Singular Value Decomposition

An Introduction to a Useful Type of Matrix Factorization

Anders Alexander Andersen

October 18, 2019

Table of Contents

Motivation - The Spectral Theorem

Theorem (The Spectral Theorem)

Let $A \in M_{n\times n}(\mathbb{R})$. Then A is diagonalizable, meaning there exists a diagonal matrix Λ (with the eigenvalues of A on the diagonal) and an orthogonal matrix U such that

$$A = U\Lambda U^{-1} = U\Lambda U^{T},$$

if and only if A is normal (meaning $AA^T = A^TA$).

Motivation

Let $A \in M_{m \times n}(\mathbb{R})$. If $m \neq n$, it no longer makes sense to ask if A can be diagonalized. However, one can raise the question of whether there exist two different orthogonal matrices U and V such that

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

and where Σ is a diagonal (but rectangular) matrix. It turns out that the answer to this question is yes, and that the specific factorization, known as the singular value decomposition, is closely related to the diagonalization of the normal matrix AA^T (or similarly $A^T A$).

Some useful propositions

Proposition

A symmetric matrix $A \in M_{n\times n}(\mathbb{R})$ is positive definite if and only if all its eigenvalues are positive. Similarly, A is positive semi-definite if and only if all its eigenvalues are non-negative.

Proof.

$(\Rightarrow)$ $\lambda \left\|x\right\|^2 = \langle \lambda x, x\rangle = \langle Ax, x\rangle > 0 \Rightarrow \lambda > 0.$
$(\Leftarrow)$ $\langle Ax, x\rangle = \langle U^T \Lambda U x, x\rangle = \langle \Lambda U x, U x\rangle = \langle \Lambda y, y\rangle =$
$\lambda_1 \left\|y_1\right\|^2 + \cdots + \lambda_n \left\|y_n\right\|^2 > 0.$ □

Lemma

For any $A \in M_{m\times n}(\mathbb{R})$ and $A \in M_{n\times m}(\mathbb{R})$, the matrices AB and BA have the same non-zero eigenvalues.

Proof.

Let $ABx = \lambda x$. Then, if $\lambda \neq 0$ (hence $Bx = 0$ is impossible), we have $BA(Bx) = B(ABx) = B\lambda x = \lambda Bx$. Vice versa for BA. □

A useful corollary

Corollary

Let $A \in M_{m\times n}(\mathbb{R})$. Then the $(n \times n)$ matrix $A^T A$ and the $(m \times m)$ matrix AA^T are symmetric with non-negative eigenvalues, and the positive eigenvalues of the two matrices coincide.

Singular Values

Definition

Let $A \in M_{m\times n}(\mathbb{R})$ have rank r. Let $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_r^2$ be the positive eigenvalues of $A^T A$. The scalars $\sigma_1, \ldots, \sigma_r$ are called the positive singular values of A.

Table of Contents

Singualr Value Decomposition

Theorem (Singualr Value Decomposition)

Suppose $A \in M_{m\times n}(\mathbb{R})$ *is of rank* r, *and let* $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$ *be the positive singular values of* A. *Let* Σ *be the* $(m \times n)$ *matrix defined by*

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

Then there exists an $(m \times m)$ *orthogonal matrix* U *and an* $(n \times n)$ *orthogonal matrix* V *such that*

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

The Proof - Part I

The matrix $A^T A$ is symmetric with positive eigenvalues $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_r^2$ and $(n-r)$ eigenvalues equal to zero. Thus, by the Spectral Theorem, there exists an $(n \times n)$ orthogonal matrix V such that

$$V^T A^T A V = (AV)^T (AV) = D,$$

where $D = \Sigma^T \Sigma$ is the $(n \times n)$ diagonal matrix with

$$D_{ii} = \sigma_i^2, \quad i = 1, \ldots r,$$

and zeros elsewhere. Note that we can see that the column vectors of AV must be pairwise orthogonal. Also, for $1 \leq j \leq r$, the j-th column vector, $(AV)_j$, has length σ_j.

The Proof - Part II

Let U' denote the $(m \times r)$ matrix with $(AV)_j/\sigma_j$ as its j-th column. Complete U' to an $(m \times m)$ orthogonal matrix U by finding an orthonormal basis (via, for example, Gram-Schmidt) for the orthogonal complement $C(U')^\perp \subseteq \mathbb{R}^n$ of (the column space of) U', and using these basis vectors as the last $(m-r)$ columns in U. We then have

$$AV = U\Sigma \quad \Longleftrightarrow \quad A = U\Sigma V^T.$$

□

Note that although there are many choices when choosing U and V, the matrix Σ is always unique.

