3 · Singular Value Decomposition

Thus far we have focused on matrix factorizations that reveal the eigenvalues of a square matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, such as the SCHUR factorization and the JORDAN canonical form. Eigenvalue-based decompositions are ideal for analyzing the behavior of dynamical systems likes $\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t)$ or $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k$. When it comes to solving linear systems of equations or tackling more general problems in data science, eigenvalue-based factorizations are often not so illuminating. In this chapter we develop another decomposition that provides deep insight into the rank structure of a matrix, showing the way to solving all variety of linear equations and exposing optimal low-rank approximations.

3.1 Singular Value Decomposition

The *singular value decomposition (SVD)* is remarkable factorization that writes a general rectangular matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$ in the form

$$\mathbf{A} = (\text{unitary matrix}) \times (\text{diagonal matrix}) \times (\text{unitary matrix})^*.$$

From the unitary matrices we can extract bases for the four fundamental subspaces $\mathcal{R}(\mathbf{A})$, $\mathcal{N}(\mathbf{A})$, $\mathcal{R}(\mathbf{A}^*)$, and $\mathcal{N}(\mathbf{A}^*)$, and the diagonal matrix will reveal much about the rank structure of $\mathbf{A}$.

We will build up the SVD in a four-step process. For simplicity suppose that $\mathbf{A} \in \mathbb{C}^{m\times n}$ with $m \geq n$. (If $m < n$, apply the arguments below to $\mathbf{A}^* \in \mathbb{C}^{n\times m}$.) Note that $\mathbf{A}^*\mathbf{A} \in \mathbb{C}^{n\times n}$ is always Hermitian positive semidefinite. (Clearly $(\mathbf{A}^*\mathbf{A})^* = \mathbf{A}^*(\mathbf{A}^*)^* = \mathbf{A}^*\mathbf{A}$, so $\mathbf{A}^*\mathbf{A}$ is Hermitian. For any $\mathbf{x} \in \mathbb{C}^n$, note that $\mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x} = (\mathbf{A}\mathbf{x})^*(\mathbf{A}\mathbf{x}) = \|\mathbf{A}\mathbf{x}\|^2 \geq 0$, so $\mathbf{A}^*\mathbf{A}$ is positive semidefinite.)

Step 1. Using the spectral decomposition of a Hermitian matrix discussed in Section 1.5, $\mathbf{A}^*\mathbf{A}$ has n eigenpairs $\{(\lambda_j, \mathbf{v}_j)\}_{j=1}^n$ with orthonormal unit eigenvectors ($\mathbf{v}_j^*\mathbf{v}_j = 1$, $\mathbf{v}_j^*\mathbf{v}_k = 0$ when $j \neq k$). We are free to pick any convenient indexing for these eigenpairs; label the eigenvalues in decreasing magnitude, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$.

Step 2. Define $s_j := \|\mathbf{A}\mathbf{v}_j\|$.
Note that $s_j^2 = \|\mathbf{A}\mathbf{v}_j\|^2 = \mathbf{v}_j^*\mathbf{A}^*\mathbf{A}\mathbf{v}_j = \lambda_j$. Since the eigenvalues $\lambda_1, \ldots, \lambda_n$ are decreasing in magnitude, so are the s_j values: $s_1 \geq s_2 \geq \cdots \geq s_n \geq 0$.

Step 3. Next, we will build a set of related orthonormal vectors in $\mathbb{C}^m$. Suppose we have already constructed such vectors $\mathbf{u}_1, \ldots, \mathbf{u}_{j-1}$.
If $s_j \neq 0$, then define $\mathbf{u}_j = s_j^{-1}\mathbf{A}\mathbf{v}_j$, so that $\|\mathbf{u}_j\| = s_j^{-1}\|\mathbf{A}\mathbf{v}_j\| = 1$.
If $s_j = 0$, then pick $\mathbf{u}_j$ to be any unit vector such that

$$\mathbf{u}_j \in \operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_{j-1}\}^\perp;$$

i.e., ensure $\mathbf{u}_j^*\mathbf{u}_k = 0$ for all $k < j$.[1]

By construction, $\mathbf{u}_j^*\mathbf{u}_k = 0$ for $j \neq k$ if s_j or s_k is zero. If both s_j and s_k are nonzero, then

$$\mathbf{u}_j^*\mathbf{u}_k = \frac{1}{s_j s_k}(\mathbf{A}\mathbf{v}_j)^*(\mathbf{A}\mathbf{v}_k) = \frac{1}{s_j s_k}\mathbf{v}_j^*\mathbf{A}^*\mathbf{A}\mathbf{v}_k = \frac{\lambda_k}{s_j s_k}\mathbf{v}_j^*\mathbf{v}_k,$$

where we used the fact that $\mathbf{v}_j$ is an eigenvector of $\mathbf{A}^*\mathbf{A}$. Now if $j \neq k$, then $\mathbf{v}_j^*\mathbf{v}_k = 0$, and hence $\mathbf{u}_j^*\mathbf{u}_k = 0$. On the other hand, $j = k$ implies that $\mathbf{v}_j^*\mathbf{v}_k = 1$, so $\mathbf{u}_j^*\mathbf{u}_k = \lambda_j/s_j^2 = 1$.

In conclusion, we have constructed a set of orthonormal vectors $\{\mathbf{u}_j\}_{j=1}^n$ with $\mathbf{u}_j \in \mathbb{C}^m$.
Step 4. For all $j = 1, \ldots, n$,

$$\mathbf{A}\mathbf{v}_j = s_j\mathbf{u}_j,$$

regardless of whether $s_j = 0$ or not. We can stack these n vector equations as columns of a single matrix equation,

$$\begin{bmatrix} | & | & & | \\ \mathbf{A}\mathbf{v}_1 & \mathbf{A}\mathbf{v}_2 & \cdots & \mathbf{A}\mathbf{v}_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ s_1\mathbf{u}_1 & s_2\mathbf{u}_2 & \cdots & s_n\mathbf{u}_n \\ | & | & & | \end{bmatrix}.$$

[1] If $s_j = 0$, then $\lambda_j = 0$, and so $\mathbf{A}^*\mathbf{A}$ has a zero eigenvalue; i.e., this matrix is singular.

Note that both matrices in this equation can be factored into the product of simpler matrices:

$$\mathbf{A}\begin{bmatrix} | & | & & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_n \end{bmatrix}.$$

Denote these matrices as $\mathbf{AV} = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}$, where $\mathbf{A} \in \mathbb{C}^{m\times n}$, $\mathbf{V} \in \mathbb{C}^{n\times n}$, $\widehat{\mathbf{U}} \in \mathbb{C}^{m\times n}$, and $\widehat{\mathbf{\Sigma}} \in \mathbb{C}^{n\times n}$.

The (j,k) entry of $\mathbf{V}^*\mathbf{V}$ is simply $\mathbf{v}_j^*\mathbf{v}_k$, and so $\mathbf{V}^*\mathbf{V} = \mathbf{I}$. Since $\mathbf{V}$ is a square matrix, we have just proved that it is unitary, and hence, $\mathbf{V}\mathbf{V}^* = \mathbf{I}$ as well. We conclude that

$$\mathbf{A} = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\mathbf{V}^*.$$

This matrix factorization is known as the *reduced singular value decomposition* or the *economy-sized singular value decomposition* (or, informally, the *skinny SVD*). It can be obtained via the MATLAB command

```
[Uhat, Sighat, V] = svd(A,'econ');
```

The Reduced Singular Value Decomposition

Theorem 3.1. *Any matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$ with $m \geq n$ can be written as*

$$\mathbf{A} = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\mathbf{V}^*,$$

where $\widehat{\mathbf{U}} \in \mathbb{C}^{m\times n}$ has orthonormal columns, $\mathbf{V} \in \mathbb{C}^{n\times n}$ is unitary, and $\widehat{\mathbf{\Sigma}} = \operatorname{diag}(s_1, \ldots, s_n) \in \mathbb{C}^{n\times n}$ has real nonnegative decreasing entries. The columns of $\mathbf{U}$ are left singular vectors, the columns of $\mathbf{V}$ are right singular vectors, and the values $s_1, \ldots, s_n$ are the singular values.

