COL758: Advanced Algorithms

Ragesh Jaiswal, CSE, IITD

Best Fit Subspaces and Singular Value Decomposition (SVD)

Problem

Given an $n \times d$ matrix A, where we interpret the rows of the matrix as points in $\mathbb{R}^d$, find a best fit line through the origin for the given n points.

- Question: How do we define best fit line?

Problem

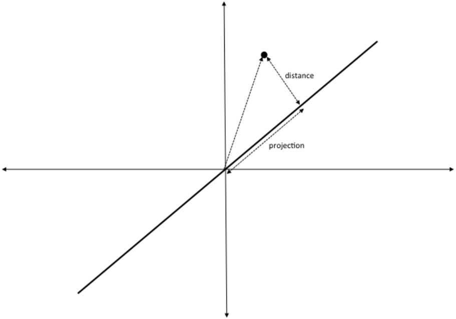
Given an $n \times d$ matrix A, where we interpret the rows of the matrix as points in $\mathbb{R}^d$, find a best fit line through the origin for the given n points.

- Question: How do we define best fit line?
  - A line that minimises the sum of squared distance of the n points to the line.

Problem

Given an $n \times d$ matrix A, where we interpret the rows of the matrix as points in $\mathbb{R}^d$, find a best fit line through the origin for the given n points.

- Question: How do we define best fit line?
  - A line that minimises the sum of squared distance of the n points to the line.
  - Claim: The best fit line maximises the sum of projections squared of the n points to the line.



Problem

Given an $n \times d$ matrix A, where we interpret the rows of the matrix as points in $\mathbb{R}^d$, find a best fit line through the origin for the given n points.

- The best fit line through the origin is one that minimises the sum of squared distance of the n points to the line.
- Let $\mathbf{v}$ denote a unit vector ($d \times 1$ matrix) in the direction of the best fit line.
- Claim: The sum of squared lengths of projections of the points onto $\mathbf{v}$ is $||A\mathbf{v}||^2$.

Problem

Given an $n \times d$ matrix A, where we interpret the rows of the matrix as points in $\mathbb{R}^d$, find a best fit line through the origin for the given n points.

- The best fit line through the origin is one that minimises the sum of squared distance of the n points to the line.
- Let $\mathbf{v}$ denote a unit vector ($d \times 1$ matrix) in the direction of the best fit line.
- Claim: The sum of squared lengths of projections of the points onto $\mathbf{v}$ is $||A\mathbf{v}||^2$.
- So, the best fit line is defined by unit vector $\mathbf{v}$ that maximises $||A\mathbf{v}||$.
- This is the first singular vector of the matrix A. So, the first singular vector is defined as:

$$\mathbf{v_1} = \arg \max_{||\mathbf{v}||=1} ||A\mathbf{v}||$$

Problem

Given an $n \times d$ matrix A, where we interpret the rows of the matrix as points in $\mathbb{R}^d$, find a best fit line through the origin for the given n points.

- The best fit line through the origin is one that minimises the sum of squared distance of the n points to the line.
- Let $\mathbf{v}$ denote a unit vector ($d \times 1$ matrix) in the direction of the best fit line.
- Claim: The sum of squared lengths of projections of the points onto $\mathbf{v}$ is $||A\mathbf{v}||^2$.
- So, the best fit line is defined by unit vector $\mathbf{v}$ that maximises $||A\mathbf{v}||$.
- This is the first singular vector of the matrix A. So, the first singular vector is defined as:

$$\mathbf{v_1} = \arg \max_{||\mathbf{v}||=1} ||A\mathbf{v}||$$

- The value $\sigma_1 = ||A\mathbf{v_1}||$ is called the first singular value of A.

Problem

Given an $n \times d$ matrix A, where we interpret the rows of the matrix as points in $\mathbb{R}^d$, find a best fit line through the origin for the given n points.

- The first singular vector is defined as:

$$\mathbf{v_1} = \arg \max_{||\mathbf{v}||=1} ||A\mathbf{v}||$$

- The value $\sigma_1 = ||A\mathbf{v_1}||$ is called the first singular value of A.
- So, σ_1^2 is equal to the sum of squared length of projections.
- Note that if all the data points are "close" to a line through the origin, then the first singular vector gives such a line.
- Question: if the data points are close to a plane (and in general close to a k-dimensional subspace), then how do we find such a plane?

Problem

Given an $n \times d$ matrix A, where we interpret the rows of the matrix as points in $\mathbb{R}^d$, find a best fit plane through the origin for the given n points.

