

# Optimization and Computational Linear Algebra for Data Science

## Lecture 7: Singular value decomposition

Léo MIOLANE · leo.miolane@gmail.com

July 10, 2019

**Warning:** *This material is not meant to be lecture notes. It only gathers the main concepts and results from the lecture, without any additional explanation, motivation, examples, figures...*

## 1 The Spectral Theorem

The main result of this section is the following “Spectral Theorem” which tells us that a symmetric matrix is diagonalizable in an orthonormal basis.

### Theorem 1.1 (*Spectral Theorem*)

Let  $A \in \mathbb{R}^{n \times n}$  be a **symmetric** matrix. Then there is a orthonormal basis of  $\mathbb{R}^n$  composed of eigenvectors of  $A$ .

Given an  $n \times n$  symmetric matrix  $A$ , Theorem 1.1 tells us that one can find an orthonormal basis  $(v_1, \dots, v_n)$  of  $\mathbb{R}^n$  and scalars  $\lambda_1, \dots, \lambda_n \in \mathbb{R}$  such that for all  $i \in \{1, \dots, n\}$ ,

$$Av_i = \lambda_i v_i.$$

Let  $P$  be the  $n \times n$  matrix whose columns are  $v_1, \dots, v_n$ . Since  $(v_1, \dots, v_n)$  is an orthonormal basis, we get that  $P$  is an orthogonal matrix. Let  $D = \text{Diag}(\lambda_1, \dots, \lambda_n)$  and compute

$$AP = A \begin{pmatrix} | & | & \cdots & | \\ v_1 & v_2 & & v_n \\ | & | & & | \end{pmatrix} = \begin{pmatrix} | & | & \cdots & | \\ Av_1 & Av_2 & & Av_n \\ | & | & & | \end{pmatrix} = \begin{pmatrix} | & | & \cdots & | \\ \lambda_1 v_1 & \lambda_2 v_2 & & \lambda_n v_n \\ | & | & & | \end{pmatrix} = PD.$$

By multiplying by  $P^\top$  on both sides, we get  $APP^\top = PDP^\top$ . Recall now that  $P$  is orthogonal, therefore  $PP^\top = \text{Id}_n$ . We conclude that  $A = PDP^\top$ .

### Theorem 1.2 (*Spectral Theorem, matrix formulation*)

Let  $A \in \mathbb{R}^{n \times n}$  be a symmetric matrix. Then there exists an orthogonal matrix  $P$  and a diagonal matrix  $D$  of sizes  $n \times n$ , such that

$$A = PDP^\top.$$

### Proposition 1.1

Let  $A$  be a  $n \times n$  symmetric matrix and let  $\lambda_1 \geq \dots \geq \lambda_n$  be its  $n$  eigenvalues and  $v_1, \dots, v_n$  be the associated orthonormal family of eigenvectors. Then

$$v_1 = \arg \max_{\|v\|=1} v^\top Av, \quad \text{and for } k = 2, \dots, n, \quad v_k = \arg \max_{\|v\|=1, v \perp v_1, \dots, v_{k-1}} v^\top Av.$$

**Remark 1.1.** Applying the proposition above to the matrix  $-A$  which is symmetric with eigenvalues  $-\lambda_n \geq \dots \geq -\lambda_1$  and associated eigenvectors  $v_n, \dots, v_1$ , we get

$$v_n = \arg \min_{\|v\|=1} v^\top Av, \quad \text{and for } k = 1, \dots, n-1 \quad v_k = \arg \min_{\|v\|=1, v \perp v_{k+1}, \dots, v_n} v^\top Av.$$

## Positive matrices

### Definition 1.1

A symmetric matrix  $A \in \mathbb{R}^{n \times n}$  is said to be positive semi-definite if

$$\forall x \in \mathbb{R}^n, x^\top A x \geq 0. \quad (1)$$

The matrix  $A$  is said to be positive definite if moreover the inequality in (1) is strict for all  $x \neq 0$ .

**Remark 1.2.** Negative semi-definite and negative definite matrices are defined analogously.

### Proposition 1.2

Let  $A \in \mathbb{R}^{n \times n}$  be a symmetric matrix, and let  $\lambda_1, \dots, \lambda_n \in \mathbb{R}$  its eigenvalues. Then

$$A \text{ is positive semi-definite} \iff \lambda_i \geq 0 \text{ for } i = 1, \dots, n,$$

and

$$A \text{ is positive definite} \iff \lambda_i > 0 \text{ for } i = 1, \dots, n.$$

**Exercise 1.1.** Let  $A \in \mathbb{R}^{n \times n}$ .

- Show that  $A^\top A$  positive semi-definite.
- Let  $M$  be a  $n \times n$  symmetric positive semi-definite matrix. Show that there exists  $A \in \mathbb{R}^{n \times n}$  such that  $M = A^\top A$ .

