The Singular Value Decomposition

We are interested in more than just sym+def matrices. But the eigenvalue decompositions discussed in the last section of notes will play a major role in solving general systems of equations

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \qquad \boldsymbol{y} \in \mathbb{R}^M, \;\; \boldsymbol{A} \text{ is } M \times N \;, \;\; \boldsymbol{x} \in \mathbb{R}^N.$$

We have seen that a symmetric positive definite matrix can be decomposed as $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}$, where $\boldsymbol{V}$ is an orthogonal matrix ($\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = \mathbf{I}$) whose columns are the eigenvectors of $\boldsymbol{A}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues of $\boldsymbol{A}$. Because both orthogonal and diagonal matrices are trivial to invert, this eigenvalue decomposition makes it very easy to solve systems of equations $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ and analyze the stability of these solutions.

The **singular value decomposition** (SVD) takes apart an arbitrary $M \times N$ matrix $\boldsymbol{A}$ in a similar manner. The SVD of a $M \times N$ matrix $\boldsymbol{A}$ with rank[1] R is

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

where

1. $\boldsymbol{U}$ is a $M \times R$ matrix

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{u}_1 & | & \boldsymbol{u}_2 & | & \cdots & | & \boldsymbol{u}_R \end{bmatrix},$$

   whose columns $\boldsymbol{u}_m \in \mathbb{R}^M$ are orthonormal. Note that while $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \mathbf{I}$, in general $\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} \neq \mathbf{I}$ when $R < M$.

[1] Recall that the rank of a matrix is the number of linearly independent columns of a matrix (which is always equal to the number of linearly independent rows).

2. $\boldsymbol{V}$ is a $N \times R$ matrix

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1 & | & \boldsymbol{v}_2 & | & \cdots & | & \boldsymbol{v}_R \end{bmatrix},$$

whose columns $\boldsymbol{v}_n \in R^N$ are orthonormal. Again, while $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \mathbf{I}$, in general $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} \neq \mathbf{I}$ when $R < N$.

3. $\boldsymbol{\Sigma}$ is a $R \times R$ diagonal matrix with positive entries:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & 0 & \cdots \\ 0 & \sigma_2 & 0 & \cdots \\ \vdots & & \ddots & \\ 0 & \cdots & \cdots & \sigma_R \end{bmatrix}.$$

We call the σ_r the **singular values** of $\boldsymbol{A}$. By convention, we will order them such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_R$.

4. The $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_R$ are eigenvectors of the positive semi-definite matrix $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$. Note 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 the singular values $\sigma_1, \ldots, \sigma_R$ are the square roots of the non-zero eigenvalues of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$.

5. Similarly,

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

and so the $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_R$ are eigenvectors of the positive semi-definite matrix $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$. Since the non-zero eigenvalues of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ are the same, the σ_r are also square roots of the eigenvalues of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$.

6. The rank R is the number of linearly independent columns of $\boldsymbol{A}$; this is the same as the number of linearly independent rows. Thus $R \leq \min(M, N)$. We say $\boldsymbol{A}$ is **full rank** if $R = \min(M, N)$.

7. As $\boldsymbol{A}$ is rank R, its rows span an R-dimensional linear subspace of $\mathbb{R}^N$. As we have seen, this is called the **row space** of $\boldsymbol{A}$:
$$\begin{aligned}\operatorname{Row}(\boldsymbol{A}) &= \operatorname{Col}(\boldsymbol{A}^{\mathrm{T}}) \\ &= \{\boldsymbol{w} \in \mathbb{R}^N \ : \ \boldsymbol{w} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{z} \text{ for some } \boldsymbol{z} \in \mathbb{R}^M\}.\end{aligned}$$
The columns of $\boldsymbol{V}$ form an orthobasis for $\operatorname{Row}(\boldsymbol{A})$.

