

# AAIL100: Introduction to Data Science

Ragesh Jaiswal, IITD

## Best Fit Subspaces and Singular Value Decomposition (SVD)

# Singular Value Decomposition (SVD)

## Theorem

Let  $A$  be any  $n \times d$  matrix with right singular vectors  $\mathbf{v}_1, \dots, \mathbf{v}_r$ , left-singular vectors  $\mathbf{u}_1, \dots, \mathbf{u}_r$ , and corresponding singular values  $\sigma_1, \dots, \sigma_r$ . Then  $A = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ .

- The decomposition  $A = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T$  is called the **Singular Value Decomposition** (or SVD in short).
- In matrix notation, we can write  $A = UDV^T$  where:
  - $U$  is a  $n \times r$  matrix where the  $i^{\text{th}}$  column is  $\mathbf{u}_i$ .
  - $D$  is a  $r \times r$  diagonal matrix with the  $i^{\text{th}}$  diagonal element  $\sigma_i$ .
  - $V$  is a  $d \times r$  matrix where the  $i^{\text{th}}$  column is  $\mathbf{v}_i$ .
- Question: How do we compute the SVD?
- Question: What are the applications of SVD?

# Singular Value Decomposition (SVD)

## Best rank- $k$ approximation

- Let  $A = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T$  be the SVD of an  $n \times d$  matrix  $A$ .
- For  $k \in \{1, \dots, r\}$  let

$$A_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T \quad (\text{i.e., sum truncated to first } k \text{ elements})$$

- Claim 1:  $A_k$  has rank  $k$ .
- Claim 2: The rows of  $A_k$  are the projections of the rows of  $A$  onto the subspace  $V_k$  spanned by the first  $k$  singular vectors of  $A$ .
- We will prove that  $A_k$  is the best **rank  $k$  approximation** to  $A$  where the error is measured in terms of the Frobenius norm.

### Theorem

*For any matrix  $B$  with rank at most  $k$ :*

$$\|A - A_k\|_F \leq \|A - B\|_F.$$

# Singular Value Decomposition (SVD)

## Best rank- $k$ approximation

### Theorem

For any matrix  $B$  with rank at most  $k$ :

$$\|A - A_k\|_F \leq \|A - B\|_F.$$

- The above theorem tells us that  $A_k$  is a good approximation for  $A$  (w.r.t. Frobenius norm).
- The approximation  $A_k$  also is good for computation of product with any vector  $\mathbf{x}$  with  $\|\mathbf{x}\| \leq 1$ .
  - Computing  $A\mathbf{x}$  would cost  $O(nd)$  multiplications.
  - However, computing  $A_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}^T$  only costs  $O(kd + nk)$  multiplications.
- Question: Is  $A_k$  best rank- $k$  approximation to  $A$  w.r.t. the computation  $A\mathbf{x}$  for an arbitrary  $\mathbf{x}$  with  $\|\mathbf{x}\| \leq 1$ ?
  - We want a rank- $k$  matrix  $B$  such that  $\max_{\|\mathbf{x}\| \leq 1} \|(A - B)\mathbf{x}\|$  is minimized.

# Singular Value Decomposition (SVD)

## Best rank- $k$ approximation

### Theorem

For any matrix  $B$  with rank at most  $k$ :

$$\|A - A_k\|_F \leq \|A - B\|_F.$$

- The above theorem tells us that  $A_k$  is a good approximation for  $A$  (w.r.t. Frobenius norm).
- The approximation  $A_k$  also is good for computation of product with any vector  $\mathbf{x}$  with  $\|\mathbf{x}\| \leq 1$ .
  - Computing  $A\mathbf{x}$  would cost  $O(nd)$  multiplications.
  - However, computing  $A_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}^T$  only costs  $O(kd + nk)$  multiplications.
- Question: Is  $A_k$  best rank- $k$  approximation to  $A$  w.r.t. the computation  $A\mathbf{x}$  for an arbitrary  $\mathbf{x}$  with  $\|\mathbf{x}\| \leq 1$ ?
  - We want a rank- $k$  matrix  $B$  such that  $\max_{\|\mathbf{x}\| \leq 1} \|(A - B)\mathbf{x}\|$  is minimized.

### Definition (Spectral norm)

The 2-norm or spectral norm of a matrix  $A$ , denoted by  $\|A\|_2$ , is defined as:  $\|A\|_2 = \max_{\|\mathbf{x}\| \leq 1} \|A\mathbf{x}\|$ .

