Lecture 13
Introduction to Principal Component Analysis

Jason J. Bramburger

In the previous lecture we were introduced to principal component analysis (PCA). Here we are going to fill in some of the details that we brushed over, while also giving a real introduction to PCA. As we saw, the SVD is just a way of factoring matrices, but it turns out to be very useful for applications. In fact, it has been discovered numerous times in a variety of different disciplines. For that reason, many similar techniques have different names, but end up just being the SVD. This includes PCA, Proper Orthogonal Decomposition (POD), Karhunen-Leove Decomposition, Hotelling transform, empirical orthogonal functions, or in general, reduced order modelling.

Reviewing Some Statistics

We need to start with a brief review of some statistics. Recall that the singular values and vectors we saw in the previous lecture were related to variance in some way. Let's remind ourselves what variance is first. Variance is the spread of data. Suppose we have a vector

$$\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix}.$$

You should recall that the mean (or average) of the data in the vector is given by

$$\mu = \frac{1}{n}\sum_{k=1}^{n} a_k.$$

Then, the square of the variance is given by

$$\sigma^2 = \frac{1}{n}\sum_{k=1}^{n}(a_k - \mu)^2.$$

We are always going to have to subtract the mean off from our data, so it would be helpful to just assume that this has already been done. Recall that we did this for the weight/height data. Furthermore, this amounts to assuming that $\mu = 0$ in the above formulas. Thus,

$$\sigma^2 = \frac{1}{n}\sum_{k=1}^{n} a_k^2.$$

Notice that this is just the 2-norm squared, i.e. $\sigma^2 = (1/n)\|\mathbf{a}\|_2^2$. We can equivalently write this as the inner product (dot product)

$$\sigma^2 = \frac{1}{n}\mathbf{a}\mathbf{a}^T.$$

We now need to make one correction. Above is the formula for calculating the variance of a population. For example, if the vector $\mathbf{a}$ contains the grades for every student in the class, we can use the above formulas to find the mean and the variance. In many real-world applications, we don't have data about the whole population. In this scenario, we just estimate it from a sample and we use the sample to try to estimate the population variance. To make this clear, go back to the homework grades example. We could randomly choose 20 students and check the variance of their grades. If we use the formula above, we will, on average, under-predict the spread of the data. In order to, on average, get the right answer, we need to divide by $n-1$ instead:

$$\sigma^2 = \frac{1}{n-1}\mathbf{a}\mathbf{a}^T.$$

On average, this will give the right population variance. This makes it an **unbiased estimator**. For a derivation of this fact see this link.

Now suppose we have two sets of data in vectors $\mathbf{a}$ and $\mathbf{b}$ of length n. For convenience, we will assume the means are already subtracted from them, meaning both vectors have mean zero. We can calculate the variance of each:

$$\sigma_a^2 = \frac{1}{n-1}\mathbf{a}\mathbf{a}^T, \qquad \sigma_b^2 = \frac{1}{n-1}\mathbf{b}\mathbf{b}^T.$$

We can further calculate the **covariance**, which measures how the variables vary with respect to each other. This is given by the formula

$$\sigma_{ab}^2 = \frac{1}{n-1}\mathbf{a}\mathbf{b}^T$$

Going back to our example, let us assume that vector $\mathbf{a}$ contains the scores on homework 1 and vector $\mathbf{b}$ contains the scores on homework 2. We would expect higher values in one vector to correspond to higher values in the other. This would lead to a positive covariance. We saw something similar in our weight and height data from last lecture. If the two variables are inversely related, then the covariance is negative. If the covariance is zero, then the two variables are **uncorrelated**. This is the case when the variables are **statistically independent**. A high covariance means they are highly correlated. This is important for data because it gives us an idea of how redundant our data is, i.e. how much information in the vector $\mathbf{b}$ can be obtained just from vector $\mathbf{a}$. If $\mathbf{a} = \mathbf{b}$ we see that we just get the variance of the vector back, which is the maximum possible covariance. This would mean that the vectors are completely redundant. When the vectors are uncorrelated, then each new variable brings brand new information. Importantly, knowing $\mathbf{a}$ doesn't tell you anything about $\mathbf{b}$. Hence, you can think of covariance as measuring how much redundancy is in the data.

Principal Components

Let us now imagine we have multiple vectors. For example, imagine there were 10 assignments and each vector contains grades from the same 20 randomly chosen students. We can put each of those row vectors into a matrix:

$$\mathbf{X} = \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \\ \mathbf{c} \\ \mathbf{d} \end{bmatrix}.$$

We can compute all the variances and covariances between the rows of $\mathbf{X}$ with one matrix multiplication:

$$\mathbf{C_X} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T = \begin{bmatrix} \sigma_a^2 & \sigma_{ab}^2 & \sigma_{ac}^2 & \sigma_{ad}^2 \\ \sigma_{ba}^2 & \sigma_b^2 & \sigma_{bc}^2 & \sigma_{bd}^2 \\ \sigma_{ca}^2 & \sigma_{cb}^2 & \sigma_c^2 & \sigma_{cd}^2 \\ \sigma_{da}^2 & \sigma_{db}^2 & \sigma_{dc}^2 & \sigma_d^2 \end{bmatrix}.$$

You should notice that $\mathbf{C_X}$ is a square symmetric matrix. Unsurprisingly, it is called the **covariance matrix**.