8. Recall that the **null space** of $\boldsymbol{A}$,
$$\operatorname{Null}(\boldsymbol{A}) = \{\boldsymbol{w} \in \mathbb{R}^N \ : \ \boldsymbol{A}\boldsymbol{w} = \mathbf{0}\},$$
is orthogonal to the row space. For $\boldsymbol{x}_1 \in \operatorname{Row}(\boldsymbol{A})$ and $\boldsymbol{x}_2 \in \operatorname{Null}(\boldsymbol{A})$, we have
$$\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle = \langle \boldsymbol{A}^{\mathrm{T}}\boldsymbol{z}, \boldsymbol{x}_2 \rangle = \langle \boldsymbol{z}, \boldsymbol{A}\boldsymbol{x}_2 \rangle = \langle \boldsymbol{z}, \mathbf{0} \rangle = 0.$$
The null space has dimension $N - R$, and so is spanned by some set of orthonormal basis vectors $\boldsymbol{v}_{R+1}, \ldots, \boldsymbol{v}_N$ that we can collect into an $N \times (N - R)$ matrix $\boldsymbol{V}_0$:
$$\boldsymbol{V}_0 = \begin{bmatrix} \boldsymbol{v}_{R+1} & | & \boldsymbol{v}_{R+2} & | & \cdots & | & \boldsymbol{v}_N \end{bmatrix}.$$
Note that $\boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{V}_0 = \mathbf{I}$ and $\boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{V} = \mathbf{0}$.

9. As $\boldsymbol{A}$ is rank R, it columns span an R-dimensional subspace of $\mathbb{R}^M$. As we have seen, this is called the **column space** of $\boldsymbol{A}$:
$$\operatorname{Col}(\boldsymbol{A}) = \{\boldsymbol{z} \in \mathbb{R}^M \ : \ \boldsymbol{z} = \boldsymbol{A}\boldsymbol{w} \text{ for some } \boldsymbol{w} \in \mathbb{R}^N\}.$$

The columns of $\boldsymbol{U}$ form an orthobasis for $\mathrm{Col}(\boldsymbol{A})$.

10. The null space of $\boldsymbol{A}^{\mathrm{T}}$, sometimes referred to as the **left null space** of $\boldsymbol{A}$,

$$\mathrm{Null}(\boldsymbol{A}^{\mathrm{T}}) = \{\boldsymbol{z} \in \mathbb{R}^M \ : \ \boldsymbol{A}^{\mathrm{T}}\boldsymbol{z} = \boldsymbol{0}\},$$

is orthogonal to the column space. For $\boldsymbol{y}_1 \in \mathrm{Range}(\boldsymbol{A})$ and $\boldsymbol{y}_2 \in \mathrm{Null}(\boldsymbol{A}^{\mathrm{T}})$, we have

$$\langle \boldsymbol{y}_1, \boldsymbol{y}_2 \rangle = \langle \boldsymbol{A}\boldsymbol{w}, \boldsymbol{y}_2 \rangle = \langle \boldsymbol{w}, \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}_2 \rangle = \langle \boldsymbol{w}, \boldsymbol{0} \rangle = 0.$$

The left null space has dimension $M - R$, and so is spanned by some set of orthonormal basis vectors $\boldsymbol{u}_{R+1}, \ldots, \boldsymbol{u}_M$ that we can collect into an $M \times (M - R)$ matrix $\boldsymbol{U}_0$:

$$\boldsymbol{U}_0 = \begin{bmatrix} \boldsymbol{u}_{R+1} & | & \boldsymbol{u}_{R+2} & | & \cdots & | & \boldsymbol{u}_M \end{bmatrix}.$$

Note that $\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{U}_0 = \mathbf{I}$ and $\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{0}$.

11. An equivalent way to write the SVD is as

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

where

$$\boldsymbol{U}_{\mathrm{full}} = \begin{bmatrix} \boldsymbol{U} & \boldsymbol{U}_0 \end{bmatrix}, \quad \boldsymbol{V}_{\mathrm{full}} = \begin{bmatrix} \boldsymbol{V} & \boldsymbol{V}_0 \end{bmatrix}, \quad \boldsymbol{\Sigma}_{\mathrm{full}} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{0}_{R\times(N-R)} \\ \boldsymbol{0}_{(M-R)\times R} & \boldsymbol{0}_{(M-R)\times(N-R)} \end{bmatrix}.$$

Now, $\boldsymbol{U}_{\mathrm{full}}$ is an $M \times M$ orthonormal matrix with $\boldsymbol{U}_{\mathrm{full}}\boldsymbol{U}_{\mathrm{full}}^{\mathrm{T}} = \mathbf{I}$, similarly $\boldsymbol{V}_{\mathrm{full}}$ is $N \times N$ with $\boldsymbol{V}_{\mathrm{full}}\boldsymbol{V}_{\mathrm{full}}^{\mathrm{T}} = \mathbf{I}$, and $\boldsymbol{\Sigma}_{\mathrm{full}}$ is $M \times N$ (the same sizes as $\boldsymbol{A}$) with a diagonal matrix in its upper left corner. In fact, this is the factorization the MATLAB command `svd` returns.

