Session 7: Spectral theorem, PCA & Singular Value Decomposition

Optimization and Computational Linear Algebra for Data Science

Marylou Gabrié (based on material by Léo Miolane)

Midterm

- The Midterm exam is in 1 week.
- **Scope:** Session 1 to Session 6 included - HW1 to HW6 included
- **Knowing is not enough!** You need to practice: review problems available on the last year's course's webpage.
- **Practice is not enough!** You need to know the definitions/theorems/propositions.
- Past years midterms also available, with solutions.
- **Important:** when working on a problem, take **at least** 10min on it before looking at the solution (in case you are stuck).
- You can bring notes, but **if you think that you need them for the exam, you are probably not prepared enough**.

Contents

1. The Spectral theorem

1.1 The Spectral theorem

Theorem

Let $A \in \mathbb{R}^{n \times n}$ be a **symmetric** matrix. Then there is a orthonormal basis of $\mathbb{R}^n$ composed of eigenvectors of A.

That means that if A is symmetric, then there exists an orthonormal basis $(v_1, \dots, v_n)$ of $\mathbb{R}^n$ and $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ such that

$$Av_i = \lambda_i v_i \qquad \text{for all} \quad i \in \{1, \dots, n\}.$$

Theorem (Matrix formulation)

Let $A \in \mathbb{R}^{n \times n}$ be a **symmetric** matrix. Then there exists an orthogonal matrix P and a diagonal matrix D of sizes $n \times n$ such that

$$A = PDP^{\mathsf{T}}.$$

The spectral orthonormal basis

Geometric interpretation

1.2 Consequences

If

$$A = P \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} P^{\mathsf{T}}$$

for some orthogonal matrix P then:

Consequence #1: $\lambda_1, \ldots, \lambda_n$ are the only eigenvalues of A, and the number of time that an eigenvalue appear on the diagonal equals its multiplicity.

Proof sketch on an example

Consider $n = 3$ and

$$A = P \begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{pmatrix} P^{\mathsf{T}} \qquad \text{where} \qquad P = \begin{pmatrix} | & | & | \\ v_1 & v_2 & v_3 \\ | & | & | \end{pmatrix}$$

is an orthogonal matrix.

Proof sketch on an example

1.2 Consequences

If

$$A = P \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} P^{\mathsf{T}}$$

for some orthogonal matrix P then:

Consequence #2: The rank of A equals to the number of non-zero λ_i's on the diagonal:

$$\mathrm{rank}(A) = \#\{i \mid \lambda_i \neq 0\}.$$

Proof

1.2 Consequences

If

$$A = P \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} P^{\mathsf{T}}$$

for some orthogonal matrix P then:

Consequence #3: A is invertible if and only if $\lambda_i \neq 0$ for all i. In such case

$$A^{-1} = P \begin{pmatrix} 1/\lambda_1 & 0 & \cdots & 0 \\ 0 & 1/\lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1/\lambda_n \end{pmatrix} P^{\mathsf{T}}$$

Proof

1.2 Consequences

If

$$A = P \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} P^{\mathsf{T}}$$

for some orthogonal matrix P then:

Consequence #4: $\mathrm{Tr}(A) = \lambda_1 + \cdots + \lambda_n.$

1.3 The Theorem behind PCA

Theorem

Let A be a $n \times n$ symmetric matrix and let $\lambda_1 \geq \cdots \geq \lambda_n$ be its n eigenvalues and $v_1, \ldots, v_n$ be an associated orthonormal family of eigenvectors. Then

$$\lambda_1 = \max_{\|v\|=1} \ v^\mathsf{T} A v \qquad \text{and} \qquad v_1 = \operatorname*{arg\,max}_{\|v\|=1} \ v^\mathsf{T} A v \,.$$

Moreover, for $k = 2, \ldots, n$:

$$\lambda_k = \max_{\|v\|=1,\, v \perp v_1, \ldots, v_{k-1}} v^\mathsf{T} A v \,, \quad \text{and} \quad v_k = \operatorname*{arg\,max}_{\|v\|=1,\, v \perp v_1, \ldots, v_{k-1}} v^\mathsf{T} A v.$$

Proof

Proof

Proof

2. Principal Component Analysis

Empirical mean and covariance

We are given a dataset of n points $a_1, \dots, a_n \in \mathbb{R}^d$

$\underline{d = 1}$

- **Mean**

$$\mu = \frac{1}{n} \sum_{i=1}^{n} a_i \qquad \in \mathbb{R}$$

- **Variance**

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (a_i - \mu)^2 \qquad \in \mathbb{R}$$

Empirical mean and covariance

We are given a dataset of n points $a_1, \ldots, a_n \in \mathbb{R}^d$

<u>$d = 1$</u>

- **Mean**

$$\mu = \frac{1}{n} \sum_{i=1}^{n} a_i \qquad \in \mathbb{R}$$

- **Variance**

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (a_i - \mu)^2 \qquad \in \mathbb{R}$$

