The singular value decomposition

While we have now derived the solution to the least squares problem (at least in certain special cases), our solution so far has not provided much insight into what kind of properties we can expect the solution to have. A particularly useful lens for thinking about least squares is through the **singular value decomposition** of the matrix $\boldsymbol{A}$ in the objective function $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$. The singular value decomposition, or SVD for short, is just one of a family of useful matrix decompositions that you might encounter in data science. In general, a matrix decomposition is where we take a matrix $\boldsymbol{A}$ and re-express it as a product (or sometimes a sum) of other simpler matrices that more clearly reveal some important structure of $\boldsymbol{A}$.

The SVD of a real-valued[1] $M \times N$ matrix $\boldsymbol{A}$ is simply a factorization of the form

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

where $\boldsymbol{U}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{V}$ satisfy a number of properties, described below. Note that these properties involve a number of concepts from linear algebra such as the notion of an **orthnormal basis**, the **rank** of a matrix, and the concept of **eigenvectors** and **eigenvalues**. If these are a little fuzzy, review the linear algebra primer at the end of these notes before proceeding.

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

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

[1] You can just as easily define the SVD for complex-valued matrices, but then every matrix transpose has to be replaced with a *Hermetian transpose*, in which you take both a transpose and compute the complex conjugate of all the entries. In this class we will only work with real-valued matrices, so we will stick with the simpler notation.

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 $R < M$. The columns of $\boldsymbol{U}$ are an orthobasis for the column space of $\boldsymbol{A}$. The columns of $\boldsymbol{U}$ are called the **left-singular vectors** of $\boldsymbol{A}$.

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

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

whose columns $\boldsymbol{v}_n \in \mathbb{R}^N$ are orthonormal. Again, while $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \mathbf{I}$, in general $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} \neq \mathbf{I}$ when $R < N$. The columns of $\boldsymbol{V}$ are an orthobasis for the row space of $\boldsymbol{A}$. The columns of $\boldsymbol{V}$ are called the **right-singular vectors** of $\boldsymbol{A}$.

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

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

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

What is important for our purposes is that *any* matrix (no matter its dimensions or entries) can be factorized in this form.

We can say a little bit more about how to interpret the SVD by recalling the notion of an **eigenvector** and its corresponding **eigenvalue**. Recall that for a square matrix $\boldsymbol{B}$, we call a vector $\boldsymbol{x}$ an eigenvector if it satisfies

$$\boldsymbol{B}\boldsymbol{x} = \lambda\boldsymbol{x},$$

for some $\lambda \in \mathbb{C}$. We call λ the eigenvalue associated with $\boldsymbol{x}$.

Note that it is possible for $\boldsymbol{B}$ to have complex eigenvalues, even if $\boldsymbol{B}$ contains only real numbers. However, an important fact is that if $\boldsymbol{B}$ is symmetric, i.e., if $\boldsymbol{B} = \boldsymbol{B}^{\mathrm{T}}$, then all of its eigenvalues are *real*. An important class of matrices that we will arise frequently throughout this course are: **positive semidefinite matrices**. When we say that a matrix is positive semidefinite, we mean that it is symmetric and all of its eigenvalues are non-negative. If a symmetric matrix has eigenvalues that are strictly greater than zero, we call it **positive definite**.

There is a lot more that we could say about eigenvectors and eigenvalues, but for now all we want to point out is that the $\boldsymbol{u}_m$ and the $\boldsymbol{v}_n$ in the SVD can be interpreted as eigenvectors of matrices related to $\boldsymbol{A}$. For $M \neq N$, $\boldsymbol{A}$ does not even have eigenvectors, but the $\boldsymbol{u}_m$ are eigenvectors of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$, and the $\boldsymbol{v}_n$ are eigenvectors of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$.

