CSE 392: Matrix and Tensor Algorithms for Data

Instructor: Shashanka Ubaru

University of Texas, Austin
Spring 2024

Lecture 5: Matrix factorizations II - eigenvalue decomposition, PCA

Outline

1. Eigenvalue problems
2. PCA
3. Eigenfaces

Eigenvalue problems

Given a square matrix $\boldsymbol{A} \in \mathbb{C}^{n \times n}$, the *eigenvalue problem:*

$$\boldsymbol{A}\boldsymbol{u} = \lambda \boldsymbol{u}.$$

λ is an *eigenvalue* and $\boldsymbol{u}$ is an *eigenvector* of $\boldsymbol{A}$.

Types of problems:

- Find the largest or the smallest eigenvalues.
- Compute all eigenvalues in region of $\mathbb{C}$.
- Compute dominant eigenvalues and eigenvectors.

Applications: Structural engineering, stability analysis, electronic structure calculations, dimensionality reduction, spectral clustering and graphs, pagerank and many more.

Eigenvalues and properties

A complex scalar λ is called an *eigenvalue* of a square matrix $\boldsymbol{A} \in \mathbb{C}^{n \times n}$ if there exists a nonzero vector $\boldsymbol{u} \in \mathbb{C}^n$ such that

$$\boldsymbol{A}\boldsymbol{u} = \lambda \boldsymbol{u}.$$

The vector $\boldsymbol{u}$ is called an *eigenvector* of $\boldsymbol{A}$ associated with λ.

- λ is an eigenvalue iff the columns of $\boldsymbol{A} - \lambda \boldsymbol{I}$ are linearly dependent.
- That is, $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$.

Eigenvalues and properties II

- The set of all eigenvalues of $\boldsymbol{A}$, denoted $\Lambda(\boldsymbol{A})$, is the *spectrum* of $\boldsymbol{A}$.
- An eigenvalue is a root of the *characteristic polynomial*:

$$p_{\boldsymbol{A}}(\lambda) = \det(\boldsymbol{A} - \lambda \boldsymbol{I})$$

- So there are n eigenvalues (counted with their multiplicities).
- The multiplicity of these eigenvalues as roots of $p_{\boldsymbol{A}}$ are called *algebraic multiplicities*.
- The *geometric multiplicity* of an eigenvalue λ_i is the number of linearly independent eigenvectors associated with λ_i.
- Geometric multiplicity is $\leq$ algebraic multiplicity.

Eigenvalues and properties III

- **Diagonalization:** Two matrices $\boldsymbol{A}, \boldsymbol{B}$ are *similar* if there exists a nonsingular matrix $\boldsymbol{X}$ such that: $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{B}\boldsymbol{X}^{-1}$.
- $\boldsymbol{A}\boldsymbol{u} = \lambda\boldsymbol{u} \Leftrightarrow \boldsymbol{B}(\boldsymbol{X}^{-1}\boldsymbol{u}) = \lambda(\boldsymbol{X}^{-1}\boldsymbol{u})$
  eigenvalues remain the same, eigenvectors transformed.
- $\boldsymbol{A}$ is diagonalizable if it is similar to a diagonal matrix.
- *Transformations that preserve eigenvectors:*
  - Shift : $\boldsymbol{B} = (\boldsymbol{A} - \eta\boldsymbol{I})$
  - Polynomial : $\boldsymbol{B} = p(\boldsymbol{A})$
  - Inverse: $\boldsymbol{B} = \boldsymbol{A}^{-1}$
  - Shift and inverse: $\boldsymbol{B} = (\boldsymbol{A} - \eta\boldsymbol{I})^{-1}$

Symmetric eigenvalue problem

- For every square symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, we can compute eigendecompostion:

$$\boldsymbol{A} = \boldsymbol{U} \Lambda \boldsymbol{U}^\top,$$

  where $\boldsymbol{U}$ is an orthogonal matrix with eigenvectors $\boldsymbol{u}_i$ as columns, and Λ is diagonal matrix with eigenvalues λ_i on the diagonal.
- $\boldsymbol{U}$ forms an orthonormal basis of eigenvectors of $\boldsymbol{A}$.
- Eigenvalues of $\boldsymbol{A}$ are real.
- When $\boldsymbol{A}$ is real, $\boldsymbol{U}$ is also real.

