Mathematics for Machine Learning

Prof Willie Brink

Applied Mathematics, Stellenbosch University

Lecture 3: Matrix Decompositions

Contents of the module

4.1 Determinant and trace

The determinant of a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is a real number $\det(\boldsymbol{A}) = |\boldsymbol{A}|$ related to the existence of an inverse: $\boldsymbol{A}$ is invertible if and only if $\det(\boldsymbol{A}) \neq 0$.

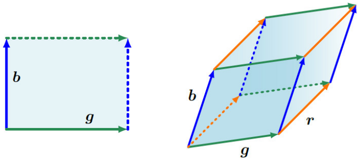
If $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$, $\det(\boldsymbol{A}) = a_{11}a_{22} - a_{12}a_{21}$

If $\boldsymbol{A} \in \mathbb{R}^{3\times 3}$,
$$\det(\boldsymbol{A}) = a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} - a_{31}a_{22}a_{13} - a_{11}a_{32}a_{23} - a_{21}a_{12}a_{33}$$
Sarrus' rule

If $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ is upper-triangular ($t_{i,j} = 0$ for $i > j$) or lower-triangular ($t_{i,j} = 0,\ i < j$),

$$\det(\boldsymbol{T}) = \prod_{i=1}^{n} t_{i,i}$$

$\det(\boldsymbol{A})$ is the signed volume of an n-dimensional parallelepiped formed by columns of $\boldsymbol{A}$.



Laplace expansion allows us to compute the determinant of an $n \times n$ matrix in terms of the determinant of an $(n-1) \times (n-1)$ matrix.

$$\text{Expansion along column } j\text{: } \det(\boldsymbol{A}) = \sum_{k=1}^{n} (-1)^{k+j} a_{k,j} \det(\boldsymbol{A}_{k,j})$$

where $\boldsymbol{A}_{k,j}$ is $\boldsymbol{A}$ with row k and column j deleted. Expansion along a row is similar.

$\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$. $\det(\boldsymbol{A}^\mathsf{T}) = \det(\boldsymbol{A})$. If $\boldsymbol{A}$ is invertible, $\det(\boldsymbol{A}^{-1}) = 1/\det(\boldsymbol{A})$.

Multiplication of a row/col by $\lambda \in \mathbb{R}$ scales $\det(\boldsymbol{A})$ by λ, hence $\det(\lambda\boldsymbol{A}) = \lambda^n \det(\boldsymbol{A})$.

A square matrix $\boldsymbol{A}$ has $\det(\boldsymbol{A}) \neq 0$ if and only if $\text{rk}(\boldsymbol{A}) = n$.
That is to say, $\boldsymbol{A}$ is invertible if and only if it is full rank.

The trace of a square matrix $\boldsymbol{A}$, $\text{tr}(\boldsymbol{A})$, is the sum of the diagonal elements of $\boldsymbol{A}$.

Trace is invariant under cyclic permutations of factors: $\text{tr}(\boldsymbol{ABC}) = \text{tr}(\boldsymbol{BCA})$

4.2 Eigenvalues and eigenvectors

Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. Then $\lambda \in \mathbb{R}$ is an eigenvalue of $\boldsymbol{A}$, with corresponding eigenvector $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$, if $\boxed{\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}}$.

Note: if $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$, then so is $c\boldsymbol{x}$ for any $c \in \mathbb{R}\backslash\{0\}$.

Eigenvalues are the roots of the characteristic polynomial of $\boldsymbol{A}$: $p_{\boldsymbol{A}}(\lambda) = \det(\boldsymbol{A} - \lambda \boldsymbol{I}_n)$
Every eigenvalue has an algebraic multiplicity.

All eigenvectors associated with an eigenvalue λ forms the eigenspace of $\boldsymbol{A}$ w.r.t. λ.
It is the solution space of the system $(\boldsymbol{A} - \lambda \boldsymbol{I}_n)\boldsymbol{x} = \mathbf{0}$, i.e. the null space of $\boldsymbol{A} - \lambda \boldsymbol{I}_n$.
Its dimension is called the geometric multiplicity of λ.

$\boldsymbol{A}$ and $\boldsymbol{A}^{\mathsf{T}}$ have the same eigenvalues, but not necessarily the same eigenvectors.

Symmetric, positive definite matrices always have positive, real eigenvalues.

Geometrically, the eigenvector corresponding to a nonzero eigenvalue points in a direction that is stretched by the linear mapping $\boldsymbol{A}$. The eigenvalue is the factor by which it is stretched (can be negative).