# Singular Value Decomposition (SVD)

## Best rank- $k$ approximation

- The approximation  $A_k$  also is good for computation of product with any vector  $\mathbf{x}$  with  $\|\mathbf{x}\| \leq 1$ .
  - Computing  $A\mathbf{x}$  would cost  $O(nd)$  multiplications.
  - However, computing  $A_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}^T$  only costs  $O(kd + nk)$  multiplications.
- Question: Is  $A_k$  best rank- $k$  approximation to  $A$  w.r.t. the computation  $A\mathbf{x}$  for an arbitrary  $\mathbf{x}$  with  $\|\mathbf{x}\| \leq 1$ ?
  - We want a rank- $k$  matrix  $B$  such that  $\max_{\|\mathbf{x}\| \leq 1} \|(A - B)\mathbf{x}\|$  is minimized.

### Definition (Spectral norm)

The 2-norm or spectral norm of a matrix  $A$ , denoted by  $\|A\|_2$ , is defined as:  $\|A\|_2 = \max_{\|\mathbf{x}\| \leq 1} \|A\mathbf{x}\|$ .

- Claim:  $\|A\|_2 = \sigma_1$ .

# Singular Value Decomposition (SVD)

## Best rank- $k$ approximation

- The approximation  $A_k$  also is good for computation of product with any vector  $\mathbf{x}$  with  $\|\mathbf{x}\| \leq 1$ .
  - Computing  $A\mathbf{x}$  would cost  $O(nd)$  multiplications.
  - However, computing  $A_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T$  only costs  $O(kd + nk)$  multiplications.
- Question: Is  $A_k$  best rank- $k$  approximation to  $A$  w.r.t. the computation  $A\mathbf{x}$  for an arbitrary  $\mathbf{x}$  with  $\|\mathbf{x}\| \leq 1$ ?
  - We want a rank- $k$  matrix  $B$  such that  $\max_{\|\mathbf{x}\| \leq 1} \|(A - B)\mathbf{x}\|$  is minimized.

### Definition (Spectral norm)

The 2-norm or spectral norm of a matrix  $A$ , denoted by  $\|A\|_2$ , is defined as:  $\|A\|_2 = \max_{\|\mathbf{x}\| \leq 1} \|A\mathbf{x}\|$ .

- Claim:  $\|A\|_2 = \sigma_1$ .
- The question can now be rephrased as:  
*Is  $A_k$  the best rank- $k$  approximation to  $A$  w.r.t. the spectral norm?*

# Singular Value Decomposition (SVD)

## Best rank- $k$ approximation

### Definition (Spectral norm)

The 2-norm or spectral norm of a matrix  $A$ , denoted by  $\|A\|_2$ , is defined as:  $\|A\|_2 = \max_{\|x\| \leq 1} \|Ax\|$ .

- Question: Is  $A_k$  the best rank- $k$  approximation to  $A$  w.r.t. the spectral norm?

### Theorem

Let  $A$  be any  $n \times d$  matrix. For any matrix  $B$  of rank at most  $k$ :

$$\|A - A_k\|_2 \leq \|A - B\|_2.$$

- First, we show that the left singular vectors  $\mathbf{u}_1, \dots, \mathbf{u}_r$  are pairwise orthogonal.

# Singular Value Decomposition (SVD)

Best rank- $k$  approximation

## Theorem

*The left singular vectors  $\mathbf{u}_1, \dots, \mathbf{u}_r$  are pairwise orthogonal.*

# Singular Value Decomposition (SVD)

## Best rank- $k$ approximation

### Theorem

*Let  $A$  be any  $n \times d$  matrix. For any matrix  $B$  of rank at most  $k$ :*

$$\|A - A_k\|_2 \leq \|A - B\|_2.$$

- First, we show that the left singular vectors  $\mathbf{u}_1, \dots, \mathbf{u}_r$  are pairwise orthogonal.

### Theorem

*The left singular vectors  $\mathbf{u}_1, \dots, \mathbf{u}_r$  are pairwise orthogonal.*

- We will also need the following theorem.

### Theorem

$$\|A - A_k\|_2^2 = \sigma_{k+1}^2.$$

# Singular Value Decomposition (SVD)

## Best rank- $k$ approximation

### Theorem

*Let  $A$  be any  $n \times d$  matrix. For any matrix  $B$  of rank at most  $k$ :*

$$\|A - A_k\|_2 \leq \|A - B\|_2.$$

### Theorem

*The left singular vectors  $\mathbf{u}_1, \dots, \mathbf{u}_r$  are pairwise orthogonal.*

### Theorem

$$\|A - A_k\|_2^2 = \sigma_{k+1}^2.$$

- Finally, we show the following:

### Theorem

*Let  $A$  be an  $n \times d$  matrix. For any matrix  $B$  of rank at most  $k$ :*

$$\|A - A_k\|_2 \leq \|A - B\|_2.$$

# Singular Value Decomposition (SVD)

- Exercise: Show that  $u_i$ 's are the right singular vectors for the matrix  $A^T$ .

