The Singular Value Decomposition

This is the capstone of the class. It's a fundamental application of linear algebra to:

- Statistics (PCA)
- Engineering
- Data Science
- etc.

Today we'll discuss the outer product form and the mechanics (plumbing?) of the SVD.

Introduction to the SVD

Thm (SVD, outer product form): Let \( A \) be an \( m \times n \) matrix of rank \( r \). Then

\[
A = \sigma_1 u_1 v_1^\top + \sigma_2 u_2 v_2^\top + \cdots + \sigma_r u_r v_r^\top
\]

where

- \( \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 \)
- \( \{u_1, \ldots, u_r\} \) is an orthonormal set in \( \mathbb{R}^m \)
- \( \{v_1, \ldots, v_r\} \) is an orthonormal set in \( \mathbb{R}^n \).

What does this mean?

Idea: columns of \( A \) are data points

Here's an informal description of what SVD says.
\( r = 1 \): Let \( u, v \in \mathbb{R}^n \) be nonzero vectors.

\[
uv^T = \begin{pmatrix}
    u_1 & \cdots & u_n
\end{pmatrix}
\begin{pmatrix}
    v_1 & \cdots & v_n
\end{pmatrix}
=
\begin{pmatrix}
    v_1u_1 & \cdots & v_nu_n
\end{pmatrix}
\]

This is an \( m \times n \) matrix of rank 1: \( \text{Col}(uv^T) = \text{Span} \{u\} \).

Let's plot the columns ("data points")

\[
\begin{pmatrix}
    3 \\
    2
\end{pmatrix}
\begin{pmatrix}
    -1 & 2 & 1 & 3 & -2
\end{pmatrix}
\]

Upshot: A rank-1 matrix encodes data points (columns) that lie on a line (dim Col(A) = 1). The SVD tells you which line \& which multiples.

\( r = 2 \):

\[
A = u_1v_1^T + u_2v_2^T
= \begin{pmatrix}
    v_1u_1 & \cdots & v_nu_1
\end{pmatrix}
+ \begin{pmatrix}
    v_2u_2 & \cdots & v_nu_2
\end{pmatrix}

= \begin{pmatrix}
    v_1u_1 + v_2u_1 & \cdots & v_nu_1 + v_2u_2
\end{pmatrix}
\]

The columns are linear combinations of \( u_1 \) \& \( u_2 \).

Let's plot the columns ("data points"):

\[
\begin{pmatrix}
    3 \\
    2
\end{pmatrix}
\begin{pmatrix}
    -1 & 2 & 1 & 3 & -2
\end{pmatrix}
\]

\[
\begin{pmatrix}
    3 \\
    2
\end{pmatrix}
\begin{pmatrix}
    3 & 1 & 2 & -1 & 0
\end{pmatrix}
\]

\( v_1 = \) weights of \( \begin{pmatrix} 3 \\ 2 \end{pmatrix} \)

\( v_2 = \) weights of \( \begin{pmatrix} 3 \\ 2 \end{pmatrix} \)
Upshot: A rank-2 matrix encodes data points that lie on a plane \((\dim \text{Col}(A)=2))\). The SVD gives you a basis \(3u_1, u_2, \bar{3}\) and the weights for each column.

But: \(\| (\frac{3}{2}) \| \gg \| (\frac{-2}{3}) \|\) so the \((-\frac{2}{3})\) direction is less important!

\[
\begin{pmatrix}
\frac{3}{2} \\
\frac{1}{2}
\end{pmatrix}
(-1 2 1 3 -2) + \begin{pmatrix}
\frac{2}{3} \\
\frac{-3}{3}
\end{pmatrix}
(3 1 2 -1 0)
\]

\[
\cong \begin{pmatrix}
\frac{3}{2} \\
\frac{1}{2}
\end{pmatrix}
(-1 2 1 3 -2) \quad \text{(to one decimal place)}
\]

We've extracted important information: our data points almost lie on a line!

