1 Low-rank approximations to a matrix using SVD

First point: we can write the SVD as a sum of rank-1 matrices, each given by left singular vector outer-product with right singular vector, weighted by singular value.

\[ A = USV^\top = s_1 u_1 v_1^\top + \cdots + s_n u_n v_n^\top \]  

Second point: to get the best (in sense of minimum squared error) low-rank approximation to a matrix \( A \), truncate after \( k \) singular vectors. Thus, for example, the best rank-1 approximation to \( A \) (also known as a separable approximation) is given by

\[ A \approx s_1 u_1 v_1^\top \]

2 Determinant

The determinant of a square matrix quantifies how that matrix changes the volume of a unit hypercube. The absolute value of the determinant of a square matrix \( A \) is equal to the product of its singular values

\[ |\det(A)| = \prod_{i=1}^{n} s_i, \]

where \( \{s_i\} \) are the singular values of \( A \).

It is easy to see this intuitively from thinking about the SVD of \( A \), which consists of a rotation (by \( V^\top \)) a stretching along the cardinal axes (by \( s_i \) for each direction), followed by a second rotation (by \( U \)). Clearly the stretching by \( S \) is the only part of \( A \) that increases or decreases volume. The determinant of an orthogonal matrix is +1 or -1 (where -1 arises from flipping the sign of an axis), since a purely rotational (length preserving) linear operation neither expands nor contracts the volume of the space.

The more general definition of the determinant is that it is equal to the product of the eigenvalues of a matrix:

\[ \det(A) = \prod_{i=1}^{n} e_i, \]

where \( \{e_i\} \) are the eigenvalues of \( A \).

For symmetric, positive semi-definite matrices, this is also equal to the product of singular values.
3 Principal Components Analysis (PCA)

Suppose someone hands you a stack of $N$ vectors, $\{\vec{x}_1, \ldots, \vec{x}_N\}$, each of dimension $d$. For example, we might imagine we have made a simultaneous recording from $d$ neurons, so each vector represents the spike counts of all recorded neurons in a single time bin, and we have $N$ time bins total in the experiment.

We suspect that these vectors not “fill” out the entire $d$-dimensional space, but instead be confined to a lower-dimensional subspace. (For example, if two neurons always emit the same number of spikes, then their responses live entirely along the 1D subspace corresponding to the $x_i = x_j$ line).

Can we make a mathematically rigorous theory of dimensionality reduction that captures how much of the “variability” in the data is captured by a low-dimensional projection? (Yes: it turns out the tool we are looking for is PCA!)

3.1 Finding the best 1D subspace

Let’s suppose we wish to find the best 1D subspace, i.e., the one-dimensional projection of the data that captures the largest amount of variability. We can formalize this as the problem of finding the unit vector $\vec{v}$ that maximizes the sum of squared linear projections of the data vectors:

$$\text{Sum of squared linear projections} = \sum_{i=1}^{N} (\vec{x}_i \cdot \vec{v})^2 = \|X\vec{v}\|^2$$

$$= (X\vec{v})^\top (X\vec{v})$$

$$= \vec{v}^\top X^\top X \vec{v}$$

$$= \vec{v}^\top (X^\top X) \vec{v}$$

$$= \vec{v}^\top C \vec{v},$$

where $C = X^\top X$, under the constraint that $\vec{v}$ is a unit vector, i.e., $\vec{v}^\top \vec{v} = 1$.

It turns out that the solution corresponds to the top eigenvector of $C$ (i.e., the eigenvector with the largest eigenvalue). Let’s look at the SVD of the $C$ matrix (which is also the eigenvalue decomposition of $C$), and look at what happens if we just restrict our choice of $\vec{v}$ to the eigenvectors of $C$.

First, remember that because $C$ is symmetric and positive-semi-definite (“psd”), its SVD is also its eigenvector decomposition:

$$C = U S U^\top,$$

where the columns of $U$ are the eigenvectors and diagonal entries of $S$ are eigenvalues.

Now let’s consider what happens if we set $\vec{v} = \vec{u}_j$, i.e., the $j$’th eigenvector of $C$. Because $U$ is an orthogonal matrix (i.e., its columns for an orthonormal basis), then $U^\top \vec{u}_j$ will be a vector of zeros.
with a 1 in the $j$’th component. We have

$$\bar{u}_j^T C \bar{u}_j = \bar{u}_j^T (USU^T) \bar{u}_j$$

$$= (\bar{u}_j U) S (U^T \bar{u}_j)$$

$$= \begin{bmatrix} 0 & \cdots & 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} s_1 \\ \vdots \\ s_j \\ \vdots \\ s_d \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

$$= s_j$$

So: plugging in the $j$’th eigenvector of $C$ gives us out $s_j$ as the sum of squared projections. Since we want to maximize this quantity (i.e., find the linear projection that maximizes it), we should clearly choose the eigenvector with largest eigenvalue, which (since SVD orders them from greatest to smallest) corresponds to the solution

$$\bar{v} = \bar{u}_1.$$ 

This vector (the “dominant” eigenvector of $C$) is the first principle component (sometimes called the “first PC” for short).