Lecture 4: Matrix Decompositions

Yi, Yung (이융)

Mathematics for Machine Learning
https://yung-web.github.io/home/courses/mathml.html
KAIST EE

April 7, 2021

Roadmap

Summary

- How to summarize matrices: determinants and eigenvalues
- How matrices can be decomposed: Cholesky decomposition, diagonalization, singular value decomposition
- How these decompositions can be used for matrix approximation

Roadmap

Determinant: Motivation (1)

KAIST EE

- For $\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, $\boldsymbol{A}^{-1} = \frac{1}{a_{11}a_{22}-a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$.
- $\boldsymbol{A}$ is invertible iff $a_{11}a_{22} - a_{12}a_{21} \neq 0$
- Let's define $\det(\boldsymbol{A}) = a_{11}a_{22} - a_{12}a_{21}$.
- Notation: $\det(\boldsymbol{A})$ or |whole matrix|
- What about 3×3 matrix? By doing some algebra (e.g., Gaussian elimination),

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23}$$
$$- a_{31}a_{22}a_{13} - a_{11}a_{32}a_{23} - a_{21}a_{12}a_{33}$$

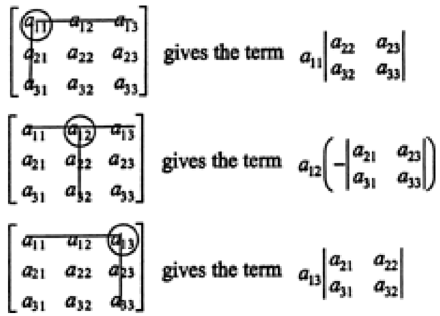
Determinant: Motivation (2)

KAIST EE

- Try to find some pattern ...

$a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23}$
$- a_{31}a_{22}a_{13} - a_{11}a_{32}a_{23} - a_{21}a_{12}a_{33} =$
$a_{11}(-1)^{1+1}\det(\boldsymbol{A}_{1,1}) + a_{12}(-1)^{1+2}\det(\boldsymbol{A}_{1,2})$
$+ a_{13}(-1)^{1+3}\det(\boldsymbol{A}_{1,3})$

- $\boldsymbol{A}_{k,j}$ is the submatrix of $\boldsymbol{A}$ that we obtain when deleting row k and column j.



source: www.cliffsnotes.com

- This is called Laplace expansion.
- Now, we can generalize this and provide the formal definition of determinant.

Determinant: Formal Definition

KAIST EE

Determinant

For a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, for all $j = 1, \ldots, n$,

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

- All expansion are equal, so no problem with the definition.
- Theorem. $\det(\boldsymbol{A}) \neq 0 \iff \text{rk}(\boldsymbol{A}) = n \iff \boldsymbol{A}$ is invertible.

Determinant: Properties

KAIST EE

(1) $\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$
(2) $\det(\boldsymbol{A}) = \det(\boldsymbol{A}^{\mathsf{T}})$
(3) For a regular $\boldsymbol{A}$, $\det(\boldsymbol{A}^{-1}) = 1/\det(\boldsymbol{A})$
(4) For two similar matrices $\boldsymbol{A}, \boldsymbol{A}'$ (i.e., $\boldsymbol{A}' = \boldsymbol{S}^{-1}\boldsymbol{AS}$ for some $\boldsymbol{S}$), $\det(\boldsymbol{A}) = \det(\boldsymbol{A}')$
(5) For a triangular matrix[1] $\boldsymbol{T}$, $\det(\boldsymbol{T}) = \prod_{i=1}^{n} T_{ii}$
(6) Adding a multiple of a column/row to another one does not change $\det(\boldsymbol{A})$
(7) Multiplication of a column/row with λ scales $\det(\boldsymbol{A})$: $\det(\lambda\boldsymbol{A}) = \lambda^n \boldsymbol{A}$
(8) Swapping two rows/columns changes the sign of $\det(\boldsymbol{A})$
   - Using (5)-(8), Gaussian elimination (reaching a triangular matrix) enables to compute the determinant.

[1] This includes diagonal matrices.

