The Least-Squares Problem

We can use the SVD to "solve" the general system of linear equations

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$

where $\boldsymbol{y} \in \mathbb{R}^M$, $\boldsymbol{x} \in \mathbb{R}^N$, and $\boldsymbol{A}$ is an $M \times N$ matrix.

Given $\boldsymbol{y}$, we want to find $\boldsymbol{x}$ in such a way that

1. when there is a unique solution, we return it;
2. when there is no solution, we return something reasonable;
3. when there are an infinite number of solutions, we choose one to return in a "smart" way.

The **least-squares** framework revolves around finding an $\boldsymbol{x}$ that minimizes the length of the residual

$$\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}.$$

That is, we want to solve the optimization problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \;\; \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2, \tag{1}$$

where $\|\cdot\|_2$ is the standard Euclidean norm. We will see that the SVD of $\boldsymbol{A}$:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}, \tag{2}$$

plays a pivotal role in solving this problem.

To start, note that we can write any $\boldsymbol{x} \in \mathbb{R}^N$ as

$$\boldsymbol{x} = \boldsymbol{V}\boldsymbol{\alpha} + \boldsymbol{V}_0\boldsymbol{\alpha}_0. \tag{3}$$

Here, $\boldsymbol{V}$ is the $N \times R$ matrix appearing in the SVD decomposition (2), and $\boldsymbol{V}_0$ is a $N \times (N-R)$ matrix whose columns are orthogonal to one another and to the columns in $\boldsymbol{V}$. We have the relations

$$\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \mathbf{I}, \quad \boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{V}_0 = \mathbf{I}, \quad \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}_0 = \mathbf{0}.$$

You can think of $\boldsymbol{V}_0$ as an orthobasis for the null space of $\boldsymbol{A}$. Of course, $\boldsymbol{V}_0$ is not unique, as there are many orthobases for $\mathrm{Null}(\boldsymbol{A})$, but any such set of vectors will serve our purposes here. The decomposition (3) is possible since $\mathrm{Range}(\boldsymbol{A}^{\mathrm{T}})$ and $\mathrm{Null}(\boldsymbol{A})$ partition $\mathbb{R}^N$ for any $M \times N$ matrix $\boldsymbol{A}$. Taking

$$\boldsymbol{\alpha} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{x}, \quad \boldsymbol{\alpha}_0 = \boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{x},$$

we see that (3) holds since

$$\boldsymbol{x} = \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{V}_0\boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{x} = (\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{V}_0\boldsymbol{V}_0^{\mathrm{T}})\boldsymbol{x} = \boldsymbol{x},$$

where we have made use of the fact that $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{V}_0\boldsymbol{V}_0^{\mathrm{T}} = \mathbf{I}$, because $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}$ and $\boldsymbol{V}_0\boldsymbol{V}_0^{\mathrm{T}}$ are ortho-projectors onto complementary subspaces[1] of $\mathbb{R}^N$. So we can solve for $\boldsymbol{x} \in \mathbb{R}^N$ by solving for the pair $\boldsymbol{\alpha} \in \mathbb{R}^R$, $\boldsymbol{\alpha}_0 \in \mathbb{R}^{N-R}$.

Similarly, we can decompose $\boldsymbol{y}$ as

$$\boldsymbol{y} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{U}_0\boldsymbol{\beta}_0, \tag{4}$$

where $\boldsymbol{U}$ is the $M \times R$ matrix from the SVD decomposition, and $\boldsymbol{U}_0$ is a $M \times (M-R)$ complementary orthogonal basis. Again,

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \mathbf{I}, \quad \boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{U}_0 = \mathbf{I}, \quad \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}_0 = \mathbf{0},$$

[1]Subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ are **complementary** in $\mathbb{R}^N$ if $\mathcal{S}_1 \perp \mathcal{S}_2$ (everything in $\mathcal{S}_1$ is orthogonal to everything in $\mathcal{S}_2$) and $\mathcal{S}_1 \oplus \mathcal{S}_2 = \mathbb{R}^N$. You can think of $\mathcal{S}_1, \mathcal{S}_2$ as a partition of $\mathbb{R}^N$ into two orthogonal subspaces.

