COL866: Foundations of Data Science

Ragesh Jaiswal, IITD

Theorem (Johnson-Lindenstrauss (JL) Theorem)

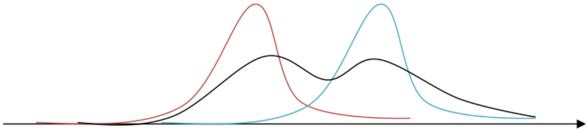
For any $0 < \varepsilon < 1$ and any integer n, let $k \geq \frac{3}{c\varepsilon^2} \ln n$ with c as in the Random Projection Theorem. For any set of n points in $\mathbb{R}^d$, the random projection $f : \mathbb{R}^d \to \mathbb{R}^k$ defined as before has the property that for all pairs of points $\mathbf{v_i}$ and $\mathbf{v_j}$, with probability at least $(1 - \frac{3}{2n})$,

$$(1 - \varepsilon)\sqrt{k}||\mathbf{v_i} - \mathbf{v_j}|| \leq ||f(\mathbf{v_i}) - f(\mathbf{v_j})|| \leq (1 + \varepsilon)\sqrt{k}||\mathbf{v_i} - \mathbf{v_j}||.$$

- Here is an application of the JL Theorem for the Nearest Neighbour (NN) problem:
  - Suppose we need to pre-process n data points $X \subseteq \mathbb{R}^d$ so that we can answer at most n' queries of the form: "find the point from X that is nearest to a given point $p \in \mathbb{R}^d$".
  - If we use a JL mapping with $k \geq \frac{3}{c\varepsilon^2} \ln(n + n')$, then we can store $f(\mathbf{x})$ for all $\mathbf{x} \in X$. For a query point $\mathbf{p}$, we just return the the point that is nearest to $f(\mathbf{p})$.

Separating Gaussians

- Mixture of Gaussians are used to model heterogenous data coming from multiple sources.
- Consider an example of height of people in a city:
  - Let $p_M(x)$ denote the Gaussian density of height of men in the city and $p_F(x)$ for women.
  - Let w_M and w_F denote the proportion of men and women in the city respectively.
  - So, the mixture model $p(x) = w_M \cdot p_M(x) + w_F \cdot p_F(x)$ is a natural way to model the density of hight of people in the city.

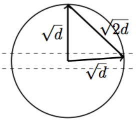


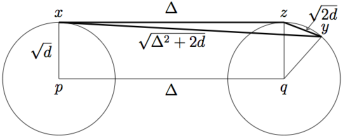
- Mixture of Gaussians are used to model heterogenous data coming from multiple sources.
- Consider an example of height of people in a city:
  - Let $p_M(x)$ denote the Gaussian density of height of men in the city and $p_F(x)$ for women.
  - Let w_M and w_F denote the proportion of men and women in the city respectively.
  - So, the mixture model $p(x) = w_M \cdot p_M(x) + w_F \cdot p_F(x)$ is a natural way to model the density of hight of people in the city.
- The parameter estimation problem is to guess the parameters of the mixture given samples from the mixture.
  - In our above example this means that we are given heights of a number of people of the city and the task is to infer w_M, w_F and the mean and variance of $p_M(x)$ and $p_F(x)$.

- Mixture of Gaussians are used to model heterogenous data coming from multiple sources.
- Consider an example of height of people in a city:
  - Let $p_M(x)$ denote the Gaussian density of height of men in the city and $p_F(x)$ for women.
  - Let w_M and w_F denote the proportion of men and women in the city respectively.
  - So, the mixture model $p(x) = w_M \cdot p_M(x) + w_F \cdot p_F(x)$ is a natural way to model the density of hight of people in the city.
- The parameter estimation problem is to guess the parameters of the mixture given samples from the mixture.
  - In our above example this means that we are given heights of a number of people of the city and the task is to infer w_M, w_F and the mean and variance of $p_M(x)$ and $p_F(x)$.
  - In the example, given the height of an individual can we infer whether it is a man or a woman?

- We will first consider the following simpler problem of separating unit variance Gaussians:
  - Given samples from a mixture of two spherical Gaussians with unit variance in $\mathbb{R}^d$, separate the samples.
- If the means of the Gaussians are too close, then it will be hard to distinguish samples from the distributions. Suppose the distance between the means is Δ.
- We will try to design an algorithm that estimates the parameters for some minimum value on Δ.

- We will first consider the following simpler problem of separating unit variance Gaussians:
  - Given samples from a mixture of two spherical Gaussians with unit variance in $\mathbb{R}^d$, separate the samples.
- If the means of the Gaussians are too close, then it will be hard to distinguish samples from the distributions. Suppose the distance between the means is Δ.
- We will try to design an algorithm that estimates the parameters for some minimum value on Δ.
- Claim 1: Let $\mathbf{x}$ and $\mathbf{y}$ be two random points sampled from the same Gaussian. Then $||\mathbf{x} - \mathbf{y}|| = \sqrt{2d} \pm O(1)$ w.h.p.
- Claim 2: Let $\mathbf{x}$ and $\mathbf{y}$ be two random points sampled from different Gaussians. Then $||\mathbf{x} - \mathbf{y}|| = \sqrt{2d + \Delta^2} \pm O(1)$ w.h.p.





