Principal Component Analysis

Lecturer:
Jiří Matas

Authors:
Ondřej Drbohlav, Jiří Matas

Centre for Machine Perception
Czech Technical University, Prague
http://cmp.felk.cvut.cz

1.1.2017

- Alternative name: Karhunene Loeve transform
- Used for: data approximation, identifying sources of variance in the data

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

Let the data be $\{\mathbf{x}_i \mid i = 1, 2, ..., N\}$, with sample mean $\overline{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n$.

Let us find the unit vector $\mathbf{u}_1$ to project to such that the variance $J(\mathbf{u}_1)$ of the projected data is *maximized*. The projection $\mathbf{x}_n^{(\mathrm{p})}$ of an $\mathbf{x}_n$ to one-dimensional subspace generated by $\mathbf{u}_1$ is given by

$$\mathbf{x}_n^{(\mathrm{p})} = \mathbf{u}_1 \left(\mathbf{u}_1^{\mathrm{T}} \mathbf{x}_n\right), \quad \mathbf{u}_1^{\mathrm{T}} \mathbf{u}_1 = 1 \,. \tag{1}$$

The variance $J(\mathbf{u}_1)$ of projected data is

$$J(\mathbf{u}_1) = \frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_n - \mathbf{u}_1^{\mathrm{T}}\overline{\mathbf{x}}\right)^2 = \frac{1}{N}\sum_{n=1}^{N} \mathbf{u}_1^{\mathrm{T}}(\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^{\mathrm{T}}\mathbf{u}_1 = \mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 \,, \tag{2}$$

where $\mathbf{S}$ is the normalized scatter matrix:

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^{\mathrm{T}} \,. \tag{3}$$

The Lagrangian of this optimization problem is

$$L(\mathbf{u}_1, \lambda_1) = J(\mathbf{u}_1) + \lambda_1 \underbrace{(1 - \mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1)}_{\text{constraint}} = \mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1)\,, \tag{4}$$

where λ_1 is the Lagrange multiplier. Taking the derivative w.r.t. the vector $\mathbf{u}_1$ and setting it to zero gives

$$\frac{\partial L(\mathbf{u}_1, \lambda_1)}{\partial \mathbf{u}_1} = \mathbf{S}\mathbf{u}_1 - \lambda_1\mathbf{u}_1 = 0\,, \tag{5}$$

and thus

$$\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1\,. \tag{6}$$

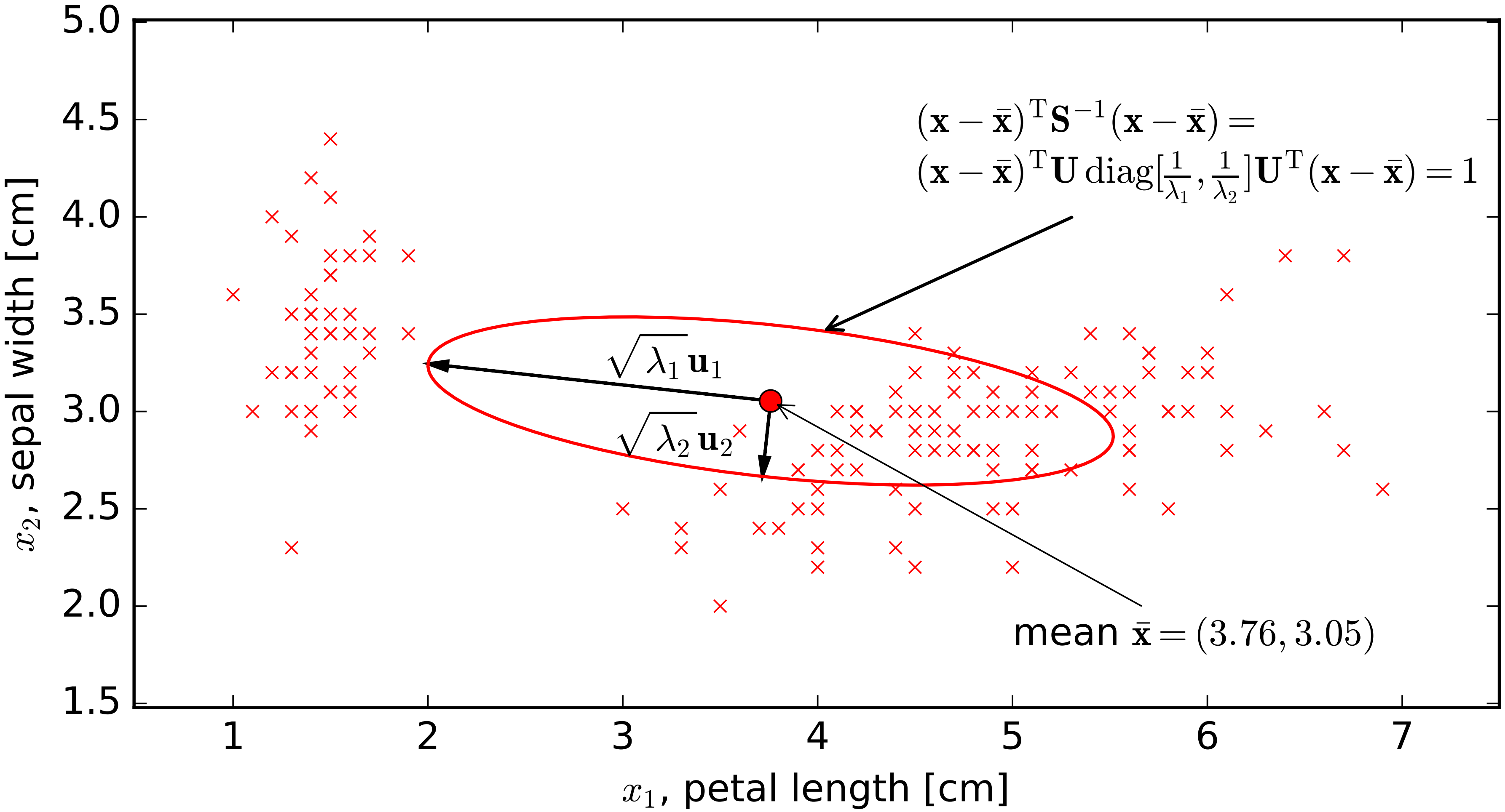
This is the characteristic equation for the covariance matrix $\mathbf{S}$. Any eigenvalue λ_1 and its corresponding eigenvector $\mathbf{v}_1$ solves this equation, with variance $J(\mathbf{u}_1)$ equal to:

$$J(\mathbf{u}_1) = \mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 = \mathbf{u}_1^{\mathrm{T}}\lambda_1\mathbf{u}_1 = \lambda_1\,. \tag{7}$$

The maximum is attained if λ_1 is the largest eigenvalue of the matrix $\mathbf{S}$ and $\mathbf{u}_1$ is its corresponding eigenvector.

Example 1 - Iris dataset

Iris dataset: feature vectors are 4-dimensional, here dimensions 2 and 3 used (petal length and sepal width). Data shown as crosses ×.



$$\mathbf{S} = \begin{bmatrix} 3.09 & -0.32 \\ -0.32 & 0.19 \end{bmatrix} = [\mathbf{u_1}, \mathbf{u_2}] \begin{bmatrix} \lambda_1 & \\ & \lambda_2 \end{bmatrix} [\mathbf{u_1}, \mathbf{u_2}]^{\mathrm{T}}$$

Eigenvectors: $[\mathbf{u}_1, \mathbf{u}_2] = \begin{bmatrix} -0.99 & -0.11 \\ 0.11 & -0.99 \end{bmatrix}$, eigenvalues: $\lambda_1 = 3.13$, $\lambda_2 = 0.15$

Variance is maximized when data are projected to direction $\mathbf{u}_1$.