and we can think of $\boldsymbol{U}_0$ as an orthogonal basis for everything in $\mathbb{R}^M$ that is not in the range of $\boldsymbol{A}$. As before, we can calculate the decomposition above using

$$\boldsymbol{\beta} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{y}, \quad \boldsymbol{\beta}_0 = \boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{y}.$$

Using the decompositions (2), (3), and (4) for $\boldsymbol{A}$, $\boldsymbol{x}$, and $\boldsymbol{y}$, we can write the residual $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}$ as

$$\begin{aligned}\boldsymbol{r} &= \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{U}_0\boldsymbol{\beta}_0 - \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}(\boldsymbol{V}\boldsymbol{\alpha} + \boldsymbol{V}_0\boldsymbol{\alpha}_0) \\ &= \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{U}_0\boldsymbol{\beta}_0 - \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\alpha} \quad (\text{since } \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \mathbf{I} \text{ and } \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}_0 = \mathbf{0}) \\ &= \boldsymbol{U}_0\boldsymbol{\beta}_0 + \boldsymbol{U}(\boldsymbol{\beta} - \boldsymbol{\Sigma}\boldsymbol{\alpha}).\end{aligned}$$

We want to choose $\boldsymbol{\alpha}$ that minimizes the energy of $\boldsymbol{r}$:

$$\begin{aligned}\|\boldsymbol{r}\|_2^2 &= \langle \boldsymbol{U}_0\boldsymbol{\beta}_0 + \boldsymbol{U}(\boldsymbol{\beta} - \boldsymbol{\Sigma}\boldsymbol{\alpha}),\ \boldsymbol{U}_0\boldsymbol{\beta}_0 + \boldsymbol{U}(\boldsymbol{\beta} - \boldsymbol{\Sigma}\boldsymbol{\alpha})\rangle \\ &= \langle \boldsymbol{U}_0\boldsymbol{\beta}_0, \boldsymbol{U}_0\boldsymbol{\beta}_0\rangle \ + \ 2\langle \boldsymbol{U}_0\boldsymbol{\beta}_0, \boldsymbol{U}(\boldsymbol{\beta} - \boldsymbol{\Sigma}\boldsymbol{\alpha})\rangle \\ &\qquad + \langle \boldsymbol{U}(\boldsymbol{\beta} - \boldsymbol{\Sigma}\boldsymbol{\alpha}), \boldsymbol{U}(\boldsymbol{\beta} - \boldsymbol{\Sigma}\boldsymbol{\alpha})\rangle \\ &= \|\boldsymbol{\beta}_0\|_2^2 + \|\boldsymbol{\beta} - \boldsymbol{\Sigma}\boldsymbol{\alpha}\|_2^2\end{aligned}$$

where the last equality comes from the facts that $\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{U}_0 = \mathbf{I}$, $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \mathbf{I}$, and $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}_0 = \mathbf{0}$. We have no control over $\|\boldsymbol{\beta}_0\|_2^2$, since it determined entirely by our observations $\boldsymbol{y}$. Therefore, our problem has been reduced to finding $\boldsymbol{\alpha}$ that minimizes the second term $\|\boldsymbol{\beta} - \boldsymbol{\Sigma}\boldsymbol{\alpha}\|_2^2$ above, which is non-negative. We can make it zero (i.e. as small as possible) by taking

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}.$$

Finally, the $\boldsymbol{x}$ which minimizes the residual (solves (1)) is

$$\hat{\boldsymbol{x}} = \boldsymbol{V}\hat{\boldsymbol{\alpha}} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y}. \tag{5}$$

Thus we can calculate the solution to (1) simply by applying the linear operator $\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}$ to the input data $\boldsymbol{y}$. There are two interesting facts about the solution $\hat{\boldsymbol{x}}$ in (5):

1. When $\boldsymbol{y} \in \operatorname{span}(\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M\})$, we have $\boldsymbol{\beta}_0 = \boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{y} = \mathbf{0}$, and so the residual $\boldsymbol{r} = \mathbf{0}$. In this case, there is at least one exact solution, and the one we choose satisfies $\boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{y}$.