The goal of principal component analysis is to find a new set of coordinates (a change of basis) so that the variables are now uncorrelated. That will mean that each variable contains completely new information, i.e. no redundancies. It would also be nice to know which variables have the largest variance because these contain the most important information about our data. Therefore, we want to diagonalize this matrix so that all off-diagonal elements (covariances) are zero:

$$\mathbf{C_X} = \mathbf{V}\Lambda\mathbf{V}^{-1}.$$

The basis of eigenvectors contained in $\mathbf{V}$ are called the principal components. They are uncorrelated since they are orthogonal. Why? Since $\mathbf{C_X}$ is a symmetric matrix, its eigenvalues are real and the corresponding eigenvectors are orthogonal. The diagonal entries of Λ, the eigenvalues of $\mathbf{C_X}$, are the variances of these new variables.

Connection to the SVD

Principal component analysis should be reminding you of the SVD. Recall that the SVD of a matrix $\mathbf{A}$ is connected to the eigenvalue decomposition of $\mathbf{A}\mathbf{A}^T$. To account for the $1/(n-1)$ factor, consider

$$\mathbf{A} = \frac{1}{\sqrt{n-1}}\mathbf{X}.$$

Then,

$$\mathbf{C_X} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T = \mathbf{A}\mathbf{A}^T.$$

Then, from our work in Lecture 11 we have that

$$\mathbf{C_X} = \mathbf{A}\mathbf{A}^T = \mathbf{U}\Sigma^2\mathbf{U}^T,$$

where $\mathbf{U}$ is the (orthogonal) matrix of left-singular vectors and Σ is the diagonal matrix of singular values. So, the eigenvalues of the covariance matrix are the squares of the (scaled) singular values. Recall that this factor of $\sqrt{n-1}$ is exactly what we scaled the singular values by to get the right units.

Recall that we used a change of basis to work in the basis of the principal components. To do this you multiply by $\mathbf{U}^{-1} = \mathbf{U}^T$. The data in the new coordinates is

$$\mathbf{Y} = \mathbf{U}^T\mathbf{X}.$$

The covariance of $\mathbf{Y}$ is

$$\mathbf{C_Y} = \frac{1}{n-1}\mathbf{Y}\mathbf{Y}^T = \frac{1}{n-1}\mathbf{U}^T\mathbf{X}\mathbf{X}^T\mathbf{U} = \mathbf{U}^T\mathbf{A}\mathbf{A}^T\mathbf{U} = \mathbf{U}^T\mathbf{U}\Sigma^2\mathbf{U}^T\mathbf{U} = \Sigma^2.$$

Since the off-diagonal elements of Σ are zero, it follows that the variables in $\mathbf{Y}$ are uncorrelated.

Spring-Mass System

Imagine the following application coming from your homework assignment. Consider an experiment with a spring-mass system. That is, a mass is hanging from a spring and the mass bobs up and down. You can use Newton's second law and Hooke's law to get a second order ordinary differential equation for the vertical displacement of the mass. We won't bother with the actual equation, but simply note that the solution is given by

$$z(t) = A\cos(\omega t + \varphi),$$

where A and φ are determined by the initial displacement and velocity of the mass and the frequency ω can be found from the constants in the differential equation. Most important to our work here is that the motion is one-dimensional.

Imagine we don't have much physics training, so we wanted to understand the motion of the mass-spring system in another way. We can set up three cameras around the system and collect video. In the frame of each video, there are two dimensions. For camera 1, we can denote them (x_a, y_a) at each instance in time. We can do the same for camera 2 and camera 3. If we make vectors of all the position at each time, we can get the matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{y}_a \\ \mathbf{x}_b \\ \mathbf{y}_b \\ \mathbf{x}_c \\ \mathbf{y}_c \end{bmatrix}.$$

So, what can the SVD do for us? First, the actual data is one-dimension, but we have collected six dimensions of data! The SVD will help us weed out redundancies. This is just like what happened in the previous lecture with the line in 3D. This may seem like a silly example - particularly, why would we use three cameras? The answer is that in most application, we don't know how many dimensions our data will have. Therefore, we just have to take measurements and analyze it. In our spring-mass system, there *should* only be one nonzero singular value (representing the fact that the motion is 1D). Of course, no video recording or data collecting in perfect, and so we expect a little bit of noise that will throw things off a bit. There might also be some movement in the other directions if the mass isn't started at a perfectly vertical trajectory. Hence, we shouldn't expect only one nonzero singular value, but we should expect that one is significantly larger than all the others, representing the dominant up-down motion of the mass. That is, the large singular values help us pick out which directions are most important. If our camera isn't shot at the right angle, the first principal component will tell us we need to rotate it. If have really small singular values, we might choose to ignore them and treat them as noise.