Lecture 05: A Refresher in Linear Algebra

Nikola Zlatanov

Innopolis University

Advanced Statistics

20-th of Feb to 27-th of Feb, 2023

Spectral Theorem

- Thm (Spectral Thm): $\forall$ symmetric $n \times n$ matrix $\boldsymbol{A}$, the i-th eigenvalue and eigenvector are defined as the unit vector, $\boldsymbol{u}_i$, and scalar λ_i that satisfy the following equality

$$\boldsymbol{A}\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i \tag{1}$$

Moreover, the following holds

  - all eigenvalues, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ are real,
  - all unit eigenvectors, $\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_n$, corresponding to eigenvalues, $\lambda_1, \lambda_2, ..., \lambda_n$, respectively, are orthonormal, and therefore they form the orthonormal basis of $\mathbb{R}^n$

$$u_i^T u_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Spectral Decomposition (SD

- Cor (Spectral Decomposition (SD)): $\forall$ symmetric $n \times n$ matrix $\boldsymbol{A}$, can be written as

$$\boldsymbol{A} = \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T, \tag{2}$$

where $r = \text{rank}(\boldsymbol{A})$ and $\lambda_i = 0$, for $r < i \leq n$
Proof: $\forall$ $\boldsymbol{x}$, can be written as $\boldsymbol{x} = \sum_{i=1}^{n} (\boldsymbol{u}_i^T \boldsymbol{x}) \boldsymbol{u}_i$, which is the basis expansion of $\boldsymbol{x}$. Multiplying both sides by $\boldsymbol{A}$, we get

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x} &= \sum_{i=1}^{n} (\boldsymbol{u}_i^T \boldsymbol{x}) \boldsymbol{A} \boldsymbol{u}_i \overset{(a)}{=} \sum_{i=1}^{n} (\boldsymbol{u}_i^T \boldsymbol{x}) \lambda_i \boldsymbol{u}_i = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i (\boldsymbol{u}_i^T \boldsymbol{x}) \\ &= \left(\sum_{i=1}^{n} (\lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T) \right) \boldsymbol{x} \overset{(a)}{=} \boldsymbol{A}\boldsymbol{x} \end{aligned} \tag{3}$$

where (a) comes from (1) and (b) holds only if (2) holds. Q.E.D.

Matrix Form of the SD

- Thm (Matrix Form of the SD): Let $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_n]$ and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$. Then,

$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \end{bmatrix}$

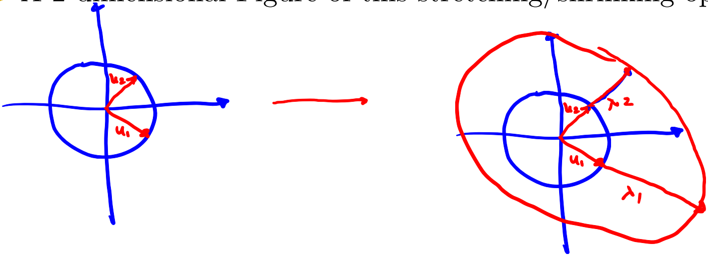
$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$$

  Proof: Prove at home.

- The matrix $\boldsymbol{U}$ is called an orthogonal matrix.
- A matrix $\boldsymbol{U}$ is orthogonal if it satisfies any of the following equivalent properties
  - $\boldsymbol{U}$ is invertible and $\boldsymbol{U}^{-1} = \boldsymbol{U}^T$
  - $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$
  - $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$
  - The columns of $\boldsymbol{U}$ are mutually orthonormal
  - The rows of $\boldsymbol{U}$ are mutually orthonormal

Geometric Form of the SD Thm

- Geometrically, we can view $\boldsymbol{A}$ as a operator that stretches or shrinks the orthonormal basis $\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_n$ by $\lambda_1, \lambda_2, ..., \lambda_n$, respectively.
- Hence, it transforms a n-dimensional unit sphere into an n-dimensional ellipsoid.
- A 2-dimensional Figure of this stretching/shrinking operation:



