Lecture 308.3

Date: 2 April 2020
Scribes: Akash Gupta
Topics: PCA intuition using Eigendecomposition and Change of bases

0.1 Introduction

The setting behind subsequent explanations rely on an example that considers a $3-D$ space wherein we have attached a mass at the end of a fixed spring which is frinctionless. We aim to measure its oscillations or frequency along the x axis. So we essentially place three viewing angles A, B, C at three axes (not knowing which axis is infact the x axis). After this we record measurements of the position of the mass over a certain amount of time. Consider that it moves with a frequency of $120hz$.

The primary function of Principal component analysis is to express the original dataset in the form of optimally chosen basis vectors that are different than the standard basis vectors. We expect this new basis to filter out noise from our dataset and reveal the most important variables in the system. This process helps us to understand in the sea of variables and data, as to which are the most important variables and which are the noisy variables. Now note that each time sample or experimental trial is treated as one sample in our dataset where we are recording the position of the mass over 10 minutes. The positions are recorded by all three camera's with respect to their $2-D$ image from their viewing angle : in essence they are recording a projection of the mass position from $3-D$ space to $2-D$ space. So the three $2-D$ measurements of the three cameras are collapsed into one vector X. Note that recording the ball position in an experimental trial for 10 minutes involves taking : $10 \times 60 \times 120 = 72000$ obersations. Our observation vector can be expressed as :-

$$\boldsymbol{X} = \begin{bmatrix} x_A \\ y_A \\ x_B \\ y_B \\ x_C \\ y_C \end{bmatrix} \tag{1}$$

We can here make some important statements like :- **Sample vector X is made up of m measurements or sample vector X is an m dimensional vector in a space that is spanned by some set of orthonormal basis vectors. Also every measurement vector in this space is essentially a linear combination of these orthonormal unit basis vectors**. Note yet another important point that since we have measure out data as per the standard basis, our coordinate vectors of sample measurements would naturally be formed by linear combinations of these basis vectors only. For an m dimensional case, we will end up forming a $m \times m$

matrix of these basis vectors :-

$$B = \begin{bmatrix} \boldsymbol{b_1} \\ \vdots \\ \boldsymbol{b_n} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \ddots & & \\ 0 & & & 1 \end{bmatrix} = I \tag{2}$$

Here each row is an orthonormal basis vector $\boldsymbol{b_i}$ with m components.

0.2 Changing the bases

Is there another set of basis vectors which is essentially a linear combination of the previous basis vectors that can re-express our dataset ?

Let X be our original dataset where each column represents a recording of the position of the mass at a certain point of time and essentially it is a 6×72000 matrix. Now let Y be another dataset that is related to the original dataset by linear transformation P. Y is our new representation of our dataset and is given by :-

$$PX = Y \tag{3}$$

Now we will define a few variables for further interpretation :-

- $\boldsymbol{p_i}$ are the rows of P
- $\boldsymbol{x_i}$ are the columns of X
- $\boldsymbol{y_i}$ are the columns of Y
- Aso remember that P is a matrix that transforms X into Y
- P can be treated as a matrix that rotates and then stretches X in the process of transformation
- The rows or $P : [\boldsymbol{p_1}, \cdots, \boldsymbol{p_m}]$ are the new basis vectors that express the coordinates of the X columns

We can see how this transformation plays out in matrices :-

$$PX = \begin{bmatrix} \boldsymbol{p_1} \\ \vdots \\ \boldsymbol{p_m} \end{bmatrix} \begin{bmatrix} \boldsymbol{x_1} & \cdots & \boldsymbol{x_n} \end{bmatrix} = \begin{bmatrix} \boldsymbol{p_1} \cdot \boldsymbol{x_1} & \cdots & \boldsymbol{p_1} \cdot \boldsymbol{x_n} \\ \vdots & \ddots & \\ \boldsymbol{p_m} \cdot \boldsymbol{x_1} & \cdots & \boldsymbol{p_m} \cdot \boldsymbol{x_n} \end{bmatrix} \tag{4}$$

Now note that each column of Y is given by the form :-

$$\boldsymbol{y_i} = \begin{bmatrix} \boldsymbol{p_1} \cdot \boldsymbol{x_1} \\ \vdots \\ \boldsymbol{p_m} \cdot \boldsymbol{x_1} \end{bmatrix} \tag{5}$$

An important point to note here is that each coefficient or coordinate of $\boldsymbol{y_i}$ is the inner product of $\boldsymbol{x_i}$ with the corresponding row of P. Note that the j^{th} **coefficient of $\boldsymbol{y_i}$ is simply the projection on to the** j^{th} **row of** P. Each column of Y is a projection on to the basis vectors of P. rows of P are the new basis vectors representing the columns of X. So we note finally that the row vectors of P which are the new basis vectors are also known as the principal components of X. Now the question is :- What is the best way to re-express X and what is a good choice of P. Also what features we would like Y to have ?