In general, the SVD will find the best-fit line, plane, 3-space, ..., r-space for our data, all at once, and tell you how good is the fit in the sense of orthogonal least squares!

(more on this later)
Why might we care?

- Data compression: $uv^\top$ is 7 numbers instead of 10 for a $2\times 5$ matrix.
- Data analysis: SVD will reveal all approximate linear relations among our data points.
- Dimension reduction: if our data in $\mathbb{R}^{1000}$ almost lie on a 1000-dimensional subspace then computers are happier to do the computations.
- Statistics: SVD finds more & less important correlations etc.

Mechanics of the SVD

Back to the statement of the SVD:

$$A = \sigma_1 u_1 v_1^\top + \sigma_2 u_2 v_2^\top + \ldots + \sigma_r u_r v_r^\top \quad r = \text{rank}(A)$$

where

- $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq 0$
- $\{u_1, \ldots, u_r\}$ is an orthonormal set in $\mathbb{R}^m$
- $\{v_1, \ldots, v_r\}$ is an orthonormal set in $\mathbb{R}^n$.

Def: $\sigma_1, \ldots, \sigma_r$ are the singular values of $A$
- $u_1, \ldots, u_r$ are the left singular vectors
- $v_1, \ldots, v_r$ are the right singular vectors
Here are some formal consequences of the statement.

**Note 1:** For any vector \( x \in \mathbb{R}^n \),
\[
Ax = (\sigma_1 u_1 v_1^T + \cdots + \sigma_r u_r v_r^T)x = \sigma_1 u_1 v_1^T x + \cdots + \sigma_r u_r v_r^T x
= \sigma_1 (v_1 \cdot x) u_1 + \cdots + \sigma_r (v_r \cdot x) u_r
\]
\[
A_x = \sigma_1 (v_1 \cdot x) u_1 + \cdots + \sigma_r (v_r \cdot x) u_r
\]

**Note 2:** Taking \( x = v_i \), we have
\[
Av_i = (\sigma_1 (v_i \cdot v_i) u_1 + \cdots + \sigma_i (v_i \cdot v_i) u_i + \cdots + \sigma_r (v_i \cdot v_i) u_r)
\]
(recall \( v_1, \ldots, v_r \) and \( s_1, \ldots, s_r \) are orthonormal).
So the singular vectors are related by
\[
Av_i = \sigma_i u_i \quad \text{and thus} \quad ||Av_i|| = \sigma_i
\]

**Note 3:** Take transposes:
\[
A^T = (\sigma_1 u_1 v_1^T + \cdots + \sigma_r u_r v_r^T)^T = \sigma_1 v_1 u_1^T + \cdots + \sigma_r v_r u_r^T
\]
Therefore, \[
A^T = \sigma_1 v_1 u_1^T + \cdots + \sigma_r v_r u_r^T
\]
is the SVD of \( A^T \)!

So \( A \) & \( A^T \) have the same
\- singular values and
\- singular vectors (switch right & left).
Note 4: \( \text{Note 2 + Note 3} \Rightarrow A^T u_i = \sigma_i v_i \) so
\( A^T A v_i = A^T (\sigma_i u_i) = \sigma_i A^T u_i = \sigma_i^2 v_i \)
\( A A^T u_i = A (\sigma_i v_i) = \sigma_i A v_i = \sigma_i^2 u_i \)

In particular,
\[ \exists u_1, \ldots, u_r \] are orthonormal eigenvectors of \( A^T A \)
with eigenvalues \( \sigma_1^2, \ldots, \sigma_r^2 \).
\[ \exists u_{r+1}, \ldots, u_n \] are orthonormal eigenvectors of \( A A^T \)
with eigenvalues \( \sigma_{r+1}^2, \ldots, \sigma_n^2 \).

This tells us how to prove/compute the SVD:
orthogonally diagonalize \( A^T A \)

**Proof of the SVD:** Pay attention to steps 1-3: they illustrate the mechanics of the SVD!

Let \( \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \geq 0 \) be the eigenvalues of \( A^T A \)
(the \( \lambda_i \)'s show up multiple times if \( A^T A \) is non-square).

