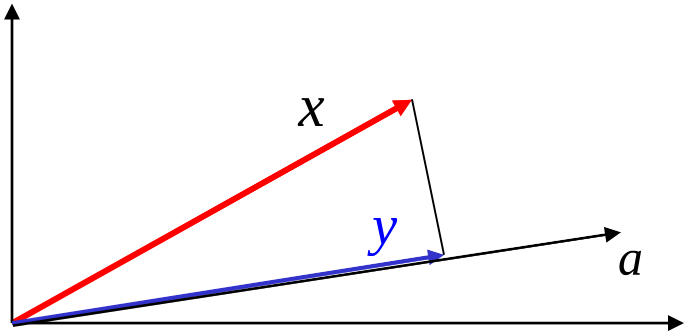
APPLIED MACHINE LEARNING

Principal Component Analysis (PCA)

Part III - Derivation

Constructing a projection

Problem: project x onto a

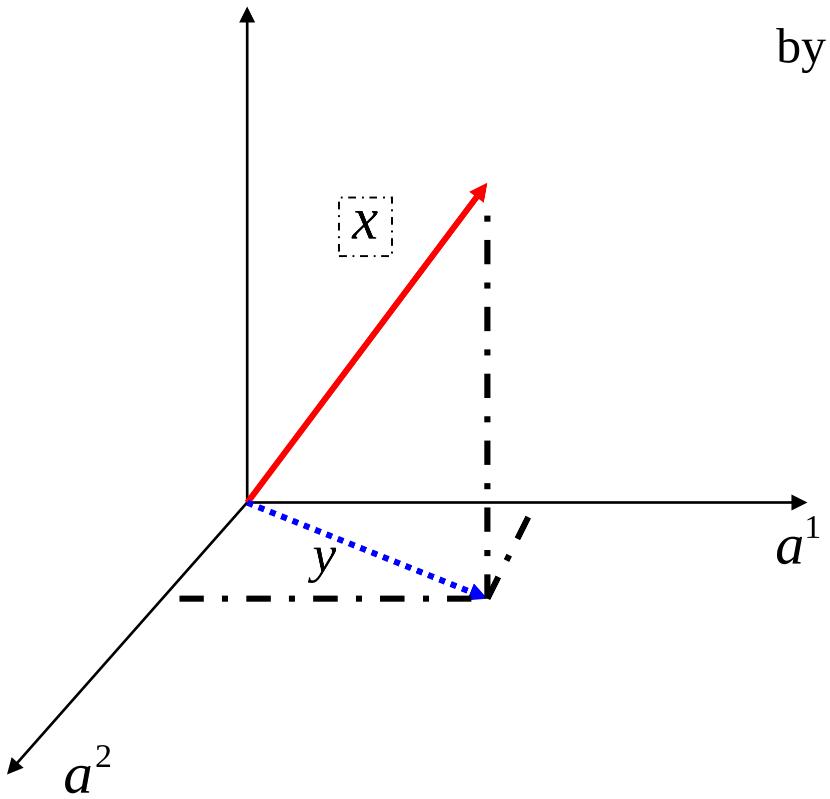


Projection vector is: a

y, the projection of x onto a is:

$$y = a^T \cdot x \frac{a}{\|a\|^2}$$

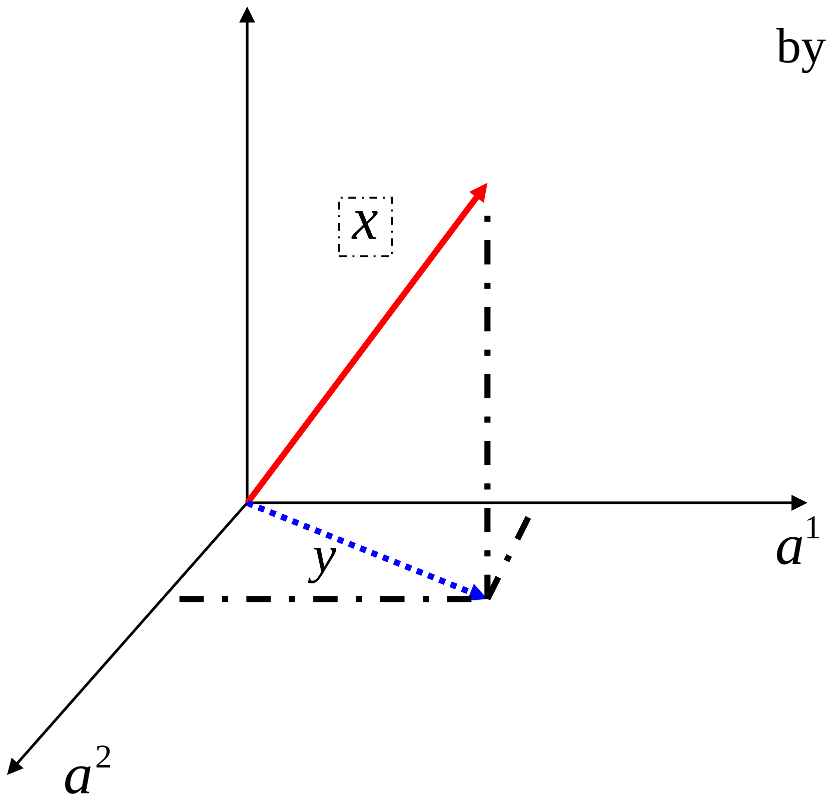
Constructing a projection



The projection y of x onto the plane formed by $\left(a^1, a^2\right)$, with $\left(a^1\right)^T a^2 = 0,$ is given by:

$$y = \underbrace{\frac{\left(a^1\right)^T \cdot x}{\left\|a^1\right\|^2}}_{\text{coordinate of y onto } a^1} a^1 + \underbrace{\frac{\left(a^2\right)^T \cdot x}{\left\|a^2\right\|^2}}_{\text{coordinate of y onto } a^2} a^2$$

Constructing a projection



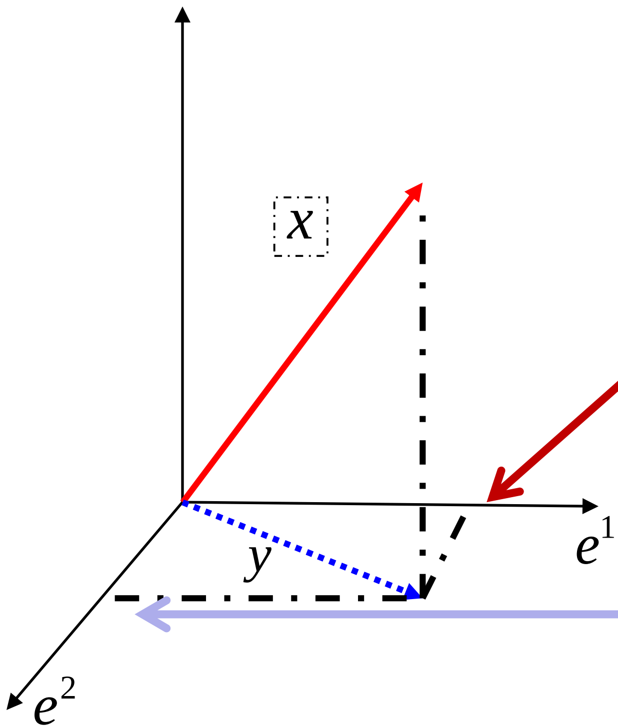
The projection y of x onto the plane formed by $\left(a^1, a^2\right)$, with $\left(a^1\right)^T a^2 = 0$, is given by:

$$y = \underbrace{\frac{\left(a^1\right)^T \cdot x}{\left\|a^1\right\|^2}}_{\text{coordinate of y onto } a^1} a^1 + \underbrace{\frac{\left(a^2\right)^T \cdot x}{\left\|a^2\right\|^2}}_{\text{coordinate of y onto } a^2} a^2$$

Normalize the projection vectors

$$e^i = \frac{a^i}{\left\|a^i\right\|}, \; i = 1, 2$$

Constructing a projection



The projection y of x onto the plane formed by $\left(e^1, e^2\right)$, with $\left(e^1\right)^T e^2 = 0$, is given by:

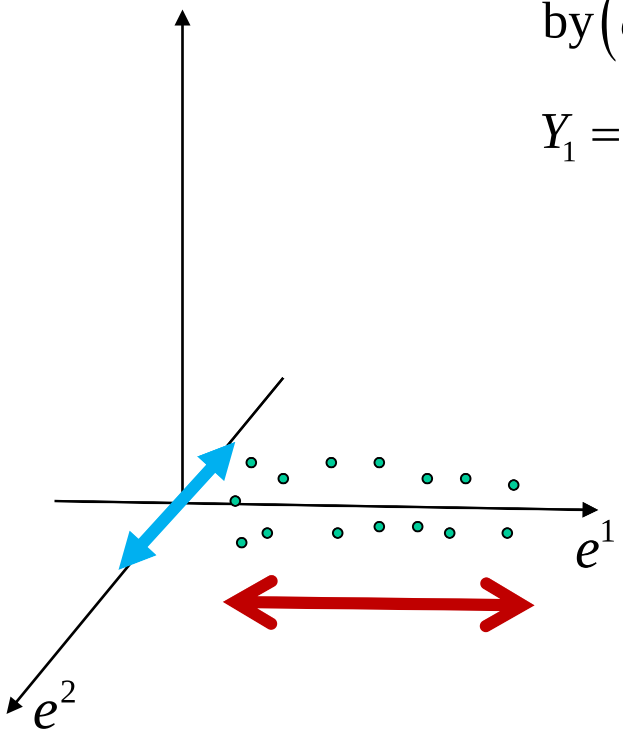
$$y = \underbrace{\left(e^1\right)^T \cdot x}_{\text{coordinate of y onto } e^1} e^1 + \underbrace{\left(e^2\right)^T \cdot x}_{\text{coordinate of y onto } e^2} e^2$$

Constructing a projection

The projection Y of dataset X onto the plane formed by $\left(e^1, e^2\right)$, with $\left(e^1\right)^T e^2 = 0$, is given by:

$$Y_1 = \underbrace{\left(e^1\right)^T \cdot X}_{\text{Norm measures amount of spread of Y onto } e^1} \; ; \; Y_2 = \underbrace{\left(e^2\right)^T \cdot X}_{\text{Norm measures amount of spread of Y onto } e^2}$$

Y_i: i-th row of Y



Finding the optimal projection

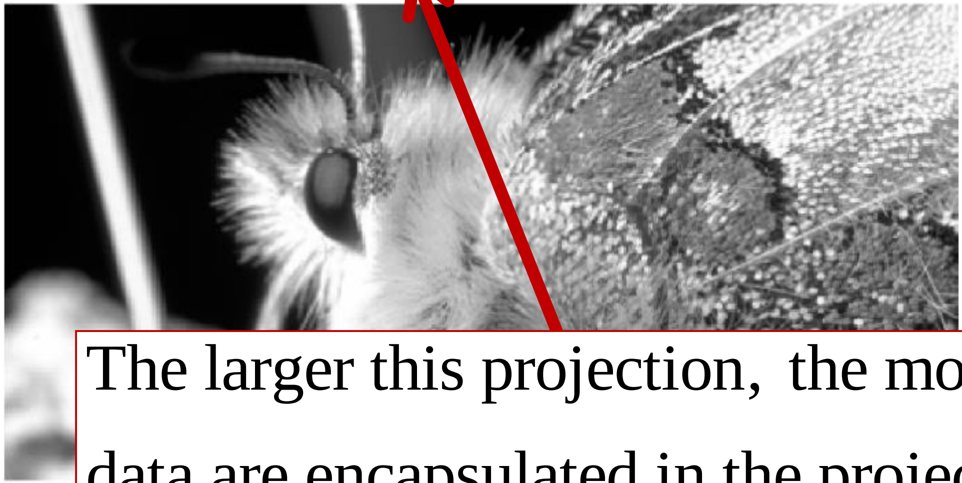
Each image is encoded in $x \in \mathbb{R}^N$.

1. Compute A but ask $A \in \mathbb{R}^{N \times N}$!

2. Project the image in $y = Ax$.

$$A = \begin{bmatrix} (e^1)^T \\ (e^2)^T \\ \cdot \\ \cdot \end{bmatrix}$$

$$\Rightarrow y = \sum_{i=1}^{N} \left(\left(e^i\right)^T x \right) e^i$$



The larger this projection, the more features in the data are encapsulated in the projection e^i.