0.3 Variance

We realise at this point that we need to keep our noise in measurements low so as to get most information out of our signal. Also, all noise is measured in relative terms to the signal. With this we get the Signal to noise ratio which is a ratio of variances :-

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}} \tag{6}$$

A high SNR would indicate precise measurement and low SNR indicates more noise in the data. Here is a figure to give an idea of how signal and noise incorporate a series of measurements :- Each camera's motion recording should be expected to have a straight line motion, so any deviation in positioning resulting from the measurements is noise. Note that the small line depicts the variance of noise and the big line depicts variance of the signal. We assume that **Directions with largest variances in the measurement space must contain the dynamics of interest**. Also we can see that the largest variance or the most dominant direction of measurement is neither along the x basis or the y basis but along the slanted axis of our measurement points. That is why we have to look for other bases because our dominant direction of highest variance does not correspond to the directions of their of the standard basis vectors.

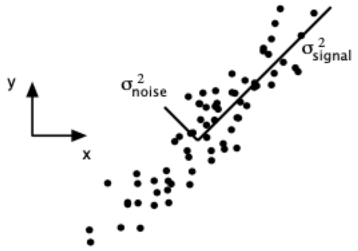

Figure 1: SNR

Note here that maximising the variance or our SNR is equivalent to appropriately rotating the standard basis vectors so as to adjust one of them along the best fit line to our data which also has the largest variance. Hence, rotating the standard basis vector along a line parallel to the dominant direction of our best fit line would reveal the direction of motion of our mass. An additional point to note is that measuring multiple variables causes redundancy or confounding. Two measure in this space seem to be correlated. As a result we can say that essentially measuring one variable is enough to express the data in a concise manner. This is the fundamental idea behind **dimensionality reduction**.

0.3.1 Covariance

In the case of 2 variables it is somewhat easy to expose confounding by fitting a best line on the data and assessing the quality of the fit. We will now generalize this notion to any dimension. Consider two sets of measurements or samples with 0 means :-

$$A = [a_1, \cdots, a_n], B = [b_1, \cdots, b_n] \tag{7}$$

The variance of A and B are given as follows :-

$$\sigma_A^2 = \frac{1}{n}\sum_{i=1}^{n} a_i^2, \sigma_B^2 = \frac{1}{n}\sum_{i=1}^{n} b_i^2 \tag{8}$$

The covariance is a measure of the degree of linear relationship between two variables. The covariance between A and B is given by :-

$$\sigma_{AB}^2 = \frac{1}{n}\sum_{i=1}^{n} a_i b_i \tag{9}$$

If A and B in the above example are considered to be row vectors we can write the covariance formula in matrix form as follows :-

$$\sigma_{ab}^2 = \frac{1}{n}\boldsymbol{a}\boldsymbol{b}^T \tag{10}$$

Further expanding this idea, if we consider the matrix X as containing many such row vectors of the form $\boldsymbol{a}$ and $\boldsymbol{b}$ then we can simultaneously find the covariance between each pair of these measurement vectors by computing the covariance matrix in the following way :-

$$C_X = \frac{1}{n}XX^T \tag{11}$$

Remember that this is a square $m \times m$ matrix with variance of each sample measurement vector along the diagonals of the matrix. The off diagonal terms are the covariance terms between various pairs of sample measurement vectors. Remember two crucial points :-

- The high values in diagonal elements reflect the interesting structure that explains the significant variance of the measurement that interests us.
- The off diagonal elements reflect the noise in the system.

We ideally want to manipulate this C_X covariance matrix to obtain C_Y which has some interesting optimal features.

0.3.2 Diagonalize the Covariance matrix

Remember that our ultimate aim is to minimize redundancy which is measure by the off diagonal terms in the covariance matrix and we wish to maximize the signal which is measure by the diagonal variance terms. We will what this optimized C_Y would look like. Remember two main points regarding the optimal form :-

- The off diagonal terms should be made 0 and in essence, the effect of the y measurements that cause confounding is decoupled from our measurements.
- Each successive dimension should be rank ordered as per variance values.

PCA comes in at this point wherein it is essentially a method of transformation the covariance matrix by diagonalizing it in the easiest way possible. That is by premultiplying the projection matrix P which is an orthonormal matrix since its columns which represent the new basis vectors are orthonormal vectors : $[\boldsymbol{p_1}, \cdots, \boldsymbol{p_m}]$. The way PCA works is that P essentially rotates the existing basis vector so that it alings with the dominant direction of the best fit line of maximum variance. The algorithm is as follows :-

- Select a normalized direction in the m dimensional space in which the variance in X is maximum. Select this direction as vector $\boldsymbol{p_1}$.
- Now find another direction along which variance is maximised but this time include only those direction to which the earlier found direction is orthongonal. This is due to the orthonormality constraints. Keep saving these vectors as $\boldsymbol{p_i}$.
- Repeat this procedure until m vectors are selected.

