

# Lecture 4: Matrix Decompositions

CSE4130: 기초머신러닝

Junsuk Choe (최준석)

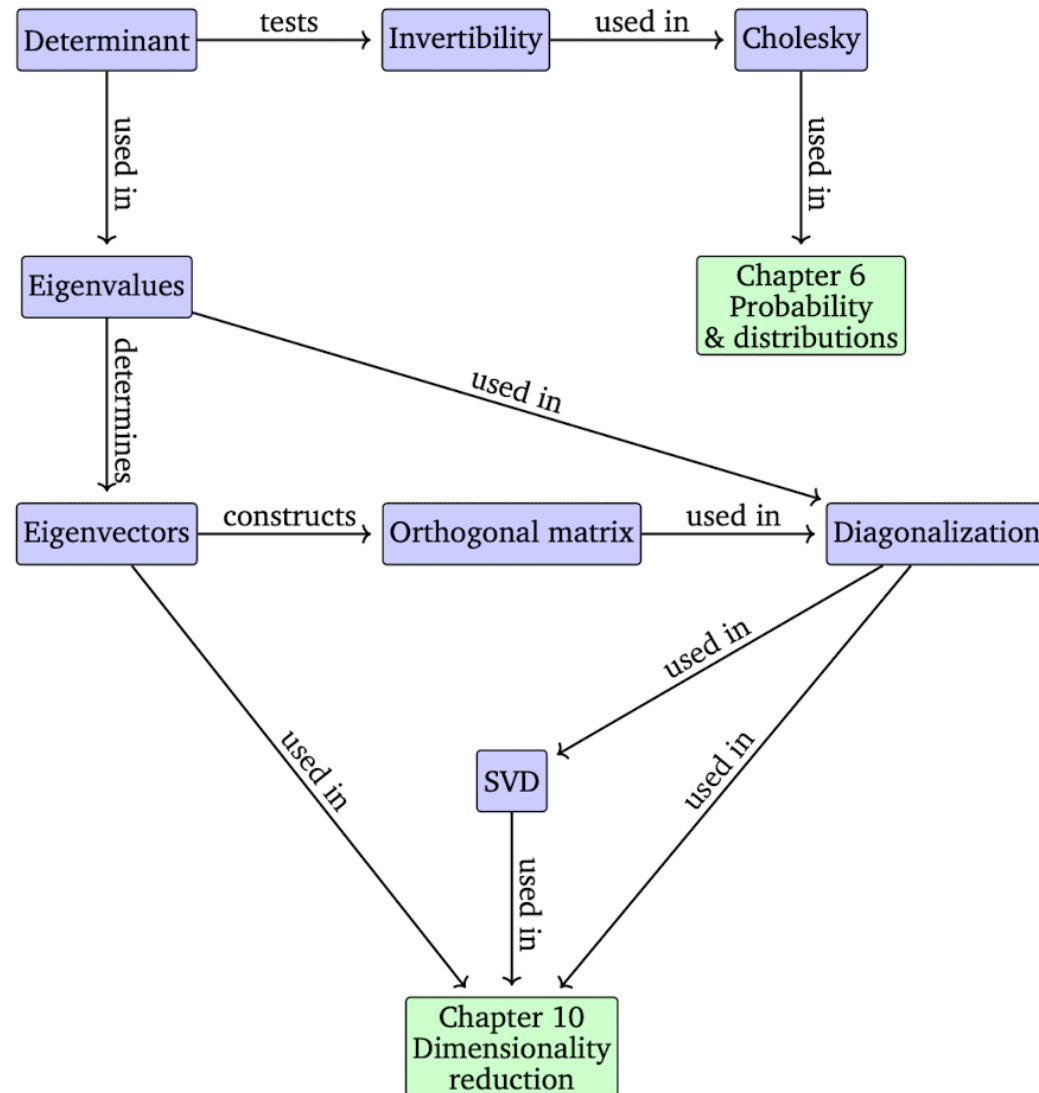
# Roadmap

- (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

# 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

# Summary



# 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

## Determinant: Motivation (1)

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

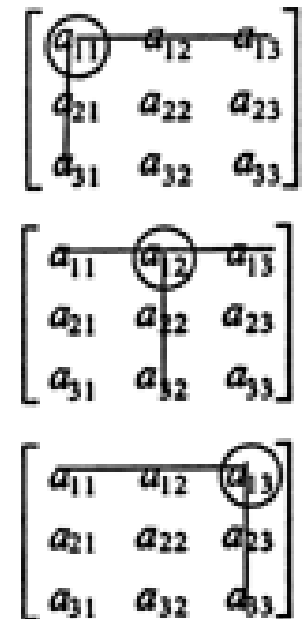
$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} \\ - a_{31}a_{22}a_{13} - a_{11}a_{32}a_{23} - a_{21}a_{12}a_{33}$$

