

# Optimization and Computational Linear Algebra for Data Science

## Lecture 7: The spectral theorem and PCA

Léo MIOLANE · leo.miolane@gmail.com

October 11, 2020

**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 an 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

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

## 1.1 Some consequences

For  $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ , we introduce the notation

$$\text{Diag}(\lambda_1, \dots, \lambda_n) \stackrel{\text{def}}{=} \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

### Proposition 1.2

Let  $A = P \text{Diag}(\lambda_1, \dots, \lambda_n) P^\top$  (where  $P \in \mathbb{R}^{n \times n}$  is orthogonal). Then

1.  $\lambda_1, \dots, \lambda_n$  are the eigenvalues of  $A$ , and the number of time that an eigenvalue appear on the diagonal equals its multiplicity.
2.  $\text{rank}(A) = \#\{i \mid \lambda_i \neq 0\}$ .
3.  $\text{Tr}(A) = \sum_{i=1}^n \lambda_i$ .
4.  $A$  is invertible if and only if  $\lambda_i \neq 0$  for all  $i$ . In such case,  $A^{-1} = P \text{Diag}(\lambda_1^{-1}, \dots, \lambda_n^{-1}) P^\top$ .

## 1.2 Positive matrices

### Definition 1.1

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

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

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 1.3

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  $A \in \mathbb{R}^{n \times n}$ .

- a. Show that  $A^\top A$  positive semi-definite.
- b. 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 Application: Principal Component Analysis (PCA)

Assume that we are given a dataset of  $n$  points  $a_1, \dots, a_n \in \mathbb{R}^d$ , with  $d$  very large. We aim at representing this dataset in lower dimension, i.e. finding  $\tilde{a}_1, \dots, \tilde{a}_n \in \mathbb{R}^k$  where  $k$  is smaller than  $d$ , such that the points  $(\tilde{a}_1, \dots, \tilde{a}_n)$  look like the original ones  $(a_1, \dots, a_n)$ . This could be for instance used

- to reduce computing time.
- to visualize an high-dimensional dataset in dimension  $k = 2$  or  $3$ .

### 2.1 Sample covariance matrix

Let  $\mu = \frac{1}{n} \sum_{i=1}^n a_i$  be the mean of the dataset. The sample covariance matrix is then defined<sup>1</sup> as:

$$S = \sum_{i=1}^n (a_i - \mu)(a_i - \mu)^\top \in \mathbb{R}^{d \times d}.$$

Assume now that the dataset is centered meaning that  $\sum_{i=1}^n a_i = 0$  (otherwise subtract the mean  $\mu$  to all the points). In that case,  $S$  can be simply written as:

$$S = \sum_{i=1}^n a_i a_i^\top = A^\top A.$$

where  $A$  is the  $n \times d$  “data matrix”:

$$A = \begin{pmatrix} - & a_1 & - \\ & \vdots & \\ - & a_n & - \end{pmatrix}.$$

### 2.2 “Maximal variance” directions

We would like to find a direction, that is a vector  $v \in \mathbb{R}^d$  of unit norm, such that the variance of the projections of the data points onto it is large. More precisely, given a point  $a_i$ , its projection onto  $\text{Span}(v)$  is

$$P_{\text{Span}(v)}(a_i) = \langle v, a_i \rangle v.$$

We aim at maximizing the variance of the coordinates  $\{\alpha_1 \stackrel{\text{def}}{=} \langle v, a_1 \rangle, \dots, \alpha_n \stackrel{\text{def}}{=} \langle v, a_n \rangle\}$  of the points of the dataset in the direction  $v$ . The mean of the  $\alpha_i$  is zero because the dataset is assumed to be centered:

$$\sum_{i=1}^n \alpha_i = \sum_{i=1}^n \langle v, a_i \rangle = \left\langle v, \sum_{i=1}^n a_i \right\rangle = \langle v, 0 \rangle = 0.$$

The variance of the  $\alpha_i$  is then simply (up to a  $1/n$  factor)

$$\sum_{i=1}^n \langle a_i, v \rangle^2 = \sum_{i=1}^n (v^\top a_i)(a_i^\top v) = \sum_{i=1}^n v^\top (a_i a_i^\top) v = v^\top \left( \sum_{i=1}^n a_i a_i^\top \right) v = v^\top S v.$$

Using Proposition 1.1 we get that the direction  $v$  that maximizes the variance is simply the eigenvector  $v_1$  associated to the largest eigenvalue  $\lambda_1$  of  $S = A^\top A$ . The variance along this direction is  $\lambda_1 \geq 0$  (because  $A^\top A$  is positive semi-definite).  $v_1$  is called the first (right) singular vector

---

<sup>1</sup>Strictly speaking, the sample covariance matrix is  $\frac{1}{n-1} S$ . However, we chose for simplicity to remove the factor  $\frac{1}{n-1}$  here, since it will not change the analysis.

of  $A$  and  $\sigma_1 = \sqrt{\lambda_1}$  is called the first singular value of  $A$ .

