COMP229: Introduction to Data Science
Lecture 27: The covariance matrix

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 covariance matrix,
- how to compute it,
- what conclusions can we derive from it.

Reminder: Eigenthings

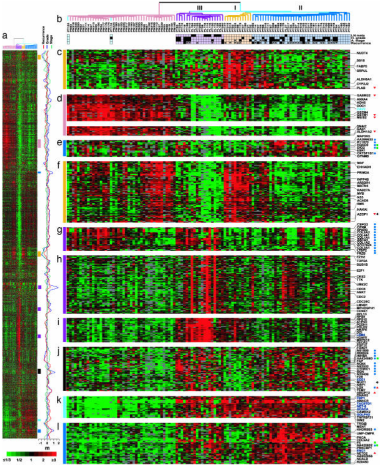
- *Eigenvalues* λ and *eigenvectors* $\vec{v}$ of A are solutions of $A\vec{v} = \lambda\vec{v}$, hence $\det(A - \lambda I) = 0$.
- Any symmetric positive-definite matrix A has an orthogonal basis of eigenvectors.

A high-dimensional data problem

A typical obstacle for high-dimensional data: understand which of many different measurements are closely related. Any human has about 20,000 genes. Genomics studies relationships between genes and diseases. Since many genes are often responsible for one disease, we need to find these correlated genes from a sample of human genes.

Real-life example

Gene expression profiling identifies clinically relevant subtypes of prostate cancer



The image visualises the dataset consisting of the expression of 5153 genes in 112 prostate cancer patients. We may say that there are 5153 features associated with every patient. Can we notice any patterns?

Let's try pairwise comparisons, similar to linear correlation.

The variance and the covariance

The **sample variance** (the squared sample standard deviation) of n values $x_1, \ldots, x_n$ sampled from a random variable X is $\operatorname{var} X = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$, where $\bar{x} = \dfrac{1}{n} \sum_{i=1}^{n} x_i$ is the sample mean.

Definition 27.1. The **sample covariance** between samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ of two random variables is $\operatorname{cov}(X, Y) = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$.

The variance of X is the covariance of X with itself.

Another formula for the covariance

Problem 27.2. Prove that $\operatorname{cov}(X, Y) = \dfrac{\sum\limits_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1}$.

Solution 27.2. Expand the brackets:

$$\operatorname{cov}(X, Y) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i y_i - \bar{x} y_i - \bar{y} x_i + \bar{x}\bar{y}) =$$

$$\sum_{i=1}^{n}(x_i y_i) - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + n\bar{x}\bar{y} =$$

$$\sum_{i=1}^{n} x_i y_i - \bar{x}(n\bar{y}) - \bar{y}(n\bar{x}) + n\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}. \qquad \square$$

Properties of the covariance

The covariance is related to the correlation r_{xy}:

$r_{xy} = \dfrac{\operatorname{cov}(X, Y)}{s_x s_y}$, where $s_x, s_y > 0$ are the sample standard deviations of X, Y.

From the correlation properties we remember that

- $\operatorname{cov}(X, Y) > 0$ means that both random variables X, Y simultaneously increase or decrease;
- $\operatorname{cov}(X, Y) < 0$ means that the variable X increases while Y decreases (and vice versa);
- $\operatorname{cov}(X, Y) = 0$ means that X, Y are 'independent'.

The covariance of a random vector

Definition 27.3. Let $X_1, \ldots, X_k$ be k random variables. The (i, j) element of the **covariance matrix** $\operatorname{cov}(X_1, \ldots, X_k)$ is $\operatorname{cov}(X_i, X_j)$ from Def 27.1.

Claim 27.4. The variance and covariance are preserved if we shift any variable by a constant.

Proof. $\operatorname{cov}(X, Y) = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$ consists of the differences that are preserved when all sample values x_i or y_i are shifted by a constant. $\square$

A sample matrix of data

Definition 27.5. Let each of k data features have $n \geqslant k$ sample values. All values are represented by the **sample** $k \times n$ **matrix** S, where each row represents one of k features and each column represents one of n data objects, e.g. s_{ij} is the j-th sample value of the i-th feature, see the table:

Subjects	student 1	stud. 2	stud. 3	stud. 4	stud. 5
Maths	3	3	2	1	1
English	2	3	2	2	1
Art	3	1	2	3	1

The covariance is a matrix product

Claim 27.6. Let each of k features X_i have the zero mean over n sample values. If S is the sample $k \times n$ matrix, then $\operatorname{cov}(X_1, \ldots, X_k) = \dfrac{SS^T}{n-1}$.