## Determinant: Motivation (2)

- Try to find some pattern ...

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

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



gives the term  $a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}$

gives the term  $a_{12} \left( - \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} \right)$

gives the term  $a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$

source: [www.cliffsnotes.com](http://www.cliffsnotes.com)

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

# Determinant: Formal Definition

## Determinant

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

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

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



## Example 4.3 (Laplace Expansion)

## Determinant: Properties

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

---

<sup>1</sup>This includes diagonal matrices.

# Trace

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

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

- $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$
- $\text{tr}(\alpha \mathbf{A}) = \alpha \text{tr}(\mathbf{A})$
- $\text{tr}(\mathbf{I}_n) = n$

## Invariant under Cyclic Permutations

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

$$\text{tr}(\mathbf{B}) = \text{tr}(\mathbf{S}^{-1} \mathbf{A} \mathbf{S}) = \text{tr}(\mathbf{A} \mathbf{S} \mathbf{S}^{-1}) = \text{tr}(\mathbf{A})$$

- **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  $\mathbf{A} \in \mathbb{R}^{n \times n}$ , the characteristic polynomial of  $\mathbf{A}$  is defined as:

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

where  $c_0 = \det(\mathbf{A})$  and  $c_{n-1} = (-1)^{n-1} \operatorname{tr}(\mathbf{A})$ .

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

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

# 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

# Eigenvalue and Eigenvector

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

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

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

## Example

- For  $\mathbf{A} = \begin{pmatrix} 4 & 2 \\ 1 & 3 \end{pmatrix}$ ,  $p_{\mathbf{A}}(\lambda) = \begin{vmatrix} 4 - \lambda & 2 \\ 1 & 3 - \lambda \end{vmatrix} = (4 - \lambda)(3 - \lambda) - 2 \cdot 1 = \lambda^2 - 7\lambda + 10$
- Eigenvalues  $\lambda = 2$  or  $\lambda = 5$ .
- Eigenvector  $E_5$  for  $\lambda = 5$   
$$\begin{pmatrix} 4 - \lambda & 2 \\ 1 & 3 - \lambda \end{pmatrix} \mathbf{x} = 0 \implies \begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0 \implies E_5 = \text{span}\left[\begin{pmatrix} 2 \\ 1 \end{pmatrix}\right]$$
- Eigenvector  $E_2$  for  $\lambda = 2$ . Similarly, we get  $E_2 = \text{span}\left[\begin{pmatrix} 1 \\ -1 \end{pmatrix}\right]$
- **Message.** Eigenvectors are not unique.



## Properties (1)

- If  $\mathbf{x}$  is an eigenvector of  $\mathbf{A}$ , so are all vectors that are collinear<sup>2</sup>.
- $E_\lambda$ : the set of all eigenvectors for eigenvalue  $\lambda$ , spanning a subspace of  $\mathbb{R}^n$ . We call this **eigenspace** of  $\mathbf{A}$  for  $\lambda$ .
- $E_\lambda$  is the solution space of  $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$ , thus  $E_\lambda = \ker(\mathbf{A} - \lambda\mathbf{I})$
- **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  $\mathbf{I}$ : one eigenvalue  $\lambda = 1$  and all vectors  $\mathbf{x} \neq \mathbf{0}$  are eigenvectors.

---

<sup>2</sup>Two vectors are collinear if they point in the same or the opposite direction.