Low values $=$ noise $\rightarrow$ can be discarded

Finding the optimal projection

Each image is encoded in $x \in \mathbb{R}^N$.

1. Compute A but ask $A \in \mathbb{R}^{N \times N}$!
2. Project the image in $y = Ax$.

$$\Rightarrow y = \sum_{i=1}^{N} \left(\left(e^i\right)^T x \right) e^i$$

Remove rows of A with smallest projections $\left(e^i\right)^T x$.

$$\Rightarrow y = \sum_{i=1}^{p} \left(\left(e^i\right)^T x \right) e^i, \quad p < N.$$

The smaller p, the more compression

Original Image

Image compressed

Finding the optimal projection

Original image is encoded in $x \in \mathbb{R}^N$.

Compressed image is $y \in \mathbb{R}^p$

$y = A_p x$, with $p = 0.1N$

A_p contains p lines of A

Original Image

Image compressed 90%

PCA as constrained-based optimization

A ensures minimal reconstruction error
- keep statistics
- minimal loss of information

Find p lines of A such that

$$\min_A \left\| A^{-1} y^* - x \right\|$$

Least-square approximation for reconstruction $y^* = \begin{bmatrix} y_{1:p} \\ 0_{N-p} \end{bmatrix}$

Requests that all projection vectors are orthonormal.

$$A = \begin{bmatrix} \left(e^1\right)^T \\ \left(e^2\right)^T \\ . \\ . \end{bmatrix} \quad \text{with} \begin{cases} \left\| e^i \right\| = 1, \ \forall i \\ \left(e^i\right)^T e_j = 0, \ \forall i \neq j \end{cases}$$

Reconstruction through error minimization

$$\min_{e^{p+1},\dots,e^{N}} \left\| \sum_{i=p+1}^{N} \left(\left(e^{i}\right)^{T} x \right) e^{i} \right\|$$

Since all projections are orthogonal $\left(e^{i}\right)^{T} e^{j} = 0,\ i \neq j$

$$\Rightarrow \min_{e^{p+1},\dots,e^{N}} \left\| \sum_{i=p+1}^{N} \left(\left(e^{i}\right)^{T} x \right) e^{i} \right\| = \min_{e^{p+1},\dots,e^{N}} \sum_{i=p+1}^{N} \left\| \left(\left(e^{i}\right)^{T} x \right) e^{i} \right\|$$

$$= \min_{e^{p+1},\dots,e^{N}} \sum_{i=p+1}^{N} \left(\left(e^{i}\right)^{T} x e^{i} \right)^{T} \overbrace{\left(\left(e^{i}\right)^{T} x e^{i} \right)}^{= e^{i}\left(e^{i}\right)^{T} x}$$

$$= \min_{e^{p+1},\dots,e^{N}} \sum_{i=p+1}^{N} \left(\left(e^{i}\right)^{T} x \right) \underbrace{\left(e^{i}\right)^{T} e^{i}}_{=1} \left(x^{T} e^{i} \right)$$

$$= \boxed{\min_{e^{p+1},\dots,e^{N}} \sum_{i=p+1}^{N} \left(e^{i}\right)^{T} x x^{T} e^{i}}$$