Optimization Form of the SD Thm:

- We can also view the SD as a series of optimization problems as

$$\lambda_1 = \max_{\boldsymbol{x}:\ ||\boldsymbol{x}||_2=1} \boldsymbol{x}^T A \boldsymbol{x}, \text{ where: } \boldsymbol{u}_1 = \arg\max_{\boldsymbol{x}:\ ||\boldsymbol{x}||_2=1} \boldsymbol{x}^T A \boldsymbol{x} \tag{4}$$

$$\lambda_2 = \max_{\boldsymbol{x}:\ \substack{||\boldsymbol{x}||_2=1 \\ \boldsymbol{x}\perp\boldsymbol{u}_1}} \boldsymbol{x}^T A \boldsymbol{x}, \text{ where: } \boldsymbol{u}_2 = \arg\max_{\boldsymbol{x}:\ \substack{||\boldsymbol{x}||_2=1 \\ \boldsymbol{x}\perp\boldsymbol{u}_1}} \boldsymbol{x}^T A \boldsymbol{x} \tag{5}$$

$$\vdots$$

$$\lambda_i = \max_{\boldsymbol{x}:\ \substack{||\boldsymbol{x}||_2=1 \\ \boldsymbol{x}\perp\{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{i-1}\}}} \boldsymbol{x}^T A \boldsymbol{x}, \text{ where: } \boldsymbol{u}_i = \arg\max_{\boldsymbol{x}:\ \substack{||\boldsymbol{x}||_2=1 \\ \boldsymbol{x}\perp\{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{i-1}\}}} \boldsymbol{x}^T A \boldsymbol{x} \tag{6}$$

$$\vdots$$

$$\lambda_r = \min_{\boldsymbol{x}:\ ||\boldsymbol{x}||_2=1} \boldsymbol{x}^T A \boldsymbol{x}, \text{ where: } \boldsymbol{u}_r = \arg\min_{\boldsymbol{x}:\ ||\boldsymbol{x}||_2=1} \boldsymbol{x}^T A \boldsymbol{x} \tag{7}$$

Singular Value Decomposition (SVD)

- All of the above holds for symmetric matrices. We need to extend this to non-symmetric matrices.
- Thm (Singular Value Decomposition (SVD)): $\forall$ general $m \times n$ matrix $\boldsymbol{A}$, can be written as

$$\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T, \text{ where } r = \text{rank}(\boldsymbol{A})$$
$$\text{and } \sigma_i = 0, \text{ for } r < i \leq \min\{m, n\} \tag{8}$$

  where $\sigma_1 \geq \sigma_2 \geq ...\sigma_r \geq 0$. The vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_m$ are mutually orthonormal and thereby form a orthonormal basis on $\mathbb{R}^m$, and the vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_n$ are mutually orthonormal and thereby form a orthonormal basis on $\mathbb{R}^n$.
- The vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_m$ are called "left singular value vectors" and the vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_n$ are called "right singular value vectors".

Informal Proof of SVD

- Since (8) holds, let's see what are $\boldsymbol{A}^T\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^T$

$$\boldsymbol{A}\boldsymbol{A}^T = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \left(\sum_{j=1}^{r} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^T \right)^T = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \sum_{j=1}^{r} \sigma_j \boldsymbol{v}_j \boldsymbol{u}_j^T$$

$$= \sum_{i=1}^{r} \sum_{j=1, j\neq 1}^{r} \sigma_i \boldsymbol{u}_i (\boldsymbol{v}_i^T \boldsymbol{v}_j) \boldsymbol{u}_j^T + \sum_{i=1}^{r} \sigma_i^2 \boldsymbol{u}_i (\boldsymbol{v}_i^T \boldsymbol{v}_j) \boldsymbol{u}_j^T = \sum_{i=1}^{r} \sigma_i^2 \boldsymbol{u}_i \boldsymbol{u}_j^T$$

