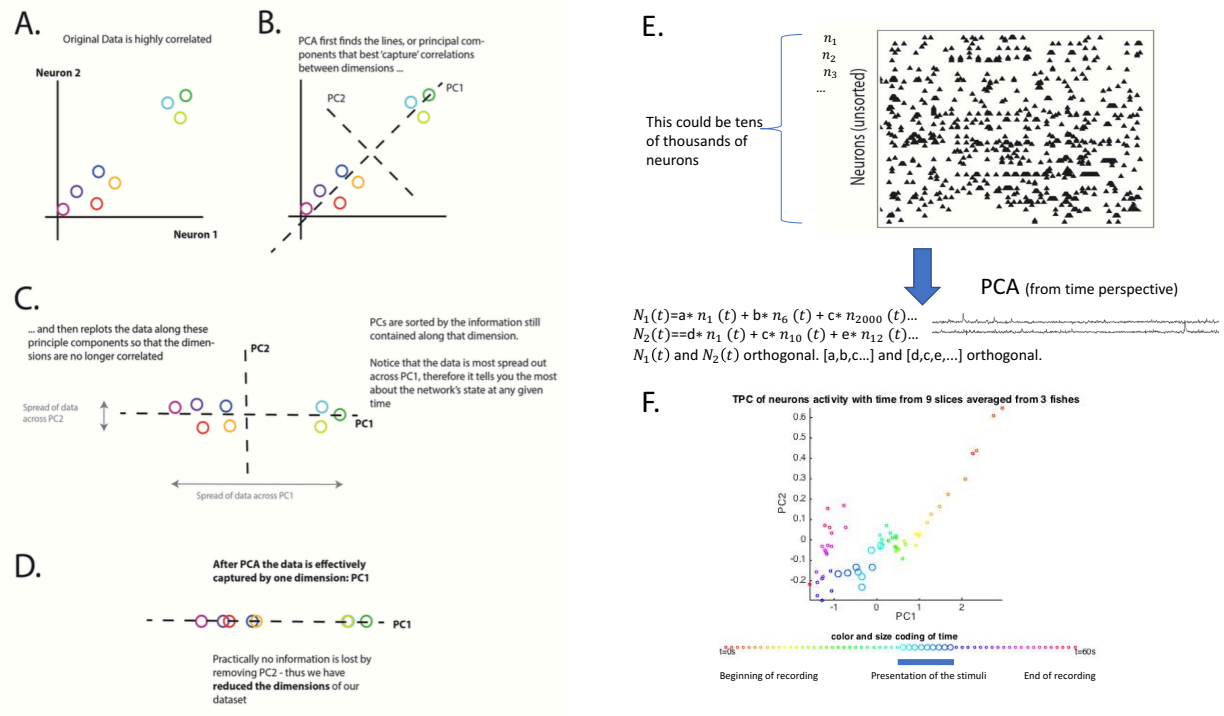
Principle Component Analysis (PCA) and Singular Value Decomposition (SVD)

In this tutorial [1], we will use neural data as an example to see how PCA can help make sense of big data—by big I mean high demensional data. We will also see how SVD connects the matrices involved in PCA.



We usually put neural data collected in a matrix of $N \times T$: N neurons' signal at T time points. N can be tens of thousands—in this case the whole data matrix looks messy and is hard to make sense of (see Fig. E for an example, which only has less than 100 neurons.) Thus, we need to a way to reduce the "dimension" of data to a low number of dimensions that can capture the greatest amount of information contained in your data. Specifically, in our example we want to combine neurons that are correlated into one ensemble, and only look at a few ensembles that are drastically different from each other. In PCA, dimensions are considered redundant when they are highly correlated to other dimensions. If two dimensions are correlated, that means you can use the value of a datapoint in one dimension to predict the value of that data in the other. Thus, having that second dimension doesnt actually provide any new information—we can get rid of it and not lose any knowledge. When N=2 in the data, it is easy to demonstrate visually what it means. Take a look at Fig. A-D.

