Lecture 25: The Covariance Matrix and Principal Component Analysis

Anup Rao

May 31, 2019

The covariance of two variables X, Y is a number, but it is often very convenient to view this number as an entry of a bigger matrix, called the *covariance matrix*. Suppose $X = (X_1, X_2, \ldots, X_k)$ is a random vector. Then the covariance matrix K is the matrix with $K_{i,j} = \text{Cov}\left[X_i, X_j\right]$. It is a $k \times k$ matrix, and the entries on the diagonal are the variances of all the variables. If $X - \mathbb{E}\left[X\right]$ is viewed as a column vector, the covariance matrix is

$$K = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])(X - \mathbb{E}\left[X\right])^\intercal\right].$$

If X is a random matrix, then its expectation is the matrix whose entries are the expectation of the corresponding random entries of X. So, $\mathbb{E}\left[X\right] = \sum_x p(x) \cdot x$, where here x is a matrix.

The matrix representation is particularly nice for working with data. Suppose $X = (X_1, \ldots, X_k)$ is a random vector sampled by setting X to be a uniformly random vector from the set $\{v_1, v_2, \ldots, v_n\}$. Let $\overline{v}$ be the average of these vectors. Then let V be the $k \times n$ matrix whose columns are $v_1 - \overline{v}, v_2 - \overline{v}, \ldots, v_n - \overline{v}$. Then it is easy to check that the covariance matrix of X is

$$K = (1/n) \cdot VV^\intercal.$$

The matrix representation allows us to easily generalize linear regression to the case that the inputs themselves are vectors. Suppose we are given data in the form of pairs $(v_1, y_1), (v_2, y_2), \ldots, (v_n, y_n)$, where $v_1, \ldots, v_n$ are each k-dimensional column vectors, and $y_1, \ldots, y_n$ are numbers. For simplicity, let us assume again that the mean of the v's is 0, and the mean of the y's is 0. Let X be a random one of the n vectors, and Y be the corresponding number. Then we seek the vector a such that $\mathbb{E}\left[(a^\intercal X - Y)^2\right]$ is minimized. We have:

$$\begin{aligned}\nabla \mathbb{E}\left[(a^\intercal X - Y)^2\right] &= \nabla \mathbb{E}\left[Y^2 - 2Ya^\intercal X + a^\intercal XX^\intercal a\right] \\ &= 2a^\intercal \mathbb{E}\left[XX^\intercal\right] - 2\,\mathbb{E}\left[YX\right],\end{aligned}$$

but here $\mathbb{E}\left[XX^\intercal\right]$ is exactly the covariance matrix. So, it turns out that the solution is to set $a^\intercal = \mathbb{E}\left[YX\right] \cdot K^{-1}$.

K may not be invertible in general, but you can still solve for a using the *pseudoinverse* of K.

Principal Component Analysis

The covariance matrix has an extremely important application in machine learning. In machine learning, we seek to find the key features of large sets of data, features that are most informative about the underlying data set.

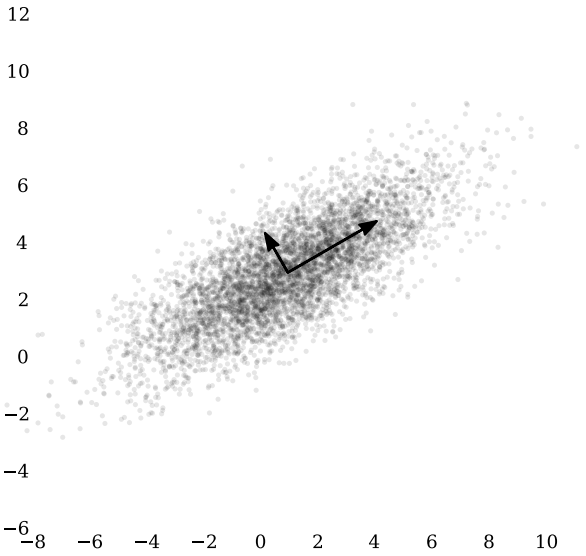

Figure 1: PCA applied to two dimensional data. Image credit: Nicoguaro, from Wikipedia.