Reconstruction through error minimization

Generalize to minimizing reconstruction error for a set of M datapoints

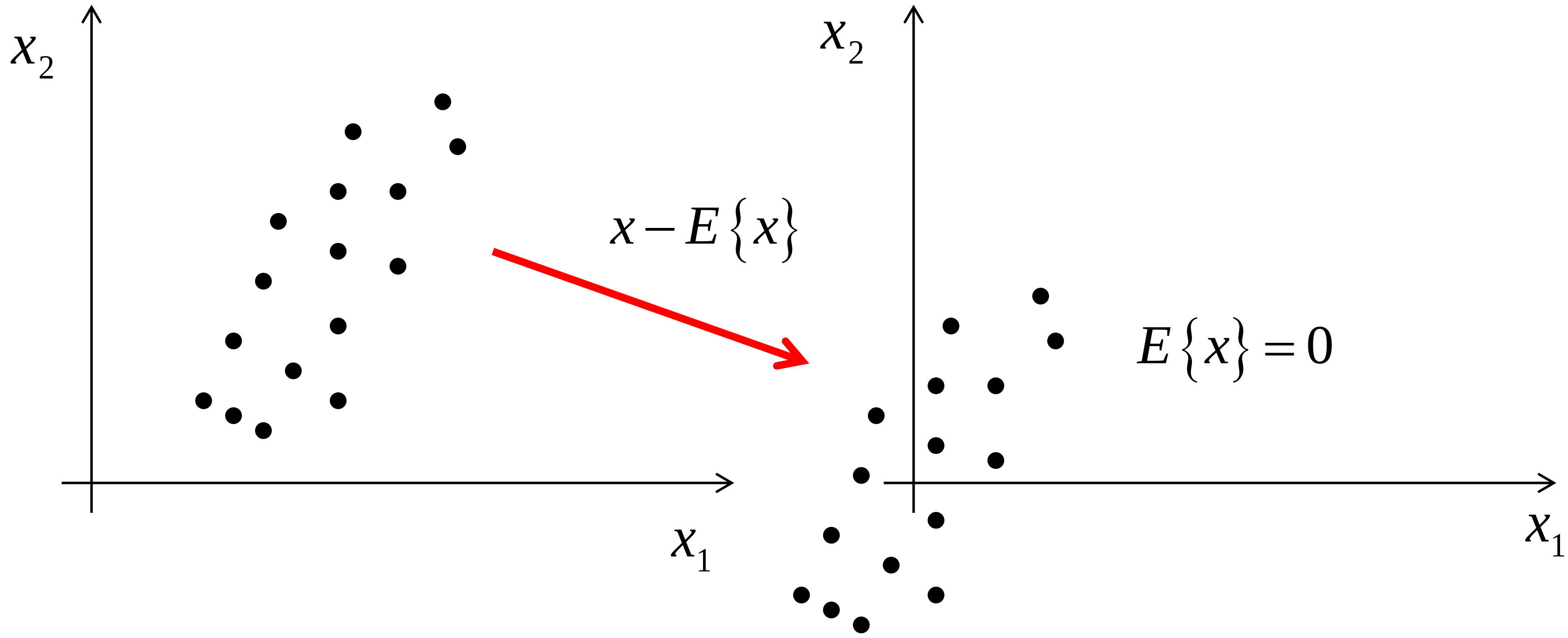
$$\min_{e^{p+1},\dots,e^N} \frac{1}{M} \sum_{i=p+1}^{N} \sum_{j=1}^{M} \left(\left(e^i\right)^T x^j e^i \right)^T \left(\left(e^i\right)^T x^j e^i \right)$$

$$= \min_{e^{p+1},\dots,e^N} \sum_{i=p+1}^{N} \left(e^i\right)^T \left(\frac{1}{M} \sum_{j=1}^{M} x^j \left(x^j\right)^T \right) e^i$$

Covariance Matrix for zero-mean data

$$C = \frac{1}{M} X X^T$$

First pre-processing step in PCA: Center the data



EPFL

PCA as constrained-based optimization

Ensure minimal reconstruction error

$$= \min_{e^{p+1},\dots,e^N} \sum_{i=p+1}^{N} \left(e^i\right)^T C e^i$$

Request that all projection vectors be orthonormal.

$$\left\| e^i \right\| = 1, \ \forall i$$

$$\left(e^i\right)^T e_j = 0, \ \forall i \neq j$$

Optimization with constraints: convex objective function under equality constraint → Lagrange method

Solution to PCA

Constrained-based optimization (solving for one projection)

Minimum of the Lagrangian: $L(e^1,\lambda)=(e^1)^T Ce^1-\lambda\left((e^1)^T e^1-1\right)$

$\lambda \geq 0$

$$\frac{\partial L(e^1,\lambda)}{\partial e^1}=Ce^1-\lambda e^1=0$$

$$\Rightarrow Ce^1=\lambda e^1$$

The solution is an eigenvector of the covariance matrix C!

All eigenvectors of the matrix C are orthonormal
➔ the *p* projections are *p* eigenvectors of *C.*

How do we choose the optimal *p* eigenvectors of *C*?

How do we choose the optimal *p* eigenvectors of *C*?

