Graph 3. Matrix Factorization with

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 know how to project word embeddings onto a 2D space for visualization (via PCA).

- We want to understand how to use a graph to propagate signals over nodes with a desirable smoothness penalty (later lectures on Laplacian Eigenmaps, Graph Convolution Networks)
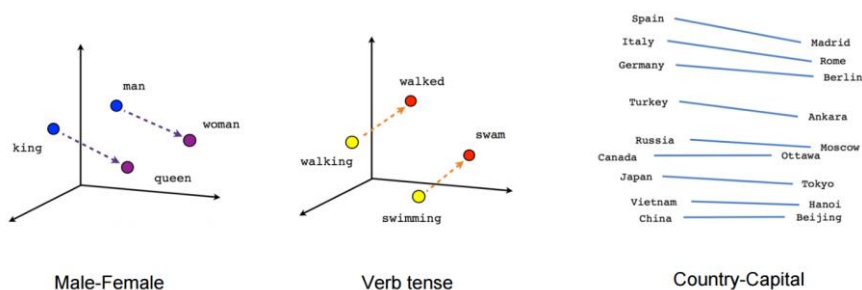
2/27/2024 @Yiming Yang, S24 Lecture on ED & SVD 3

3

Visualizing Word Embeddings in 2D or 3D

- Visualizing word embeddings in 2D allows us to see some interesting (analogical) patterns about their relationship.

Male-Female   Verb tense   Country-Capital

2/27/2024 @Yiming Yang, S24 Lecture on ED & SVD 4

4

Word Embedding Matrix

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

Features (factors)

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

5

Principal Component Analysis (PCA)
https://en.wikipedia.org/wiki/Principal_component_analysis

- PCA reduces the dimensionality of data (matrix X)

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

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

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

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

6

Page 3

3

Steps of PCA & Intuition

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

- **Preprocessing:** make each row of raw input $X^{(0)}$ to the centroid of its row vectors (the data points)

$$\mu := \frac{1}{n}\sum_{i=1}^{n} X_{i:}^{(0)}, \qquad X := X_0 - \hat{1}\mu$$

- **Computing the top-k eigenvectors of** $X^T X = V\Lambda V^T$ as

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

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

$$T_k = XV_k$$

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

$$Var(v_1; X) \geq var(v_2; X) \geq \cdots \geq var(v_m; X)$$
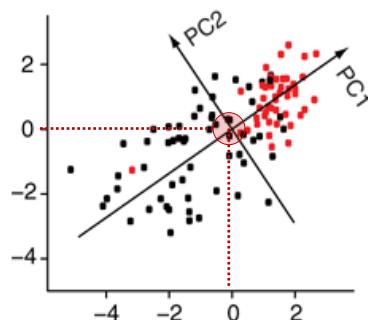
2/27/2024   @Yiming Yang, S24 Lecture on ED & SVD   7

7

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.
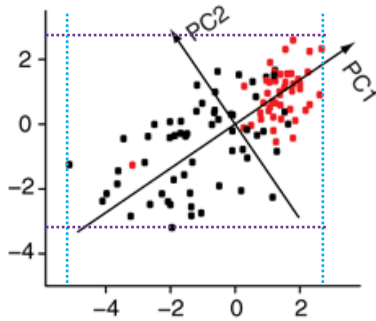
2/27/2024   @Yiming Yang, S24 Lecture on ED & SVD   8

8

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

$$var(X_2) = \frac{1}{(n-1)} \sum_{i=1}^{n} x_{i1}^2$$

- Total variance in data

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

9

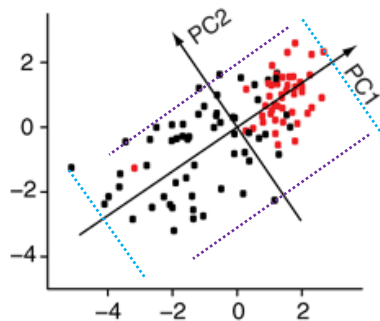# Variance on the new basis (PC1 and PC2)

- 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)

$$var(Z_2) = \frac{1}{(n-1)} \sum_{i=1}^{n} z_{i1}^2$$

- Total variance in data (does not change)

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

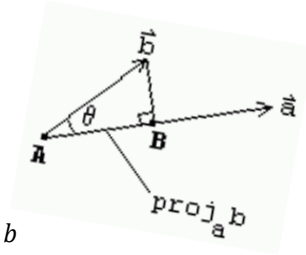- Z_1 is the direction preserving the maximal variance in data!

10

Page 5

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 $(x_i \cdot v)$.

11

First Component (PC1)

- Sample variance of X on new direction 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 $f(v, \lambda) = v X^T X v + \lambda(1 - v^T v)$, 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 $X^T X$.

12

First and Other Components

- Necessary Condition
  - v_1 must be an eigenvector of X^TX 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 $B = X^TX$ until convergence.
- How do we find next v_t for $t = 2, 3, \ldots$?

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

13

Distortions by Au
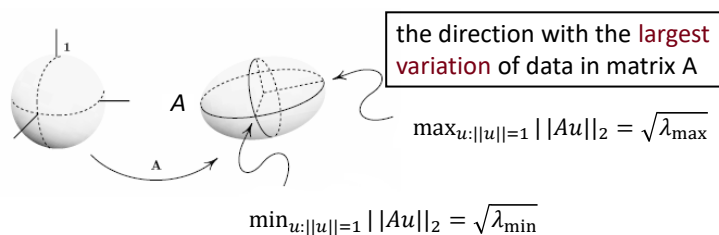(a figure from the book by Carl Mayer, Ch 5, p 281)

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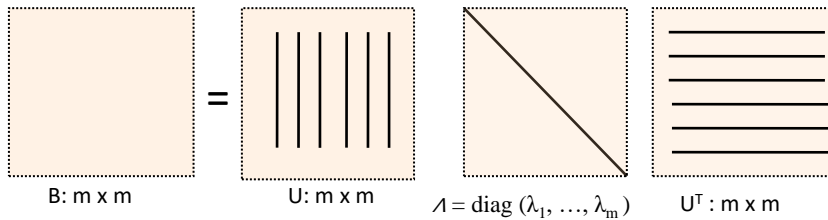
- Eigenvalues $\lambda_1, \ldots, \lambda_m$ are sorted by absolute value in decreasing order.
- Eigenvectors (the columns of matrix U) are ordered accordingly.

17

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

18

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:

$$I v = v$$

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

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

2/27/2024   @Yiming Yang, S24 Lecture on ED & SVD   19

19

Truncated Eigendecomposition

- Full dimension: $\quad B_{m\times m} = U_{m\times m} \Lambda_{m\times m} U_{m\times m}^T$

- Dimension reduced:

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

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

2/27/2024   @Yiming Yang, S24 Lecture on ED & SVD   20

20

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   V^T: r x m

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

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

21

Truncated SVD (with $k < r$)

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

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

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

22

SVD and ED

- Rectangular matrix
$$A = U\Sigma V^T$$
- Squared matrices
$$B_a = A^T A = V\Sigma^T U^T U\Sigma V^T = V\ \Sigma^2 V^T = V\Lambda V^T$$
$$B_h = AA^T = U\Sigma V^T V\Sigma U^T = U\Sigma^2 U^T = U\Lambda U^T$$
$$\Sigma = diag(\sigma_1, \dots, \sigma_r)$$
$$\Lambda = diag(\lambda_1, \dots, \lambda_r),\ \lambda_i = \sigma_i^2\ \forall i$$
- 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."

25

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.

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

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

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

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

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

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

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

26

26