Graph 3. Matrix Factorization via

Eigendecompostion (ED) and Singular Value Decomposition (SVD)

1

Outline

- Motivation with PCA
- ED and SVD
- Convergence of HITS

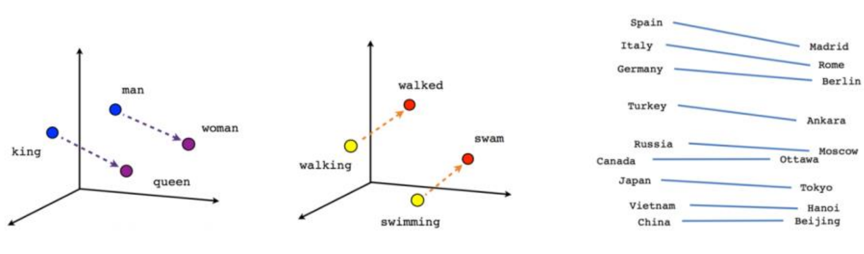
2

Why do we care about ED and SVD?

- We want to know why HITS converges and where does it converge.
- We want to understand how to visualize word embeddings in a 2D or 3D space (via PCA projection of high-dimensional vectors).
- We want to understand how to use a graph to propagate signals over nodes smoothly (later lectures on Laplacian Eigenmaps & Graph Convolution Networks)

Visualizing Word Embeddings in 2D or 3D

- Allowing us to see some interesting (analogical) patterns



Male-Female

Verb tense

Country-Capital

Consider a word embedding matrix

$$X_{n\times m} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & x_{ij} & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \text{ n words}$$

m features (factors)

- Each **row** is the embedding of a word.
- Each **column** is a latent feature of words.
- Each **cell** is the feature weight of the corresponding word.

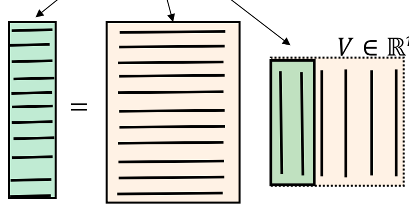
Principal Component Analysis (PCA)

https://en.wikipedia.org/wiki/Principal_component_analysis

- Projecting high-dimensional data (row vectors in X) to k-dimensional

$$T_k = XV_k \qquad (k < m)$$

$V \in \mathbb{R}^{m\times m},$



$T_k \in \mathbb{R}^{n\times k},$ $X \in \mathbb{R}^{n\times m},$ $V_k \in \mathbb{R}^{m\times k},$

$V = (v_1, v_2, \cdots, v_m)$ are the eigenvectors of $X^T X$.

Steps and Intuition in PCA
https://en.wikipedia.org/wiki/Principal_component_analysis

- **Preprocessing:** Make each row of input matrix $X^{(0)}$ to the centroid of all the row vectors

$$\mu := \frac{1}{n}\sum_{i=1}^{n} X_{i:}^{(0)}, \qquad X := X^{(0)} - \hat{1}\mu$$

- **Obtaining the top-k eigenvectors of matrix** $X^T X$

$$V_k = (v_1, \dots, v_k), \qquad k < m$$

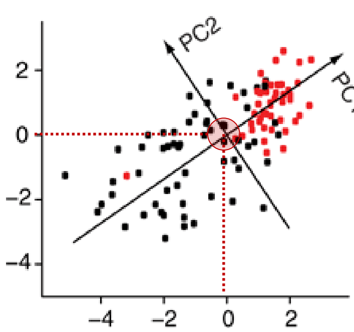
- **Projecting** the row vectors of X onto the top-k eigenvectors

$$T_k = XV_k$$

- **Intuition**: Preserving most variance in data with

$$Var(v_1; X) \geq var(v_2; X) \geq \cdots \geq var(v_m; X)$$

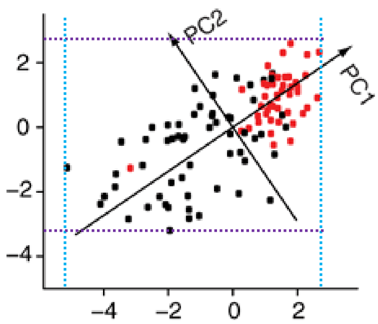
PCA of 2D Data



By rotating the orthogonal axes, we can see

- The 1st eigenvector (PC1=v_1) identifies the direction with the maximum variance in data.
- The 2nd eigenvector (PC2 =v_2) identifies the direction with the secondary variance in data.

