# COMP229: Introduction to Data Science
# Lecture 28: Principal Component Analysis

Olga Anosova, O.Anosova@liverpool.ac.uk
Autumn 2023, Computer Science department
University of Liverpool, United Kingdom

# Lecture plan & learning outcomes

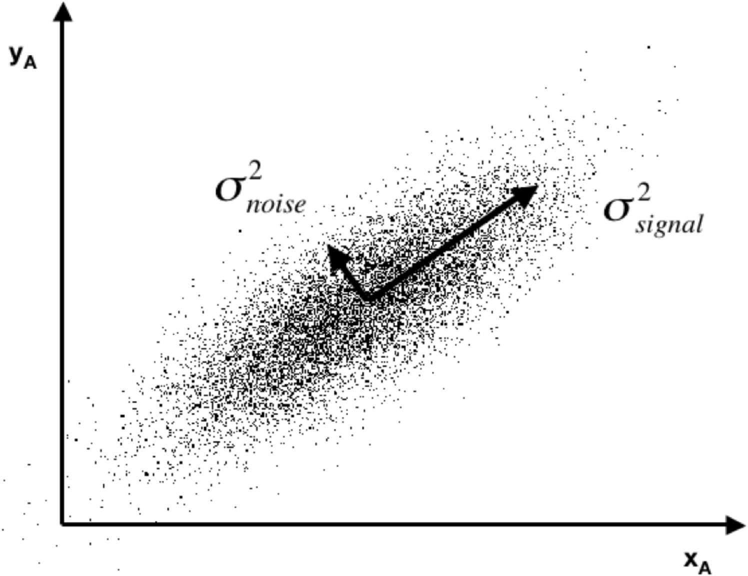
On this lecture we should learn

- what is the Principal Component Analysis,
- the assumptions behind the PCA,
- three main steps of PCA.

# Reminder:

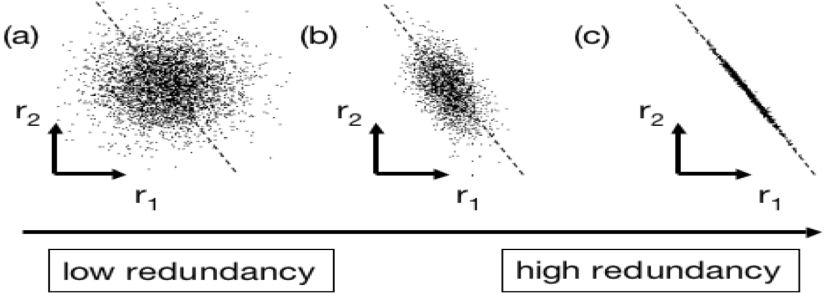
- The *sample covariance* of variables $X, Y$ is $\operatorname{cov}(X, Y) = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$.
- $\operatorname{cov}(X_1, \ldots, X_k)$ is symmetric, positive-definite matrix equal to $\dfrac{SS^T}{n-1}$ (if all sample means are zeros), where $s_{ij}$ is the $j$-th sample value (measurement) of the $i$-th feature (variable).
- Any symmetric positive-definite matrix $A$ has an orthogonal basis of eigenvectors.

# How to find a signal in noisy data



# The signal-to-noise ratio

The *signal-to-noise ratio* $SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}$ is a relative characteristic that helps find a signal in noisy data.



A high *SNR* means that data can be *redundant* : in the data on the right one can keep one of $r_1, r_2$.

# A naive approach to find a signal

The pictures above show that the data is distributed along one direction $\vec{v}_1$ with a highest variance.

The next interesting direction $\vec{v}_2$ has a highest variance after the part parallel to $\vec{v}_1$ is ignored.

If all directions $\vec{v}_1, \vec{v}_2, \dots$ are orthogonal, then the variability along these *principal* directions are *not correlated* (or have the sample covariance 0).

So the key idea of PCA is to find the directions that *diagonalise* the sample covariance matrix, i.e. turn all pairwise correlations into 0.

# PCA definition

Principal Component Analysis is an **orthogonal linear transformation** that transforms the data to a new coordinate system such that
in the new coordinates the greatest variance by some scalar projection of the data lies on the first coordinate (called the first principal component),
the second greatest variance lies on the second coordinate, etc.

# Assumptions (limitations) of PCA

**Linearity**: given data are near a *linear subspace*.

**Sufficiency** of the mean and variance: the data is *normally distributed* near the linear subspace above.

**The signal-to-noise ratio** is high, i.e. the data points are distributed along a linear signal with a large variance, while the noise has a low variance.