1 2
3 4
5 6
7 8
9 10
11 12
13 14
15 16
17 18
19 20
21 22
23 24
25 26
27 28
29 30

Maximum variance formulation (3/3)

Recall: The variance of a 1-D projection is maximized when data are projected to the direction of the eigenvector of $\mathbf{S}$ corresponding to the largest eigenvalue.

$\mathbf{S}$ is symmetric and positive semidefinite. The eigenvectors corresponding to different eigenvalues are orthogonal.

It follows that the D-dimensional subspace maximizing the variance of the data is the one formed by D eigenvectors of $\mathbf{S}$ corresponding the the D largest eigenvalues.

Note: "Variance" in the above sentence is the sum of variances in individual orthogonal directions. For a 2-D subspace,

$$J(\mathbf{u}_1, \mathbf{u}_2) = \frac{1}{N} \sum_{n=1}^{N} [\mathbf{u}_1^{\mathrm{T}}(\mathbf{x}_n - \overline{\mathbf{x}})]^2 + [\mathbf{u}_2^{\mathrm{T}}(\mathbf{x}_n - \overline{\mathbf{x}})]^2 \,. \tag{8}$$

Equivalence to Minimum error formulation (1/2)

Consider the complete orthogonal basis $\{\mathbf{u}_i\}$ where $i = 1, \ldots, D$. Thus

$$\mathbf{u}_i^{\mathrm{T}} \mathbf{u}_j = \delta_{ij} \tag{9}$$

Each point can be represented as

$$\mathbf{x}_n = \sum_{i=1}^{D} \alpha_{ni} \mathbf{u}_i \,, \tag{10}$$

and

$$\mathbf{x}_n = \sum_{i=1}^{D} (\mathbf{x}_n^{\mathrm{T}} \mathbf{u}_i) \mathbf{u}_i \,. \tag{11}$$

This is just expressing $\mathbf{x}_n$ in a rotated coordinate system given by orthonormal system $\{\mathbf{u}_i\}$. Let us create an approximation to each $\mathbf{x}_n$ by truncating this expansion to only M components, the remaining $D - M$ components approximated by constants b_i. The approximation $\tilde{\mathbf{x}}_n$:

$$\tilde{\mathbf{x}}_n = \sum_{i=1}^{M} (\mathbf{x}_n^{\mathrm{T}} \mathbf{u}_i) \mathbf{u}_i + \sum_{i=M+1}^{D} b_i \mathbf{u}_i \tag{12}$$

$$\tilde{\mathbf{x}}_n = \sum_{i=1}^{M} (\mathbf{x}_n^{\mathrm{T}} \mathbf{u}_i) \mathbf{u}_i + \sum_{i=M+1}^{D} b_i \mathbf{u}_i \tag{12}$$

Clearly,

$$b_i = \overline{\mathbf{x}}^{\mathrm{T}} \mathbf{u}_i, i = M+1, \ldots, D \tag{13}$$

The task is to find the optimal orthonormal basis $\{\mathbf{u}_i\}$ which produces the best approximation measured by

$$J(\{\mathbf{u}_i\}) = \frac{1}{N} \sum_{n=1}^{N} \|\mathbf{x}_n - \tilde{\mathbf{x}}_n\|^2 \tag{14}$$

The minimum error criterion is the complement of the maximum variance criterion, and thus the solution to the set $\{\mathbf{u}_i\}$ is the same.

Recall that the ML estimate of the Multivariate Normal Distribution is defined by sample mean $\overline{\mathbf{x}}$ and sample covariance matrix $\mathbf{S}$. The model is

$$p(\mathbf{x} \mid \overline{\mathbf{x}}, \mathbf{S}) = \frac{1}{\sqrt{|2\pi\mathbf{S}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}-\overline{\mathbf{x}})\right\} \tag{15}$$

Denote stacked eigenvectors in descending order of their eigenvalues as $\mathbf{U}$,

$$\mathbf{U} = \{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_D\} \tag{16}$$

Therefore (characteristic equation)

$$\mathbf{S}\mathbf{U} = \mathbf{U}\boldsymbol{\Lambda} = \mathbf{U}\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_D \end{bmatrix}, \tag{17}$$

and

$$\mathbf{S} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}. \tag{18}$$

We approximate the data, as before, by projecting to first M eigenvectors. Thus, given data point $\mathbf{x}$ we have

$$\mathbf{x} - \overline{\mathbf{x}} = (\delta_1, \delta_2, ..., \delta_M, \delta_{M+1}, ..., \delta_D) \tag{19}$$

Note that we only can compute δ_1 .. δ_M, as often we don't or can't store all eigenvectors for computing all δ's. However, we can easily compute

$$\Delta = \delta_{M+1}^2 + \delta_{M+2}^2 + ... + \delta_D^2 = \|\mathbf{x} - \overline{\mathbf{x}}\|^2 - \delta_1^2 - \delta_2^2 - ... - \delta_M^2 \tag{20}$$

and the exponent is then approximated as

$$-\frac{1}{2}(\mathbf{x} - \overline{\mathbf{x}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x} - \overline{\mathbf{x}}) \simeq -\frac{1}{2}\left(\frac{\delta_1^2}{\lambda_1} + \frac{\delta_2^2}{\lambda_2} + \frac{\delta_3^2}{\lambda_3} + ... \frac{\delta_M^2}{\lambda_M} + \frac{\Delta}{\lambda}\right) \tag{21}$$

Common choice: $\lambda = \lambda_{M+1}$

High-dimensional data (1/2)

Dimensionality of data can be high, and even higher than number of samples.

Consider dimensionality $D = 1\mathrm{M}$ (one million) and number of samples $N = 100$. All analysis still applies, but it would be wasteful to compute eigenvectors for the 1Mx1M matrix, as its rank will anyway be at most N (thus 100). Let us define $\mathbf{X}$ to be a matrix formed by stacking all the data vectors (after having subtracted the mean from them): $\mathbf{X} = [\mathbf{x}_1 - \overline{\mathbf{x}}, \mathbf{x}_2 - \overline{\mathbf{x}}, ..., \mathbf{x}_N - \overline{\mathbf{x}}]$.

Thus,

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^{\mathrm{T}} = \frac{1}{N}\mathbf{X}\mathbf{X}^{\mathrm{T}}\,. \tag{22}$$

The characteristic equation is then

$$\frac{1}{N}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{u} = \lambda\mathbf{u}\,. \tag{23}$$

Left-multiplying both sides by $\mathbf{X}^{\mathrm{T}}$ gives

$$\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{X}\overbrace{(\mathbf{X}^{\mathrm{T}}\mathbf{u})}^{\mathbf{w}} = \lambda\overbrace{(\mathbf{X}^{\mathrm{T}}\mathbf{u})}^{\mathbf{w}}\,. \tag{24}$$

Thus, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, which is only 100×100, has exactly the same set of eigenvalues:

$$\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} = \lambda\mathbf{w}\,. \tag{25}$$