To see this, note that

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

since $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \mathbf{I}$. Note that since $\boldsymbol{\Sigma}$ is a diagonal matrix, $\boldsymbol{\Sigma}^2$ is just the same diagonal matrix, but where the entries along the diagonal are squared: σ_r^2. Now consider $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{u}_m$. If we let $\boldsymbol{e}_m$ denote the m^{th} "standard basis element", i.e., the vector of all zeros with a single 1 in the m^{th} entry, then note that:

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{A}^{T}\boldsymbol{u}_m &= \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^{\mathrm{T}}\boldsymbol{u}_m \\
&= \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{e}_m \\
&= \boldsymbol{U}\sigma_m^2\boldsymbol{e}_m \\
&= \sigma_m^2\boldsymbol{u}_m.
\end{aligned}$$

This shows that $\boldsymbol{u}_m$ is an eigenvector of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$. Moreover, σ_m^2 is the corresponding eigenvalue. Thus, the singular values $\sigma_1, \ldots, \sigma_R$ are the square roots of the non-zero eigenvalues of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$. This also shows that $\boldsymbol{A}\boldsymbol{A}^T$, in addition to being symmetric (since $(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{\mathrm{T}} = \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$), is in fact positive semidefinite, since σ_m^2 is always positive.

By essentially the same argument, you can show that $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_R$ are eigenvectors of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$, and that the singular values $\sigma_1, \ldots, \sigma_R$ are *also* the square roots of the non-zero eigenvalues of the positive semidefinite matrix $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$.

The rank R is the dimension of the space spanned by the columns of $\boldsymbol{A}$, this is the same as the dimension of the space spanned by the rows. Thus $R \leq \min(M, N)$. We say $\boldsymbol{A}$ is **full rank** if $R = \min(M, N)$.

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

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

Alternative forms of the SVD

In most popular software packages, if you ask it to compute the SVD of a matrix $\boldsymbol{A}$, it will return matrices $\boldsymbol{U}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{V}$, but the dimensions will differ from what I have written above. In particular, the default output will be:

- an $M \times M$ matrix $\widetilde{\boldsymbol{U}}$,
- an $M \times N$ matrix $\widetilde{\boldsymbol{\Sigma}}$,
- and $N \times N$ matrix $\widetilde{\boldsymbol{V}}$.

So how does this output relate to what we have given above? Quite simply, $\boldsymbol{U}$ is the first R columns of $\widetilde{\boldsymbol{U}}$, $\boldsymbol{\Sigma}$ is the first R columns and rows of $\widetilde{\boldsymbol{\Sigma}}$, and $\boldsymbol{V}$ is the first R columns of $\widetilde{\boldsymbol{V}}$.

This raises the question of what is happening in the remaining columns of $\widetilde{\boldsymbol{U}}$, $\widetilde{\boldsymbol{\Sigma}}$, and $\widetilde{\boldsymbol{V}}$. For $\widetilde{\boldsymbol{\Sigma}}$, the additional rows and columns are all zeros. In the case where $R < M$, so that $\boldsymbol{U} \neq \widetilde{\boldsymbol{U}}$, we can write

$$\widetilde{\boldsymbol{U}} = \left[\, \boldsymbol{U} \mid \boldsymbol{U}_0 \, \right],$$

where $\boldsymbol{U}_0$ is an $M \times (M - R)$ matrix whose columns are an orthonormal basis for the part of $\mathbb{R}^M$ that is orthogonal to the columns of $\boldsymbol{U}$. If $R < N$, so that $\boldsymbol{V} \neq \widetilde{\boldsymbol{V}}$, we can similarly write

$$\widetilde{\boldsymbol{V}} = \left[\, \boldsymbol{V} \mid \boldsymbol{V}_0 \, \right],$$

where $\boldsymbol{V}_0$ is an $N \times (N - R)$ matrix whose columns are an orthonormal basis for the part of $\mathbb{R}^N$ orthogonal to the columns of $\boldsymbol{V}$.