## 2 Singular value decomposition

### Theorem 2.1 (Singular value decomposition (SVD))

Let  $A \in \mathbb{R}^{n \times m}$ . Then there exists two orthogonal matrices  $U \in \mathbb{R}^{n \times n}$  and  $V \in \mathbb{R}^{m \times m}$  and a matrix  $\Sigma \in \mathbb{R}^{n \times m}$  such that  $\Sigma_{1,1} \geq \Sigma_{2,2} \geq \dots \geq 0$  and  $\Sigma_{i,j} = 0$  for  $i \neq j$

$$A = U \Sigma V^\top.$$

The columns  $u_1, \dots, u_n$  of  $U$  (respectively the columns  $v_1, \dots, v_m$  of  $V$ ) are called the left (resp. right) singular vectors of  $A$ . The non-negative numbers  $\Sigma_{i,i}$  are the singular values of  $A$ . Moreover  $\text{rank}(A) = \#\{i \mid \Sigma_{i,i} \neq 0\}$ .

Notice that the singular vectors (similarly to the eigenvectors) are not uniquely defined: if  $A = U \Sigma V^\top$  is a SVD of  $A$ , then  $A = (-U) \Sigma (-V)^\top$  is also a SVD of  $A$ . However, with a slight abuse of language, we will often refer  $v_i$  as the  $i^{\text{th}}$  right singular vector of  $A$ .

### 2.1 Properties of the SVD

Let  $A \in \mathbb{R}^{n \times m}$  and let  $U \Sigma V^\top$  be a singular value decomposition of  $A$  as in Theorem 2.1. Let  $u_1, \dots, u_n$  be the left singular vectors (i.e. the columns of  $U$ ) and  $v_1, \dots, v_m$  be the right singular vectors (i.e. the columns of  $V$ ). Let  $\sigma_i = \Sigma_{i,i}$  be the singular values of  $A$ .

**Proposition 2.1**

For  $i = 1, \dots, \text{rank}(A)$  we have

$$Av_i = \sigma_i u_i \quad \text{and} \quad A^\top u_i = \sigma_i v_i.$$

The most important property of the singular vectors for us is the following:

**Proposition 2.2**

We have

$$v_1 = \arg \max_{\|v\|=1} \|Av\| \quad \text{and} \quad \sigma_1 = \max_{\|v\|=1} \|Av\|. \quad (2)$$

It holds also that

$$v_2 = \arg \max_{\|v\|=1, v \perp v_1} \|Av\| \quad \text{and} \quad \sigma_2 = \max_{\|v\|=1, v \perp v_1} \|Av\| \quad (3)$$

and more generally:

$$v_k = \arg \max_{\|v\|=1, v \perp v_1, \dots, v_{k-1}} \|Av\|. \quad \text{and} \quad \sigma_k = \max_{\|v\|=1, v \perp v_1, \dots, v_{k-1}} \|Av\|. \quad (4)$$

**Remark 2.1.** Considering  $A^\top$  leads to an analogous result for the left singular vectors  $u_k$ :

$$u_k = \arg \max_{\|u\|=1, u \perp u_1, \dots, u_{k-1}} \|A^\top u\|. \quad \text{and} \quad \sigma_k = \max_{\|u\|=1, u \perp u_1, \dots, u_{k-1}} \|A^\top u\|. \quad (5)$$

**Proof.** Compute  $A^\top A = V \Sigma^\top \Sigma V^\top = V D V^\top$  where the matrix  $D \stackrel{\text{def}}{=} \Sigma^\top \Sigma$  is diagonal with  $D_{i,i} = \sigma_i^2$ . The family  $(v_1, \dots, v_m)$  is therefore an orthonormal family of eigenvectors of the symmetric matrix  $A^\top A$  and  $\sigma_1^2 \geq \dots \geq \sigma_m^2$  are the corresponding eigenvalues. The result follows then from Proposition 1.1 applied to  $A^\top A$ , noticing that  $v^\top A^\top A v = \|Av\|^2$ .  $\square$

**2.2 Proof of Theorem 2.1**

We apply the Spectral Theorem (Theorem 1.1) to the  $m \times m$  matrix  $A^\top A$ : there exists an orthonormal basis  $(v_1, \dots, v_m)$  of  $\mathbb{R}^m$  of eigenvectors of  $A^\top A$  associated to eigenvalues  $\lambda_1 \geq \dots \geq \lambda_m$  that are all non-negative because  $A^\top A$  is non-negative. Let  $V \in \mathbb{R}^{m \times m}$  be the orthogonal matrix whose columns are  $(v_1, \dots, v_m)$ .