2. Note that if $R < N$, then the solution is not unique. In this case, $\boldsymbol{V}_0$ has at least one column, and any part of a vector $\boldsymbol{x}$ in the range of $\boldsymbol{V}_0$ is not seen by $\boldsymbol{A}$, since

$$\boldsymbol{A}\boldsymbol{V}_0\boldsymbol{\alpha}_0 = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}_0\boldsymbol{\alpha}_0 = \mathbf{0} \quad (\text{since } \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}_0 = \mathbf{0}).$$

   As such,

$$\boldsymbol{x}' = \hat{\boldsymbol{x}} + \boldsymbol{V}_0\boldsymbol{\alpha}_0$$

   for *any* $\boldsymbol{\alpha}_0 \in \mathbb{R}^{N-R}$ will have exactly the same residual, since $\boldsymbol{A}\boldsymbol{x}' = \boldsymbol{A}\hat{\boldsymbol{x}}$. In this case, our solution $\hat{\boldsymbol{x}}$ is the solution with smallest norm, since

$$\begin{aligned}
\|\boldsymbol{x}'\|_2^2 &= \langle \hat{\boldsymbol{x}} + \boldsymbol{V}_0\boldsymbol{\alpha}_0,\ \hat{\boldsymbol{x}} + \boldsymbol{V}_0\boldsymbol{\alpha}_0\rangle \\
&= \langle \hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}\rangle + 2\langle \hat{\boldsymbol{x}}, \boldsymbol{V}_0\boldsymbol{\alpha}_0\rangle + \langle \boldsymbol{V}_0\boldsymbol{\alpha}, \boldsymbol{V}_0\boldsymbol{\alpha}\rangle \\
&= \|\hat{\boldsymbol{x}}\|_2^2 + 2\langle \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y}, \boldsymbol{V}_0\boldsymbol{\alpha}_0\rangle + \|\boldsymbol{\alpha}_0\|_2^2 \quad (\text{since } \boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{V}_0 = \mathbf{I}) \\
&= \|\hat{\boldsymbol{x}}\|_2^2 + \|\boldsymbol{\alpha}_0\|_2^2 \quad (\text{since } \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}_0 = \mathbf{0})
\end{aligned}$$

   which is minimized by taking $\boldsymbol{\alpha}_0 = \mathbf{0}$.

To summarize, $\hat{\boldsymbol{x}} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y}$ has the desired properties stated at the beginning of this module, since

1. when $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ has a unique exact solution, it must be $\hat{\boldsymbol{x}}$,
2. when an exact solution is not available, $\hat{\boldsymbol{x}}$ is the solution to (1),

3. when there are an infinite number of minimizers to (1), $\hat{\boldsymbol{x}}$ is the one with smallest norm.

Because the matrix $\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}$ gives us such an elegant solution to this problem, we give it a special name: the **pseudo-inverse**.

The Pseudo-Inverse

The **pseudo-inverse** of a matrix $\boldsymbol{A}$ with singular value decomposition (SVD) $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ is

$$\boldsymbol{A}^{\dagger} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}. \tag{6}$$

Other names for $\boldsymbol{A}^{\dagger}$ include **natural inverse**, **Lanczos inverse**, and **Moore-Penrose inverse**.

Given an observation $\boldsymbol{y}$, taking $\hat{\boldsymbol{x}} = \boldsymbol{A}^{\dagger}\boldsymbol{y}$ gives us the **least squares** solution to $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$. The pseudo-inverse $\boldsymbol{A}^{\dagger}$ always exists, since every matrix (with rank R) has an SVD decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ with $\boldsymbol{\Sigma}$ as an $R \times R$ diagonal matrix with $\Sigma[r, r] > 0$.

