Optimization and Computational Linear Algebra for Data Science
Lecture 7: The spectral theorem and PCA

Léo Miolane · leo.miolane@gmail.com

November 7, 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 $D = \mathrm{Diag}(\lambda_1, \dots, \lambda_n)$ and compute

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

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

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

$$A = PDP^\mathsf{T}.$$

Proposition 1.1

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

$$v_1 = \mathop{\arg\max}_{\|v\|=1} v^\mathsf{T} Av\,, \qquad \text{and for } k = 2, \dots n, \qquad v_k = \mathop{\arg\max}_{\|v\|=1,\, v\perp v_1,\dots,v_{k-1}} v^\mathsf{T} Av.$$

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

$$v_n = \mathop{\arg\min}_{\|v\|=1} v^\mathsf{T} Av\,, \qquad \text{and for } k = 1, \dots, n-1 \qquad v_k = \mathop{\arg\min}_{\|v\|=1,\, v\perp v_{k+1},\dots,v_n} v^\mathsf{T} Av.$$

Positive matrices

Definition 1.1

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

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

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

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 $\widetilde{a}_1, \dots, \widetilde{a}_n \in \mathbb{R}^k$ where k is smaller than d, such that the points $(\widetilde{a}_1, \dots, \widetilde{a}_n)$ 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[1] 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 μ to all the points). In that case, S can be simply written as:

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

[1] 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.

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 $\mathrm{Span}(v)$ is

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

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 α_i is zero because the dataset is assumed to be centered:

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

The variance of the α_i is then simply

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

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 λ_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 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

$$\text{maximize} \;\; v^\top S v, \qquad \text{subject to} \;\; \|v\| = 1, \;\; v \perp v_1.$$

Again, from Proposition 1.1, we know that the solution of this problem is given by v_2, the eigenvector of $S = A^\top A$ associated with the second largest eigenvalue λ_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

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

From Proposition 1.1, we know that $v_1, \dots, v_k$ will be eigenvectors of $S = A^\top A$ 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^{th} (right) singular vector of A and $\sigma_j = \sqrt{\lambda_j}$ is called the j^{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^{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 $\lambda_i = \sigma_i^2$.

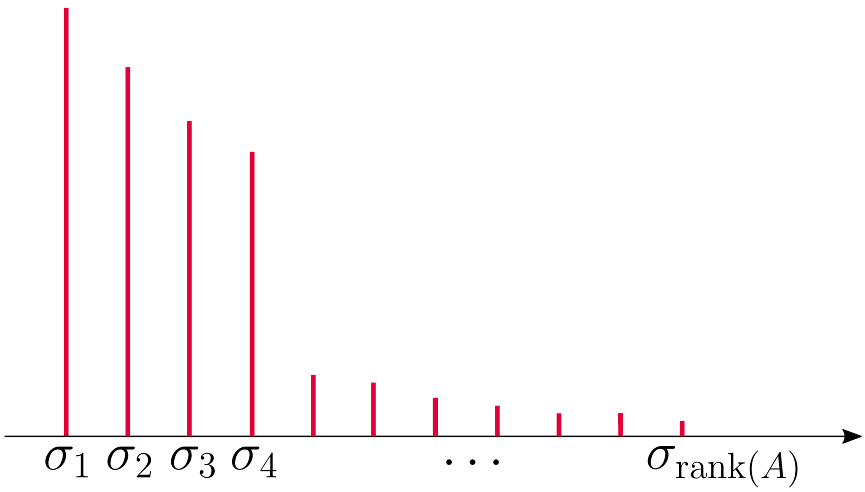

Figure 1: Singular values of A, ranked in decreasing order.

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.

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 $a_i = (\text{size of object } i \text{ in cm}, \text{ weight of object } i \text{ in kg})$. 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 σ_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 σ_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 $\Sigma_{1,1} \geq \Sigma_{2,2} \geq \cdots \geq 0$ and $\Sigma_{i,j} = 0$ for $i \neq j$

$$A = U\Sigma V^{\mathsf{T}}.$$

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

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 $A = U\Sigma V^{\mathsf{T}}$ is a SVD of A, then $A = (-U)\Sigma(-V)^{\mathsf{T}}$ is also a SVD of A. However, with a slight abuse of language, we will often refer v_i as the i^{th} right singular vector of A.

3.1 Properties of the SVD

Let $A \in \mathbb{R}^{n\times m}$ and let $U\Sigma V^{\mathsf{T}}$ 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 $i = 1, \dots, \operatorname{rank}(A)$ we have

$$A v_i = \sigma_i u_i \qquad \text{and} \qquad A^{\mathsf{T}} u_i = \sigma_i v_i.$$

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

Proposition 3.2

We have

$$v_1 = \operatorname*{arg\,max}_{\|v\|=1} \|Av\| \qquad \text{and} \qquad \sigma_1 = \max_{\|v\|=1} \|Av\|. \tag{2}$$

It holds also that

$$v_2 = \operatorname*{arg\,max}_{\|v\|=1,\, v\perp v_1} \|Av\| \qquad \text{and} \qquad \sigma_2 = \max_{\|v\|=1,\, v\perp v_1} \|Av\| \tag{3}$$

and more generally:

$$v_k = \operatorname*{arg\,max}_{\|v\|=1,\, v\perp v_1,\dots,v_{k-1}} \|Av\|. \qquad \text{and} \qquad \sigma_k = \max_{\|v\|=1,\, v\perp v_1,\dots,v_{k-1}} \|Av\|. \tag{4}$$

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

$$u_k = \underset{\|u\|=1,\, u\perp u_1,\dots,u_{k-1}}{\arg\max} \|A^\top u\|. \qquad \textit{and} \qquad \sigma_k = \max_{\|u\|=1,\, u\perp u_1,\dots,u_{k-1}} \|A^\top u\|. \tag{5}$$

Proof. Compute $A^\top A = V\Sigma^\top \Sigma V^\top = VDV^\top$ where the matrix $D \overset{\text{def}}{=} \Sigma^\top\Sigma$ 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 $\sigma_1^2 \geq \cdot \geq \sigma_m^2$ 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. Σ corresponds to a scaling.
3. U is another rotation/reflection: length, dot products, angles are preserved.