The visualization should give you an idea that PCA is a powerful tool. In Fig F, I showed a picture of reducing about 700 neurons to only 2 ensembles using PCA. You can see a trajectory of neurons' states at the beginning of the experiment (0s) to the end of the experiment (60s), and when the stimuli (smell) is present during the 10s in the middle of the experiment.

But how to derive the direction for the greatest variance? We start with a $N \times T$ matrix of data, $\mathbf{M}$. Assume without loss of generality that $T \geq N$. Also assume we subtracted the means off each row (from the neuron perspective) or column (from the time perspective) of $\mathbf{M}$. Before we dive in, I need to introduce the covariance matrix as a preparation.

1. Covariance Matrices

[1] Materials adapted from Emily Mackevicius for Computational Neuroscience Woods Hole Summer Course, Strang 4E & 5E, and materials from Kristian Herrera for Systems Neuroscience (Harvard MCB105)

We can calculate covariance from the neuron point of view. Suppose $\vec{m}(t)$ is an observed pattern of neuronal firing at time t (a column of the matrix $\mathbf{M}$). Then the covariance between neuron i and neuron j is[2]:

$$\mathbf{C}_{ij}^{\text{neurons}} = \frac{1}{T}\sum_{t=1}^{T} M_{it}M_{jt}$$

In matrix notation:

$$\underbrace{\mathbf{C}^{\text{neurons}}}_{\text{N x N}} = \frac{1}{T}\underbrace{\mathbf{M}}_{\text{N x T}}\underbrace{\mathbf{M}^T}_{\text{T x N}}$$

Alternatively, we can calculate covariance from the neuron point of view:

$$\underbrace{\mathbf{C}^{\text{time}}}_{\text{T x T}} = \frac{1}{N}\underbrace{\mathbf{M}^T}_{\text{T x N}}\underbrace{\mathbf{M}}_{\text{N x T}}$$

where $\mathbf{C}_{ij}^{\text{time}}$ is the covariance between time bins i and j

2. **PCA: How to find the direction of the greatest variance?**
What direction captures most of $\mathbf{M}$'s variance? This works similarly from either the time perspective or the neuron perspective. First we show from the time perspective. For each neuron's data $\vec{M}_{i,:}$, the variance across the time domain along an arbitrary direction defined by the unit vector $\vec{v}$ is $\|\vec{M}_{i,:} \cdot \vec{v}\|$. Thus for all of the neurons at once, we have:

$$\begin{aligned}\sigma^2_{\vec{v}\text{ time}} &= \|M\vec{v}\|^2 \\ &= (M\vec{v})^T(\vec{v}M) = (\vec{v}M)^T(M\vec{v}) \\ &= \vec{v}^T M^T M\vec{v} \\ &\propto \vec{v}^T\mathbf{C}^{\text{time}}\vec{v}\end{aligned}$$

The neuron perspective is similar. For each time bin's data $\vec{M}_{:,i}$, the variance across the time domain along an arbitrary direction defined by the unit vector $\vec{v}$ is $\|\vec{M}_{:,i} \cdot \vec{v}\|$. Thus for all of the neurons at once, we have (simply replace M with M^T):

$$\begin{aligned}\sigma^2_{\vec{v}\text{ neuron}} &= \|M^T\vec{v}\|^2 \\ &= (M^T\vec{v})^T(\vec{v}M^T) = (\vec{v}M^T)^T(M^T\vec{v}) \\ &= \vec{v}^T MM^T\vec{v} \\ &\propto \vec{v}^T\mathbf{C}^{\text{neuron}}\vec{v}\end{aligned}$$

Regardless of which perspective, we end up at the form of $\vec{v}^T\mathbf{C}\vec{v}$.