Note that, since $\widetilde{\boldsymbol{\Sigma}}$ is just a zero-padded version of $\boldsymbol{\Sigma}$, it is not hard to show that

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

The zeros in $\widetilde{\Sigma}$ simply eliminate the contribution from $\boldsymbol{U}_0$ and $\boldsymbol{V}_0$, so we are equally entitled to write $\boldsymbol{A} = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{V}}^{\mathrm{T}}$. Just as was the case with $\boldsymbol{U}$ and $\boldsymbol{V}$, we can write

$$\widetilde{\boldsymbol{U}}^{\mathrm{T}}\widetilde{\boldsymbol{U}} = \mathbf{I} \qquad \widetilde{\boldsymbol{V}}^{\mathrm{T}}\widetilde{\boldsymbol{V}} = \mathbf{I}.$$

However, in this case, since $\widetilde{\boldsymbol{U}}$ and $\widetilde{\boldsymbol{V}}$ are square matrices, we also have that

$$\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^{\mathrm{T}} = \mathbf{I} \qquad \widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^{\mathrm{T}} = \mathbf{I}.$$

This is a consequence of a general fact about square matrices that you will prove on the homework. An equivalent way to think about this fact is that $\widetilde{\boldsymbol{U}}$ is an orthonormal basis for all of $\mathbb{R}^M$ and $\widetilde{\boldsymbol{V}}$ is an orthonormal basis for all of $\mathbb{R}^N$. Please remember, however, that in general this is not true for $\boldsymbol{U}$ and $\boldsymbol{V}$.

The SVD and least squares

So what does the SVD have to do with least squares? Recall that we had previously argued that if the matrix $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ is invertible, then the optimization problem

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

has solution

$$\widehat{\boldsymbol{x}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}. \tag{2}$$

Now what can we say about this formula using the fact that we can write $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$? First, note that

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

Next, we use the fact that

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

This is easy to check:

$$\boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^{\mathrm{T}}(\boldsymbol{V}\boldsymbol{\Sigma}^{-2}\boldsymbol{V}^{\mathrm{T}}) = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{\Sigma}^{-2}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = \mathbf{I}.$$

Recall that, in general, we cannot always conclude that $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = \mathbf{I}$. However, here we assume that the $N \times N$ matrix $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ is invertible. This matrix can only be invertible when it is full-rank, meaning that $R = N$, in which case, we also have $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = \mathbf{I}$.

Returning to the least squares problem, using the above, we have that

$$\begin{aligned}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}} &= \boldsymbol{V}\boldsymbol{\Sigma}^{-2}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}} \\ &= \boldsymbol{V}\boldsymbol{\Sigma}^{-2}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}} \\ &= \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}.\end{aligned}$$

Thus, we arrive at an alternative way to express the solution to (1), now in terms of the SVD of $\boldsymbol{A}$:

$$\widehat{\boldsymbol{x}} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y}. \tag{3}$$

This way of writing the least squares solution has a few nice advantages. First, recall that our previous formula in (2) only applied when $\boldsymbol{A}^T\boldsymbol{A}$ was invertible. It is not too hard to show (although we will not do so here) that (3) provides the solution to (1) regardless of whether $\boldsymbol{A}^T\boldsymbol{A}$ is invertible or not.

Moreover, it also provides the solution to a related problem. Suppose that we have a system $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ which is *underdetermined*, meaning

that $M < N$. In this case, the least squares problem in (1) has infinitely many solutions. An alternative approach that might make more sense in this context (which can also be thought of as a kind of "least squares") is to, from all $\boldsymbol{x}$ that satisfy $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, choose the one that is "smallest" in the Euclidean norm sense. As an optimization problem we could write this as:

$$\underset{\boldsymbol{x} \in \mathbb{R}^N}{\text{minimize}} \ \|\boldsymbol{x}\|_2^2 \ \text{ subject to } \ \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}. \tag{4}$$

It turns out that $\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y}$ is also the solution to (4).

Finally, it is relatively straightforward to show that in the case where $\boldsymbol{A}$ is square and invertible ($M = N = R$), then $\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}$ is simply $\boldsymbol{A}^{-1}$.