While the matrix $\widehat{\mathbf{U}}$ has orthonormal columns, *it is not a unitary matrix* when $m > n$. In particular, we have $\widehat{\mathbf{U}}^*\widehat{\mathbf{U}} = \mathbf{I} \in \mathbb{C}^{n\times n}$, but

$$\widehat{\mathbf{U}}\widehat{\mathbf{U}}^* \in \mathbb{C}^{m\times m}$$

cannot be the identity unless $m = n$. (To see this, note that $\widehat{\mathbf{U}}\widehat{\mathbf{U}}^*$ is an orthogonal projection onto $\mathcal{R}(\widehat{\mathbf{U}}) = \operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$. Since $\dim(\mathcal{R}(\widehat{\mathbf{U}})) = n$, this projection cannot equal the m-by-m identity matrix when $m > n$.)

Though $\widehat{\mathbf{U}}$ is not unitary, it is *subunitary*. We can construct $m-n$ additional column vectors to append to $\widehat{\mathbf{U}}$ to make it unitary. Here is the recipe: For $j = n+1, \ldots, m$, pick

$$\mathbf{u}_j \in \operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_{j-1}\}^{\perp}$$

with $\mathbf{u}_j^* \mathbf{u}_j = 1$. Then define

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

Confirm that $\mathbf{U}^*\mathbf{U} = \mathbf{U}\mathbf{U}^* = \mathbf{I} \in \mathbb{C}^{m\times m}$, showing that $\mathbf{U}$ is unitary.

We wish to replace the $\widehat{\mathbf{U}}$ in the reduced SVD with the unitary matrix $\mathbf{U}$. To do so, we also need to replace $\widehat{\mathbf{\Sigma}}$ by some $\mathbf{\Sigma}$ in such a way that $\widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}} = \mathbf{U}\mathbf{\Sigma}$. The simplest approach constructs $\mathbf{\Sigma}$ by appending zeros to the end of $\widehat{\mathbf{\Sigma}}$, thus ensuring there is no contribution when the new entries of $\mathbf{U}$ multiply against the new entries of $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} = \begin{bmatrix} \widehat{\mathbf{\Sigma}} \\ \mathbf{0} \end{bmatrix} \in \mathbb{C}^{m\times n}.$$

The factorization $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ is called the *full singular value decomposition*.

The Full Singular Value Decomposition

Theorem 3.2. *Any matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$ can be written in the form*

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*,$$

where $\mathbf{U} \in \mathbb{C}^{m\times m}$ and $\mathbf{V} \in \mathbb{C}^{n\times n}$ are unitary matrices and $\mathbf{\Sigma} \in \mathbb{C}^{m\times n}$ is zero except for the main diagonal;
The columns of $\mathbf{U}$ are left singular vectors, the columns of $\mathbf{V}$ are right singular vectors, and the values $s_1, \ldots, s_{\min\{m,n\}}$ are the singular values.

A third version of the singular value decomposition is often very helpful. Start with the reduced SVD $\mathbf{A} = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\mathbf{V}^*$ in Theorem 3.1 Multiply $\widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}$ together to get

$$\mathbf{A} = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\mathbf{V}^* = \begin{bmatrix} | & & | \\ s_1\mathbf{u}_1 & \cdots & s_n\mathbf{u}_n \\ | & & | \end{bmatrix} \begin{bmatrix} — & \mathbf{v}_1^* & — \\ & \vdots & \\ — & \mathbf{v}_n^* & — \end{bmatrix} = \sum_{j=1}^{n} s_j \mathbf{u}_j \mathbf{v}_j^*,$$

which renders $\mathbf{A}$ as the *sum of outer products* $\mathbf{u}_j\mathbf{v}_j^* \in \mathbb{C}^{m\times n}$, *weighted by nonnegative numbers* s_j. Let r denote the number of nonzero singular values, so that if $r < n$, then

$$s_{r+1} = \cdots = s_n = 0.$$

Thus $\mathbf{A}$ can be written as

$$\mathbf{A} = \sum_{j=1}^{r} s_j \mathbf{u}_j \mathbf{v}_j^*, \tag{3.1}$$

known as the *dyadic form of the SVD*.

The Dyadic Form of the Singular Value Decomposition

Theorem 3.3. *For any matrix* $\mathbf{A} \in \mathbb{C}^{m\times n}$, *there exists some* $r \in \{1,\dots,n\}$ *such that*

$$\mathbf{A} = \sum_{j=1}^{r} s_j \mathbf{u}_j \mathbf{v}_j^*,$$

where $s_1 \geq s_2 \geq \cdots \geq s_r > 0$ *and* $\mathbf{u}_1,\dots,\mathbf{u}_r \in \mathbb{C}^m$ *are orthonormal, and* $\mathbf{v}_1,\dots,\mathbf{v}_r \in \mathbb{C}^n$ *are orthonormal.*

Corollary 3.4. *The rank of a matrix equals its number of nonzero singular values.*

The singular value decomposition also gives an immediate formula for the 2-norm of a matrix.

Theorem 3.5. *Let* $\mathbf{A} \in \mathbb{C}^{m\times n}$ *have singular value decomposition* (3.1). *Then*

$$\|\mathbf{A}\| = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} = s_1.$$

Proof. The proof follows from the construction of the SVD at the start of this chapter. Note that

$$\|\mathbf{A}\|^2 = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|^2}{\|\mathbf{x}\|^2} = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x}}{\mathbf{x}^*\mathbf{x}},$$

and so $\|\mathbf{A}\|^2$ is the maximum Rayleigh quotient of the Hermitian matrix $\mathbf{A}^*\mathbf{A}$. By Theorem 2.2, this maximum value is the largest eigenvalue of

$\mathbf{A}^*\mathbf{A}$, i.e., $\lambda_1 = s_1^2$ in "Step 2" of the construction on page 92. Thus $\|\mathbf{A}\| = \sqrt{\lambda_1} = s_1$. ■

3.1.1 Inductive Proof of the SVD

3.1.2 The SVD and the Four Fundamental Subspaces

3.2 The SVD: Undoer of Many Knots

In the introduction to this chapter, we claimed that eigenvalue-based decompositions were the right tool for handling systems that involved *dynamics*. The SVD, in turn, is the perfect tool for handling *static* systems, i.e., systems that do not change with time. We identify three canonical static problems:

1. Find the unique solution $\mathbf{x}$ to $\mathbf{A}\mathbf{x} = \mathbf{b}$, where $\mathbf{A}$ is an invertible square matrix.

2. Find the solution $\mathbf{x}$ of minimum norm that solves the *underdetermined* system $\mathbf{A}\mathbf{x} = \mathbf{b}$, where $\mathbf{A}$ is a matrix with a nontrivial null space and $\mathbf{b} \in \mathcal{R}(\mathbf{A})$.

