf{L} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_M \end{pmatrix}$$

In this subspace the features are now *decorrelated*.

Whitening

- An extension of PCA by modifying the projections:

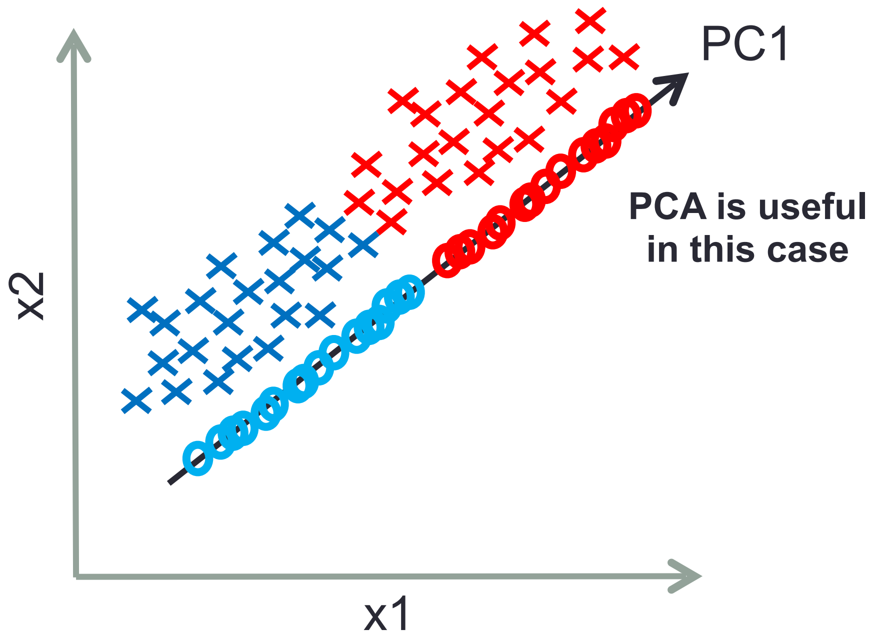
$$\mathbf{y}_n = \mathbf{L}^{-\frac{1}{2}}\mathbf{U}^T(\mathbf{x}_n - \boldsymbol{\mu}_x)$$

1. Data is mean centered, i.e., zero mean.
2. Features are decorrelated (as in PCA).
3. Features have the same variance (spherical covariance matrix).

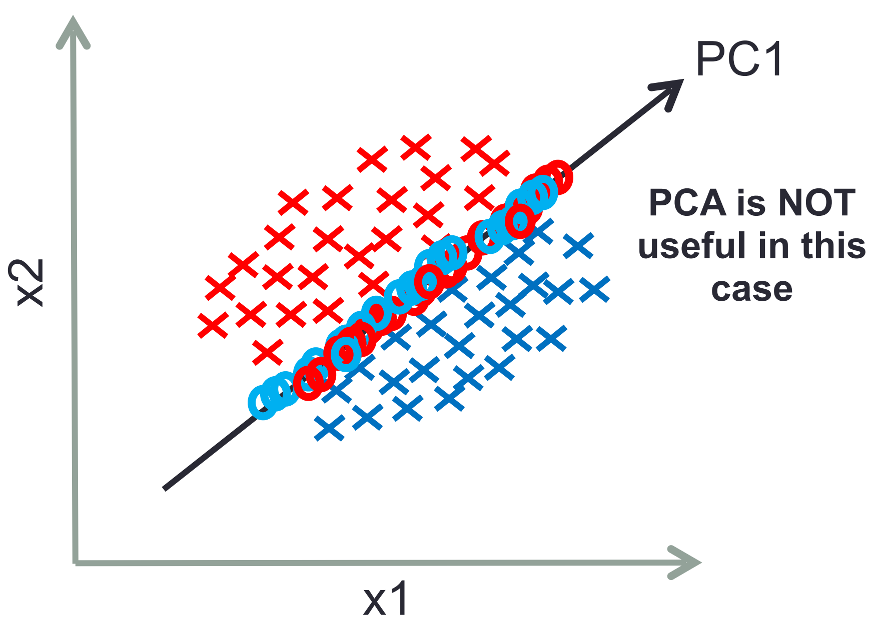
$$\begin{aligned}
\boldsymbol{\Sigma}_y &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n\mathbf{y}_n^T \\
&= \mathbf{L}^{-\frac{1}{2}}\mathbf{U}^T \underbrace{\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \mathbf{m}_x)(\mathbf{x}_n - \mathbf{m}_x)^T}_{\boldsymbol{\Sigma}_x} \mathbf{U}\mathbf{L}^{-\frac{1}{2}} \\
&= \mathbf{L}^{-\frac{1}{2}} \underbrace{\mathbf{U}^T\boldsymbol{\Sigma}_x\mathbf{U}}_{\mathbf{L}} \mathbf{L}^{-\frac{1}{2}} = \mathbf{I}
\end{aligned}$$