The min-max theorem (Courant-Fischer)

Label eigenvalues decreasingly: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

The eigenvalues of a Hermitian matrix $\boldsymbol{A}$ are characterized by the relation

$$\lambda_k = \max_{\boldsymbol{S}, \dim(\boldsymbol{S})=k} \min_{\boldsymbol{x} \in \boldsymbol{S}, \boldsymbol{x} \neq 0} \frac{\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \rangle}{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$$

or

$$\lambda_k = \min_{\boldsymbol{S}, \dim(\boldsymbol{S})=n-k+1} \max_{\boldsymbol{x} \in \boldsymbol{S}, \boldsymbol{x} \neq 0} \frac{\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \rangle}{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$$

- $\frac{\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \rangle}{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$ is called the Rayleigh–Ritz quotient of $\boldsymbol{A}$.
- $\lambda_1 = \max_{\boldsymbol{x} \neq 0} \frac{\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \rangle}{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$ and $\lambda_n = \min_{\boldsymbol{x} \neq 0} \frac{\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \rangle}{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$.

Question: Use min-max theorem to show that $\sigma_1 = \|\boldsymbol{A}\|_2$.

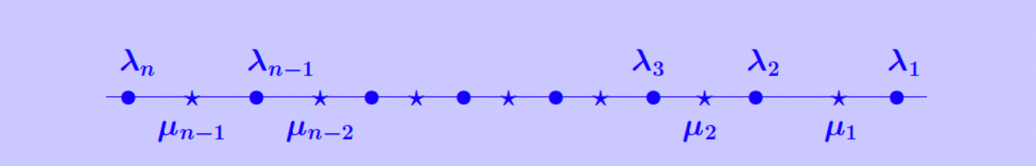
Interlacing Theorem

Suppose $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric. Let $\boldsymbol{B} \in \mathbb{R}^{m\times m}$ with $m < n$ be a principal submatrix (obtained by deleting both i-th row and i-th column for some values of i).
Suppose $\boldsymbol{A}$ has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$, and $\boldsymbol{B}$ has eigenvalues $\mu_1 \geq \cdots \geq \mu_m$. Then

$$\lambda_k \geq \mu_k \geq \lambda_{n+k-m} \text{ for } k = 1, \ldots, m$$

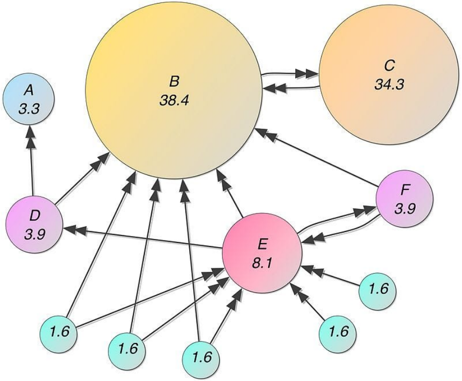
and if $m = n - 1$,

$$\lambda_1 \geq \mu_1 \geq \lambda_2 \geq \mu_2 \geq \cdots \geq \mu_{n-1} \geq \lambda_n$$



PageRank

- **PageRank** is the first Google algorithm developed to evaluate the quality and importance of web pages.
- Webgraph - created by all World Wide Web pages as nodes and hyperlinks as edges.
- Likelihood that a person randomly clicking on links will arrive at any particular page.



PageRank

- PageRank value of a page is given as:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)},$$

  $p_1, p_2, ..., p_N$ are the pages, $M(p_i)$ = set of pages that link to p_i, $L(p_j)$ = number of outbound links on page p_j, N = total number of pages, and d = damping factor.

- The values are the entries of the dominant right eigenvector of the modified adjacency matrix rescaled so that each column adds up to one.

$$\mathbf{r} = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_N) \end{bmatrix}$$

- $\mathbf{r}$ is the solution of the equation

$$\mathbf{r} = \begin{bmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{bmatrix} + d \begin{bmatrix} \ell(p_1, p_1) & \ell(p_1, p_2) & \cdots & \ell(p_1, p_N) \\ \ell(p_2, p_1) & \ddots & & \vdots \\ \vdots & & \ell(p_i, p_j) & \\ \ell(p_N, p_1) & \cdots & & \ell(p_N, p_N) \end{bmatrix} \mathbf{r}$$

  the adjacency function $\ell(p_i, p_j)$ is the ratio between number of links outbound from page j to page i to the total number of outbound links of page j.

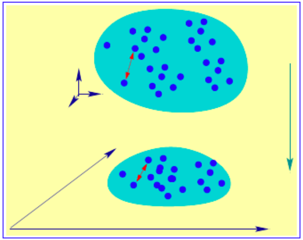
- $$\sum_{i=1}^{N} \ell(p_i, p_j) = 1,$$

  The matrix is a stochastic matrix. Closely related to the problem of finding the stationary points of Markov processes. It is also a variant of the eigenvector centrality measure used commonly in network analysis.