3. Find the minimum-norm vector $\mathbf{x}$ that minimizes $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|$, for any given $\mathbf{b}$.

Notice that this last problem subsumes the first two. (If $\mathbf{A}\mathbf{x} = \mathbf{b}$ has a solution $\mathbf{x}$, then $\|\mathbf{A}\mathbf{x} - \mathbf{b}\| = 0$ is minimal; if there are multiple solutions, one then finds the one having smallest norm.) Thus, our discussion will focus on problem 3.

Let $\mathbf{A} \in \mathbb{C}^{m\times n}$ have rank r and let $\mathbf{u}_1, \ldots, \mathbf{u}_m \in \mathbb{C}^m$ be a full set of *left* singular vectors, giving an orthonormal basis for $\mathbb{C}^m$ with

$$\begin{aligned} \mathcal{R}(\mathbf{A}) &= \operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_r\} \\ \mathcal{N}(\mathbf{A}^*) &= \operatorname{span}\{\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m\}. \end{aligned}$$

Expand $\mathbf{b} \in \mathbb{C}^m$ as a linear combination of the left singular vectors:

$$\mathbf{b} = \beta_1\mathbf{u}_1 + \cdots + \beta_m\mathbf{u}_m.$$

We can find the coefficients β_j very easily, since the orthonormality of the singular vectors gives

$$\begin{aligned} \mathbf{u}_j^*\mathbf{b} &= \beta_1\mathbf{u}_j^*\mathbf{u}_1 + \cdots + \beta_m\mathbf{u}_j^*\mathbf{u}_m \\ &= \beta_j. \end{aligned}$$

Now for any $\mathbf{x} \in \mathbb{C}^n$, notice that $\mathbf{A}\mathbf{x} \in \mathcal{R}(\mathbf{A})$, and so

$$\begin{aligned}\mathbf{A}\mathbf{x} - \mathbf{b} &= \Big(\mathbf{A}\mathbf{x} - (\beta_1\mathbf{u}_1 + \cdots + \beta_r\mathbf{u}_r)\Big) - \Big(\beta_{r+1}\mathbf{u}_{r+1} + \cdots \beta_m\mathbf{u}_m\Big) \\ &= (\mathbf{A}\mathbf{x} - \mathbf{b}_R) - \mathbf{b}_N,\end{aligned}$$

where

$$\begin{aligned}\mathbf{b}_R &:= (\beta_1\mathbf{u}_1 + \cdots + \beta_r\mathbf{u}_r) \in \mathcal{R}(\mathbf{A}) \\ \mathbf{b}_N &:= (\beta_{r+1}\mathbf{u}_{r+1} + \cdots \beta_m\mathbf{u}_m) \in \mathcal{N}(\mathbf{A}^*).\end{aligned}$$

Since the $\mathcal{R}(\mathbf{A}) \perp \mathcal{N}(\mathbf{A}^*)$ (by the Fundamental Theorem of Linear Algebra), $(\mathbf{A}\mathbf{x} - \mathbf{b}_R) \perp \mathbf{b}_N$, and hence, by the Pythagorean Theorem,

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 = \|(\mathbf{A}\mathbf{x} - \mathbf{b}_R) - \mathbf{b}_N\|^2 = \|\mathbf{A}\mathbf{x} - \mathbf{b}_R\|^2 + \|\mathbf{b}_N\|^2. \tag{3.2}$$

Inspect this expression. The choice of $\mathbf{x}$ does not affect $\|\mathbf{b}_N\|^2$: $\mathbf{b}_N$ is the piece of $\mathbf{b}$ that is beyond the reach of $\mathbf{A}\mathbf{x}$. (If $\mathbf{A}\mathbf{x} = \mathbf{b}$ has a solution, then $\mathbf{b}_N = \mathbf{0}$.) To minimize (3.2), the best we can do is find $\mathbf{x}$ such that $\|\mathbf{A}\mathbf{x} - \mathbf{b}_R\| = 0$. The dyadic form of the SVD,

$$\mathbf{A} = \sum_{j=1}^{r} s_j \mathbf{u}_j \mathbf{v}_j^*, \tag{3.3}$$

makes quick work of this problem. Expand any $\mathbf{x} \in \mathbb{C}^n$ in the *right* singular vectors,

$$\mathbf{x} = \xi_1\mathbf{v}_1 + \cdots + \xi_n\mathbf{v}_n,$$

so that (via orthonormality of the singular vectors),

$$\mathbf{A}\mathbf{x} = \sum_{j=1}^{r} s_j \mathbf{u}_j \mathbf{v}_j^*\mathbf{x} = \sum_{j=1}^{r} s_j \xi_j \mathbf{u}_j.$$

We can equate this expression with

$$\begin{aligned}\mathbf{b}_R &= \beta_1\mathbf{u}_1 + \cdots + \beta_r\mathbf{u}_r \\ &= (\mathbf{u}_1^*\mathbf{b})\mathbf{u}_1 + \cdots + (\mathbf{u}_r^*\mathbf{b})\mathbf{u}_r\end{aligned} \tag{3.4}$$

by simply matching the coefficients in the $\mathbf{u}_1, \ldots, \mathbf{u}_r$ directions:

$$s_k\xi_k = \mathbf{u}_k^*\mathbf{b}, \qquad k = 1, \ldots, r$$

giving

$$\xi_k = \frac{\mathbf{u}_k^* \mathbf{b}}{s_k}, \qquad k = 1, \ldots, r.$$

What about $\xi_{r+1}, \ldots, \xi_n$? *They can take any value!* To confirm this fact, define

$$\mathbf{x} = \sum_{k=1}^{r} \frac{\mathbf{u}_k^* \mathbf{b}}{s_k} \mathbf{v}_k + \sum_{k=r+1}^{n} \xi_k \mathbf{v}_k \tag{3.5}$$

for any $\xi_{r+1}, \ldots, \xi_n$, and verify that $\mathbf{A}\mathbf{x} = \mathbf{b}_R$:

$$\begin{aligned}
\mathbf{A}\mathbf{x} &= \left(\sum_{j=1}^{r} s_j \mathbf{u}_j \mathbf{v}_j^* \right) \left(\sum_{k=1}^{r} \frac{\mathbf{u}_k^* \mathbf{b}}{s_k} \mathbf{v}_k + \sum_{k=r+1}^{n} \xi_k \mathbf{v}_k \right) \\
&= \sum_{j=1}^{r} \sum_{k=1}^{r} s_j \frac{\mathbf{u}_k^* \mathbf{b}}{s_k} \mathbf{u}_j \mathbf{v}_j^* \mathbf{v}_k + \sum_{j=1}^{r} \sum_{k=r+1}^{n} s_j \xi_k \mathbf{u}_j \mathbf{v}_j^* \mathbf{v}_k \\
&= \sum_{j=1}^{r} (\mathbf{u}_j^* \mathbf{b}) \mathbf{u}_j + 0 \\
&= \mathbf{b}_R,
\end{aligned}$$

according to the expansion (3.4) for $\mathbf{b}_R$. Thus when $r < n$, equation (3.5) thus expresses the *infinitely many solutions* that minimize $\|\mathbf{A}\mathbf{b} - \mathbf{x}\|$. From all these solutions, we might naturally select the one that minimizes $\|\mathbf{x}\|$, the one that contains nothing extra. You might well suspect that this is the $\mathbf{x}$ we get from setting

$$\xi_{r+1} = \cdots = \xi_n = 0.$$

To confirm this fact, apply the Pythagorean theorem to (3.5) to get

$$\|\mathbf{x}\|^2 = \sum_{k=1}^{r} \frac{|\mathbf{u}_k^* \mathbf{b}|^2}{s_k^2} + \sum_{k=r+1}^{n} |\xi_k|^2,$$

making it obvious that the *unique* norm-minimizing solution is

$$\mathbf{x} = \sum_{j=1}^{r} \frac{\mathbf{u}_j^* \mathbf{b}}{s_j} \mathbf{v}_j = \left(\sum_{j=1}^{r} \frac{1}{s_j} \mathbf{v}_j \mathbf{u}_j^* \right) \mathbf{b}. \tag{3.6}$$

Take a moment to savor this beautiful formula, arguably one of the most important in matrix theory! In fact, the formula is so useful that the matrix on the right-hand side deserves some special nomenclature.

