Mathematical Tools for Neuroscience (NEU 314)
Princeton University, Fall 2021
Jonathan Pillow

Notes: Principal Components Analysis (PCA)

(Lecture 11)

1 Setup

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.

Let's think of the data arranged in an $N \times d$ matrix that we'll call X. Each row of this matrix is a data vector representing the response from d neurons to a single stimulus:

$$X = \begin{bmatrix} — \ \vec{x}_1 \ — \\ — \ \vec{x}_2 \ — \\ \vdots \\ — \ \vec{x}_N \ — \end{bmatrix}$$

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 "variance" in the data is captured by a low-dimensional projection? (Yes: it turns out the tool we are looking for is PCA!)

2 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:

$$\begin{aligned}\text{Sum of squared linear projections} = \sum_{i=1}^{N} (\vec{x}_i \cdot \vec{v})^2 &= ||X\vec{v}||^2 \\ &= (X\vec{v})^\top (X\vec{v}) \\ &= \vec{v}^\top X^\top X \vec{v} \\ &= \vec{v}^\top (X^\top X)\vec{v} \\ &= \vec{v}^\top C \vec{v},\end{aligned}$$

where $C = X^\top X$, under the constraint that $\vec{v}^\top \vec{v} = 1$.

It turns out that the solution corresponds to the top eigenvector of C (i.e., the eivenvector 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:

$$C = USU^\top,$$

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

$$\begin{aligned}\vec{u}_j^\top C \vec{u}_j &= \vec{u}_j^\top (USU^\top)\vec{u}_j \\ &= (\vec{u}_j U) S (U^\top \vec{u}_j) \\ &= \begin{bmatrix}0 & \cdots & 1 & \cdots & 0\end{bmatrix} \begin{bmatrix} s_1 & & & & \\ & \ddots & & & \\ & & s_j & & \\ & & & \ddots & \\ & & & & s_d \end{bmatrix} \begin{bmatrix}0 \\ \vdots \\ 1 \\ \vdots \\ 0\end{bmatrix} \\ &= s_j\end{aligned}$$

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

$$\vec{v} = \vec{u}_1$$

.

This vector (the "dominant" eigenvector of C) is the *first principle component* (sometimes called the "first PC" for short).

3 Finding best k-dimensional subspace

PCA finds an orthonormal basis for the k-dimensional subspace that maximizes the sum-of-squares of the projected data. The solution is given by the singular value decomposition (which is also the eigenvector decomposition) of $X^\top X$:

$$(X^\top X) = USU^\top,$$

The first k columns of U are the first k principal components: $\{\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_k\}$.

The singular values correspond to the sum-of-squares of the data vectors projected into the corresponding principal component:

$$s_j = \sum_{i=1}^{N} (\vec{u}_j \cdot \vec{x}_i)^2$$

3.1 Fraction of variance

The squared Frobenius norm of X is (surprisingly!) equal to the sum of the singular values:

$$||X||_F^2 = \sum_{i=1}^{N} ||\vec{x}_i||^2 = \sum_{j=1}^{D} s_j$$

The *fraction* of the total variance accounted for by the first k principal components is therefore given by:

$$\frac{s_1 + \cdots + s_k}{s_1 + \cdots + s_k + \cdots + s_d}.$$

3.2 Fitting an ellipse to your data

PCA is equivalent to fitting an ellipse to your data: the eigenvectors $\vec{u}_i$ give the dominant axes of the ellipse, while the s_i gives the elongation of the ellipse along each axis, and is equal sum of squared projections (what we've been calling "variability" above) of the data along that axis.

4 Zero-centering

So far we've assumed we wanted to maximize the sum of squared projections of the vectors $\{\vec{x}_i\}$ onto some subspace, which is equivalent to using an ellipse centered at the origin to describe the data. In most applications, we want to consider an ellipse *centered on the data*, and find principal components that describe the spread of the datapoints relative to the mean.

To "center" the dataset at zero, we can simply subtract off the mean from each data vector. The mean is given by

$$\bar{x} = \frac{1}{N} \sum \vec{x}_i$$

.

Then the zero-centered data matrix can be formed as by placing $\vec{z}_i = \vec{x}_i - \bar{x}$ on each row:

$$Z = \begin{bmatrix} — & \vec{z}_1 & — \\ & \vdots & \\ — & \vec{z}_N & — \end{bmatrix}$$

Then by taking the SVD of $(Z^\top Z)$ we will be obtaining principal components of the centered data. Note: this the *standard* definition of PCA! It is less common to do PCA on uncentered data.

Python implementation

In python, we can achieve zero-centering (and division by N) with the function `np.cov`. That is, `np.cov(X)` will return

$$\tfrac{1}{N-1}(Z^\top Z),$$

Appendix: Derivation for PCA

In the lectures on PCA we showed that *if* we restricted ourselves to considering eigenvectors of the $X^\top X$, then the eigenvector with largest eigenvalue captured the largest projected-sum-of-squares of the vectors in X. But we didn't show that eigenvectors themselves correspond to optimal solution.

To recap briefly, we want to find the maximum of

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

where $C = X^\top X$ is the (scaled) covariance of zero-centered data vectors $\{\vec{x}_i\}$, subject to the constraint that $\vec{v}$ is a unit vector ($\vec{v}^\top \vec{v} = 1$).

We can solve this kind of optimization problem using the method of Lagrange multipliers. The basic idea is that we minimize a function that is our original function plus a lagrange multiplier λ times an expression that is zero when our constraint is satisfied. For this problem we can define the Lagrangian:

$$L = \vec{v}^\top C \vec{v} + \lambda(\vec{v}^\top \vec{v} - 1). \tag{1}$$

We will want solutions for which

$$\frac{\partial}{\partial \vec{v}}\ L = 0 \tag{2}$$

$$\frac{\partial}{\partial \lambda}\ L = 0. \tag{3}$$

Note that the second of these is satisfied if and only if $\vec{v}$ is a unit vector (which is reassuring).

The first equation gives us:

$$\frac{\partial}{\partial \vec{v}}\ L = \frac{\partial}{\partial \vec{v}}\ \vec{v}C\vec{v} + \lambda(\vec{v}^\top \vec{v} - 1) = 2C\vec{v} - 2\lambda\vec{v} \quad = \quad 0, \tag{4}$$

which implies

$$C\vec{v} = -\lambda\vec{v}. \tag{5}$$

What is this? It's the eigenvector equation! This implies that the derivative of the Lagrangian is zero when $\vec{v}$ is an eigenvector of C. So this establishes, combined with the argument from last week, that the unit vector that captures the greatest squared projection of the raw data is the top eigenvector of C.

Objective functions for PCA

Formally, we can write the principal components as the columns of a $d \times k$ matrix B that maximizes the Frobenious norm of the data projected onto B:

$$\hat{B}_{pca} = \arg\max_{B} ||XB||_F^2$$

such that $B^\top B = I$.

An equivalent definition is

$$\hat{B}_{pca} = \arg\min_{B} ||X - XBB^\top||_F^2$$

such that $B^\top B = I$. This objective function says that the principal components define an orthonormal basis such that the distance between the original data and the data projected onto that subspace is minimal. It shouldn't take to much effort to see that that the rows of $XBB^\top$ correspond to the rows of X reconstructed in the basis defined by columns of B.