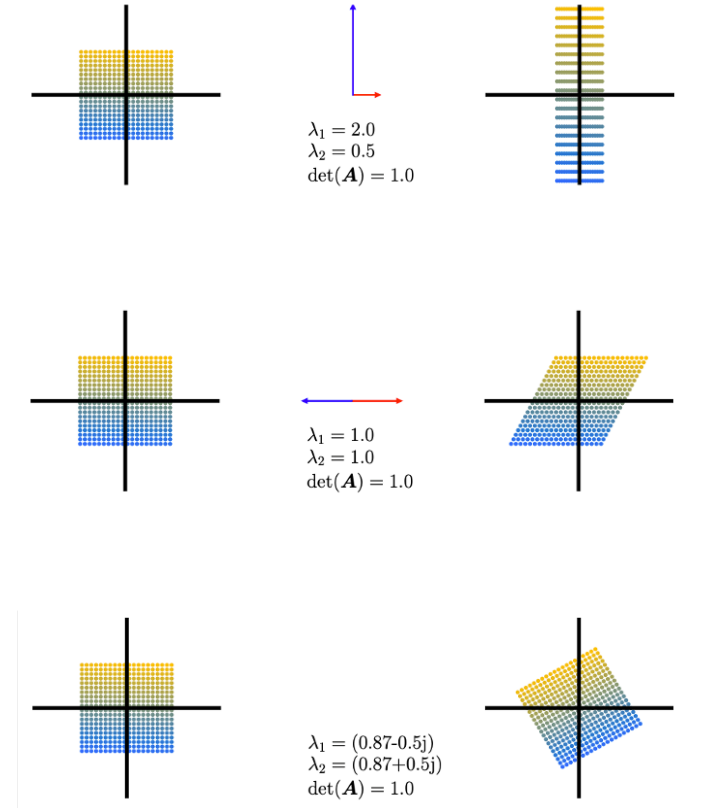
## Properties (2)

- $\mathbf{A}$  and  $\mathbf{A}^T$  share the eigenvalues, but not necessarily eigenvectors.
- For two similar matrices  $\mathbf{A}, \mathbf{A}'$  (i.e.,  $\mathbf{A}' = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$  for some  $\mathbf{S}$ ), they possess the same eigenvalues.
  - Meaning: A linear mapping  $\Phi$  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)

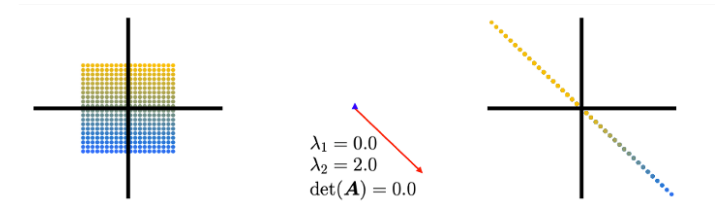
1.  $\mathbf{A} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{pmatrix}$ ,  $\det(\mathbf{A}) = 1$ 
  - $\lambda_1 = \frac{1}{2}, \lambda_2 = 2$
  - eigenvectors: canonical basis vectors
  - area preserving
2.  $\mathbf{A} = \begin{pmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{pmatrix}$ ,  $\det(\mathbf{A})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



## Examples for Geometric Interpretation (2)

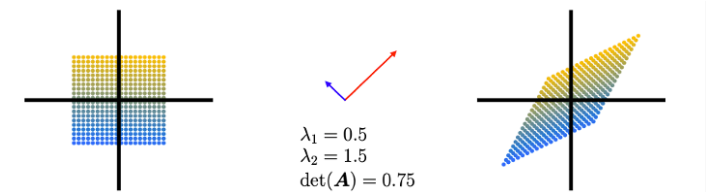
4.  $\mathbf{A} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$ ,  $\det(\mathbf{A}) = 0$

- $\lambda_1 = 0, \lambda_2 = 2$
- Mapping that collapses a 2D onto 1D
- area collapses



5.  $\mathbf{A} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$ ,  $\det(\mathbf{A}) = 3/4$

- $\lambda_1 = 0.5, \lambda_2 = 1.5$
- area scales by 75%, shearing and stretching



## Properties (3)

- For  $\mathbf{A} \in \mathbb{R}^{n \times n}$ ,  $n$  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  $\lambda_i$  of  $\mathbf{A} \in \mathbb{R}^{n \times n}$ ,