Pseudoinverse

Definition 3.6. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$ be a matrix of rank r with singular value decomposition (3.3). Then the pseudoinverse (or Moore–Penrose pseudoinverse) of $\mathbf{A}$ is given by*

$$\mathbf{A}^+ = \sum_{j=1}^{r} \frac{1}{s_j} \mathbf{v}_j \mathbf{u}_j^*. \tag{3.7}$$

We emphasize that $\mathbf{x} = \mathbf{A}^+\mathbf{b}$ in (3.6) solves all three of the problems formulated at the beginning of this section.

1. If $\mathbf{A}$ is invertible, then $m = n = r$ and
$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = \left(\sum_{j=1}^{n} \frac{1}{s_j} \mathbf{v}_j \mathbf{u}_j^* \right) \mathbf{b}.$$
In particular, whenever $\mathbf{A}$ is invertible, $\mathbf{A}^+ = \mathbf{A}^{-1}$.

2. If $\mathbf{b} \in \mathcal{R}(\mathbf{A})$, then $\mathbf{b}_R = \mathbf{b}$, and
$$\mathbf{x} = \mathbf{A}^+\mathbf{b} = \left(\sum_{j=1}^{r} \frac{1}{s_j} \mathbf{v}_j \mathbf{u}_j^* \right) \mathbf{b}$$
solves $\mathbf{A}\mathbf{x} = \mathbf{b}$. If $r < n$, then there exist infinitely many solutions to $\mathbf{A}\mathbf{x} = \mathbf{b}$ that all have the form $\mathbf{x} = \mathbf{A}^+\mathbf{b} + \mathbf{n}$ for $\mathbf{n} \in \mathcal{N}(\mathbf{A}) = \operatorname{span}\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$. Among all these solutions, $\mathbf{x} = \mathbf{A}^+\mathbf{b}$ is the unique solution of smallest norm.

3. When $\mathbf{b} \notin \mathcal{R}(\mathbf{A})$, no $\mathbf{x}$ will solve $\mathbf{A}\mathbf{x} = \mathbf{b}$, but $\mathbf{x} = \mathbf{A}^+\mathbf{b}$ will minimize $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|$. If $r < n$, there will be infinitely many $\mathbf{x}$ that minimize $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|$, each having the form $\mathbf{x} = \mathbf{A}^+\mathbf{b} + \mathbf{n}$ for $\mathbf{n} \in \mathcal{N}(\mathbf{A})$. Among all these solutions $\mathbf{x} = \mathbf{A}^+\mathbf{b}$ is the unique minimizer of $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|$ having smallest norm.

3.3 Optimal Low-Rank Approximations

Many applications call for use to approximate $\mathbf{A}$ with some low-rank matrix. Suppose we have the SVD

$$\mathbf{A} = \sum_{j=1}^{r} s_j \mathbf{u}_j \mathbf{v}_j^*, \tag{3.8}$$

and we seek some rank-k approximation to $\mathbf{A}$, for some $1 \leq k < r$. Since the singular values decay monotonically, $s_1 \geq s_2 \geq \cdots \geq s_r > 0$, one might naturally grab the leading k terms from the decomposition (3.8):

$$\mathbf{A}_k = \sum_{j=1}^{k} s_j \mathbf{u}_j \mathbf{v}_j^*.$$

How good an approximation is $\mathbf{A}_k$ to $\mathbf{A}$? Notice that

$$\mathbf{A} - \mathbf{A}_k = \sum_{j=k+1}^{r} s_j \mathbf{u}_j \mathbf{v}_j^*$$

is a singular value decomposition for the error $\mathbf{A} - \mathbf{A}_k$, and hence by Theorem **??** its norm equals the largest singular value in the decomposition:

$$\|\mathbf{A} - \mathbf{A}_k\| = s_{k+1}. \tag{3.9}$$

Can we construct a better rank-k approximation? The following theorem, one of the most important in matrix theory, says that this is not the case. (This is known as the Schmidt–Eckart–Young–Mirsky Theorem, after its various discovers.)

Optimal Low-Rank Approximation

Theorem 3.7. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$ be a rank-r matrix with singular value decomposition*

$$\mathbf{A} = \sum_{j=1}^{r} s_j \mathbf{u}_j \mathbf{v}_j^*,$$

and let $k < r$. Then

$$\min_{\text{rank}(\mathbf{X})} \|\mathbf{A} - \mathbf{X}\| = s_{k+1}$$

and this minimization is attained by

$$\mathbf{A}_k = \sum_{j=1}^{k} s_j \mathbf{u}_j \mathbf{v}_j^*. \tag{3.10}$$

Proof. We know from (3.9) that $\|\mathbf{A} - \mathbf{A}_k\| = s_{k+1}$, and so this proof must show that for any rank-k matrix $\mathbf{X}$, we have $\|\mathbf{A} - \mathbf{X}\| \geq s_{k+1}$.

Let $\mathbf{X} \in \mathbb{C}^{m\times n}$ be an arbitrary rank-k matrix. To show $\|\mathbf{A}-\mathbf{X}\| \geq s_{k+1}$, it suffices to identify some unit vector $\mathbf{z} \in \mathbb{C}^n$ for which $\|(\mathbf{A}-\mathbf{X})\mathbf{z}\| \geq s_{k+1}$, since

$$\|\mathbf{A}-\mathbf{X}\| = \max_{\|\mathbf{v}\|=1} \|(\mathbf{A}-\mathbf{X})\mathbf{v}\| \geq \|(\mathbf{A}-\mathbf{X})\mathbf{z}\|.$$

Since $\mathbf{X} \in \mathbb{C}^{m\times n}$ has rank-k, we must have

$$\dim(\mathcal{N}(\mathbf{X})) = n-k.$$

By the orthogonality of the right singular vectors of $\mathbf{A}$, we have

$$\dim(\text{span}\{\mathbf{v}_1,\ldots,\mathbf{v}_{k+1}\} = k+1.$$

The sum of the dimensions of these two spaces exceeds n, since $(n-k)+(k+1) = n+1$, and hence the intersection of these spaces must have dimension 1 or greater.

$$\dim\Big(\mathcal{N}(\mathbf{X}) \cap \text{span}\{\mathbf{v}_1,\ldots,\mathbf{v}_{k+1}\}\Big) \geq 1.$$

Thus we can find some unit vector in this intersection,

$$\mathbf{z} \in \mathcal{N}(\mathbf{X}) \cap \text{span}\{\mathbf{v}_1,\ldots,\mathbf{v}_{k+1}\}$$

with $\|\mathbf{z}\| = 1$. We will use the fact that $\mathbf{z}$ is in each of these spaces in turn. First, $\mathbf{z} \in \mathcal{N}(\mathbf{X})$ we must have $\mathbf{X}\mathbf{z} = \mathbf{0}$, so

$$\|\mathbf{A}-\mathbf{X}\| \geq \|(\mathbf{A}-\mathbf{X})\mathbf{z}\| = \|\mathbf{A}\mathbf{z}\|.$$