$\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is defective if it possesses fewer than n linearly independent eigenvectors. The eigenvectors of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with n distinct eigenvals are linearly independent.

From a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ we can always form a symmetric, positive semidefinite matrix $\boldsymbol{S} \in \mathbb{R}^{n \times n}$ as $\boldsymbol{S} = \boldsymbol{A}^T \boldsymbol{A}$. If $\text{rk}(\boldsymbol{A}) = n$, $\boldsymbol{S}$ will be symmetric, positive definite.

Spectral theorem: if $\boldsymbol{A}^{n \times n}$ is symmetric, there exists an ONB of the vector space consisting of eigenvectors of $\boldsymbol{A}$, and each eigenvalue is real.

For any matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with (possibly repeating) eigenvalues λ_i,

$$\det(\boldsymbol{A}) = \prod_{i=1}^{n} \lambda_i \quad \text{and} \quad \text{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} \lambda_i$$

Google's PageRank algorithm

Importance of a web page is defined by the importance of the pages that link to it.

Express the web as a huge directed graph of which pages linking to which.

PageRank will compute an importance $x_i \geq 0$ for each page i.

Count the number of web pages pointing to i and model a user's navigation by a transition matrix $\boldsymbol{A}$, with columns summing to 1 and $a_{i,j}$ the probability of navigating from page i to page j.

$\boldsymbol{A}$ has the property that $\boldsymbol{A}\boldsymbol{x}$, $\boldsymbol{A}^2\boldsymbol{x}$, $\boldsymbol{A}^3\boldsymbol{x}$, $\ldots$ converges to vector $\boldsymbol{x}^*$. It satisfies $\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{x}^*$, that is, $\boldsymbol{x}^*$ is an eigenvector of $\boldsymbol{A}$ corresponding to eigenvalue 1.

Normalising $\boldsymbol{x}^*$ (such that $\|\boldsymbol{x}^*\| = 1$) gives the PageRank of all pages as probabilities.

4.3 Cholesky decomposition

A symmetric, positive definite matrix $\boldsymbol{A}$ can be factorised uniquely as $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^{\mathsf{T}}$, where $\boldsymbol{L}$ is lower-triangular with positive diagonal elements.

Various algorithms for computing $\boldsymbol{L}$, including a modification of Gaussian elimination.

Note that $\det(\boldsymbol{A}) = \det(\boldsymbol{L})\det(\boldsymbol{L}^T) = \det(\boldsymbol{L})^2$.

Since $\boldsymbol{L}$ is lower-triangular, $\det(\boldsymbol{A})$ is the square of the product of $\boldsymbol{L}$'s diagonal elements.

Applications in machine learning:

- Cholesky decomposition of a covariate matrix allows us to generate samples from a multivariate Gaussian
- used in deep stochastic models (e.g. VAEs) to compute gradients

4.4 Eigendecomposition and diagonalisation

A diagonal matrix $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ has zero on all off-diagonal elements.

- $\det(\boldsymbol{D})$ is the product of the diagonal elements;
- $\boldsymbol{D}^k$ is given by each diagonal element to the power k;
- $\boldsymbol{D}^{-1}$ is the reciprocals of the diagonal elements if they are all nonzero.

Matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is diagonalisable if there exists an invertible matrix $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ such that $\boldsymbol{D} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}$ is diagonal.

Note that if $\boldsymbol{D}$ has the eigenvalues of $\boldsymbol{A}$ on its diagonal, and $\boldsymbol{P}$ the corresponding eigenvectors of $\boldsymbol{A}$ as columns, then $\boldsymbol{A}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{D}$. So for $\boldsymbol{A}$ to be diagonalisable, it must have n linearly independent eigenvectors (so that the inverse of $\boldsymbol{P}$ exists).

From the spectral theorem we have that every symmetric matrix is diagonalisable.

The eigendecomposition of $\boldsymbol{A}$ is $\boldsymbol{A} = \boldsymbol{PDP}^{-1}$, where $\boldsymbol{D}$ is diagonal with the eigenvalues of $\boldsymbol{A}$ on its diagonal and $\boldsymbol{P}$ the corresponding eigenvectors of $\boldsymbol{A}$ as its columns.

If $\boldsymbol{A}$ is symmetric, $\boldsymbol{P}$ will be orthogonal so that $\boldsymbol{A} = \boldsymbol{PDP}^T$.

Geometrically, transformations with $\boldsymbol{A}$ would be the same as:

1. performing a basis change from the standard basis to the eigenbasis $(\boldsymbol{P}^{-1})$
2. scaling along those axes by the eigenvalues $(\boldsymbol{D})$
3. transforming back into the standard coordinates $(\boldsymbol{P})$

If it exists, the eigendecomposition allows for efficient computation of matrix powers and the determinant.

