Optimization and Data Science

Lecture 8: Singular Value Decomposition

Prof. Dr. Thomas Slawig

Kiel University - CAU Kiel
Dep. of Computer Science

Summer 2020

Contents

Contents

Singular Value Decomposition

- What is it?
  Decomposition of an arbitrary matrix into two orthogonal matrices and a diagonal matrix
  Method to solve linear regression problems and for data analysis and compression
- Why are we studying this?
  Alternative to normal equation method for linear regression
  Basis of the Principal Component Analysis, widely used in data science
- How does it work?
  Mathematical result on the existence of the decomposition
  Exploitation of properties of orthogonal matrices
- What if we can use it?
  Solve linear regression problems
  Detect structures in data
  Detect dominant “modes” in data
  Reduce data size

Contents

Singular value decomposition (SVD)

Theorem (Singular value decomposition)

Every matrix $A \in \mathbb{R}^{m\times n}$, $m \geq n$, can be decomposed in the form

$$A = U \underbrace{\begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \\ 0 & \cdots & 0 \end{pmatrix}}_{=\Sigma} V^\top$$

where

- *$U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices,*
- *$\Sigma \in \mathbb{R}^{m\times n}$ with $\Sigma_{ij} = \left\{ \begin{array}{ll} \sigma_j \geq 0, & i = j, \\ 0, & i \neq j \end{array} \right\}, i = 1, \dots, m, j = 1, \dots, n.$*
- *The σ_j are ordered by magnitude, i.e., $\sigma_j \geq \sigma_{j+1}$ for all j.*
- *The decomposition also exists for $m < n$.*

Orthogonal matrices

Definition

A matrix $U \in \mathbb{R}^{n\times n}$ is called **orthogonal**, if its rows and columns build an orthonormal system of vectors, i.e., it holds

$$U_{i*}^\top U_{j*} = U_{*i}^\top U_{*j} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}, i,j = 1,\dots,n.$$

- Rotation in $\mathbb{R}^2$ by angle α:

$$U = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}, \quad \begin{array}{l} U_{i*}^\top U_{i*} = U_{*i}^\top U_{*i} = \cos^2\alpha + \sin^2\alpha = 1, i = 1,2, \\ U_{i*}^\top U_{j*} = U_{*i}^\top U_{*j} = 0, i \neq j. \end{array}$$

- Reflection in $\mathbb{R}^2$ around x_1-axis:

$$U = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Properties of orthogonal matrices

Lemma

An orthogonal matrix $U \in \mathbb{R}^{n\times n}$ *satisfies* $U^\top U = UU^\top = I$, *i.e.,* $U^{-1} = U^\top$.

Proof.

- For $E = (E_{ij})_{i,j=1}^n := AU$ we have with the rules of matrix multiplication $E_{ij} = A_{i*}^\top U_{*j}$.
- For $A = U^\top$ we have $A_{i*} = U_{*i}$, i.e.,

$$E_{ij} = A_{i*}^\top U_{*j} = U_{*i}^\top U_{*j} = \begin{cases} 0, & i \neq j, \\ 1, & i = j, \end{cases}$$

  because U has orthonormal columns.
- Thus, E is the identity matrix.

$\square$

Properties of orthogonal matrices

Lemma

For a orthogonal matrix $U \in \mathbb{R}^{n\times n}$ we have

$$\left.\begin{array}{rcl} (Ux)^\top Uy & = & x^\top y \\ \|Ux\|_2 & = & \|x\|_2 \end{array}\right\} \quad \textit{for all } x, y \in \mathbb{R}^n,$$

i.e., it preserves length of vectors and angles between vectors. If U is orthogonal, so is $U^\top$.

Proof.