As before, we will often times find it useful to write the SVD as the sum of R rank-1 matrices:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} = \sum_{r=1}^{R} \sigma_r \, \boldsymbol{u}_r \boldsymbol{v}_r^{\mathrm{T}}.$$

When $\boldsymbol{A}$ is **overdetermined** ($M > N$), the decomposition looks like this

$$\begin{bmatrix} \\ \\ \boldsymbol{A} \\ \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \\ \boldsymbol{U} \\ \\ \\ \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_R \end{bmatrix} \begin{bmatrix} & \boldsymbol{V}^{\mathrm{T}} & \\ & & \end{bmatrix}$$

When $\boldsymbol{A}$ is **underdetermined** ($M < N$), the SVD looks like this

$$\begin{bmatrix} & & & \\ & & \boldsymbol{A} & & \\ & & & \end{bmatrix} = \begin{bmatrix} & \\ \boldsymbol{U} \\ & \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_R \end{bmatrix} \begin{bmatrix} & & & \\ & & \boldsymbol{V}^{\mathrm{T}} & & \\ & & & \end{bmatrix}$$

When $\boldsymbol{A}$ is **square** and full rank ($M = N = R$), the SVD looks like

$$\begin{bmatrix} & \\ \boldsymbol{A} \\ & \end{bmatrix} = \begin{bmatrix} & \\ \boldsymbol{U} \\ & \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_N \end{bmatrix} \begin{bmatrix} & \\ \boldsymbol{V}^{\mathrm{T}} \\ & \end{bmatrix}$$

The SVD and Least-Squares

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.

Recall our **least-squares** framework that revolves the optimization program

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

where $\|\cdot\|_2$ is the standard Euclidean norm. Given $\boldsymbol{y}$ and $\boldsymbol{A}$, solving (1) has the advantages that

1. when there is a $\boldsymbol{x}$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$, unique solution, it is one of the solutions;
2. when there is no solution, we return something reasonable.

When there are an infinite number, we need a procedure for choosing one of them. In this case, we will return the solution with smallest norm; we have seen before that this corresponds to solving

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \ \|\boldsymbol{x}\|_2^2 \quad \text{subject to} \quad \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}.$$

We will see that the SVD of $\boldsymbol{A}$:

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

immediately reveals the solution to 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[2]

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

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,

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

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

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

[2]In short, the decomposition (3) is possible since $\mathrm{Row}(\boldsymbol{A})$ and $\mathrm{Null}(\boldsymbol{A})$ are orthogonal complements in $\mathbb{R}^N$ for any $M \times N$ matrix $\boldsymbol{A}$. Every vector in $\mathbb{R}^N$ can be written as a sum of components from $\mathrm{Row}(\boldsymbol{A})$ and $\mathrm{Null}(\boldsymbol{A})$, 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

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

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, $\hat{\boldsymbol{x}}_{\mathrm{ls}} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y}$ 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 $\hat{\boldsymbol{x}}_{\mathrm{ls}}$,
2. when an exact solution is not available, $\hat{\boldsymbol{x}}_{\mathrm{ls}}$ is a minimizer of (1),
3. when there are an infinite number of minimizers to (1), $\hat{\boldsymbol{x}}_{\mathrm{ls}}$ 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

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

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

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

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

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

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

Technical Details: Existence of the SVD

In this section we will prove that any $M \times N$ matrix $\boldsymbol{A}$ with $\operatorname{rank}(\boldsymbol{A}) = R$ can be written as

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

where $\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}$ have the five properties listed at the beginning of the last section.

Since $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ is symmetric positive semi-definite, we can write:

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \sum_{n=1}^{N} \lambda_n \boldsymbol{v}_n \boldsymbol{v}_n^{\mathrm{T}},$$

where the $\boldsymbol{v}_n$ are orthonormal and the λ_n are real and non-negative. Since $\operatorname{rank}(\boldsymbol{A}) = R$, we also have $\operatorname{rank}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}) = R$, and so $\lambda_1, \ldots, \lambda_R$ are all strictly positive above, and $\lambda_{R+1} = \cdots = \lambda_N = 0$.

