Data Science Theory and Practices

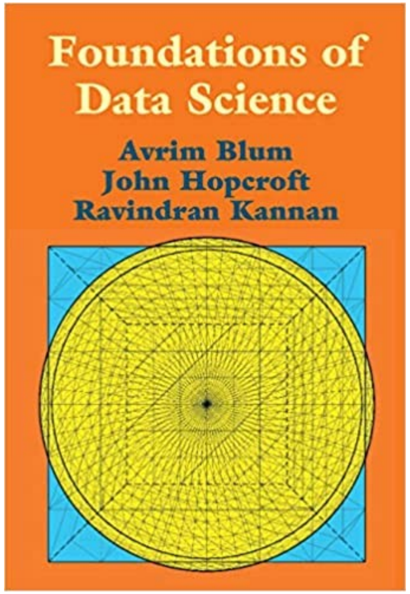
Foundations on Principal Component Analysis

Joseph Chuang-Chieh Lin
Dept. CSIE, Tamkang University

19th April, 2021

Outline

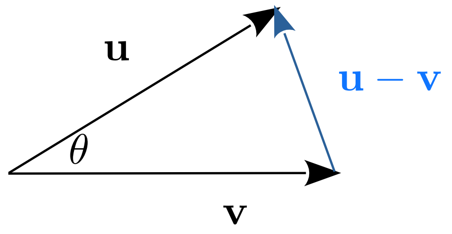
- Matrices & Overview
- Projection
- Singular Vectors
- Singular Value Decomposition (SVD)
- Best Rank-*k* Approximation
- Principal Component Analysis (PCA)



Refer to https://tinyurl.com/3rnuvb7e

Inner Products and Projection

$$||\mathbf{u}|| = \langle \mathbf{u}, \mathbf{u} \rangle^{\frac{1}{2}}$$



Low of Cosines:

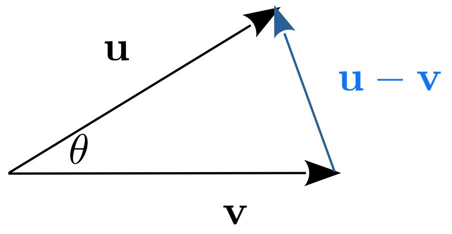
$$||\mathbf{u} - \mathbf{v}||^2 = ||\mathbf{u}||^2 + ||\mathbf{v}||^2 - 2||\mathbf{u}||\,||\mathbf{v}|| \cos\theta$$

And,

$$\langle \mathbf{u} - \mathbf{v},\ \mathbf{u} - \mathbf{v} \rangle = ||\mathbf{u}||^2 + ||\mathbf{v}||^2 - 2\langle \mathbf{u}, \mathbf{v} \rangle.$$

Inner Products and Projection

$$||\mathbf{u}|| = \langle \mathbf{u}, \mathbf{u} \rangle^{\frac{1}{2}}$$



Low of Cosines:

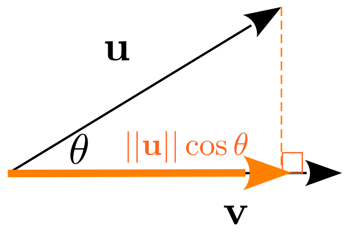
$$||\mathbf{u} - \mathbf{v}||^2 = ||\mathbf{u}||^2 + ||\mathbf{v}||^2 - 2||\mathbf{u}||\,||\mathbf{v}|| \cos\theta$$

And,

$$\langle \mathbf{u} - \mathbf{v},\ \mathbf{u} - \mathbf{v} \rangle = ||\mathbf{u}||^2 + ||\mathbf{v}||^2 - 2\langle \mathbf{u}, \mathbf{v} \rangle.$$

$$\therefore \langle \mathbf{u}, \mathbf{v} \rangle = ||\mathbf{u}|| \cdot ||\mathbf{v}|| \cos\theta$$

Inner Products and Projection



$$||\mathbf{u}|| = \langle \mathbf{u}, \mathbf{u} \rangle^{\frac{1}{2}}$$

Low of Cosines:

$$||\mathbf{u} - \mathbf{v}||^2 = ||\mathbf{u}||^2 + ||\mathbf{v}||^2 - 2||\mathbf{u}|| \, ||\mathbf{v}|| \cos\theta$$

And,

$$\langle \mathbf{u} - \mathbf{v}, \, \mathbf{u} - \mathbf{v} \rangle = ||\mathbf{u}||^2 + ||\mathbf{v}||^2 - 2\langle \mathbf{u}, \mathbf{v} \rangle.$$

$$\therefore \langle \mathbf{u}, \mathbf{v} \rangle = ||\mathbf{u}|| \cdot ||\mathbf{v}|| \cos\theta$$

$$\cos\theta = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{||\mathbf{u}|| \, ||\mathbf{v}||}, \quad ||\mathbf{u}|| \cos\theta = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{||\mathbf{v}||} = \frac{\mathbf{u}^\top \mathbf{v}}{||\mathbf{v}||}.$$

Matrices

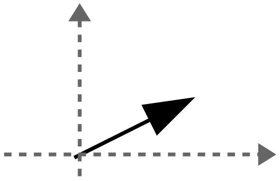
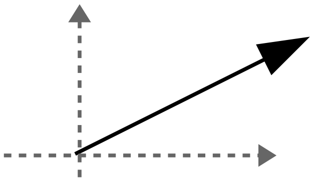
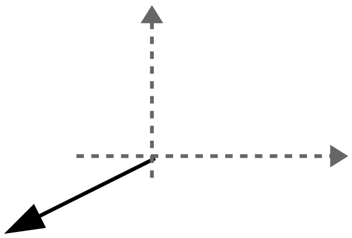
- Eigenvalues and eigenvectors.

$$A\mathbf{v} = \lambda\mathbf{v}$$

Matrices

- Eigenvalues and eigenvectors.

