Least squares with the SVD

One can solve least-squares problem also with the (thin) SVD.
Same derivation as with QR:

$$\|A\mathbf{x}-\mathbf{y}\| = \|USV^T\mathbf{x}-\mathbf{y}\| = \|S\underbrace{V^T\mathbf{x}}_{=\mathbf{z}} - U^T\mathbf{y}\|$$

$$= \left\| \begin{bmatrix} \sigma_1 z_1 \\ \sigma_2 z_2 \\ \vdots \\ \sigma_n z_n \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} \mathbf{u}_1^T\mathbf{y} \\ \mathbf{u}_2^T\mathbf{y} \\ \vdots \\ \mathbf{u}_n^T\mathbf{y} \\ \mathbf{u}_{n+1}^T\mathbf{y} \\ \vdots \\ \mathbf{u}_m^T\mathbf{y} \end{bmatrix} \right\|$$

If all the σ_n are different from 0, the minimum is when $z_i = \frac{\mathbf{u}_i^T\mathbf{y}}{\sigma_i}$.
Then $\mathbf{x} = V\mathbf{z} = V\Sigma_0^{-1}U_0^T\mathbf{y}$. The minimum value is $U_c^T\mathbf{y}$.

Least squares with the SVD

Putting everything together, one gets

$$\mathbf{x} = \sum_{i=1}^{n} \mathbf{v}_i \frac{\mathbf{u}_i^T \mathbf{y}}{\sigma_i} = V \begin{bmatrix} \frac{1}{\sigma_1} & & & \\ & \frac{1}{\sigma_2} & & \\ & & \ddots & \\ & & & \frac{1}{\sigma_n} \end{bmatrix} U^T \mathbf{y}.$$

Note that the small σ_i's contribute more to the solution (unless also $\mathbf{u}_i^T \mathbf{y} \approx 0$).

The expression in red gives a formula for A^+ in terms of the SVD. Note that we need only the thin SVD to compute it: $A^+ = V \Sigma_0^{-1} U_0^T$.

Full rank and the SVD

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

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

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

(Also, you may recall that $r = \text{rank}(A)$ is the number of nonzero singular values).

Zero singular values

What happens if $r < n$, i.e., $\sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_n = 0$?
From the first slide: in those rows we get $-\mathbf{u}_i^T \mathbf{y}$, independent of z_i.
All choices of z_i are valid solutions (minima).

(Recall: $A^T A$ is only positive semidefinite, so the quadratic function is not strongly convex and the minimizer is not unique.)

"But I want **one** solution": a possibility is taking $z_i = 0$ when $\sigma_i = 0$. This gives the solution with minimum norm $\|\mathbf{z}\| = \|\mathbf{x}_*\|$:

$$\mathbf{x}_* = \arg \min_{\mathbf{x} \in \arg\min \|A\mathbf{x} - \mathbf{y}\|} \|\mathbf{x}\|.$$

Essentially, this means replacing $\frac{1}{\sigma_i}$ with 0 in the previous formulas whenever $\sigma_i = 0$.

The definition of pseudoinverse can be extended to the case of a rank-deficient A, with $\mathbf{x}_* = A^+ \mathbf{y}$ returning the minimum-norm solution (see exercises).

Rank-deficient least-squares problems

Zero singular values $\iff$ redundant models: for instance,

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

would be redundant. (Only linear dependencies cause singularity.)

Problem: exact dependencies are very rarely encountered.

More often, one will see approximate dependencies. This is caused also by two effects:

- Noise in your data: e.g., $\begin{bmatrix} 0.1 & 0.2 \\ 0.2 & 0.4 \\ 0.3 & 0.599999 \end{bmatrix}$ is not an exact dependency, $\sigma_n \neq 0$.
- Inexact computation: even with an exact dependency, computer arithmetic often produces $\sigma_n \neq 0$. We will see more in the following, but the effect of machine arithmetic (with backward stable algorithms) is comparable to a (relative) error of order $\mathrm{u} \approx 10^{-16}$ in your data.

Theorem

Let σ_i be the singular values of A, and $\tilde{\sigma}_i$ those of $A + E$. Then, $\|\sigma_i - \tilde{\sigma}_i\| \leq \|E\|$.