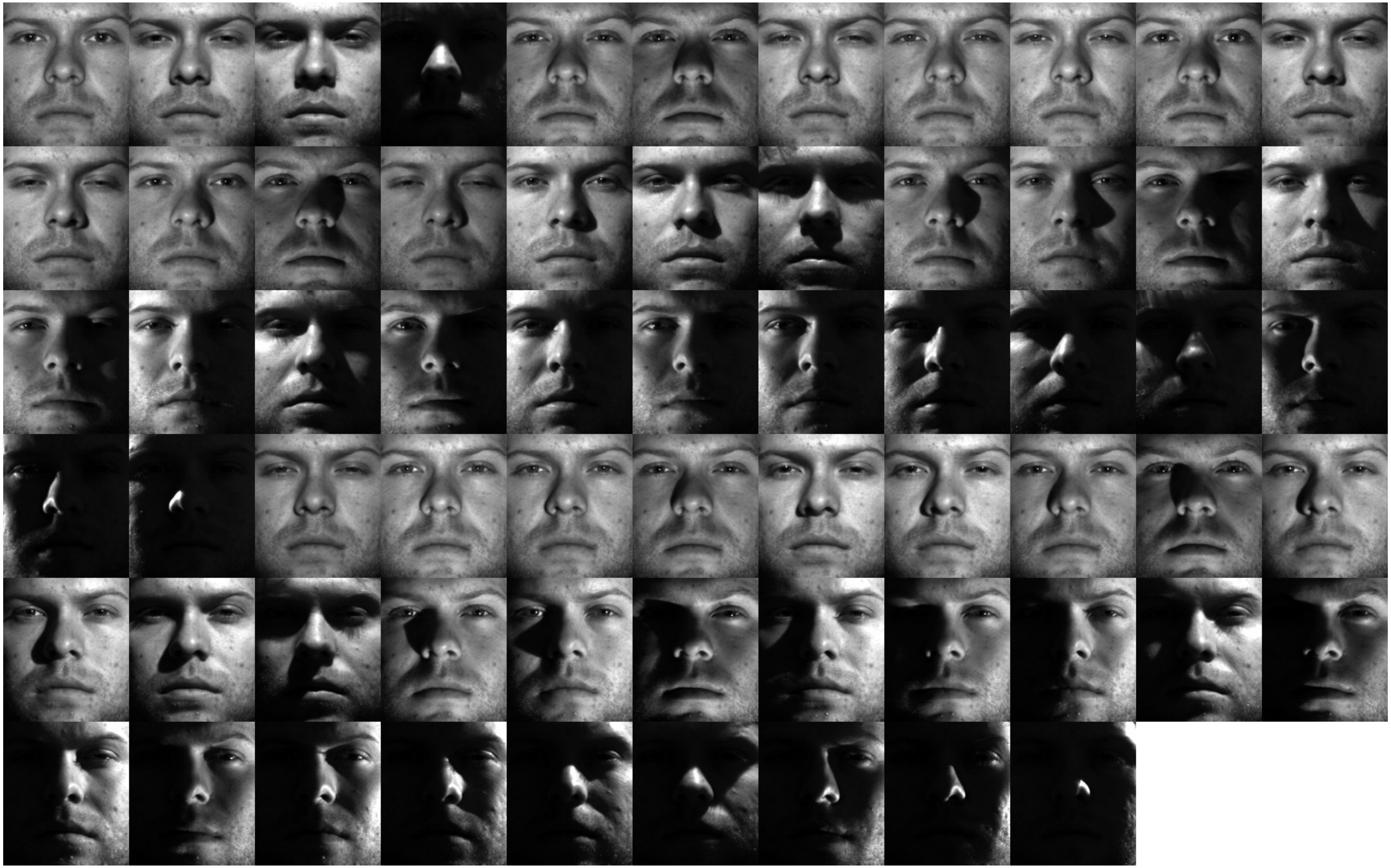
Left-multiplying now by $\mathbf{X}$, we get

$$\frac{1}{N}\mathbf{X}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{w}) = \lambda(\mathbf{X}\mathbf{w})\,. \tag{26}$$

Conclusion: If $D \gg N$, form the matrix $\mathbf{T} = \frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and compute its eigenvalues λ's and eigenvectors $\mathbf{w}$. Compute the eigenvectors of $\mathbf{S} = \frac{1}{N}\mathbf{X}\mathbf{X}^{\mathrm{T}}$ as

$$\mathbf{v} = \frac{\mathbf{X}\mathbf{w}}{\|\mathbf{X}\mathbf{w}\|}\,. \tag{27}$$

images of 38 subjects, each under 64 different illumination conditions:



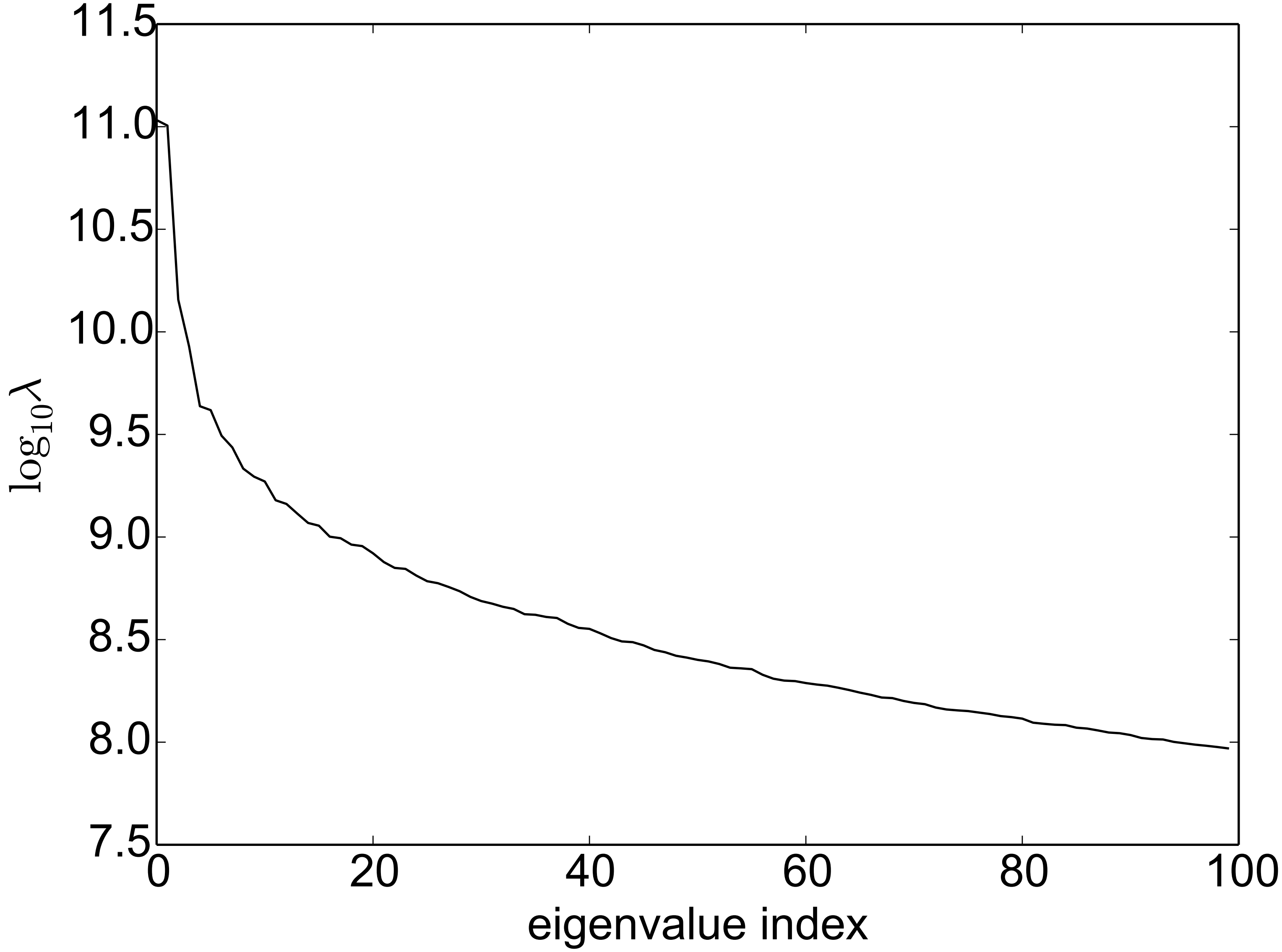
Subject 1, 64 illumination conditions

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

images of 38 subjects, each under 64 different illumination conditions:

38 subjects

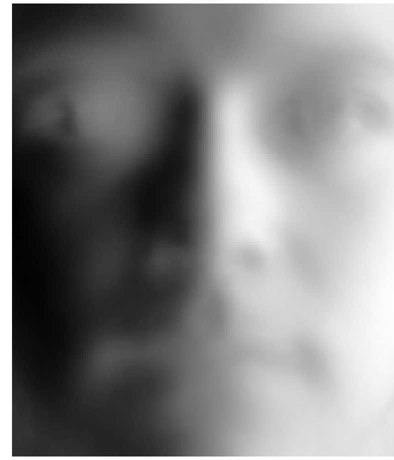
images of 38 subjects, each under 64 different illumination conditions. Thus, there is $38 \times 64 = 2432$ images in total. Each of them is a feature vector with $192 \times 168 = 32256$ dimensions (pixels). PCA gives the following eigenvalues:



Example 2 - Yale database (4/5)

mean

1st ev



2nd ev

3rd ev

first 72 eigenvectors

Example 2 - Yale database (5/5)

Reconstruction of original vector using eigenvectors

original

mean and 3 evs

mean and 10 evs

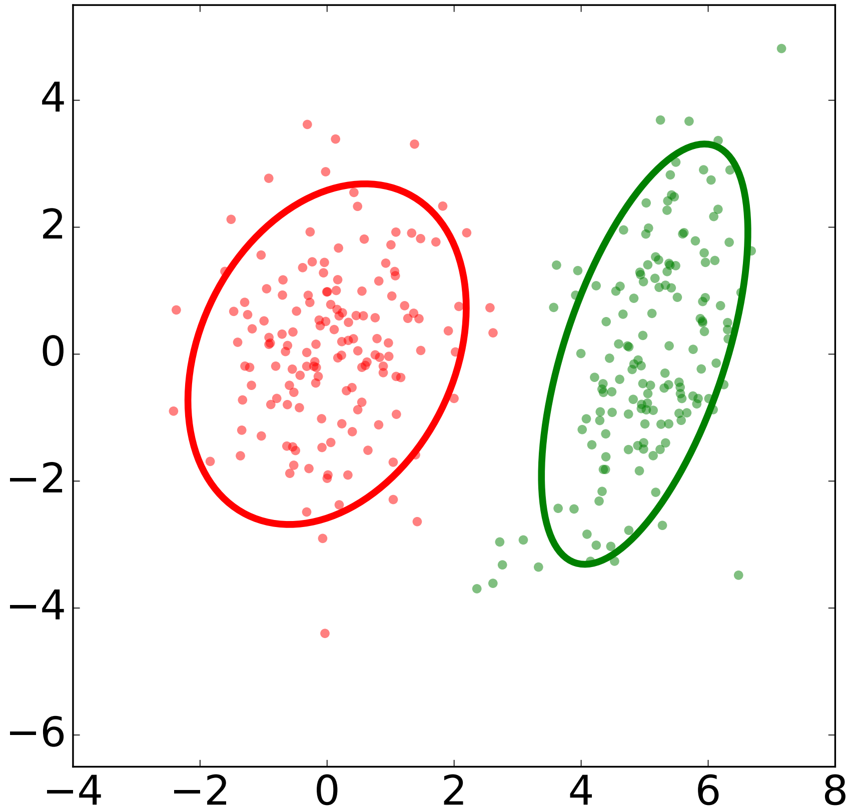
mean and 50 evs

mean and 100 evs

mean and 300 evs

Setting: Classification, training set: N_1 points (class 1) and N_2 points (class 2)
Goal: Project data to a 1D subspace such that a low-error classifier can be constructed.
Approach: Find a direction to project the data to such that the two classes are well separated in this projection.