Now since $\mathbf{z} \in \text{span}\{\mathbf{v}_1,\ldots,\mathbf{v}_{k+1}\}$ we can expand

$$\mathbf{z} = c_1\mathbf{v}_1 + \cdots + c_{k+1}\mathbf{v}_{k+1}.$$

By the Pythagorean Theorem (and the orthonormality of the $\{\mathbf{v}_j\}$)

$$1 = \|\mathbf{z}\|^2 = |c_1|^2 + \cdots + |c_{k+1}|^2. \tag{3.11}$$

The orthonormality of the $\{\mathbf{v}_j\}$ also implies that

$$\mathbf{A}\mathbf{z} = \Big(\sum_{j=1}^{r} s_j\mathbf{u}_j\mathbf{v}_j^*\Big)\Big(\sum_{\ell=1}^{k+1} c_\ell\mathbf{v}_\ell\Big) = \sum_{j=1}^{r}\sum_{\ell=1}^{k+1} s_j c_\ell \mathbf{u}_j\mathbf{v}_j^*\mathbf{v}_\ell = \sum_{j=1}^{k+1} s_j c_j \mathbf{u}_j.$$

Again by the Pythagorean Theorem (now with orthonormality of the $\{\mathbf{u}_j\}$) we have

$$\|\mathbf{A}\mathbf{z}\|^2 \;=\; \sum_{j=1}^{k+1} s_j^2|c_j|^2 \;\geq\; s_{k+1}^2 \sum_{j=1}^{k+1}|c_j|^2 \;=\; s_{k+1}^2,$$

where the inequality follows from the decaying magnitudes of the singular values, $s_1 \geq \cdots \geq s_k \geq s_{k+1}$, and the last step uses the formula (3.11). We conclude that

$$\|\mathbf{A} - \mathbf{X}\| \;=\; \max_{\|\mathbf{v}\|=1} \|(\mathbf{A} - \mathbf{X})\mathbf{v}\| \;\geq\; \|(\mathbf{A} - \mathbf{X})\mathbf{z}\| \;=\; \|\mathbf{A}\mathbf{z}\| \;\geq\; s_{k+1},$$

thus establishing the theorem. ■

One must wonder if the partial sum

$$\mathbf{A}_k = \sum_{j=1}^{k} s_j \mathbf{u}_j \mathbf{v}_j^*$$

delivers a *unique* best rank-k approximation. That is not generally the case: if $k < r$ there are *infinitely many*. To see this, define

$$\widehat{\mathbf{A}}_k = \sum_{j=1}^{k} \widehat{s}_j \mathbf{u}_j \mathbf{v}_j^*,$$

for some $\widehat{s}_1, \ldots, \widehat{s}_k$, so that

$$\mathbf{A} - \widehat{\mathbf{A}}_k = \sum_{j=1}^{k} (s_j - \widehat{s}_j)\mathbf{u}_j \mathbf{v}_j^* + \sum_{j=k+1}^{r} s_j \mathbf{u}_j \mathbf{v}_j^*.$$

This is a like a singular value decomposition for $\mathbf{A} - \widehat{\mathbf{A}}$, except the "singular values" $s_j - \widehat{s}_j$ could potentially be negative. One can show that the norm of the misfit $\mathbf{A} - \widehat{\mathbf{A}}_k$ is the largest magnitude of these quantities and the untouched s_j for $j = k+1, \ldots, r$:

$$\|\mathbf{A} - \widehat{\mathbf{A}}_k\| = \max\Big\{|s_1 - \widehat{s}_1|, \ldots, |s_k - \widehat{s}_k|, s_{k+1}\Big\}.$$

So long as we pick $\widehat{s}_1, \ldots, \widehat{s}_k$ so that

$$\max_{1 \leq j \leq k} |s_j - \widehat{s}_j| \leq s_{k+1},$$

then $\|\mathbf{A} - \widehat{\mathbf{A}}_k\| = s_{k+1}$, and $\widehat{\mathbf{A}}_k$ is another best rank-k approximation to $\mathbf{A}$. (However, $\mathbf{A}_k$ is the only one of these that leaves a rank $r-k$ misfit $\mathbf{A} - \mathbf{A}_k$.)

3.4 The Polar Decomposition

Throughout our studies we have fluently translated basic properties of scalars into their matrix analogues: $a \pm b$ effortlessly becomes $\mathbf{A} \pm \mathbf{B}$, the multiplication ab becomes $\mathbf{AB}$, but with the caveat that the product does not in general commute; the inversion $1/b$ becomes $\mathbf{B}^{-1}$, and division by $b = 0$ corresponds to singular $\mathbf{B}$. Even magnitude generalizes: $|a|$ becomes $\|\mathbf{A}\|$.

The singular value decomposition allows one more construction in the same vein. Any $a \in \mathbb{C}$ can be written in the *polar form* $a = re^{i\theta}$ for $r \geq 0$ and $\theta \in [0, 2\pi)$: that is, a is the product of a nonnegative number r and a number with magnitude 1, since $|e^{i\theta}| = 1$.

Suppose that $\mathbf{A} \in \mathbb{C}^{m\times n}$ with $m \geq n$. Into the skinny SVD

$$\mathbf{A} = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\mathbf{V}^*$$

splice the identity matrix in the form $\mathbf{V}^*\mathbf{V} = \mathbf{I}$,

$$\mathbf{A} = \widehat{\mathbf{U}}(\mathbf{V}^*\mathbf{V})\widehat{\mathbf{\Sigma}}\mathbf{V}^* = (\widehat{\mathbf{U}}\mathbf{V}^*)(\mathbf{V}\widehat{\mathbf{\Sigma}}\mathbf{V}^*) =: \mathbf{ZR},$$

where $\mathbf{Z} := \widehat{\mathbf{U}}\mathbf{V}^* \in \mathbb{C}^{m\times n}$ is subunitary and $\mathbf{R} := \mathbf{V}\widehat{\mathbf{\Sigma}}\mathbf{V}^* \in \mathbb{C}^{n\times n}$ is Hermitian positive semidefinite. (Note that $\mathbf{R}$ is constructed via a unitary diagonalization whose central matrix $\widehat{\mathbf{\Sigma}}$ has nonnegative diagonal entries: hence $\mathbf{R}$ must be Hermitian positive semidefinite.) In the polar form $a = re^{i\theta}$, the unit-length scalar $e^{i\theta}$ generalizes to the subunitary $\mathbf{Z}$, while the nonnegative scalar $r > 0$ generalizes to the Hermitian semidefinite $\mathbf{R}$.

Polar Decomposition

Theorem 3.8. *Any matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$, $m \geq n$, can be written as*

$$\mathbf{A} = \mathbf{ZR}$$

for subunitary $\mathbf{Z} \in \mathbb{C}^{m\times n}$ and Hermitian positive semidefinite $\mathbf{R} \in \mathbb{C}^{n\times n}$.