$$A\mathbf{v} = \lambda\mathbf{v}$$

eigenvector   or 

other vectors → rotations

Data: *d*-dimensional *n* points

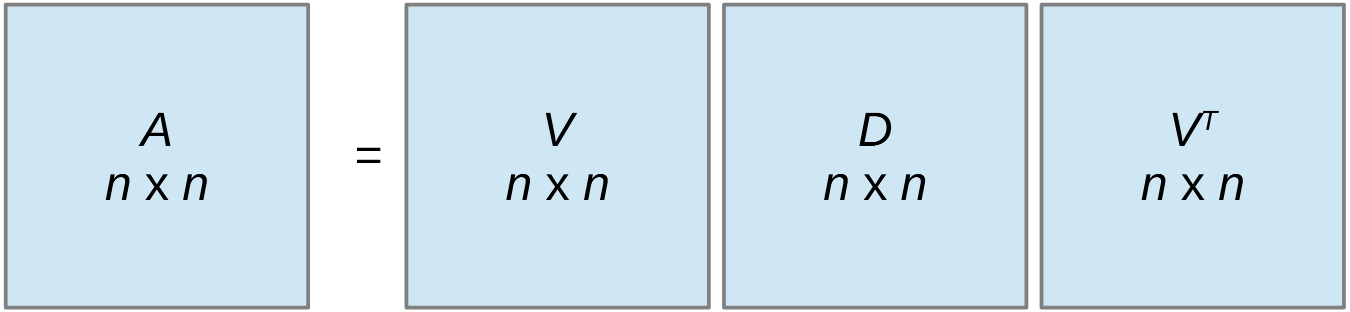
	feature 1	feature 2	......	feature *d*-1	feature *d*
record 1	$a_{1,1}$	$a_{1,2}$	......	$a_{1,d\text{-}1}$	$a_{1,d\text{-}1}$
record 2	$a_{2,1}$	$a_{2,2}$	......	$a_{2,d\text{-}1}$	$a_{2,d\text{-}1}$
⋮	⋮	⋮	......	⋮	⋮
record *n*-1	$a_{n\text{-}1,\ 1}$	$a_{n\text{-}1,2}$	......	$a_{n\text{-}1,d\text{-}1}$	$a_{n\text{-}1,d\text{-}1}$
record *n*	$a_{n,\ 1}$	$a_{n,2}$	......	$a_{n,d\text{-}1}$	$a_{n,d\text{-}1}$

Eigenvalue Decomposition (Overview)

- A: square matrix
- If A is symmetric,

$$A = VDV^\top$$

D: diagonal



Singular Value Decomposition (Overview)

A
n x *d*

=

U
n x *r*

D
r x *r*

V$^{\mathrm{T}}$
r x *d*

$\mathbf{u}_i : \quad n \times 1 \text{ matrix}$

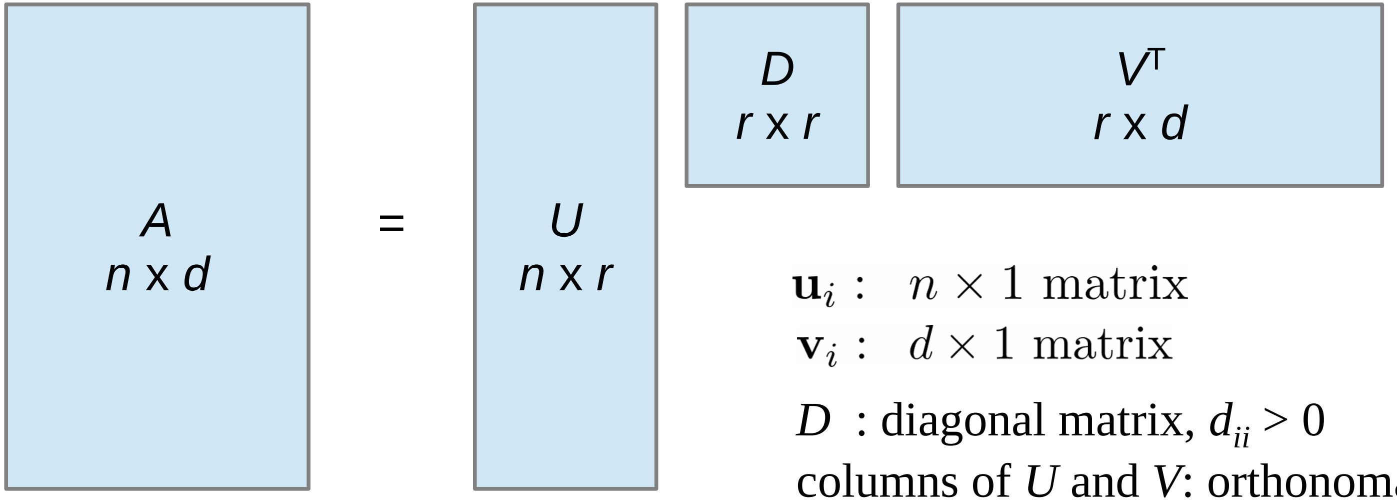
$\mathbf{v}_i : \quad d \times 1 \text{ matrix}$

D : diagonal matrix, $d_{ii} > 0$

columns of U and V: orthonomal

$$A = UDV^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.$$

Singular Value Decomposition (Overview)



$\mathbf{u}_i : \quad n \times 1$ matrix

$\mathbf{v}_i : \quad d \times 1$ matrix

D : diagonal matrix, $d_{ii} > 0$

columns of U and V: orthonomal

$$A = UDV^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top .$$

$$A\mathbf{v}_i = d_{ii}\mathbf{u}_i \quad \text{and} \quad A^\top \mathbf{u}_i = d_{ii}\mathbf{v}_i .$$

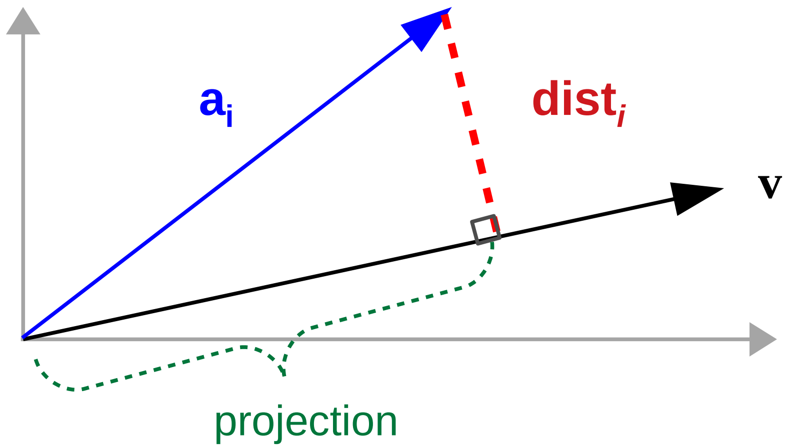
$$A^\top A\mathbf{v}_i = d_{ii}^2 \mathbf{v}_i .$$

Projection

Project a point (vector)

$$\mathbf{a_i} = (a_{i_1}, a_{i_2}, \ldots, a_{i_d})$$

onto a line (vector) $\mathbf{v}$.



Pythagorean Theorem:

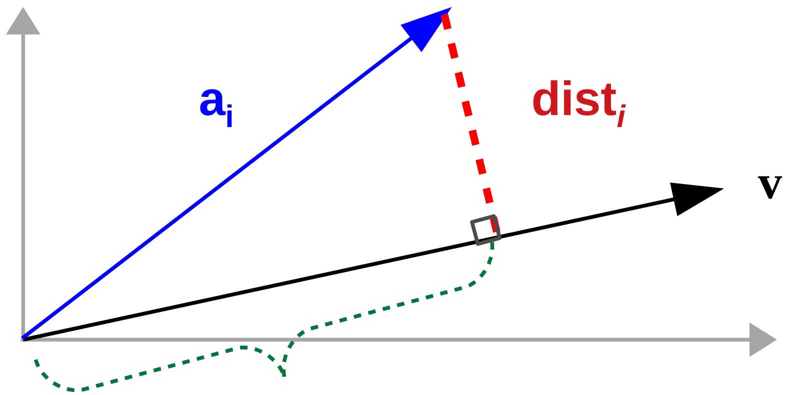
$$\begin{aligned} a_{i_1}^2 + a_{i_2}^2 + \cdots + a_{i_d}^2 &= (\text{length of projection})^2 \\ &+ (\text{distance of point to line})^2. \end{aligned}$$

Projection

Project a point (vector)

$$\mathbf{a_i} = (a_{i_1}, a_{i_2}, \ldots, a_{i_d})$$

onto a line (vector) $\mathbf{v}$.