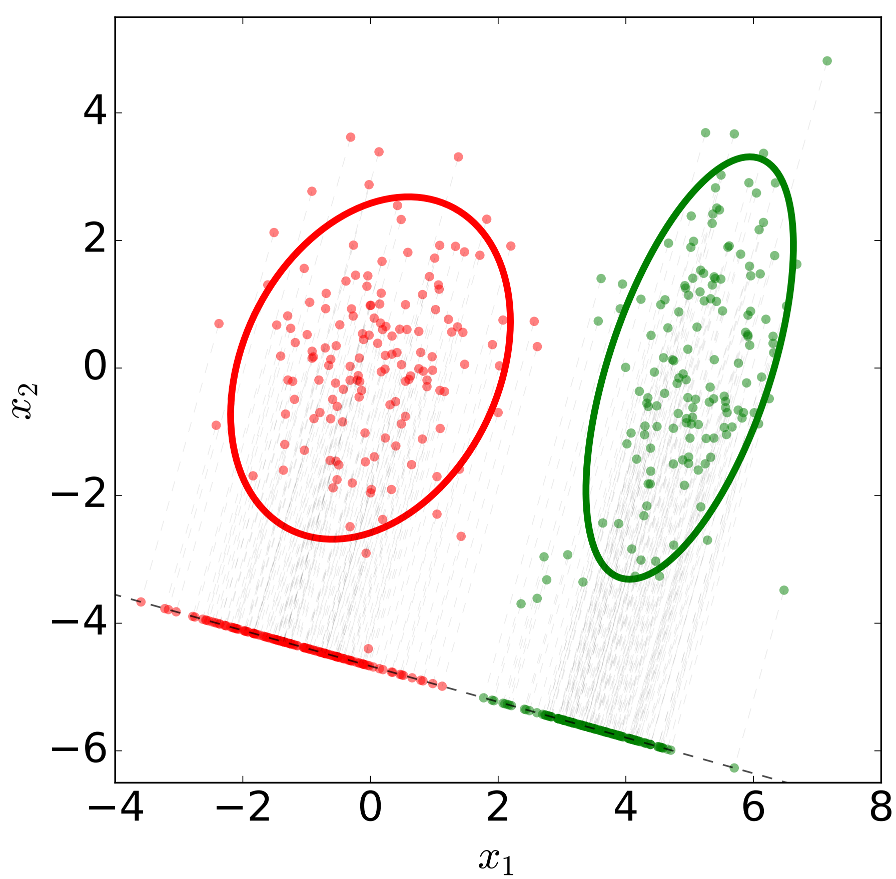
Example: Data as shown

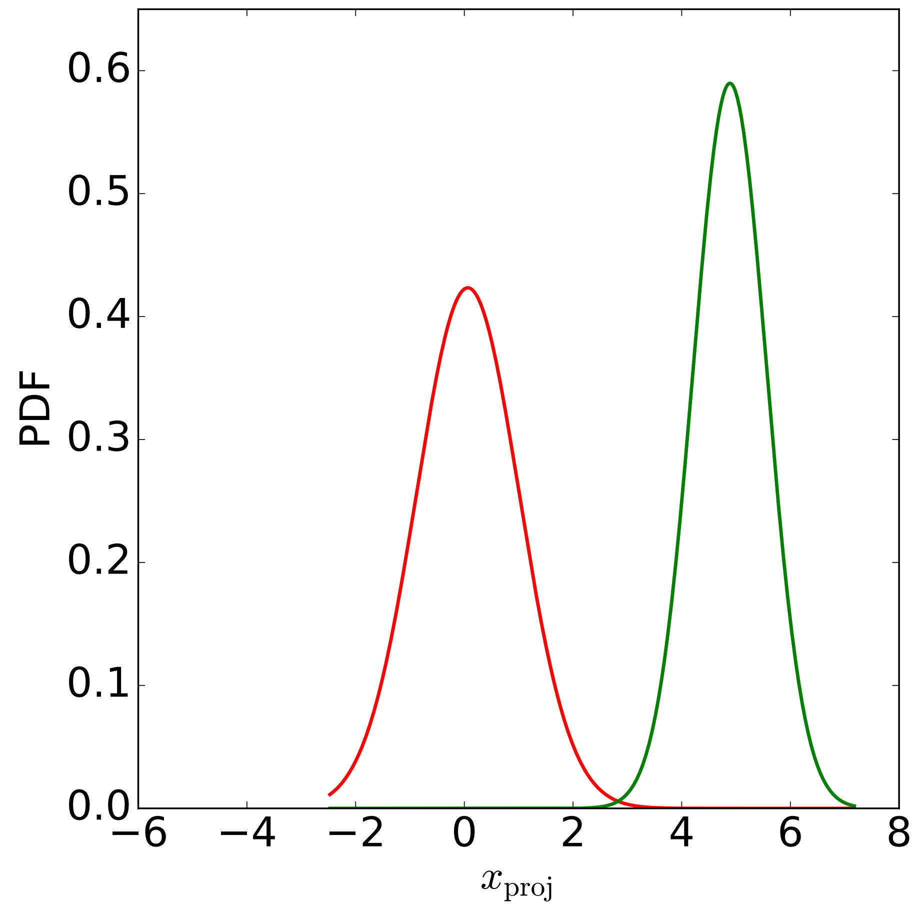


m p 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

Setting: Classification, training set: N_1 points (class 1) and N_2 points (class 2)
Goal: Project data to a 1D subspace such that a low-error classifier can be constructed.
Approach: Find a direction to project the data to such that the two classes are well separated in this projection.

Example: Projection direction producing good separation

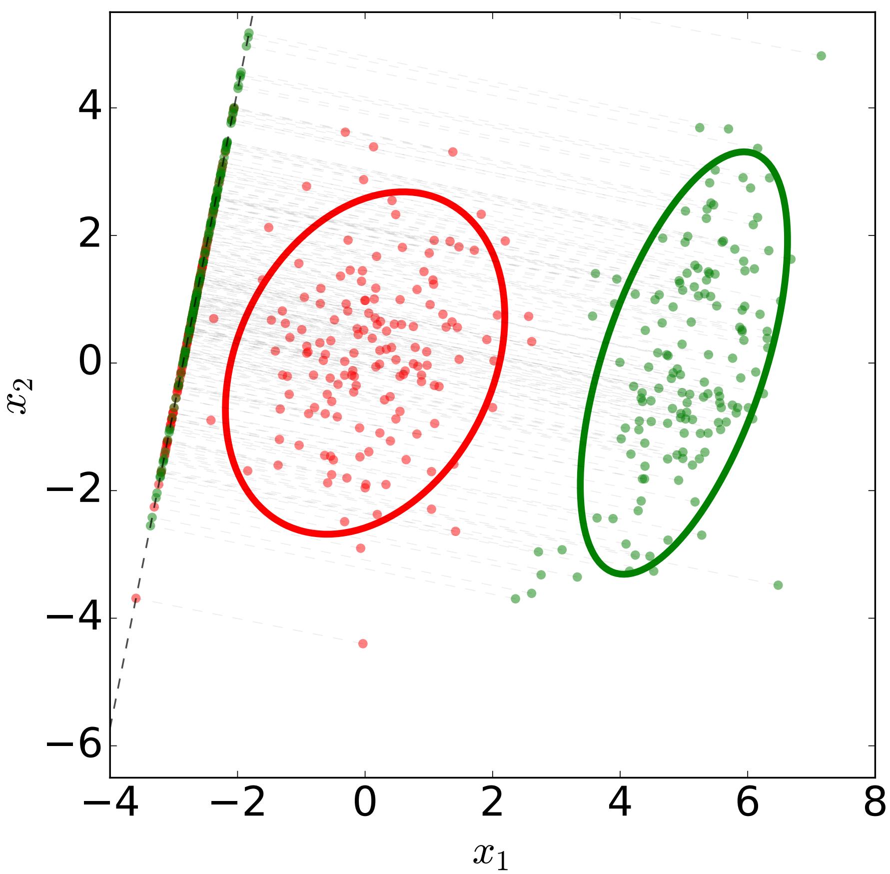


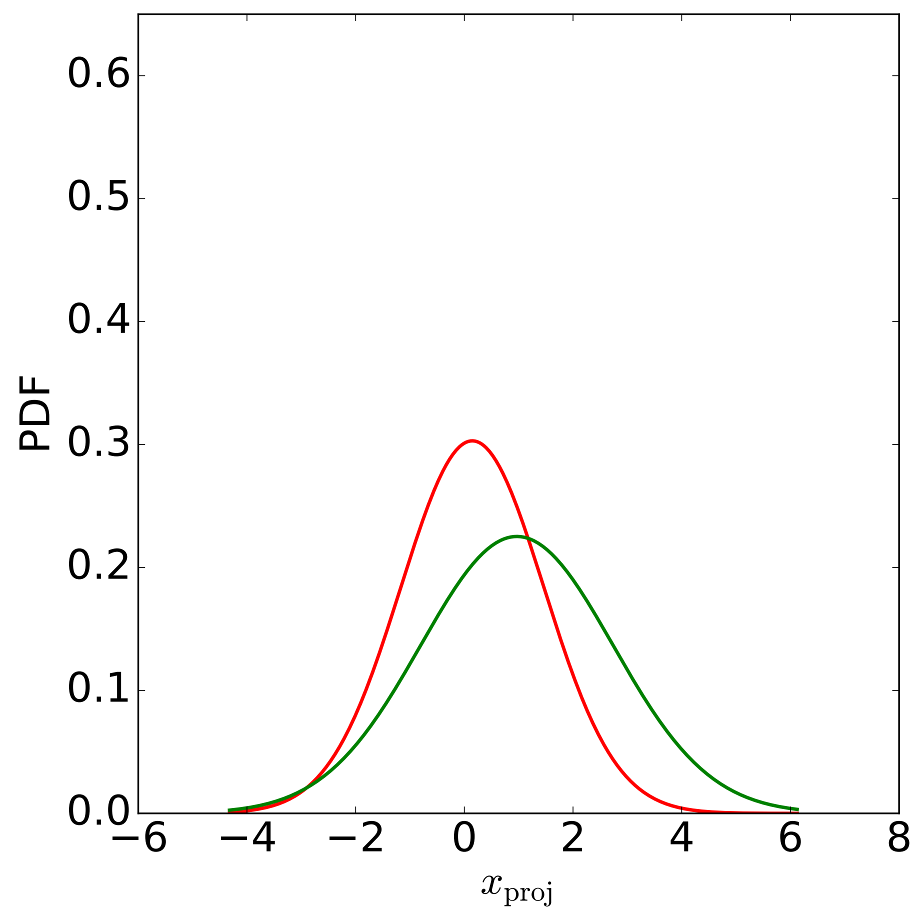


Linear Discriminant Analysis (LDA)

Setting: Classification, training set: N_1 points (class 1) and N_2 points (class 2)
Goal: Project data to a 1D subspace such that a low-error classifier can be constructed.
Approach: Find a direction to project the data to such that the two classes are well separated in this projection.