When $\boldsymbol{A}$ is full rank ($R = \min(M, N)$), then we can calculate the pseudo-inverse without using the SVD. There are three cases:

- When $\boldsymbol{A}$ is square and invertible ($R = M = N$), then

  $$\boldsymbol{A}^{\dagger} = \boldsymbol{A}^{-1}.$$

  This is easy to check, as here

  $$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \quad \text{where both } \boldsymbol{U}, \boldsymbol{V} \text{ are } N \times N,$$

and since in this case $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \mathbf{I}$ and $\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \mathbf{I}$,

$$\begin{aligned}\boldsymbol{A}^{\dagger}\boldsymbol{A} &= \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \\ &= \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \\ &= \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} \\ &= \mathbf{I}.\end{aligned}$$

Similarly, $\boldsymbol{A}\boldsymbol{A}^{\dagger} = \mathbf{I}$, and so $\boldsymbol{A}^{\dagger}$ is both a left and right inverse of $\boldsymbol{A}$, and thus $\boldsymbol{A}^{\dagger} = \boldsymbol{A}^{-1}$.

- When $\boldsymbol{A}$ more rows than columns and has full column rank ($R = N \leq M$), then $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ is invertible, and

$$\boldsymbol{A}^{\dagger} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}. \tag{7}$$

This type of $\boldsymbol{A}$ is "tall and skinny"

$$\begin{bmatrix} \; \boldsymbol{A} \; \\ \\ \\ \end{bmatrix},$$

and its columns are linearly independent. To verify equation (7), recall that

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{V}\boldsymbol{\Sigma}^{2}\boldsymbol{V}^{\mathrm{T}},$$

and so

$$(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{V}\boldsymbol{\Sigma}^{-2}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}},$$

which is exactly the content of (6).

- When $\boldsymbol{A}$ has more columns than rows and has full row rank ($R = M \leq N$), then $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ is invertible, and

$$\boldsymbol{A}^{\dagger} = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}. \tag{8}$$

This occurs when $\boldsymbol{A}$ is "short and fat"

$$\left[\quad\quad \boldsymbol{A} \quad\quad\right],$$

and its rows are linearly independent. To verify equation (8), recall that

$$\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^{\mathrm{T}},$$

and so

$$\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{\Sigma}^{-2}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}},$$

which again is exactly (6).

$\boldsymbol{A}^{\dagger}$ is as close to an inverse of $\boldsymbol{A}$ as possible

As discussed in the last section, when $\boldsymbol{A}$ is square and invertible, $\boldsymbol{A}^{\dagger}$ is exactly the inverse of $\boldsymbol{A}$. When $\boldsymbol{A}$ is not square, we can ask if there is a better right or left inverse. We will argue that there is not.

Left inverse Given $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, we would like $\boldsymbol{A}^{\dagger}\boldsymbol{y} = \boldsymbol{A}^{\dagger}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}$ for any $\boldsymbol{x}$. That is, we would like $\boldsymbol{A}^{\dagger}$ to be a *left inverse* of $\boldsymbol{A}$: $\boldsymbol{A}^{\dagger}\boldsymbol{A} = \mathbf{I}$. Of course, this is not always possible, especially when $\boldsymbol{A}$ has more columns than rows, $M < N$. But we can ask if any other matrix $\boldsymbol{H}$ comes closer to being a left inverse than $\boldsymbol{A}^\dagger$. To find the "best" left-inverse, we look for the matrix which minimizes

$$\min_{\boldsymbol{H}\in\mathbb{R}^{N\times M}} \|\boldsymbol{H}\boldsymbol{A} - \mathbf{I}\|_F^2. \tag{9}$$