Pythagorean Theorem:

$$\begin{aligned} a_{i_1}^2 + a_{i_2}^2 + \cdots + a_{i_d}^2 &= (\text{length of projection})^2 \\ &+ \boxed{(\text{distance of point to line})^2}. \end{aligned}$$

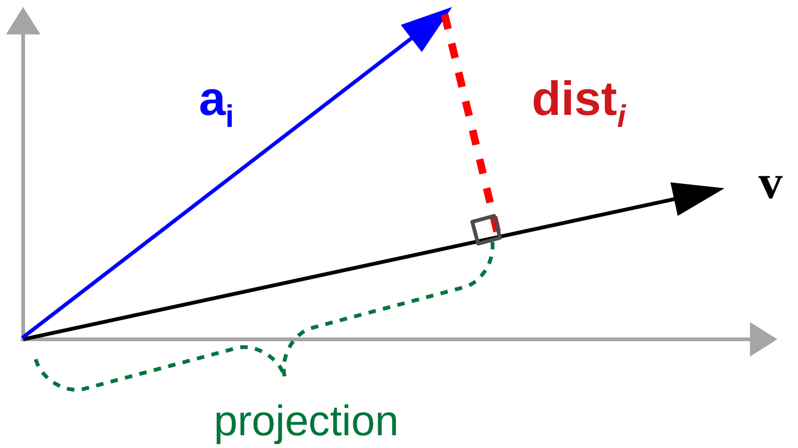
← Minimize this!

Projection

Project a point (vector)

$$\mathbf{a_i} = (a_{i_1}, a_{i_2}, \ldots, a_{i_d})$$

onto a line (vector) $\mathbf{v}$.



Pythagorean Theorem:

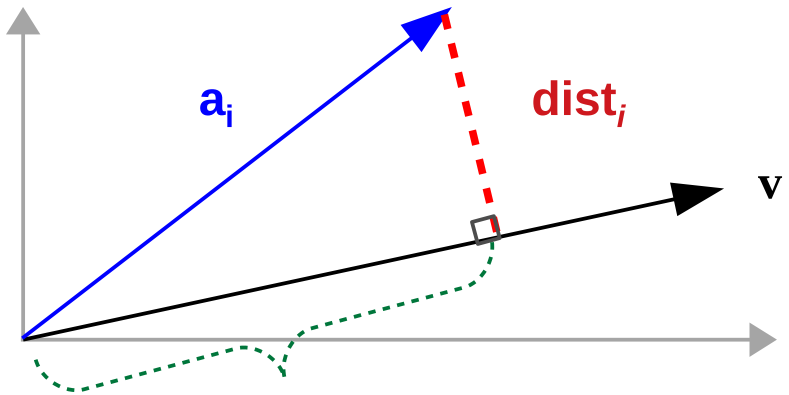
$$\begin{aligned} a_{i_1}^2 + a_{i_2}^2 + \cdots + a_{i_d}^2 &= \boxed{(\text{length of projection})^2} \quad \longleftarrow \text{Maximize this!} \\ &+ (\text{distance of point to line})^2. \end{aligned}$$

Projection

Project a point (vector)

$$\mathbf{a_i} = (a_{i_1}, a_{i_2}, \ldots, a_{i_d})$$

onto a line (vector) $\mathbf{v}$.



Pythagorean Theorem:

$$a_{i_1}^2 + a_{i_2}^2 + \cdots + a_{i_d}^2 = \boxed{(\text{length of projection})^2} + (\text{distance of point to line})^2.$$

← Maximize this!

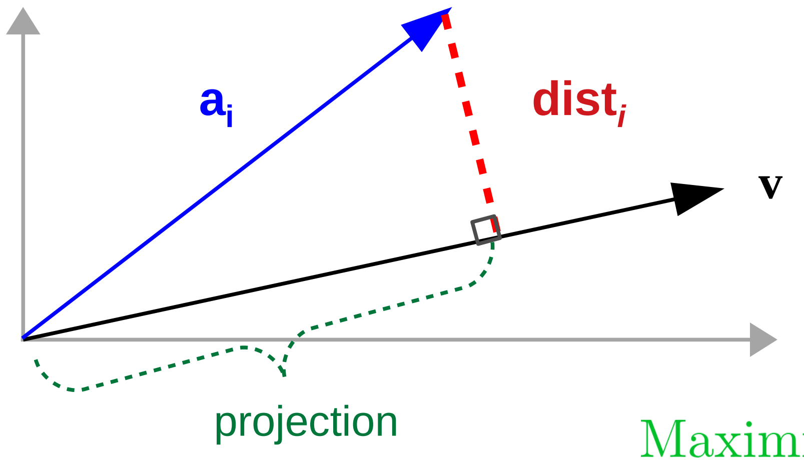
Find a $\mathbf{v}$ such that the sum of the projection lengths is maximum!

Projection

Project a point (vector)

$$\mathbf{a_i} = (a_{i_1}, a_{i_2}, \ldots, a_{i_d})$$

onto a line (vector) $\mathbf{v}$.



Maximize $|A\mathbf{v}|^2$

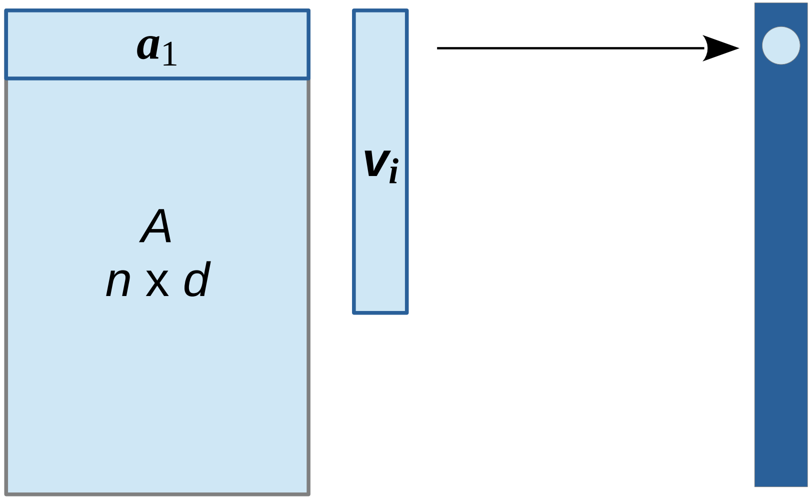
Pythagorean Theorem:

$$a_{i_1}^2 + a_{i_2}^2 + \cdots + a_{i_d}^2 = \boxed{(\text{length of projection})^2} + (\text{distance of point to line})^2.$$