Percentage of the dataset covered by each projection: $\frac{\left\| X^T e^i \right\|}{\left\| \sum_j X^T e^j \right\|}$

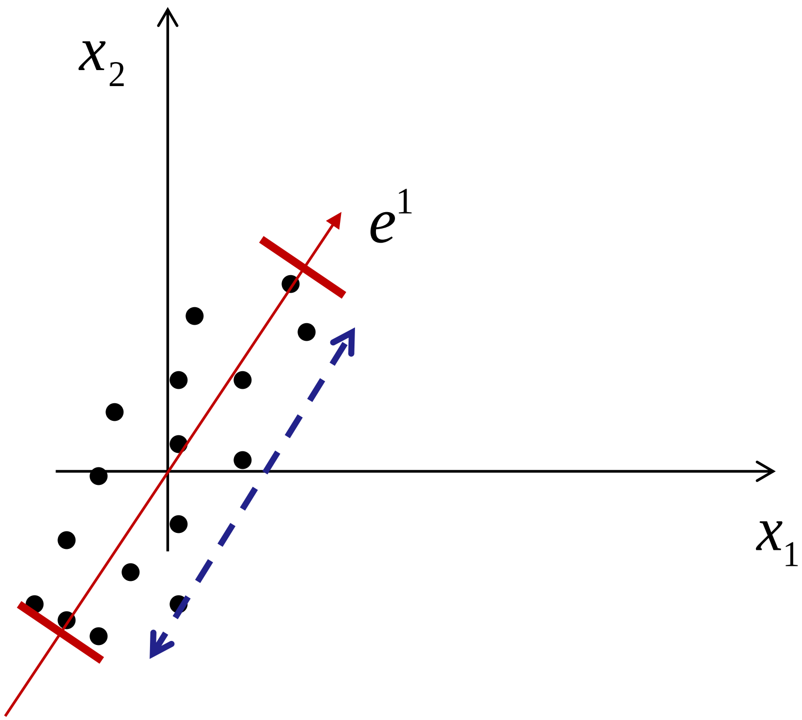
$$\left(X^T e^i\right)^T X^T e^i = \left(e^i\right)^T X X^T e^i = M \left(e^i\right)^T \lambda_i e^i = M \lambda_i .$$

$$\left(e^i\right)^T e^j = 0 \Rightarrow \left\| \sum_j X^T e^j \right\| = \sum_j \left\| X^T e^j \right\|$$

$$\Rightarrow \frac{\left\| X^T e^i \right\|}{\left\| \sum_j X^T e^j \right\|} = \frac{\lambda_i}{\sum_j \lambda_j} .$$

The eigenvalues give a measure of the variance of the distribution of X on each projection.

PCA: Maximize Variance



$$\arg\max_{j \in 1,\dots p} \left(e^j\right)^T C e^j$$

under constraint $\left\| e^j \right\| = 1.$

$$\Leftrightarrow C e^j = \lambda e^j$$

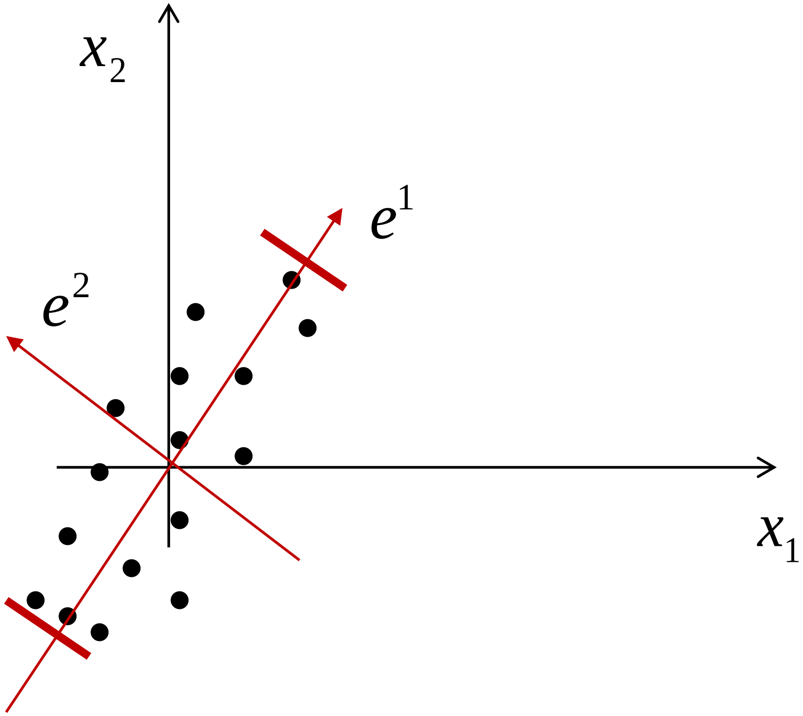
The solution is also an eigenvector of the Covariance matrix.