Let us write  $\sigma_i = \sqrt{\lambda_i}$  and let  $r = \max\{i | \sigma_i > 0\}$ . Define for  $i = 1, \dots, r$

$$u_i = \frac{1}{\sigma_i} A v_i \in \mathbb{R}^n. \quad (6)$$

**Lemma 2.1**

The family  $(u_1, \dots, u_r)$  is orthonormal.

**Proof.** Let  $i, j \in \{1, \dots, r\}$ .

$$\langle u_i, u_j \rangle = \left( \frac{1}{\sigma_i} A v_i \right)^\top \left( \frac{1}{\sigma_j} A v_j \right) = \frac{1}{\sigma_i \sigma_j} v_i^\top A^\top A v_j = \frac{\sigma_i}{\sigma_j} v_i^\top v_j = \mathbb{1}_{i=j},$$

since  $A^\top A v_i = \sigma_i^2 v_i$ .  $\square$

If  $r < n$  we let  $(u_{r+1}, \dots, u_n)$  be an orthonormal family of vectors of  $\mathbb{R}^n$  that are orthogonal to  $u_1, \dots, u_r$ . The family  $(u_1, \dots, u_n)$  is then an orthonormal basis of  $\mathbb{R}^n$ . Let  $U \in \mathbb{R}^{n \times n}$  be the orthogonal matrix whose columns are  $(u_1, \dots, u_n)$ .

**Lemma 2.2**

For  $i = r + 1, \dots, m$ ,  $Av_i = 0$ .

**Proof.** We compute for  $i = r + 1, \dots, m$ :

$$\|Av_i\|^2 = v_i^\top A^\top A v_i = v_i^\top (\lambda_i v_i) = \sigma_i^2 = 0.$$

□

Finally, we let  $\Sigma \in \mathbb{R}^{n \times m}$  defined by:

$$\Sigma_{i,j} = \begin{cases} \sigma_i & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

It remains to verify that  $A = U\Sigma V^\top$ . Compute for  $i = 1, \dots, m$ , using the definition (6) and Lemma 2.2:

$$Av_i = \begin{cases} \sigma_i u_i & \text{if } i \leq r \\ 0 & \text{otherwise.} \end{cases}$$

By orthogonality of  $V$  and the construction of  $\Sigma$  one verifies easily that

$$U\Sigma V^\top v_i = \begin{cases} \sigma_i u_i & \text{if } i \leq r \\ 0 & \text{otherwise.} \end{cases}$$

We conclude that for all  $i \in \{1, \dots, m\}$ ,  $Av_i = U\Sigma V^\top v_i$ . Since a linear transformation is uniquely determined by the image of a basis, we conclude that  $A = U\Sigma V^\top$ .

It remains to show:

**Lemma 2.3**

$\text{rank}(A) = r$ .

**Proof.** The family  $(u_1, \dots, u_r)$  is orthonormal, hence linearly independent. By definition  $u_i \in \text{Im}(A)$  which implies that  $\text{rank}(A) = \dim(\text{Im}(A)) \geq r$ . To prove the converse inequality, notice that by Lemma 2.2  $v_i \in \text{Ker}(A)$  for  $i = r + 1, \dots, m$ . The vectors  $(v_{r+1}, \dots, v_m)$  are orthonormal, hence linearly independent. This implies that  $\dim(\text{Ker}(A)) \geq m - r$ . We conclude by applying the rank Theorem:

$$\text{rank}(A) = m - \dim(\text{Ker}(A)) \leq m - (m - r) = r.$$

□

## 3 Interpretation and applications of the SVD

### 3.1 Geometric interpretation

### 3.2 “Maximal variance” interpretation

Let  $a_1, \dots, a_n \in \mathbb{R}^d$  be  $n$  points in  $d$  dimensions. We assume that this points are centered, meaning that

$$\sum_{i=1}^n a_i = 0.$$

Let  $A$  be the  $n \times d$  matrix whose rows are  $a_1, \dots, a_n$  and let  $(v_1, \dots, v_n)$  be its right singular vectors. By Proposition 2.2,  $v_1$ , the first right singular vector of  $A$ , maximizes

$$v \mapsto \|Av\|^2 = \sum_{i=1}^n \langle a_i, v \rangle^2$$

over the unit sphere. This quantity is the variance of the coordinates of the points  $a_1, \dots, a_n$  along the direction  $\text{Span}(v)$ .

The first right singular vector  $v_1$  gives therefore the direction along which the variance of the data is maximal. Proposition 2.2 gives also that

$$v_k = \arg \max_{\|v\|=1, v \perp v_1, \dots, v_{k-1}} \|Av\|^2. \quad (7)$$

Hence  $v_2$  gives the direction orthogonal to  $v_1$  that maximizes the variance and so on...