Example of computation

We will find the SVD of A, where

$$A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}.$$

We have that

$$A^T A = \begin{bmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} = \begin{bmatrix} 13 & 12 & 2 \\ 12 & 13 & -2 \\ 2 & -2 & 8 \end{bmatrix},$$

which has the eigenvalues $\sigma_1^2 = 25$, $\sigma_2^2 = 9$, and $\sigma_3^2 = 0$.

Example of computation

To find the eigenvectors, we compute

$$\sigma_1 : \begin{bmatrix} 13-25 & 12 & 2 \\ 12 & 13-25 & -2 \\ 2 & -2 & 8-25 \end{bmatrix} \sim \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & -1 & -\frac{17}{2} \end{bmatrix} \Rightarrow v_1 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \\ 0 \end{bmatrix}$$

$$\sigma_2 : \begin{bmatrix} 13-9 & 12 & 2 \\ 12 & 13-9 & -2 \\ 2 & -2 & 8-9 \end{bmatrix} \sim \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & \frac{1}{4} \\ 1 & 0 & -\frac{1}{4} \end{bmatrix} \Rightarrow v_2 = \begin{bmatrix} \frac{\sqrt{2}}{6} \\ -\frac{\sqrt{2}}{6} \\ \frac{2\sqrt{2}}{3} \end{bmatrix}$$

$$\sigma_3 : \begin{bmatrix} 13 & 12 & 2 \\ 12 & 13 & -2 \\ 2 & -2 & 8 \end{bmatrix} \sim \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -2 \\ 1 & 0 & 2 \end{bmatrix} \Rightarrow v_3 = \begin{bmatrix} \frac{2}{3} \\ -\frac{2}{3} \\ -\frac{1}{3} \end{bmatrix}$$

Example of computation

The matrix V is built from the eigenvectors v_1, v_2 and v_3, i.e.

$$V = [v_1 \mid v_2 \mid v_3] = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{6} & \frac{2}{3} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{6} & -\frac{2}{3} \\ 0 & \frac{2\sqrt{2}}{3} & -\frac{1}{3} \end{bmatrix},$$

and

$$\Sigma = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix}.$$

We compute U as follows:

$$U = \left[\frac{Av_1}{\|Av_1\|} \mid \frac{Av_2}{\|Av_2\|} \right] = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix}.$$

Example of computation

Finally, we get

$$A = U\Sigma V^T = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{2}}{6} & -\frac{\sqrt{2}}{6} & \frac{2\sqrt{2}}{3} \\ \frac{2}{3} & -\frac{2}{3} & -\frac{1}{3} \end{bmatrix}.$$

It should be noted that one usually never compute this by hand because computers can.

Table of Contents

Some basic results

Proposition

If $U\Sigma V^T$ is a SVD of A with
$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq \sigma_{r+1} = \ldots = \sigma_{\min(\{m,n\})} = 0$,
$V = [v_1 \mid \ldots \mid v_n]$, and $U = [u_1 \mid \ldots \mid u_m]$, we have the following statements:

- $\text{rank}(A) = r$;
- $\text{null}(A) = \text{span}(\{v_{r+1}, \ldots v_n\})$;
- $\text{range}(A) = \text{span}(\{u_1, \ldots u_r\})$;
- $A = \Sigma_{j=1}^{r} \sigma_j u_j v_j^T$;
- $\|A\|_F = \sqrt{\text{trace}(A^T A)} = \sqrt{\sigma_1^2 + \cdots + \sigma_p^2}$;
- $\|A\|_\infty = \sigma_1$;
- $\sigma_j = \sqrt{\lambda_j(A^T A)}$; and
- The v_i's are orthonormalized vectors of $A^T A$ and the u_i's are orthonormalized vectors of AA^T.

Polar decompsition

Corollary

For any square matrix A, there exists a unitary matrix W and a positive semi-definite matrix P such that

$$A = WP.$$

Proof.

$$A = U\Sigma V^T = U(V^T V)\Sigma V^T = (UV^T)(V\Sigma V^T) = WP.$$

□

Image Processing

Suppose a satellite takes a picture, and wants to send it to earth. The picture may contain 1000×1000 pixels - a million little squares each with a definite color. We can code the colors, and send back 1 000 000 numbers. However, it is more convenient if we can find the essential information, and send only this.

Suppose we know the SVD, and specifically the matrix of singular values Σ. Typically, some of the σ's are significant, whereas others are extremely small. If we keep, say, 20 singular values, and discard the remaining 980, then we need only send the corresponding 20 columns of U and V . Thus, if only 20 singular values are kept, we send 20×2000 numbers rather than a million (and this is a 25 to 1 compression).

There is, of course, the additional cost of computing the SVD. This has become quite efficient, but is still expensive for big matrices.