If we want to reduce the dimension of the dataset from  $d$  to 1, then we are basically done. The “dimensionally reduced” dataset will simply be the coordinates along  $v_1$ :  $\langle v_1, a_1 \rangle, \langle v_1, a_2 \rangle, \dots, \langle v_1, a_n \rangle$ . The inner-product  $\langle v_1, a_i \rangle$  is called the first principal component of  $a_i$ . But in general, we might be interested to obtain a dataset of dimension  $k > 1$ , so we need to find other directions of “large variance”.

In the spirit of what we did above, it is very natural to look for  $v \in \mathbb{R}^d, \|v\| = 1$ , orthogonal to  $v_1$ , along which the variance of the coordinates of the dataset is maximal. That is we want to

$$\text{maximize } v^T S v, \quad \text{subject to } \|v\| = 1, \quad v \perp v_1.$$

Again, from Proposition 1.1, we know that the solution of this problem is given by  $v_2$ , the eigenvector of  $S = A^T A$  associated with the second largest eigenvalue  $\lambda_2$ .  $v_2$  is called the second (right) singular vector of  $A$  and  $\sigma_2 = \sqrt{\lambda_2}$  is called the second singular value of  $A$ .

If our goal was to reduce our dataset to 2 dimensions, then we are done. The new data points with simply be

$$\begin{pmatrix} \langle v_1, a_1 \rangle \\ \langle v_2, a_1 \rangle \end{pmatrix}, \begin{pmatrix} \langle v_1, a_2 \rangle \\ \langle v_2, a_2 \rangle \end{pmatrix}, \begin{pmatrix} \langle v_1, a_3 \rangle \\ \langle v_2, a_3 \rangle \end{pmatrix} \cdots \begin{pmatrix} \langle v_1, a_n \rangle \\ \langle v_2, a_n \rangle \end{pmatrix}.$$

Otherwise, we can continue the same process and construct  $v_3, \dots, v_k$  such that for all  $j \in \{3, \dots, k\}$ ,  $v_j$  is solution of

$$\text{maximize } v^T S v, \quad \text{subject to } \|v\| = 1, \quad v \perp v_1, v \perp v_2, \dots, v \perp v_{j-1}.$$

From Proposition 1.1, we know that  $v_1, \dots, v_k$  will be eigenvectors of  $S = A^T A$  associated with the  $k$  largest eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k$ . For  $j \in \{1, \dots, k\}$  the vector  $v_j$  is called the  $j^{\text{th}}$  (right) singular vector of  $A$  and  $\sigma_j = \sqrt{\lambda_j}$  is called the  $j^{\text{th}}$  singular value of  $A$ . The dimensionally reduced dataset is then

$$\begin{pmatrix} \langle v_1, a_1 \rangle \\ \langle v_2, a_1 \rangle \\ \vdots \\ \langle v_k, a_1 \rangle \end{pmatrix}, \begin{pmatrix} \langle v_1, a_2 \rangle \\ \langle v_2, a_2 \rangle \\ \vdots \\ \langle v_k, a_2 \rangle \end{pmatrix}, \begin{pmatrix} \langle v_1, a_3 \rangle \\ \langle v_2, a_3 \rangle \\ \vdots \\ \langle v_k, a_3 \rangle \end{pmatrix} \cdots \begin{pmatrix} \langle v_1, a_n \rangle \\ \langle v_2, a_n \rangle \\ \vdots \\ \langle v_k, a_n \rangle \end{pmatrix}.$$

