CSE 392: Matrix and Tensor Algorithms for Data	Spring 2024
Lecture 5 — 31 Jan '24	
Instructor: Shashanka Ubaru	*Scribe: Jenil Shah*

1 Eigenvalues and Eigenvectors

1.1 Introduction

Definition. Eigenvalue/eigenvector Given a square matrix A $\in \mathbb{C}^{n\times n}$, all vectors $\boldsymbol{u}$ and real $\lambda \in \mathbb{C}$ that satisfy

$$A\boldsymbol{u} = \lambda\boldsymbol{u} \tag{1}$$

are eigenvectors of A, and λ are their corresponding eigenvalues.

Typically, the dominant eigenvalues or the largest and smallest eigenvalues are of the most interest. Eigenvalues and eigenvectors find application in electronic structure calculations, Google's pagerank, etc.

1.2 Properties

1. λ is an eigenvalue $\implies$ det(A-λI) = 0.
2. Eigenvalues are the roots of the characteristic polynomial, $p_A(\lambda) = det(A - \lambda I)$.
3. The multiplicity of eigenvalues as roots of p_A is called the algebraic multiplicty; whereas, the number of linearly independent eigenvectors corresponding to an eigenvalue is called its geometric multiplicity.
4. Geometric multiplicity $\leq$ Algebraic multiplicity.
5. **Similarity**: Two matrices A, B are similar if $\exists$ X st, A = XBX^{-1}. A and B have the same eigenvalues, whereas the eigenvectors of a are X$^{-1}\boldsymbol{u_B}$
6. A is said to be *diagonalizable* if it is similar to a diagonal matrix.
7. The following transformations preserve eigenvectors
   (a) *Shift*: $A - \eta I$
   (b) *Polynomial*: p(A)
   (c) *Inverse*: A^{-1}
   (d) *Shift and inverse*: $(A - \eta I)^{-1}$
8. $\forall$ square symmetric matrices A $\in \Re^{n\times n}, \exists$ orthogonal U s.t.,

   $$A = U\Lambda U^T$$

   , where columns of U are the eigenvectors and Λ is a diagonal matrix containing the eigenvalues (which are real).

1.3 Min-max theorem

$\frac{<\mathbf{A}\vec{x},\vec{x}>}{\vec{x},\vec{x}} = RR$ is called the Rayleigh Ritz quotient of A and non-zero x. Then, the min-max theorem states that,

$$\lambda_k = \max_{S,dim(S)=k} \min_{x \in S, x \neq 0} RR = \min_{S,dim(S)=n-k+1} \max_{x \in S, x \neq 0} RR \tag{2}$$

where λ_k is the kth largest eigenvalue of A.

1.4 Interlacing theorem

For a principal submatrix $\mathbf{B} \in \Re^{m \times m}$ with eigenvalues $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m$ of a symmetric matrix $\mathbf{A} \in \Re^{n \times n}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, the interlacing theorem states

$$\lambda_k \geq \mu_k \geq \lambda_{n+k-m} \ \forall \ k \in \{1, \cdots, m\} \tag{3}$$

2 Page Rank

Pagerank is an algorithm developed by Google for optimizing their search engine results. Each webpage is viewed as a node, and the rank denotes the "importance" of the webpages, ie, the likelihood that a person randomly clicking will arrive at that particular webapage, and it is given by:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \tag{4}$$

where p_i denote the webpages, $M(p_i)$ are the set of pages linking to p_i, $L(p_j)$ is the number of outbound links, N is the total number of pages and d is the damping factor.

The solution, $r = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \cdots \\ PR(p_n) \end{bmatrix}$ is the solution of the below equation:

$$\mathbf{r} = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \cdots \\ PR(p_n) \end{bmatrix} + d \begin{bmatrix} l(p_1,p_1) & l(p_1,p_2) & \cdots \\ l(p_2,p_1) & \cdots & \cdots \\ \cdots & l(p_i,p_j) & \cdots \\ l(p_N,p_1) & \cdots & l(p_N,p_N) \end{bmatrix} \mathbf{r} \tag{5}$$

where the elements of the adjacency matrix $l(p_i,p_j)$ are the number of outbound links from p_i to p_j, and $\sum_i l(p_i,p_j) = 1$. This is a stochastic matrix, and the solution to this system is closely related to finding the stationary points in a Markov process.

3 Dimensionality Reduction

The idea is to find a map $\Phi : \mathbf{x} \in \Re^d \to y \in \Re^k$ where $k \ll d$, to reduce noise and discover patterns in data.

3.1 Projection based Dimensionality Reduction

Map $\mathbf{X} \in \Re^{n \times d}$ to a $\mathbf{Y} \in \Re^{k \times d}$ with explicit mapping:

$$\mathbf{y_i} = \mathbf{V^T x_i} \tag{6}$$

Different projections based on the constraints on $\mathbf{y_i}$.

Eg: **Principal component analysis (PCA)**. In PCA, the constraint is to find an orthogonal map $\mathbf{V}$ st the projection $\mathbf{V^T X}$ captures the maximum variance. This means:

$$\begin{aligned} V = argmax_k \sum_i ||\mathbf{y_{ik}} - \frac{\sum_j \mathbf{y_{jk}}}{n}||_2 = \sum_i ||\mathbf{V_k x_{ik}} - \frac{\sum_j \mathbf{V_k x_{ik}}}{n}||_2 = \sum_i ||\mathbf{V_k}(\mathbf{x_{ik}} - \mathbf{\bar{x}_{ik}})||_2 \\ = ||\mathbf{V_k^T \bar{X}}||_F = Tr[\mathbf{V_k^T \bar{X} \bar{X}^T V_k}] \end{aligned} \tag{7}$$

The solution to this is columns of $\mathbf{V}, \mathbf{v_i}$ being the i^{th} top singular columns of $\mathbf{\bar{X}}$.

3.2 Low rank approximation

For a data matrix $\mathbf{X} \in \Re^{n \times d}$ the best rank-K approximation of X is given by Eckhart-Young-Mirsky as:

$$\mathbf{X}_k = \mathbf{U}_k \Sigma_k \mathbf{V}_k = \mathbf{U}_k \mathbf{U}_k^T \mathbf{X} = \mathbf{X} \mathbf{V}_k \mathbf{V}_k^T \tag{8}$$

Also,

$$\min_{X_k} ||\mathbf{X} - \mathbf{X}_k||_F = \sum_{k+1}^{n} \sigma_i^2$$

.