Example: A Classification Problem

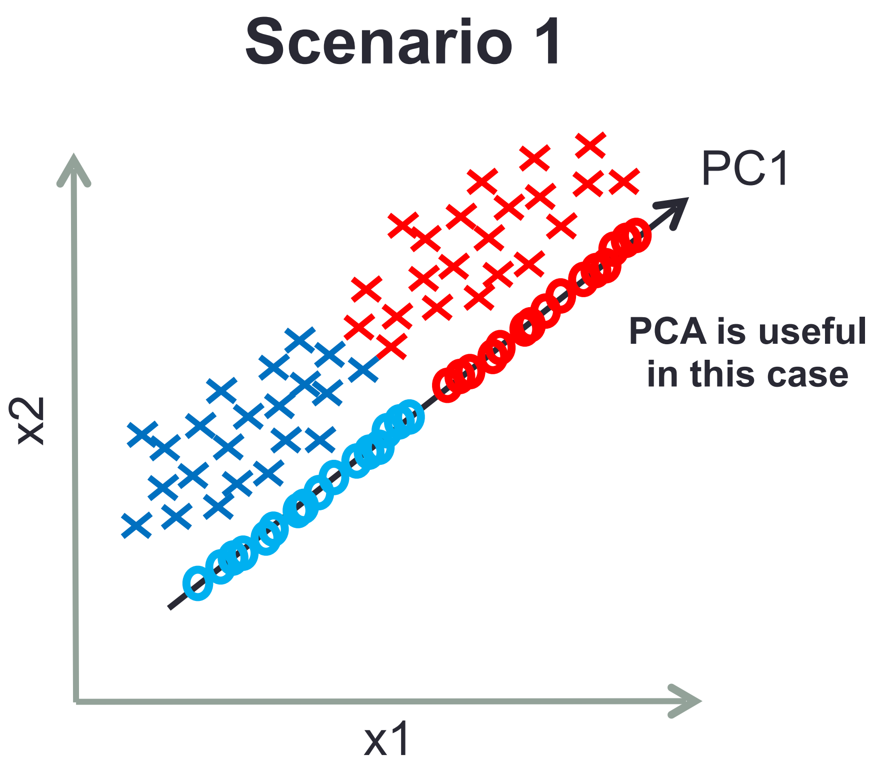
Scenario 1

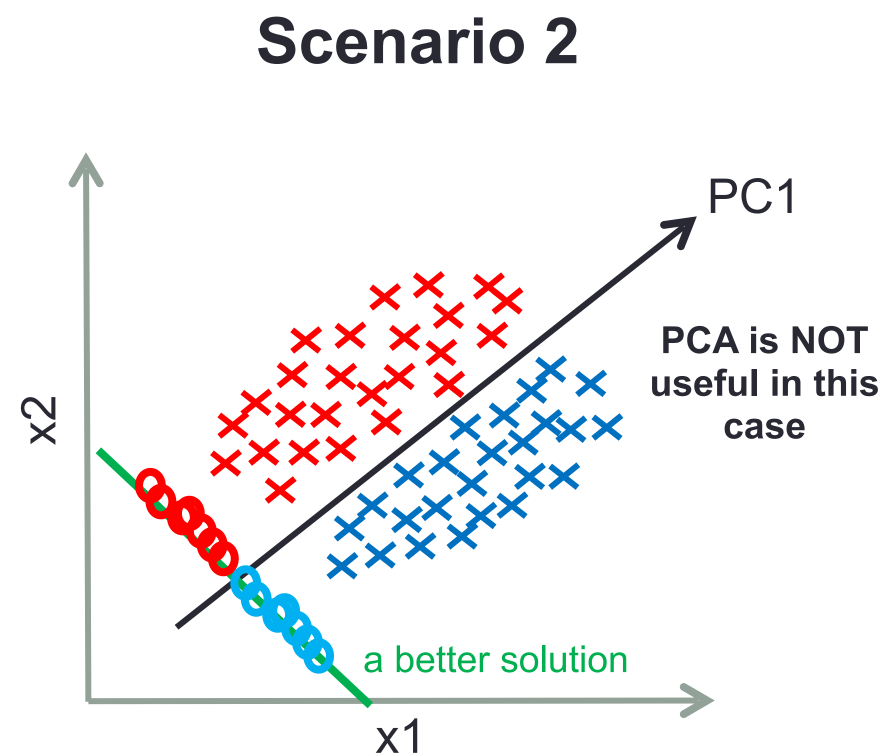


Scenario 2



Example: A Classification Problem





PCA does not use and loses class relevant information.
We can consider Linear Discriminant Analysis (LDA) to learn linear transformations for better class-separability.