Example

```
>> M = dlmread('salaries.csv', ',', 1, 1);
>> A = M(:, 1:3);
>> A(:,4) = A(:,1) + A(:,3);
ans =
>> svd(A)
   2.8060e+04
   3.2171e+03
   8.7262e+02
   1.5007e-12
>> rank(A'*A)
ans =
     3
>> svd(A + 0.01*rand(size(A)))
     2.8060e+04
     3.2172e+03
     8.7264e+02
     6.7068e-02
```

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
```

Note that with `eig` the smallest eigenvalue 0 is affected by a perturbation of $10^{-8} \approx \mathsf{u}\|A^T A\| = \mathsf{u}\lambda_1$, while with `svd` the smallest singular value 0 is affected by a perturbation of $10^{-12} \approx \mathsf{u}\|A\| = \mathsf{u}\sigma_1$. So `svd` is more accurate than `eig`.

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

If you know for certain that $\sigma_4 = 0$, you can stop the sum early and compute the minimum-norm solution as

$$\mathbf{x}_* = \sum_{i=1}^{r} \mathbf{v}_i \frac{\mathbf{u}_i^T \mathbf{y}}{\sigma_i}.$$

Small singular values

A related issue is the one of small singular values. Many real-world matrices have decaying singular values, e.g.,

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

This makes it even more difficult to tell when a model is exactly singular.

Truncated SVD

The exact solution $\mathbf{x}$ varies wildly depending on the exact value of the small σ_i.

This has a large impact on the computed solution, since σ_i appears in the denominator:

$$\mathbf{x} = \sum_{i=1}^{n} \mathbf{v}_i \frac{\mathbf{u}_i^T \mathbf{y}}{\sigma_i}.$$

However, in many applications the most meaningful features correspond to the large singular values; recall: eigenfaces, image compression.

One often gets a better solution (from the point of view of the application) by ignoring the contribution of small singular values:

$$\mathbf{x}_{reg} = \sum_{i=1}^{k} \mathbf{v}_i \frac{\mathbf{u}_i^T \mathbf{y}}{\sigma_i}, \quad \text{(for a certain } k < r.\text{)}$$

This $\mathbf{x}_{reg}$ is not the solution of $\min \|A\mathbf{x} - \mathbf{y}\|$, but sometimes it gives better application results.

Example (not the best one)

With the previous $A, \mathbf{y}$ from the basketball analytics problem:

```
>> AA = A + 0.01*rand(size(A));
>> AA \ y
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)'*y
ans =
   5.6843e+03
  -2.6577e+04
   1.9155e+03
   7.6007e+03
```

This is a better approximation of the (inaccessible) true solution `A \ y`.

Alternative: Tikhonov regularization / ridge regression

A different solution to the problem of what to do when there are tiny singular values: change your problem, and look for

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|A\mathbf{x}-\mathbf{y}\|^2 + \alpha^2\|\mathbf{x}\|^2$$

(for a given $\alpha > 0$). The second term discourages solutions with large norm. This is a classical strategy in optimization: penalty terms.

We can rewrite the objective function as

$$\|A\mathbf{x}-\mathbf{y}\|^2 + \alpha^2\|\mathbf{x}\|^2 = \left\| \begin{bmatrix} A \\ \alpha I \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} \right\|^2.$$

Tikhonov / ridge — formula

Thanks to this expression, we can give an explicit solution formula:

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

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

Tikhonov / ridge and SVD

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

$$\mathbf{x}_\alpha = \sum_{i=1}^{n} \mathbf{v}_i \frac{\sigma_i}{\sigma_i^2 + \alpha^2} \mathbf{u}_i^T \mathbf{y}.$$

This function $f(\sigma) = \frac{\sigma}{\sigma^2+\alpha^2}$ approximates a truncated SVD:

When $\sigma \gg \alpha$, $f(\sigma) \approx \frac{1}{\sigma}$: similar to the true LS solution.

When $\sigma \ll \alpha$, $f(\sigma) \approx \frac{\sigma}{\alpha} \approx 0$: approximately ignoring small singular values.

Choice of α

How to choose α? Difficult to motivate the choice mathematically: we are asking "how to modify the problem", not "how to solve the problem".

Sometimes it makes sense to take $\alpha \approx$ magnitude of the noise/uncertainty in your data (if you know it!). Sometimes, there are application-specific choices; you will see more in ML / AI courses. ML people love grid searches: throw processing power at it and learn from similar problems.

Similar arguments hold for the choice of k in truncated SVD.

Book references: Demmel, ch. 3.5. Trefethen-Bau, just some quick remarks on p. 143. Eldén, ch. 6.7 and 7 (best).

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. Show that the matrix denoted with Σ^+ above is, indeed, the pseudoinverse of Σ.

Exercises

1. Let A be a matrix that does not have full column rank, and $A = U\Sigma V^T$ be its SVD, with rank r and singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_n = 0$. Show that the solution of $\arg\min_{\mathbf{x} \in \arg\min \|A\mathbf{x} - \mathbf{y}\|} \|\mathbf{x}\|$ is $\mathbf{x}_* = A^+\mathbf{y}$, where

$$A^+ = V \begin{bmatrix} \frac{1}{\sigma_1} & & & & & & & \\ & \frac{1}{\sigma_2} & & & & & & \\ & & \ddots & & & & & \\ & & & \frac{1}{\sigma_r} & & & & \\ & & & & 0 & & & \\ & & & & & 0 & & \\ & & & & & & \ddots & \\ & & & & & & & 0 \end{bmatrix} U^T.$$

   This formula can be taken as a definition of the pseudoinverse A^+ for a matrix A that does not have full column rank.