Here, $\|\cdot\|_F$ is the *Frobenius norm*, defined for an $N \times M$ matrix $\boldsymbol{Q}$ as the sum of the squares of the entires:

$$\|\boldsymbol{Q}\|_F^2 = \sum_{n=1}^{M}\sum_{n=1}^{N} |Q[m,n]|^2$$

(It is also true, and you can and should prove this at home, that $\|\boldsymbol{Q}\|_F^2$ is the sum of the squares of the singular values of $\boldsymbol{Q}$: $\|\boldsymbol{Q}\|_F^2 = \lambda_1^2 + \cdots + \lambda_p^2$.) With (9), we are finding $\boldsymbol{H}$ such that $\boldsymbol{H}\boldsymbol{A}$ is as close to the identity as possible in the least-squares sense.

The pseudo-inverse $\boldsymbol{A}^\dagger$ minimizes (9). To see this, recognize (see the exercise below) that the solution $\hat{\boldsymbol{H}}$ to (9) must obey

$$\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\hat{\boldsymbol{H}}^{\mathrm{T}} = \boldsymbol{A}. \tag{10}$$

We can see that this is indeed true for $\hat{\boldsymbol{H}} = \boldsymbol{A}^\dagger$:

$$\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}^{\dagger\mathrm{T}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{A}.$$

So there is no $N \times M$ matrix that is closer to being a left inverse than $\boldsymbol{A}^\dagger$.

Right inverse If we re-apply $\boldsymbol{A}$ to our solution $\hat{\boldsymbol{x}} = \boldsymbol{A}^\dagger\boldsymbol{y}$, we would like it to be as close as possible to our observations $\boldsymbol{y}$. That is, we would like $\boldsymbol{A}\boldsymbol{A}^\dagger$ to be as close to the identity as possible. Again, achieving this goal exactly is not always possible, especially if $\boldsymbol{A}$ has more rows that columns. But we can attempt to find the "best" right inverse, in the least-squares sense, by solving

$$\underset{\boldsymbol{H}\in\mathbb{R}^{N\times M}}{\text{minimize}} \ \|\boldsymbol{A}\boldsymbol{H} - \mathbf{I}\|_F^2. \tag{11}$$

The solution $\hat{\boldsymbol{H}}$ to (11) (see the exercise below) must obey

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\hat{\boldsymbol{H}} = \boldsymbol{A}^{\mathrm{T}}. \tag{12}$$

Again, we show that $\boldsymbol{A}^\dagger$ satisfies (12), and hence is a minimizer to (11):

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^\dagger = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{A}^{\mathrm{T}}.$$

Moral:
$\boldsymbol{A}^\dagger = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}$ is as close (in the least-squares sense) to an inverse of $\boldsymbol{A}$ as you could possibly have.

Exercise:

1. Show that the minimizer $\hat{\boldsymbol{H}}$ to (9) must obey (10). Do this by using the fact that the derivative of the functional $\|\boldsymbol{H}\boldsymbol{A} - \mathbf{I}\|_F^2$ with respect to an entry $H[k,\ell]$ in $\boldsymbol{H}$ must obey

$$\frac{\partial\|\boldsymbol{H}\boldsymbol{A} - \mathbf{I}\|_F^2}{\partial H[k,\ell]} = 0, \quad \text{for all } 1 \le k \le N,\ 1 \le \ell \le M,$$

to be a solution to (9). Do the same for (11) and (12).

Stability Analysis of the Pseudo-Inverse

We have seen that if we make indirect observations $\boldsymbol{y} \in \mathbb{R}^M$ of an unknown vector $\boldsymbol{x}_0 \in \mathbb{R}^N$ through a $M \times N$ matrix $\boldsymbol{A}$, $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0$, then applying the pseudo-inverse of $\boldsymbol{A}$ gives us the least squares estimate of $\boldsymbol{x}_0$:

$$\hat{\boldsymbol{x}}_{\text{ls}} = \boldsymbol{A}^\dagger \boldsymbol{y} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y},$$

where $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ is the singular value decomposition (SVD) of $\boldsymbol{A}$.

