# 1 7th of November 2018 — F. Poloni

## 1.1 Least squares problem with SVD

**Do you recall?**

**Tall thin SVD:** A matrix $A$ can be written as $A = USV^T$, where $U$ is orthogonal, $S$ is a diagonal matrix and $V$ is orthogonal as well. In the case of a tall, thin $A$ the decomposition has the following shape:

$$\begin{pmatrix} & & & \\ U^1 & U^2 & \cdots & U^m \\ & & & \end{pmatrix} \Sigma \begin{pmatrix} & & & \\ V^1 & V^2 & \cdots & V^n \\ & & & \end{pmatrix}^T$$

where

$$\Sigma = \begin{pmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ & & & & & \sigma_n \end{pmatrix}$$

And if we denote $U_1$ the matrix obtained as the first $n$ columns of $U$ we have the tall, thin SVD: $A = U_1 \Sigma V^T$

We would like to see how we can solve a least squares problem through $SVD$:

$$
\begin{aligned}
\|Ax-b\| &= \left\|USV^Tx-b\right\| && \leftarrow \text{ Def. of SVD} \\
&= \left\|U^T(USV^Tx-b)\right\| && \leftarrow U^T\text{is orthogonal} \\
&= \left\|SV^Tx-U^Tb\right\| && \leftarrow \text{Distributivity + orthogonality} \\
&= \left\|Sy-U^Tb\right\| && \leftarrow y=V^Tx \\
&= \left\| \begin{pmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ & & & & \\ & & & & \\ & & 0 & & \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} - \begin{pmatrix} {U^1}^T b \\ {U^2}^T b \\ \vdots \\ {U^n}^T b \\ {U^{n+1}}^T b \\ \vdots \\ {U^m}^T b \end{pmatrix} \right\| \\
&= \left\| \begin{pmatrix} \sigma_1 y_1 - {U^1}^T b \\ \sigma_2 y_2 - {U^2}^T b \\ \vdots \\ \sigma_n y_n - {U^n}^T b \\ \sigma_{n+1} y_{n+1} - {U^{n+1}}^T b \\ \vdots \\ \sigma_m y_m - {U^m}^T b \end{pmatrix} \right\|
\end{aligned}
\tag{1.1}
$$

Where the first $n$ rows may be assigned to 0 iff $y_i = -\frac{{U^i}^T b}{\sigma_1}$ (if $\sigma_i \neq 0 \forall i$), while the latter $m-n$ do not depend on $y$. This process produces a solution $y$, but the variable change may

be inverted, so

$$
\begin{aligned}
x &= Vy && \leftarrow \text{Orthogonality of } V \\
&= V^1 y_1 + V^2 y_2 + \cdots + V^n y_n \\
&= V^1 \frac{1}{\sigma_1} U^{1T} b + V^2 \frac{1}{\sigma_2} U^{2T} b + \cdots + V^n \frac{1}{\sigma_n} U^{nT} b \\
&= V \begin{pmatrix} \frac{1}{\sigma_1} & & & & \\ & \frac{1}{\sigma_2} & & & \\ & & \ddots & & 0 \\ & & & \ddots & \\ & & & & \frac{1}{\sigma_n} \end{pmatrix} U^T b \\
&= V \begin{pmatrix} \frac{1}{\sigma_1} & & & & \\ & \frac{1}{\sigma_2} & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \frac{1}{\sigma_n} \end{pmatrix} {U_1}^T b
\end{aligned}
\tag{1.2}
$$

Which depends only on the tall, thin SVD.

**Fact 1.1.** *The $\sigma_i$ are different from 0 iff $A$ has full column rank.*

*Proof.* $A$ has full column rank

$$\Updownarrow$$

$A^T A$ is invertible

$$\Updownarrow$$

$$(USV^T)^T (USV^T) = VS^T U^T USV^T = V \begin{pmatrix} \sigma_1^2 & & & & \\ & \sigma_2^2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \sigma_n^2 \end{pmatrix} V^T \text{ is invertible}$$

$$\Updownarrow$$

$\forall i \ \sigma_i \neq 0$

□

**Observation 1.1.** *This lemma also proves that the factorization is also a $QR$ factorization.*

**Note on Matlab syntax**

`svd(A, 0)` and `qr(A, 0)` express that we are only interested in the parts of the factorization without zeros, in case of a tall, thin matrix $A$.