Example: Projection direction producing bad separation





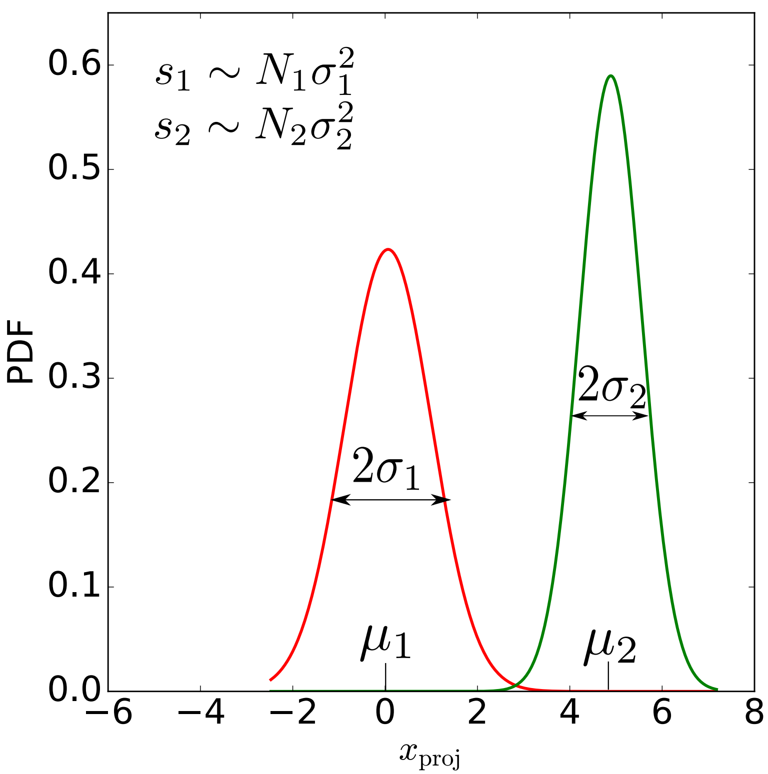
LDA: What makes a good separation?

Training set: $\mathbf{x}_1^1, \ldots \mathbf{x}_{N_1}^1$ (class 1), $\mathbf{x}_1^2, \ldots \mathbf{x}_{N_2}^2$ (class 2).
Separation is higher when:

- the means of projected data are farther apart, and/or
- the scatters of the projected data are smaller.

These two observations combined suggest the following criterion to optimize:

$$\frac{(\mu_1 - \mu_2)^2}{s_1 + s_2} \rightarrow \text{max} \tag{28}$$



μ_1, μ_2: mean of projected data

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbf{v}^{\mathrm{T}} \mathbf{x}_i^k \qquad (k = 1, 2) \tag{29}$$

s_1, s_2: scatter of projected data

$$s_k = \sum_{i=1}^{N_k} (\mathbf{v}^{\mathrm{T}} \mathbf{x}_i^k - \mu_k)^2 \qquad (k = 1, 2) \tag{30}$$

$$\frac{(\mu_1 - \mu_2)^2}{s_1 + s_2} \rightarrow \text{max}, \quad \mu_k = \mathbf{v}^{\mathrm{T}}\overline{\mathbf{x}}_k, \quad s_k = \sum_{i=1}^{N_k}(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i^k - \mu_k)^2 \ (k = 1, 2) \tag{31}$$

Let us rewrite the criterion in terms of unprojected entities. The nominator:

$$(\mu_1 - \mu_2)^2 = [\mathbf{v}^{\mathrm{T}}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)]^2 = \mathbf{v}^{\mathrm{T}} \underbrace{(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^{\mathrm{T}}}_{\mathbf{S}_b} \mathbf{v} \tag{32}$$

The scatters:

$$s_1 = \sum_{i=1}^{N_1}(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i - \mathbf{v}^{\mathrm{T}}\overline{\mathbf{x}}_1)^2 = \sum_{i=1}^{N_1}\mathbf{v}^{\mathrm{T}}(\mathbf{x}_i - \overline{\mathbf{x}}_1)(\mathbf{x}_i - \overline{\mathbf{x}}_1)^{\mathrm{T}}\mathbf{v} \tag{33}$$

$$= \mathbf{v}^{\mathrm{T}} \underbrace{\left(\sum_{i=1}^{N_1}(\mathbf{x}_i - \overline{\mathbf{x}}_1)(\mathbf{x}_i - \overline{\mathbf{x}}_1)^{\mathrm{T}}\right)}_{\mathbf{S}_1} \mathbf{v} \tag{34}$$

$$s_2 = \mathbf{v}^{\mathrm{T}}\mathbf{S}_2\mathbf{v} \qquad \mathbf{S}_1, \mathbf{S}_2 : \text{scatter matrices for classes 1, 2} \tag{35}$$

$$\frac{(\mu_1 - \mu_2)^2}{s_1 + s_2} \to \text{max}, \quad \mu_k = \mathbf{v}^{\mathrm{T}}\overline{\mathbf{x}}_k, \quad s_k = \sum_{i=1}^{N_k} (\mathbf{v}^{\mathrm{T}}\mathbf{x}_i^k - \mu_k)^2 \;\; (k = 1, 2) \tag{36}$$

Therefore, the criterion can be rewritten as

$$\frac{(\mu_1 - \mu_2)^2}{s_1 + s_2} = \frac{\mathbf{v}^{\mathrm{T}}\mathbf{S}_b\mathbf{v}}{\mathbf{v}^{\mathrm{T}}(\mathbf{S}_1 + \mathbf{S}_2)\mathbf{v}} = \frac{\mathbf{v}^{\mathrm{T}}\mathbf{S}_b\mathbf{v}}{\mathbf{v}^{\mathrm{T}}\mathbf{S}_w\mathbf{v}}, \tag{37}$$

where everything except the to-be-found vector $\mathbf{v}$ is computed from the training data:

$$\mathbf{S}_b : \text{between-class scatter matrix, } \mathbf{S}_b = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^{\mathrm{T}} \tag{38}$$

$$\mathbf{S}_w : \text{within-class scatter matrix, } \mathbf{S}_w = \mathbf{S}_1 + \mathbf{S}_2 \tag{39}$$

$$\mathbf{S}_k = \sum_{i=1}^{N_k} (\mathbf{x}_i^k - \overline{\mathbf{x}}_k)(\mathbf{x}_i^k - \overline{\mathbf{x}}_k)^{\mathrm{T}}, \qquad (k = 1, 2) \tag{40}$$

Let us now solve the maximization task:

$$\mathbf{v}_1 = \underset{\mathbf{v}}{\text{argmax}} \frac{\mathbf{v}^{\mathrm{T}}\mathbf{S}_b\mathbf{v}}{\mathbf{v}^{\mathrm{T}}\mathbf{S}_w\mathbf{v}} \tag{41}$$

Note that there is no need to contrain $\mathbf{v}$ to e.g. unit length, as the scaling in denominator and nominator cancels out.

$$\mathbf{v}_1 = \underset{\mathbf{v}}{\text{argmax}} \frac{\mathbf{v}^\mathrm{T}\mathbf{S}_b\mathbf{v}}{\mathbf{v}^\mathrm{T}\mathbf{S}_w\mathbf{v}} \tag{42}$$

Note that $\mathbf{S}_b$ is symmetric, positive semi-definite (rank 1) matrix.
Matrix $\mathbf{S}_w$ is symmetric, positive semi-definite.