We want to find the vector $\vec{v}$ with maximal variance, subject to the constraint that $\vec{v}$ is of unit length [3]. That is:

$$\max_{\vec{v}} \sigma^2_{\vec{v}} \text{ such that } ||\vec{v}||^2 = 1$$

[2]Here we are normalizing by T rather than $T-1$ as it is all of the data, not a sample. In our exploration of SVD and PCA, this is a minor issue: you can ignore the difference and still have a good understanding of SVD and PCA. And you will soon see that this would not affect the result of SVD.

[3]This section is a more straightforward version of page 376 of Strang 5E. I recommended reading page 376 nevertheless.

We use the Lagrange multiplier technique[4], so set $\sigma_{\vec{v}}^2$'s derivative parallel to $\vec{v}$:

$$\begin{aligned}
\nabla_{\vec{v}} \vec{v}^T \mathbf{C} \vec{v} &= \lambda \nabla_{\vec{v}} \vec{v}^T \vec{v} \\
\vec{v}^T \frac{\partial \mathbf{C}\vec{v}}{\partial \vec{v}} + \frac{\partial \vec{v}^T}{\partial \vec{v}} (\mathbf{C}\vec{v}) &= 2\lambda \vec{v} \qquad (\text{“Product Rule” } \frac{\partial \mathbf{u}^\top \mathbf{v}}{\partial \mathbf{x}} = \mathbf{u}^\top \frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}^\top \frac{\partial \mathbf{u}}{\partial \mathbf{x}}) \\
\mathbf{C}^T \vec{v} + \mathbf{C}\vec{v} &= 2\lambda \vec{v} \\
2\mathbf{C}\vec{v} &= 2\lambda \vec{v} \qquad \mathbf{C}^T + \mathbf{C} = 2\mathbf{C} \text{ for } \mathbf{C} \text{ symmetric} \\
\mathbf{C}\vec{v} &= \lambda \vec{v}
\end{aligned}$$

Notice that this is the eigenvector equation for the covariance matrix $\mathbf{C}$. That is, eigenvectors of $\mathbf{C}$ are the directions of the highest variance or the lowest variance—with the eigenvector associated with the largest eigenvalue having the highest variance and vice versa (substitute $\mathbf{C}\vec{v} = \lambda\vec{v}$ into the objective function $\vec{v}^T\mathbf{C}\vec{v}$, you see λ). This is the goal of PCA—to find directions to project data in terms of the greatest/lowest variance, and we have reached its answer. One comment before we leave PCA. Since both covariance matrices are symmetric, their eigenvectors are orthogonal automatically: this gives uncorrelated directions of difference variance size.

The two perspectives to calculate the covariance are not stand-alone—they are closely connected—by SVD. Before we go into the next section for SVD, let us write down the eigen-decompositions that give us the answer of the PCA problem.

Since both $\mathbf{C}$ are symmetric, given the Spectral Theorem (Chapter 6.4 on Strang), we can decompose $\mathbf{C} = Q\Lambda Q^{-1}$. Specifically, we denote the two perspectives as the following:

$$\frac{1}{T} \underbrace{\mathbf{M}}_{\text{N x T}} \underbrace{\mathbf{M}^T}_{\text{T x N}} = \underbrace{\mathbf{C}^{\text{neurons}}}_{\text{N x N}} = \underbrace{\mathbf{U}}_{\text{N x N}} \underbrace{\mathbf{\Lambda}^{\text{neurons}}}_{\text{N x N}} \underbrace{\mathbf{U}^T}_{\text{N x N}}$$

where $\mathbf{U}$ is the matrix of eigenvectors of $\mathbf{C}^{\text{neurons}}$, and $\mathbf{\Lambda}^{\text{neurons}}$ is the diagonal matrix of eigenvalues.

$$\frac{1}{N} \underbrace{\mathbf{M}^T}_{\text{T x N}} \underbrace{\mathbf{M}}_{\text{N x T}} = \underbrace{\mathbf{C}^{\text{time}}}_{\text{T x T}} = \underbrace{\mathbf{V}}_{\text{T x T}} \underbrace{\mathbf{\Lambda}^{\text{time}}}_{\text{T x T}} \underbrace{\mathbf{V}^T}_{\text{T x T}}$$

where $\mathbf{V}$ is the matrix of eigenvectors of $\mathbf{C}^{\text{time}}$, and $\mathbf{\Lambda}^{\text{time}}$ is the diagonal matrix of eigenvalues. Note that if $T > N$, there are at most N non-zero eigenvalues.