Variance in data on the original basis (X_1, X_2)



- Sample variance in direction X_1

$$var\ (X_1) = \frac{1}{(n-1)} \sum_{i=1}^{n} x_{i1}^2$$

- Sample variance in direction X_2

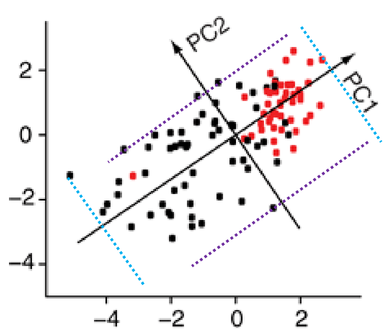
$$var(X_2) = \frac{1}{(n-1)} \sum_{i=1}^{n} x_{i2}^2$$

- Total variance in data

$$var(X) = var(X_1) + var(X_2)$$

9

Variance on the new basis (PC1 and PC2)



- Sample variance in direction Z_1 (PC1)

$$var(Z_1) = \frac{1}{(n-1)} \sum_{i=1}^{n} z_{i1}^2$$

- Sample variance in direction Z_2 (PC2)

$$var(Z_2) = \frac{1}{(n-1)} \sum_{i=1}^{n} z_{i2}^2$$

- Total variance in data (does not change)

$$var(X) = var(Z_1) + var(Z_2)$$

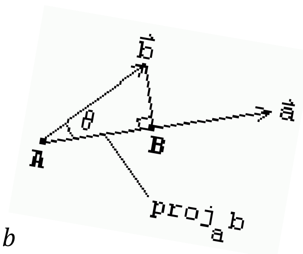
- Z_1 is the direction preserving the maximal variance in data!

10

Vector Projection

Given $a, b \in R^m$

- $a \cdot b \triangleq \sum_{i=1}^{n} a_i b_i$
- $\cos(a,b) \triangleq \frac{a \cdot b}{\|a\| \times \|b\|}$
- Projection of vector b onto a



$$\|b\|\cos(\theta) = \|b\| \frac{a \cdot b}{\|a\| \times \|b\|} = \frac{a \cdot b}{\|a\|} = \frac{a}{\|a\|} \cdot b$$

- Projection of any row vector $x_i \in X$ onto unit vector v is $(x_i \cdot v)$.

10/8/2024 @Yiming Yang, Lecture on PCA, SVD and LSI 11

11

First Component (PC1)

- Sample variance of X on any unit vector v is

$$\text{sample-var}(v; X) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i \cdot v)^2$$

- To maximize the variance, we define PC1 as

$$v_1 = \text{argmax}_{\|v\|=1} \left\{ \sum_i \, (x_i \cdot v)^2 \right\}$$

$$= \text{argmax}_{\|v\|=1} \{\|Xv\|^2\} = \text{argmax}_{\|v\|=1} \{v^T X^T X v\}$$

- Equivalently, we can solve $(v_1, \lambda_1) \;\; = \text{argmax}_{v,\lambda} \{v^T X^T X v + \lambda(1 - v^T v)\}$
- Defining $f(v,\lambda) = v^T X^T X v + \lambda(1 - v^T v)$, we have the its mode(s) with

$$0 = \nabla_v f(v,\lambda) = 2X^T X v - 2\lambda v, \;\; i.e., X^T X v = \lambda v$$

- Clearly, v and λ must be an eigenvector/eigenvalue of $X^T X$.

10/8/2024 @Yiming Yang, Lecture on PCA, SVD and LSI 12

12

First and Other Components

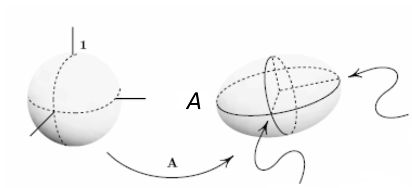
- Necessary Condition
  - v_1 must be an eigenvector of $X^T X$ and λ must be the corresponding eigenvalue.
- Sufficient condition
  - v_1 must be the eigenvector corresponding to $\lambda_1 = argmax_{\lambda}(v^T X^T X v)$
- How do we find v_1?
  - Power Iteration with $B = X^T X$ until convergence.
- How do we find next $v_{2,}\ v_3, \ldots$?

$$\hat{X}_k := X - \sum_{j=1}^{k-1} X v_j v_j^T\ , \qquad v_k = argmax_{\|v\|=1}\left\{\frac{v^T \hat{X}_k^T \hat{X}_k v}{v^\wedge v}\right\}$$

13

Distortions by Au

(Carl Meyer, "Matrix Analysis and Applied Linear Algebra" http://www.matrixanalysis.com, Ch 5, p 281)



the direction with the largest variation of data in matrix A

$\max_{u:||u||=1} ||Au||_2 = \sqrt{\lambda_{\max}}$

$\min_{u:||u||=1} ||Au||_2 = \sqrt{\lambda_{\min}}$

The eigenvalues of $B = A^T A$ reflect *how much distortions may occur* when applying A to an arbitrary vector u on the surface of the unit ball.