Least squares in noise

The real advantage to thinking about least squares through the lens of the SVD is that it tells us exactly what to expect when our measurements are corrupted with *noise*. Specifically, suppose we observe

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

where $\boldsymbol{e} \in \mathbb{R}^M$ is an (unknown) perturbation of our measurements. Suppose that we then form the least squares estimate:

$$\widehat{\boldsymbol{x}}_{\text{ls}} = \boldsymbol{A}^{\dagger}\boldsymbol{y} = \boldsymbol{A}^{\dagger}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}^{\dagger}\boldsymbol{e}. \tag{5}$$

We would like to understand the effect of the noise vector $\boldsymbol{e}$ on our estimate of $\boldsymbol{x}$. One way to understand this is to compare $\widehat{\boldsymbol{x}}_{\text{ls}}$ to what

we would have obtained if we had been able to use least squares on the noise-free observations $\boldsymbol{y}_{\text{clean}} = \boldsymbol{A}\boldsymbol{x}$. The noise-free reconstruction is

$$\widehat{\boldsymbol{x}}_{\text{clean}} = \boldsymbol{A}^{\dagger}\boldsymbol{y} = \boldsymbol{A}^{\dagger}\boldsymbol{A}\boldsymbol{x}.$$

Combining this with (5) we see that the additional error due to noise can be measured as

$$\|\widehat{\boldsymbol{x}}_{\text{ls}} - \widehat{\boldsymbol{x}}_{\text{clean}}\|_2^2 = \|\boldsymbol{A}^{\dagger}\boldsymbol{e}\|_2^2 = \|\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2. \tag{6}$$

Now suppose for a moment that the error has unit norm, $\|\boldsymbol{e}\|_2 = 1$. Just how bad can the error be? The worst case for (6) in this case is given by

$$\underset{\boldsymbol{e}\in\mathbb{R}^M}{\text{maximize}} \ \|\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{e}\|_2 = 1. \tag{7}$$

Note that $\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}$ is a vector in $\mathbb{R}^R$, and that for any vector $\boldsymbol{z} \in \mathbb{R}^R$, we have

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

since the columns of $\boldsymbol{V}$ are orthonormal. Thus, we can simplify (7) to

$$\underset{\boldsymbol{e}\in\mathbb{R}^M}{\text{maximize}} \ \|\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{e}\|_2 = 1. \tag{8}$$

We can simplify a bit further by noticing that, since the columns of $\boldsymbol{U}$ are orthonormal, $\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2 \leq \|\boldsymbol{e}\|_2^2$. To see why this is the case, recall from our discussion of alternative forms of the SVD that when $R < M$, we can form the matrix

$$\widetilde{\boldsymbol{U}} = \left[\ \boldsymbol{U} \mid \boldsymbol{U}_0\ \right],$$

where $\widetilde{\boldsymbol{U}}$ is an $M \times M$ orthonormal matrix. It follows immediately that

$$\|\widetilde{\boldsymbol{U}}^{\mathrm{T}}\boldsymbol{e}\|_2^2 = \boldsymbol{e}^{\mathrm{T}}\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^{\mathrm{T}}\boldsymbol{e} = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} = \|\boldsymbol{e}\|_2^2,$$

since $\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^{\mathrm{T}} = \mathbf{I}$. But it should also be clear that since $\widetilde{\boldsymbol{U}}^{\mathrm{T}}\boldsymbol{e}$ is just the vector of inner products between the columns in $\widetilde{\boldsymbol{U}}$ and $\boldsymbol{e}$, and $\widetilde{\boldsymbol{U}}$ is just the concatenation of $\boldsymbol{U}$ and $\boldsymbol{U}_0$,

$$\|\widetilde{\boldsymbol{U}}^{\mathrm{T}}\boldsymbol{e}\|_2^2 = \|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2 + \|\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{e}\|_2^2,$$

and hence we must have $\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}\|_2^2 \leq \|\boldsymbol{e}\|_2^2$.

With this fact in hand, we can consider a slight "relaxation" of (8) to

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

We are not explicitly enforcing the fact that $\boldsymbol{\beta}$ should be a linear combination of the columns of $\boldsymbol{U}$ here, which is why I am calling this a relaxation of the original problem. However, this problem has a simple solution that you will verify in the homework. Specifically the worst case $\boldsymbol{\beta}$ 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 $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_R$.)

Note that, even though we ignored the constraint that $\boldsymbol{\beta}$ should be a linear combination of the columns of $\boldsymbol{U}$ above, it is easy to find an $\boldsymbol{e}$ such that $\boldsymbol{\beta} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}$ gives us a 1 in the entry corresponding to σ_R.