3. **SVD: connecting U and V.**
We start from

$$\frac{1}{N} M^T M \vec{v_i} = \lambda_i^{\text{time}} \vec{v_i} \tag{1}$$

First, multiply v_i^T on both sides of the equation 1: $\vec{v_i}^T M^T M \vec{v_i} = N\lambda_i^{\text{time}} \vec{v_i}^T \vec{v_i}$. This gives $\|M\vec{v_i}\|^2 = N\lambda_i^{\text{time}}$.
Now we multiply M on both sides of the equation 1: $MM^T M\vec{v_i} = N\lambda_i^{\text{time}} M\vec{v_i}$. To see where it can lead to, add in parentheses: $MM^T(M\vec{v_i}) = N\lambda_i^{\text{time}}(M\vec{v_i})$. This means that $M\vec{v_i}$ is an eigenvector of MM^T. Since we have a matrix of eigenvectors in $\mathbf{U}$, let's pick u_i **(when needed it is $-u_i$)** so that

$$\vec{u_i} = M\vec{v_i}/\|M\vec{v_i}\| = M\vec{v_i}/\sqrt{N\lambda_i^{\text{time}}}. \tag{2}$$

when λ_i^{time} is not zero. This tentatively gives:

$$\underbrace{\mathbf{U}}_{\text{N x N}} \underbrace{\Sigma^1}_{\text{N x T}} = \underbrace{M}_{\text{N x T}} \underbrace{\mathbf{V}}_{\text{T x T}}$$

[4] An Appendix is available at the end of the document for a refresher

Now we need to figure out what is in the Σ^1 exactly. If M has rank r, then it can be easily shown that M^TM also has rank r. We can also see that M^TM has $N-r$ nonzero eigenvalues—let's fill in the The first N rows/columns of Σ^1 with these nonzero eigenvalues on the diagonal. Check that they satisfy equation 2. What to fill in the rest $T-N$ columns of Σ^1? Knowledge from earlier of the semester about matrix spaces come into play: You should note that (1) the first r columns of $\mathbf{V}$ span the row space of M so that the first r columns of $\mathbf{U}$ are not zero vectors (in the nullspace of M) but in the column space of M. (2) Given the orthorgonality of the 4 subspaces, you can fill in that the last T-r columns of $\mathbf{V}$ span the nullspace of M; the last N-r columns of $\mathbf{U}$ span the left nullspace of M. (3) This then gives that the last T-r columns of Σ^1 are zero column vectors to force zero vectors in the last T-r columns of $M\mathbf{V}$. It also gives us that the last N-r rows of Σ^1 are zero row vectors, as you cannot map either row space or nullspace to left nullspace. Thus we have filled in Σ^1.

Writing the equation just above in another way, we have

$$M = \mathbf{U}\Sigma^1\mathbf{V}^T$$

This is called the singular value decomposition: nonzero positive diagonal entries in Σ^1 are called the singular values. **The singular values are forced to be positive, if a negative sign occur, move it into either that column of U or V.** Do you see another connection to PCA now? It is easier seen with the form $M\mathbf{V} = \mathbf{U}\Sigma$. Thus the biggest Principle Component (PC) from the perspective of time is simply the first column of U times σ_1. And PC's (time perspective) are simply in $\mathbf{U}$'s columns. What's more, all the PC's are also orthogonal to each other.