Pseudoinverse

Definition

Let V and W be finite-dimensional inner product spaces over the same field, and let $T\colon V \to W$ be a linear transformation. Let $L\colon \text{null}(T)^{\perp} \to \text{range}(T)$ be the linear transformation defined by $L(x) = T(x)$ for all $x \in \text{null}(T)^{\perp}$. The pseudoinverse of T, denoted T^{+}, is defined as the unique linear transformation from W to V such that

$$T^{+}(y) = \begin{cases} L^{-1}(y), & \text{if } y \in \text{range}(T) \\ 0, & \text{if } y \in \text{range}(T)^{\perp}. \end{cases}$$

Note that if T is invertible, then $T^{+} = T^{-1}$, because $\text{null}(T)^{\perp} = V$ and L coincides with T.

Pseudoinverse

Theorem

Let $A \in M_{m\times n}(\mathbb{R})$ *have rank* r *and singular value decomposition* $A = U\Sigma V^T$, *where* $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$ *are the positive singular values of* A. *Let* Σ^+ *be the* $(n \times m)$ *matrix*

$$\Sigma^+ = \begin{cases} \frac{1}{\sigma_i}, & \text{if } i = j \leq r \\ 0, & \text{otherwise.} \end{cases}$$

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

Why care about pseudoinverses?

Consider the system of linear equations

$$Ax = b.$$

There are three scenarios:

1. A invertible $\implies x = A^{-1}b = A^{+}b$;
2. A consistent (i.e. $b \in \text{range}(A)$) $\implies z = A^{+}b$ is the unique solution to the system having minimum norm (i.e. for any z' also solving $Az' = b$, we have $\|z'\| > \|z\|$); and
3. A inconsistent (i.e. $b \notin \text{range}(A)$) $\implies z = A^{+}b$ is the unique best approximation to a solution having minimum norm. That is

$$\|Az - b\| \leq \|Ay - b\| \quad \text{for any } y \in \mathbb{R}^n,$$

with equality if and only if $Ay = Az$. Moreover, if $Ay = Az$, then $\|z\| \leq \|y\|$, with equality if and only if $z = y$.

Lemma

Lemma

Let V and W be finite-dimensional inner product spaces, and let $T\colon V \to W$ be linear. Then

1. T^+T is the orthogonal projection of V on $\mathrm{null}(T)^\perp$; and
2. TT^+ is the orthogonal projection of W on $\mathrm{range}(T)$.

Proof.

As above, we define $L\colon \mathrm{null}(T)^\perp \to \mathrm{range}(T)$ by $L(x) = T(x)$ for $x \in \mathrm{null}(T)^\perp$. So,

$$\begin{cases} x \in \mathrm{null}(T)^\perp & \Longrightarrow \quad T^+T(x) = L^{-1}L(x) = x, \\ x \in \mathrm{null}(T) & \Longrightarrow \quad T^+T(x) = T(0) = 0. \end{cases}$$

Consequently, T^+T is the orthogonal projection of V on $\mathrm{null}(T)^\perp$. This proves part (1). Part (2) is proved similarly. $\square$

Pseudoinverses - proof of 2

Let $z = A^+b$, and $Ax = b$ be consistent, i.e. $b \in \text{range}(A)$. Then

$$Az = AA^+b = TT^+b = LL^{-1}b = b,$$

by point 1 of the lemma. Thus, z is a solution to the system $Ax = b$. Now let y be any solution to the system. Then

$$A^+Ay = A^+b = z.$$

Thus, z is the orthogonal projection of y on $\text{null}(T)^\perp$. We also have $y = z + v$ with $v \in \text{null}(T)$ (since $V = \text{null}(T) \oplus \text{null}(T)^\perp$), and

$$\|y\|^2 = \|z\|^2 + \|v\|^2 .$$

It follows that $\|y\| > \|z\|$ unless $v = 0$ and $y = z$.
□

Pseudoinverses - proof of 3

Let $z = A^+b$, and $Ax = b$ be inconsistent, i.e. $b \notin \text{range}(A)$. Then

$$Az = AA^+b = TT^+b,$$

(by the lemma) is the orthogonal projection of b on range(T). Therefore, Az is the vector in range(T) nearest b. If Ay is any other vector in range(T), then necessarily

$$\|Az - b\| \leq \|Ay - b\|,$$

with equality if and only if $Az = Ay$. Finally, suppose that $y \in \mathbb{R}^n$ such that $Az = Ay = c$. Then

$$A^+c = A^+Az = A^+AA^+b = A^+b = z,$$

since $A^+AA^+ = A^+$. Hence, by previous slide, we can conclude that $\|y\| \geq \|z\|$ with equality if and only if $y = z$.

$\square$

Thank you for your attention :)