## Power Method for Singular Value Decomposition (SVD)

# Singular Value Decomposition (SVD)

## Power method for SVD

- Let  $B = A^T A$
- Question: Can you point out some interesting properties of  $B$ ?

# Singular Value Decomposition (SVD)

## Power method for SVD

- Let  $B = A^T A$
- $B = \sum_{i=1}^r \sigma_i^2 \mathbf{v}_i \mathbf{v}_i^T$

# Singular Value Decomposition (SVD)

## Power method for SVD

- Let  $B = A^T A$
- $B = \sum_{i=1}^r \sigma_i^2 \mathbf{v}_i \mathbf{v}_i^T$
- Question: Can we obtain a similar expression for  $B^2$  and in general  $B^k$ ?

# Singular Value Decomposition (SVD)

## Power method for SVD

- Let  $B = A^T A$
- $B = \sum_{i=1}^r \sigma_i^2 \mathbf{v}_i \mathbf{v}_i^T$
- Question: Can we obtain a similar expression for  $B^2$  and in general  $B^k$ ?
- $B^k = \sum_{i=1}^r \sigma_i^{2k} \mathbf{v}_i \mathbf{v}_i^T$
- So, if  $\sigma_1 > \sigma_2$ , then normalizing the first column of  $B^k$  should give a good estimate for  $\mathbf{v}_1$ .

# Singular Value Decomposition (SVD)

## Power method for SVD

- Let  $B = A^T A$
- $B = \sum_{i=1}^r \sigma_i^2 \mathbf{v}_i \mathbf{v}_i^T$
- Question: Can we obtain a similar expression for  $B^2$  and in general  $B^k$ ?
- $B^k = \sum_{i=1}^r \sigma_i^{2k} \mathbf{v}_i \mathbf{v}_i^T$
- So, if  $\sigma_1 > \sigma_2$ , then normalizing the first column of  $B^k$  should give a good estimate for  $\mathbf{v}_1$ .
- A faster method:
  - Computing  $B^k$  may be costly.
  - Select a random vector  $\mathbf{x} = \sum_{i=1}^d c_i \mathbf{v}_i$ .
  - Claim:  $B^k \mathbf{x} \approx \sigma_1^{2k} c_1 \mathbf{v}_1$
  - So, normalizing  $B^k \mathbf{x}$  approximates  $\mathbf{v}_1$ .

# Singular Value Decomposition (SVD)

## Power method for SVD

- A faster method:
  - Computing  $B^k$  may be costly.
  - Select a random vector  $\mathbf{x} = \sum_{i=1}^d c_i \mathbf{v}_i$ .
  - Claim:  $B^k \mathbf{x} \approx \sigma_1^{2k} c_1 \mathbf{v}_1$
  - So, normalizing  $B^k \mathbf{x}$  approximates  $\mathbf{v}_1$ .
- Given  $\min_{i < j} \log(\frac{\sigma_i}{\sigma_j}) \geq \lambda$ , the following algorithm estimates (within  $\varepsilon$  error with probability  $\geq (1 - \delta)$ ) the first singular value and singular vectors.

### Algorithm

1. Generate  $\mathbf{x}_0$  from a spherical gaussian with mean 0 and variance 1.
2.  $s \leftarrow \log\left(\frac{8d \log(2d/\delta)}{\varepsilon \delta}\right) / 2\lambda$
3. For  $i = 1$  to  $s$
4.  $\mathbf{x}_i \leftarrow (A^T A) \mathbf{x}_{i-1}$
5.  $\mathbf{v}_1 \leftarrow \mathbf{x}_i / \|\mathbf{x}_i\|$
6.  $\sigma_1 \leftarrow \|A \mathbf{v}_1\|$
7.  $\mathbf{u}_1 \leftarrow A \mathbf{v}_1 / \sigma_1$
8. return( $\sigma_1, \mathbf{u}_1, \mathbf{v}_1$ )