Note \( \lambda_n \geq 0 \) because \( A^T A \) is positive-semidefinite.

**Step 1:** I claim \( \lambda_1 = \ldots = \lambda_r = 0 \).
- \( \text{Nul} (A^T A) = \text{Nul} (A) \) has dimension \( n-r \).
- \( \text{Nul} (A^T A) = \) the \( 0 \)-eigenspace of \( A^T A \).
• $A(M(0)) = GM(0) \Rightarrow ATA$
  
  because $ATA$ is symmetric $\Rightarrow$ diagonalizable

So $n-r$ of the $\lambda_i$'s are $= 0$

$\Rightarrow \lambda_{r+1} = \cdots = \lambda_n = 0$

Now: $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$ are the nonzero eigenvalues of $ATA$.

Set:

• $\sigma_1 = \sqrt{\lambda_1}, \ldots, \sigma_n = \sqrt{\lambda_n}$

• Let $V_1, \ldots, V_r$ be orthonormal eigenvectors
  
  with $ATAV_i = \lambda_i V_i$.

• Let $U_1 = \frac{1}{\sigma_1} AV_1, \ldots, U_r = \frac{1}{\sigma_r} AV_r$

Step 2: I claim $\{U_1, \ldots, U_r\}$ is orthonormal. Check:

$U_i \cdot U_j = U_i^T U_j = (\frac{1}{\sigma_i} AV_i)^T (\frac{1}{\sigma_j} AV_j) = \frac{1}{\sigma_i \sigma_j} (AV_i)^T (AV_j)$

$= \frac{1}{\sigma_i \sigma_j} (V_i^T A^T AV_j) = \frac{1}{\sigma_i \sigma_j} V_i^T (ATAV_j) = \frac{1}{\sigma_i \sigma_j} V_i^T (\lambda_j V_j)$

$= \frac{\sigma_i^2}{\sigma_i \sigma_j} V_i^T V_j = \frac{\sigma_i}{\sigma_j} V_i \cdot V_j$

Since $\{V_1, \ldots, V_r\}$ is orthonormal:

• If $i = j$ this is $U_i \cdot U_i = \frac{\sigma_i}{\sigma_i} V_i \cdot V_i = \|V_i\|^2 = 1$

• If $i \neq j$ this is $\frac{\sigma_i}{\sigma_j} V_i \cdot V_j = 0$
Step 3: I claim \( \{v_{1}, \ldots, v_{r}\} \) is a basis for \( \text{Row}(A) \)
- \( v_{i} = \frac{1}{\sigma_{i}} A^{T}Av_{i} = A^{T}(\frac{1}{\sigma_{i}} Av_{i}) \in \text{Col}(A^{T}) = \text{Row}(A) \)
- \( \dim \text{Row}(A) = r \) and \( \{v_{1}, \ldots, v_{r}\} \) is orthonormal
  \[ \Rightarrow \] linearly independent
  
  So the Basis Theorem \( \Rightarrow \) \( \text{Row}(A) = \text{Span}\{v_{1}, \ldots, v_{r}\} \)

Step 4: Verify \( A = \sum_{i=1}^{r} \sigma_{i} u_{i} v_{i}^{T} \).

Let \( B = \sum_{i=1}^{r} \sigma_{i} u_{i} v_{i}^{T} \), so we want to show \( A \preceq B \).

Recall \( A \preceq B \) if \( Ax = Bx \) for all \( x \in \mathbb{R}^{n} \).