Trace

- Definition. The trace of a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is defined as

$$\mathrm{tr}(\boldsymbol{A}) := \sum_{i=1}^{n} a_{ii}$$

- $\mathrm{tr}(\boldsymbol{A}+\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A}) + \mathrm{tr}(\boldsymbol{B})$
- $\mathrm{tr}(\alpha\boldsymbol{A}) = \alpha\,\mathrm{tr}(\boldsymbol{A})$
- $\mathrm{tr}(\boldsymbol{I}_n) = n$

Invariant under Cyclic Permutations

- $\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA})$ for $\boldsymbol{A} \in \mathbb{R}^{n\times k}$ and $\boldsymbol{B} \in \mathbb{R}^{k\times n}$
- $\mathrm{tr}(\boldsymbol{AKL}) = \mathrm{tr}(\boldsymbol{KLA})$, for $\boldsymbol{A} \in \mathbb{R}^{a\times k}$, $\boldsymbol{K} \in \mathbb{R}^{k\times l}$, $\boldsymbol{L} \in \mathbb{R}^{l\times a}$
- $\mathrm{tr}(\boldsymbol{xy}^\mathsf{T}) = \mathrm{tr}(\boldsymbol{y}^\mathsf{T}\boldsymbol{x}) = \boldsymbol{y}^\mathsf{T}\boldsymbol{x} \in \mathbb{R}$
- A linear mapping $\Phi : V \mapsto V$, represented by a matrix $\boldsymbol{A}$ and another matrix $\boldsymbol{B}$.
  - $\boldsymbol{A}$ and $\boldsymbol{B}$ use different bases, where $\boldsymbol{B} = \boldsymbol{S}^{-1}\boldsymbol{AS}$

$$\mathrm{tr}(\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{S}^{-1}\boldsymbol{AS}) = \mathrm{tr}(\boldsymbol{ASS}^{-1}) = \mathrm{tr}(\boldsymbol{A})$$

  - Message. While matrix representations of linear mappings are basis dependent, but their traces are not.

Background: Characteristic Polynomial

- Definition. For $\lambda \in \mathbb{R}$ and a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, the characteristic polynomial of $\boldsymbol{A}$ is defined as:

$$\begin{aligned} p_{\boldsymbol{A}}(\lambda) &:= \det(\boldsymbol{A} - \lambda\boldsymbol{I}) \\ &= c_0 + c_1\lambda + c_2\lambda^2 + \cdots + c_{n-1}\lambda^{n-1} + (-1)^n\lambda^n, \end{aligned}$$

where $c_0 = \det(\boldsymbol{A})$ and $c_{n-1} = (-1)^{n-1}\,\mathrm{tr}(\boldsymbol{A})$.

- Example. For $\boldsymbol{A} = \begin{pmatrix} 4 & 2 \\ 1 & 3 \end{pmatrix}$,

$$p_{\boldsymbol{A}}(\lambda) = \begin{vmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{vmatrix} = (4-\lambda)(3-\lambda) - 2\cdot 1$$

Roadmap

Eigenvalue and Eigenvector

KAIST EE

- Definition. Consider a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. Then, $\lambda \in \mathbb{R}$ is an eigenvalue of $\boldsymbol{A}$ and $\boldsymbol{x} \in \mathbb{R}^n \setminus \{0\}$ is the corresponding eigenvector of $\boldsymbol{A}$ if

$$\boxed{\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}}$$

- Equivalent statements
  - λ is an eigenvalue.
  - $(\boldsymbol{A} - \lambda \boldsymbol{I}_n)\boldsymbol{x} = 0$ can be solved non-trivially, i.e., $\boldsymbol{x} \neq 0$.
  - $\text{rk}(\boldsymbol{A} - \lambda \boldsymbol{I}_n) < n$.
  - $\det(\boldsymbol{A} - \lambda \boldsymbol{I}_n) = 0 \iff$ The characteristic polynomial $p_{\boldsymbol{A}}(\lambda) = 0$.

Example

KAIST EE

- For $\boldsymbol{A} = \left(\begin{smallmatrix} 4 & 2 \\ 1 & 3 \end{smallmatrix}\right)$, $p_{\boldsymbol{A}}(\lambda) = \begin{vmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{vmatrix} = (4-\lambda)(3-\lambda) - 2 \cdot 1 = \lambda^2 - 7\lambda + 10$
- Eigenvalues $\lambda = 2$ or $\lambda = 5$.
- Eigenvector E_5 for $\lambda = 5$

$$\begin{pmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{pmatrix} \boldsymbol{x} = 0 \implies \begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0 \implies E_5 = \text{span}[\begin{pmatrix} 2 \\ 1 \end{pmatrix}]$$

- Eigenvector E_2 for $\lambda = 2$. Similarly, we get $E_2 = \text{span}[\begin{pmatrix} 1 \\ -1 \end{pmatrix}]$
- Message. Eigenvectors are not unique.

