**Statistical Modeling and Analysis of Neural Data (NEU 560)**
Princeton University, Spring 2018
Jonathan Pillow

**Lecture 5 notes:** PCA, part II
Tues, 2.20

# 1 Principal Components Analysis (PCA)

Review of basic setup:

- $N$ vectors, $\{\vec{x}_1, \ldots \vec{x}_N\}$, each of dimension $d$.
- find $k$-dimensional subspace that captures the most "variance".
- To be more explicit: find the projection such that the sum-of-squares of all projected datapoints is maximized.
- Let's think of the data arranged in an $N \times d$ matrix, where each row is a data vector:

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

## 1.1 Frobenius norm

The *Frobenius norm* of a matrix $X$ is a measure of the "length" of a matrix. It behaves like the Euclidean norm but for matrices: it's equal to the square-root of the sum of all squared elements in a matrix. It's written:

$$||X||_F = \sqrt{\sum_{ij} X_{ij}^2},$$

where $i$ and $j$ range over all entries in the matrix $X$. The Frobenius norm gives the same quantity as if we stacked all of the columns of $X$ on top of each other in order to form a single vector out of the matrix.

An equivalent way to write the Frobenius norm using matrix operation is using the trace of $X^\top X$:

$$||X||_F = \sqrt{\mathrm{Tr}[X^\top X]}.$$

## 1.2 PCA solution: 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:

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

### 1.3 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:

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

### 1.4 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.

### 1.5 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 uncommon to do PCA on uncentered data.

### 1.6 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

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

## 2 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 $\lambda$ 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 $C$.

## 2.1 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 $B$.