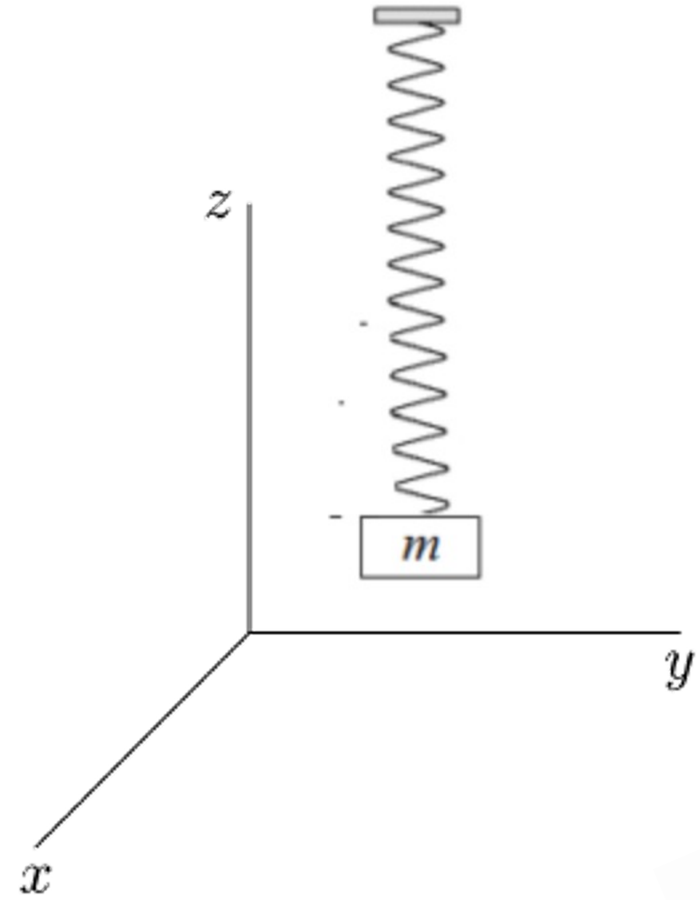
Principal Component Analysis (PCA)

Dr. Umberto Michelucci

umberto.michelucci@toelt.ai

Measurement of the position of the mass *m* at different time points t_i

For each t_i 2 measurements
In the plane of the objective



For each t_i 2 measurements
In the plane of the objective

For each t_i 2 measurements
In the plane of the objective

Only 1 measurement (the distance of the mass m from a point along the spring) is necessary.

PCA – An intuitive understanding

PCA does nothing else than expressing the original data on a new basis that has been obtained by a linear combination of the original basis vectors.

Note: PCA does *NOT* reduce the number of dimensions!!

Change of basis

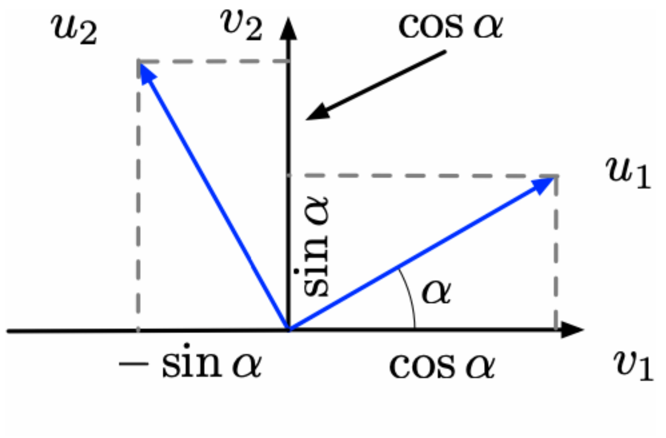
The measure of intelligence is the ability to change

--Albert Einstein.

Change of basis

In general, to make a linear change of basis, it suffices to multiply our identity matrix (our original basis) by a transformation matrix T.

- Let us consider the Euclidean space $\mathbb{R}^2$ with basis vectors $v_1 = (1{,}0)$ and $\mathrm{v}_2 = (0{,}1)$.
- Let us rotate the axis of an angle α.
- The new basis u_1, u_2 are given by



$$u_1 = (\cos\alpha, \sin\alpha)$$
$$v_1 = (-\sin\alpha, \cos\alpha)$$

Change of basis

- Let us define the basis as a matrix (with each vector v_1 and v_2 as column vectors)

$$V = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

- we can define the transformation matrix T

$$T = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$$

- The new basis expressed as a matrix U is given by

$$U = \begin{pmatrix} u_{1,x} & u_{2,x} \\ u_{1,y} & u_{2,y} \end{pmatrix}$$

$$U = TV$$

Change of basis

- Let us consider a generic vector

$$P = p_1 v_1 + p_2 v_2 = p_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + p_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$$

- P in the new basis can be expressed as

$$P_u = TP = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} p_1 \cos\alpha - p_2 \sin\alpha \\ p_1 \sin\alpha + p_2 \cos\alpha \end{pmatrix}$$