As a summary till now, we started from equation 1, which involves M^TM and its eigenvectors, and derived the SVD decomposition. But can we start from MM^T?

We can start from the MM^T. Paralleling what we did to equation 1, we start from equation 3 below

$$\frac{1}{T}MM^T\vec{u_i} = \lambda_i^{\text{neuron}}\vec{u_i} \tag{3}$$

and we obtain the same form for M:

$$M = \mathbf{U}\Sigma_2\mathbf{V}^T$$

where $\Sigma_2 = \sqrt{T\lambda_i^{\text{neuron}}}$. We conclude that $\Sigma_1 = \Sigma_2$, which is saying $T\lambda_i^{\text{neuron}} = N\lambda_i^{\text{time}}$. [5]

$$\mathbf{\Sigma} = \underbrace{\begin{pmatrix} \sqrt{T\lambda_1^{\text{neurons}}} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{T\lambda_2^{\text{neurons}}} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \sqrt{T\lambda_r^{\text{neurons}}} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}}_{\text{N x T}} = \underbrace{\begin{pmatrix} \sqrt{N\lambda_1^{\text{time}}} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{N\lambda_2^{\text{time}}} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \sqrt{N\lambda_r^{\text{time}}} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}}_{\text{N x T}}$$

If we do not scale our covariance matrices with N or T (which can be seen as N=T=1), then we obtain the mathematically elegant form with $\lambda_i^{\text{neuron}} = \lambda_i^{\text{time}} = \lambda_i$, that is, the eigenvalues of the two matrices are the same. This gives the SVD form in our book:

$$M = \mathbf{U}\Sigma\mathbf{V}^T = \sigma_1\vec{u_1}\vec{v_1}^T + ... + \sigma_r\vec{u_r}\vec{v_r}^T$$

where the Σ is a diagonal matrix with entries σ_i being the squre root of the eigenvalue associated with the eigenvector $\vec{v_i}$ of the covariance matrix M^TM (or with the the eigenvector $\vec{u_i}$ of the covariance

[5] The Σ^1 or Σ^2 is exactly the SVD's Σ. Note how the Example 1 in Strang 5E (on page 383) is solved wrong (the factor missing). Note that in the book it uses $T = 6 - 1 = 5$ rather than 6 to normalize the covariance matrix$_{\text{student}}$.) You should notice that whatever 5 or 6, the Σ is not affected as normalizing factors are multiplied back in SVD's Σ.

matrix MM^T.) In general, we sort the columns of **U** and **V** with the descending order of the value of $\sigma's$—large σ means large λ, which means more variance in the direction of the eigenvector with that variance λ.

Reading This example shows you how to hand calculate the SVD of a small matrix: Strang5E 7.2-Example 3

Exercise When is $A = U\Sigma V^T = X\Lambda X^{-1}$?

Exercise (Strang 7.2-17) (The singular values are all positive.) Suppose A is a 2 by 2 symmetric matrix with unit eigenvectors u_1 and u_2. If its eigenvalues are $\lambda_1 = 3$ and $\lambda_2 = -2$. What is the SVD of A?

Exercise (Strang 7.2-23) If Q is an orthogonal matrix, why do all its singular values equal 1?

Exercise (Strang 7.2-16) Suppose A has orthogonal columns w_1, w_2,...,w_n of length σ_1, σ_2,..., σ_n. What are U, Σ, and V in the SVD of A?

4. **Building intuition on the result of PCA/SVD**
   Reading Strang5E 7.2-"An Extreme Matrix" on Page 374.

Practice By Yourself

Chapter 7.1: 2.

Chapter 7.2: Worked Problem B, 13, 3, 21 (The singlular values of $A + I$ are not $\sigma_j + 1$), **10**, 24 (Once you reached the matrix as $\begin{bmatrix}1 & 4\end{bmatrix}$, you should see that the one big variance is alone the direction of (1,4). The variance is 1+16=17.), 25.