But this decomposition requires the matrix $\boldsymbol{A}$ to be square...

4.5 Singular value decomposition

The SVD of a rectangular matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ of rank $r \leq \min\{m, n\}$, is of the form

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$

where $\boldsymbol{U} \in \mathbb{R}^{m \times m}$ is orthogonal, with the left-singular vectors of $\boldsymbol{A}$ as columns $(\boldsymbol{u}_i)$

$\boldsymbol{\Sigma} \in \mathbb{R}^{m \times n}$ contains the singular values of $\boldsymbol{A}$ on the diagonal and zeros elsewhere

$\boldsymbol{V} \in \mathbb{R}^{n \times n}$ is orthogonal, with the right-singular vectors of $\boldsymbol{A}$ as columns $(\boldsymbol{v}_j)$

The singular values are non-negative, and by convention in non-decreasing order:

$$\sigma_1 \geq \sigma_2 \geq \ldots \sigma_{\min\{m,n\}} \geq 0$$

The SVD exists for any matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$.

Geometric intuition: a basis change via $\boldsymbol{V}^\top$, followed by a scaling and augmentation (or reduction) in dimensionality via $\boldsymbol{\Sigma}$, and then a second basis change via $\boldsymbol{U}$.

Construction of the SVD

From any $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ we can construct a symmetric, positive definite matrix $\boldsymbol{A}^\top \boldsymbol{A}$ with eigendecomposition: $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top \quad (\star)$

Assuming $\boldsymbol{A}$ can be written in the form $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$,

$$\boldsymbol{A}^\top \boldsymbol{A} = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{\Sigma}^\top \boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma}\boldsymbol{V}^\top \text{ since } \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_m$$

Compare with $(\star)$: $\boldsymbol{V} = \boldsymbol{P}$ and $\boldsymbol{\Sigma}^T \boldsymbol{\Sigma} = \boldsymbol{D}$

The diagonal elements of $\boldsymbol{\Sigma}$ are the positive square roots of the eigenvalues of $\boldsymbol{A}^T \boldsymbol{A}$.

The columns of $\boldsymbol{V}$ are the eigenvectors of $\boldsymbol{A}^T \boldsymbol{A}$ (ordered appropriately).

Similarly, from the eigendecomposition of the symmetric, positive definite matrix $\boldsymbol{A}\boldsymbol{A}^T$ we find that the columns of $\boldsymbol{U}$ are the eigenvectors of $\boldsymbol{A}\boldsymbol{A}^T$.

Consider again an $m \times n$ matrix $\boldsymbol{A}$ of rank r. Because of the many zeros in $\boldsymbol{\Sigma}$, some columns of $\boldsymbol{U}$ or rows in $\boldsymbol{V}^\mathsf{T}$ may be redundant (in certain applications).

If $r < \min\{m, n\}$, even more columns and rows can be removed.

The reduced SVD is $\boldsymbol{A} = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r^\mathsf{T}$, where

$\boldsymbol{U}_r$ is an $m \times r$ matrix consisting of the first r columns of $\boldsymbol{U}$,

$\boldsymbol{\Sigma}_r$ is an $r \times r$ diagonal matrix with $\sigma_1, \ldots, \sigma_r$ on the diagonal,

and $\boldsymbol{V}_r$ is an $n \times r$ matrix consisting of the first r columns of $\boldsymbol{V}$.

Applications of the SVD in machine learning:

- solving general linear systems, also in the least-squares sense
- low-rank matrix approximation for dimensionality reduction, topic modelling, data compression, clustering

Finding structure in movie ratings

n viewers rate m movies out of 5. We encode this in a data matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, and SVD.

The columns $\boldsymbol{u}_i$ of $\boldsymbol{U}$ are stereotypical movies, and the columns $\boldsymbol{v}_j$ of $\boldsymbol{V}$ stereotypical viewers.