**The principal directions** (also called *components* of a given data signal) are orthogonal to each other.

# Assumption of independence

In the previous lecture a data sample was given as a $k \times n$ matrix $S$, where $s_{ij}$ is the $j$-th sample value of the $i$-th feature (or a descriptor). The data below has $k = 3$ features (subjects) and $n = 5$ students.

| subjects/students | $s_{i1}$ | $s_{i2}$ | $s_{i3}$ | $s_{i4}$ | $s_{i5}$ |
|---|---|---|---|---|---|
| Maths | 3 | 3 | 2 | 1 | 1 |
| English | 2 | 3 | 2 | 2 | 1 |
| Art | 3 | 1 | 2 | 3 | 1 |

Assume that *all rows are linearly independent*.

# Step 1: centralisation

**Step 1**. Subtract the sample means $\mu_i$ so that each row in the new sample data matrix has mean 0.

**Problem 28.1**. Apply Step 1 to the data above.

**Solution 28.1**. Each row (feature) has the mean $\mu_i = 2$ (averaged over $n = 5$ students), $i = 1, 2, 3$.

| subjects | $\mu_i$ | $s_{i1} - \mu_i$ | $s_{i2} - \mu_i$ | $s_{i3} - \mu_i$ | $s_{i4} - \mu_i$ | $s_{i5} - \mu_i$ |
|---|---|---|---|---|---|---|
| Maths | 2 | 1 | 1 | 0 | -1 | -1 |
| English | 2 | 0 | 1 | 0 | 0 | -1 |
| Art | 2 | 1 | -1 | 0 | 1 | -1 |

# Step 2: covariance

**Step 2**. Compute the sample covariance matrix as $M = \dfrac{SS^T}{n-1}$ by Claim 27.6 (with all means 0).

**Problem 28.2**. Apply Step 2 to the same data.

**Solution 28.2**. $4M_{ij} = SS^T$ is computed as follows:

$$S = \begin{pmatrix} 1 & 1 & 0 & -1 & -1 \\ 0 & 1 & 0 & 0 & -1 \\ 1 & -1 & 0 & 1 & -1 \end{pmatrix}, \; M = \frac{1}{4}\begin{pmatrix} 4 & 2 & 0 \\ 2 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix}.$$

If $M$ was diagonal, we could keep only several those features (say, two) that have a highest variance.

# The idea of principal directions

We aim to find a new basis (of the principal directions) that makes the covariance matrix diagonal. Then we can project all data to a few directions that have higher variances, i.e. forget about remaining directions with lower variances.

The original covariance matrix $M$ is symmetric positive-definite by Claim 27.7.
By Th 26.12, any symmetric positive-definite matrix has an orthogonal basis of eigenvectors, hence can be diagonalised, i.e. there is a transition matrix $B$ such that the transition matrix $B$ (whose columns are eigenvectors) gives the diagonal $D = B^{-1}MB$.

Eigenvalues are invariants of linear maps, hence $D$ consists of eigenvalues.

# An orthogonal transition matrix

By Claim 21.8, a linear map $A$ is orthogonal if and only if $A^T A = I$.

By definition, the eigenvectors of $M$ are orthogonal to each other, and those are the columns of the transition matrix $B$, hence $B$ is orthogonal and $B^{-1} = B^T$.

Indeed, the $(i, j)$ element of $B^T B$ is the scalar (dot) product of the $i$-th and $j$-th columns of $B$ (equal to 1 for $i = j$, 0 otherwise), so $B^T B = I$, $B^{-1} = B^T$.

Now we will rewrite the data points from the initial sample matrix $S$ with new coordinates in a new matrix $A$ that should have zero pairwise correlations.

# The new matrix $A$ of the data $S$

The $k \times n$ matrix $A = B^T S$ has $a_{ij} = \sum_{l=1}^{k} b_{li} s_{lj}$ equal to the scalar product of the $i$-th eigenvector and the $j$-th data point. For any fixed $i$, $\{a_{i1}, \dots, a_{in}\}$ are the coordinates of the $n$ given points projected to the line of the $i$-th eigenvector $\vec{v}_i = (b_{1i}, \dots, b_{ki})$.

