Recall \( A = [U_1, U_2]\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} = U_1 \Sigma V_1^T \)

We also saw \( \|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sigma_1 \),

which \( x \) achieves this maximum? \( \text{It's } v_1 \) (first right singular vec.)

\[ A v_1 = U_1 \Sigma V_1^T v_1 = U_1 \Sigma e_1 = U_1 \sigma_1 e_1 = \sigma_1 u_1 \]

\[ \frac{\|Av_1\|}{\|v_1\|} = \frac{\sigma_1 \|u_1\|}{\|v_1\|} = \sigma_1 \quad \text{(since } u_1, v_1 \text{ are normalized)} \]

In general, we have:

\[ A v_i = \sigma_i u_i \quad \text{and} \quad A^T u_i = \sigma_i v_i \quad \forall i. \]

This is related to notion of eigenvalues:

\[ A^T A v_i = A^T (\sigma_i u_i) = \sigma_i^2 v_i \Rightarrow \sigma_i^2 \text{ is eigenvalue of } A^T A \]

\[ A A^T u_i = A (\sigma_i v_i) = \sigma_i^2 u_i \Rightarrow \sigma_i^2 \text{ is eigenvalue of } A A^T \]

(This is one way to compute SVD.)
Directly, we have:  
\[ A^T A = (U \Sigma V^T)^T (U \Sigma V^T) = U \Sigma U^T U \Sigma V^T \]

eigenvalue decomposition:  
\[ \Rightarrow = U \Sigma \Sigma V^T \]

\[ \Rightarrow (\Sigma, 0)^T (\Sigma, 0) = (\sigma_1^2, \sigma_2^2, 0) \]

Similarly,  
\[ A A^T = U \Sigma \Sigma^T U^T \]

\[ \star \text{ if } A \text{ is symmetric and positive semidefinite, } \lambda_i = \sigma_i \]
(eigenvalues are the same as singular values).

\[ \star \text{ in general, a square matrix } B \]
\[ \rightarrow \text{ might have complex eigenvalues} \]
\[ \rightarrow \text{ might not have orthogonal eigenvectors (} P D P^{-1} \text{ instead of) } \]
\[ \rightarrow \text{ might not be diagonalizable at all! (Jordan form) } P D P^{-1} \]

whereas every matrix (even non-square) has an SVD and real positive singular values.

---

**Geometric interpretation of SVD,**

\[ y = A x \]

Instead:

\[ y = U \Sigma V^T x \]

\[ R^n \rightarrow V^T \rightarrow \Sigma \rightarrow U \rightarrow R^m \]
\( \mathbb{R}^n \xrightarrow{\mathbf{V}^T} \mathbb{R}^n \) (rotation)

\[ \sum \] scaling some components and discarding others

\( \mathbb{R}^m \xrightarrow{\mathbf{U}} \mathbb{R}^m \) (rotation)

**m 2D \rightarrow 2D**

\( \mathbb{R}^n \) (\( n=2 \))

\( \mathbb{R}^m \) (\( m=2 \))

\( \mathbf{A} \)

\( \mathbf{v}_1 \rightarrow \sigma_1 \mathbf{u}_1 \)

\( \mathbf{v}_2 \rightarrow \sigma_2 \mathbf{u}_2 \)

circle maps to ellipse.
(of course, \( \mathbb{R}^2 \rightarrow \mathbb{R}^2 \) as well).

**m 2D \rightarrow 3D**

\( \mathbb{R}^n \) (\( n=2 \))

\( \mathbb{R}^m \) (\( m=2 \))

\( \mathbf{A} \)

\( \mathbf{v}_3 \rightarrow \mathbf{N}(\mathbf{A}) \) (goes nowhere)

\( \{\mathbf{v}_1, \mathbf{v}_2\} \rightarrow \mathbf{N}(\mathbf{A})^\perp \) (unreachable)

\( \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\} \rightarrow \mathbf{U}_1 \mathbf{U}_2 \mathbf{U}_3 \rightarrow \mathbf{U}(\mathbf{A}) \) (reachable).
Finding the closest line to a set of points. Example

\[ a_i \in \mathbb{R}^d \quad i = 1, \ldots, n \]

are points and we want to minimize

\[ \sum_{i=1}^{n} d_i^2 \]

(sum of squares of distances to the line).

\[ A \] This is different from regression! Here, our line is coordinate-independent, rotating the points also rotates the line. In regression, this is not the case.

\[ B \] why not \( \sum_{i=1}^{n} d_i \) (sum of distances) or \( \max d_i \) (max dist)? It's a choice! We will see these other types of distance measurements later in class.