Properties (1)

KAIST EE

- If $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$, so are all vectors that are collinear[2].
- E_λ: the set of all eigenvectors for eigenvalue λ, spanning a subspace of $\mathbb{R}^n$. We call this eigensapce of $\boldsymbol{A}$ for λ.
- E_λ is the solution space of $(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{x} = 0$, thus $E_\lambda = \ker(\boldsymbol{A} - \lambda \boldsymbol{I})$
- Geometric interpretation
  - The eigenvector corresponding to a nonzero eigenvalue points in a direction stretched by the linear mapping.
  - The eigenvalue is the factor of stretching.
- Identity matrix $\boldsymbol{I}$: one eigvenvalue $\lambda = 1$ and all vectors $\boldsymbol{x} \neq 0$ are eigenvectors.

[2] Two vectors are collinear if they point in the same or the opposite direction.

Properties (2)

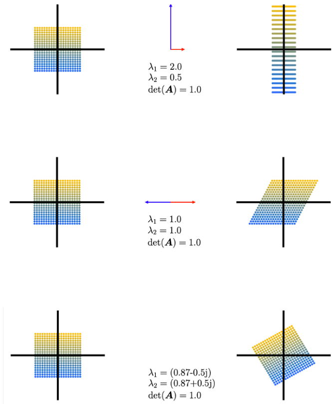
KAIST EE

- $\boldsymbol{A}$ and $\boldsymbol{A}^\mathsf{T}$ share the eigenvalues, but not necessarily eigenvectors.
- For two similar matrices $\boldsymbol{A}, \boldsymbol{A}'$ (i.e., $\boldsymbol{A}' = \boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}$ for some $\boldsymbol{S}$), they possess the same eigenvalues.
  - Meaning: A linear mapping Φ has eigenvalues that are independent of the choice of basis of its transformation matrix.
  - Symmetric, positive definite matrices always have positive, real eigenvalues.

> determinant, trace, eigenvalues: all invariant under basis change

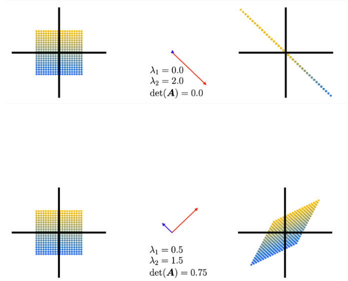
Examples for Geometric Interpretation (1)

1. $\boldsymbol{A} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{pmatrix}$, $\det(\boldsymbol{A}) = 1$
   - $\lambda_1 = \frac{1}{2}, \lambda_2 = 2$
   - eigenvectors: canonical basis vectors
   - area preserving, just vertical horizontal) stretching.
2. $\boldsymbol{A} = \begin{pmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{pmatrix}$, $\det(\boldsymbol{A}) = 1$
   - $\lambda_1 = \lambda_2 = 1$
   - eigenvectors: colinear over the horiontal line
   - area preserving, shearing
3. $\boldsymbol{A} = \begin{pmatrix} \cos(\frac{\pi}{6}) & -\sin(\frac{\pi}{6}) \\ \sin(\frac{\pi}{6}) & \cos(\frac{\pi}{6}) \end{pmatrix}$, $\det(\boldsymbol{A}) = 1$
   - Rotation by $\pi/6$ counter-clockwise
   - only complex eigenvalues (no eigenvectors)
   - area preserving



Examples for Geometric Interpretation (2)

4. $\boldsymbol{A} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$, $\det(\boldsymbol{A}) = 0$
   - $\lambda_1 = 0, \lambda_2 = 2$
   - Mapping that collapses a 2D onto 1D
   - area collapses
5. $\boldsymbol{A} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$, $\det(\boldsymbol{A}) = 3/4$
   - $\lambda_1 = 0.5, \lambda_2 = 1.5$
   - area scales by 75%, shearing and stretching



Properties (3)

