Least squares with the SVD

One can solve least-squares problem also with (thin) $A = USV^T$.
Same derivation as with QR:

$$\|Ax - b\| = \|USV^T x - b\| = \|S \underbrace{V^T x}_{=y} - U^T b\|$$

$$= \left\| \begin{bmatrix} \sigma_1 y_1 \\ \sigma_2 y_2 \\ \vdots \\ \sigma_n y_n \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} u_1^T b \\ u_2^T b \\ \vdots \\ u_n^T b \\ u_{n+1}^T b \\ \vdots \\ u_m^T b \end{bmatrix} \right\|$$

If all the σ_n are different from 0, the minimum is when $y_i = \frac{u_i^T b}{\sigma_i}$ (and then $x = Vy$).

$$\|Ax-b\| = \|USV^Tx - b\| = \|U^T(USV^Tx-b)\| =$$

$$= \|S\underbrace{V^Tx}_{y} - U^Tb\| = \|Sy - U^Tb\| = \left\| \begin{bmatrix} \sigma_1 y_1 \\ \sigma_2 y_2 \\ \vdots \\ \sigma_n y_n \\ \hline 0 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} u_1^T b \\ u_2^T b \\ \vdots \\ \hline \vdots \\ u_m^T b \end{bmatrix} \right\| =$$

$$Sy = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & 0 & \end{pmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\left(U = \begin{bmatrix} u_1 & u_2 & \cdots & u_m \end{bmatrix} \right)$$

first n entries: $\sigma_i y_i - u_i^T b$

last m-n entries: $-u_i^T b$

If we choose $y_i = \frac{1}{\sigma_i} u^T b$, the first n entries become 0 $\quad i=1,\dots,n$

$$\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \hline -u_{n+1}^T b \\ \vdots \\ -u_m^T b \end{bmatrix}$$

$$y = \begin{bmatrix} \frac{1}{\sigma_1} u_1^T b \\ \vdots \\ \frac{1}{\sigma_n} u_n^T b \end{bmatrix}$$

$$V^T x = y \iff \underbrace{V V^T}_{I} x = V y$$

$$x = V y =$$

$$x = v_1 y_1 + v_2 y_2 + \dots + v_n y_n =$$

$$= v_1 \frac{1}{\sigma_1} u_1^T b + v_2 \frac{1}{\sigma_2} u_2^T b + \dots + \frac{1}{\sigma_n} u_n^T b$$

$$V = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}$$

usually this term is larger, because $\sigma_n \leq \sigma_{n-1} \leq \dots \leq \sigma_1$

(unless $u_n^T b$ is small)

Least squares with the SVD

Putting everything together, one gets → need only $u_1, u_2, \ldots u_n$

$$x = \sum_{i=1}^{n} v_i \frac{u_i^T b}{\sigma_i}.$$

Again, we need only the thin SVD to compute it.

Note that the small σ_i's contribute more to the solution (unless also $u_i^T b \approx 0$).

$$A = \Big[\underbrace{U_1}_{n} \Big| \underbrace{U_2}_{m-n}\Big] \begin{bmatrix} S_2 \\ \hline 0 \end{bmatrix}\begin{matrix} \}\, n \\ \}\, m-n \end{matrix} \big[\, V \,\big]^T = U_1 S_2 V^T$$

Full rank and the SVD

A^TA singular $\iff$ one or more σ_i is zero
↓

Question: when are all $\sigma_i \neq 0$? Note that

$$A^TA = (USV^T)^T(USV^T) = VS^TSV^T = V \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_n^2 \end{bmatrix} V^T,$$

hence A has full column rank $\iff$ A^TA is invertible $\iff$ $\sigma_i \neq 0$ for all i.

(Also, you may recall that we said that $r = \text{rank}(A)$ is the number of nonzero $\sigma_i \ldots$).

$\text{rk}\,A = n \iff$ all n $\sigma_1, \sigma_2, \ldots \sigma_n$ are nonzero

$$A^TA = (USV^T)^T USV^T = VS^T \underbrace{U^TU}_{I} SV^T$$

$$= V \left[\begin{array}{cccc|c} \sigma_1 & & & & \\ & \sigma_2 & & & 0 \\ & & \ddots & & \\ & & & \sigma_n & \end{array}\right] \left[\begin{array}{ccc} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ \hline & 0 & \end{array}\right] V^T =$$

$$= V \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{bmatrix} V^T \leftarrow \text{eigendecomposition of } A^TA$$

Remarkable fact: from the SVD, we get for free also the eigendecomposition of A^TA

Zero singular values

What happens if $r < n$, i.e., $\sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_n = 0$?
Go back to the computation in the first slide: in those rows we get $u_i^T b$, independent of y_i. All choices of y_i are valid solutions (minima).

(Recall: the solution was given by the minimum of a quadratic function $x^T A^T A x + \ldots$. $A^T A$ is only positive semidefinite, so the solution is not unique.)