$$C = \begin{bmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_2, x_1) \\ \mathrm{cov}(x_1, x_2) & \mathrm{var}(x_2) \end{bmatrix}$$

$$\lambda_1 = \left(e^1\right)^T X X^T e^1 \sim \mathrm{var}\left(\left(e^1\right)^T x\right)$$

The eigenvector is aligned with the direction of covariance.

PCA: Decomposition



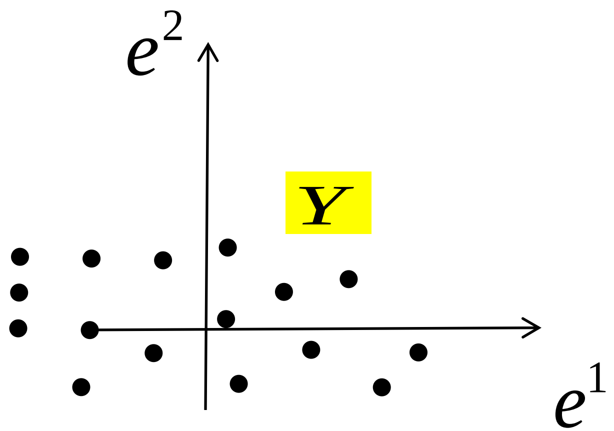
Eigendecomposition of C

$$C = V \Lambda V^T, \quad V = [e^1 \; e^2]$$

$$e^1, e^2 : \text{ orthogonal}$$

Project onto eigenvectors

$$Y = AX \quad A = V^T, \quad V = [e^1 \; e^2]$$



Compute Covariance matrix in projected space

$$C_Y = YY^T$$

$$\Rightarrow C_Y = \Lambda$$

It is diagonal
➔ The projections are uncorrelated!

Summary: Properties of PCA Projections

1. All the projections form an orthonormal basis.

2. The projections of the data onto each axis are uncorrelated.

3. PCA gives an optimal (in the mean-square sense) linear reduction of the dimensionality.

4. The first PCA projection determines the direction (vector) along which the variance of the data is maximal.

PCA Algorithm

Algorithm:

1) Substract the mean: $x \rightarrow x - E\left\{X\right\}$

2) Compute Covariance matrix: $C = E\left\{XX^T\right\}$

3) Compute eigenvalues using $\det\left(C - \lambda I\right) = 0.$

4) Compute eigenvectors using $Ce^i = \lambda_i e^i$.

5) Choose first p < N eigenvectors: $e^1, ... e^p$ with $\lambda_1 \geq \lambda_2 \geq ... \lambda_p$

6) Project data onto new basis: $Y = A_p X,\ A_p = \begin{pmatrix} e_1^1 ...... e_N^1 \\ .. \\ e_1^p ...... e_N^p \end{pmatrix}$