Change of basis

- Considering a generic dataset X, it can be expressed in a new basis as Y as

$$TX = \begin{pmatrix} — \boldsymbol{t}_1 — \\ \vdots \\ — \boldsymbol{t}_m — \end{pmatrix} \begin{pmatrix} | & & | \\ \boldsymbol{x_1} & \cdots & \boldsymbol{x_n} \\ | & & | \end{pmatrix} = \begin{pmatrix} \boldsymbol{t}_1 \cdot \boldsymbol{x}_1 & \cdots & \boldsymbol{t}_1 \cdot \boldsymbol{x}_n \\ \vdots & \ddots & \vdots \\ \boldsymbol{t}_m \cdot \boldsymbol{x}_1 & \cdots & \boldsymbol{t}_m \cdot \boldsymbol{x}_n \end{pmatrix}$$

Example of transformations

Geometrical Transformation	**Transformation Matrix**
Rotation of an angle α	$\begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$
Reflection through the x-axis	$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$
Reflection through the y-axis	$\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$
Reflection through the origin	$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

Open Questions

1. What is the best way to re-write our dataset X? Or in other words, how can we decide what is a good and what is a bad representation (or in other words a good or bad basis)?
2. What is a good choice for T?

PCA **Key Assumptions**

1. The directions along which the data show the largest variance are the ones that contain the most relevant and interesting information.
2. Features that are correlated to each other are redundant, or in other words not useful in describing the phenomena we are studying.

PCA **Key Assumptions**

1. The directions along which the data show the largest variance are the ones that contain the most relevant and interesting information.
2. Features that are correlated to each other are redundant, or in other words not useful in describing the phenomena we are studying.

Hypothesis 1 can be grasped by considering that if there is minimal variance along a certain direction, it implies that the data remain largely unchanged in that direction, suggesting a lack of significant information.

PCA **Key Assumptions**

1. The directions along which the data show the largest variance are the ones that contain the most relevant and interesting information.
2. Features that are correlated to each other are redundant, or in other words not useful in describing the phenomena we are studying.

Hypothesis 1 can be grasped by considering that if there is minimal variance along a certain direction, it implies that the data remain largely unchanged in that direction, suggesting a lack of significant information.

Covariance Matrix

Never think that lack of variability is stability. Don't confuse lack of volatility with stability, ever.

-- Nassim Nicholas Taleb

Covariance Matrix

- Let us consider two set of measurements

$$X = \{x_i\}_{i=1}^N \qquad Y = \{y_i\}_{i=1}^N$$

- Covariance between X and Y is given by

$$\sigma_{XY}^2 = \frac{1}{N} \sum_{i=1}^{N} x_i y_i$$

- It measures the degree of relationship between the two variables. A large positive value indicates positively correlated data (if one grows, so does the other). A negative value indicates negatively correlated data (if one grows, the other decreases)

Covariance Matrix

- By writing X and Y as matrices with dimensions $(1, N)$ we can write

$$\sigma^2_{XY} = \frac{1}{N}\boldsymbol{X}\boldsymbol{Y}^T$$

- In general we can generalize the definition for a generic dataset $\boldsymbol{X}$ (where each row contains a complete measurement, all the fatures, while each column all measurements of a specific feature)

$$\boldsymbol{X} = \begin{pmatrix} - & \boldsymbol{x}_1 & - \\ & \vdots & \\ - & \boldsymbol{x}_m & - \end{pmatrix}$$

Covariance Matrix

- In this case we can write the covariance matrix $\boldsymbol{C}$ as

$$\boldsymbol{C} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^T$$

- Note that $\boldsymbol{C}$ is a square matrix.
- Its **diagonal terms** are the variances of the features and the **off-diagonal terms** the covariance between pairs of different features.

Covariance Matrix and PCA

- Remembering our hypothesis, that what is interesting happens along direc- tions that have large variance and small co-variance (low redundancy), we can make the following statements:

1. Large diagonal elements indicate interesting features
2. large off-diagonal terms indicate large redundancy

You may, at this point, see where we are going with this…

What we must do

find a new basis (diagonalise the co-variance matrix) to

1. maximise the variance (so the diagonal elements)
2. minimise redundancy (the off-diagonal elements)

In other words, we must finding a basis that diagonalise the matrix $\boldsymbol{C}$.

Note: To make PCA easy to use and calculate, PCA assumes that the new basis will be an orthonormal matrix, or in other words that the vectors of the new basis are orthogonal to each other.

PCA – The **Key** Assumptions

1. *Linearity*: the change of basis we are searching for is a linear transformation (there are other approaches that lift this assumption, like t-SNE (t-distributed Stochastic Neighbour Embedding)).

2. *Large variance is important*: this assumption is often safe to make, but it is a strong assumption that is not always correct. For example, if you are studying the transverse oscillations of a spring due to imperfections in your system, PCA may simply ignore those effects since the variance is minimal along the transverse direction.

3. *The new basis is orthogonal*: this makes using PCA fast and easy, but it does not work every time. It may well be that there are directions where the interesting phenomena occur that are not orthogonal to each other.

PCA – How to diagonalise $\boldsymbol{C}$

We will not prove this but it is enough to choose as the transformation matrix T one that has for each row an eigenvector of the matrix

$$\frac{1}{N}\boldsymbol{X}\boldsymbol{X}^T$$

PCA – Dimensionality Reduction

PCA does **not** do dimensionality reduction. This is achieved by keeping the «top» 2 (or 3,4, etc.) of the new *dimensions.*