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

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

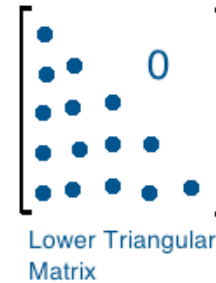
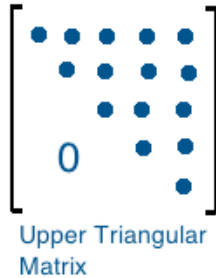
$$\text{tr}(\mathbf{A}) = \sum_{i=1}^n \lambda_i$$

- **Message.**  $\det(\mathbf{A})$  is the **area scaling** and  $\text{tr}(\mathbf{A})$  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.

$$\bullet \quad (\mathbf{E}_k \mathbf{E}_{k-1} \cdot \mathbf{E}_1) \mathbf{A} = \mathbf{U} \implies \mathbf{A} = \underbrace{(\mathbf{E}_1^{-1} \cdots \mathbf{E}_{k-1}^{-1} \mathbf{E}_k^{-1})}_{\mathbf{L}} \mathbf{U}$$

## Example 4.10 (Cholesky Factorization)

$L_4(3)$



# Cholesky Decomposition

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

# 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

# Diagonal Matrix and Diagonalization

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

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

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

# Power of Diagonalization

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

# Diagonalizability, Algebraic/Geometric Multiplicity

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

# Orthogonally Diagonalizable and Symmetric Matrix

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

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

## Example 4.8

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

$$E_1 = \text{span}\left[\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}\right], \quad E_7 = \text{span}\left[\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\right]$$

- $(111)^T$  is perpendicular to  $(-110)^T$  and  $(-101)^T$
- $\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$  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  $\mathbf{A} \in \mathbb{R}^{n \times n}$  can be factorized into  $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$ , where  $\mathbf{P} \in \mathbb{R}^{n \times n}$  and  $\mathbf{D}$  is the diagonal matrix whose diagonal entries are eigenvalues of  $\mathbf{A}$ .
  - (b) The eigenvectors of  $\mathbf{A}$  form a basis of  $\mathbb{R}^n$  (i.e., The  $n$  eigenvectors of  $\mathbf{A}$  are linearly independent)
- The above implies the columns of  $\mathbf{P}$  are the  $n$  eigenvectors of  $\mathbf{A}$  (because  $\mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{D}$ )
- $\mathbf{P}$  is an orthogonal matrix, so  $\mathbf{P}^T = \mathbf{P}^{-1}$
- $\mathbf{A}$  is symmetric, then (b) holds (Spectral Theorem).



## Example of Orthogonal Diagonalization (1)

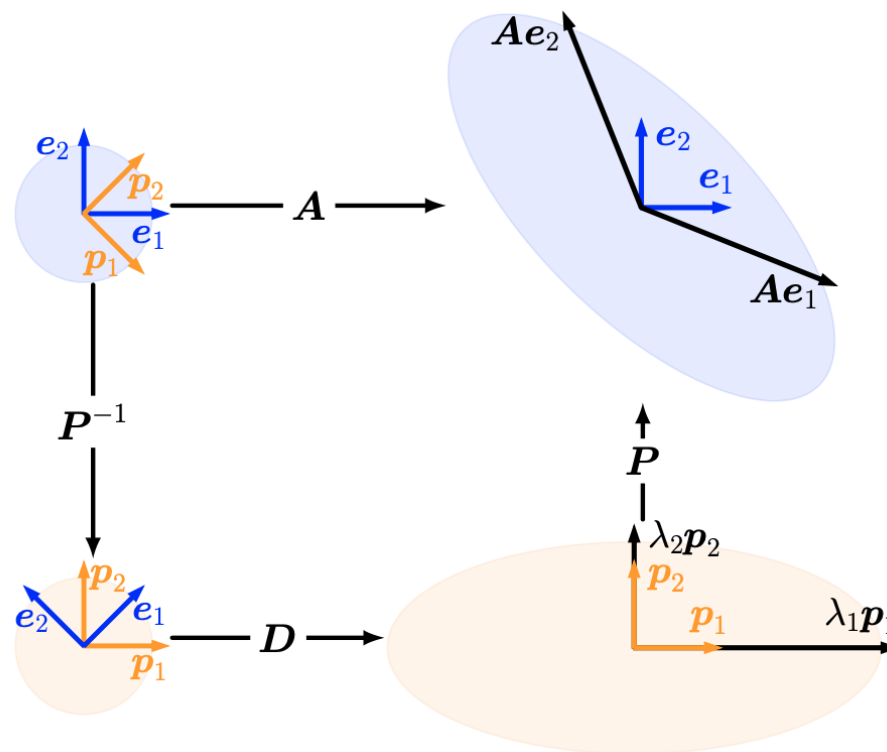
- Eigendecomposition for  $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$
- Eigenvalues:  $\lambda_1 = 1, \lambda_2 = 3$
- (normalized) eigenvectors:  $\mathbf{p}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \mathbf{p}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$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