$\sigma_i^2 = \lambda(AA^T)$

Similairy,

$$\boldsymbol{A}^T\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i^2 \boldsymbol{v}_i \boldsymbol{v}_j^T$$ $\to \sigma_i^2 = \lambda(A^TA)$

Hence, $\sigma_i^2 = \lambda_i(\boldsymbol{A}\boldsymbol{A}^T) = \lambda_i(\boldsymbol{A}^T\boldsymbol{A})$. Vectors $\boldsymbol{u}_i$'s and $\boldsymbol{v}_i$'s are the eigenvectors of $\boldsymbol{A}\boldsymbol{A}^T$ and $\boldsymbol{A}^T\boldsymbol{A}$, respectively. Now finish the proof.

Matrix form of the SVD

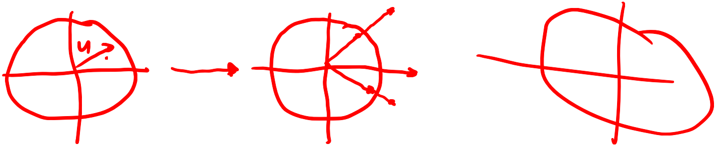
- Thm (Matrix form of the SVD): Let $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_m]$, $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_n]$ and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \sigma_2, ..., \sigma_{\min\{m,n\}})$, where $\boldsymbol{\Sigma}$ is a diagonal $m \times n$ matrix. Then,

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

Proof: Prove at home.

Geometric Form of the SVD Thm

- Geometric Form of the SVD Thm: We can see $\boldsymbol{A}$ as a operator that first rotates $\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_r$ by $\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_r$, respectively, and then stretches/shrinks the corresponding rotated vectors by $\sigma_1, \sigma_2, ..., \sigma_n$, respectively.
- Hence, it rotates a n-dimensional unit sphere to another sphere and then transforms it into an n-dimensional ellipsoid.
- A 2-dimensional Figure of this stretching/shrinking operation:



Optimization Form of the SVD Thm:

- Optimization Form of the SVD Thm:

$$\sigma_1 = \max_{\boldsymbol{x}:\ ||\boldsymbol{x}||_2=1} ||\boldsymbol{A}\boldsymbol{x}||_2, \text{ where: } \boldsymbol{v}_1 = \arg\max_{\boldsymbol{x}:\ ||\boldsymbol{x}||_2=1} ||\boldsymbol{A}\boldsymbol{x}||_2 \tag{9}$$

$$\sigma_2 = \max_{\substack{\boldsymbol{x}:\ ||\boldsymbol{x}||_2=1 \\ \boldsymbol{x}\perp\boldsymbol{v}_1}} ||\boldsymbol{A}\boldsymbol{x}||_2, \text{ where: } \boldsymbol{v}_2 = \arg\max_{\substack{\boldsymbol{x}:\ ||\boldsymbol{x}||_2=1 \\ \boldsymbol{x}\perp\boldsymbol{v}_1}} ||\boldsymbol{A}\boldsymbol{x}||_2 \tag{10}$$

$$\vdots$$

$$\sigma_i = \max_{\substack{\boldsymbol{x}:\ ||\boldsymbol{x}||_2=1 \\ \boldsymbol{x}\perp\{\boldsymbol{v}_1,\dots,\boldsymbol{v}_{i-1}\}}} ||\boldsymbol{A}\boldsymbol{x}||_2, \text{ where: } \boldsymbol{v}_i = \arg\max_{\substack{\boldsymbol{x}:\ ||\boldsymbol{x}||_2=1 \\ \boldsymbol{x}\perp\{\boldsymbol{v}_1,\dots,\boldsymbol{v}_{i-1}\}}} ||\boldsymbol{A}\boldsymbol{x}||_2 \tag{11}$$

$$\vdots$$

$$\sigma_r = \min_{\boldsymbol{x}:\ ||\boldsymbol{x}||_2=1} ||\boldsymbol{A}\boldsymbol{x}||_2, \text{ where: } \boldsymbol{v}_r = \arg\min_{\boldsymbol{x}:\ ||\boldsymbol{x}||_2=1} ||\boldsymbol{A}\boldsymbol{x}||_2 \tag{12}$$

