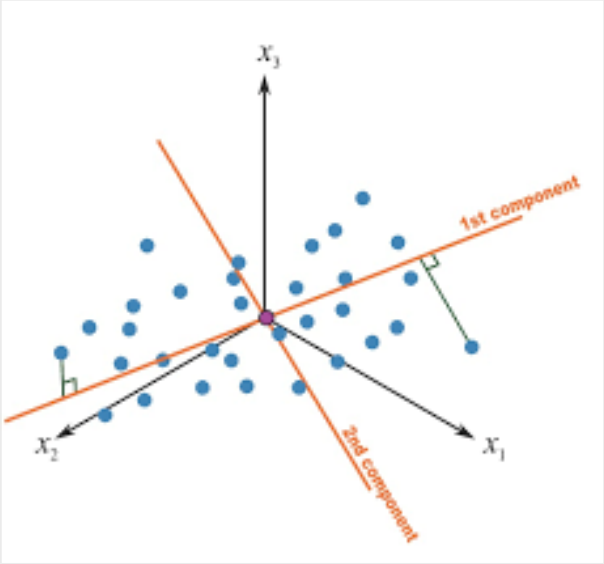
PCA - Principal Component Analysis

Idea principal
PCA es una técnica estádistica de aprendizaje no supervisado usada para examinar la interrelación entre un set de variables en orden de identificar estructuras subyacentes entre esas variables.
PCA también es conocido como analísis factorial.

Cuando la regresión entrena la mejor linea para ajustarse al set de datos, PCA determina diferentes lineas ortogonales que mejor se ajustan al dataset.

Ortogonal significa “ángulos rectos”, y estas líneas son perpendiculares entre si mismas en un espacio n-dimensional.
Un espacio n-dimensional es el espacio de muestreo de la variable, y hay tantas dimensiones como variables, por lo que en un dataset de 4 variables habrán 4 dimensiones.

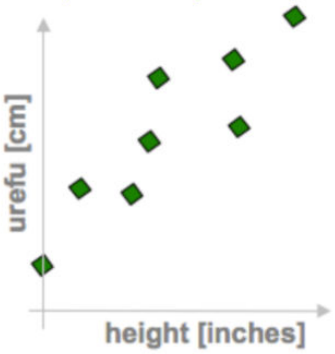


Ej de cálculo de lineas ortogonales

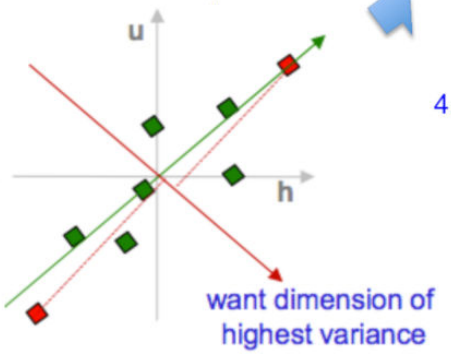
Más en profundidad:

PCA in a nutshell

1. correlated hi-d data
("urefu" means "height" in Swahili)



2. center the points



3. compute covariance matrix

$$\begin{array}{c} \\ h \\ u \end{array}\begin{array}{c} \begin{array}{cc} h & u \end{array} \\ \begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \end{array} \rightarrow \mathrm{cov}(h,u) = \frac{1}{n}\sum_{i=1}^{n} h_i u_i$$

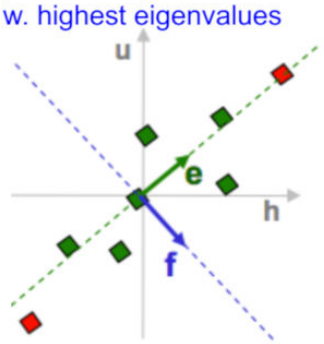
4. eigenvectors + eigenvalues

$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} e_h \\ e_u \end{pmatrix} = \lambda_e \begin{pmatrix} e_h \\ e_u \end{pmatrix}$$

$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} f_h \\ f_u \end{pmatrix} = \lambda_f \begin{pmatrix} f_h \\ f_u \end{pmatrix}$$

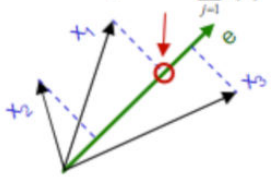
`eig(cov(data))`

5. pick m<d eigenvectors
w. highest eigenvalues



6. project data points to
those eigenvectors

$$x_e^{'} = x^T e = \sum_{j=1}^{d} x_{ij} e_j$$



7. uncorrelated low-d data