We will now discuss what happens if our measurements contain *noise* — the analysis here will be very similar to when we looked at the stability of solving square sym+def systems, and in fact this is one of the main reasons we introduced the SVD.

Suppose we observe

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{e},$$

where $\boldsymbol{e} \in \mathbb{R}^M$ is an unknown perturbation. Say that we again apply the pseudo-inverse to $\boldsymbol{y}$ in an attempt to recover $\boldsymbol{x}$:

$$\hat{\boldsymbol{x}}_{\text{ls}} = \boldsymbol{A}^\dagger \boldsymbol{y} = \boldsymbol{A}^\dagger \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{A}^\dagger \boldsymbol{e}$$

What effect does the presence of the noise vector $\boldsymbol{e}$ had on our estimate of $\boldsymbol{x}_0$? We answer this question by comparing $\hat{\boldsymbol{x}}_{\text{ls}}$ to the reconstruction we would obtain if we used standard least-squares on perfectly noise-free observations $\boldsymbol{y}_{\text{clean}} = \boldsymbol{A}\boldsymbol{x}_0$. This noise-free reconstruction can be written as

$$\begin{aligned}
\boldsymbol{x}_{\text{pinv}} = \boldsymbol{A}^\dagger \boldsymbol{y}_{\text{clean}} &= \boldsymbol{A}^\dagger \boldsymbol{A}\boldsymbol{x}_0 \\
&= \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{x}_0 \\
&= \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{x}_0 \\
&= \sum_{r=1}^{R} \langle \boldsymbol{x}_0, \boldsymbol{v}_r \rangle \boldsymbol{v}_r.
\end{aligned}$$

The vector $\boldsymbol{x}_{\text{pinv}}$ is the orthogonal projection of $\boldsymbol{x}_0$ onto the row space (everything orthogonal to the null space) of $\boldsymbol{A}$. If $\boldsymbol{A}$ has full column rank ($R = N$), then $\boldsymbol{x}_{\text{pinv}} = \boldsymbol{x}_0$. If not, then the application of $\boldsymbol{A}$ destroys the part of $\boldsymbol{x}_0$ that is not in $\boldsymbol{x}_{\text{pinv}}$, and so we only attempt to recover the "visible" components. In some sense, $\boldsymbol{x}_{\text{pinv}}$ contains all of the components of $\boldsymbol{x}_0$ that $\boldsymbol{A}$ does not completely remove, and has them preserved perfectly.

The reconstruction error (relative to $\boldsymbol{x}_{\text{pinv}}$ is)

$$\|\hat{\boldsymbol{x}}_{\text{ls}} - \boldsymbol{x}_{\text{pinv}}\|_2^2 = \|\boldsymbol{A}^\dagger \boldsymbol{e}\|_2^2 = \|\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2. \tag{13}$$

Now suppose for a moment that the error has unit norm, $\|\boldsymbol{e}\|_2^2 = 1$. Then the worst case for (13) is given by

$$\underset{\boldsymbol{e}\in\mathbb{R}^M}{\text{maximize}} \ \ \|\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{e}\|_2 = 1.$$

Since the columns of $\boldsymbol{U}$ are orthonormal, $\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2 \leq \|\boldsymbol{e}\|_2^2$, and the above is equivalent to

$$\max_{\boldsymbol{\beta}\in\mathbb{R}^R : \|\boldsymbol{\beta}\|_2 = 1} \ \|\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\|_2^2. \tag{14}$$

Also, for any vector $\boldsymbol{z} \in \mathbb{R}^R$, we have

$$\|\boldsymbol{V}\boldsymbol{z}\|_2^2 = \langle \boldsymbol{V}\boldsymbol{z}, \boldsymbol{V}\boldsymbol{z}\rangle = \langle \boldsymbol{z}, \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{z}\rangle = \langle \boldsymbol{z}, \boldsymbol{z}\rangle = \|\boldsymbol{z}\|_2^2,$$

since the columns of $\boldsymbol{V}$ are orthonormal. So we can simplify (14) to