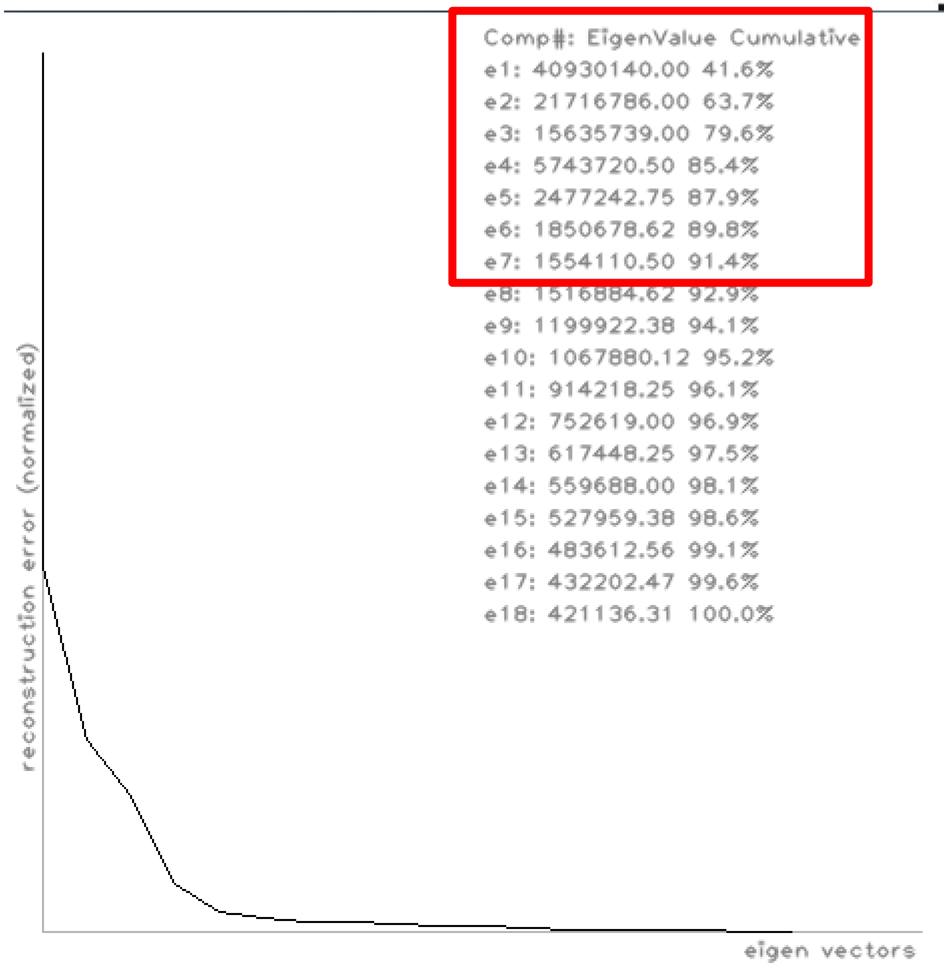
How much information is lost?

Original Image

Image compressed 90%

$$\frac{\sum_{j=p+1}^{N} \lambda_j}{\sum_{i=1}^{N} \lambda_i} = ?$$

How do we choose the optimal *p* eigenvectors of *C*?



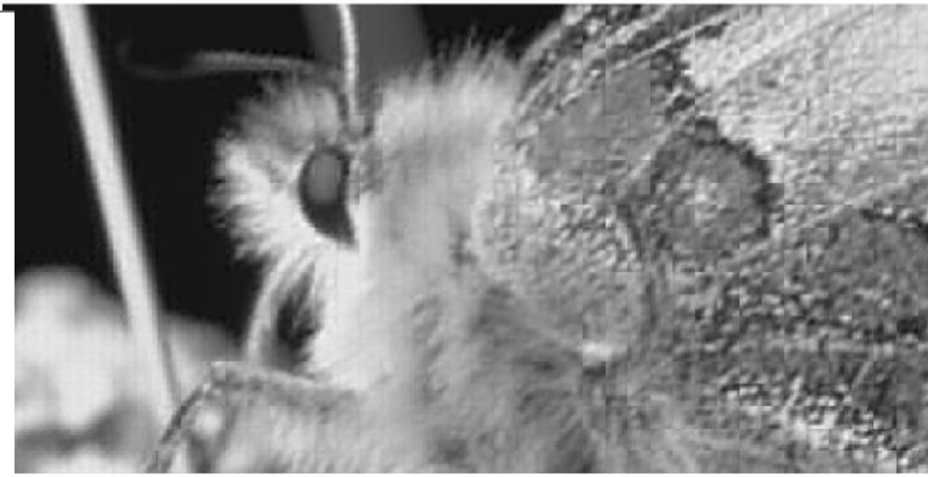


Image compressed 90%

$$\frac{\sum_{j=p+1}^{N} \lambda_j}{\sum_{i=1}^{N} \lambda_i} = 0.1$$