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.

Our analysis starts by showing how a vector in $\mathbb{R}^N$ can be decomposed in an orthobasis related to the right singular vectors of $\boldsymbol{A}$. For any $\boldsymbol{x} \in \mathbb{R}^N$, we can write

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

where $\boldsymbol{V}$ 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 an orthobasis for the null space of $\boldsymbol{A}$. We have the relations[1]

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

Using these, we can compute the $\boldsymbol{\alpha}, \boldsymbol{\alpha}_0$ using

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

and we have that

$$\|\boldsymbol{x}\|_2^2 = \|\boldsymbol{V}\boldsymbol{\alpha}\|_2^2 + \|\boldsymbol{V}_0\boldsymbol{\alpha}_0\|_2^2 = \|\boldsymbol{\alpha}\|_2^2 + \|\boldsymbol{\alpha}_0\|_2^2.$$

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 $M \times (M-R)$ orthogonal basis for the left null space of $\boldsymbol{A}$ (everything in $\mathbb{R}^M$ that is orthogonal to the range of $\boldsymbol{A}$). Again,

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

We can calculate the decomposition above using

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

and we have that

$$\|\boldsymbol{y}\|_2^2 = \|\boldsymbol{U}\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{U}_0\boldsymbol{\beta}_0\|_2^2 = \|\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{\beta}_0\|_2^2.$$

[1] On short, the decomposition (3) is possible since Range($\boldsymbol{A}^{\mathrm{T}}$) and Null($\boldsymbol{A}$) partition $\mathbb{R}^N$ for any $M \times N$ matrix $\boldsymbol{A}$. Every vector in $\mathbb{R}^N$ can be written as a sum of components from Range($\boldsymbol{A}^{\mathrm{T}}$) and Null($\boldsymbol{A}$), and these two components will be orthogonal to one another.

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

$$\begin{aligned}\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} &= \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{U}_0\boldsymbol{\beta}_0 - \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}(\boldsymbol{V}\boldsymbol{\alpha} + \boldsymbol{V}_0\boldsymbol{\alpha}_0) \\ &= \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{U}_0\boldsymbol{\beta}_0 - \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\alpha} \\ &= \boldsymbol{U}_0\boldsymbol{\beta}_0 + \boldsymbol{U}(\boldsymbol{\beta} - \boldsymbol{\Sigma}\boldsymbol{\alpha}).\end{aligned}$$

(The second equality above follows from $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \mathbf{I}$ and $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}_0 = \mathbf{0}$.) The size of the residual is:

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

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 $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}_0 = \mathbf{0}$.

Thus we can solve (1) by solving

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^R,\ \boldsymbol{\alpha}_0\in\mathbb{R}^{(N-R)}}{\text{minimize}} \quad \|\boldsymbol{\beta}_0\|_2^2 + \|\boldsymbol{\beta} - \boldsymbol{\Sigma}\boldsymbol{\alpha}\|_2^2, \tag{5}$$

and then taking $\hat{\boldsymbol{x}} = \boldsymbol{V}\hat{\boldsymbol{\alpha}} + \boldsymbol{V}_0\hat{\boldsymbol{\alpha}}_0$.

Note the following:

1. We have no control over the $\|\boldsymbol{\beta}_0\|_2^2$ term in (5), this term is determined entirely by the observation $\boldsymbol{y}$.

2. Since $\boldsymbol{\Sigma}$ is invertible (diagonal with $\sigma_r > 0$), we make the second term in (5) zero by taking

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

3. The vector $\boldsymbol{\alpha}_0$, representing the component in the null space of $\boldsymbol{A}$, plays no role in the optimization program (5). This means that the solution to our original least-squares problem (1) is not unique unless $R = N$ (i.e. $\boldsymbol{A}$ only has $\mathbf{0}$ in its null space). Combining this with the note above, we see that every vector of the form

$$\widetilde{\boldsymbol{x}} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{V}_0\boldsymbol{\alpha}_0, \tag{6}$$

is a minimizer of (1). When $R = N$, there is no $\boldsymbol{V}_0$ matrix, and the minimizer is unique.

4. The solutions in (6) all have the minimal residual value of

$$\|\boldsymbol{y} - \boldsymbol{A}\widetilde{\boldsymbol{x}}\|_2^2 = \|\boldsymbol{\beta}_0\|_2^2 = \|\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{y}\|_2^2.$$

When $R = M$ (i.e. $\boldsymbol{A}^{\mathrm{T}}$ has only $\mathbf{0}$ in its null space), there is no $\boldsymbol{U}_0$ matrix, and $\|\boldsymbol{y} - \boldsymbol{A}\widetilde{\boldsymbol{x}}\|_2^2 = 0$ for all minimizers. That is, we can always find at least one $\boldsymbol{x}$ such that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ exactly.

5. The solutions in (6) have size

$$\|\widetilde{\boldsymbol{x}}\|_2^2 = \|\boldsymbol{V}\hat{\boldsymbol{\alpha}}\|_2^2 + \|\boldsymbol{V}_0\hat{\boldsymbol{\alpha}}_0\|_2^2 = \|\hat{\boldsymbol{\alpha}}\|_2^2 + \|\hat{\boldsymbol{\alpha}}_0\|_2^2.$$

Thus we can choose the **minimum norm solution** of (1) by taking $\hat{\boldsymbol{\alpha}}_0 = \mathbf{0}$, i.e. by taking

$$\hat{\boldsymbol{x}}_{\mathrm{ls}} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y}.$$

Taking $\hat{\boldsymbol{\alpha}}_0 = \mathbf{0}$ also ensures that $\hat{\boldsymbol{x}}_{\mathrm{ls}}$ is in the row space of $\boldsymbol{A}$.