- For $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, n distinct eigenvalues $\implies$ eigenvectors are linearly independent, which form a basis of $\mathbb{R}^n$.
  - Converse is not true.
  - Example of n linearly independent eigenvectors for less than n eigenvalues???
- Determinant. For (possibly repeated) eigenvalues λ_i of $\boldsymbol{A} \in \mathbb{R}^{n \times n}$,

$$\det(\boldsymbol{A}) = \prod_{i=1}^{n} \lambda_i$$

- Trace. For (possibly repeated) eigenvalues λ_i of $\boldsymbol{A} \in \mathbb{R}^{n \times n}$,

$$\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} \lambda_i$$

- Message. $\det(\boldsymbol{A})$ is the area scaling and $\mathrm{tr}(\boldsymbol{A})$ is the circumference scaling

Roadmap

(1) Determinant and Trace

(2) Eigenvalues and Eigenvectors

(3) Cholesky Decomposition

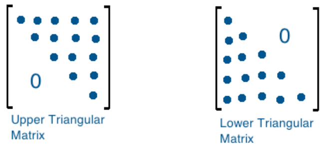
(4) Eigendecomposition and Diagonalization

(5) Singular Value Decomposition

(6) Matrix Approximation

(7) Matrix Phylogeny

LU Decomposition



Source: `http://mathonline.wikidot.com/`

- The Gaussian elimination is the processing of reaching an upper triangular matrix
- Gaussian elimination: multiplying the matrices corresponding to two elementary operations ((i) row multiplication by a and (ii) adding two rows downward)
- The above elementary operations are the low triangular matrices (LTM), and their inverses and their product are all LTMs.
- $(\boldsymbol{E}_k \boldsymbol{E}_{k-1} \cdot \boldsymbol{E}_1)\boldsymbol{A} = \boldsymbol{U} \implies \boldsymbol{A} = \underbrace{(\boldsymbol{E}_1{}^{-1} \cdots \boldsymbol{E}_{k-1}{}^{-1} \boldsymbol{E}_k{}^{-1})}_{\boldsymbol{L}} \boldsymbol{U}$

Cholesky Decomposition

- A real number: decomposition of two identical numbers, e.g., $9 = 3 \times 3$
- Theorem. For a symmetric, positive definite matrix $\boldsymbol{A}$, $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^{\mathsf{T}}$, where
  - $\boldsymbol{L}$ is a lower-triangular matrix with positive diagonals
  - Such a $\boldsymbol{L}$ is unique, called Cholesky factor of $\boldsymbol{A}$.
- Applications
  (a) factorization of covariance matrix of a multivariate Gaussian variable
  (b) linear transformation of random variables
  (c) fast determinant computation: $\det(\boldsymbol{A}) = \det(\boldsymbol{L})\det(\boldsymbol{L}^{\mathsf{T}}) = \det(\boldsymbol{L})^2$, where $\det(\boldsymbol{L}) = \prod_i l_{ii}$. Thus, $\det(\boldsymbol{A}) = \prod_i l_{ii}^2$.

Roadmap

Diagonal Matrix and Diagonalization

- Diagonal matrix. zero on all off-diagonal elements, $\boldsymbol{D} = \begin{pmatrix} d_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & d_n \end{pmatrix}$

$$\boldsymbol{D}^k = \begin{pmatrix} d_1^k & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & d_n^k \end{pmatrix}, \quad \boldsymbol{D}^{-1} = \begin{pmatrix} 1/d_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 1/d_n \end{pmatrix}, \quad \det(\boldsymbol{D}) = d_1 d_2 \cdots d_n$$

- Definition. $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is diagonalizable if it is similar to a diagonal matrix $\boldsymbol{D}$, i.e., $\exists$ an invertible $\boldsymbol{P} \in \mathbb{R}^{n \times n}$, such that $\boldsymbol{D} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}$.
- Definition. $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is orthogonally diagonalizable if it is similar to a diagonal matrix $\boldsymbol{D}$, i.e., $\exists$ an orthogonal $\boldsymbol{P} \in \mathbb{R}^{n \times n}$, such that $\boldsymbol{D} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} = \boldsymbol{P}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{P}$.

Power of Diagonalization

- $\boldsymbol{A}^k = \boldsymbol{P}\boldsymbol{D}^k\boldsymbol{P}^{-1}$
- $\det(\boldsymbol{A}) = \det(\boldsymbol{P})\det(\boldsymbol{D})\det(\boldsymbol{P}^{-1}) = \det(\boldsymbol{D}) = \prod_i d_{ii}$
- Many other things ...
- Question. Under what condition is $\boldsymbol{A}$ diagonalizable (or orthogonally diagonalizable) and how can we find $\boldsymbol{P}$ (thus $\boldsymbol{D}$)?