Dimensionality Reduction

Dimensionality Reduction

- **Dimensionality Reduction** (DR) techniques pervasive to many data applications.
- Reduce computational cost; but also more often :
  - reduce noise and redundancy in data, and
  - discover patterns.
- Given $\boldsymbol{x} \in \mathbb{R}^d$, and $k \ll d$, find the mapping $\Phi : \boldsymbol{x} \in \mathbb{R}^d \longrightarrow \boldsymbol{y} \in \mathbb{R}^k$.

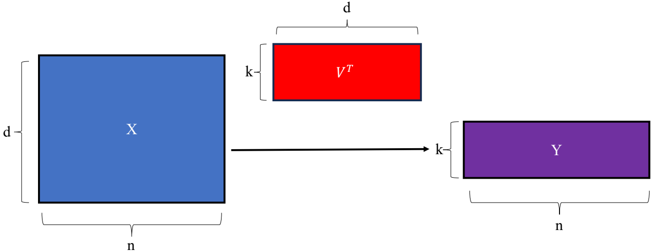


Projection-based Dimensionality Reduction

- Given dataset $\boldsymbol{X} = [\boldsymbol{x}_i, \ldots, \boldsymbol{x}_n]$, and dimension k, find the reduced set $\boldsymbol{Y}$.
- *Projection method*: Explicit mapping to the lower dimension

$$\boldsymbol{y} = \boldsymbol{V}^\top \boldsymbol{x}$$

  with $\boldsymbol{V} \in \mathbb{R}^{d \times k}$.
- Projection-based Dimensionality Reduction : $\boldsymbol{Y} = \boldsymbol{V}^\top \boldsymbol{X}$. Find the best such mapping (optimization) given that the $\boldsymbol{y}_i$'s must satisfy certain constraints.



Principal Component Analysis

- **Principal Component Analysis (PCA)** : find (orthogonal) $\boldsymbol{V}$ so that projected data $\boldsymbol{Y} = \boldsymbol{V}^\top \boldsymbol{X}$ has maximum variance.
- Maximize over all orthogonal $d \times k$ matrices $\boldsymbol{U}$:

$$\sum_i \|\boldsymbol{y}_i - \frac{1}{n}\sum_j \boldsymbol{y}_j\|_2^2 = \cdots = \mathrm{Tr}[\boldsymbol{V}^\top \bar{\boldsymbol{X}} \bar{\boldsymbol{X}}^\top \boldsymbol{V}],$$

  where $\bar{\boldsymbol{X}} = [\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_n]$ with $\bar{\boldsymbol{x}}_i = \boldsymbol{x}_i - \boldsymbol{\mu}$, and $\boldsymbol{\mu}$ =mean.
- *Solution:* $\boldsymbol{V}$ = dominant k eigenvectors of the covariance matrix. Top k singular vectors of $\bar{\boldsymbol{X}}$.

Exercises

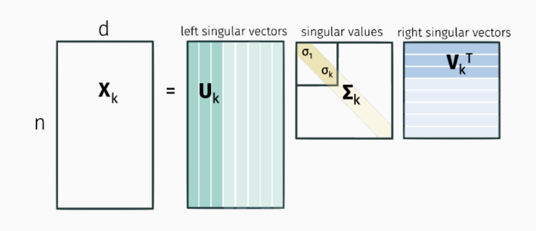
- Show that $\bar{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{I} - \frac{1}{n}\boldsymbol{e}\boldsymbol{e}^\top)$ (here $\boldsymbol{e}$ = vector of all ones). What does the projector $(\boldsymbol{I} - \frac{1}{n}\boldsymbol{e}\boldsymbol{e}^\top)$ do?
- Show that solution $\boldsymbol{V}$ also minimizes reconstruction error:

$$\sum_i \|\bar{\boldsymbol{x}}_i - \boldsymbol{V}\boldsymbol{V}^\top \bar{\boldsymbol{x}}_i\|^2 = \sum_i \|\bar{\boldsymbol{x}}_i - \boldsymbol{V}\bar{\boldsymbol{y}}_i\|^2$$

- It also maximizes $\sum_{i,j} \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2$

Low rank approximation

- Given a data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and integer k, find a rank-k approximation of $\boldsymbol{X}$.
- $\boldsymbol{X}_k = \boldsymbol{U}_k \Sigma_k \boldsymbol{V}_k^\top = \boldsymbol{U}_k \boldsymbol{U}_k^\top \boldsymbol{X} = \boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^\top$.



$$\boldsymbol{U}_k = \arg\min_{\boldsymbol{U}\in\mathbb{R}^{n\times k}} \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{X}\|_F^2 = \arg\max_{\boldsymbol{U}\in\mathbb{R}^{n\times k}} \|\boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{X}\|_F^2.$$

$$\|\boldsymbol{X} - \boldsymbol{X}_k\|_F^2 = \sum_{i=k+1}^{n} \sigma_i^2.$$

Eigenfaces