Notice that $m \geq n$ was necessary for $\mathbf{Z}$ to be subunitary, since

$$\mathbf{Z}^*\mathbf{Z} = \mathbf{V}\widehat{\mathbf{U}}^*\widehat{\mathbf{U}}\mathbf{V}^* = \mathbf{VV}^*$$

can only equal the identity if $\mathbf{V}$ is a square unitary matrix. (Also note that $\mathbf{Z} \in \mathbb{C}^{m\times n}$, and one cannot have a matrix with n orthonormal columns of length m if $n > m$.) If $m \leq n$, we can alternatively write

$$\mathbf{A} = \mathbf{U}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^* = \mathbf{U}\widehat{\mathbf{\Sigma}}(\mathbf{U}^*\mathbf{U})\widehat{\mathbf{V}}^* = (\mathbf{U}\widehat{\mathbf{\Sigma}}\mathbf{U}^*)(\mathbf{U}\widehat{\mathbf{V}}^*) =: \mathbf{RZ}^*,$$

now with $\mathbf{R} := \mathbf{U}\widehat{\mathbf{\Sigma}}\mathbf{U}^* \in \mathbb{C}^{m\times m}$ Hermitian positive semidefinite and $\mathbf{Z} := \widehat{\mathbf{V}}\mathbf{U}^* \in \mathbb{C}^{n\times m}$ subunitary.

3.5 Variational Characterization of Singular Values

Since the singular values are square roots of the eigenvalues of the Hermitian matrices $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}\mathbf{A}^*$, the singular values inherit the variational characterizations that were explored in Section 2.2. For example,

$$\sigma_1 = \max_{\mathbf{v}\in\mathbb{C}^n} \left(\frac{\mathbf{v}^*\mathbf{A}^*\mathbf{A}\mathbf{v}}{\mathbf{v}^*\mathbf{v}}\right)^{1/2} = \max_{\mathbf{u}\in\mathbb{C}^m} \left(\frac{\mathbf{u}^*\mathbf{A}\mathbf{A}^*\mathbf{u}}{\mathbf{u}^*\mathbf{u}}\right)^{1/2},$$

with the leading right and left singular vectors $\mathbf{v}_1$ and $\mathbf{u}_1$ being unit vectors that attain these maxima.

However, the singular values also satisfy a subtler variational property that incorporates both left and right singular vectors at the same time. Consider, for unit vectors $\mathbf{u} \in \mathbb{C}^m$ and $\mathbf{v} \in \mathbb{C}^n$, the quantity

$$|\mathbf{u}^*\mathbf{A}\mathbf{v}| \leq \|\mathbf{u}\|\|\mathbf{A}\|\|\mathbf{v}\| = \sigma_1,$$

using the Cauchy–Schwarz inequality and the definition of the induced matrix 2-norm. On the other hand, if $\mathbf{u}_1$ and $\mathbf{v}_1$ unit vectors that give

$$|\mathbf{u}_1^*\mathbf{A}\mathbf{v}_1| = |\mathbf{u}_1^*(\sigma_1\mathbf{u}_1)| = \sigma_1.$$

Thus

$$\sigma_1 = \max_{\mathbf{u}\in\mathbb{C}^m,\mathbf{v}\in\mathbb{C}^n} \frac{|\mathbf{u}^*\mathbf{A}\mathbf{v}|}{\|\mathbf{u}\|\,\|\mathbf{v}\|}.$$

We can characterize subsequent singular values the same way. Recall the dyadic version of the singular value decomposition,

$$\mathbf{A} = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^*$$

for $\mathbf{A}$ of rank r. If we restrict unit vectors $\mathbf{u}$ and $\mathbf{v}$ such to be orthogonal to $\mathbf{u}_1$ and $\mathbf{v}_1$, then

$$\mathbf{u}^*\mathbf{A}\mathbf{v} = \sum_{j=1}^{r} \sigma_j \mathbf{u}^*\mathbf{u}_j\mathbf{v}_j^*\mathbf{v} = \sum_{j=2}^{r} \sigma_j \mathbf{u}^*\mathbf{u}_j\mathbf{v}_j^*\mathbf{v} = \mathbf{u}^*\left(\sum_{j=2}^{r} \sigma_j \mathbf{u}_j\mathbf{v}_j^*\right)\mathbf{v}.$$

Hence

$$|\mathbf{u}^*\mathbf{A}\mathbf{v}| \leq \sigma_2,$$

with the inequality attained when $\mathbf{u} = \mathbf{u}_1$ and $\mathbf{v} = \mathbf{v}_1$. Continuing this process gives the following analogue of Theorem 2.2.

Theorem 3.9. *For any* $\mathbf{A} \in \mathbb{C}^{m\times n}$,

$$\sigma_k = \max_{\substack{\mathbf{u}\perp\text{span}\{\mathbf{u}_1,\ldots,\mathbf{u}_{k-1}\}\\ \mathbf{v}\perp\text{span}\{\mathbf{v}_1,\ldots,\mathbf{v}_{k-1}\}}} \frac{|\mathbf{u}^*\mathbf{A}\mathbf{v}|}{\|\mathbf{u}\|\|\mathbf{v}\|}.$$

3.6 Principal Component Analysis

Matrix theory enables the analysis of the volumes of data that now so commonly arise from applications ranging from basic science to public policy. Such measured data often depends on many factors, and we seek to identify those that are most critical. Within this realm of multivariate statistics, *principal component analysis* (PCA) is a fundamental tool.

Linear algebraists often say, "PCA is the SVD" – in this section, we will explain what this means, and some of the subtleties involved.

3.6.1 Variance and covariance

To understand principal component analysis, we need some basic notions from statistics, described in any basic textbook. For a general description of PCA along with numerous applications, see the text by Jolliffe [Jol02], whose presentation shaped parts of our discussion here.

The *expected value*, or *mean*, of a random variable X is denoted $\mathsf{E}[X]$. The expected value is a linear function, so for any constants $\alpha, \beta \in \mathbb{R}$, $\mathsf{E}[\alpha X + \beta] = \alpha\mathsf{E}[X] + \beta$.

The *variance* of X describes how much X is expected to deviate from its mean,

$$\mathsf{Var}(X) = \mathsf{E}[(X - \mathsf{E}[X])^2],$$

which, using linearity of the expected value, takes the equivalent form

$$\mathsf{Var}(X) = \mathsf{E}[X^2] - \mathsf{E}[X]^2.$$

The *covariance* between two (potentially correlated) random variables X and Y is

$$\begin{aligned}\mathsf{Cov}(X,Y) &= \mathsf{E}[(X - \mathsf{E}[X])(Y - \mathsf{E}[Y])]\\ &= \mathsf{E}[XY] - \mathsf{E}[X]\mathsf{E}[Y].\end{aligned}$$

with $\mathsf{Cov}(X,X) = \mathsf{Var}(X)$. These definitions of variance and covariance are the bedrock concepts underneath PCA, for with them we can understand the variance present in a linear combination of several random variables.

