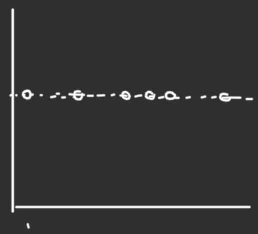
PCA

principal component analysis - extracts essential info by identifying useful dimensions in multi-dimensional data

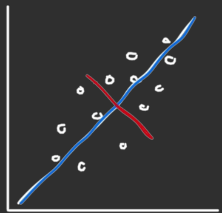
with 2D data, need 2n numbers to represent n data points.

σ^2 in y is 0.



* even though this is 2D we could represent the data just as effectively using n points due to no variance in y thus taking out the need for an "excess" dimension.

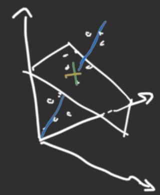
another scenario:



first principal comp.

second principal comp.

direction of the arrows are the principal components, since they are the directions of max variance and thus max information now in 3D



first
second
third
} max variance direc.

the point:

using those principal component directions we can create a new set of axes x', y', z'.

thus every point can be represented by (x', y', z'). we know that x has the most variance then y then z. thus we can get rid of z and store much of the same information in (x', y').

huge for sets with lots of dimensions (i.e. $1000 \to 20$).

how to calculate PCA

Data matrix: $\begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \\ z_1 & z_2 & \cdots & z_n \end{bmatrix}$

this is a 3D matrix
could be more columns

Calculate mean:

$$\mu_x = \frac{1}{n} \sum_{i}^{n} x_i$$

(μ_x, μ_y, μ_z) 3D

$$\mu_y = \frac{1}{n} \sum_{i}^{n} y_i$$

$(\mu_A, \mu_B, \ldots \mu_M)$ M Dimensional

$$\mu_z = \frac{1}{n} \sum_{i}^{n} z_i$$

To find the direction of maximum variance we can calculate the covariance matrix:

nxn matrix

$$C = MM^T$$

such that
$$M = \begin{bmatrix} (x_1-\mu_x) & (x_2-\mu_x) & \cdots & (x_n-\mu_x) \\ (y_1-\mu_y) & (y_2-\mu_y) & \cdots & (y_n-\mu_y) \\ \vdots & \vdots & \ddots & \end{bmatrix}$$

The diagonal of the covariance matrix are the variances along the X, Y, and Z axes. The off diagonal represent covariance between two dimensions (X/Y, Y/Z, X/Z).

$C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\quad$ $C = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ $\quad$ $C = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$

The eigenvectors of the covariance matrix are the principal components, in order of size.

eigenvalues & eigenvectors:

$$Av = \lambda v$$

matrix $\quad$ scalar

λ is the eigenvalue for the eigenvector, v

ex: consider a 3×3 matrix

$$A = \begin{bmatrix} 1.04 & 1.101 & 0.88 \\ 1.10 & 1.47 & 1.10 \\ 0.83 & 1.00 & 0.8931 \end{bmatrix}$$

now consider, v

$$v = \begin{bmatrix} -0.1233 \\ 0.6644 \\ -0.7371 \end{bmatrix}$$

notice that $$\underset{\uparrow \atop A}{A}v = \begin{bmatrix} -0.0051 \\ 0.0274 \\ -0.0304 \end{bmatrix} = \underset{\uparrow \atop \lambda}{0.0412} \begin{bmatrix} -0.01233 \\ 0.6644 \\ -0.7371 \end{bmatrix}$$

λ is the eigenvalue, and v is the eigen vector

throwbacks to multi.