## Example of Orthogonal Diagonalization (2)

- $\mathbf{A} = \begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{pmatrix}$
- Eigenvalues:  $\lambda_1 = -1, \lambda_2 = 5$   
( $\alpha_1 = 2, \alpha_2 = 1$ )
- $E_{-1} = \text{span}\left[\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}\right] \xrightarrow{\text{Gram-Schmidt}}$   
 $\text{span}\left[\frac{1}{\sqrt{2}}\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \frac{1}{\sqrt{6}}\begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}\right]$

- $E_5 = \text{span}\left[\frac{1}{\sqrt{3}}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\right]$
- $\mathbf{P} = \begin{pmatrix} -1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 0 & 2/\sqrt{6} & 1/\sqrt{3} \end{pmatrix}$
- $\mathbf{D} = \mathbf{P}^T \mathbf{A} \mathbf{P} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 5 \end{pmatrix}$

# Eigendecomposition: Geometric Interpretation



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

# 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

# Storyline

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

# Singular Value Decomposition

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

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \quad \left| \quad \begin{array}{c} n \\ \boxed{\mathbf{A}} \end{array} = \begin{array}{c} m \\ \boxed{\mathbf{U}} \end{array} \begin{array}{c} n \\ \boxed{\mathbf{\Sigma}} \end{array} \begin{array}{c} n \\ \boxed{\mathbf{V}^T} \end{array} \right.$$

with an orthogonal matrix  $\mathbf{U} = (\mathbf{u}_1 \cdots \mathbf{u}_m) \in \mathbb{R}^{m \times m}$  and an orthogonal matrix  $\mathbf{V} = (\mathbf{v}_1 \cdots \mathbf{v}_n) \in \mathbb{R}^{n \times n}$ . Moreover,  $\mathbf{\Sigma}$  is an  $m \times n$  matrix with  $\Sigma_{ii} = \sigma_i \geq 0$  and  $\Sigma_{ij} = 0$ ,  $i \neq j$ , which is uniquely determined for  $\mathbf{A}$ .

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

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

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

$$\mathbf{A}^T \mathbf{A} = \mathbf{V} \mathbf{D} \mathbf{V}^T.$$

- $\mathbf{D} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$  and an orthogonal matrix  $\mathbf{V} = (\mathbf{v}_1 \cdots \mathbf{v}_n)$ , where  $\lambda_1 \geq \cdots \geq \lambda_r \geq \lambda_{r+1} = \cdots = \lambda_n = 0$  are the eigenvalues of  $\mathbf{A}^T \mathbf{A}$  and  $\{\mathbf{v}_i\}$  are orthonormal.
- All  $\lambda_i$  are positive

For all eigenvectors  $\mathbf{x}$  of  $\mathbf{A}^T \mathbf{A}$ ,

$$\|\mathbf{Ax}\|^2 = (\mathbf{Ax})^T \mathbf{Ax} = \mathbf{x}^T \mathbf{A}^T \mathbf{Ax} = \lambda_i \|\mathbf{x}\|^2$$

- $\text{rk}(\mathbf{A}) = \text{rk}(\mathbf{A}^T \mathbf{A}) = \text{rk}(\mathbf{D}) = r$
- Choose  $\mathbf{U}' = (\mathbf{u}_1 \cdots \mathbf{u}_r)$ , where