Lagrange Multipliers and the Rayleigh Quotient

Our goal is to find the extremum of the function $f(x) : R^n \to R$ subject to the constraints $h_i(x) = 0$ for $i = 1 \dots m$ ($h_i : R^n \to R$).

Theorem (Lagrange multipliers): If x^* is an extremum of f subject to the constraints $h_i(x) = 0$ there exist scalars $\lambda_1, \dots, \lambda_m$ such that

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) = 0 \tag{1}$$

where ∇f is the gradient of f. In other words, if x^* is an extremum subject to the constraints then

$$\frac{\partial f(x^*)}{\partial x_j} + \sum_{i=1}^{m} \lambda_i \frac{\partial h_i(x^*)}{x_j} = 0 \text{ for } j = 1 \dots n \tag{2}$$

$\lambda_1, \dots, \lambda_m$ are called Lagrange multipliers.

Example: The Rayleigh Quotient

$$\max_x \frac{x^T A x}{x^T x} \tag{3}$$

where A is symmetric.

Notice that $\frac{x'^T A x'}{x'^T x'} = \frac{x^T A x}{x^T x}$ for $x' = cx$ and $c \neq 0 \in R$, therefore we will solve for x with a unit norm $\|x\|_2^2 = 1$.

$$\begin{aligned} &\max x^T A x \\ &\text{s.t. } x^T x = 1 \end{aligned} \tag{4}$$

The Lagrangian is

$$L(x) = x^T A x + \lambda(x^T x - 1) \tag{5}$$

Taking the derivative with respect to x:

$$\frac{\partial L(x)}{\partial x} = x^T(A + A^T) + 2\lambda x^T \tag{6}$$
$$\frac{\partial L(x)}{\partial \lambda} = x^T x - 1 \text{ (the original constraint)} \tag{7}$$
$$\frac{\partial L(x)}{\partial x} = 0 \Rightarrow \tag{8}$$
$$x^T(A + A^T) = -2\lambda x^T \Rightarrow \tag{9}$$
$$(A + A^T)x = -2\lambda x \Rightarrow \text{(A is symmetric)} \tag{10}$$
$$Ax = \tilde{\lambda} x \text{ where } (\tilde{\lambda} = -2\lambda) \tag{11}$$

Hence the maximum and the minimum are obtained for x an eigenvector of A (the Lagrange multipliers provide a necessary condition. The extremum is indeed obtained because $x^T Ax$ is a continuous function and the unit sphere is a compact set). For x an eigenvector of A with unit norm, $x^T Ax = x^T \lambda x = \lambda x^T x = \lambda$. Therefore the maximum is obtained at the eigenvector corresponding to the largest eigenvalue of A.

The Generalized Rayleigh Quotient is:

$$\max_x \frac{x^T Ax}{x^T Bx} \tag{12}$$

For A, B symmetric and positive definite. Again, to choose a certain solution we will constrain x:

$$\begin{aligned} &\max_x x^T x \\ &\text{s.t. } x^T Bx = 1 \end{aligned} \tag{13}$$

We will solve the Generalized Rayleigh Quotient by reduction to the Rayleigh Quotient.

Define $B = D^T D$, $C = D^{-T} AD^{-1}$ and $y = Dx$. Notice that $C \in PSDN$.

$$\frac{x^T Ax}{x^T Bx} = \frac{x^T D^T D^{-T} AD^{-1} Dx}{x^T D^T Dx} = \frac{y^T Cy}{y^T y} \tag{14}$$

This is the Rayleigh Quotient with the symmetric matrix C and the unit vector y ($y^T y = x^T D^T Dx = x^T BX = 1$). The solution is the first eigenvector of C. Notice that the first eigenvalue of C and $B^{-1}A$ is the same (substitute $y = Dx$ in $D^{-T} AD^{-1} y = \lambda y$), but their first eigenvector is different.