Suppose we have a set of real-valued random variables $X_1, \ldots, X_n$ in which we suspect there may be some redundancy. Perhaps some of these variables can be expressed as linear combinations of the others – either exactly, or nearly so. At the other extreme, there may be some way to combine $X_1, \ldots, X_n$ that captures much of the variance in one (or a few) aggregate random variables. In particular, we shall seek scalars $\gamma_1, \ldots, \gamma_n$ such that

$$\sum_{j=1}^{n} \gamma_j X_j$$

has the largest possible variance. The definitions of variance and covariance, along with the linearity of the expected value, lead to a formula for the variance of a linear combination of random variables:

$$\mathsf{Var}\Big(\sum_{j=1}^{n} \gamma_j X_j\Big) = \sum_{j=1}^{n}\sum_{k=1}^{n} \gamma_j \gamma_k \,\mathsf{Cov}(X_j, X_k). \tag{3.12}$$

You have seen double sums like this before. If we define the *covariance matrix* $\mathbf{C} \in \mathbb{C}^{n\times n}$ having (j,k) entry

$$c_{j,k} = \mathsf{Cov}(X_j, X_k),$$

and let $\mathbf{v} = [\gamma_1, \ldots, \gamma_n]^{\mathrm{T}}$, then the variance of the combined variable is just a Rayleigh quotient:

$$\mathsf{Var}\Big(\sum_{j=1}^{n} \gamma_j X_j\Big) = \mathbf{v}^* \mathbf{C} \mathbf{v}.$$

Since the covariance function is symmetric: $\mathsf{Cov}(X,Y) = \mathsf{Cov}(Y,X)$, the matrix $\mathbf{C}$ is Hermitian; it is also positive semidefinite. Why? Variance, by its definition as the expected value of the square of a real random variable, is always nonnegative. Thus the formula (3.12), which derives from the linearity of the expected value, ensures that $\mathbf{v}^*\mathbf{C}\mathbf{v} \geq 0$. (Under what circumstances can this quantity be zero?)

We can write $\mathbf{C}$ in another convenient way. Collect the random variables into the vector

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}.$$

Then the (j,k) entry of $\mathsf{E}[\mathbf{X}\mathbf{X}^*] - \mathsf{E}[\mathbf{X}]\mathsf{E}[\mathbf{X}^*]$ is

$$\mathsf{E}[X_j X_k] - \mathsf{E}[X_j]\mathsf{E}[X_k] = \mathsf{Cov}(X_j, X_k) = c_{j,k},$$

and so

$$\mathbf{C} = \mathsf{E}[\mathbf{X}\mathbf{X}^*] - \mathsf{E}[\mathbf{X}]\mathsf{E}[\mathbf{X}^*].$$

3.6.2 Derived variables that maximize variance

Return now to the problem of *maximizing* the variance of $\mathbf{v}^*\mathbf{C}\mathbf{v}$. Without constraint on $\mathbf{v}$, this quantity can be arbitrarily large (assuming $\mathbf{C}$ is nonzero); thus we shall require that $\sum_{j=1}^k \gamma_j^2 = \|\mathbf{v}\|^2 = 1$. With this normalization, you immediately see how to maximize the variance $\mathbf{v}^*\mathbf{C}\mathbf{v}$: $\mathbf{v}$ should be a unit eigenvector associated with the largest magnitude eigenvalue of $\mathbf{C}$; call this vector $\mathbf{v}_1$. The associated variance, of course, is the largest eigenvalue of $\mathbf{C}$; call it

$$\lambda_1 = \mathbf{v}_1^*\mathbf{C}\mathbf{v}_1 = \max_{\mathbf{v}\in\mathbb{C}^n} \frac{\mathbf{v}^*\mathbf{C}\mathbf{v}}{\mathbf{v}^*\mathbf{v}}.$$

The eigenvector $\mathbf{v}_1$ encodes the way to combine $X_1, \ldots, X_n$ to maximize variance. The new variable – *the leading principal component* – is

$$\mathbf{v}_1^*\mathbf{X} = \sum_{j=1}^n \gamma_j X_j.$$

You are already suspecting that a unit eigenvector associated with the second largest eigenvalue, $\mathbf{v}_2$ with $\lambda_2 = \mathbf{v}_2^*\mathbf{C}\mathbf{v}_2$, must encode the second-largest way to maximize variance.

Let us explore this intuition. To find the second-best way to combine the variables, we should insist that the next new variable, for now call it $\mathbf{w}^*\mathbf{X}$, should be *independent* of the first, i.e.,

$$\mathsf{Cov}(\mathbf{v}_1^*\mathbf{X}, \mathbf{w}^*\mathbf{X}) = 0.$$

However, using linearity of expectation and the fact that, e.g., $\mathbf{w}^*\mathbf{X} = \mathbf{X}^*\mathbf{w}$ for real vectors,

$$\begin{aligned}
\mathsf{Cov}(\mathbf{v}_1^*\mathbf{X}, \mathbf{w}^*\mathbf{X}) &= \mathsf{E}[(\mathbf{v}_1^*\mathbf{X})(\mathbf{w}^*\mathbf{X})] - \mathsf{E}[\mathbf{v}_1^*\mathbf{X}]\mathsf{E}[\mathbf{w}^*\mathbf{X}] \\
&= \mathsf{E}[(\mathbf{v}_1^*\mathbf{X}\mathbf{X}^*\mathbf{w}] - \mathsf{E}[\mathbf{v}_1^*\mathbf{X}]\mathsf{E}[\mathbf{X}^*\mathbf{w}] \\
&= \mathbf{v}_1^*\mathsf{E}[\mathbf{X}\mathbf{X}^*]\mathbf{w} - \mathbf{v}_1^*\mathsf{E}[\mathbf{X}]\mathsf{E}[\mathbf{X}^*]\mathbf{w} \\
&= \mathbf{v}_1^*(\mathsf{E}[\mathbf{X}\mathbf{X}^*] - \mathsf{E}[\mathbf{X}]\mathsf{E}[\mathbf{X}^*])\mathbf{w} \\
&= \mathbf{v}_1^*\mathbf{C}\mathbf{w} = \lambda_1\mathbf{v}_1^*\mathbf{w}.
\end{aligned}$$

Hence (assuming $\lambda_1 \neq 0$), for the combined variables $\mathbf{v}_1^*\mathbf{X}$ and $\mathbf{w}^*\mathbf{X}$ to be independent, the vectors $\mathbf{v}_1$ and $\mathbf{w}$ must be *orthogonal*, perfectly confirming your intuition: the second-best way to combine the variables is to pick $\mathbf{w}$

to be a unit eigenvector $\mathbf{v}_2$ of $\mathbf{C}$ corresponding to the second largest eigenvalue – a direct result of the variational characterization of eigenvalues in Theorem 2.2. The associated variance of $\mathbf{v}_2^*\mathbf{X}$ is

$$\lambda_2 = \max_{\mathbf{w} \perp \text{span}\{\mathbf{u}_1\}} \frac{\mathbf{w}^*\mathbf{C}\mathbf{w}}{\mathbf{w}^*\mathbf{w}}.$$

Of course, in general, the kth best way to combine the variables is given by the eigenvector $\mathbf{v}_k$ of $\mathbf{C}$ associated with the kth largest eigenvalue.

We learn much about our variables from the relative size of the variances (eigenvalues)

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0.$$

If some of the latter eigenvalues are very small, that indicates that the set of n random variables can be well approximated by a fewer number of aggregate variables. These aggregate variables are the *principal components* of $X_1, \ldots, X_n$.

3.6.3 Approximate PCA from empirical data

In practical situations, one often seeks to analyze empirical data drawn from some unknown distribution: the expected values and covariances are not available. Instead, we will *estimate* these from the measured data.