- We will first consider the following simpler problem of separating unit variance Gaussians:
  - Given samples from a mixture of two spherical Gaussians with unit variance in $\mathbb{R}^d$, separate the samples.
- If the means of the Gaussians are too close, then it will be hard to distinguish samples from the distributions. Suppose the distance between the means is Δ.
- We will try to design an algorithm that estimates the parameters for some minimum value on Δ.
- Claim 1: Let $\mathbf{x}$ and $\mathbf{y}$ be two random points sampled from the same Gaussian. Then $||\mathbf{x} - \mathbf{y}|| = \sqrt{2d} \pm O(1)$ w.h.p.
- Claim 2: Let $\mathbf{x}$ and $\mathbf{y}$ be two random points sampled from different Gaussians. Then $||\mathbf{x} - \mathbf{y}|| = \sqrt{2d + \Delta^2} \pm O(1)$ w.h.p.
- So, we can distinguish points from the same/different Gaussians based on the pairwise distance as long as $\sqrt{2d} + O(1) \leq \sqrt{2d + \Delta^2} - O(1)$ which implies that $\Delta = \omega(d^{1/4})$.
  - Since we want this for almost all point pairs there is an extra factor of $O(\sqrt{\log n})$ in Δ.

- We will first consider the following simpler problem of separating unit variance Gaussians:
  - Given n samples from a mixture of two spherical Gaussians with unit variance in $\mathbb{R}^d$, separate the samples.
- Let the distance between the means be $\Delta = \Omega(d^{1/4}\sqrt{\log n})$.
- Here is an algorithm for separating points from the two Gaussians.

Algorithm

- Calculate pairwise distance between all pairs of points
- The cluster of smallest pairwise distances must come from the same Gaussian. Remove these points.
- The remaining points come from the second Gaussian.

- We will first consider the following simpler problem of separating unit variance Gaussians:
  - Given n samples from a mixture of two spherical Gaussians with unit variance in $\mathbb{R}^d$, separate the samples.
- The parameter estimation problem was to estimate the parameters of the Gaussian that the data points are sampled.
- Since, we now have an algorithm for separating points, we should think of how to fit a spherical Gaussian to the given data.

- Given samples $\mathbf{x}_1, ..., \mathbf{x}_n$ in a d-dimensional space, we want to find the spherical Gaussian that best fits the points.
- Let f be an unknown Gaussian with mean μ and variance σ^2 in each direction.
- The probability density of picking these points from this Gaussian is given by $c \cdot exp\left(-\frac{||\mathbf{x}_1-\mu||^2+...+||\mathbf{x}_n-\mu||^2}{2\sigma^2}\right)$.
- The Maximum Likelihood Estimator (MLE) of f, given the samples $\mathbf{x}_1, ..., \mathbf{x}_n$ is the f that maximizes the above probability density.

Theorem

The maximum likelihood spherical Gaussian for a set of samples is the Gaussian with the center equal to the sample mean and standard deviation equal to the standard deviation of the sample from the true mean.

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

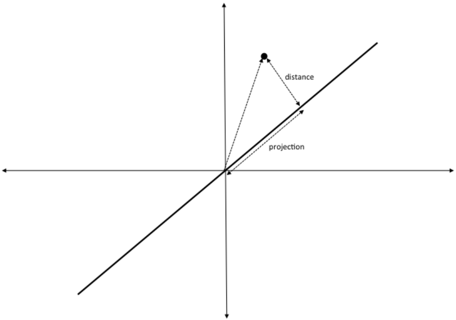
Given an $n \times d$ matrix A, where we interpret the rows of the matrix as points in $\mathbb{R}^d$, find a best fit line through the origin for the given n points.

- 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.
- Claim 1: There exists an orthonormal basis $(\mathbf{w_1}, \mathbf{w_2})$ of W such that $\mathbf{w_2}$ is perpendicular to $\mathbf{v_1}$.
- Claim 2: $||A\mathbf{w_1}||^2 \leq ||A\mathbf{v_1}||^2$.
- Claim 3: $||A\mathbf{w_2}||^2 \leq ||A\mathbf{v_2}||^2$.
- This gives $||A\mathbf{w_1}||^2 + ||A\mathbf{w_2}||^2 \leq ||A\mathbf{v_1}||^2 + ||A\mathbf{v_2}||^2$. $\square$

- 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 $\mathbf{v}_1, ..., \mathbf{v}_r$. $\square$

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?

End