Diagonalizablity, Algebraic/Geometric Multiplicity

- Definition. For a matrix $\boldsymbol{A} \in realnn$ with an eigenvalue λ_i,
  - the algebraic multiplicity α_i of λ_i is the number of times the root appears in the characteristic polynomial.
  - the geometric multiplicity ζ_i of λ_i is the number of linearly independent eigenvectors associated with λ_i (i.e., the dimension of the eigenspace spanned by the eigenvectors of λ_i)
- Example. The matrix $\boldsymbol{A} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$ has two repeated eigenvalues $\lambda_1 = \lambda_2 = 2$, thus $\alpha_1 = 2$. However, it has only one distinct unit eigenvector $\boldsymbol{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, thus $\zeta_1 = 1$.
- Theorem. $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is diagonalizable $\iff \sum_i \alpha_i = \sum_i \zeta_i = n$.

Orthogonally Diagonaliable and Symmetric Matrix

Theorem. $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is orthogonally diagonalizable $\iff$ $\boldsymbol{A}$ is symmetric.

- Question. . How to find $\boldsymbol{P}$ (thus $\boldsymbol{D}$)?
- Spectral Theorem. If $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric,
  (a) the eigenvalues are all real
  (b) the eigenvectors to different eigenvalues are perpendicular.
  (c) there exists an orthogonal eigenbasis
- For (c), from each set of eigenvectors, say $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ associated with a particular eigenvalue, say λ_j, we can construct another set of eigenvectors $\{\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_k\}$ that are orthonormal, using the Gram-Schmidt process.
- Then, all eigenvectors can form an orthornormal basis.

Example

- Example. $\boldsymbol{A} = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{pmatrix}$. $p_{\boldsymbol{A}}(\lambda) = -(\lambda-1)^2(\lambda-7)$, thus $\lambda_1 = 1, \lambda_2 = 7$

  $E_1 = \text{span}[\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}]$, $\quad E_7 = \text{span}[\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}]$
  - $(111)^{\mathsf{T}}$ is perpendicular to $(-110)^{\mathsf{T}}$ and $(-101)^{\mathsf{T}}$
  - $\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} -1/2 \\ -1/2 \\ 1 \end{pmatrix}$ (for $\lambda = 1$) and $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ (for $\lambda = 7$) are the orthogonal basis in $\mathbb{R}^3$.
  - After normalization, we can make the orthonormal basis.

Eigendecomposition

- Theorem. The following is equivalent.
  - (a) A square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ can be factorized into $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$, where $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{D}$ is the diagonal matrix whose diagonal entries are eigenvalues of $\boldsymbol{A}$.
  - (b) The eigenvectors of $\boldsymbol{A}$ form a basis of $\mathbb{R}^n$ (i.e., The n eigenvectors of $\boldsymbol{A}$ are linearly independent)
- The above implies the columns of $\boldsymbol{P}$ are the n eigenvectors of $\boldsymbol{A}$ (because $\boldsymbol{AP} = \boldsymbol{PD}$)
- $\boldsymbol{P}$ is an orthogonal matrix, so $\boldsymbol{P}^{\mathsf{T}} = \boldsymbol{P}^{-1}$
- $\boldsymbol{A}$ is symmetric, then (b) holds (Spectral Theorem).

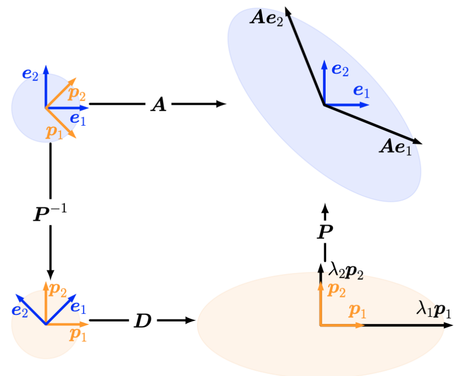
Example of Orthogonal Diagonalization (1)