We may observe that the computational complexity is $O(15n^3)$ for square matrices, while it's $O(mn^2)$ in the tall, thin case.

### 1.1.1 Behaviour in case of zeros as singular values

What happens when there are some zeros as singular values?

**Do you recall?**

We may recall that the singular values are ordered on the diagonal in decreasing order (the largerst in top left position). From this assumption, we may say that if there are some $\sigma_i = 0$ then they are in the bottom right part of the matrix.

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0$$

We also recall the following $\|Ax - b\| = \left\| \begin{pmatrix} \sigma_1 y_1 - U^{1T} b \\ \sigma_2 y_2 - U^{2T} b \\ \vdots \\ \sigma_n y_n - U^{nT} b \\ 0 \cdot y_{n+1} - U^{n+1T} b \\ \vdots \\ 0 \cdot y_m - U^{mT} b \end{pmatrix} \right\|$.

No matter what value we choose for $y_{r+1} \cdots y_n$, the value doesn't change since it's multiplied by 0. Therefore we get infinite solutions of the form $y = \begin{pmatrix} -\frac{U^{1T} b}{\sigma_1} \\ -\frac{U^{2T} b}{\sigma_2} \\ \vdots \\ -\frac{U^{rT} b}{\sigma_r} \\ * \end{pmatrix}$

We would like to make the solution unique, so we can modify the problem:

- taking the value that minimize the norm:

$$\min_{x \in \arg\min(\|Ax-b\|)} \|x\| . \tag{P2}$$

  Note that $\|x\| = \|y\|$, because $x = Vy$. It follows from the expression of $y$ that the choice that minimizes its norm is $y_{r+1} = \cdots = y_n = 0$.

-

The solution of $P2$ is given by

$$Vy = \begin{pmatrix} V^1 & V^2 & \cdots & V^n \end{pmatrix} \begin{pmatrix} -\frac{U^{1^T}b}{\sigma_1} \\ -\frac{U^{2^T}b}{\sigma_2} \\ \vdots \\ -\frac{U^{r^T}b}{\sigma_r} \\ 0 \end{pmatrix} = V^{1^T}\frac{1}{\sigma_1}U^{1^T}b + \cdots + V^{r^T}\frac{1}{\sigma_r}U^{r^T}b$$

What happens when working with machine precision? Let's make an example where $r = n-1$, so only the last singular value is 0. If the check of $\sigma_n = 0$ fails, $\frac{1}{\sigma_n}$ becomes very big. A way to circumvent this problem is to find the linear dependencies between the columns, so that the algorithm works correctly.

## 1.2 Truncated SVD

In many real world setups first singular components correspond to the most prominent features of the dataset, while the smallest ones are fine details and noise. Note, though, that in the sum $\sum_{i=1}^{n} V^i \frac{U^{i^T}b}{\sigma_i}$ the small singular values may have a large impact, because $\sigma_i$ is in the denominator.

We can modify the solution to cope with real world data problems:

$x = \sum_{i=1}^{n} V^i \frac{U^{i^T}b}{\sigma_i} \quad \longrightarrow \quad x_{trunc} = \sum_{i=1}^{k} V^i \frac{U^{i^T}b}{\sigma_i}$

for a certain $k$, ignoring small singular values.

Another way to modify the problem is the following.

## 1.3 Tikhonov regularization / ridge regression

The Tikhonov regularization is a smoother version of truncated SVD.

$$x_{Tik} = \arg\min_{x \in \mathbb{R}^n} \|Ax - b\|^2 + \alpha^2 \|x\|^2$$

**Fact 1.2.** *The Tikhonov regularization is equivalent to*

$$x_{Tik} = \arg\min_{x \in \mathbb{R}^n} \left\| \begin{pmatrix} A \\ \alpha I \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|^2 \tag{1.3}$$

We show that the two objective functions coincide.