This follows from

$$(Ux)^\top Uy = x^\top U^\top Uy = x^\top U^{-1} Uy = x^\top y.$$

The angle between two vectors is defined by the scalar product. The results for $U^\top$ follows directly from the definition. $\square$

Examples

Rotations and reflections in $\mathbb{R}^n$ can be described by orthogonal matrices:

- Rotation in $\mathbb{R}^2$ by angle α:

$$
\begin{aligned}
U &= \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}, \\
UU^\top &= \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \\
&= \begin{pmatrix} \cos^2\alpha + \sin^2\alpha & 0 \\ 0 & \cos^2\alpha + \sin^2\alpha \end{pmatrix} = I
\end{aligned}
$$

- Reflection in $\mathbb{R}^2$ around x_1-axis:

$$
U = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad U = U^\top, \quad UU = I.
$$

Singular values and vectors

Definition

In the singular value decomposition of $A \in \mathbb{R}^{m\times n}$,

$$\begin{pmatrix} & & \\ & A & \\ & & \end{pmatrix} = \begin{pmatrix} & & \\ & U & \\ & & \end{pmatrix} \underbrace{\begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \\ 0 & \cdots & 0 \end{pmatrix}}_{=\Sigma} \begin{pmatrix} & & \\ & V^\top & \\ & & \end{pmatrix}$$

- the $\sigma_j \geq 0, j = 1, \ldots, n$, are called **singular values**,
- the columns of $U \in \mathbb{R}^{m\times m}$ are called **left singular vectors**, and
- the rows of $V \in \mathbb{R}^{n\times n}$ (i.e., the columns of $V^\top$) are called **right singular vectors** of A.

Contents

Linear regression problems

Definition

Let $z \in \mathbb{R}^m$ be data and $y = Ax \in \mathbb{R}^m$ a given linear model, i.e., y depends linearly on some parameters $x \in \mathbb{R}^n$. The problem to find

$$x^* = \operatorname*{argmin}_{x \in \mathbb{R}^n} \|Ax - z\|_2^2$$

is called **linear regression problem** or **linear least-squares problem**.

Theorem

A solution $x \in \mathbb{R}^n$ to the normal equations

$$A^\top A x = A^\top z$$

is a solution of the linear regression problem with data $z \in \mathbb{R}^m$ and matrix $A \in \mathbb{R}^{m \times n}$. If A has full rank, the solution is unique.

Disadvantages and problems when solving the normal equations

- Effort to compute $A^\top A$.
- Entries in the matrix $A^\top A$ may have wide spread in magnitude:

$$\sum_{k=1}^{m} t_k^{2n} \gg m.$$

- $\rightsquigarrow$ Solution of the linear system will be sensitive to errors.
- Nearly linear dependent rows in matrix $\rightsquigarrow$ matrix “nearly” singular $\rightsquigarrow$ result (solution of linear system) may be inaccurate.

$\rightsquigarrow$ Use alternative method to solve

$$\min_{x\in\mathbb{R}^n} \|Ax - z\|_2.$$

Solving the linear regression problem with SVD

- We have $A = U\Sigma V^\top$:

$$\|Ax - z\|_2 = \|U\Sigma V^\top x - z\|_2$$

- U orthogonal (Lemma above) $\Rightarrow U^\top$ (Lemma above) $\Rightarrow \|U^\top y\|_2 = \|y\|_2$ gives:

$$\|U\Sigma V^\top x - z\|_2 = \|U^\top U\Sigma V^\top x - U^\top z\|_2$$

- Lemma above: $U^\top U = I$ gives:

$$\|U^\top U\Sigma V^\top x - U^\top z\|_2 = \|\Sigma V^\top x - U^\top z\|_2.$$

- Define $\hat{x} = V^\top x, \hat{z} = U^\top z$:

$$\|Ax - z\|_2 = \|\Sigma V^\top x - U^\top z\|_2 = \|\Sigma\hat{x} - \hat{z}\|_2.$$

