Lecture 5

Friday, 29 January 2021 15:48

How eigenvectors/values can be used to explain your data

If data is normally distributed, mean +- SD represents about 68% of your data

- However, variance only explains variance seen along a certain axis
- Covariance matrix can be used is variance is diagonal

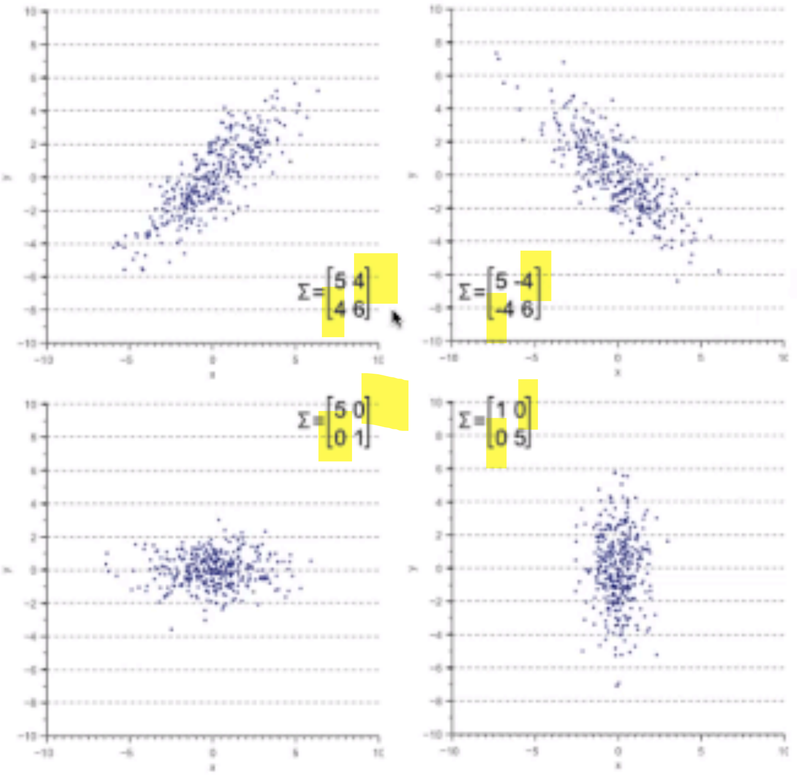


Figure 3. The covariance matrix defines the shape of the data. Diagonal spread is captured by the covariance, while axis-aligned spread is captured by the variance.

$$\sum = \left[\begin{matrix} Va \\ Co \end{matrix} \right.$$

+Ve correlation = +ve cov

-ve corr = -ve cov

No covariance = no corr (no variance on diagonal)

Top 2: eigenvalues are the same as the variances as no covariance

X and y aces are fine to explain variation

Bottom 2: eigenvectors (blue and pink lines) not the same as variances

- X and y axes aren't sufficient, eigenvectors as axes more effectively convey variation

$$\begin{matrix} ar(x) & Cov(x,y) \\ v(y,x) & Var(y) \end{matrix}\Bigg]$$

convey variation

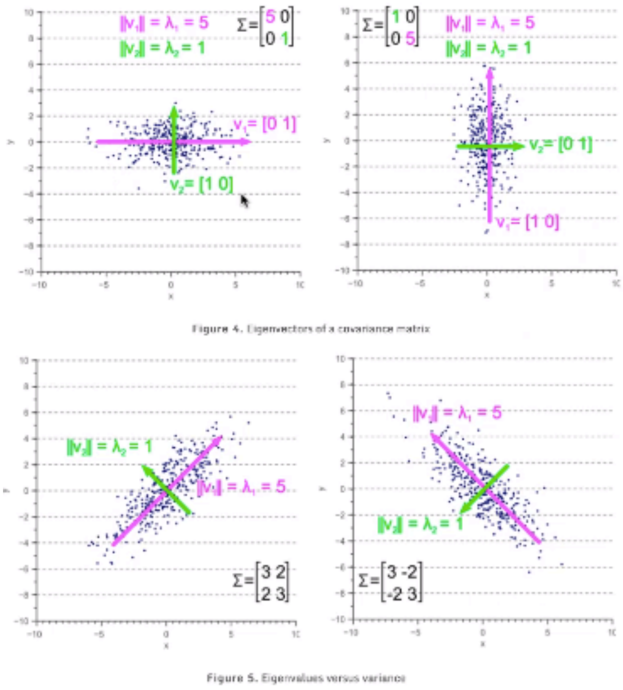


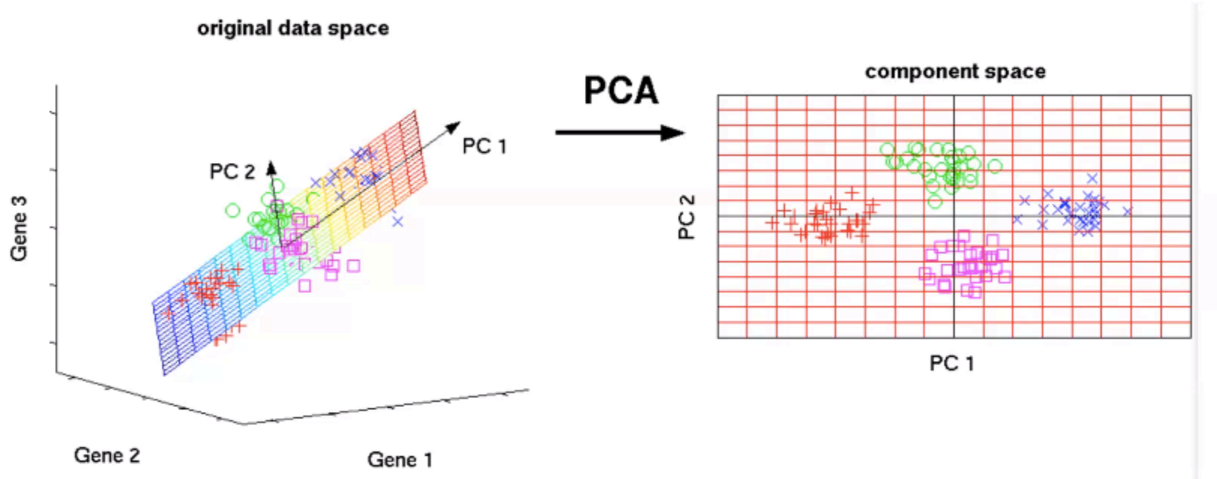
Figure 4. Eigenvectors of a covariance matrix

Figure 5. Eigenvalues versus variance

$$\sum = \begin{bmatrix} Var(x) & Cov(x,y) \\ Cov(y,x) & Var(y) \end{bmatrix}$$

$$\Sigma\vec{v} = \lambda\vec{v}$$

Principle Component Analysis



If have very large datasets, will often have many parameters which is hard to interpret

- Num params = num dimensions

Principle component analysis turns 3 dimensional graph on left into 2d on the right, aims to minimise dimension while preserving variability

PC1 crosses through multidimensional mean, and is the line that has the lowest RSS that captures the most variance
PC2 is the line that captures the second most variance

PC1 is eigenvector that gives most variance = highest eigenvector. PC2 =

PC1 is eigenvector that gives most variance = highest eigenvector, PC2 eigenvector with second highest = second highest number