<u>$d \geq 2$</u>

- **Mean**

$$\mu = \frac{1}{n} \sum_{i=1}^{n} a_i \qquad \in \mathbb{R}^d$$

- **Covariance matrix**

$$S = \frac{1}{n} \sum_{i=1}^{n} (a_i - \mu)(a_i - \mu)^{\mathsf{T}} \qquad \in \mathbb{R}^{d \times d}$$

$$= \frac{1}{n} \sum_{i=1}^{n} a_i a_i^{\mathsf{T}} \qquad \text{if } \mu = 0.$$

PCA

- We are given a dataset of n points $a_1, \ldots, a_n \in \mathbb{R}^d$, where d is «large».
- **Goal:** represent this dataset in lower dimension, i.e. find $\widetilde{a}_1, \ldots, \widetilde{a}_n \in \mathbb{R}^k$ where $k \ll d$.
- Assume that the dataset is centered: $\sum_{i=1}^n a_i = 0$.
- Then, S can be simply written as:

$$S = \sum_{i=1}^{n} a_i a_i^\mathsf{T} = A^\mathsf{T} A.$$

where A is the $n \times d$ "data matrix":

$$A = \begin{pmatrix} - \; a_1^\mathsf{T} \; - \\ \vdots \\ - \; a_n^\mathsf{T} \; - \end{pmatrix}.$$

Direction of maximal variance

Direction of maximal variance

Direction of maximal variance

Good news: $S = A^\mathsf{T} A$ is symmetric.
Spectral Theorem: let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of S and $(v_1, \ldots, v_n)$ an associated orthonormal basis of eigenvectors.

2nd direction of maximal variance

j^{th} direction of maximal variance

- The « j^{th} direction of maximal variance » is v_j since v_j is solution of

$$\text{maximize } v^{\mathsf{T}} S v, \qquad \text{subject to } \|v\| = 1,\ v \perp v_1, v \perp v_2, \ldots, v \perp v_{j-1}.$$

- The dimensionally reduced dataset of in k-dimensions is then

$$\begin{pmatrix} \langle v_1, a_1 \rangle \\ \langle v_2, a_1 \rangle \\ \vdots \\ \langle v_k, a_1 \rangle \end{pmatrix}, \begin{pmatrix} \langle v_1, a_2 \rangle \\ \langle v_2, a_2 \rangle \\ \vdots \\ \langle v_k, a_2 \rangle \end{pmatrix}, \begin{pmatrix} \langle v_1, a_3 \rangle \\ \langle v_2, a_3 \rangle \\ \vdots \\ \langle v_k, a_3 \rangle \end{pmatrix} \cdots \begin{pmatrix} \langle v_1, a_n \rangle \\ \langle v_2, a_n \rangle \\ \vdots \\ \langle v_k, a_n \rangle \end{pmatrix}.$$

Recap

How to conpute reduced dimensional dataset?

Which value of k should we take?

Which value of k should we take?

3. Singular Value Decomposition

PCA

- Data matrix $A \in \mathbb{R}^{n \times m}$
- "Covariance matrix" $S = A^\mathsf{T} A \in \mathbb{R}^{m \times m}$.
- S is symmetric positive semi-definite.
- **Spectral Theorem:** there exists an orthonormal basis $v_1, \ldots, v_m$ of $\mathbb{R}^m$ such that the v_i's are eigenvectors of S associated with the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$.

Singular values/vectors

For $i = 1, \ldots, m$:

- we define $\sigma_i = \sqrt{\lambda_i}$, called the i^{th} **singular value** of A.
- we call v_j the i^{th} **right singular vector** of A.

For $i = 1, \ldots, r$:

- we call $u_i = \frac{1}{\sigma_i} A v_i$ the i^{th} **left singular vector** of A.

If $r < n$, we add $u_{r+1}, \cdots u_n$ such that $u_1, \cdots u_n$ is an orthonormal basis of $\mathbb{R}^n$.

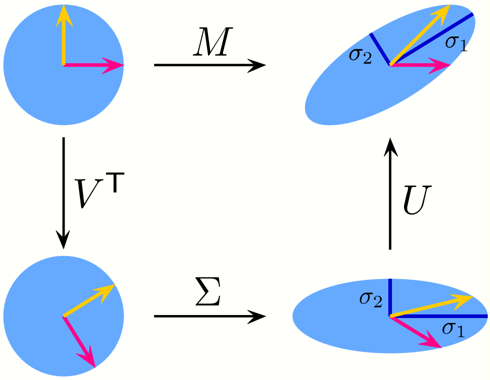
Singular Value decomposition

Theorem

Let $A \in \mathbb{R}^{n\times m}$. Then there exists two orthogonal matrices $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$ and a matrix $\Sigma \in \mathbb{R}^{n\times m}$ such that $\Sigma_{1,1} \geq \Sigma_{2,2} \geq \cdots \geq 0$ and $\Sigma_{i,j} = 0$ for $i \neq j$, that verify

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

Geometric interpretation of $U\Sigma V^\top$



$$M = U\Sigma V^\top$$

Questions?

Questions?