Assume that $\mathbf{S}_w$ has full rank, thus $\mathbf{S}_w^{-1}$ exists. Let $\mathbf{S}_w^{\frac{1}{2}}$ be the symmetric, positive-definite matrix such that $\mathbf{S}_w = \mathbf{S}_w^{\frac{1}{2}}\mathbf{S}_w^{\frac{1}{2}}$. Let its inverse be denoted $\mathbf{S}_w^{-\frac{1}{2}}$.
Define a substitution

$$\mathbf{z} = \mathbf{S}_w^{\frac{1}{2}}\mathbf{v}\,. \tag{43}$$

Symmetric, positive definite $\mathbf{S}$:
$\mathbf{S} = \mathbf{U}\,\text{diag}[\lambda_1, ..., \lambda_D]\,\mathbf{U}^\mathrm{T}$
$\mathbf{U}$: orthogonal, unit columns
$\mathbf{S}^{\frac{1}{2}} = \mathbf{U}\,\text{diag}[\sqrt{\lambda_1}, ..., \sqrt{\lambda_D}]\,\mathbf{U}^\mathrm{T}$
$\mathbf{S}^{-\frac{1}{2}} = \mathbf{U}\,\text{diag}[\frac{1}{\sqrt{\lambda_1}}, .., \frac{1}{\sqrt{\lambda_D}}]\,\mathbf{U}^\mathrm{T}$
$\mathbf{S}^{-1} = \mathbf{U}\,\text{diag}[\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_D}]\,\mathbf{U}^\mathrm{T}$

Using the variable $\mathbf{z}$, the criterion becomes

$$\frac{\mathbf{v}^\mathrm{T}\mathbf{S}_b\mathbf{v}}{\mathbf{v}^\mathrm{T}\mathbf{S}_w\mathbf{v}} = \frac{\mathbf{z}^\mathrm{T}\mathbf{S}_w^{-\frac{1}{2}}\mathbf{S}_b\mathbf{S}_w^{-\frac{1}{2}}\mathbf{z}}{\mathbf{z}^\mathrm{T}\mathbf{z}} \tag{44}$$

Let us fix the length of $\mathbf{z}$ to 1 ($\mathbf{z}^\mathrm{T}\mathbf{z} = 1$). The denomimator is then a constant, and the criterion is maximized when the nominator is maximized. The latter achieves maximum for the largest eigenvalue λ_1 of matrix $\mathbf{S}_w^{-\frac{1}{2}}\mathbf{S}_b\mathbf{S}_w^{-\frac{1}{2}}$ and the corresponding eigenvector $\mathbf{z}_1$:

$$\mathbf{S}_w^{-\frac{1}{2}}\mathbf{S}_b\mathbf{S}_w^{-\frac{1}{2}}\mathbf{z}_1 = \lambda_1\mathbf{z}_1 \tag{45}$$

(copied from previous slide:)

$$\mathbf{S}_w^{-\frac{1}{2}}\mathbf{S}_b\mathbf{S}_w^{-\frac{1}{2}}\mathbf{z}_1 = \lambda_1\mathbf{z}_1 \tag{46}$$

Taking this $\mathbf{z}_1$, and substituting back, gives the solution $\mathbf{v}_1 = \mathbf{S}_w^{-\frac{1}{2}}\mathbf{z}_1$. But left-multiplying the previous equation by $\mathbf{S}_w^{-\frac{1}{2}}$, we see that

$$\mathbf{S}_w^{-1}\mathbf{S}_b(\mathbf{S}_w^{-\frac{1}{2}}\mathbf{z}_1) = \lambda_1(\mathbf{S}_w^{-\frac{1}{2}}\mathbf{z}_1), \quad \Rightarrow \quad \mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{v}_1 = \lambda_1\mathbf{v}_1\,. \tag{47}$$

Thus $\mathbf{v}_1$ can be computed directly as the eigenvector of $\mathbf{S}_w^{-1}\mathbf{S}_b$ corresponding to the highest eigenvalue, λ_1 (note that $\mathbf{S}_w^{-1}\mathbf{S}_b$ and $\mathbf{S}_w^{-\frac{1}{2}}\mathbf{S}_b\mathbf{S}_w^{-\frac{1}{2}}$ share the eigenvalues).

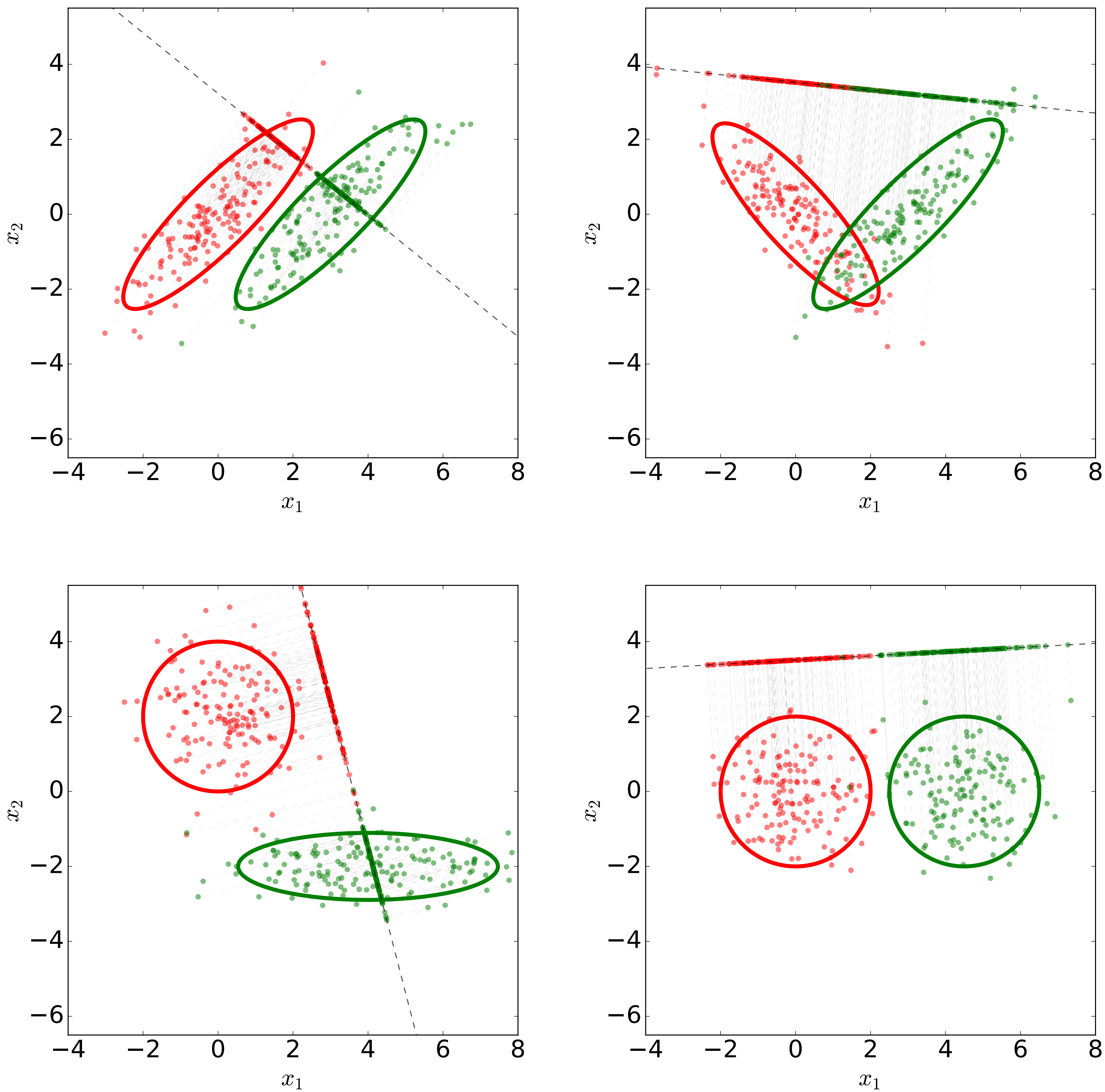
Moreover, $\mathbf{S}_b$ has rank 1. There holds $\mathbf{S}_b = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^{\mathrm{T}}$, and

$$\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{z} = {\mathbf{S}_w}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)\underbrace{(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^{\mathrm{T}}\mathbf{z}}_{\text{a scalar}}, \tag{48}$$