Note that the resultant ordered set of vectors $\boldsymbol{p_i}$ are precisely what are known as **principal components**. The importance of rank ordering these dirction vectors is that we can obtain easily the importance of each direction.

0.4 Solve PCA using Eigen Decomposition

Our goal for original dataset X is as follows :- **Find an orthonormal matrix P in the relation $PX = Y$ such that the covariance matrix $C_Y = \frac{1}{n}YY^T$ is a diagonal matrix. And that the rows of P are the principal components or new basis vectors of X** Here is how we find C_Y :-

$$C_Y = \frac{1}{n}YY^T \tag{12}$$

$$C_Y = \frac{1}{n}(PX)(PX)^T \tag{13}$$

$$C_Y = \frac{1}{n}PXX^TP^T \tag{14}$$

$$C_Y = P(\frac{1}{n}XX^T)P^T \tag{15}$$

$$C_Y = PC_XP^T \tag{16}$$

Additionally we note that any square symmetric matrix is infact diagonalized by an orthogonal matrix of its eigenvectors. This relation of similarity transformation or diagonalization is given by :-

$$A = EDE^T \tag{17}$$

Where A is a symmetric matrix and D is a diagonal matrix containing the eigenvalues on its diagonals. E is a matrix of eigenvectors where the eigenvectors are the orthogonal columns of matrix E. Now we select a matrix P such that its columns $\boldsymbol{p_i}$ are the eigenvectors of the X covariance matrix given by :-

$$\frac{1}{n}XX^T \tag{18}$$

Recall that for a projection matrix $P^{-1} = P^T, PP^{-1} = I$ and that we are choosing $P = E^T$. Now we can re-evaluate C_Y as :-

$$C_Y = PC_XP^T \tag{19}$$

$$C_Y = P(EDE^T)P^T \tag{20}$$

$$C_Y = P(P^TDP)P^T \tag{21}$$

$$C_Y = (PP^T)D(PP^T) \tag{22}$$

$$C_Y = (PP^{-1})D(PP^{-1}) \tag{23}$$

$$C_Y = D \tag{24}$$

And hence we find that with that particular choice of P our C_Y has been diagonalized. Our results can be summarized as follows :-

- The principal components of X are the eigenvectors of $C_X = \frac{1}{n}XX^T$.
- The i^{th} diagonal value of C_Y is the variance of X along the principal component $\boldsymbol{p_i}$.
- The process involves first centering the values of X and then finding the eigenvectors of C_X.

0.5 Using the SVD

Let X be a $n \times m$ matrix and X^TX be a $m \times m$ square symmetric matrix of rank r. We then note the following :-

- $[\boldsymbol{v_1}, \cdots, \boldsymbol{v_r}]$ is the set of orthonormal eigenvectors of the X^TX matrix with corresponding eigenvalues as $\lambda_1 \cdots, \lambda_r$. We can write the characteristic equation as :-

  $$(X^TX)\boldsymbol{v_i} = \lambda_i\boldsymbol{v_i} \tag{25}$$

- $\sigma_i = \sqrt{\lambda_i}$ are the positive singular values

- $[\boldsymbol{u_1}, \cdots, \boldsymbol{u_r}]$ are vectors such that :-

$$\boldsymbol{u_i} = \frac{1}{\sigma_i} X \boldsymbol{v_i} \tag{26}$$

- $\boldsymbol{u}.\boldsymbol{u_j} = 1 \; or \; 0$ if $i = j$ or $i \neq j$.
- $X\boldsymbol{v_i} = \sigma_i \boldsymbol{u_i}$

Note that the sets of eigenvectors v and the vectors u are orthonormal vectors in r dimensional space. Also Σ is a diagonal matrix as previously shown with diagonal elements containing the singular values in a rank ordered manner. Finally when we fill up all the v and u vectors into matrices we get the decomposition of X as follows :-

$$X = U\Sigma V^T \tag{27}$$

This essentially means the the orthonormal matrix U first rotates the X matrix, the Σ matrix stretches it and then V^T matrix again rotates it. In the equation $XV = \Sigma U$ we can think of the columns of V as input vectors and columns of U as ouput vectors wherein these vectors span the input and output spaces respectively. Now we present this manipulation :-

$$X = U\Sigma V^T \tag{28}$$

$$U^T X = \Sigma V^T \tag{29}$$

$$U^T X = Z \tag{30}$$

Where $Z = \Sigma V^T$. We note that in this last equation U^T can be essentially seen as a change of basis matrix such that X is re-expressed as Z. Note that the columns of V in this case are the principal components of X.

0.6 References

- Jonathon Schlens - A Tutorial on Principal Component Analysis