Optimization Form of the SVD Thm:

- Moreover

$$\sigma_1 = \max_{\boldsymbol{x}:\ ||\boldsymbol{x}||_2=1} ||\boldsymbol{x}^T \boldsymbol{A}||_2, \text{ where: } \boldsymbol{u}_1 = \arg\max_{\boldsymbol{x}:\ ||\boldsymbol{x}||_2=1} ||\boldsymbol{x}^T \boldsymbol{A}||_2 \tag{13}$$

$$\sigma_2 = \max_{\boldsymbol{x}:\ \substack{||\boldsymbol{x}||_2=1 \\ \boldsymbol{x}\perp\boldsymbol{v}_1}} ||\boldsymbol{x}^T \boldsymbol{A}||_2, \text{ where: } \boldsymbol{u}_2 = \arg\max_{\boldsymbol{x}:\ \substack{||\boldsymbol{x}||_2=1 \\ \boldsymbol{x}\perp\boldsymbol{u}_1}} ||\boldsymbol{x}^T \boldsymbol{A}||_2 \tag{14}$$

$$\vdots$$

$$\sigma_i = \max_{\boldsymbol{x}:\ \substack{||\boldsymbol{x}||_2=1 \\ \boldsymbol{x}\perp\{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{i-1}\}}} ||\boldsymbol{x}^T \boldsymbol{A}||_2, \text{ where: } \boldsymbol{u}_i = \arg\max_{\boldsymbol{x}:\ \substack{||\boldsymbol{x}||_2=1 \\ \boldsymbol{x}\perp\{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{i-1}\}}} ||\boldsymbol{x}^T \boldsymbol{A}||_2 \tag{15}$$

$$\vdots$$

$$\sigma_r = \min_{\boldsymbol{x}:\ ||\boldsymbol{x}||_2=1} ||\boldsymbol{x}^T \boldsymbol{A}||_2, \text{ where: } \boldsymbol{u}_r = \arg\min_{\boldsymbol{x}:\ ||\boldsymbol{x}||_2=1} ||\boldsymbol{x}^T \boldsymbol{A}||_2 \tag{16}$$

Optimization Form of the SVD Thm:

- In all previous optimization problems, we can remove the restriction of searching only over unit vectors, i.e., $||\boldsymbol{x}||_2 = 1$, by optimizing $\frac{||\boldsymbol{Ax}||_2}{||\boldsymbol{x}||_2}$ instead of $||\boldsymbol{Ax}||_2$ and by optimizing $\frac{||\boldsymbol{x}^T\boldsymbol{A}||_2}{||\boldsymbol{x}||_2}$ instead of $||\boldsymbol{x}^T\boldsymbol{A}||_2$
- Hence,

$$\sigma_1 = \max_{\boldsymbol{x}} \frac{||\boldsymbol{Ax}||_2}{||\boldsymbol{x}||_2},$$

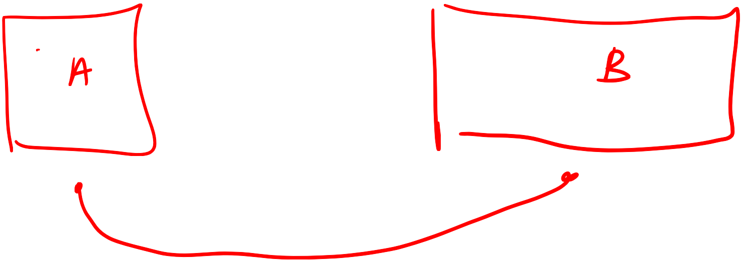
$$\sigma_r = \min_{\boldsymbol{x}} \frac{||\boldsymbol{Ax}||_2}{||\boldsymbol{x}||_2},$$