For example, if I were to try and figure out what kinds of movies a particular student in 312 likes, I might ask the student to tell me how much they like movies from the following categories: action, romantic comedies, horror, ... But are these the right questions to ask? Should I be asking whether they like animated movies or not, or whether they like slapstick humor? One of the key developments in machine learning is the use of mathematics to eliminate the subjectivity here—data is used to generate the most informative questions.

How can we bring the math to bear in the this scenario? Let us associate every student in the class with a column vector—say a k dimensional vector x, where k is the number of movies on Netflix, and each coordinate indicates whether or not the student likes the movie. Let X be the vector corresponding to a uniformly random student in the class. If I want to boil down a students preferences to a single number, it is natural to consider a number that is a linear function of their vector. So, I want a unit length vector a such that the inner product

$$a^\intercal x = a_1 x_1 + a_2 x_2 + \cdots + a_k x_k$$

is the most informative. Then a natural approach is to pick a to maximize $\mathrm{Var}[a^\intercal X]$. Geometrically, this corresponds to projecting all the k dimensional vectors to a single direction, in such a way that the variance is maximized.

There is a very nice web demo here: `http://setosa.io/ev/principal-component-analysis/`.

To understand how to find this direction a, we need to use the concept of eigenvalues and eigenvectors from linear algebra. Given a

square matrix M, a number $\lambda \in \mathbb{R}$ is an eigenvalue of M if there is a vector v such that $Mv = \lambda v$. In general, eigenvalues can be complex numbers, even if the original matrix is only real valued. However, we have:

Fact 1. *If M is a $k \times k$ symmetric matrix, then it has exactly k real valued eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$, and the corresponding eigenvectors form an orthonormal basis $v_1, \ldots, v_k$ for $\mathbb{R}^k$.*

Returning to our problem above, we seek the vector a that maximizes $\mathrm{Var}[a^\intercal X]$. For simplicity, let us as usual assume that $\mathbb{E}[X] = 0$, since if this is not the case, we can consider the random variable $X' = X - \mathbb{E}[X]$ instead, and this does not change the choice of a. The assumption that the mean is 0 ensures that $\mathbb{E}[a^\intercal X] = a_1 \mathbb{E}[X_1] + \cdots + a_n \mathbb{E}[X_n] = 0$. Then we have

$$\mathrm{Var}[a^\intercal X] = \mathbb{E}\left[(a^\intercal X)^2\right] = \mathbb{E}[a^\intercal X X^\intercal a] = a^\intercal \mathbb{E}[XX^\intercal]\, a = a^\intercal K a,$$

where K is the covariance matrix of X.

Since K is a symmetric matirx, Fact 1 applies. Let $\lambda_1, \ldots, \lambda_k, v_1, \ldots, v_k$ be as in the fact. Then we can express $a = \alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_k v_k$. The fact that a has length 1 corresponds to

$$\sqrt{\alpha_1^2 + \cdots + \alpha_k^2} = 1 \Rightarrow \alpha_1^2 + \cdots + \alpha_k^2 = 1.$$

We have:

$$\begin{aligned} & a^\intercal K a \\ &= (\alpha_1 v_1 + \cdots + \alpha_k v_k)^\intercal K (\alpha_1 v_1 + \cdots + \alpha_k v_k) \\ &= (\alpha_1 v_1 + \cdots + \alpha_k v_k)^\intercal (\lambda_1 \alpha_1 v_1 + \cdots + \lambda_k \alpha_k v_k) \\ &= \lambda_1 \alpha_1^2 + \cdots + \lambda_k \alpha_k^2. \end{aligned}$$

Since v_i, v_j are orthonormal, for $i \neq j$, we have $v_i^\intercal v_j = 0$, and $v_i^\intercal v_i = 1$.

Since λ_1 is the largest eigenvalue, and $\alpha_1^2 + \ldots + \alpha_k^2 = 1$, this quantity is maximized when $\alpha_1 = 1$, and $\alpha_2 = \alpha_3 = \cdots = \alpha_k = 0$. So, the largest variance is achieved when $a = v_1$, the eigenvector with the largest eigenvalue! This eigenvector is called the *principal component*.

In general, if you want the 5 most informative numbers, you should pick the eigenvectors corresponding to the top 5 eigenvalues, and ask for the corresponding linear combinations.