Set

$$\boldsymbol{u}_m = \frac{1}{\sqrt{\lambda_m}} \boldsymbol{A}\boldsymbol{v}_m, \quad \text{for } m = 1, \ldots, R, \qquad \boldsymbol{U} = \begin{bmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_R \end{bmatrix}.$$

Notice that these $\boldsymbol{u}_m$ are orthonormal, as

$$\langle \boldsymbol{u}_m, \boldsymbol{u}_\ell \rangle = \frac{1}{\sqrt{\lambda_m \lambda_\ell}} \boldsymbol{v}_\ell^{\mathrm{T}} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{v}_m = \sqrt{\frac{\lambda_m}{\lambda_\ell}}\, \boldsymbol{v}_\ell^{\mathrm{T}} \boldsymbol{v}_m = \begin{cases} 1, & m = \ell, \\ 0, & m \neq \ell. \end{cases}$$

These $\boldsymbol{u}_m$ also happen to be eigenvectors of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$, as

$$\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{u}_m = \frac{1}{\sqrt{\lambda_m}} \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{v}_m = \sqrt{\lambda_m}\boldsymbol{A}\boldsymbol{v}_m = \lambda_m \boldsymbol{u}_m.$$

Now let $\boldsymbol{u}_{R+1}, \ldots, \boldsymbol{u}_M$ be an orthobasis for the null space of $\boldsymbol{U}^{\mathrm{T}}$ — concatenating these two sets into $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M$ forms an orthobasis for all of $\mathbb{R}^M$.

Let

$$\boldsymbol{V} = \begin{bmatrix}\boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_R\end{bmatrix}, \quad \boldsymbol{V}_0 = \begin{bmatrix}\boldsymbol{v}_{R+1} & \boldsymbol{v}_{R+2} & \cdots & \boldsymbol{v}_N\end{bmatrix}, \quad \boldsymbol{V}_{\text{full}} = \begin{bmatrix}\boldsymbol{V} & \boldsymbol{V}_0\end{bmatrix}$$

and

$$\boldsymbol{U}_0 = \begin{bmatrix}\boldsymbol{u}_{R+1} & \boldsymbol{u}_{R+2} & \cdots & \boldsymbol{u}_M\end{bmatrix}, \quad \boldsymbol{U}_{\text{full}} = \begin{bmatrix}\boldsymbol{U} & \boldsymbol{U}_0\end{bmatrix}.$$

It should be clear that $\boldsymbol{V}_{\text{full}}$ is an $N \times N$ orthonormal matrix and $\boldsymbol{U}_{\text{full}}$ is a $M \times M$ orthonormal matrix. Consider the $M \times N$ matrix $\boldsymbol{U}_{\text{full}}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_{\text{full}}$ — the entry in the mth rows and nth column of this matrix is

$$(\boldsymbol{U}_{\text{full}}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_{\text{full}})[m,n] = \boldsymbol{u}_m^{\mathrm{T}}\boldsymbol{A}\boldsymbol{v}_n = \begin{cases}\sqrt{\lambda_n}\,\boldsymbol{u}_m^{\mathrm{T}}\boldsymbol{u}_n & n = 1,\ldots,R\\ 0, & n = R+1,\ldots,N.\end{cases}$$
$$= \begin{cases}\sqrt{\lambda_n}, & m = n = 1,\ldots,R\\ 0, & \text{otherwise.}\end{cases}$$

Thus

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

where

$$\Sigma_{\text{full}}[m,n] = \begin{cases}\sqrt{\lambda_n}, & m = n = 1,\ldots,R\\ 0, & \text{otherwise.}\end{cases}$$

Since $\boldsymbol{U}_{\text{full}}\boldsymbol{U}_{\text{full}}^{\mathrm{T}} = \mathbf{I}$ and $\boldsymbol{V}_{\text{full}}\boldsymbol{V}_{\text{full}}^{\mathrm{T}} = \mathbf{I}$, we have

$$\boldsymbol{A} = \boldsymbol{U}_{\text{full}}\boldsymbol{\Sigma}_{\text{full}}\boldsymbol{V}_{\text{full}}^{\mathrm{T}}.$$

Since $\boldsymbol{\Sigma}_{\text{full}}$ is non-zero only in the first R locations along its main diagonal, the above reduces to

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}, \quad \boldsymbol{\Sigma} = \begin{bmatrix}\sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_R}\end{bmatrix}.$$