In particular, $\boldsymbol{e} = \boldsymbol{u}_R$ gives us precisely this $\boldsymbol{\beta}$, and so even though we ignored this constraint, the result would not have changed had we included this requirement.

Returning to the reconstruction error (6), we now see that

$$\|\widehat{\boldsymbol{x}}_{\text{ls}} - \widehat{\boldsymbol{x}}_{\text{clean}}\|_2^2 \;=\; \|\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\text{T}}\boldsymbol{e}\|_2^2 \;\leq\; \frac{1}{\sigma_R^2}\|\boldsymbol{e}\|_2^2.$$

While this is a worst-case bound, notice that if σ_R is small, the worst case reconstruction error can be **very bad**. Later in this course we will discuss some strategies to mitigate this fact.

Linear Algebra Review II: Inner products, orthonormal bases, and eigenvectors

We have already encountered the notion of a general vector space with an abstract norm. One other key concept, which we have already treated informally, is the notion of an abstract *inner product*.

Inner products

Definition: An **inner product** on a real-valued[2] vector space $\mathcal{S}$ is a mapping

$$\langle \cdot, \cdot \rangle \ : \ \mathcal{S} \times \mathcal{S} \to \mathbb{R}$$

that obeys

1. $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$

2. For any $a, b \in \mathbb{R}$

$$\langle a\boldsymbol{x} + b\boldsymbol{y}, \boldsymbol{z} \rangle \ = \ a\langle \boldsymbol{x}, \boldsymbol{z} \rangle + b\langle \boldsymbol{y}, \boldsymbol{z} \rangle$$

3. $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0 \quad \text{and} \quad \langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{0}$

Standard Examples:

- $\mathcal{S} = \mathbb{R}^N$,

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \ = \ \sum_{n=1}^{N} x_n y_n \ = \ \boldsymbol{y}^{\mathrm{T}} \boldsymbol{x}$$

[2] By real-valued, we mean the scalar associated with scalar multiplication is a real-number. You can easily extend these definitions to a complex-valued vector space, but since we will not need this level of generality in this course we will stick with the simpler real-valued case.

- $\mathcal{S} = L_2([a,b])$,

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \;=\; \int_a^b x(t)y(t)\;\mathrm{d}t$$

Slightly less standard example:

- $\mathcal{S} = \mathbb{R}^{M \times N}$ (the set of $M \times N$ matrices with real entries)

$$\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \mathrm{trace}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X}) = \sum_{m=1}^{M}\sum_{n=1}^{N} X_{m,n}Y_{m,n}$$

  (Recall that $\mathrm{trace}(\boldsymbol{X})$ is the sum of the entries on the diagonal of $\boldsymbol{X}$.) This is called the *trace inner product* or *Frobenius inner product* or *Hilbert-Schmidt inner product*.

A vector space equipped with an inner product is called an **inner product space**. Inner products allow us to think about angles between vectors in arbitrary vector spaces – most importantly, we can generalize the notion of orthogonality from Euclidean space to an arbitrary inner product space: two vectors $\boldsymbol{x}, \boldsymbol{y}$ are **orthogonal** if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$.

Induced norms

A valid inner product induces a valid norm by

$$\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$$

(You can check that $\sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$ indeed satisfies the properties required of a norm on your own as an exercise.)

It is not hard to see that in $\mathbb{R}^N$, the standard inner product induces the ℓ_2 norm, i.e., $\sqrt{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}} = \|\boldsymbol{x}\|_2^2$.

Properties of induced norms

In addition to the triangle inequality,

$$\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|,$$

which all norms must obey, induced norms obey some very handy inequalities. Below I give two of the most famous and useful ones. Note that these are not necessarily true for norms in general, only for norms induced by an inner product:

- **Pythagorean Theorem**

  $$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0 \;\Rightarrow\; \|\boldsymbol{x} + \boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2$$

  The left-hand side above also implies that
  $\|\boldsymbol{x} - \boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2$.