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^R}{\text{maximize}} \ \|\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_2 = 1.$$

The worst case $\boldsymbol{\beta}$ (you should verify this at home) will have a 1 in the entry corresponding to the largest entry in $\boldsymbol{\Sigma}^{-1}$, and will be zero everywhere else. Thus

$$\max_{\boldsymbol{\beta}\in\mathbb{R}^R:\|\boldsymbol{\beta}\|_2=1} \|\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\|_2^2 \quad = \quad \max_{r=1,\ldots,R} \sigma_r^{-2} \quad = \quad \frac{1}{\sigma_R^2}.$$

(Recall that by convention, we order the singular values so that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_R$.)

Returning to the reconstruction error (13), we now see that

$$\|\hat{\boldsymbol{x}}_{\mathrm{ls}} - \boldsymbol{x}_{\mathrm{pinv}}\|_2^2 \ = \ \|\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2 \ \leq \ \frac{1}{\sigma_R^2}\|\boldsymbol{e}\|_2^2.$$

Since $\boldsymbol{U}$ is an $M \times R$ matrix, it is possible when $R < M$ that the reconstruction error is zero. This happens when $\boldsymbol{e}$ is orthogonal to every column of $\boldsymbol{U}$, i.e. $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e} = \boldsymbol{0}$. Putting this together with the work above means

$$0 \ \leq \ \frac{1}{\sigma_1^2}\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2 \ \leq \ \|\hat{\boldsymbol{x}}_{\mathrm{ls}} - \boldsymbol{x}_{\mathrm{pinv}}\|_2^2 \ \leq \ \frac{1}{\sigma_R^2}\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2 \ \leq \ \frac{1}{\sigma_R^2}\|\boldsymbol{e}\|_2^2.$$

Notice that if σ_R is small, the worst case reconstruction error can be **very bad**.

We can also relate the "average case" error to the singular values. Say that $\boldsymbol{e}$ is additive Gaussian white noise, that is each entry $e[m]$ is a random variable independent of all the other entries, and distributed

$$e[m] \ \sim \ \mathrm{Normal}(0, \nu^2).$$

Then, as we have argued before, the average measurement error is

$$\mathrm{E}[\|\boldsymbol{e}\|_2^2] = M\nu^2,$$

and the average reconstruction error[2] is

$$\begin{aligned}\mathrm{E}\left[\|\boldsymbol{A}^\dagger\boldsymbol{e}\|_2^2\right] = \nu^2 \cdot \mathrm{trace}(\boldsymbol{A}^{\dagger\mathrm{T}}\boldsymbol{A}^\dagger) &= \nu^2 \cdot \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \cdots + \frac{1}{\sigma_R^2}\right) \\ &= \frac{1}{M}\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \cdots + \frac{1}{\sigma_R^2}\right) \cdot \mathrm{E}[\|\boldsymbol{e}\|_2^2].\end{aligned}$$

Again, if σ_R is tiny, $1/\sigma_R^2$ will dominate the sum above, and the average reconstruction error will be quite large.

[2]We are using the fact that if $\boldsymbol{e}$ is vector of iid Gaussian random variables, $\boldsymbol{e} \sim \mathrm{Normal}(\boldsymbol{0}, \nu^2\mathbf{I})$, then for any matrix $\boldsymbol{M}$, $\mathrm{E}[\|\boldsymbol{M}\boldsymbol{e}\|_2^2] = \nu^2\,\mathrm{trace}(\boldsymbol{M}^\mathrm{T}\boldsymbol{M})$. We will argue this carefully as part of the next homework.

Decomposition of the estimation error

In the previous section, we compared the estimate $\hat{\boldsymbol{x}}_{\text{ls}}$ to $\boldsymbol{x}_{\text{pinv}}$, the projection of $\boldsymbol{x}_0$ onto the row space of $\boldsymbol{A}$. In this section, let's see how to compare $\hat{\boldsymbol{x}}_{\text{ls}}$ to $\boldsymbol{x}_o$ itself.