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

1. when $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ has a unique exact solution, it must be $\hat{\boldsymbol{x}}_{\text{ls}}$,
2. when an exact solution is not available, $\hat{\boldsymbol{x}}_{\text{ls}}$ is a minimizer of (1),
3. when there are an infinite number of minimizers to (1), $\hat{\boldsymbol{x}}_{\text{ls}}$ is the one with smallest norm.

Because the matrix $\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\text{T}}$ 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) $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\text{T}}$ is

$$\boldsymbol{A}^{\dagger} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\text{T}}. \tag{7}$$

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 $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\text{T}}$ 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

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

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

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

and its columns are linearly independent. To verify equation (8), 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 (7).

- 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

$$\boldsymbol{A}^{\dagger} = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}. \tag{9}$$

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

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

  and its rows are linearly independent. To verify equation (9), 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 (7).

$\boldsymbol{A}^{\dagger}$ 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

$$\min_{\boldsymbol{H} \in \mathbb{R}^{N \times M}} \|\boldsymbol{H}\boldsymbol{A} - \mathbf{I}\|_F^2. \tag{10}$$

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 (10), 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 (10). To see this, recognize (see the exercise below) that the solution $\hat{\boldsymbol{H}}$ to (10) must obey

$$\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\hat{\boldsymbol{H}}^{\mathrm{T}} = \boldsymbol{A}. \tag{11}$$

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

$$\underset{\boldsymbol{H} \in \mathbb{R}^{N \times M}}{\text{minimize}} \ \|\boldsymbol{A}\boldsymbol{H} - \mathbf{I}\|_F^2. \tag{12}$$

The solution $\hat{\boldsymbol{H}}$ to (12) (see the exercise below) must obey

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\hat{\boldsymbol{H}} = \boldsymbol{A}^{\mathrm{T}}. \tag{13}$$

Again, we show that $\boldsymbol{A}^\dagger$ satisfies (13), and hence is a minimizer to (12):

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

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 (10) must obey (11). 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

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

to be a solution to (10). Do the same for (12) and (13).

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 $\boldsymbol{x}_0$:

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

where $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\text{T}}$ 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 recon-

struction 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.

We will start by studying the reconstruction error relative to $\boldsymbol{x}_{\text{pinv}}$:

$$\|\hat{\boldsymbol{x}}_{\text{ls}} - \boldsymbol{x}_{\text{pinv}}\|_2^2 = \|\boldsymbol{A}^{\dagger}\boldsymbol{e}\|_2^2 = \|\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2. \tag{14}$$

We want to understand how this error scales with the size of the noise $\|\boldsymbol{e}\|_2^2$.

We start by noting

$$\|\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2 \;=\; \|\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2,$$

which follows simply from the fact that $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \mathbf{I}$. Using the same reasoning as when we discussed solving symmetric systems, we know

$$\frac{1}{\sigma_1^2}\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2 \;\le\; \|\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2 \;\le\; \frac{1}{\sigma_R^2}\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2$$

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

It is always the case that $\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2 \leq \|\boldsymbol{e}\|_2^2$, as

$$\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2 = \sum_{m=1}^{R} |\langle \boldsymbol{e}, \boldsymbol{u}_m \rangle|^2 \leq \sum_{m=1}^{M} |\langle \boldsymbol{e}, \boldsymbol{u}_m \rangle|^2 = \|\boldsymbol{e}\|_2^2,$$

where again the $\boldsymbol{u}_{R+1}, \ldots, \boldsymbol{u}_M$ are the columns of the complementary $\boldsymbol{U}_0$ matrix (equality will have above when $\boldsymbol{e} \in \mathrm{Range}(\boldsymbol{A})$). So the worst case error is

$$\|\hat{\boldsymbol{x}}_{\mathrm{ls}} - \boldsymbol{x}_{\mathrm{pinv}}\|_2^2 \;\leq\; \frac{1}{\sigma_R^2}\|\boldsymbol{e}\|_2^2,$$

and is achieved when $\boldsymbol{e}$ is aligned with $\boldsymbol{u}_r$.

For the lower bound, it is entirely possible that $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e} = \boldsymbol{0}$, even when $\boldsymbol{e} \neq \boldsymbol{0}$ is arbitrarily large. Indeed, any $\boldsymbol{e} \in \mathrm{Null}(\boldsymbol{A}^{\mathrm{T}})$ will completely vanish once the pseudo-inverse is applied. Thus is is possible that $\hat{\boldsymbol{x}}_{\mathrm{ls}} = \boldsymbol{x}_{\mathrm{pinv}}$ even when $\boldsymbol{e}$ is not zero. In general, we have the bounds

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

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.

The expected error above is like an average of the $1/\sigma_r^2$, but if $R < M$, we include $M - R$ zeros in this average. This is because $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}$ is immediately destroying the $M - R$ components of $\boldsymbol{e}$ that lie in $\mathrm{Null}(\boldsymbol{A}^{\mathrm{T}})$.

[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 $\boldsymbol{M}$, $\mathrm{E}[\|\boldsymbol{M}\boldsymbol{e}\|_2^2] = \nu^2 \,\mathrm{trace}(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M})$.

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.