- **Cauchy-Schwarz Inequality**

  $$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \;\leq\; \|\boldsymbol{x}\|\,\|\boldsymbol{y}\|$$

  Equality is achieved above when (and only when) $\boldsymbol{x}$ and $\boldsymbol{y}$ are **colinear**:

  $$\exists\, a \in \mathbb{R} \quad \text{such that} \quad \boldsymbol{y} = a\boldsymbol{x}.$$

The last remaining set of concepts from linear algebra that we need to review relate to the notion of a **basis** (and if we have an inner product, an **orthonormal basis**). To get offer formal definitions of these terms, we first need to review a few basic ideas.

Linear combinations and spans

Let $\mathcal{M} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N\}$ be a set of vectors in a real-valued vector space $\mathcal{S}$.

Definition: A **linear combination** of vectors in $\mathcal{M}$ is a sum of the form

$$a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_2 + \cdots + a_N\boldsymbol{v}_N$$

for some $a_1, a_2, \ldots, a_N \in \mathbb{R}$.

Definition: The **span** of $\mathcal{M}$ is the set of all linear combinations of $\mathcal{M}$. We write this as

$$\text{span}(\mathcal{M}) \;=\; \text{span}(\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N\})$$

Example:
Suppose

$$\mathcal{S} = \mathbb{R}^3, \qquad \boldsymbol{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \qquad \boldsymbol{v}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

In this case,

$$\text{span}(\{\boldsymbol{v}_1, \boldsymbol{v}_2\}) \;=\; \text{the } (x_1, x_2) \text{ plane.}$$

In other words, the span of $\boldsymbol{v}_1, \boldsymbol{v}_2$ is the set of any $\boldsymbol{x} \in \mathbb{R}^3$ of the form

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ 0 \end{bmatrix}.$$

It should be clear that for any such $\boldsymbol{x}$, we can write $\boldsymbol{x}$ as a linear combination of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$:

$$\boldsymbol{x} = x_1\boldsymbol{v}_1 + (x_2 - x_1)\boldsymbol{v}_2.$$

Linear dependence

A set of vectors $\{\boldsymbol{v}_j\}_{j=1}^N$ is said to be **linearly dependent** if there exists scalars $a_1, a_2, \ldots, a_N$, not all equal to 0, such that

$$\sum_{n=1}^{N} a_n \boldsymbol{v}_n = \boldsymbol{0}.$$

Note that in such a case, we can write (at least) one of the vectors $\boldsymbol{v}_n$ as a linear combination of the others.

Likewise, if $\sum_n a_n\boldsymbol{v}_n = \boldsymbol{0}$ only when all the $a_j = 0$, then $\{\boldsymbol{v}_n\}_{n=1}^N$ is said to be **linearly independent**.

Example:

$$\mathcal{S} = \mathbb{R}^3, \quad \boldsymbol{v}_1 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \quad \boldsymbol{v}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \quad \boldsymbol{v}_3 = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$$

It is too hard to find an a_1, a_2, a_3 such that

$$a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_2 + a_3\boldsymbol{v}_3 = \boldsymbol{0}.$$

Note that any two of the vectors above are linearly independent:

$$\begin{aligned} \text{span}(\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3\}) &= \text{span}(\{\boldsymbol{v}_1, \boldsymbol{v}_2\}) \\ &= \text{span}(\{\boldsymbol{v}_1, \boldsymbol{v}_3\}) \\ &= \text{span}(\{\boldsymbol{v}_2, \boldsymbol{v}_3\}). \end{aligned}$$

Bases

Definition: A **basis** of a finite-dimensional[3] vector space $\mathcal{S}$ is a set of vectors $\mathcal{B}$ such that

1. $\text{span}(\mathcal{B}) = \mathcal{S}$
2. $\mathcal{B}$ is linearly independent

The second condition ensures that all bases of $\mathcal{S}$ will have the same number of elements.

The **dimension** of $\mathcal{S}$ is the number of elements required in a basis for $\mathcal{S}$.

Examples:

1. $\mathbb{R}^N$ with

$$\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N\} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \cdots, \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \right\}$$

   This is the **standard basis** for $\mathbb{R}^N$.
   The dimension of $\mathbb{R}^N$ is N.

2. $\mathbb{R}^N$ with any set of N linearly independent vectors.

