Lecture 38: The Singular Value Decomposition

Let $A$ be an $m \times n$ matrix, $T : \mathbb{R}^n \to \mathbb{R}^m$ the linear transformation $\vec{x} \to A\vec{x}$. Recall the picture

\[ \mathbb{R}^n \xrightarrow{T} \mathbb{R}^m \]

\[ \text{Col}(A^T) = \text{Row}(A) \]

\[ \text{Nul}(A) \xrightarrow{T} \mathbb{R}^n \]

in which

- $T$ restricts to an isomorphism $\text{Row}(A) \cong \text{Col}(A)$, and so they have the same dimension $r$.
- $\text{Col}(A) \perp \text{Nul}(A^T)$ and $\text{Col}(A^T) \perp \text{Nul}(A)$

Null space of $A$ is vectors $\perp$ to its rows, i.e., columns of $A^T$.

- $\text{Col}(A) \cap \text{Nul}(A^T)$ span $\mathbb{R}^m$.
- $\text{Col}(A^T) \cap \text{Nul}(A)$ span $\mathbb{R}^n$.

Recovers Rank-Nullity:

\[ \dim \text{Nul}(A^T) = m - r \]
\[ \dim \text{Nul}(A) = n - r \]

If $\vec{b} \in \text{Col}(A)$, then there is a unique solution to $A\vec{x} = \vec{b}$ in $\text{Col}(A^T)$, and many more if we add an arbitrary vector in
If $\hat{b} \notin \text{Col}(A)$, then $A\hat{x} = \hat{b}$ is inconsistent and the best we can do is minimize $A\hat{x} - \hat{b}$, which is done by a least-squares solution $\hat{x}$: error $\hat{e}$.

To minimize $\hat{e}$, we take $\hat{b} = \text{proj}_\text{Col}(A) \hat{b}$ and solve $A\hat{x} = \hat{b}$ for $\hat{x}$; then $\hat{e} = \hat{b} - A\hat{x} = \hat{b} - \hat{b} = 0 \in \text{Col}(A^\perp) = \text{Null}(A^T)$. And so $A^T(\hat{b} - A\hat{x}) = 0$ recovers the normal equations.
\[ A^T A \mathbf{x} = A^T \mathbf{b}. \]

Now let's talk about bases. \( A^T A \) is symmetric, and also positive-semidefinite, since (for any \( \mathbf{x} \in \mathbb{R}^n \))
\[ \mathbf{x}^T A^T A \mathbf{x} = \mathbf{x}^T A \mathbf{x} = \| A \mathbf{x} \|^2 \geq 0. \]
So \( A^T A \) has an orthonormal eigenspace \( \mathbf{B} = \{ \mathbf{e}_1, ..., \mathbf{e}_n \} \) with all \( \lambda_i \geq 0 \). The 0-eigenspace \( \mathbf{E}_0 \) is \( \text{Nul}(A^T A) = \text{Nul}(A^T) \) (since \( A^T A \mathbf{x} = 0 \Rightarrow \mathbf{x}^T A^T A \mathbf{x} = 0 \Rightarrow \| A \mathbf{x} \|^2 = 0 \Rightarrow A \mathbf{x} = 0 \)), and the other eigenspaces are orthogonal to \( \mathbf{E}_0 \) hence contained in \( \text{Row}(A) \). This means (after reordering) we have
\[ \{ \mathbf{v}_1, ..., \mathbf{v}_r \} \subset \text{Row}(A) \quad \text{both (o.n.) bases}, \]
\[ \{ \mathbf{v}_{r+1}, ..., \mathbf{v}_n \} \subset \text{Nul}(A) \]
while \( \lambda_1, ..., \lambda_r > 0 \) and \( \lambda_{r+1} = \cdots = \lambda_n = 0 \).

**Definition:** The singular values of \( A \) are \( \sigma_1, ..., \sigma_r \), where \( \sigma_i := \sqrt{\lambda_i} \). (We'll also write \( \sigma_r = \cdots = \sigma_n = 0 \).)

So \( A^T A \mathbf{v}_i = \sigma_i^2 \mathbf{v}_i \) \( (1 \leq i \leq r) \Rightarrow \mathbf{v}_i^T A^T A \mathbf{v}_i = \sigma_i^2 \mathbf{v}_i^T \mathbf{v}_i \Rightarrow \| A \mathbf{v}_i \|^2 = \sigma_i^2 \| \mathbf{v}_i \|^2 = \sigma_i^2 \) (since \( \mathbf{v}_i \) is normalized) \( \Rightarrow \mathbf{w}_i := \frac{1}{\sigma_i} A \mathbf{v}_i \) has unit length \( \Rightarrow \)
\( \{ \mathbf{w}_1, ..., \mathbf{w}_r \} \subset \text{Col}(A) \) is an o.n. basis.

Letting \( \{ \mathbf{u}_{r+1}, ..., \mathbf{u}_n \} \subset \text{Nul}(A) \) be an o.n. basis,
$C := \{\hat{e}_1, \ldots, \hat{e}_m\}$ is an orthonormal basis of $\mathbb{R}^n$.

Now define matrices as follows:

- $U = \begin{pmatrix} \hat{e}_1 & \cdots & \hat{e}_m \end{pmatrix}$ is an $n \times m$ orthogonal matrix.
- $V = \begin{pmatrix} \hat{v}_1 & \cdots & \hat{v}_n \end{pmatrix}$ is an $n \times n$ orthogonal matrix.
- $\Sigma = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$ is an $m \times n$ matrix.