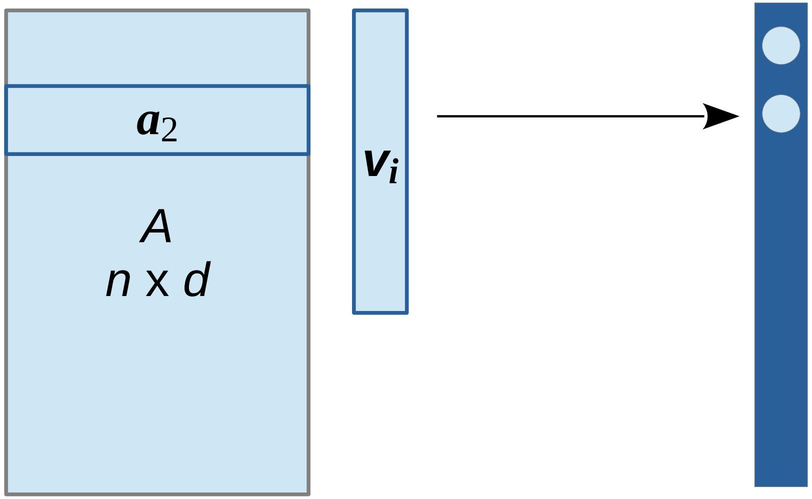
← Maximize this!

Find a $\mathbf{v}$ such that the sum of the projection lengths is maximum!

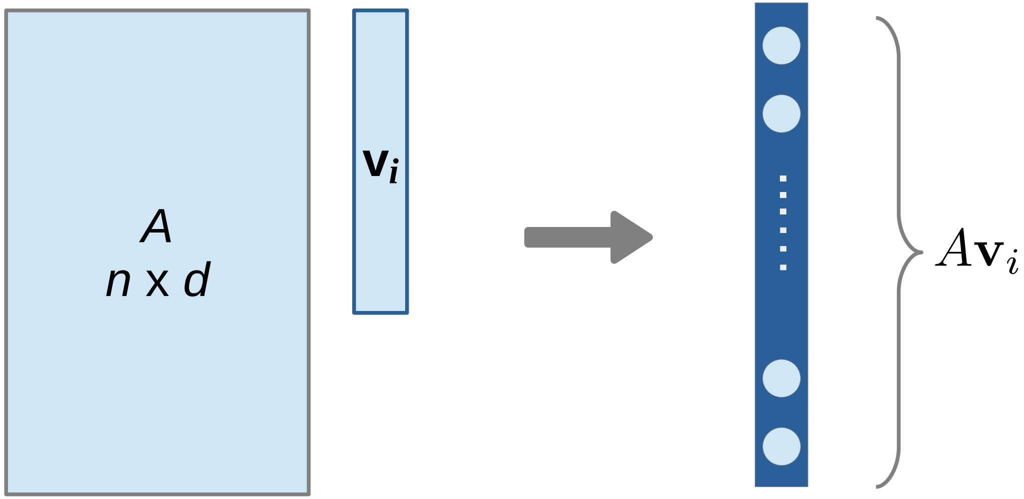
Projection



Projection



Projection



Singular Vectors

- The first singular vector $\mathbf{v}_1$ of A:

$$\mathbf{v}_1 = \arg\max_{|\mathbf{v}|=1} |A\mathbf{v}|^2$$

$\sigma_1 = |A\mathbf{v}|$: first singular value of A.

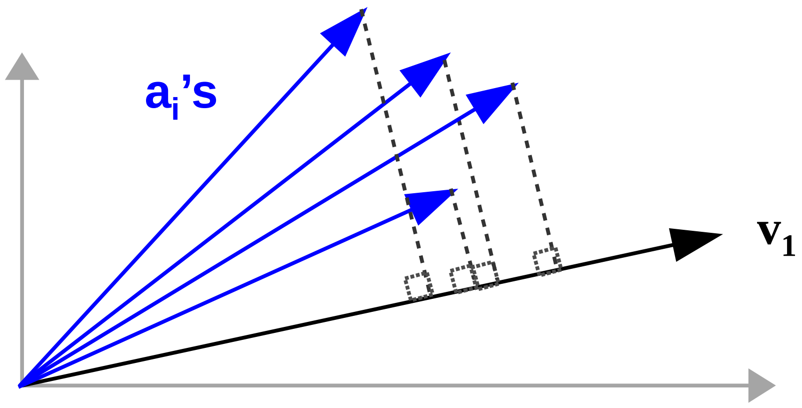
Singular Vectors

- The first singular vector $\mathbf{v}_1$ of A:

$$\mathbf{v}_1 = \arg\max_{|\mathbf{v}|=1} |A\mathbf{v}|^2$$

$$\sigma_1 = |A\mathbf{v}| : \text{first singular value of } A.$$

$$\sigma_1^2 = |A\mathbf{v}|^2 = \sum_{i=1}^{n} (\mathbf{a}_i \cdot \mathbf{v}_i)^2$$



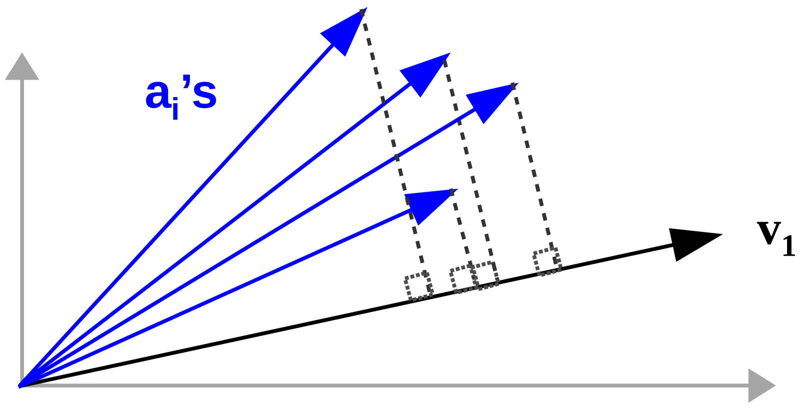
Singular Vectors

- The first singular vector $\mathbf{v}_1$ of A:

$$\mathbf{v}_1 = \arg\max_{|\mathbf{v}|=1} |A\mathbf{v}|^2$$

$$\sigma_1 = |A\mathbf{v}| : \text{first singular value of } A.$$

$$\sigma_1^2 = |A\mathbf{v}|^2 = \sum_{i=1}^{n} (\mathbf{a}_i \cdot \mathbf{v}_i)^2$$



- Sometimes we are not lucky enough so that data points are not close to "one line", but close to a *k*-dimension space.