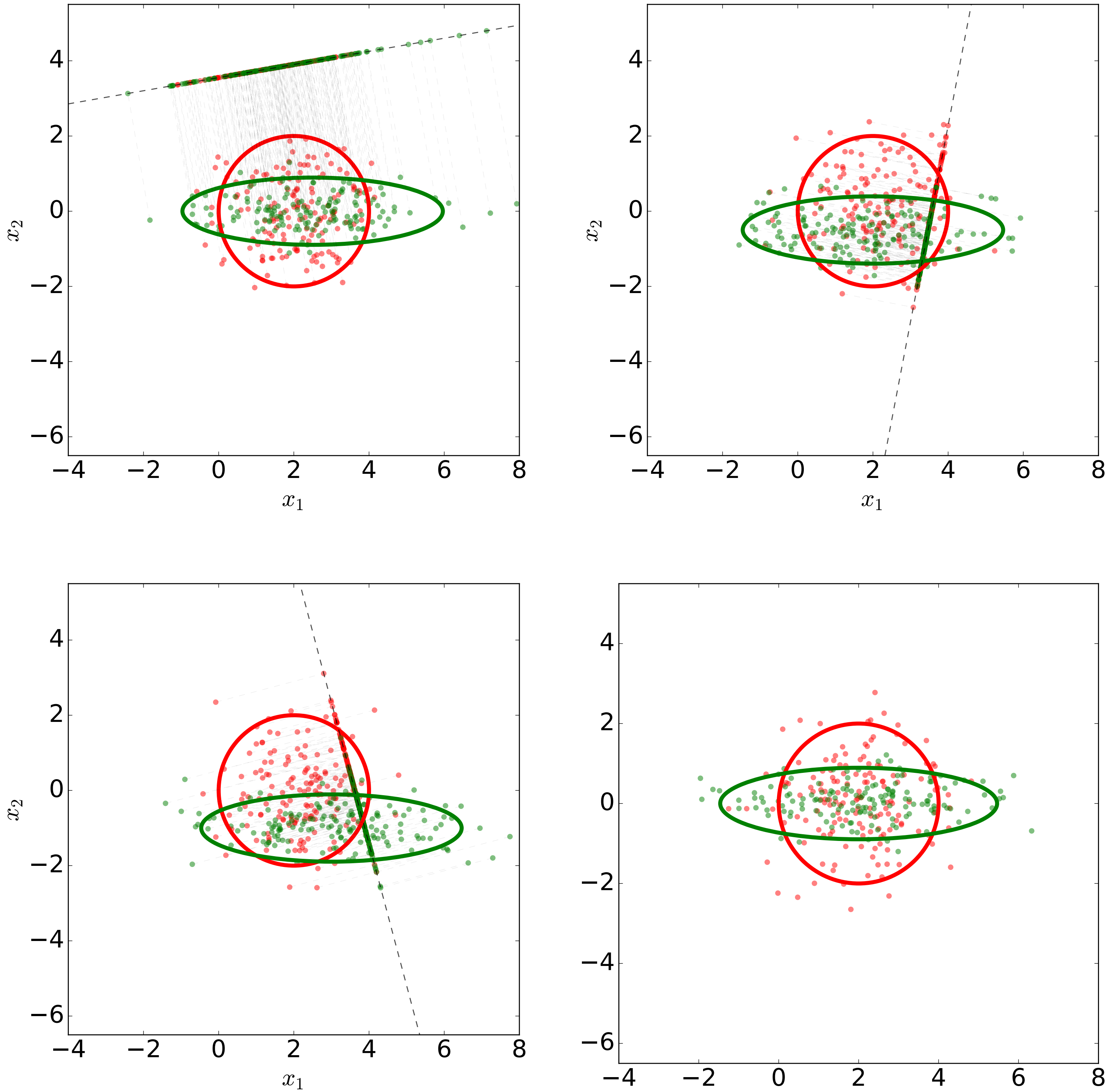
thus the dominant eigenvector (the only one with non-zero eigenvalue) must be

$$\boxed{\mathbf{v}_1 = \mathbf{S}_w^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)\,.} \tag{49}$$

LDA: Examples (1)



LDA: Examples (2)



LDA: Invariance to linear transformations

Consider the case that the data points $\mathbf{x}$'s are transformed by a non-singular linear transformation $\mathbf{A}$. The entities appearing in formulation and solution of LDA are then transformed as follows:

	points	scatter matrix	inv. scatter m.
original	$\mathbf{x}$	$\mathbf{S}$	$\mathbf{S}^{-1}$
transformed	$\mathbf{A}\mathbf{x}$	$\mathbf{A}\mathbf{S}\mathbf{A}^{\mathrm{T}}$	$\mathbf{A}^{-\mathrm{T}}\mathbf{S}^{-1}\mathbf{A}^{-1}$

Thus, $\mathbf{v}_1 = \mathbf{S}_w^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$ transforms to

$$\mathbf{v}_1' = \mathbf{A}^{-\mathrm{T}}\mathbf{S}_w^{-1}\mathbf{A}^{-1}\mathbf{A}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) = \mathbf{A}^{-\mathrm{T}}\mathbf{S}_w^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)\,. \tag{50}$$

The original projected coordinates are

$$\mathbf{v}_1^{\mathrm{T}}\mathbf{x} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^{\mathrm{T}}\mathbf{S}_w^{-1}\mathbf{x}\,, \tag{51}$$

and do not change under $\mathbf{A}$, as

$$\mathbf{v'}_1^{\mathrm{T}}\mathbf{x}' = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^{\mathrm{T}}\mathbf{S}_w^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{x} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^{\mathrm{T}}\mathbf{S}_w^{-1}\mathbf{x} = \mathbf{v}_1^{\mathrm{T}}\mathbf{x}\,. \tag{52}$$

1 2
3 4
5 6
7 8
9 10
11 12
13 14
15 16
17 18
19 20
21 22
23 24
25 26
27 28
29 30

Multiple Discriminant Analysis (MDA)

Generalization of LDA to multiple classes K

Define:

$$\mathbf{S}_w = \sum_{i=1}^{K} \mathbf{S}_i \quad \text{(sum of class scatters)} \tag{53}$$

$$\mathbf{S}_b = \sum_{i=1}^{K} N_i (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})^{\mathrm{T}} \tag{54}$$

$$\overline{\mathbf{x}}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbf{x}_i^k \quad \text{(mean of class } k \text{ data)} \tag{55}$$

$$\overline{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \quad \text{(mean of all data)} \tag{56}$$

Goal: find matrix $\mathbf{V}$ stacking $L < K$ vectors such that

$$\frac{\det(\mathbf{V}^{\mathrm{T}} \mathbf{S}_b \mathbf{V})}{\det(\mathbf{V}^{\mathrm{T}} \mathbf{S}_w \mathbf{V})} \to \max \tag{57}$$

Solution: L most significant eigenvectors for the generalized eigenvalue problem:

$$\mathbf{S}_b\mathbf{v} = \lambda\mathbf{S}_w\mathbf{v} \tag{58}$$

Note: $\mathbf{S}_b$ can have rank at most $K-1$, thus at most $K-1$ projection directions will be produced.

Employing MDA:

Useful e.g. when the number of classes K and/or number of data is very high and thus the only information about data which can be used is stored in means and scatters of classes. These are computed in incremental fashion.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30