We have then

$$U \Sigma = \begin{pmatrix} \hat{e}_1 & \cdots & \hat{e}_m \end{pmatrix} \begin{pmatrix} \hat{v}_1 & \cdots & \hat{v}_n \end{pmatrix}^T = \begin{pmatrix} \hat{v}_1 & \cdots & \hat{v}_n \end{pmatrix} \begin{pmatrix} \hat{e}_1 & \cdots & \hat{e}_m \end{pmatrix} = A \Sigma V^T.$$

and multiplying by $V^T$ on the right gives $U \Sigma V^T = A V V^T$.

But $V V^T = I_n$, since $V$ is orthogonal, yielding the

**Singular Value Decomposition (SVD)** $A = U \Sigma V^T$

When $A$ is itself square and symmetric, we can take $U = V = P$ and $\Sigma = D$ — i.e. $A = P D P^T$ is the decomposition. But the SVD extends this to matrices which might not be symmetric, or diagonalizable, or even square (since we didn't require $m = n$). **And the $\delta_i$ are its eigenvalues**.
The SVD has been called the "fundamental theorem of matrix algebra". Its efficient computer implementation for large matrices has been the subject of countless articles in numerical analysis. While these algorithms are too complicated to describe here, they are more efficient (and accurate) than the orthogonal diagonalization of $A^*A$, which are implemented in things like MATLAB.

**Application #1: Principal Component Analysis**

The SVD reads

$$A = \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_m \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{pmatrix} \begin{pmatrix} \mathbf{v}_1 \cdots \mathbf{v}_n \end{pmatrix} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T.$$  

Let's say $A$ is a big fat matrix of data. For instance, in a country of $n = 10,000,000$ internet users, you want to target ads (or fake news) based on online behavior, where the # of possible clicks or purchases is also
very large. Assume however that only $k$ (cultural, demographic, etc.)
attributes of a person generally determine this behavior, that is, there exists an $n \times k$ “attribute
matrix” and a $k \times m$ “behavior matrix” whose product is
roughly $A$, the $n \times m$ matrix of raw user data. (In
other words, we expect $A$ to be well-approximated by
a rank $k$ matrix, with $k$ much smaller than $n$ or $m$.)
Alternatively, $A$ might be a matrix of numbers controlling
brightness of pixels in an image. Either way, it will
typically be the case that most of the $\sigma_i$ are very small
compared to the first few. So we get

$$A \approx \sigma_1 \vec{v}_1 \vec{v}_1^T + \ldots + \sigma_r \vec{v}_r \vec{v}_r^T$$

when $k$ is maybe 5. Then instead of $1,000,000^2$ numbers,
you only need $2 \times 5 = 1,000,000 + 5$ numbers. And this will
typically be a very good approximation, much better than
Faurier analysis would give.

Application #2: Least squares and the “pseudo-inverse”

Define a transformation

$$T^* : \mathbb{R}^m \rightarrow \mathbb{R}^n$$

by sending

$$\vec{v}_i \rightarrow \frac{1}{\sigma_i} \vec{u}_i \quad (i=1, \ldots, r)$$
$$\vec{v}_i \rightarrow 0 \quad (i=r+1, \ldots, n).$$
It's restriction to $\text{Col}(A)$ inverts the restriction of $T$ to $\text{Row}(A)$, since $T$ sends $\vec{v}_i \mapsto \sigma_i \vec{v}_i$ ($i = 1, \ldots, r$):

$$\begin{array}{ccc}
\mathbb{R}^n & \xrightarrow{T} & \text{Row}(A) \\
\text{Null}(A) & \xrightarrow{T} & \text{Col}(A) \\
\text{Null}(A^T) & \xleftarrow{T^*} & \mathbb{R}_r^n
\end{array}$$

The matrix of this pseudo-inverse is

$$A^* := V \tilde{\Sigma} U^T$$

where $\tilde{\Sigma} := \begin{pmatrix} \tilde{\sigma} & 0 \\ 0 & 0 \end{pmatrix}$ is $n \times n$. (Why?)

**Claim:** $\hat{x} := A^* \hat{b}$ gives the unique least-squares solution to $Ax = \hat{b}$ in $\text{Row}(A)$ (called the "minimally least-squares solution").

**Why does it work?** In this case, a picture is worth a thousand math symbols:

...
That is, \((\mathbf{a} = \mathbf{A}^\top \mathbf{b}) = \mathbf{A}^\top \mathbf{b}\) and \(\mathbf{A}^\top \mathbf{b}\) maps \(\mathbf{A}\) on \(\text{Col}(\mathbf{A})\).

\[
\mathbf{A} \hat{\mathbf{x}} = \hat{\mathbf{b}}.
\]

Let's finish with a numerical example:

\[
\mathbf{A} = \begin{pmatrix}
1 & -1 \\
1 & -1 \\
2 & -2
\end{pmatrix}
\]

The SVD is

\[
\mathbf{A} = \begin{pmatrix}
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}}
\end{pmatrix}
\begin{pmatrix}
\sqrt{12} & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}}
\end{pmatrix}
\]

And so the pseudo-inverse is

\[
\mathbf{A}^+ = \begin{pmatrix}
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}}
\end{pmatrix}
\begin{pmatrix}
\sqrt{12} & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}}
\end{pmatrix}
\]
\[ A^* = \begin{pmatrix}
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
-\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\
\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{3}}
\end{pmatrix}
\]

\[ = \frac{1}{\sqrt{2}} \begin{pmatrix}
1 & 1 & 2 \\
-1 & -1 & -2
\end{pmatrix} \]

So the minimal least-squares solution to \( Ax = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \) is

\[ \hat{x} = A^* \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \end{pmatrix} \]

Doesn't seem like the easiest way to do this example (the normal equations), but it is numerically far superior for large matrices.

Friday we'll review for the exam.