We can separate this error into two separate errors:

$$\hat{\boldsymbol{x}}_{\text{ls}} - \boldsymbol{x}_0 = \hat{\boldsymbol{x}}_{\text{ls}} - \boldsymbol{x}_{\text{pinv}} + \boldsymbol{x}_{\text{pinv}} - \boldsymbol{x}_0. \tag{15}$$

The first error above, $\hat{\boldsymbol{x}}_{\text{ls}} - \boldsymbol{x}_{\text{pinv}}$, is what we analyzed above. Let's take a close look at the second error $\boldsymbol{x}_{\text{pinv}} - \boldsymbol{x}_0$.

The singular vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_R \in \mathbb{R}^N$ are an orthobasis for the row space (the span of the rows) of $\boldsymbol{A}$. Let $\boldsymbol{v}_{R+1}, \ldots, \boldsymbol{v}_N \in \mathbb{R}^N$ be an orthobasis for the null space of $\boldsymbol{A}$ (the linear subspace of all vectors $\boldsymbol{x}$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$) — these would be the columns of the $N \times (N-R)$ matrix $\boldsymbol{V}_0$ in the last set of notes. Together, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$ form an orthobasis for all of $\mathbb{R}^N$.

We can decompose $\boldsymbol{x}_0$ as

$$\begin{aligned}
\boldsymbol{x}_0 &= \sum_{n=1}^{N} \langle \boldsymbol{x}_0, \boldsymbol{v}_n \rangle \boldsymbol{v}_n \\
&= \sum_{n=1}^{R} \langle \boldsymbol{x}_0, \boldsymbol{v}_n \rangle \boldsymbol{v}_n + \sum_{n=R+1}^{N} \langle \boldsymbol{x}_0, \boldsymbol{v}_n \rangle \boldsymbol{v}_n \\
&= \boldsymbol{x}_{\text{pinv}} + \boldsymbol{x}_{\text{null}}.
\end{aligned}$$

The vector $\boldsymbol{x}_{\text{pinv}}$ is the projection of $\boldsymbol{x}_0$ onto the row space of $\boldsymbol{A}$ (i.e. $\boldsymbol{x}_{\text{pinv}}$ is the closest point in the subspace of $\mathbb{R}^N$ formed by taking all linear combinations of the rows of $\boldsymbol{A}$). The vector $\boldsymbol{x}_{\text{null}}$ is the projection of $\boldsymbol{x}_0$ onto the null space of $\boldsymbol{A}$. Since the $\boldsymbol{v}_n$ are all orthogonal to one another, it should be clear that $\boldsymbol{x}_{\text{pinv}} \perp \boldsymbol{x}_{\text{null}}$.

Returning to (15), we can write

$$\hat{\boldsymbol{x}}_{\text{ls}} - \boldsymbol{x}_0 \;=\; (\hat{\boldsymbol{x}}_{\text{ls}} - \boldsymbol{x}_{\text{pinv}}) - \boldsymbol{x}_{\text{null}}.$$

As we saw above, $\hat{\boldsymbol{x}}_{\text{ls}} - \boldsymbol{x}_{\text{pinv}}$ is also in the span of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_R$, and so it is also orthogonal to $\boldsymbol{x}_{\text{null}}$. Thus we can write

$$\begin{aligned} \|\hat{\boldsymbol{x}}_{\text{ls}} - \boldsymbol{x}_0\|_2^2 &= \|\hat{\boldsymbol{x}}_{\text{ls}} - \boldsymbol{x}_{\text{pinv}}\|_2^2 + \|\boldsymbol{x}_{\text{null}}\|_2^2 \\ &= \|\text{Noise error}\|_2^2 + \|\text{Nullspace error}\|_2^2. \end{aligned}$$

The null space error has nothing to do with the noise added to the observations. It is simply the part of $\boldsymbol{x}_0$ that $\boldsymbol{A}$ removes completely.