### 3.3 Application: Principal Component Analysis (PCA)

#### 3.4 Best-fitting subspace

Let  $a_1, \dots, a_n \in \mathbb{R}^d$  be  $n$  points in  $d$  dimensions. We consider the problem of finding the  $k$ -dimensional subspace (for  $k = 1, \dots, n$ ) that fits “the best” these  $n$  data points. By “best”, we mean here the  $k$ -dimensional subspace  $S$  that minimize the sum of the square distances to the  $n$  points:

$$\text{minimize } \sum_{i=1}^n d(a_i, S)^2 \text{ with respect to } S \text{ subspace of dimension } k. \quad (8)$$

Let  $A$  be the  $n \times d$  matrix whose rows are  $a_1, \dots, a_n$ . The goal of this section is to prove:

#### Theorem 3.1

*Let  $v_1, \dots, v_n$  be right singular vectors of  $A$ . Then for all  $k \in \{1, \dots, n\}$ , the subspace  $\text{Span}(v_1, \dots, v_k)$  is a solution of (8).*

In this case we have for all  $i \in \{1, \dots, n\}$ ,

$$d(a_i, S)^2 = \|a_i - P_S(a_i)\|^2 = \|a_i\|^2 - \|P_S(a_i)\|^2,$$

by Pythagorean Theorem (recall that  $P_S(a_i) \perp (a_i - P_S(a_i))$ ). Since  $v_1$  is of unit norm,  $P_S(a_i) = \langle v_1, a_i \rangle v_1$ , hence:

$$d(a_i, S)^2 = \|a_i\|^2 - \langle v_1, a_i \rangle^2.$$

Minimizing (8) is therefore equivalent to maximize

$$\sum_{i=1}^n \|P_S(a_i)\|^2. \quad (9)$$

Let us fix an orthonormal basis  $(s_1, \dots, s_k)$  of  $S$ . Then for all  $x \in \mathbb{R}^d$ ,  $P_S(x) = \langle s_1, x \rangle s_1 + \dots + \langle s_k, x \rangle s_k$ , hence

$$\sum_{i=1}^n \|P_S(a_i)\|^2 = \sum_{i=1}^n \sum_{j=1}^k \langle a_i, s_j \rangle^2 = \|As_1\|^2 + \dots + \|As_k\|^2, \quad (10)$$

Consequently, minimizing (8) is equivalent to maximizing (10) over all orthonormal families  $(s_1, \dots, s_k)$ .

For  $k = 1$ , Proposition 2.2 tells us that a subspace of dimension 1 that minimizes (8) is  $\text{Span}(v_1)$  because

$$v_1 = \arg \max_{\|v\|=1} \|Av\|. \quad (11)$$

If we now want to solve the problem for  $k = 2$ , a natural candidate for the subspace  $S$  would be  $S = \text{Span}(v_1, v_2)$  since by Proposition 2.2

$$v_2 = \arg \max_{\|v\|=1, v \perp v_1} \|Av\|. \quad (12)$$

We can follow this greedy strategy for  $k = 3, \dots, n$ ,  $S = \text{Span}(v_1, \dots, v_k)$  is a natural candidate for being solution of (8).

It is not a priori obvious (except for  $k = 1$ ) that  $S = \text{Span}(v_1, \dots, v_k)$  is a minimizer of (8) over all the subspaces of dimension  $k$ . We need the following lemma.

**Lemma 3.1**

Let  $k \in \{2, \dots, n\}$ . Assume that  $(v_1, \dots, v_{k-1})$  is an orthonormal family that maximizes (10). Define

$$v_k = \arg \max_{\|v\|=1, v \perp \text{Span}(v_1, \dots, v_{k-1})} \|Av\|.$$

Then  $(v_1, \dots, v_k)$  is an orthonormal family and  $\text{Span}(v_1, \dots, v_k)$  minimizes (8), i.e.  $(v_1, \dots, v_k)$  maximizes (10).

**Proof.** Let  $S$  be a subspace of dimension  $k$ . Let  $(w_1, \dots, w_k)$  be an orthonormal basis of  $S$  such that  $w_k \perp \text{Span}(v_1, \dots, v_{k-1})$ . By definition of  $v_k$ , we have  $\|Aw_k\| \leq \|Av_k\|$ . We also assumed that  $(v_1, \dots, v_k)$  maximizes (10), so

$$\|Av_1\|^2 + \dots + \|Av_{k-1}\|^2 \geq \|Aw_1\|^2 + \dots + \|Aw_{k-1}\|^2.$$

We conclude that

$$\|Av_1\|^2 + \dots + \|Av_k\|^2 \geq \|Aw_1\|^2 + \dots + \|Aw_k\|^2,$$

so  $(v_1, \dots, v_k)$  maximizes (10). □

Theorem 3.1 follows then by induction.