Proof. The i-th row $(s_{i1}, \ldots, s_{in})$ of S is the sample vector of X_i. When the means of all rows are zeros, by Definition 27.1 $(n-1)\operatorname{cov}(X_i, X_j) = \sum\limits_{l=1}^{n} s_{il}s_{jl}$ in terms of linear algebra is the scalar product of the i-th and j-th rows of S, hence

$$\operatorname{cov}(X_i, X_j) = \frac{\sum\limits_{l=1}^{n} s_{il}s_{jl}}{n-1} = \frac{\sum\limits_{l=1}^{n} s_{il}(S^T)_{lj}}{n-1} = \frac{(SS^T)_{ij}}{n-1}. \qquad \square$$

The covariance is positive-definite

Claim 27.7. If a sample matrix S has *linearly independent* rows, then the covariance matrix $\mathrm{cov}\,(X_1, \dots, X_k)$ is symmetric positive-definite.

Proof The symmetry follows from Definition 27.1 as $\mathrm{cov}\,(X, Y) = \mathrm{cov}\,(Y, X)$. By Claim 27.6 we now check that the matrix SS^T is positive-definite:

Lemma 21.7 says that $\vec{v}^T S = (S^T \vec{v})^T$. For $\vec{v} \neq \vec{0}$, $\vec{v}^T (SS^T)\vec{v} = (S^T \vec{v})^T (S^T \vec{v}) =$

The covariance is positive-definite

Claim 27.7. If a sample matrix S has *linearly independent* rows, then the covariance matrix $\operatorname{cov}(X_1, \ldots, X_k)$ is symmetric positive-definite.

Proof The symmetry follows from Definition 27.1 as $\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$. By Claim 27.6 we now check that the matrix SS^T is positive-definite:

Lemma 21.7 says that $\vec{v}^T S = (S^T \vec{v})^T$. For $\vec{v} \neq \vec{0}$, $\vec{v}^T (SS^T) \vec{v} = (S^T \vec{v})^T (S^T \vec{v}) = |S^T \vec{v}|^2 > 0$, because a linear independence of rows in S means that any combination of columns $S^T \vec{v} \neq \vec{0}$ for $\vec{v} \neq 0$. □

Now, remebering the previous Lecture 26, is it clear why the covariance matrix is important?

Step-by-step example

Step1: subtract sample means.

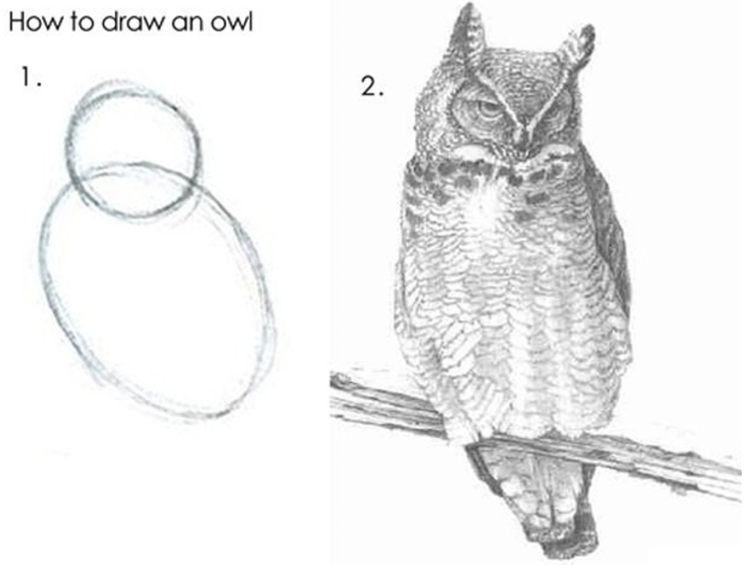
The table of original marks (grades):

subjects/students	s_{i1}	s_{i2}	s_{i3}	s_{i4}	s_{i5}	sum	mean μ_i
Maths	3	3	2	1	1	10	2
English	2	3	2	2	1	10	2
Art	3	1	2	3	1	10	2

The table after subtracting means of rows:

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

Step2: draw the owl



Step2: compute the covariance

For $S = \begin{pmatrix} 1 & 1 & 0 & -1 & -1 \\ 0 & 1 & 0 & 0 & -1 \\ 1 & -1 & 0 & 1 & -1 \end{pmatrix}$, the product SS^T is $\begin{pmatrix} 4 & 2 & 0 \\ 2 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix}$. Then $\frac{SS^T}{n-1} = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ is the sample covariance matrix $\mathrm{cov}\,(X_1, X_2, X_3)$,

whose diagonal contains $\mathrm{var}\,(X_i)$, e.g. the maths and art marks are more variable than English marks.

Time to revise and ask questions

- 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).

Problem 27.8. In the problem above what can you say about dependence of marks in different subjects?