- Let $\mathbf{v_1}$ denote the first singular vector of A.
- Idea: Find a unit vector $\mathbf{v}$ perpendicular to $\mathbf{v_1}$ that maximises $||A\mathbf{v}||$. Output the plane through the origin defined by vectors $\mathbf{v_1}$ and $\mathbf{v}$.
- Claim: The plane defined above indeed maximises sum of squared distances of all the points.
- The second singular vector is defined as:

$$\mathbf{v_2} = \arg\max_{||\mathbf{v}||=1, \mathbf{v} \perp \mathbf{v_1}} ||A\mathbf{v}||.$$

- The value $\sigma_2 = ||A\mathbf{v_2}||$ is called the second singular value of A.

Problem

Given an $n \times d$ matrix A, where we interpret the rows of the matrix as points in $\mathbb{R}^d$, find a best fit plane through the origin for the given n points.

- Let $\mathbf{v_1}$ denote the first singular vector of A.
- The second singular vector is defined as:

$$\mathbf{v_2} = \arg\max_{||\mathbf{v}||=1, \mathbf{v} \perp \mathbf{v_1}} ||A\mathbf{v}||.$$

- The value $\sigma_2 = ||A\mathbf{v_2}||$ is called the second singular value of A.

Theorem

For any matrix A, the plane spanned by $\mathbf{v_1}$ and $\mathbf{v_2}$ is the best fit plane.

- The first singular vector is defined as: $\mathbf{v_1} = \arg\max_{||\mathbf{v}||=1} ||A\mathbf{v}||$.
- The second singular vector is defined as: $\mathbf{v_2} = \arg\max_{||\mathbf{v}||=1, \mathbf{v} \perp \mathbf{v_1}} ||A\mathbf{v}||$.

Theorem

For any matrix A, the plane spanned by $\mathbf{v_1}$ and $\mathbf{v_2}$ is the best fit plane.

Proof sketch

- Let W denote the best fit plane for A.
- <u>Claim 1</u>: There exists an orthonormal basis $(\mathbf{w_1}, \mathbf{w_2})$ of W such that $\mathbf{w_2}$ is perpendicular to $\mathbf{v_1}$.
- <u>Claim 2</u>: $||A\mathbf{w_1}||^2 \leq ||A\mathbf{v_1}||^2$.
- <u>Claim 3</u>: $||A\mathbf{w_2}||^2 \leq ||A\mathbf{v_2}||^2$.
- This gives $||A\mathbf{w_1}||^2 + ||A\mathbf{w_2}||^2 \leq ||A\mathbf{v_1}||^2 + ||A\mathbf{v_2}||^2$. □

- The first singular vector and first singular value is defined as:

$$\mathbf{v_1} = \arg\max_{||\mathbf{v}||=1} ||A\mathbf{v}|| \quad \text{and} \quad \sigma_1 = ||A\mathbf{v_1}||$$

- The second singular vector and second singular value is defined as:

$$\mathbf{v_2} = \arg\max_{||\mathbf{v}||=1, \mathbf{v} \perp \mathbf{v_1}} ||A\mathbf{v}|| \quad \text{and} \quad \sigma_2 = ||A\mathbf{v_2}||.$$

- The third singular vector and third singular value is defined as:

$$\mathbf{v_3} = \arg\max_{||\mathbf{v}||=1, \mathbf{v} \perp \mathbf{v_1}, \mathbf{v_2}} ||A\mathbf{v}|| \quad \text{and} \quad \sigma_3 = ||A\mathbf{v_3}||.$$

- ...and so on.
- Let r be the smallest positive integer such that: $\max_{||\mathbf{v}||=1, \mathbf{v} \perp \mathbf{v_1}, ..., \mathbf{v_r}} ||A\mathbf{v}|| = 0$. Then A has r singular vectors $\mathbf{v_1}, ..., \mathbf{v_r}$.

Theorem

Let A be any $n \times d$ matrix with r singular vectors $\mathbf{v_1}, ..., \mathbf{v_r}$. For $1 \leq k \leq r$, let V_k be the subspace spanned by $\mathbf{v_1}, ..., \mathbf{v_k}$. For each k, V_k is the best-fit k-dimensional subspace for A.

- The first singular vector and first singular value is defined as:

$$\mathbf{v_1} = \arg\max_{||\mathbf{v}||=1} ||A\mathbf{v}|| \quad \text{and} \quad \sigma_1 = ||A\mathbf{v_1}||$$

- The second singular vector and second singular value is defined as:

$$\mathbf{v_2} = \arg\max_{||\mathbf{v}||=1, \mathbf{v} \perp \mathbf{v_1}} ||A\mathbf{v}|| \quad \text{and} \quad \sigma_2 = ||A\mathbf{v_2}||.$$

- The third singular vector and third singular value is defined as:

$$\mathbf{v_3} = \arg\max_{||\mathbf{v}||=1, \mathbf{v} \perp \mathbf{v_1}, \mathbf{v_2}} ||A\mathbf{v}|| \quad \text{and} \quad \sigma_3 = ||A\mathbf{v_3}||.$$

- ...and so on.
- Let r be the smallest positive integer such that: $\max_{||\mathbf{v}||=1, \mathbf{v} \perp \mathbf{v_1}, ..., \mathbf{v_r}} ||A\mathbf{v}|| = 0$. Then A has r singular vectors $\mathbf{v_1}, ..., \mathbf{v_r}$.
- The vectors $\mathbf{v_1}, ..., \mathbf{v_r}$ are more specifically called the right singular vectors.

