COMP229: Introduction to Data Science
Lecture 22: the covariance matrix

Vitaliy Kurlin, vitaliy.kurlin@liverpool.ac.uk
Autumn 2018, Computer Science department
University of Liverpool, United Kingdom

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.

In this lecture real data are modelled by a vector in $\mathbb{R}^n$ with random (usually normal) coordinates.

The variance and the covariance

Definition 22.1. The *sample variance* (the squared standard deviation) of n values $x_1, \ldots, x_n$ is $\operatorname{var} X = \dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^2$, where $\bar{x} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i$ is the sample mean, compare with Definition 13.6.

Definition 22.2. 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 X.

Properties of the covariance

The covariance is related to the correlation r_{xy}, see Definition 14.2: $r_{xy} = \dfrac{\operatorname{cov}(X, Y)}{s_x s_y}$, where $s_x, s_y > 0$ are the sample standard deviations.

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 X, Y are "independent".

The covariance of a random vector

Definition 22.3. 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 $\operatorname{cov}(X_i, X_j)$ in Def 22.2.

Claim 22.4. The variance and covariance are preserved if we shift variables by a constant.

Proof. The formula from Definition 22.2 $\operatorname{cov}(X, Y) = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$ contains the differences that are preserved when all sample values x_i or y_i are shifted by a constant. □

A sample matrix of data

Definition 22.5. Let each of k data features have $n \geq k$ 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	student 2	student 3	4	5
Maths	3	3	2	1	1
English	2	3	2	2	1
Art	3	1	2	3	1

The covariance via the sample matrix

Claim 22.6. Let each of k features X_i have the zero mean over n sample values. If S is the sample $k \times n$ matrix, $\operatorname{cov}(X_1, \ldots, X_k) = \dfrac{SS^T}{n-1}$.

Proof. The variable X_l has the sample vector $(s_{l1}, \ldots, s_{ln})$. We compare the elements from row l and column j in the required matrix identity by Definition 22.2 (when the means are zeros):

$$\operatorname{cov}(X_l, X_j) = \frac{\sum\limits_{i=1}^{n} s_{li} s_{ji}}{n-1} = \frac{\sum\limits_{i=1}^{n} s_{li} (S^T)_{ij}}{n-1} = \frac{(SS^T)_{lj}}{n-1}. \qquad \square$$

The covariance is positive-definite

Linear independence of rows in S means that any combination of columns $S^T\vec{v} \neq \vec{0}$ for $\vec{v} \neq 0$.

Claim 22.7. 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 22.2 and $\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$. By Claim 22.6 we check that the matrix SS^T is positive-definite, i.e.

$$\vec{v}^T SS^T\vec{v} = (S^T\vec{v})^T(S^T\vec{v}) = ||S^T\vec{v}||^2 > 0,\ \vec{v} \neq 0. \qquad \square$$

Subtract sample means

The table of original marks (grades):

Subjects	1	2	3	4	5	sum	mean
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:

Subjects	1	2	3	4	5
Maths	1	1	0	–1	–1
English	0	1	0	0	–1
Art	1	–1	0	1	–1

Computing 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 $\begin{pmatrix} 4 & 2 & 0 \\ 2 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix}$. With the factor $\frac{1}{n-1} = \frac{1}{4}$ the sample covariance matrix is $\begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

The diagonal shows variances: maths and art marks are more variable than English marks.

Your questions and the quiz

To benefit from the lecture, now you could

- ask or submit your anonymous questions to the COMP229 folder after the lecture;
- write down your summary in 2–3 phrases, e.g. list key concepts you have learned;
- talk to your classmates to revise the lecture.

Question. Looking at the covariance matrix above, what are dependencies of the subject marks?

Answer to the quiz and summary

Answer. The marks for maths and English seem to be correlated with each other. The art marks seem to be independent of the other marks.

- The *sample covariance matrix* of variables is $\operatorname{cov}(X, Y) = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$, which is always symmetric and positive-definite.
- If all sample means are zeros, then the covariance is $\dfrac{SS^T}{n-1}$, where s_{ij} is the j-th sample value of the i-th feature (variable).