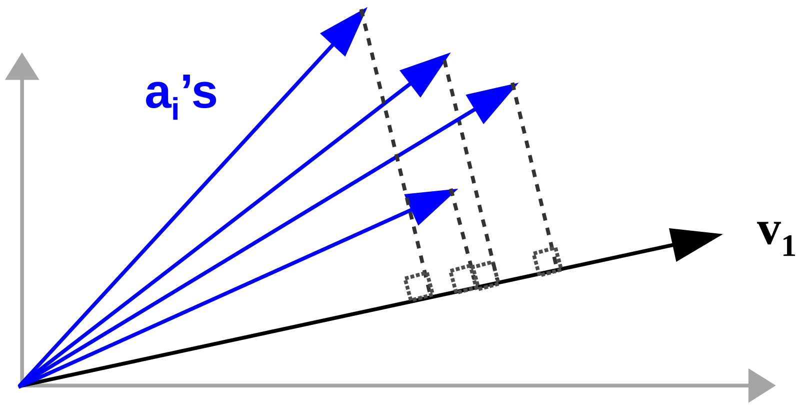
Singular Vectors

- The first singular vector $\mathbf{v}_1$ of A:

$$\mathbf{v}_1 = \arg\max_{|\mathbf{v}|=1} |A\mathbf{v}|^2$$

$$\sigma_1 = |A\mathbf{v}| : \text{first singular value of } A.$$

$$\sigma_1^2 = |A\mathbf{v}|^2 = \sum_{i=1}^{n} (\mathbf{a}_i \cdot \mathbf{v}_i)^2$$



- How about the second, the third, ..., the *k*th singular vectors?

Singular Vectors

- The second singular vector: the best-fit line perpendicular to $\mathbf{v}_1$.

$$\mathbf{v}_2 = \arg\max_{\substack{\mathbf{v}\perp\mathbf{v}_1 \\ |\mathbf{v}|=1}} |A\mathbf{v}|^2. \qquad \sigma_2(A) = |A\mathbf{v}_2| \text{ : second singular value of } A$$

Singular Vectors

- The second singular vector: the best-fit line perpendicular to $\mathbf{v}_1$.

$$\mathbf{v}_2 = \arg \max_{\substack{\mathbf{v} \perp \mathbf{v}_1 \\ |\mathbf{v}|=1}} |A\mathbf{v}|^2. \qquad \sigma_2(A) = |A\mathbf{v}_2| : \text{second singular value of } A$$

- The third singular vector: the best-fit line perpendicular to $\mathbf{v}_1$ and $\mathbf{v}_2$.

$$\mathbf{v}_3 = \arg \max_{\substack{\mathbf{v} \perp \mathbf{v}_1, \mathbf{v}_2, \\ |\mathbf{v}|=1}} |A\mathbf{v}|^2. \qquad \sigma_3(A) = |A\mathbf{v}_3| : \text{third singular value of } A$$

Singular Vectors

- The second singular vector: the best-fit line perpendicular to $\mathbf{v}_1$.

$$\mathbf{v}_2 = \arg \max_{\substack{\mathbf{v} \perp \mathbf{v}_1 \\ |\mathbf{v}|=1}} |A\mathbf{v}|^2. \qquad \sigma_2(A) = |A\mathbf{v}_2| : \text{second singular value of } A$$

- The third singular vector: the best-fit line perpendicular to $\mathbf{v}_1$ and $\mathbf{v}_2$.

$$\mathbf{v}_3 = \arg \max_{\substack{\mathbf{v} \perp \mathbf{v}_1, \mathbf{v}_2, \\ |\mathbf{v}|=1}} |A\mathbf{v}|^2. \qquad \sigma_3(A) = |A\mathbf{v}_3| : \text{third singular value of } A$$

$\vdots$

- Find singular vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ and singular values $\sigma_1, \sigma_2, \ldots, \sigma_r$

$$\max_{\substack{\mathbf{v} \perp \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r \\ |\mathbf{v}|=1}} |A\mathbf{v}|^2 = 0.$$

The Greedy Algorithm Find the Best-Fit Subspace!

- Theorem. Let A be an n x d matrix with singular vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$. For $1 \le k \le r$, let V_k be the subspace spanned by $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$. For each k, V_k is the best-fit k-dimensional subspace for A.

Sum of squares of singular values

$$\begin{aligned}
\sum_{j=1}^{n} |\mathbf{a}_j|^2 &= \sum_{j=1}^{n}\sum_{i=1}^{r} (\mathbf{a}_j \cdot \mathbf{v}_i)^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n} (\mathbf{a}_j \cdot \mathbf{v}_i)^2 \\
&= \sum_{i=1}^{r} |A\mathbf{v}_i|^2 \\
&= \sum_{i=1}^{r} \sigma_i^2(A).
\end{aligned}$$

Sum of squares of singular values

$$
\begin{aligned}
\sum_{j=1}^{n} |\mathbf{a}_j|^2 &= \sum_{j=1}^{n} \boxed{\sum_{i=1}^{r} (\mathbf{a}_j \cdot \mathbf{v}_i)}^2 \\
&= \sum_{i=1}^{r} \sum_{j=1}^{n} (\mathbf{a}_j \cdot \mathbf{v}_i)^2 \\
&= \sum_{i=1}^{r} |A\mathbf{v}_i|^2 \\
&= \sum_{i=1}^{r} \sigma_i^2(A).
\end{aligned}
$$

linear combination of $\boldsymbol{a}_j$ w.r.t. v_i's

Sum of squares of singular values

$$
\begin{aligned}
\boxed{\sum_{j=1}^{n} |\mathbf{a}_j|^2} &= \sum_{j=1}^{n}\sum_{i=1}^{r} (\mathbf{a}_j \cdot \mathbf{v}_i)^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n} (\mathbf{a}_j \cdot \mathbf{v}_i)^2 \\
&= \sum_{i=1}^{r} |A\mathbf{v}_i|^2 \\
&= \boxed{\sum_{i=1}^{r} \sigma_i^2(A).}
\end{aligned}
$$

Sum of squares of singular values

$$
\begin{aligned}
\boxed{\sum_{j=1}^{n} |\mathbf{a}_j|^2} &= \sum_{j=1}^{n}\sum_{i=1}^{r} (\mathbf{a}_j \cdot \mathbf{v}_i)^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n} (\mathbf{a}_j \cdot \mathbf{v}_i)^2 \\
&= \sum_{i=1}^{r} |A\mathbf{v}_i|^2 \\
&= \boxed{\sum_{i=1}^{r} \sigma_i^2(A).}
\end{aligned}
$$

The sum of squares of all the entries of A.

||

The sum of squares of singular values of A.