Solving the linear regression problem with SVD

- We have with $\hat{x} = V^\top x, \hat{z} = U^\top z$:

$$\|Ax - z\|_2 = \|\Sigma\hat{x} - \hat{z}\|_2$$

  and thus also

$$\|Ax - z\|_2^2 = \|\Sigma\hat{x} - \hat{z}\|_2^2$$

- Use the structure of Σ:

$$\Sigma\hat{x} = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_n \\ \hat{x}_{n+1} \\ \vdots \\ \hat{x}_m \end{pmatrix} = \begin{pmatrix} \sigma_1\hat{x}_1 \\ \vdots \\ \sigma_n\hat{x}_n \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \Rightarrow \quad \Sigma\hat{x} - \hat{z} = \begin{pmatrix} \sigma_1\hat{x}_1 - \hat{z}_1 \\ \vdots \\ \sigma_n\hat{x}_n - \hat{z}_n \\ -\hat{z}_{n+1} \\ \vdots \\ -\hat{z}_m \end{pmatrix}$$

Solving the linear regression problem with SVD

- We get

$$\|Ax - z\|_2^2 = \|\Sigma \hat{x} - \hat{z}\|_2^2 = \left\| \begin{pmatrix} \sigma_1 \hat{x}_1 - \hat{z}_1 \\ \vdots \\ \sigma_n \hat{x}_n - \hat{z}_n \\ -\hat{z}_{n+1} \\ \vdots \\ -\hat{z}_m \end{pmatrix} \right\|_2^2 = \sum_{i=1}^{n} (\sigma_i \hat{x}_i - \hat{z}_i)^2 + \sum_{i=n+1}^{m} \hat{z}_i^2$$

- The first term on the right is minimal $(= 0)$, if

$$\hat{x}_i = \frac{\hat{z}_i}{\sigma_i}, \quad i = 1, \ldots, n.$$

- The second one is independent of the choice of $\hat{x}$. It is the misfit.

Algorithm: Solution of the linear regression problem with SVD

- Given $z \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times n}$.
- Compute SVD of $A \rightsquigarrow U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{n \times n}, \sigma_i \geq 0, i = 1, \ldots, n$.
- Compute
$$\hat{z} := U^\top z \in \mathbb{R}^m.$$
- Compute
$$\hat{x}_i := \frac{\hat{z}_i}{\sigma_i}, \quad i = 1, \ldots, n.$$
- Compute solution
$$x^* := V\hat{x} \in \mathbb{R}^n$$
- ... and value of f (misfit):
$$\|Ax^* - z\|_2^2 = \sum_{i=n+1}^{m} \hat{z}_i^2.$$
- Implementations: linear algebra packages (LAPACK), built-in functions in Python, octave

Contents

Second look on the SVD

- Let us now assume that $A \in \mathbb{R}^{m \times n}$ is an arbitrary data set.
- SVD gives

$$
A = U \underbrace{\begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \\ 0 & \cdots & 0 \end{pmatrix}}_{=\Sigma} V^\top
$$

- We get by the rules for matrix-matrix multiplication:

$$
\left(\Sigma V^\top\right)_{ij} = \sigma_i (V^\top)_{ij} = \begin{cases} \sigma_i V_{ji}, & i = 1, \dots, n \\ 0, & i = n+1, \dots, m. \end{cases}
$$

Compression with SVD

- Using

$$\left(\Sigma V^\top\right)_{ij} = \begin{cases} \sigma_i V_{ji}, & i = 1, \ldots, n \\ 0, & i = n+1, \ldots, m \end{cases}$$

  we get

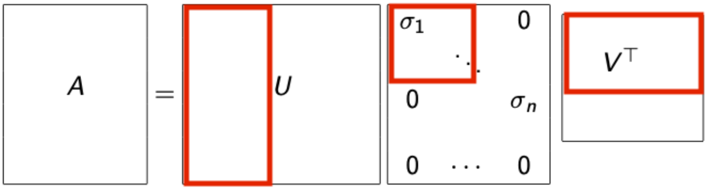
$$A_{kj} = \sum_{i=1}^{m} U_{ki} \left(\Sigma V^\top\right)_{ij} = \sum_{i=1}^{n} \sigma_i U_{ki} V_{ji}.$$

- Theory of the SVD $\rightsquigarrow$ the σ_i are ordered by magnitude, i.e., $\sigma_i \geq \sigma_{i+1}$ for all $i = 1, \ldots, n$.
- We may omit "very small" values of σ_i and corresponding parts in the data:

$$A_{kj} \approx \sum_{i:\sigma_i > \epsilon} \sigma_i U_{ki} V_{ji} \quad \text{with some given } \epsilon > 0.$$

Compression with SVD

- If we omit small values of σ_i, we only use parts of the matrices



… to obtain a compressed version of A.