In this scenario, \( d_i^2 = \|a_i - \text{proj}_x a_i\|^2 \)
recall \( \text{proj}_x a = \frac{x^T a}{x^T x} x \)

Note: we can interchange matrix-vector-scalar multiplication!

\[
\text{e.g. } x^T a x = \frac{xx^T a}{x^T} = \frac{x x^T a}{\text{vector}} \quad \frac{x x^T a}{\text{matrix}} \quad \frac{x x^T a}{\text{vector}} (\ast)
\]

also, \( (a^T x)^2 = (a^T x)(a^T x) = a^T (x x^T) a = x^T (a a^T) x \) (quadratic forms, scalar-scalar multiplication)

\[
a_i^2 = \| a_i - \text{proj}_x a_i \|^2
\]

\[
= \| a_i - \frac{x^T a_i}{x^T x} x \|^2 \quad \text{use trick (\ast)}
\]

\[
= \| a_i - \frac{1}{x^T x} (x x^T) a_i \|^2 \quad \text{factor } a_i \text{ from the right.}
\]

\[
= \| (I - \frac{1}{x^T x} (x x^T)) a_i \|^2 \quad \text{use fact that } \|v\|^2 = v^T v.
\]

\[
= a_i^T (I - \frac{1}{x^T x} (x x^T))(I - \frac{1}{x^T x} (x x^T)) a_i
\]

\[
= a_i^T (I - \frac{1}{x^T x} (x x^T) - \frac{1}{x^T x} (x x^T) + \frac{1}{(x^T x)^2} x x^T x x^T) a_i
\]

\[
= a_i^T (I - \frac{1}{x^T x} x x^T) a_i
\]
\[
d_i^2 = a_i^T \left( I - \frac{1}{x^T x} x x^T \right) a_i \\
= a_i^T a_i - \frac{1}{x^T x} a_i^T x^T a_i \quad \text{using trick on prev page.} \\
= a_i^T a_i - \frac{1}{x^T x} x^T (a_i a_i^T) x.
\]

Now minimizing \( d_i^2 \) is equivalent to maximizing \( \frac{1}{x^T x} x^T (a_i a_i^T) x \) because of the negative sign and because \( a_i^T a_i \) is constant.

\[
\max_{x \neq 0} \frac{1}{x^T x} \sum_{i=1}^n x^T (a_i a_i^T) x
\]

\[
= \max_{x \neq 0} \frac{1}{x^T x} x^T \left( \sum_{i=1}^n a_i a_i^T \right) x
\]

\[
= \max_{x \neq 0} \frac{x^T A^T A x}{x^T x}
\]

\[
= \max_{x \neq 0} \frac{\| A x \|^2}{\| x \|^2} \quad \text{where } x_{\text{opt}} = v_1, \text{ the first right singular vector!}
\]

\[
\tilde{\sigma}_1^2
\]

so the best line is \( x = v_1 \) where \( v_1 \) is first col. of \( V \) in the SVD \( A = U \Sigma V^T \) and \( A = \begin{bmatrix} a_1^T \\ \vdots \\ a_n^T \end{bmatrix} \) is the data matrix.
In general, the best approximation to a set of points by a \( k \)-dimensional subspace can be found in a similar fashion. Here, instead of choosing a direction \( x \), we choose an orthonormal basis \( \{w_1, w_2, \ldots, w_k\} \). Replace \( \text{proj}_x a_i \) with \( \text{proj}_W a_i \) where \( W = [w_1, w_2, \ldots, w_k] \),

\[
\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n} ||a_i - \text{proj}_W a_i||^2 = \sum_{i=1}^{n} ||a_i - WW^T a_i||^2 = \sum_{i=1}^{n} a_i^T (I - WW^T) a_i = \sum_{i=1}^{n} a_i^T (I - WW^T) a_i = (\sum_{i=1}^{n} a_i^T a_i) - \sum_{i=1}^{n} a_i^T WW^T a_i = (\sum_{i=1}^{n} a_i^T a_i) - \text{trace}\left[ W^T (\sum_{i=1}^{n} a_i a_i^T) W \right].
\]

so minimizing \( \sum_{i=1}^{n} a_i^2 \) is equivalent to maximizing \( \text{trace}(W^T A^T A W) \)

\[
\Rightarrow \text{maximize } ||AW||_F^2 \quad \text{if we let } A = U_1 \Sigma_1 V_1^T
\]

We choose \( W \in \mathbb{R}^{n \times k} \), Worthington

\[
||AW||_F^2 = ||U_1 \Sigma_1 V_1^T W||_F^2 = ||\Sigma_1 V_1^T W||_F^2
\]

maximized when \( W = [V_1, V_2, \ldots, V_k] \).

i.e. \( \text{span} [V_1, ..., V_k] \) is best \( k \)-dimensional subspace. These are called the principal components.