**Claim 28.3.** The new covariance matrix of the transformed data $A = B^T S$ is a diagonal matrix.

*Proof.* $\frac{AA^T}{n-1} = \frac{(B^T S)(B^T S)^T}{n-1} = \frac{(B^T S)(S^T B)}{n-1} = \frac{B^{-1}(SS^T)B}{n-1} = B^{-1}MB = D$, which is diagonal by the choice of $B$. □

# Step 3: eigenvectors of covariance

**Step 3**. Find eigenvectors of the original covariance $M$, which will be the columns of the transition matrix $B$.

**Problem 28.4**. Find the eigenvalues, eigenvectors of the covariance matrix $M = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

**Solution 28.4**. Solve the characteristic equation

$$0 = \det(M - \lambda I) = \det \begin{pmatrix} 1-\lambda & 0.5 & 0 \\ 0.5 & 0.5-\lambda & 0 \\ 0 & 0 & 1-\lambda \end{pmatrix}$$

# Solving a characteristic equation

$$\det(M - \lambda I) = \det \begin{pmatrix} 1-\lambda & 0.5 \\ 0.5 & 0.5-\lambda \end{pmatrix} (1-\lambda) =$$

$$(1-\lambda)((1-\lambda)(0.5-\lambda)-0.25) = (1-\lambda)(\lambda^2-1.5\lambda+0.25) = 0.$$

Order eigenvalues: $\lambda_1 = \dfrac{3+\sqrt{5}}{4} > \lambda_2 = 1 > \lambda_3 = \dfrac{3-\sqrt{5}}{4}$.

The diagonal matrix has all *ordered* eigenvalues on the diagonal:

$$\begin{pmatrix} \frac{3+\sqrt{5}}{4} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{3-\sqrt{5}}{4} \end{pmatrix}$$

Eigenvectors: for each $\lambda$ such that $\det(M - \lambda I) = 0$, the equation $(M - \lambda I)\vec{v} = \vec{0}$ has infinitely many solutions.

# Finding one eigenvector (of many)

For example, for $\lambda_2 = 1$, the matrix $M - I$ is

$$\begin{pmatrix} 0 & 0.5 & 0 \\ 0.5 & -0.5 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Due to the 3rd column of zeros $(M - I)\vec{v} = \vec{0}$ has the solution $\vec{v}_2 = (0, 0, 1)$, or any vector parallel to $\vec{v}$.

Similarly, any other eigenvector is not unique, so we talk about principal **directions** (not vectors).

# Finding other eigenvectors

For $\lambda_1 = \frac{3+\sqrt{5}}{4}$, the equation $(M - \lambda_1 I)\vec{v}_1 = \vec{0}$ is

$$\begin{pmatrix} \frac{1-\sqrt{5}}{4} & 0.5 & 0 \\ 0.5 & -\frac{1+\sqrt{5}}{4} & 0 \\ 0 & 0 & \frac{1-\sqrt{5}}{4} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

so $z = 0$ and $(1-\sqrt{5})x + 2y = 0 = 2x - (1+\sqrt{5})y$.

One of infinitely many solutions to the last two equations are $x = 1 + \sqrt{5}$, $y = 2$. The eigenvector can be (any parallel to) $\vec{v}_1 = (\sqrt{5}+1, 2, 0)$.

Similarly find an eigenvector for $\lambda_3 = \frac{3-\sqrt{5}}{4}$.

Eigenvectors can be chosen as vectors parallel to $\vec{v}_1 = (\sqrt{5}+1, 2, 0)$ for $\lambda_1$, $\vec{v}_2 = (0, 0, 1)$ for $\lambda_2 = 1$, $\vec{v}_3 = (-2, \sqrt{5}+1, 0)$ for $\lambda_3$.

# 1-dimensional PCA approximation

To find the 1-dimensional approximation (reconstruction) to our example data by the PCA, take the linear approximation along the 1st principal direction $\vec{v}_1 = (\sqrt{5}+1, 2, 0)$.

The straight line should pass via the mean point $(2, 2, 2)$, so $x = 2 + (\sqrt{5}+1)t$, $y = 2 + 2t$, $z = 2$ for $t \in \mathbb{R}$.

# Time to revise and ask questions

- Step 1: subtract the means so that the rows of the sample $k \times n$ matrix $S$ have mean 0.
- Step 2: find the covariance matrix $M = \dfrac{SS^T}{n-1}$.
- Step 3: a few eigenvectors of $M$ with largest eigenvalues span an approximating subspace.

**Problem 28.5**. What are the advantages and the drawbacks of the PCA?