Sum of squares of singular values

$$\boxed{\sum_{j=1}^{n} |\mathbf{a}_j|^2} = \sum_{j=1}^{n}\sum_{i=1}^{r} (\mathbf{a}_j \cdot \mathbf{v}_i)^2$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{n} (\mathbf{a}_j \cdot \mathbf{v}_i)^2$$

$$= \sum_{i=1}^{r} |A\mathbf{v}_i|^2$$

$$= \boxed{\sum_{i=1}^{r} \sigma_i^2(A).}$$

The sum of squares of all the entries of A.

||

The sum of squares of singular values of A.

The singular values summarizes the whole matrix in some sense.

Left and Right-Singular Vectors

- $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$: **right**-singular vectors.
- To find the left-singular vectors:

Left and Right-Singular Vectors

- $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$: **right**-singular vectors.
- To find the left-singular vectors:

$$\mathbf{u}_i = \frac{1}{\sigma_i(A)} A\mathbf{v}_i.$$

Left and Right-Singular Vectors

- $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$: **right**-singular vectors.
- To find the left-singular vectors:

$$\mathbf{u}_i = \frac{1}{\boxed{\sigma_i(A)}} A\mathbf{v}_i .$$

Concept of normalization

Singular Value Decomposition

- Theorem. Let A be an $n \times d$ matrix with right-singular vectors $\mathbf{v}_1$, $\mathbf{v}_2$, ... , $\mathbf{v}_r$, left-singular vectors $\mathbf{u}_1$, $\mathbf{u}_2$, ... , $\mathbf{u}_r$ and corresponding singular values σ_1, σ_2, ... , σ_r. Then

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top .$$

Singular Value Decomposition

- Theorem. Let A be an n x d matrix with right-singular vectors $\mathbf{v}_1$, $\mathbf{v}_2$, ... , $\mathbf{v}_r$, left-singular vectors $\mathbf{u}_1$, $\mathbf{u}_2$, ... , $\mathbf{u}_r$ and corresponding singular values σ_1, σ_2, ... , σ_r. Then

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top .$$

- *Proof*:

$$\sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \mathbf{v}_j = \sigma_j \mathbf{u}_j = A\mathbf{v}_j \quad \text{for each } j$$

Singular Value Decomposition

- Theorem. Let A be an n x d matrix with right-singular vectors $\mathbf{v}_1$, $\mathbf{v}_2$, ... , $\mathbf{v}_r$, left-singular vectors $\mathbf{u}_1$, $\mathbf{u}_2$, ... , $\mathbf{u}_r$ and corresponding singular values σ_1, σ_2, ... , σ_r. Then

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top .$$

- *Proof*:

$$\sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \mathbf{v}_j = \sigma_j \mathbf{u}_j = A\mathbf{v}_j \quad \text{for each } j$$

Any vector $\mathbf{v}$ can be expressed as a linear combination of $\mathbf{v}_i$'s plus a vector perpendicular to the $\mathbf{v}_i$'s.

$\boxed{A = B \Leftrightarrow Av = Bv \text{ for each } v}$

Best Rank-k Approximations

- The sum truncated after k terms.

$$A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top .$$

- 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.

Best Rank-k Approximations

- The sum truncated after k terms.

$$A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top .$$

- 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.

$$\text{projection of vector } \mathbf{a} \text{ onto } V_k : \quad \sum_{i=1}^{k} (\mathbf{a} \cdot \mathbf{v}_i) \mathbf{v}_i^\top .$$

$$\therefore \sum_{i=1}^{k} A \mathbf{v}_i \mathbf{v}_i^\top = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top = A_k .$$

On Left Singular Vectors

- The left singular vectors are pairwise **orthogonal**.

On Left Singular Vectors

- The left singular vectors are pairwise **orthogonal**.

  i : the smallest integer such that $\mathbf{u}_i$ is not orthogonal to some other $\mathbf{u}_j$.

  Assume that $\mathbf{u}_i^\top \mathbf{u}_j = \delta > 0$.

  For $\epsilon > 0$, let $\mathbf{v}'_i = \dfrac{\mathbf{v}_i + \epsilon \mathbf{v}_j}{|\mathbf{v}_i + \epsilon \mathbf{v}_j|}$. $\quad A\mathbf{v}'_i = \dfrac{\sigma_i \mathbf{u}_i + \epsilon \sigma_j \mathbf{u}_j}{\sqrt{1+\epsilon^2}}$.

On Left Singular Vectors

- The left singular vectors are pairwise **orthogonal**.

  i : the smallest integer such that $\mathbf{u}_i$ is not orthogonal to some other $\mathbf{u}_j$.

  Assume that $\mathbf{u}_i^\top \mathbf{u}_j = \delta > 0$.

  $$\text{For } \epsilon > 0, \text{ let } \mathbf{v}_i' = \frac{\mathbf{v}_i + \epsilon \mathbf{v}_j}{|\mathbf{v}_i + \epsilon \mathbf{v}_j|}. \quad A\mathbf{v}_i' = \frac{\sigma_i \mathbf{u}_i + \epsilon \sigma_j \mathbf{u}_j}{\sqrt{1+\epsilon^2}}.$$

  Note that $\mathbf{v}_i + \epsilon \mathbf{v}_j \perp \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{i-1}$. $(j > i)$

On Left Singular Vectors

- The left singular vectors are pairwise **orthogonal**.

  i : the smallest integer such that $\mathbf{u}_i$ is not orthogonal to some other $\mathbf{u}_j$.

  Assume that $\mathbf{u}_i^\top \mathbf{u}_j = \delta > 0$.

  For $\epsilon > 0$, let $\mathbf{v}_i' = \dfrac{\mathbf{v}_i + \epsilon \mathbf{v}_j}{|\mathbf{v}_i + \epsilon \mathbf{v}_j|}$. $\boxed{A\mathbf{v}_i'} = \dfrac{\boxed{\sigma_i \mathbf{u}_i} + \epsilon \sigma_j \mathbf{u}_j}{\sqrt{1+\epsilon^2}}$.

  Note that $\mathbf{v}_i + \epsilon \mathbf{v}_j \perp \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{i-1}$.

  Clearly, $|A\mathbf{v}_i'| \geq \mathbf{u}_i^\top \cdot (A\mathbf{v}_i')$.

On Left Singular Vectors

- The left singular vectors are pairwise **orthogonal**.

i : the smallest integer such that $\mathbf{u}_i$ is not orthogonal to some other $\mathbf{u}_j$.

Assume that $\mathbf{u}_i^\top \mathbf{u}_j = \delta > 0$.

For $\epsilon > 0$, let $\mathbf{v}_i' = \dfrac{\mathbf{v}_i + \epsilon \mathbf{v}_j}{|\mathbf{v}_i + \epsilon \mathbf{v}_j|}$. $\quad A\mathbf{v}_i' = \dfrac{\sigma_i \mathbf{u}_i + \epsilon \sigma_j \mathbf{u}_j}{\sqrt{1+\epsilon^2}}$.