As above,

\[
Bx = \sum_{i=1}^{r} \sigma_{i} (v_{i}^{T}x) u_{i} \quad \text{and} \quad Bv_{i} = \sum_{i=1}^{r} \sigma_{i} (v_{i}^{T}v_{i}) u_{i} = \sigma_{i} u_{i}
\]

(1) If \( x \in \text{Null}(A) \) then \( Ax = 0 \) and

\[
Bx = \sum_{i=1}^{r} \sigma_{i} (v_{i}^{T}x) u_{i} = 0 = Ax
\]

because \( v_{1}, \ldots, v_{r} \in \text{Row}(A) = \text{Null}(A)^{\perp} \)

(2) If \( x \in \text{Row}(A) \) then we can solve \( x = x_{1}v_{1} + \cdots + x_{r}v_{r} \) by Step 3. Then

\( Ax = A(x_{1}v_{1} + \cdots + x_{r}v_{r}) = x_{1}Av_{1} + \cdots + x_{r}Av_{r} \)

\( (u_{i} = \frac{1}{\sigma_{i}} Av_{i}) \) = \( x_{1}\sigma_{1}u_{1} + \cdots + x_{r}\sigma_{r}u_{r} \)
\[ Bx = B(x_1 v_1 + \ldots + x_r v_r) = x_1 Bv_1 + \ldots + x_r Bv_r \]
\[ (Bv_i = \sigma_i u_i) = x_1 \sigma_1 u_1 + \ldots + x_r \sigma_r u_r = Ax \]

(3) Any \( x \in \mathbb{R}^n \) has an orthogonal decomposition \( x = x_v + x_{v^*} \) \( x_v \in \text{Row}(A) \) \( x_{v^*} \in \text{Null}(A) \).
\[
\Rightarrow Ax = A(x_v + x_{v^*}) = A x_v + A x_{v^*}
\]
\[ = B x_v + B x_{v^*} = B(x_v + x_{v^*}) = Bx \]

**NB:** \( A^TA \) and \( AA^T \) have the same non-zero eigenvalues \( \sigma_1^2, \ldots, \sigma_r^2 \). (We showed in the proof that the other eigenvalues are \( = 0 \).)

→ What about the \( 0 \) eigenvalue?
→ What if \( A \) is a tall matrix with FCR?

**NB:** We showed in the proof that \( \{ u_1, \ldots, u_r \} \) is a basis for \( \text{Row}(A) \).
Replace \( A \) by \( A^T \) \( \Rightarrow \)
\( \{ u_1, \ldots, u_r \} \) is a basis for \( \text{Row}(A^T) = \text{Col}(A) \).
Mechanics of the SVD: Summary

A: an $m \times n$ matrix of rank $r$

SVD: $A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \ldots + \sigma_r u_r v_r^T$

$Ax = \sigma_1 (v_1 \cdot x) u_1 + \ldots + \sigma_r (v_r \cdot x) u_r$

$\sigma_1 \geq \ldots \geq \sigma_r > 0$: singular values

$\sigma_1^2 \geq \ldots \geq \sigma_r^2$: nonzero eigenvalues of $A^T A$ and $AA^T$

$\{u_1, \ldots, u_r\}$: left singular vectors

- orthonormal eigenvectors of $A A^T$
  \[ A A^T u_i = \sigma_i^2 u_i \]
  - orthonormal basis for $\text{Col}(A)$

$\{v_1, \ldots, v_r\}$: right singular vectors

- orthonormal eigenvectors of $A^T A$
  \[ A^T A v_i = \sigma_i^2 v_i \]
  - orthonormal basis for $\text{Row}(A)$

$A v_i = \sigma_i u_i \implies ||A v_i|| = \sigma_i$

SVD: $A^T = \sigma_1 v_1 u_1^T + \ldots + \sigma_r v_r u_r^T$

$A^T u_i = \sigma_i v_i \implies ||A^T u_i|| = \sigma_i$
This also gives us a procedure to compute the SVD. It is not the algorithm used in practice!

Efficient computation of the SVD is a difficult problem!

Naive Schoolbook Procedure to Compute the SVD:

Let $A$ be an $m \times n$ matrix of rank $r$.

1. Compute the nonzero eigenvalues of $A^T A$:
   \[ \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0 \]
   (where $\lambda_i$ appears multiple times if $A M > 1$)
   \( \rightarrow \) There are automatically $r$ of them, and they're positive.