3. $\mathcal{S} = \{\text{polynomials of degree at most } p\}$.
   A basis for $\mathcal{S}$ is $\mathcal{B} = \{1, t, t^2, \ldots, t^p\}$.
   The dimension of $\mathcal{S}$ is $p + 1$.

[3] Things get a bit more complicated in infinite-dimensional vector spaces. We won't worry about those details in this class.

Column space, row space, and rank

Consider the matrix $\boldsymbol{V}$ which has the vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \ldots, \boldsymbol{v}_N \in \mathbb{R}^M$ as columns:

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

Note that this is an $M \times N$ matrix.

The **column space**, also commonly called the **range**, of $\boldsymbol{V}$ is just the span of the columns of $\boldsymbol{V}$, i.e., $\text{span}(\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N\})$. This is often denoted by $\mathcal{R}(\boldsymbol{V})$. Note that $\mathcal{R}(\boldsymbol{V}) \subseteq \mathbb{R}^M$ (since the $\boldsymbol{v}_j$ are vectors in $\mathbb{R}^M$). It may be possible to have $\mathcal{R}(\boldsymbol{V}) = \mathbb{R}^M$ – this will occur if $\boldsymbol{V}$ has M linearly independent columns. Note that $\boldsymbol{V}$ cannot have more than M linearly independent columns: since $\mathbb{R}^M$ is an M-dimensional space, you cannot have more than M linearly independent vectors within that space. Also, note that if $N < M$, then the columns of $\boldsymbol{V}$ cannot possibly span all of $\mathbb{R}^M$. Moreover, even if $N > M$, it is still possible that $\mathcal{R}(\boldsymbol{V})$ is a strict subspace of $\mathbb{R}^M$.

The **row space** is defined similarly, and is simply the span of the *rows* of $\boldsymbol{V}$, or equivalently, the row space is the column space of $\boldsymbol{V}^{\mathrm{T}}$. Note that the row space is a subspace of $\mathbb{R}^N$.

It is a fact that for any matrix $\boldsymbol{V}$, the column space and row space have the same dimension – that is, the maximum number of linearly independent columns is the same as the maximum number of linearly independent rows. This dimension is called the **rank** of $\boldsymbol{V}$.

Orthogonal bases

A collection of vectors $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N\}$ in a finite dimensional vector space $\mathcal{S}$ is called an **orthogonal basis** if

1. $\text{span}(\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N\}) = \mathcal{S}$,
2. $\langle \boldsymbol{v}_j, \boldsymbol{v}_k \rangle = 0$ for all $j \neq k$.

If in addition the vectors are normalized (under the induced norm),

$$\|\boldsymbol{v}_n\| = 1, \quad \text{for} \;\; n = 1, \ldots, N,$$

we will call it an **orthonormal basis** or **orthobasis**.

In the case where $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N$ are vectors in $\mathbb{R}^M$, we can form an $M \times N$ matrix $\boldsymbol{V}$ with them as columns:

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

In this case, another way to express that $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N$ are othornormal is that

$$\boldsymbol{V}^{\mathrm{T}} \boldsymbol{V} = \mathbf{I}.$$

To see this, note that $\boldsymbol{V}^{\mathrm{T}} \boldsymbol{v}_j$ computes the inner product between $\boldsymbol{v}_j$ and all N vectors in the basis:

$$\begin{bmatrix} — & \boldsymbol{v}_1^{\mathrm{T}} & — \\ & \vdots & \\ — & \boldsymbol{v}_N^{\mathrm{T}} & — \end{bmatrix} \begin{bmatrix} | \\ \boldsymbol{v}_j \\ | \end{bmatrix} = \begin{bmatrix} \langle \boldsymbol{v}_j, \boldsymbol{v}_1 \rangle \\ \vdots \\ \langle \boldsymbol{v}_j, \boldsymbol{v}_N \rangle \end{bmatrix}.$$

These inner products will all be zero except for the j^{th} element. Repeating this for $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots \boldsymbol{v}_N$ results in $\boldsymbol{V}^{\mathrm{T}} \boldsymbol{V} = \mathbf{I}$.