- Eigendecomposition for $\boldsymbol{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$
- Eigenvalues: $\lambda_1 = 1, \lambda_2 = 3$
- (normalized) eigenvectors: $\boldsymbol{p}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$, $\boldsymbol{p}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$.
- $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ linearly independent, so A is diagonalizable.
- $\boldsymbol{P} = \begin{pmatrix} \boldsymbol{p}_1 & \boldsymbol{p}_2 \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$
- $\boldsymbol{D} = \boldsymbol{P}^{-1}\boldsymbol{AP} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$. Finally, we get $\boldsymbol{A} = \boldsymbol{PDP}^{-1}$

Example of Orthogonal Diagonalization (2)

- $\boldsymbol{A} = \begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{pmatrix}$
- Eigenvalues: $\lambda_1 = -1, \lambda_2 = 5$ $(\alpha_1 = 2, \alpha_2 = 1)$
- $E_{-1} = \text{span}[\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}] \xrightarrow{\text{Gram-Schmidt}}$ $\text{span}[\frac{1}{\sqrt{2}}\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \frac{1}{\sqrt{6}}\begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}]$
- $E_5 = \text{span}[\frac{1}{\sqrt{3}}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}]$
- $\boldsymbol{P} = \begin{pmatrix} -1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 0 & 2/\sqrt{6} & 1/\sqrt{3} \end{pmatrix}$
- $\boldsymbol{D} = \boldsymbol{P}^{\mathsf{T}}\boldsymbol{AP} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 5 \end{pmatrix}$

Eigendecomposition: Geometric Interpretation



Question. Can we generalize this beautiful result to a general matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$?

Roadmap

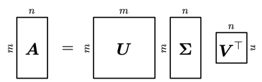
Storyline

- Eigendecomposition (also called EVD: EigenValue Decomposition): (Orthogoanl) Diagonalization for symmetric matrices $\boldsymbol{A} \in \mathbb{R}^{n \times n}$.
- Extensions: Singular Value Decomposition (SVD)
  1. First extension: diagonalization for non-symmetric, but still square matrices $\boldsymbol{A} \in \mathbb{R}^{n \times n}$
  2. Second extension: diagonalization for non-symmeric, and non-square matrices $\boldsymbol{A} \in \mathbb{R}^{m \times n}$
- Background. For $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, a matrix $\boldsymbol{S} := \boldsymbol{A}^\mathsf{T}\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is always symmetric, positive semidefinite.
  - Symmetric, because $\boldsymbol{S}^\mathsf{T} = (\boldsymbol{A}^\mathsf{T}\boldsymbol{A})^\mathsf{T} = \boldsymbol{A}^\mathsf{T}\boldsymbol{A} = \boldsymbol{S}$.
  - Positive semidefinite, because $\boldsymbol{x}^\mathsf{T}\boldsymbol{S}\boldsymbol{x} = \boldsymbol{x}^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{A}\boldsymbol{x})^\mathsf{T}(\boldsymbol{A}\boldsymbol{x}) \geq 0$.
  - If $\mathrm{rk}(\boldsymbol{A}) = n$, then symmetric and positive definite.

Singular Value Decomposition

- Theorem. $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ with rank $r \in [0, \min(m,n)]$. The SVD of $\boldsymbol{A}$ is a decomposition of the form

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



with an orthogonal matrix $\boldsymbol{U} = \begin{pmatrix}\boldsymbol{u}_1 & \cdots & \boldsymbol{u}_m\end{pmatrix} \in \mathbb{R}^{m \times m}$ and an orthogonal matrix $\boldsymbol{V} = \begin{pmatrix}\boldsymbol{v}_1 & \cdots & \boldsymbol{v}_n\end{pmatrix} \in \mathbb{R}^{n \times n}$. Moreoever, Σ s an $m \times n$ matrix with $\Sigma_{ii} = \sigma_i \geq 0$ and $\Sigma_{ij} = 0, \ i \neq j$, which is uniquely determined for $\boldsymbol{A}$.

- Note
  - The diagonal entries $\sigma_i, \ i = 1, \ldots, r$ are called singular values.
  - $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ are called left and right singular vectors, respectively.