Suppose, as before, that we are considering n random variables, $X_1, \ldots, X_n$, with m samples of each:

$$x_{j,k}, \quad k = 1, \ldots, m,$$

i.e., $x_{j,k}$ is the kth sample of the random variable X_j. The expected value has a the familiar *unbiased estimate*

$$\mu_j = \frac{1}{m} \sum_{j=1}^{m} x_{j,k}.$$

Similarly, we can approximate the covariance

$$\mathsf{Cov}(X_j, X_k) = \mathsf{E}[(X_j - \mathsf{E}[X_j])(X_k - \mathsf{E}[X_k)].$$

One might naturally estimate this as

$$\frac{1}{m} \sum_{\ell=1}^{m} (x_{j,\ell} - \mu_j)(x_{k,\ell} - \mu_k).$$

However, replacing the true expected values $\mathsf{E}[X_j]$ and $\mathsf{E}[X_k]$ with the empirical estimates μ_j and μ_k introduces some slight bias into this estimate. This bias can be removed by scaling [Cra46, Sect. 27.6], replacing $1/m$ by $1/(m-1)$ to get the *unbiased estimate*

$$s_{j,k} = \frac{1}{m-1}\sum_{\ell=1}^{m}(x_{j,\ell}-\mu_j)(x_{k,\ell}-\mu_k), \quad j,k = 1,\dots,n.$$

If we let

$$\mathbf{x}_j = \begin{bmatrix} x_{j,1} \\ \vdots \\ x_{j,m} \end{bmatrix}, \quad j = 1,\dots,n,$$

then each covariance estimate is just an inner product

$$s_{j,k} = \frac{1}{m-1}(\mathbf{x}_j - \mu_j)^*(\mathbf{x}_k - \mu_k).$$

Thus, if we center the samples of each variable about its empirical mean, we can write the empirical covariance matrix $\mathbf{S} = [s_{j,k}]$ as a matrix product. Let

$$\mathfrak{X} := [\,(\mathbf{x}_1-\mu_1) \quad (\mathbf{x}_2-\mu_2) \quad \cdots \quad (\mathbf{x}_n-\mu_n)\,] \in \mathbb{R}^{m\times n},$$

so that

$$\mathbf{S} = \frac{1}{m-1}\mathfrak{X}^*\mathfrak{X}.$$

Now conduct principal component analysis just as before, but with the empirical covariance matrix $\mathbf{S}$ replacing the true covariance matrix $\mathbf{C}$. The eigenvectors of $\mathbf{S}$ now lead to *sample principal components*. Note that there is no need to explicitly form the matrix $\mathbf{S}$: instead, we can simply perform the singular value decomposition of the data matrix $\mathfrak{X}$. This is why some say, "PCA is just the SVD." We summarize the details step-by-step.

1. Collect m samples of each of n random variables, $x_{j,k}$ for $j = 1,\dots,m$ and $k = 1,\dots,n$. (We need $m > 1$, and, generally expect $m \gg n$.)

2. Compute the empirical means of each column, $\mu_k = (\sum_{j=1}^{m} x_{j,k})/m$.

3. Stacking the samples of the kth variable in the vector $\mathbf{x}_k \in \mathbb{R}^m$, construct the mean-centered data matrix

$$\mathfrak{X} = [\,(\mathbf{x}_1-\mu_1) \quad (\mathbf{x}_2-\mu_2) \quad \cdots \quad (\mathbf{x}_n-\mu_n)\,] \in \mathbb{R}^{m\times n}.$$

4. Compute the (skinny) singular value decomposition $\mathfrak{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$, with $\mathbf{U} \in \mathbb{R}^{m\times n}$, $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n\times n}$, and $\mathbf{V} = [\mathbf{v}_1 \ \cdots \ \mathbf{v}_n] \in \mathbb{R}^{n\times n}$.

5. The kth sample principal component is given by $\mathbf{v}_k^*\mathbf{X}$, where $\mathbf{X} = [X_1, \ldots, X_n]^*$ is the vector of random variables.

A word of caution: when conducting principal component analysis, the *scale* of each column matters. For example, if the random variables sampled in each column of $\mathfrak{X}$ are measurements of physical quantities, they can differ considerably in magnitude depending on the units of measurement. By changing units of measurement, you can significantly alter the principal components.

References

[Ber33] Daniel Bernoulli. Theoremata de oscillationibus corporum filo flexili connexorum et catenae verticaliter suspensae. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 6:108–122, 1733. Reprinted and translated in [CD81].

[Ber34] Daniel Bernoulli. Demonstrationes theorematum suorum de oscillationibus corporum filo flexili connexorum et catenae verticaliter suspensae. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 7:162–173, 1734.

[BS72] R. H. Bartels and G. W. Stewart. Solution of the matrix equation $AX+XB = C$. *Comm. ACM*, 15:820–826, 1972.

[BS09] Robert E. Bradley and C. Edward Sandifer. *Cauchy's Cours d'analyse: An Annotated Translation*. Springer, Dordrecht, 2009.

[Cau21] Augustin-Louis Cauchy. *Cours d'Analayse de l'École Royale Polytechnique*. Debure Frères, Paris, 1821.

[Cay58] Arthur Cayley. Memoir on the theory of matrices. *Phil. Trans. Royal Soc. London*, 148:17–37, 1858.

[CD81] John T. Cannon and Sigalia Dostrovsky. *The Evolution of Dynamics: Vibration Theory from 1687 to 1782*. Springer-Verlag, New York, 1981.

[Cra46] Harald Cramér. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, 1946.

[ES] Mark Embree and D. C. Sorensen. *An Introduction to Model Reduction for Linear and Nonlinear Differential Equations*. In preparation.

[FS83] R. Fletcher and D. C. Sorensen. An algorithmic derivation of the Jordan canonical form. *Amer. Math. Monthly*, 90:12–16, 1983.

[GW76] G. H. Golub and J. H. Wilkinson. Ill-conditioned eigensystems and the computation of the Jordan canonical form. *SIAM Review*, 18:578–619, 1976.

[HJ85] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

[HJ13] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, second edition, 2013.

[HK71] Kenneth M. Hoffman and Ray Kunze. *Linear Algebra*. Prentice Hall, Englewood Cliffs, N.J., second edition, 1971. check this.

[Jol02] I. T. Jolliffe. *Principal Component Analysis*. Springer, New York, second edition, 2002.

[Kat80] Tosio Kato. *Perturbation Theory for Linear Operators*. Springer-Verlag, Berlin, corrected second edition, 1980.

[Lak76] Imre Lakatos. *Proofs and Refutations: The Logical of Mathematical Discovery*. Cambridge University Press, Cambridge, 1976.

[Par98] Beresford N. Parlett. *The Symmetric Eigenvalue Problem*. SIAM, Philadelphia, SIAM Classics edition, 1998.

[Ray78] Lord Rayleigh (John William Strutt). *The Theory of Sound*. Macmillan, London, 1877, 1878. 2 volumes.

[RS80] Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics I: Functional Analysis*. Academic Press, San Diego, revised and enlarged edition, 1980.

[SM03] Endre Süli and David Mayers. *An Introduction to Numerical Analysis*. Cambridge University Press, Cambridge, 2003.

[Ste04] J. Michael Steele. *The Cauchy–Schwarz Master Class*. Cambridge University Press, Cambridge, 2004.

[Str93] Gilbert Strang. The fundamental theorem of linear algebra. *Amer. Math. Monthly*, 100:848–855, 1993.

[Tru60] C. Truesdell. *The Rational Mechanics of Flexible or Elastic Bodies, 1638–1788*. Leonhardi Euleri Opera Omnia, Introduction to Volumes X and XI, Second Series. Orell Füssli, Zürich, 1960.

[vNW29] J. von Neumann and E. Wigner. Über das Verhalten von Eigenwerten bei Adiabatischen Prozessen. *Physikalische Zeit.*, 30:467–470, 1929.

[You88] Nicholas Young. *An Introduction to Hilbert Space*. Cambridge University Press, Cambridge, 1988.