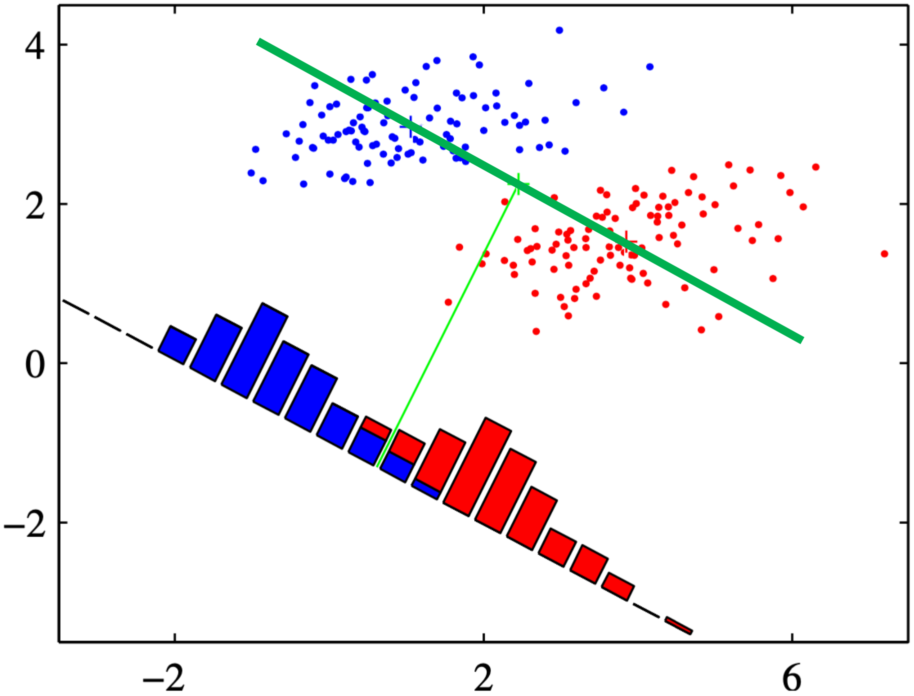
Linear Discriminant Analysis (LDA)

- Supervised method
- Transformation of data to increase class separability by considering labels

- Case I: maximizing the distance between class means after projection.
- Case II (Fisher LDA): maximizing distance between projected class means while within-class covariance of projected data is as small as possible.

Illustrating the two cases

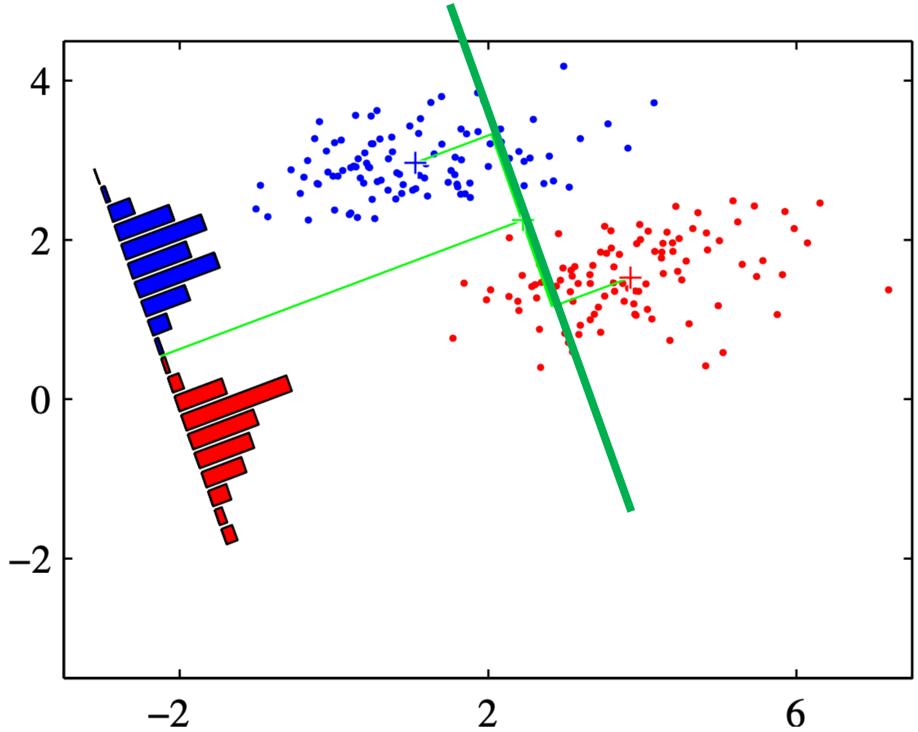
Case I: Maximizing the distance between class means



$$\mathbf{u} \propto (\boldsymbol{\mu}_{x,2} - \boldsymbol{\mu}_{x,1})$$

$\mathbf{u}$ is in the direction of the line joining the class means

Case II: Fisher LDA



$$\mathbf{u} \propto \boldsymbol{\Sigma}_w^{-1}(\boldsymbol{\mu}_{x,2} - \boldsymbol{\mu}_{x,1})$$