14

Distortion Factor: *Rayleigh Quotient*

$$\max_x \left\{\frac{||Ax||}{||x||}\right\} = \sqrt{\lambda_1(B)} = \sigma_1(A)$$

- λ_1 is the largest eigenvalue of matrix $B = A^T A$
- σ_1 is the largest singular value of matrix A.

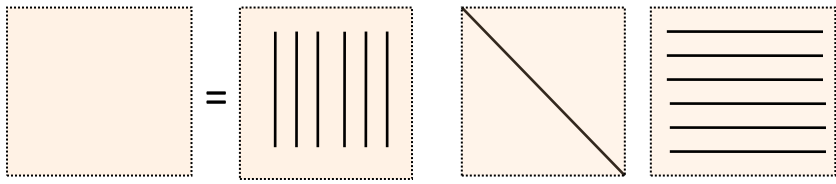
15

Outline

- Motivation with PCA
- ED and SVD
- Convergence of HITS

16

Eigendecompostion (Squared Matrix)



B: m x m   U: m x m $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_m)$ U^T : m x m

- Eigenvalues $\lambda_1, \ldots, \lambda_m$ are sorted by absolute value in decreasing order.
- Eigenvectors (the columns of matrix U) are ordered accordingly.

Properties of eigenvectors

- If u is an eigenvector, then $v = cu$ is also an eigenvector for any $c \in R$.

$$Bu = \lambda u \quad \rightarrow \quad B\underbrace{cu}_{v} = \lambda\underbrace{cu}_{v}$$

Thus, we only focus on unit-length eigenvectors.

- If u and v are both eigenvectors, $u + v$ is also an eigenvector.

$$Bu = \lambda u, \quad Bv = \lambda v, \quad \rightarrow \quad B\underbrace{(u+v)}_{w} = \lambda\underbrace{(u+v)}_{w}$$

Thus, we focus on linearly independent eigenvectors, especially on orthonormal eigenvectors

Properties of eigenvectors

- B (m-by-m) has at most m linearly independent eigenvectors.
- The number of distinct eigenvalues may be equal or smaller than the number of the linearly independent eigenvectors.
- For example, identity matrix I has m independent eigenvectors (each column vector is orthogonal from other columns) but only one distinct eigenvalue $\lambda = 1$, as shown below:

$$\mathrm{I}v = v$$

or $$0 = det(B - \lambda I) = (1-\lambda)^m \det(I) \rightarrow \lambda = 1$$

- We are interested in more general cases other than identity matrix.

Truncated Eigendecomposition

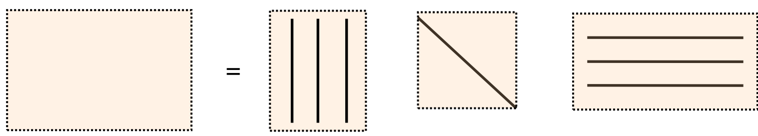
- Full dimension: $B_{m\times m} = U_{m\times m}\Lambda_{m\times m}U^T_{m\times m}$
- Dimension reduced:

$$\hat{B}_{m\times m} = \underbrace{U_{m\times k}}_{(u_1,\, u_2,\, \cdots, u_k)} \underbrace{\Lambda_{k\times k}}_{\uparrow \atop diag\{\lambda_1,\cdots,\lambda_k\}} U^T_{k\times m} \qquad k < m$$

$$= \sum_{i=1}^{k} \lambda_i\, u_i\, u_i{}^T$$

Singular Value Decomposition (SVD) of any matrix of rank $r \leq \min(n, m)$

X: n x m   U: n x r   Σ: r x r $\mathrm{V^T}$: r x m

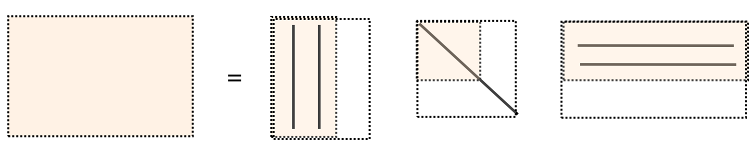


$$X = U\Sigma\mathrm{V^T} = (\vec{u}_1 \quad \vec{u}_2 \quad \cdots \quad \vec{u}_r)\begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{pmatrix}\begin{pmatrix} \vec{v}_1{}^T \\ \vec{v}_2{}^T \\ \vdots \\ \vec{v}_r{}^T \end{pmatrix} = \sum_{j=1}^{r} \sigma_j \vec{u}_j \vec{v}_j{}^T$$

