Optimization and Computational Linear Algebra for Data Science
Lecture 6: Singular value decomposition

Léo Miolane · leo.miolane@gmail.com

May 15, 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 Eigenvalues and eigenvectors

Definition 1.1

*Let $A \in \mathbb{R}^{n \times n}$. A **non-zero** vector $v \in \mathbb{R}^n$ is said to be an* eigenvector *of A is there exists $\lambda \in \mathbb{R}$ such that*

$$Av = \lambda v.$$

The scalar λ is called the eigenvalue (of A) associated to v.

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

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

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*

$$v_n = \operatorname*{arg\,min}_{\|v\|=1} v^\top A v\,, \qquad \textit{and for } k = 1, \dots, n-1 \qquad v_k = \operatorname*{arg\,min}_{\|v\|=1,\, v \perp v_{k+1}, \dots, v_n} v^\top A v.$$

2 Singular value decomposition

Let $a_1, \dots, a_n \in \mathbb{R}^d$ be n points in d dimension.

The goal of Singular Value Decomposition (SVD) is to find 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:

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

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 (1) is therefore equivalent to maximize

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

Let us fix an orthonormal basis $(v_1, \dots, v_k)$ of S. Then for all $x \in \mathbb{R}^d$, $P_S(x) = \langle v_1, x \rangle v_1 + \cdots + \langle v_k, x \rangle v_k$, hence

$$\sum_{i=1}^{n} \|P_S(a_i)\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{k} \langle a_i, v_j \rangle^2 = \|Av_1\|^2 + \cdots + \|Av_k\|^2, \tag{3}$$

where A is the $n \times d$ matrix whose rows are $a_1, \dots, a_n$. Consequently, minimizing (1) is equivalent to maximizing (3) over all orthonormal families $(v_1, \dots, v_k)$.

For $k = 1$, a subspace of dimension 1 that minimizes (1) is therefore $\mathrm{Span}(v_1)$ where

$$v_1 \stackrel{\text{def}}{=} \operatorname*{arg\,max}_{\|v\|=1} \|Av\|. \tag{4}$$

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

$$v_2 \stackrel{\text{def}}{=} \operatorname*{arg\,max}_{\|v\|=1,\, v \perp v_1} \|Av\|. \tag{5}$$

We can follow this greedy strategy for $k = 3, \dots, n$ and define recursively

$$v_k \stackrel{\text{def}}{=} \operatorname*{arg\,max}_{\|v\|=1,\, v \perp v_1, \dots, v_{k-1}} \|Av\|. \tag{6}$$

Definition 2.1

- *The vectors* $v_1, \dots, v_n$ *are called* singular vectors *of the matrix* A.
- *The non-negative numbers* $\sigma_k \stackrel{\text{def}}{=} \|Av_k\|$ *are called the* singular values *of* A.

Of course (4)-(6) admits many other maximizers (for instance $-v_k$), so **the singular vectors are not uniquely defined**.

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

Lemma 2.1

Let $k \in \{2, \dots, k\}$. *Assume that* $(v_1, \dots, v_{k-1})$ *is an orthonormal family that maximizes* (3). *Define*

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

Then $(v_1, \dots, v_k)$ *is an orthonormal family and* $\operatorname{Span}(v_1, \dots v_k)$ *minimizes* (1), *i.e.* $(v_1, \dots, v_k)$ *maximizes* (3).

Proof. Let S be a subspace of dimension k. Let $(w_1, \dots, w_k)$ be an orthonormal basis of S such that $w_k \perp \operatorname{Span}(v_1, \dots, v_{k-1})$. By definition of v_k, we have $\|Aw_k\| \leq \|Av_k\|$. We also assumed that $(v_1, \dots, v_k)$ maximizes (3), 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 $(v_1, \dots, v_k)$ maximizes (3). $\square$

Using Lemma 2.1 we get by induction:

Proposition 2.1

Let $v_1, \dots, v_n$ *be singular vectors of* A *defined by* (4)-(6). *Then for all* $k \in \{1, \dots, n\}$, *the subspace* $\operatorname{Span}(v_1, \dots, v_k)$ *is a solution of* (1).