# Singular Value Decomposition (SVD)

## Power method for SVD

- Let  $B = A^T A$
- $B = \sum_{i=1}^r \sigma_i^2 \mathbf{v}_i \mathbf{v}_i^T$
- Question: Can we obtain a similar expression for  $B^2$  and in general  $B^k$ ?
- $B^k = \sum_{i=1}^r \sigma_i^{2k} \mathbf{v}_i \mathbf{v}_i^T$
- So, if  $\sigma_1 > \sigma_2$ , then normalizing the first column of  $B^k$  should give a good estimate for  $\mathbf{v}_1$ .
- A faster method:
  - Computing  $B^k$  may be costly.
  - Select a random vector  $\mathbf{x} = \sum_{i=1}^d c_i \mathbf{v}_i$ .
  - Claim:  $B^k \mathbf{x} \approx \sigma_1^{2k} c_1 \mathbf{v}_1$
  - So, normalizing  $B^k \mathbf{x}$  approximates  $\mathbf{v}_1$ .
- The above approximations are with respect to the fact that  $\sigma_1$  is significantly larger than  $\sigma_2$ . What if this is not true?

# Singular Value Decomposition (SVD)

## Power method for SVD

- Let  $B = A^T A$
- $B = \sum_{i=1}^r \sigma_i^2 \mathbf{v}_i \mathbf{v}_i^T$
- Question: Can we obtain a similar expression for  $B^2$  and in general  $B^k$ ?
- $B^k = \sum_{i=1}^r \sigma_i^{2k} \mathbf{v}_i \mathbf{v}_i^T$
- So, if  $\sigma_1 > \sigma_2$ , then normalizing the first column of  $B^k$  should give a good estimate for  $\mathbf{v}_1$ .
- A faster method:
  - Computing  $B^k$  may be costly.
  - Select a random vector  $\mathbf{x} = \sum_{i=1}^d c_i \mathbf{v}_i$ .
  - Claim:  $B^k \mathbf{x} \approx \sigma_1^{2k} c_1 \mathbf{v}_1$
  - So, normalizing  $B^k \mathbf{x}$  approximates  $\mathbf{v}_1$ .
- The above approximations are with respect to the fact that  $\sigma_1$  is significantly larger than  $\sigma_2$ . What if this is not true?

### Theorem

Let  $A$  be an  $n \times d$  matrix and  $\mathbf{x}$  a unit length vector in  $\mathbb{R}^d$  with  $\mathbf{x}_t^T \mathbf{v}_1 \geq \delta$ , where  $\delta > 0$ . Let  $V$  be the space spanned by the right singular vectors of  $A$  corresponding to singular values greater than  $(1 - \epsilon)\sigma_1$ . Let  $\mathbf{w}$  be the unit vector after  $k = \frac{\ln 1/\epsilon\delta}{2\epsilon}$  iterations of the power method, namely  $\mathbf{w} = \frac{(A^T A)^k \mathbf{x}}{\|(A^T A)^k \mathbf{x}\|}$ . Then  $\mathbf{w}$  has a component of at most  $\epsilon$  perpendicular to  $V$ .

## Singular vectors and Eigenvectors

# Singular Value Decomposition (SVD)

## Singular vectors and Eigenvectors

- For a square matrix  $B$ , if  $B\mathbf{x} = \lambda\mathbf{x}$ , then  $\mathbf{x}$  is said to be an *eigenvector* of  $B$  and  $\lambda$  the corresponding *eigenvalue*.
- For a rectangular  $n \times d$  matrix  $A$ , let  $B = A^T A$ .
- Claim 1: The right singular vectors of  $A$  are the eigenvectors of  $B$ .
- Claim 2: The left singular vectors of  $A$  are the eigenvectors of  $AA^T$ .

# Singular Value Decomposition (SVD)

## Singular vectors and Eigenvectors

- For a square matrix  $B$ , if  $B\mathbf{x} = \lambda\mathbf{x}$ , then  $\mathbf{x}$  is said to be an *eigenvector* of  $B$  and  $\lambda$  the corresponding *eigenvalue*.
- For a rectangular  $n \times d$  matrix  $A$ , let  $B = A^T A$ .
- Claim 1: The right singular vectors of  $A$  are the eigenvectors of  $B$ .
- Claim 2: The left singular vectors of  $A$  are the eigenvectors of  $AA^T$ .
- Claim 3: For every  $\mathbf{x}$ ,  $\mathbf{x}^T B \mathbf{x} \geq 0$ . Such matrices are called *positive semidefinite*.
- Fact: Any positive semidefinite matrix can be decomposed into a product  $A^T A$ .
- So, *eigen decomposition* of any positive semidefinite matrix may be obtained from SVD of  $A$ .

# Singular Value Decomposition (SVD)

## Applications of SVD

- The line **through the origin** that minimizes the sum of squared distances is defined by the first singular vector.
- Question: What if we drop the constraint that the line should pass through the origin? This general case is more relevant in data analysis.

End