SVD: How It Works (for $\boldsymbol{A} \in \mathbb{R}^{n \times n}$)

- $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with rank $r \leq n$. Then, $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$ is symmetric.
- Orthogonal diagonalization of $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$:

$$\boldsymbol{A}^\mathsf{T}\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}.$$

- $\boldsymbol{D} = \begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{pmatrix}$ and an orthogonal matrix $\boldsymbol{V} = \begin{pmatrix}\boldsymbol{v}_1 \cdots \boldsymbol{v}_n\end{pmatrix}$, where $\lambda_1 \geq \cdots \geq \lambda_r \geq \lambda_{r+1} = \cdots \lambda_n = 0$ are the eigenvalues of $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$ and $\{\boldsymbol{v}_i\}$ are orthonormal.
- All λ_i are positive
  $\forall \boldsymbol{x} \in \mathbb{R}^n, \|\boldsymbol{A}\boldsymbol{x}\|^2 = \boldsymbol{A}\boldsymbol{x}^\mathsf{T}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\boldsymbol{x} = \lambda_i \|\boldsymbol{x}\|^2$

- $\mathrm{rk}(\boldsymbol{A})$ $=\mathrm{rk}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})$ $= \mathrm{rk}(D)$ $=r$
- Choose $\boldsymbol{U}' = \begin{pmatrix}\boldsymbol{u}_1 & \cdots \boldsymbol{u}_r\end{pmatrix}$, where

$$\boldsymbol{u}_i = \frac{\boldsymbol{A}\boldsymbol{v}_i}{\sqrt{\lambda_i}}, \ 1 \leq i \leq r.$$

- We can construct $\{\boldsymbol{u}_i\}, \ i = r+1, \cdots, n$, so that $\boldsymbol{U} = \begin{pmatrix}\boldsymbol{u}_1 & \cdots \boldsymbol{u}_n\end{pmatrix}$ is an orthonormal basis of $\mathbb{R}^n$.
- Define $\Sigma = \begin{pmatrix}\sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_n}\end{pmatrix}$
- Then, we can check that $\boldsymbol{U}\Sigma = \boldsymbol{A}\boldsymbol{V}$.
- Similar arguments for a general $\boldsymbol{A}\mathbb{R}^{m \times n}$ (see pp. 104)

Example

- $\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{pmatrix}$
- $\boldsymbol{A}^\top \boldsymbol{A} = \begin{pmatrix} 5 & -2 & 1 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top$,

$\boldsymbol{D} = \begin{pmatrix} 6 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, $\boldsymbol{V} = \begin{pmatrix} \frac{5}{\sqrt{30}} & \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{30}} \\ 0 & \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{pmatrix}$

- $\mathrm{rk}(\boldsymbol{A}) = 2$ because we have two singular values $\sigma_1 = \sqrt{6}$ and $\sigma_2 = 1$
- $\Sigma = \begin{pmatrix} \sqrt{6} & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$
- $\boldsymbol{u}_1 = \boldsymbol{A}\boldsymbol{v}_1/\sigma_1 = \begin{pmatrix} \frac{1}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} \end{pmatrix}$
- $\boldsymbol{u}_2 = \boldsymbol{A}\boldsymbol{v}_2/\sigma_2 = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{pmatrix}$
- $\boldsymbol{U} = (\boldsymbol{u}_1 \;\; \boldsymbol{u}_2) = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}$
- Then, we can see that $\boldsymbol{A} = \boldsymbol{U}\Sigma V^\top$.

EVD ($\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$) vs. SVD ($\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^\top$)

- SVD: always exists, EVD: square matrix and exists if we can find a basis of eigenvectors (such as symmetric matrices)
- $\boldsymbol{P}$ in EVD is not necessarily orthogonal (only true for symmetric $\boldsymbol{A}$), but $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal (so representing rotations)
- Both EVD and SVD: (i) basis change in the domain, (ii) independent scaling of each new basis vector and mapping from domain to codomain, (iii) basis change in the codomain. The difference: for SVD, different vector spaces of domain and codomain.
- SVD and EVD are closely related through their projections
  - The left-singular (resp. right-singular) vectors of $\boldsymbol{A}$ are eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\top$ (resp. $\boldsymbol{A}^\top\boldsymbol{A}$)
  - The singular values of $\boldsymbol{A}$ are the square roots of eigenvalues of $\boldsymbol{A}\boldsymbol{A}^\top$ and $\boldsymbol{A}^\top\boldsymbol{A}$
  - When $\boldsymbol{A}$ is symmetric, EVD = SVD (from spectral theorem)

Different Forms of SVD

- When $\mathrm{rk}(\boldsymbol{A}) = r$, we can construct SVD as the following with only non-zero diagonal entries in $\boldsymbol{\Sigma}$:

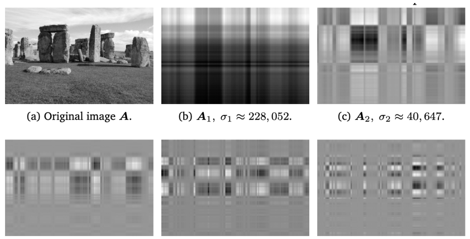
$$\boldsymbol{A} = \overbrace{\boldsymbol{U}}^{m\times r} \overbrace{\boldsymbol{\Sigma}}^{r\times r} \overbrace{\boldsymbol{V}^\top}^{r\times n}$$

- We can even truncate the decomposed matrices, which can be an approximation of $\boldsymbol{A}$: for $k < r$

$$\boldsymbol{A} \approx \overbrace{\boldsymbol{U}}^{m\times k} \overbrace{\boldsymbol{\Sigma}}^{k\times k} \overbrace{\boldsymbol{V}^\top}^{k\times n}$$

We will cover this in the next slides.

Matrix Approximation via SVD



(a) Original image $\boldsymbol{A}$. (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) $\boldsymbol{A}_5$, $\sigma_5 \approx 15,436$.

- $\boldsymbol{A} = \sum_{i=1}^r \sigma_i \overbrace{\boldsymbol{u}_i \boldsymbol{v}_i^\top}^{\boldsymbol{A}_i}$, where $\boldsymbol{A}_i$ is the outer product[3] of $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$
- Rank k-approximation: $\hat{\boldsymbol{A}}(k) = \sum_{i=1}^k \sigma_i \boldsymbol{A}_i$, $k < r$

[3] If $\boldsymbol{u}$ and $\boldsymbol{v}$ are both nonzero, then the outer product matrix $\boldsymbol{u}\boldsymbol{v}^\top$ always has matrix rank 1. Indeed, the columns of the outer product are all proportional to the first column.

How Close $\hat{\boldsymbol{A}}(k)$ is to $\boldsymbol{A}$?

- Definition. Spectral Norm of a Matrix. For $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\|\boldsymbol{A}\|_2 := \max_{\boldsymbol{x}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}$
  - As a concept of length of $\boldsymbol{A}$, it measures how long any vector $\boldsymbol{x}$ can at most become, when multiplied by $\boldsymbol{A}$
- Theorem. Eckart-Young. For $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ of rank r and $\boldsymbol{B} \in \mathbb{R}^{m \times n}$ of rank k, for any $k \le r$, we have:

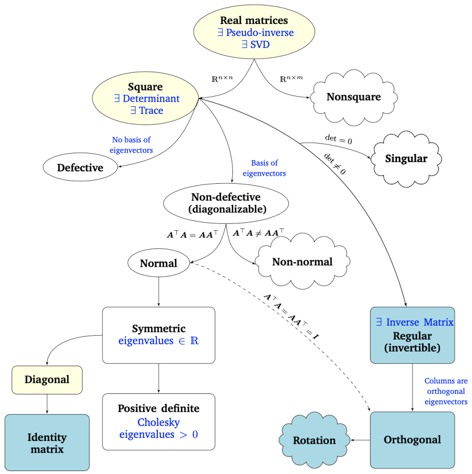
$$\hat{\boldsymbol{A}}(k) = \arg \min_{\mathrm{rk}(\boldsymbol{B})=k} \|\boldsymbol{A} - \boldsymbol{B}\|_2, \quad \text{and} \quad \left\|\boldsymbol{A} - \hat{\boldsymbol{A}}(k)\right\|_2 = \sigma_{k+1}$$

  - Quantifies how much error is introduced by the SVD-based approximation
  - $\hat{\boldsymbol{A}}(k)$ is optimal in the sense that such SVD-based approximation is the best one among all rank-k approximations.
  - In other words, it is a projection of the full-rank matrix $\boldsymbol{A}$ onto a lower-dimensional space of rank-at-most-k matrices.

Roadmap

(1) Determinant and Trace
(2) Eigenvalues and Eigenvectors
(3) Cholesky Decomposition
(4) Eigendecomposition and Diagonalization
(5) Singular Value Decomposition
(6) Matrix Approximation
(7) Matrix Phylogeny

Phylogenetic Tree of Matrices



Questions?

Review Questions

1)