2. Find an orthonormal eigenbasis for each eigenspace: get an orthonormal set $\mathbf{v}_1, \ldots, \mathbf{v}_r$ with $\mathbf{A}^T \mathbf{A} \mathbf{v}_i = \lambda_i \mathbf{v}_i$.

3. Set $s_i = \sqrt{\lambda_i}$; $\mathbf{u}_i = \frac{1}{s_i} \mathbf{A} \mathbf{v}_i$.

Then $\{ \mathbf{u}_1, \ldots, \mathbf{u}_r \}$ is orthonormal and
\[ A = s_1 \mathbf{u}_1 \mathbf{v}_1^T + s_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + s_r \mathbf{u}_r \mathbf{v}_r^T. \]

NB: It may be easier to compute SVD of $A^T$!

(if $A$ is wide: $m < n$, $A^T A$ is $n \times n$, but $A A^T$ is $m \times m$)
\[ \mathbf{A} = \begin{pmatrix} 3 & 0 \\ 4 & 5 \end{pmatrix} \]  
\text{NB: } r = 2 \ (2 \text{ pivots})

(1) \[ \mathbf{A}^T \mathbf{A} = \begin{pmatrix} 25 & 20 \\ 20 & 25 \end{pmatrix} \]  
\[ \rho \left( \mathbf{A}^T \mathbf{A} \right) = \lambda^2 - 50 \lambda + 225 = (\lambda - 45)(\lambda - 5) \]

\[ \lambda_1 = 45 \quad \lambda_2 = 5 \]

(2) Compute eigenspaces:

\[ \mathbf{A}^T \mathbf{A} - 45 \mathbf{I}_2 = \begin{pmatrix} -20 & 20 \\ 20 & -20 \end{pmatrix} \quad \Rightarrow \quad \mathbf{v}_1 = \frac{1}{\sqrt{2}} (1, 1) \]

\[ \mathbf{A}^T \mathbf{A} - 5 \mathbf{I}_2 = \begin{pmatrix} 15 & 20 \\ 20 & 15 \end{pmatrix} \quad \Rightarrow \quad \mathbf{v}_2 = \frac{1}{\sqrt{2}} (1, -1) \]

(3) \[ \sigma_1 = \sqrt{\lambda_1} = \sqrt{45} = 3 \sqrt{5} \quad \sigma_2 = \sqrt{\lambda_2} = \sqrt{5} \]

\[ \mathbf{u}_1 = \frac{1}{\sigma_1} \mathbf{A} \mathbf{v}_1 = \frac{1}{3 \sqrt{5}} \begin{pmatrix} 3 \\ 4 \end{pmatrix} \frac{1}{\sqrt{2}} (1, 1) = \frac{1}{3 \sqrt{10}} (3, -2, -2) \]

\[ \mathbf{u}_2 = \frac{1}{\sigma_2} \mathbf{A} \mathbf{v}_2 = \frac{1}{\sqrt{5}} \begin{pmatrix} 3 \\ 4 \end{pmatrix} \frac{1}{\sqrt{2}} (1, -1) = \frac{1}{\sqrt{10}} (3, 2, 2) \]

\text{SVD:}

\[ \begin{pmatrix} 3 & 0 \\ 4 & 5 \end{pmatrix} = 3 \sqrt{5} \cdot \frac{1}{\sqrt{10}} (3, 1) \cdot \frac{1}{\sqrt{2}} (1, 1) + \sqrt{5} \cdot \frac{1}{\sqrt{10}} (3, 1) \cdot \frac{1}{\sqrt{2}} (-1, -1) \]

\text{Check: } \| \mathbf{u}_1 \| = \frac{1}{\sqrt{10}} \sqrt{3^2 + 3^2} = 1 \quad \| \mathbf{u}_2 \| = \frac{1}{\sqrt{10}} \sqrt{3^2 + 2^2} = 1 \quad \mathbf{u}_1 \cdot \mathbf{u}_2 = 0 \quad \checkmark