*Proof.*

$$\begin{aligned}\left\|\begin{pmatrix}A\\ \alpha I\end{pmatrix}x-\begin{pmatrix}b\\ 0\end{pmatrix}\right\|^2 &= \left\|\begin{pmatrix}Ax\\ \alpha x\end{pmatrix}-\begin{pmatrix}b\\ 0\end{pmatrix}\right\|^2\\ &= \left\|\begin{pmatrix}Ax-b\\ \alpha x\end{pmatrix}\right\|^2\\ &= \|Ax-b\|^2+\|\alpha x\|^2\\ &= \|Ax-b\|^2+\alpha^2\|x\|^2\end{aligned} \tag{1.4}$$

□

**Fact 1.3.** *The solution of the Tikhonov regularization is given by the formula* $x_{Tik} = (A^TA+\alpha^2 I)^{-1}A^Tb$.

*Proof.* We start by writing the explicit solution of Equation (1.3) using the pseudoinverse

$$\begin{aligned}\begin{pmatrix}A\\ \alpha I\end{pmatrix}^+\begin{pmatrix}b\\ 0\end{pmatrix} &= \left(\begin{pmatrix}A\\ \alpha I\end{pmatrix}^T\begin{pmatrix}A\\ \alpha I\end{pmatrix}\right)^{-1}\begin{pmatrix}A\\ \alpha I\end{pmatrix}^T\begin{pmatrix}b\\ 0\end{pmatrix}\\ &= \left(\begin{pmatrix}A^T & \alpha I\end{pmatrix}\begin{pmatrix}A\\ \alpha I\end{pmatrix}\right)^{-1}\begin{pmatrix}A^T & \alpha I\end{pmatrix}\begin{pmatrix}b\\ 0\end{pmatrix}\\ &= (A^TA+\alpha^2 I)^{-1}A^Tb.\end{aligned} \tag{1.5}$$

□

**Fact 1.4.** *We can observe that* $(A^TA+\alpha^2 I)$ *is positive definite.*

*Proof.* $z^T\cdot(A^TA+\alpha^2 I)\cdot z = z^TA^TAz+\alpha^2 z^Tz \overset{(1)}{=} \alpha^2 z^Tz = \alpha^2\|x\|^2 > 0$. The equality (1) is obtained because $z^TA^TAz \geq 0$, since $A^TA$ is positive semidefinite. □

**Exercise 1.1.** *Show using the SVD of A that the Tikhonov / Ridge solution can be written as*

$$x = \sum_{i=1}^{n} V^i \frac{\sigma_i}{\sigma_i^2+\alpha^2} U^{iT} b.$$

*When* $\sigma_i \gg \alpha$, $\frac{\sigma_i}{\sigma_i^2+\alpha^2} \approx \frac{1}{\sigma_i}$*: similar to the 'true' solution.*
*When* $\sigma_i \ll \alpha$, $\frac{\sigma_i}{\sigma_i^2+\alpha^2} \approx \frac{\sigma_i}{\alpha^2} \approx 0$*: approximately ignoring small singular values.*

How can we choose $k$?

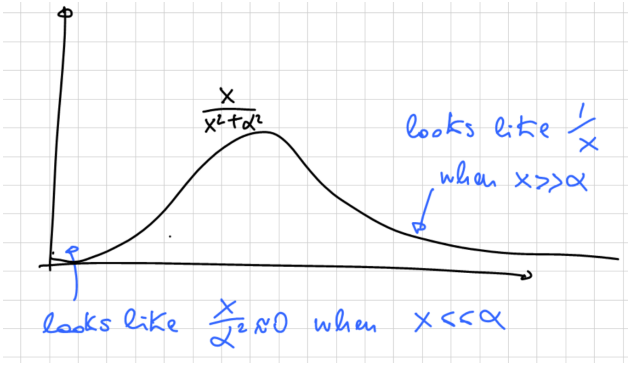


FIGURE 1.1: Here is the shape of the formula for the singular values.