- The first singular vector and first singular value is defined as:

$$\mathbf{v_1} = \arg\max_{||\mathbf{v}||=1} ||A\mathbf{v}|| \quad \text{and} \quad \sigma_1 = ||A\mathbf{v_1}||$$

- The second singular vector and second singular value is defined as:

$$\mathbf{v_2} = \arg\max_{||\mathbf{v}||=1, \mathbf{v} \perp \mathbf{v_1}} ||A\mathbf{v}|| \quad \text{and} \quad \sigma_2 = ||A\mathbf{v_2}||.$$

- The third singular vector and third singular value is defined as:

$$\mathbf{v_3} = \arg\max_{||\mathbf{v}||=1, \mathbf{v} \perp \mathbf{v_1}, \mathbf{v_2}} ||A\mathbf{v}|| \quad \text{and} \quad \sigma_3 = ||A\mathbf{v_3}||.$$

- ...and so on.
- Let r be the smallest positive integer such that: $\max_{||\mathbf{v}||=1, \mathbf{v} \perp \mathbf{v_1}, ..., \mathbf{v_r}} ||A\mathbf{v}|| = 0$. Then A has r singular vectors $\mathbf{v_1}, ..., \mathbf{v_r}$.
- The vectors $\mathbf{v_1}, ..., \mathbf{v_r}$ are more specifically called the right singular vectors.
- For any singular vector $\mathbf{v_i}$, $\sigma_i = ||A\mathbf{v_i}||$ may be interpreted as the component of the matrix A along $\mathbf{v_i}$.
- Given this interpretation, the "*the components should add up to give the whole content of A*".

- Let r be the smallest positive integer such that: $\max_{||\mathbf{v}||=1, \mathbf{v} \perp \mathbf{v_1}, ..., \mathbf{v_r}} ||A\mathbf{v}|| = 0$. Then A has r singular vectors $\mathbf{v_1}, ..., \mathbf{v_r}$.
- The vectors $\mathbf{v_1}, ..., \mathbf{v_r}$ are more specifically called the right singular vectors.
- For any singular vector $\mathbf{v_i}$, $\sigma_i = ||A\mathbf{v_i}||$ may be interpreted as the component of the matrix A along $\mathbf{v_i}$.
- Given this interpretation, the "*the components should add up to give the whole content of A*".
- For any row a_j in the matrix A, we can write $||a_j||^2 = \sum_{i=1}^{r} (a_j \cdot \mathbf{v_i})^2$. This further gives:

$$\sum_{j=1}^{n} ||a_j||^2 = \sum_{j=1}^{n} \sum_{i=1}^{r} (a_j \cdot \mathbf{v_i})^2 = \sum_{i=1}^{r} ||A\mathbf{v_i}||^2 = \sum_{i=1}^{r} \sigma_i^2.$$

- Let r be the smallest positive integer such that: $\max_{||\mathbf{v}||=1, \mathbf{v} \perp \mathbf{v_1}, ..., \mathbf{v_r}} ||A\mathbf{v}|| = 0$. Then A has r singular vectors $\mathbf{v_1}, ..., \mathbf{v_r}$.
- The vectors $\mathbf{v_1}, ..., \mathbf{v_r}$ are more specifically called the right singular vectors.
- For any singular vector $\mathbf{v_i}$, $\sigma_i = ||A\mathbf{v_i}||$ may be interpreted as the component of the matrix A along $\mathbf{v_i}$.
- Given this interpretation, the "*the components should add up to give the whole content of A*".
- For any row a_j in the matrix A, we can write $||a_j||^2 = \sum_{i=1}^{r} (a_j \cdot \mathbf{v_i})^2$. This further gives:

$$\sum_{j=1}^{n} ||a_j||^2 = \sum_{j=1}^{n} \sum_{i=1}^{r} (a_j \cdot \mathbf{v_i})^2 = \sum_{i=1}^{r} ||A\mathbf{v_i}||^2 = \sum_{i=1}^{r} \sigma_i^2.$$

- The LHS of the above equation may be interpreted as "*content of the matrix*" defines the Frobenius Norm of the matrix A.

Definition (Frobenius Norm)

The Frobenius norm of a given $n \times d$ matrix A, denoted by $||A||_F$, is defined as: $||A||_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{d} A_{i,j}^2}$.