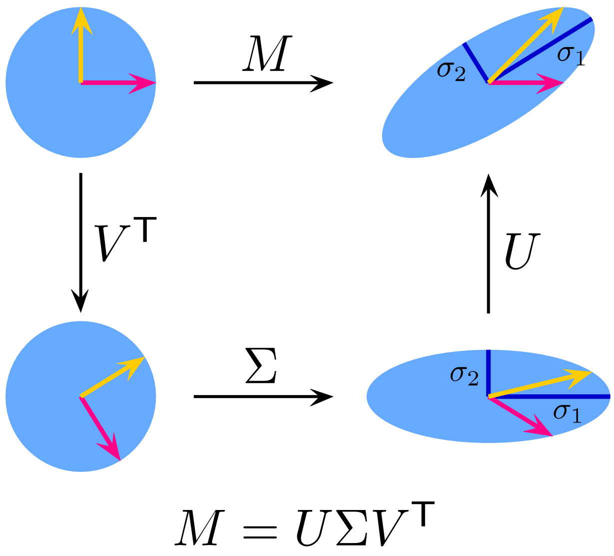

Figure 2: Geometric interpretation of SVD

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:

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

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 $\mathrm{Span}(v_1, \dots, v_k)$ *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

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

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

$$\sum_{i=1}^{n} \|P_S(a_i)\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{k} \langle a_i, s_j \rangle^2 = \|As_1\|^2 + \cdots + \|As_k\|^2. \tag{8}$$

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 $S = \mathrm{Span}(s_1)$. In this case, (8) is simply:

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

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

$$v_1 = \operatorname*{arg\,max}_{\|v\|=1} \|Av\|. \tag{10}$$

If we now want to solve the problem for $k = 2$, a natural candidate for the subspace S would be $S = \mathrm{Span}(v_1, v_2)$ since by Proposition 3.2

$$v_2 = \operatorname*{arg\,max}_{\|v\|=1,\, v \perp v_1} \|Av\|. \tag{11}$$

We can follow this greedy strategy for $k = 3, \dots, n$, $S = \mathrm{Span}(v_1, \dots, v_k)$ is a natural candidate for being solution of (6).

It is not a priori obvious (except for $k = 1$) that $S = \mathrm{Span}(v_1, \dots, v_k)$ 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*

$$v_k = \operatorname*{arg\,max}_{\|v\|=1,\, v \perp \mathrm{Span}(v_1, \dots, v_{k-1})} \|Av\|.$$

Then $(v_1, \dots, v_k)$ is an orthonormal family and $\mathrm{Span}(v_1, \dots v_k)$ *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 $w_k \perp \mathrm{Span}(v_1, \dots, v_{k-1})$. 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

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

We conclude that

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

so $(v_1, \ldots, v_k)$ 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 $A^\mathsf{T} A$: there exists an orthonormal basis $(v_1, \ldots, v_m)$ of $\mathbb{R}^m$ of eigenvectors of $A^\mathsf{T} A$ associated to eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m$ that are all non-negative because $A^\mathsf{T} A$ is non-negative. Let $V \in \mathbb{R}^{m \times m}$ be the orthogonal matrix whose columns are $(v_1, \ldots, v_m)$.

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

$$u_i = \frac{1}{\sigma_i} A v_i \in \mathbb{R}^n. \tag{12}$$

Lemma 4.2

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

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

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

since $A^\mathsf{T} A v_i = \sigma_i^2 v_i$. □

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

Lemma 4.3

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

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

$$\|Av_i\|^2 = v_i^\mathsf{T} A^\mathsf{T} A^\mathsf{T} v_i = v_i^\mathsf{T} (\lambda_i v_i) = \sigma_i^2 = 0.$$

□

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

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

It remains to verify that $A = U \Sigma V^\mathsf{T}$. Compute for $i = 1, \ldots, m$, using the definition (12) and Lemma 4.3:

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

By orthogonality of V and the construction of Σ one verifies easily that

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

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

It remains to show:

Lemma 4.4

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

Proof. The family $(u_1, \ldots, u_r)$ is orthonormal, hence linearly independent. By definition $u_i \in \mathrm{Im}(A)$ which implies that $\mathrm{rank}(A) = \dim(\mathrm{Im}(A)) \geq r$. To prove the converse inequality, notice that by Lemma 4.3 $v_i \in \mathrm{Ker}(A)$ for $i = r+1, \ldots, m$. The vectors $(v_{r+1}, \ldots, v_m)$ are orthonormal, hence linearly independent. This implies that $\dim(\mathrm{Ker}(A)) \geq m - r$. We conclude by applying the rank Theorem:

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

□