- a vector in $\text{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ might be a particular viewer's preferences
- a vector in $\text{span}(\boldsymbol{u}_1, \dots, \boldsymbol{u}_m)$ might be a particular movie's likeability

$$\begin{array}{r} \\ \text{Star Wars} \\ \text{Blade Runner} \\ \text{Amelie} \\ \text{Delicatessen} \end{array} \begin{array}{c} \begin{matrix} \text{A} & \text{B} & \text{C} \end{matrix} \\ \begin{bmatrix} 5 & 4 & 1 \\ 5 & 5 & 0 \\ 0 & 0 & 5 \\ 1 & 0 & 4 \end{bmatrix} \end{array} = \begin{bmatrix} -0.6710 & 0.0236 & 0.4647 & -0.5774 \\ -0.7197 & 0.2054 & -0.4759 & 0.4619 \\ -0.0939 & -0.7705 & -0.5268 & -0.3464 \\ -0.1515 & -0.6030 & 0.5293 & -0.5774 \end{bmatrix} \begin{bmatrix} 9.6438 & 0 & 0 \\ 0 & 6.3639 & 0 \\ 0 & 0 & 0.7056 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -0.7367 & -0.6515 & -0.1811 \\ 0.0852 & 0.1762 & -0.9807 \\ 0.6708 & -0.7379 & -0.0743 \end{bmatrix}$$

$\boldsymbol{u}_1$ has large values for the first two movies, grouping a type of user with a specific set of movies (sci-fi).

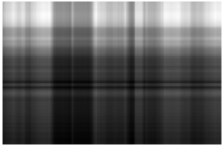
$\boldsymbol{v}_1$ shows large values for users A and B, suggesting the notion of a science fiction lover.

$\boldsymbol{u}_2$ captures an art-house theme, and $\boldsymbol{v}_2$ indicates that C is close to an idealised lover of such movies.

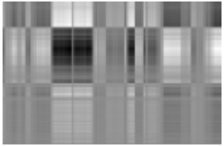
4.6 Matrix approximation

The reduced SVD can be expressed as $\boldsymbol{A} = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}^\mathsf{T} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\mathsf{T} = \sum_{i=1}^{r} \sigma_i \boldsymbol{A}_i$ where $\boldsymbol{A}_i = \boldsymbol{u}_i \boldsymbol{v}_i^\mathsf{T}$ is a rank-1 matrix.

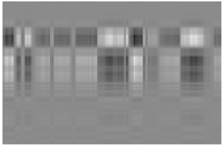
(a) Original image $\boldsymbol{A}$. 1432 × 1910



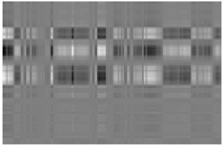
(b) $\boldsymbol{A}_1,\ \sigma_1 \approx 228,052$.



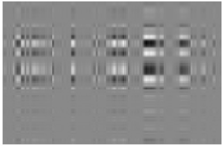
(c) $\boldsymbol{A}_2,\ \sigma_2 \approx 40,647$.



(d) $\boldsymbol{A}_3,\ \sigma_3 \approx 26,125$.



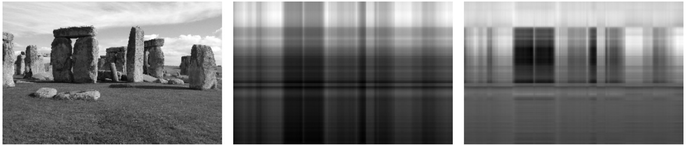
(e) $\boldsymbol{A}_4,\ \sigma_4 \approx 20,232$.



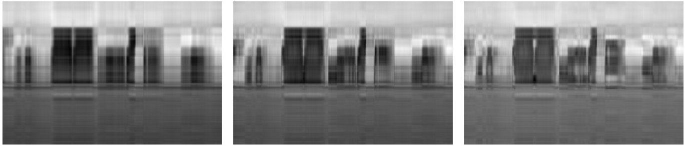
(f) $\boldsymbol{A}_5,\ \sigma_5 \approx 15,436$.

By summing only the first $k < r$ terms, we obtain a rank-k approximation $\widehat{\boldsymbol{A}}_k$.

It turns out that $\widehat{\boldsymbol{A}}_k$ is the closest rank-k matrix to $\boldsymbol{A}$, in terms of the spectral norm*.



(a) Original image $\boldsymbol{A}$. 1432×1910 (b) Rank-1 approximation $\widehat{\boldsymbol{A}}(1)$. (c) Rank-2 approximation $\widehat{\boldsymbol{A}}(2)$



d) Rank-3 approximation $\widehat{\boldsymbol{A}}(3)$. (e) Rank-4 approximation $\widehat{\boldsymbol{A}}(4)$. (f) Rank-5 approximation $\widehat{\boldsymbol{A}}(5)$

* $\|\boldsymbol{A}\|_2 = \max_x \|\boldsymbol{A}\boldsymbol{x}\|_2 / \|\boldsymbol{x}\|_2 = \sigma_1$. The spectral norm of $\boldsymbol{A}$ is equal to its largest singular value.