- And this can be looked at as the maximum and minimum possible distortion that the transformation $\boldsymbol{A}$ can impose on any vector $\boldsymbol{x}$. Example:

$$\sigma_r||\boldsymbol{x}||_2 \leq ||\boldsymbol{Ax}||_2 \leq \sigma_1||\boldsymbol{x}||_2, \ \forall \boldsymbol{x}$$

Matrix Norms

- Later on we would like to know how close is matrix $\boldsymbol{A}$ to $\boldsymbol{B}$
- Figure:



- Hence, we need some metric that measures distances between two matrices
- There are many such natural metrics. The most easy ones are the Frobenius norm and the Operator norm

Frobenius Norm

- Def: Frobenius (a.k.a. Hilbert Schmidt) norm of matrix $\boldsymbol{A}$, denoted by matrix $||\boldsymbol{A}||_F^2$ is
$$||\boldsymbol{A}||_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2 = \mathrm{tr}(A^T A) \tag{17}$$
- Hence, Frobenius norm of matrix $\boldsymbol{A}$ is simply a sum of all of its squared elements.
- Note that Frobenius norm does not distinguish between rows and columns. It pretends that the matrix is a vector.
- Def: Frobenius inner (dot) product between matrix $\boldsymbol{A}$ and $\boldsymbol{B}$, denoted by $\langle \boldsymbol{A}, \boldsymbol{B}\rangle_F$, is
$$\langle \boldsymbol{A}, \boldsymbol{B}\rangle_F = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij} b_{ij} = \mathrm{tr}(\underbrace{A^T B}_{c}) \tag{18}$$

Hence,
$$\langle \boldsymbol{A}, \boldsymbol{A}\rangle_F = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2 = \mathrm{tr}(A^T A) = ||\boldsymbol{A}||_F^2 \tag{19}$$

Frobenius Norm

Properties:

- Orthogonal Invariance: For any orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$

$$||\boldsymbol{U}\boldsymbol{A}||_F = ||\boldsymbol{A}\boldsymbol{V}||_F = ||\boldsymbol{A}||_F^2$$

  Proof: $\mathrm{tr}(A) = \mathrm{tr}(A^T)$

$$||\boldsymbol{U}\boldsymbol{A}||_F^2 = \langle \boldsymbol{U}\boldsymbol{A}, \boldsymbol{U}\boldsymbol{A}\rangle_F = \mathrm{tr}(\boldsymbol{A}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{A}) = ||\boldsymbol{U}\boldsymbol{A}||_F^2$$

- Euclidean norm of singular values: $||\boldsymbol{A}||_F^2 = \sum_{i=1}^{r} \sigma_i^2$

  Proof: $\mathrm{tr}(AA^T)$

$$||\boldsymbol{A}||_F^2 = ||\boldsymbol{U}^T\boldsymbol{\Sigma}\boldsymbol{V}||_F^2 = \mathrm{tr}(\boldsymbol{V}^T\boldsymbol{\Sigma}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{\Sigma}\boldsymbol{V}) = \mathrm{tr}(\boldsymbol{V}^T\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\boldsymbol{V})$$
$$= \mathrm{tr}(\boldsymbol{\Sigma}^T\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{\Sigma}) = \mathrm{tr}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}) = \sum_{i=1}^{r} \sigma_i^2 \qquad (20)$$

Operator Norm

- Def: Operator norm is the largest singular value i.e.,

$$||\boldsymbol{A}||_{op} = \sigma_1 = \max_{\boldsymbol{x}} \frac{||\boldsymbol{A}\boldsymbol{x}||_2}{||\boldsymbol{x}||_2}, \tag{21}$$

- Properties:

$$||\boldsymbol{A}||_{op} \le ||\boldsymbol{A}||_F \le \sqrt{n}||\boldsymbol{A}||_{op} \tag{22}$$

Proof:

$$\sigma_1^2 \le \sum_{i=1}^{r} \sigma_i^2 \le n\sigma_1^2 \tag{23}$$

taking square root, we obtain the desired result.