"But I want to return **one** solution": a possibility is taking $y_i = 0$ when $\sigma_i = 0$. This gives the solution with minimum $\|y\| = \|x\|$:

$$\min_{x \text{ s.t. } \|Ax-b\| \text{ is minimized}} \|x\|.$$

Most of the time, though, this means "go back and check your model". For instance:

$$(\text{salary}) \approx (\text{rebounds})x_1 + (\text{fouls})x_2 + (\text{points})x_3 + (\text{points+rebounds})x_4.$$

$$\begin{bmatrix} \sigma_1 y_1 \\ \vdots \\ \sigma_n y_n \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} u_1^T b \\ \vdots \\ u_n^T b \\ u_{n+1}^T b \\ \vdots \\ u_m^T b \end{bmatrix}$$

$\leftarrow$ if $\sigma_n = 0$, this entry contains $-u_n^T b$ regardless of y_n

$$\min \|Ax - b\|^2 = \min \underbrace{x^T A^T A x} + \dots$$

Hessian: $2A^TA$

Example

```
>> M = dlmread('salaries.csv', ',', 1, 1);
>> A = M(:, 1:3); A(:);
>> A(:,4) = A(:,1) + A(:,3);
>> svd(A)
ans =
   2.8060e+04
   3.2171e+03
   8.7262e+02
   1.5007e-12
>> rank(A'*A)
ans =
     3
```

Eigenvalues and singular values

```
>> eig(A'*A)
ans =
   5.7662e-08
   7.6146e+05
   1.0350e+07
   7.8736e+08
>> svd(A).^2
ans =
   7.8736e+08
   1.0350e+07
   7.6146e+05
   2.2520e-24
```

(Why is the smallest eigenvalue different between the two computations? Which one is the most accurate?)

```
>> A \ b
Warning: Rank deficient, rank = 3, tol = 1.956415e-09.
ans =
   3.7690e+03
  -2.6578e+04
            0
   9.5162e+03
```

Note that the formula to find this solution is

$$x = \sum_{i=1}^{r} v_i \frac{u_i^T b}{\sigma_i}.$$

Small singular values

A different, related issue is the one of small singular values. More frequently than exact zeros, we will encounter small singular values, e.g.,

```
ans =
   1.9307e-04
   3.7276e-03
   7.1969e-02
   1.3895e+00
   2.6827e+01
   5.1795e+02
```

One of the reasons is noisy data: we will see more in future (if we have time), but essentially a perturbation of norm δ may "move" each singular value by an amount δ.
Example in the next slide.

```
>> S1 = svd(A)
S1 =
   2.8060e+04
   3.2171e+03
   8.7262e+02
   1.5007e-12
>> S2 = svd(A + 0.01*rand(size(A)))
S2 =
   2.8060e+04
   3.2172e+03
   8.7264e+02
   6.7068e-02
>> S2 - S1
ans =
   1.3315e-01
   6.2690e-03
   2.6129e-02
   6.7068e-02
```

Truncated SVD

Also, in many applications the 'most meaningful' features correspond to the largest singular values (recall: eigenfaces, image compression).
Unfortunately, when solving least-squares problems small singular values count more, not less: recall

$$x = \sum_{i=1}^{n} v_i \frac{u_i^T b}{\sigma_i}.$$

In some contexts, it makes sense to ignore the contribution of small singular values:

$$x_{reg} = \sum_{i=1}^{r} v_i \frac{u_i^T b}{\sigma_i}.$$

Example (not the best one)

```
>> AA = A + 0.01*rand(size(A));
>> AA \ b
ans =
   9.1286e+07
  -2.9669e+04
   9.1282e+07
  -9.1272e+07
>> [U, S, V] = svd(AA);
>> V(:,1:3) / S(1:3, 1:3) * U(:, 1:3)'*b
ans =
   5.6843e+03
  -2.6577e+04
   1.9155e+03
   7.6007e+03
```

Better (in a suitable sense) approximation of the 'true' solution `A \ b`.

Alternative: Tikhonov regularization / ridge regression

A different solution to the problem of 'what to do when there are tiny singular values': find

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

(for some $\alpha > 0$). "Discourages" solutions with large norm. Some similar strategies used in optimization.
It can be rewritten as

$$\min_{x \in \mathbb{R}^n} \left\| \begin{bmatrix} A \\ \alpha I \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|^2 .$$

Solution: $\begin{bmatrix} A \\ \alpha I \end{bmatrix}^+ \begin{bmatrix} b \\ 0 \end{bmatrix}$.

Tikhonov / ridge — formula

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

Note that $z^T(A^T A + \alpha^2 I)z \geq \alpha^2 z^T z > 0$ for all $z \neq 0 \implies \begin{bmatrix} A \\ \alpha I \end{bmatrix}$ has full column rank.

Tikhonov / ridge and SVD

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

$$x = \sum_{i=1}^{n} v_i \frac{\sigma_i}{\sigma_i^2 + \alpha^2} u_i^T b.$$

When $\sigma_i \gg \alpha$, $\frac{\sigma_i}{\sigma_i^2+\alpha^2} \approx \frac{1}{\sigma_i}$: similar to the 'true' solution.

When $\sigma_i \ll \alpha$, $\frac{\sigma_i}{\sigma_i^2+\alpha^2} \approx \frac{\sigma_i}{\alpha} \approx 0$: approximately ignoring small singular values.

Choice of α

How to choose α? Here it is difficult to motivate a mathematically sound solution — we are discussing "how to modify the problem", not "how to solve the problem".

Often, $\alpha \approx$ amount of 'noise' in the data (when it is known). Otherwise, there are application-specific strategies — you will see more in ML / AI courses.

Exercises

1. Let $A \in \mathbb{R}^{m\times n}$, $m \geq n$, be a matrix with full column rank, and let $A = U\Sigma V^T$ be its SVD, with $\sigma_i = (\Sigma)_{ii}$ as usual. Show that $A^+ = V\Sigma^+ U^T$, where Σ^+ is the $n \times m$ matrix such that

$$\Sigma^+ = \begin{bmatrix} \frac{1}{\sigma_1} & & & & \\ & \frac{1}{\sigma_2} & & & \\ & & \ddots & & \\ & & & \frac{1}{\sigma_n} & \end{bmatrix}.$$

   (As usual, elements not shown are zeros). Hint: use $A^+ = (A^T A)^{-1} A^T$.
2. Could one have obtained the same result also from the formula at the top of Slide 2?
3. Show that the matrix denoted with Σ^+ above is, indeed, the pseudoinverse of Σ.