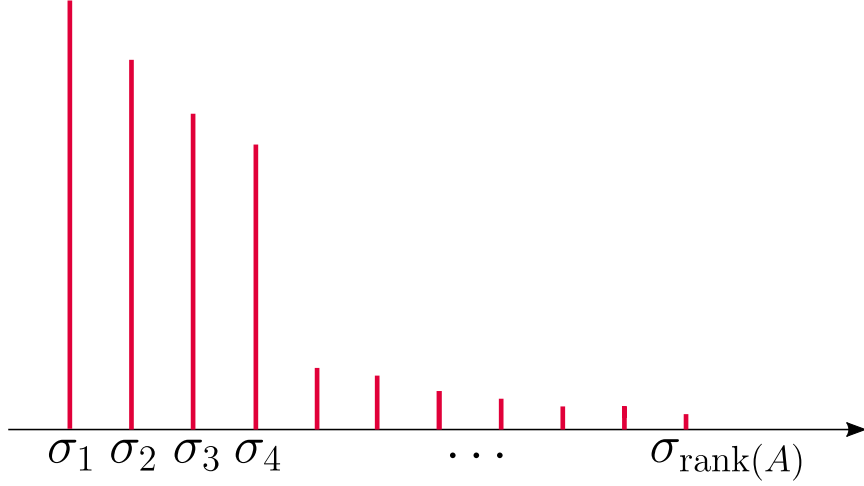
The inner-product  $\langle v_j, a_i \rangle$  is called the  $j^{\text{th}}$  principal component of the vector  $a_i$ .

**How do we chose  $k$  ?** The dimension  $k$  of the dimensionally-reduced data can be chosen by looking at the singular values of  $A$ . Let  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{\min(n,d)}$  be the singular values of  $A$ . As we have seen above, the variance of the dataset along the direction  $v_i$  is  $\lambda_i = \sigma_i^2$ .

A simple way to chose  $k$  is therefore to plot the square singular values as on Figure 1 and look for a good cut-off ( $k = 4$  on Figure 1). Doing so, one captures a fraction

$$\frac{\sum_{i=1}^k \sigma_i^2}{\sum_{i=1}^{\min(n,d)} \sigma_i^2}$$

of the total variance.



**Figure 1:** Singular values of  $A$ , ranked in decreasing order.

**Should we “normalize” the dataset?** It depends, but in general the answer is yes, especially if you have data from heterogeneous types. Imagine that you have measured the size and the weight of  $n$  objects, and stored the information in vectors  $a_i = (\text{size of object } i \text{ in cm, weight of object } i \text{ in kg})$ . If I change the weighting unit from kilograms to grams, this multiply the variance along the second coordinates by  $10^6$ , leading to very different principal components. Normalizing the dataset (i.e. dividing the columns of the data matrix by their standard deviation) allows to be unaffected by a change of units.

However, this decreases a lot the variance of columns which high variance and amplify a lot the variance of columns with low variance. Hence you may not always want to normalize the columns.

### 3 Singular value decomposition

Given a matrix  $A$ , we can follow the procedure of Section 2.2 to compute the right-singular vectors  $v_j$  and the associated singular values  $\sigma_j$ . When  $\sigma_j \neq 0$  we can define  $u_j$  the left-singular vector  $u_j$  of  $A$  by

$$u_j = \frac{1}{\sigma_j} A v_j.$$

It turns out that the vectors  $u_j$ ,  $v_j$  and the  $\sigma_j$  allow to obtain a decomposition of the matrix  $A$ , called the singular value decomposition:

**Theorem 3.1 (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  $i \neq j$

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

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  $\text{rank}(A) = \#\{i \mid \Sigma_{i,i} \neq 0\}$ .

Theorem 3.1 is proved at the end of these notes.

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  $A$ .

### 3.1 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 3.1. 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 3.1

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 3.2

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

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

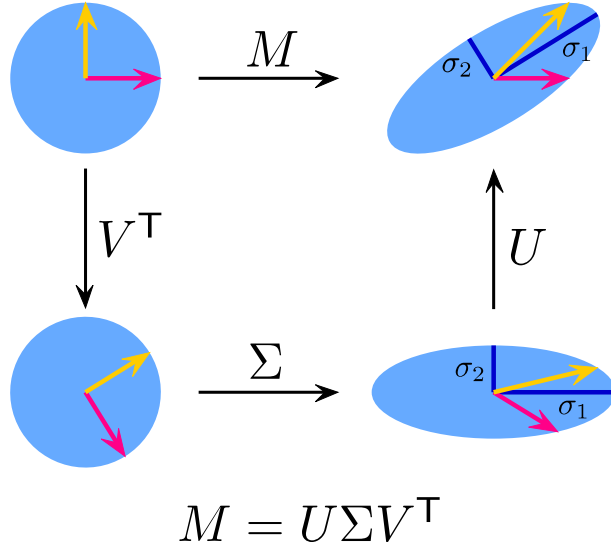
**Proof.** Compute  $A^\top A = V\Sigma^\top \Sigma V^\top = V D V^\top$  where the matrix  $D \stackrel{\text{def}}{=} \Sigma^\top \Sigma$  is diagonal with  $D_{i,i} = \sigma_i^2$ . The family  $(v_1, \dots, v_m)$  is therefore an orthonormal family of eigenvectors of the symmetric matrix  $A^\top A$  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  $v^\top A^\top A v = \|Av\|^2$ .  $\square$

