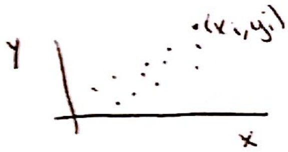
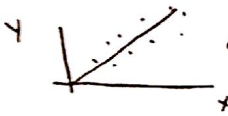
Why do the eigenvectors of the covariance matrix capture the direction along which the data varies the most?

⇒ To answer this, let's consider a simple case - like 2-D data.



Here's the accompanying data matrix: $X = \begin{bmatrix} x_1 & \cdots & x_n \\ y_1 & \cdots & y_n \end{bmatrix}$

Let's just think about what happens when we project this data onto 1-dimension. To capture the most structure in the data, we want this line to be in the direction with the most variance:



← like this

when projected:

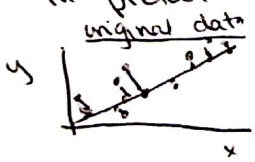
The accompanying equation for this transformation (from) to)) is:

$\underbrace{[u_1 \cdots u_n]}$ = $\underbrace{\begin{bmatrix} p_1 \\ p_2 \end{bmatrix}}_{\text{line that we projected data onto}} \underbrace{\begin{bmatrix} x_1 & \cdots & x_n \\ y_1 & \cdots & y_n \end{bmatrix}}_{\text{original data}}$

data along u axis

*Here I am showing data slightly off the u axis so you can see it, but really these points lie directly on the line.

* a quick note on what projection means in picture form:



We find the shortest distance from each data point to the line, and take that location on the line as the "projected" data point.

Okay, so we have: $\underset{1\times n}{[u_1 \cdots u_n]} = \underset{1\times 2}{[p_1\ p_2]} \underset{2\times n}{\begin{bmatrix} x_1 & \cdots & x_n \\ y_1 & \cdots & y_n \end{bmatrix}}$. We want the new data points $[u_1 \cdots u_n]$ to have maximum variance (think about this).

That means, we want to maximize the variance of $[p_1\ p_2]\begin{bmatrix} x_1 & \cdots & x_n \\ y_1 & \cdots & y_n \end{bmatrix}$, or pX. From class, recall that the variance of a vector $\vec{a}$ is

$\text{var} = \frac{1}{n}\sum_i a_i^2 = \frac{1}{n}a^T a$, if a is a column vector, or $\frac{1}{n}aa^T$ if a is a row vector.

Scanned by CamScanner

So, the variance of pX (also a vector) is $\frac{1}{n}(pX)(pX)^T$.

$$\Rightarrow \frac{1}{n}(pX)(pX)^T = \frac{1}{n} pXX^T p^T = pC_X p^T,$$ where C_X is the covariance matrix of X.

Now, what we want to do is maximize $pC_X p^T$. (Remember that this is just a number). Recall that because C_X is symmetric, we can decompose this into: $C_X = EDE^T$. Then,

$pC_X p^T = pEDE^T p^T$. Let $E = [e_1\ e_2]$ and $D = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$. Then, (where $\lambda_1 \geq \lambda_2$)

$$pEDE^T p^T = p[e_1\ e_2]\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}\begin{bmatrix} e_1^T \\ e_2^T \end{bmatrix} p^T$$

$$= p[e_1\ e_2]\begin{bmatrix} \lambda_1 e_1^T \\ \lambda_2 e_2^T \end{bmatrix} p^T$$

$$= \underbrace{p(\lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T)p^T}_{\text{This is still } pC_Xp^T\text{, the thing we want to maximize}} \leq p(\lambda_1 e_1 e_1^T + \lambda_1 e_2 e_2^T)p^T$$

Because $\lambda_1 \geq \lambda_2$, this expression (where $\lambda_2 = \lambda_1$) represents the maximum ~~[illegible]~~ of this function

So: $\lambda_1 p(e_1 e_1^T + e_2 e_2^T)p^T$ is the maximum value of pC_Xp^T. But this isn't super clear, so let's keep simplifying.

$$\Rightarrow \lambda_1 p(e_1 e_1^T + e_2 e_2^T)p^T = \lambda_1 p\left([e_1\ e_2]\begin{bmatrix} e_1^T \\ e_2^T \end{bmatrix}\right)p^T$$

$$= \lambda_1 p(EE^T)p^T$$

$$= \lambda_1 pp^T$$

Let's assume $pp^T = 1$ (that p is normalized such that its length is 1).

Now, λ_1 is the maximum value of pC_Xp^T, or the variance of our projected data. So, now the question is, for what value of p does $pC_Xp^T = \lambda_1$?

Let's return to the expression $p(\lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T)p^T$. Rewriting this

$\Rightarrow \lambda_1 p e_1 e_1^T p^T + \lambda_2 p e_2 e_2^T p^T$. If $p = e_1^T$, then:

$\Rightarrow \lambda_1 e_1^T e_1 e_1^T e_1 + \lambda_2 e_1^T e_2 e_2^T e_1$. Because $e_1^T e_1 = 1$ and $e_1^T e_2 = 0$, this equals λ_1!

So: $p = e_1^T$ maximizes the variance of our projected data, and e_1^T is the eigenvector with the largest eigenvalue of the covariance matrix of our data!