$\Sigma = (\sigma_1, \sigma_2, \ldots, \sigma_r)$ are the **singular values ("spectrum")**;
$U = (u_1, u_2, \ldots, u_r)$ are the **left singular vectors** (orthonormal);
$V = (v_1, v_2, \ldots, v_r)$ are the **right singular vectors** (orthonormal).

Truncated SVD (with $k < r$)

X: n x m U_k : n x k   Σ : k x k V_k^T : k x m



$$X_k = U_k\Sigma_\mathrm{k}\mathrm{V}_k^\mathrm{T} = (\vec{u}_1 \quad \cdots \quad \vec{u}_k)\begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_k \end{pmatrix}\begin{pmatrix} \vec{v}_1{}^T \\ \vec{v}_2{}^T \\ \vdots \\ \vec{v}_k{}^T \end{pmatrix} = \sum_{j=1}^{k} \sigma_j \vec{u}_j \vec{v}_j{}^T$$

$(\sigma_1, \sigma_2, \ldots, \sigma_k)$ are the **top−k singular values**;
U_k and V_k contains the **top-k left/right singular vectors**, respectively.

SVD and ED

- Rectangular matrix

$$A = U\Sigma V^T$$

- Squared matrices

$$B_a = A^T A = V\Sigma^{\mathrm{T}}\boxed{U^T U}\Sigma V^T = V\ \Sigma^2 V^T = V\Lambda V^T$$

$$B_h = AA^T = U\Sigma\boxed{V^T V}\Sigma U^T = U\Sigma^2 U^T = U\Lambda U^T$$

$$\Sigma = diag(\sigma_1, \dots, \sigma_r)$$

$$\Lambda = diag(\lambda_1, \dots, \lambda_r),\ \lambda_i = \sigma_i^2\ \forall i$$

- U consists of the eigenvectors of B_h;
- V consists of the eigenvectors of B_a.

23

Outline

- Motivation with PCA
- ED and SVD
- Convergence of HITS

24

Recap the Statement about HITS' Convergence (in the lecture about HITS)

- https://en.wikipedia.org/wiki/Power_iteration

"If we assume the matrix has an eigenvalue that is strictly greater in magnitude than its other eigenvalues and the starting vector has a nonzero component in the direction of an eigenvector associated with the dominant eigenvalue, then a subsequence converges to the eigenvector associated with the dominant eigenvalue."

Convergence of Power Iteration in HITS (my sketch proof)

For precise proof see https://en.wikipedia.org/wiki/Power_iteration, which uses the Jordan Normal Decomposition instead of SVD.

$$B_a = A^T A \text{ where } A = U\Sigma V^T, \quad A \epsilon \{0,1\}^{n\times n},\ U \epsilon \mathbb{R}^{n\times r},\ V \epsilon \mathbb{R}^{n\times r}, r = rank(A)$$

$$B_a{}^k = (A^T A)^k = (V\Sigma \cancel{U^T U}\Sigma V^T)^k = V\Sigma^2 \cdots \Sigma^2 V^T = V\Sigma^{2k} V^T$$

$$B_a{}^k z = V\Sigma^{2k} \underbrace{V^T z}_{w} = V\Sigma^{2k} w\ , \quad z \in R^n, w \in R^r$$

$$= [\vec{v}_1 \cdots \vec{v}_r] \begin{bmatrix} \sigma_1{}^{2k} & & \\ & \ddots & \\ & & \sigma_r{}^{2k} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_r \end{bmatrix}$$

$$= w_1 \sigma_1{}^{2k} \vec{v}_1 + \cdots + w_r \sigma_r{}^{2k} \vec{v}_r$$ ← a linear combination of the eigenvectors

$$= \sigma_1{}^{2k} \left(w_1 \vec{v}_1 + w_2 \left(\frac{\sigma_2}{\sigma_1}\right)^{2k} \vec{v}_2 + w_3 \left(\frac{\sigma_3}{\sigma_1}\right)^{2k} \vec{v}_3 \cdots \right)$$

The 1st term dominates when k is large if w1 is non-zero and if $|\sigma_1| > |\sigma_2|$.