## 4 Interpretations of the SVD

### 4.1 Geometric interpretation

The decomposition  $M = U\Sigma V^\top$  gives that the linear transformation associated to the matrix  $M$  is the composition of three linear transformations:

1.  $V^\top$  is a rotation/reflection: length, dot products, angles are preserved.
2.  $\Sigma$  corresponds to a scaling.
3.  $U$  is another rotation/reflection: length, dot products, angles are preserved.



**Figure 2:** Geometric interpretation of SVD

## 4.2 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:

$$\text{minimize } \sum_{i=1}^n d(a_i, S)^2 \text{ with respect to } S \text{ subspace of dimension } k. \quad (6)$$

Recall that the distance of a vector  $x$  to the subspace  $S$  is defined as  $d(x, S) = \|x - P_S x\|$ . 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 4.1

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 (6).

We start by noticing that 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  $P_S(a_i) \perp (a_i - P_S(a_i))$ ). Minimizing (6) is therefore equivalent to maximize

$$\sum_{i=1}^n \|P_S(a_i)\|^2. \quad (7)$$

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

$$\sum_{i=1}^n \|P_S(a_i)\|^2 = \sum_{i=1}^n \sum_{j=1}^k \langle a_i, s_j \rangle^2 = \|A s_1\|^2 + \dots + \|A s_k\|^2. \quad (8)$$

Consequently, minimizing (6) is equivalent to maximizing (8) over all orthonormal families  $(s_1, \dots, s_k)$ .

For simplicity, we start by considering the case  $k = 1$ , in which case  $S = \text{Span}(s_1)$ . In this case, (8) is simply:

$$\sum_{i=1}^n \|P_S(a_i)\|^2 = \|As_1\|^2. \quad (9)$$

Proposition 3.2 tells us that a subspace of dimension 1 that maximizes (9) and hence that minimizes (6) is  $\text{Span}(v_1)$  because

$$v_1 = \arg \max_{\|v\|=1} \|Av\|. \quad (10)$$

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 3.2

$$v_2 = \arg \max_{\|v\|=1, v \perp v_1} \|Av\|. \quad (11)$$

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 (6).

It is not a priori obvious (except for  $k = 1$ ) that  $S = \text{Span}(v_1, \dots, v_k)$  is a minimizer of (6) over all the subspaces of dimension  $k$ . We need the following lemma.

**Lemma 4.1**

Let  $k \in \{2, \dots, k\}$ . Assume that  $(v_1, \dots, v_{k-1})$  is an orthonormal family that maximizes (8). 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 (6), i.e.  $(v_1, \dots, v_k)$  maximizes (8).

**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  $(v_1, \dots, v_{k-1})$  maximizes (8), 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 (8). □

Theorem 4.1 follows then by induction.

## Proof of Theorem 3.1

We apply the Spectral Theorem (Theorem 1.1) to the  $m \times m$  matrix  $A^T A$ : there exists an orthonormal basis  $(v_1, \dots, v_m)$  of  $\mathbb{R}^m$  of eigenvectors of  $A^T A$  associated to eigenvalues  $\lambda_1 \geq \dots \geq \lambda_m$  that are all non-negative because  $A^T A$  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  $i = 1, \dots, r$

$$u_i = \frac{1}{\sigma_i} Av_i \in \mathbb{R}^n. \quad (12)$$



**Lemma 4.2**

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  $A^\top A v_i = \sigma_i^2 v_i$ . □

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 4.3**

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

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

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

□

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 (12) and Lemma 4.3:

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

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  $i \in \{1, \dots, m\}$ ,  $A v_i = U \Sigma V^\top v_i$ . 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 4.4**

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

**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 4.3  $v_i \in \text{Ker}(A)$  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:

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

□