Note that $\mathbf{v}_i + \epsilon \mathbf{v}_j \perp \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{i-1}$.

Clearly, $|A\mathbf{v}_i'| \geq \mathbf{u}_i^\top \cdot (A\mathbf{v}_i')$.

$\Rightarrow$ $|A\mathbf{v}_i'| > |A\mathbf{v}_i| \quad (\Rightarrow\Leftarrow)$

$$\begin{aligned} \mathbf{u}_i^\top \left(\frac{\sigma_i \mathbf{u}_i + \epsilon \sigma_j \mathbf{u}_j}{\sqrt{1+\epsilon^2}} \right) &> (\sigma_i + \epsilon \sigma_j \delta)\left(1 - \epsilon^2/2\right) \\ &> \sigma_i. \end{aligned}$$

Analog of eigenvalues and eigenvectors

$$A\mathbf{v}_i = \sigma_i \mathbf{u}_i \text{ and } A^\top \mathbf{u}_i = \sigma_i \mathbf{v}_i.$$

Analog of eigenvalues and eigenvectors

$$A\mathbf{v}_i = \sigma_i \mathbf{u}_i \text{ and } A^\top \mathbf{u}_i = \sigma_i \mathbf{v}_i.$$

Proof for the second:

From the SVD, we have

Analog of eigenvalues and eigenvectors

$$A\mathbf{v}_i = \sigma_i \mathbf{u}_i \text{ and } A^\top \mathbf{u}_i = \sigma_i \mathbf{v}_i.$$

Proof for the second:

From the SVD, we have

$$A^\top \mathbf{u}_i = \sum_j (\sigma_j \mathbf{u}_j \mathbf{v}_j^\top)^\top \mathbf{u}_i$$

Analog of eigenvalues and eigenvectors

$$A\mathbf{v}_i = \sigma_i \mathbf{u}_i \text{ and } A^\top \mathbf{u}_i = \sigma_i \mathbf{v}_i.$$

Proof for the second:

From the SVD, we have

$$\begin{aligned} A^\top \mathbf{u}_i &= \sum_j (\sigma_j \mathbf{u}_j \mathbf{v}_j^\top)^\top \mathbf{u}_i \\ &= \sum_j \sigma_j \mathbf{v}_j \mathbf{u}_j^\top \mathbf{u}_i \\ &= \sigma_i \mathbf{v}_i \mathbf{u}_i^\top \mathbf{u}_i = \sigma_i \mathbf{v}_i. \end{aligned}$$

A_k: best rank-k 2-norm approximation

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

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

A_k: best rank-k 2-norm approximation

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

Note: $||A||_2 = \max\limits_{|\mathbf{x}|=1} |A\mathbf{x}|$.

Let $A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. Then $A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. $A - A_k = \sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$.

A_k: best rank-k 2-norm approximation

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

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

$$\text{Let } A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top. \text{ Then } A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top. \; A - A_k = \sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.$$

$\mathbf{v}$: top singular vector of $A - A_k$.

$$\text{Express } \mathbf{v} : \quad \mathbf{v} = \sum_{j=1}^{r} c_j \mathbf{v}_j.$$

A_k: best rank-k 2-norm approximation

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

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

Let $A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. Then $A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. $A - A_k = \sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$.

$\mathbf{v}$: top singular vector of $A - A_k$.

Express $\mathbf{v}$: $\mathbf{v} = \sum_{j=1}^{r} c_j \mathbf{v}_j$.

$$
\begin{aligned}
|(A - A_k)\mathbf{v}| &= \left| \sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \sum_{j=1}^{r} c_j \mathbf{v}_j \right| \\
&= \left| \sum_{i=k+1}^{r} c_i \sigma_i \mathbf{u}_i \mathbf{v}_i^T \mathbf{v}_i \right| \\
&= \left| \sum_{i=k+1}^{r} c_i \sigma_i \mathbf{u}_i \right| = \sqrt{\sum_{i=k+1}^{r} c_i^2 \sigma_i^2}.
\end{aligned}
$$

A_k: best rank-k 2-norm approximation

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

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

Let $A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.$ Then $A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.$ $A - A_k = \sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.$

$\mathbf{v}$: top singular vector of $A - A_k$.

Express $\mathbf{v}$: $\mathbf{v} = \sum_{j=1}^{r} c_j \mathbf{v}_j.$

$$
\begin{aligned}
|(A - A_k)\mathbf{v}| &= \left| \sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \sum_{j=1}^{r} c_j \mathbf{v}_j \right| \\
&= \left| \sum_{i=k+1}^{r} c_i \sigma_i \mathbf{u}_i \mathbf{v}_i^T \mathbf{v}_i \right| \\
&= \left| \sum_{i=k+1}^{r} c_i \sigma_i \mathbf{u}_i \right| = \boxed{\sqrt{\sum_{i=k+1}^{r} c_i^2 \sigma_i^2}.}
\end{aligned}
$$

Try to maximize it!

$\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_r^2.$

$|\mathbf{v}|^2 = \sum_{i=1}^{r} c_i^2 = 1.$

Set $c_{k+1} = 1$ and the rest to be zero.

A_k: best rank-k 2-norm approximation

- Theorem. Let A be an n x d matrix. For any matrix B of rank $\leq k$,

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

A_k: best rank-k 2-norm approximation

- <u>Theorem</u>. Let A be an n x d matrix. For any matrix B of rank ≤ k,

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

  - Assume A is of rank > k.

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

A_k: best rank-k 2-norm approximation

- <u>Theorem</u>. Let A be an n x d matrix. For any matrix B of rank $\leq k$,

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

  - Assume A is of rank > k.

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

$\dim(\mathrm{Null}(B)) \geq d - k.$

$\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k+1}$: first $k+1$ singular vectors of A.

A_k: best rank-k 2-norm approximation

- Theorem. Let A be an n x d matrix. For any matrix B of rank $\leq k$,

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

  - Assume A is of rank $> k$.

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

$\dim(\mathrm{Null}(B)) \geq d - k$.

$\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k+1}$: first $k+1$ singular vectors of A.

$\exists \mathbf{z} \neq \mathbf{0},\ |\mathbf{z}| = 1,\ \mathbf{z} \in \mathrm{Null}(B) \cap \mathrm{span}(\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\})$.

A_k: best rank-k 2-norm approximation

- Theorem. Let A be an n x d matrix. For any matrix B of rank $\leq k$,

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

  - Assume A is of rank $> k$.

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

$\dim(\mathrm{Null}(B)) \geq d - k.$

$\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k+1}$: first $k+1$ singular vectors of A.

$\exists \mathbf{z} \neq \mathbf{0}, \ |\mathbf{z}| = 1, \ \mathbf{z} \in \mathrm{Null}(B) \cap \mathrm{span}(\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\}).$

$||A - B||_2^2 \geq |(A - B)\mathbf{z}|^2 \geq |A\mathbf{z}|^2 \quad (\because B\mathbf{z} = 0)$

A_k: best rank-k 2-norm approximation

1 x d $\mathbf{v}_i^T$ $\mathbf{z}$ d x 1

- <u>Theorem</u>. Let A be an n x d matrix. For any matrix B of rank $\leq k$,

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

  - Assume A is of rank $> k$.