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

- We can construct  $\{\mathbf{u}_i\}$ ,  $i = r+1, \dots, n$ , so that  $\mathbf{U} = (\mathbf{u}_1 \cdots \mathbf{u}_n)$  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  $\mathbf{U} \Sigma = \mathbf{A} \mathbf{V}$ .
- Similar arguments for a general  $\mathbf{A} \in \mathbb{R}^{m \times n}$  (see pp. 104-106)

## Example 4.13

- $\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{pmatrix}$
- $\mathbf{A}^\top \mathbf{A} = \begin{pmatrix} 5 & -2 & 1 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} = \mathbf{V} \mathbf{D} \mathbf{V}^\top,$
- $\mathbf{D} = \begin{pmatrix} 6 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \mathbf{V}^\top = \begin{pmatrix} \frac{5}{\sqrt{30}} & \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{30}} \\ 0 & \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{pmatrix}$
- $\text{rk}(\mathbf{A}) = 2$  because we have two singular values  $\sigma_1 = \sqrt{6}$  and  $\sigma_2 = 1$
- $\Sigma = \begin{pmatrix} \sqrt{6} & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$

- $\mathbf{u}_1 = \mathbf{A} \mathbf{v}_1 / \sigma_1 = \begin{pmatrix} \frac{1}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} \end{pmatrix}$
- $\mathbf{u}_2 = \mathbf{A} \mathbf{v}_2 / \sigma_2 = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{pmatrix}$
- $\mathbf{U} = (\mathbf{u}_1 \ \mathbf{u}_2) = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}$
- Then, we can see that  $\mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^\top$ .



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

- SVD: **always** exists, EVD: **square** matrix and exists if we can find **a basis of eigenvectors** (such as symmetric matrices)
- $\mathbf{P}$  in EVD is **not necessarily orthogonal** (only true for symmetric  $\mathbf{A}$ ), but  $\mathbf{U}$  and  $\mathbf{V}$  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  $\mathbf{A}$  are eigenvectors of  $\mathbf{A}\mathbf{A}^T$  (resp.  $\mathbf{A}^T\mathbf{A}$ )
  - The singular values of  $\mathbf{A}$  are the square roots of eigenvalues of  $\mathbf{A}\mathbf{A}^T$  and  $\mathbf{A}^T\mathbf{A}$
- When  $\mathbf{A}$  is symmetric, EVD = SVD (from spectral theorem)

## Different Forms of SVD

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

$$\mathbf{A} = \overbrace{\mathbf{U}}^{m \times r} \overbrace{\Sigma}^{r \times r} \overbrace{\mathbf{V}^T}^{r \times n}$$

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

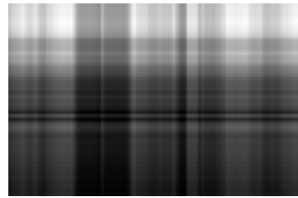
$$\mathbf{A} \approx \overbrace{\mathbf{U}}^{m \times k} \overbrace{\Sigma}^{k \times k} \overbrace{\mathbf{V}^T}^{k \times n}$$

We will cover this in the next slides.

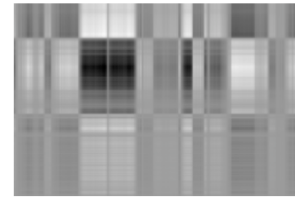
# Matrix Approximation via SVD



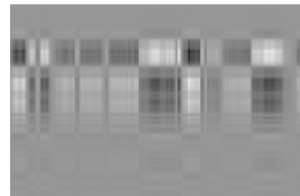
(a) Original image  $\mathbf{A}$ .



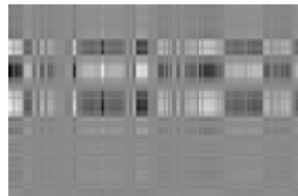
(b)  $\mathbf{A}_1$ ,  $\sigma_1 \approx 228,052$ .



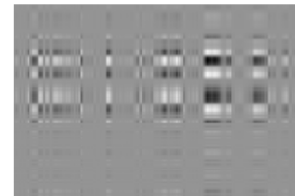
(c)  $\mathbf{A}_2$ ,  $\sigma_2 \approx 40,647$ .