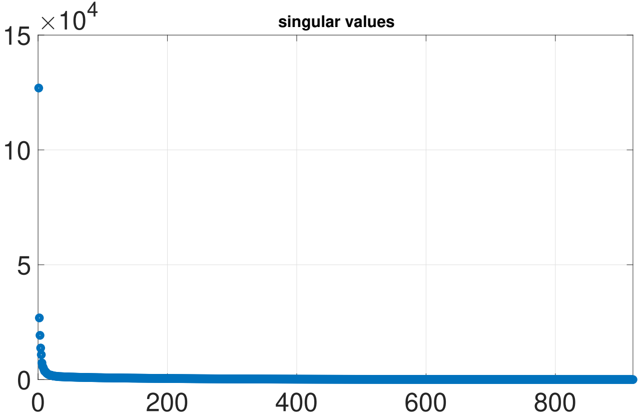
- If we take $k < n$ singular values, we have to store

$$\left.\begin{array}{l} \text{for } U : (m \times k) \\ \text{for } V : (n \times k) \\ k \text{ singular values} \end{array}\right\} = (m+n+1) \times k \text{ values instead of } (m \times n) \text{ values for } A.$$

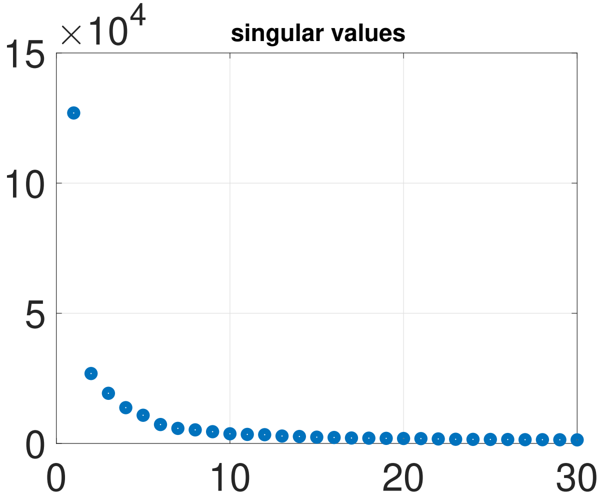
- In many cases, the σ_i decrease rapidly with i and only the first ones are dominant.

Example: Image compression

original size: 1060760



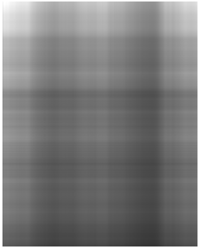
Example: Image compression



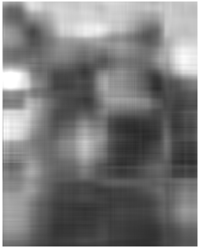
Example: Image compression (using 1, 3, 10, 20, 50, 100 singular values)

original size: 1060760

reduced size: 2074



reduced size: 10370



reduced size: 20740

reduced size: 41480

reduced size: 103700

reduced size: 207400

Example: Image compression (using 100 singular values)

original size: 1060760

reduced size: 207400

What is important?

- Singular Value Decomposition is the decomposition of an arbitrary matrix in the product of two orthogonal matrices and a diagonal matrix.
- The diagonal matrix contains the non-negative singular values.
- The SVD can be used to solve linear regression problems.
- In this case, it avoids the computation of $A^\top A$ and $A^\top z$ ...
- ... and is less sensitive to data errors or perturbations.
- However, the SVD itself requires additional effort of $\mathcal{O}(n^3)$.
- It can be also used for data compression ...
- ... and data analysis (we will see this later on).