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

$\dim(\mathrm{Null}(B)) \geq d - k.$

$\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k+1}$: first $k+1$ singular vectors of A.

$\exists \mathbf{z} \neq \mathbf{0},\ |\mathbf{z}| = 1,\ \mathbf{z} \in \mathrm{Null}(B) \cap \mathrm{span}(\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\}).$

$||A - B||_2^2 \geq |(A - B)\mathbf{z}|^2 \geq |A\mathbf{z}|^2 \quad (\because B\mathbf{z} = 0)$

$$\begin{aligned}
|A\mathbf{z}|^2 &= \left|\sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \mathbf{z}\right|^2 \\
&= \sum_{i=1}^{n} \sigma_i^2 (\mathbf{v}^T \mathbf{z})^2 \\
&= \sum_{i=1}^{k+1} \sigma_i^2 (\mathbf{v}_i^T \mathbf{z})^2 \\
&\geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} (\mathbf{v}_i^T \mathbf{z})^2 \\
&\geq \sigma_{k+1}^2.
\end{aligned}$$

Remark

- Singular Value Decomposition (SVD) is used as core routine in the **Principal Component Analysis (PCA)**.
  - Unsupervised learning paradigm.
  - A dimensionality reduction algorithm.

Application: Principle Component Analysis

- Movie recommendation:
  - n customers and d movies.
  - $a_{i,j}$: the amount how customer i likes movie j.

Application: Principle Component Analysis

- Movie recommendation:
  - n customers and d movies.
  - $a_{i,j}$: the amount how customer i likes movie j.
- Hypothesis:
  - Only k underlying factors that determine a customer likes a movie.
  - $k << n, d$.

Application: Principle Component Analysis

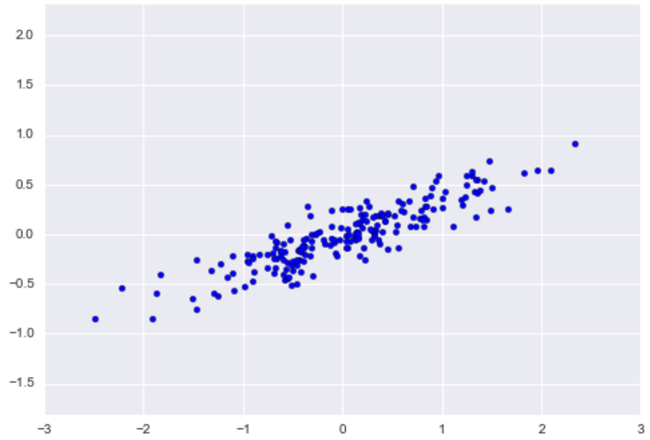
- Movie recommendation:
  - n customers and d movies.
  - $a_{i,j}$: the amount how customer i likes movie j.
- Hypothesis:
  - Only k underlying factors that determine a customer likes a movie.
  - $k << n, d$.
- U: $n \times k$ describing the customers.
- V^T: $d \times k$ matrix, describing the movies.

Application: Principle Component Analysis

- Movie recommendation:
  - n customers and d movies.
  - $a_{i,j}$: the amount how customer i likes movie j.
- Hypothesis:
  - Only k underlying factors that determine a customer likes a movie.
  - $k \ll n, d$.
- U: $n \times k$ describing the customers.
- V^T: $d \times k$ matrix, describing the movies.
- Goal: Find the best rank-k approximation A_k.

In scikit-learn

- Suppose we have a dataset ***X*** of two-dimension.



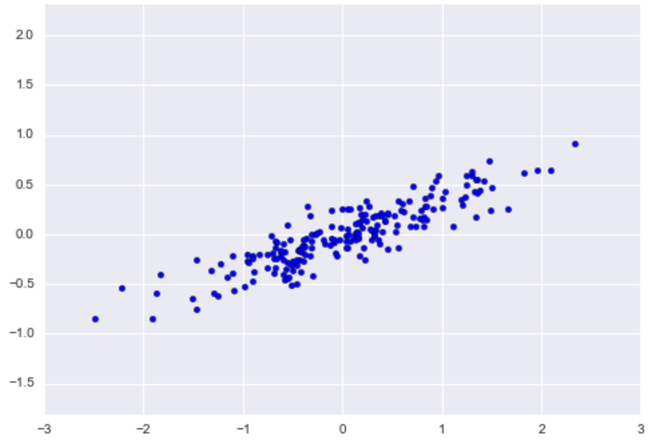
A set of 200 randomly generated 2-D points

In scikit-learn

- Suppose we have a dataset ***X*** of two-dimension.
- Import PCA as below:

```
from sklearn.decomposition import PCA
pca = PCA(n_components=2)
pca.fit(X)
```

n_component=2 ⟺ $k = 2$



A set of 200 randomly generated 2-D points

In scikit-learn

- Suppose we have a dataset ***X*** of two-dimension.
- Import PCA as below:

```
from sklearn.decomposition import PCA
pca = PCA(n_components=2)
pca.fit(X)
```

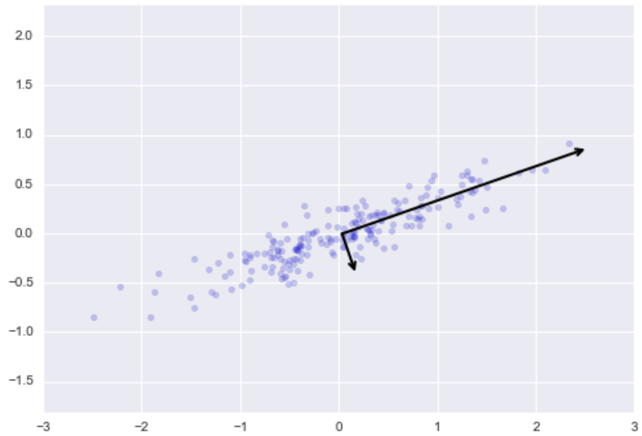
n_component=2 ⟺ $k = 2$

```
print(pca.components_)
```

[[0.94446029 0.32862557]
 [0.32862557 -0.94446029]]

```
print(pca.explained_variance_)
```

[0.75871884 0.01838551]



A set of 200 randomly generated 2-D points + components as axis

In scikit-learn

- Dimension reduction:

```
pca = PCA(n_components=1)
pca.fit(X)
X_pca = pca.transform(X)
```

```
X_new = pca.inverse_transform(X_pca)
plt.scatter(X[:, 0], X[:, 1], alpha=0.2)
plt.scatter(X_new[:, 0], X_new[:, 1], alpha=0.8)
```

In scikit-learn

- Refer to the webpage for more detail.