- For any row a_j in the matrix A, we can write $||a_j||^2 = \sum_{i=1}^{r}(a_j \cdot \mathbf{v_i})^2$. This further gives:

$$\sum_{j=1}^{n} ||a_j||^2 = \sum_{j=1}^{n}\sum_{i=1}^{r}(a_j \cdot \mathbf{v_i})^2 = \sum_{i=1}^{r} ||A\mathbf{v_i}||^2 = \sum_{i=1}^{r} \sigma_i^2.$$

- The LHS of the above equation may be interpreted as "*content of the matrix*" defines the Frobenius Norm of the matrix A.

Definition (Frobenius Norm)

The Frobenius norm of a given $n \times d$ matrix A, denoted by $||A||_F$, is defined as: $||A||_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d} A_{i,j}^2}$.

Theorem

For any matrix A, the sum of squares of the right singular values equals the square of the Frobenius norm of the matrix.

- Let $\mathbf{v}_1, ..., \mathbf{v}_r$ be the right singular vectors and $\sigma_1, ..., \sigma_r$ be the corresponding singular values of matrix A.
- The left singular vectors are defined as $\mathbf{u}_i = \frac{1}{\sigma_i} A\mathbf{v}_i$.
- $\sigma_i \mathbf{u}_i$ may be interpreted as a vector whose components are the projections of the rows of A onto $\mathbf{v}_i$.

- Let $\mathbf{v}_1, ..., \mathbf{v}_r$ be the right singular vectors and $\sigma_1, ..., \sigma_r$ be the corresponding eigenvalues of matrix A.
- The left singular vectors are defined as $\mathbf{u}_i = \frac{1}{\sigma_i} A\mathbf{v}_i$.
- $\sigma_i \mathbf{u}_i$ may be interpreted as a vector whose components are the projections of the rows of A onto $\mathbf{v}_i$.
- $\sigma_i \mathbf{u}_i \mathbf{v}_i^T$ is a rank one matrix whose rows can be interpreted as component of rows of A along $\mathbf{v}_i$.
- Given this, the following decomposition of A into rank one matrices should make sense (we will prove this): $A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$.

Theorem

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

Singular Value Decomposition (SVD)

Theorem

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

Proof sketch

- Lemma: Matrices A and B are identical iff for all vectors $\mathbf{v}$, $A\mathbf{v} = B\mathbf{v}$.
- Let $B = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$.
- For any j, $A\mathbf{v}_j = \sigma_j \mathbf{u}_j$ from the definition of u_j.
- $B\mathbf{v}_j = \left(\sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T\right) \mathbf{v}_j = \sigma_j \mathbf{u}_j$ from orthonormality.
- Fact: Any vector $\mathbf{v}$ can be written as a linear combination of right eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_r$ and a vector perpendicular to $\mathbf{v}_1, ..., \mathbf{v}_r$. □

Singular Value Decomposition (SVD)

Theorem

Let A be any $n \times d$ matrix with right singular vectors $\mathbf{v}_1, ..., \mathbf{v}_r$, left-singular vectors $\mathbf{u}_1, ..., \mathbf{u}_r$, and corresponding singular values $\sigma_1, ..., \sigma_r$. 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^{th} column is $\mathbf{u}_i$.
  - D is a $r \times r$ diagonal matrix with the i^{th} diagonal element σ_i.
  - V is a $d \times r$ matrix where the i^{th} column is $\mathbf{v}_i$.
- Question: How do we compute the SVD?
- Question: What are the applications of SVD?

- 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 $k \in \{1, ..., r\}$ 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:

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

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.

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: $||A||_2 = \max_{||\mathbf{x}|| \leq 1} ||A\mathbf{x}||$.

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

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

Definition (Spectral norm)

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

- 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 $\mathbf{u}_1, ..., \mathbf{u}_r$ are pairwise orthogonal.

Theorem

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

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 $\mathbf{u}_1, ..., \mathbf{u}_r$ are pairwise orthogonal.

Theorem

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

- We will also need the following theorem.

Theorem

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

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 $\mathbf{u}_1, ..., \mathbf{u}_r$ 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:

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

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

Power Method for Singular Value Decomposition (SVD)

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

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

- 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?

- 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 $\mathbf{v}_1$.

- 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 $\mathbf{v}_1$.

- 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 $\min_{i<j} \log\left(\frac{\sigma_i}{\sigma_j}\right) \geq \lambda$, the following algorithm estimates (within ε 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. $\quad \mathbf{x}_i \leftarrow (A^T A)\mathbf{x}_{i-1}$
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$(\sigma_1, \mathbf{u}_1, \mathbf{v}_1)$

- 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 σ_1 is significantly larger than σ_2. What if this is not true?

- 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 σ_1 is significantly larger than σ_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 $\mathbf{x}_t \mathbf{v}_1 \geq \delta$, 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 ϵ perpendicular to V.

End