Lecture 3.1: The rank

Optimization and Computational Linear Algebra for Data Science

Léo Miolane

Rank of a family of vectors

Definition

We define the rank of a family $x_1, \ldots, x_k$ of vectors of $\mathbb{R}^n$ as the dimension of its span:

$$\operatorname{rank}(x_1, \ldots, x_k) \stackrel{\text{def}}{=} \dim(\operatorname{Span}(x_1, \ldots, x_k)).$$

Rank of a matrix

Definition

Let $M \in \mathbb{R}^{n \times m}$. Let $c_1, \ldots, c_m \in \mathbb{R}^n$ be its columns. We define

$$\operatorname{rank}(M) \stackrel{\text{def}}{=} \operatorname{rank}(c_1, \ldots, c_m) = \dim(\operatorname{Im}(M)).$$

Example

« Rank of columns = rank of rows »

Proposition

Let $M \in \mathbb{R}^{n\times m}$. Let $r_1, \ldots, r_n \in \mathbb{R}^m$ be the rows of M and $c_1, \ldots, c_m \in \mathbb{R}^n$ be its columns. Then we have

$$\operatorname{rank}(r_1, \ldots, r_n) = \operatorname{rank}(c_1, \ldots, c_m) = \operatorname{rank}(M).$$

The rank in Data Science

Consider a matrix M of size 1000×500:

$$M = \begin{pmatrix} - & r_1 & - \\ & \vdots & \\ - & r_{1000} & - \end{pmatrix}$$

What does it mean to say that « $\mathrm{rank}(M) = 5$ » ?

The rank in Data Science

Imagine now that

- The rows of M corresponds to Netflix's users.
- The columns of M corresponds to Netflix's movies.
- The entry $M_{i,j}$ is rating of the movie j by the user i, assuming that all the users have rated all the movies.

Claim: the rank of M is "small".

- The ratings of a user can be obtained as a linear combination of a small number of « profiles ».
- In practice, we do not have access to the full matrix, so we can use this assumption to predict the missing entries.