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



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



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

- $\mathbf{A} = \sum_{i=1}^r \sigma_i \underbrace{\mathbf{u}_i \mathbf{v}_i^T}_{\mathbf{A}_i}$ , where  $\mathbf{A}_i$  is the outer product<sup>3</sup> of  $\mathbf{u}_i$  and  $\mathbf{v}_i$
- Rank  $k$ -approximation:  $\hat{\mathbf{A}}(k) = \sum_{i=1}^k \sigma_i \mathbf{A}_i$ ,  $k < r$

---

<sup>3</sup>If  $\mathbf{u}$  and  $\mathbf{v}$  are both nonzero, then the outer product matrix  $\mathbf{u}\mathbf{v}^T$  always has matrix rank 1. Indeed, the columns of the outer product are all proportional to the first column.

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

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

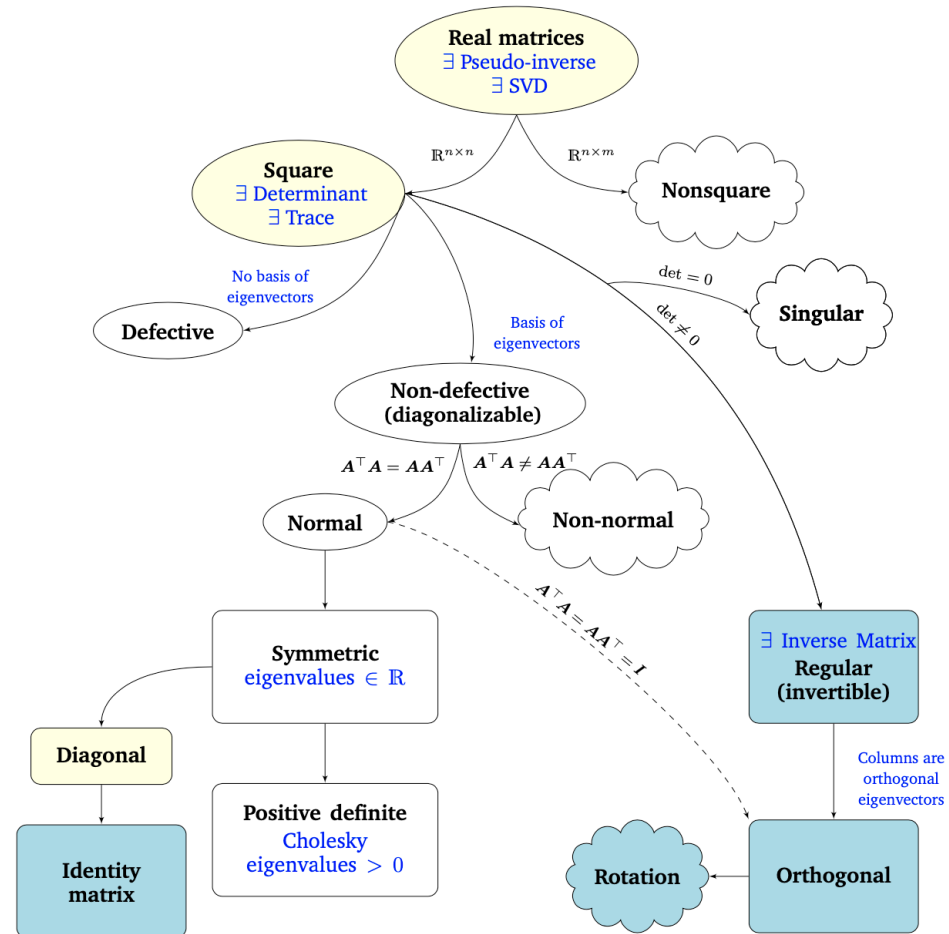
$$\hat{\mathbf{A}}(k) = \arg \min_{\text{rk}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_2, \quad \text{and} \quad \left\| \mathbf{A} - \hat{\mathbf{A}}(k) \right\|_2 = \sigma_{k+1}$$

- Quantifies how much error is introduced by the SVD-based approximation
- $\hat{\mathbf{A}}(k)$  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  $\mathbf{A}$  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?

## References

- [1] This lecture slide is mainly based upon <https://yung-web.github.io/home/courses/mathml.html> (made by Prof. Yung Yi, KAIST EE)
- [2] Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. Mathematics for machine learning. Cambridge University Press, 2020.