Repasemos:

Dados unos datos en $\mathbb{R}$

$$A = \begin{pmatrix} - a_1 - \\ \cdots \\ - a_m - \end{pmatrix} \in \mathbb{R}^{m \times n}$$

En cada fila a_i tenemos n características de m individuos.

Calculamos $\bar{a} = \frac{1}{m} \sum_{i=1}^{m} a_i$

donde $a_i \in \mathbb{K}^n$ es un vector $\Rightarrow$

$\bar{a} \in \mathbb{K}^n$ también es un vector

$$(\bar{a})_j = \frac{1}{m} \sum (a_i)_j \rightarrow$$ promedio de las n caracter. de los indiv.

coord j de a_i es la característica j del individuo a_i

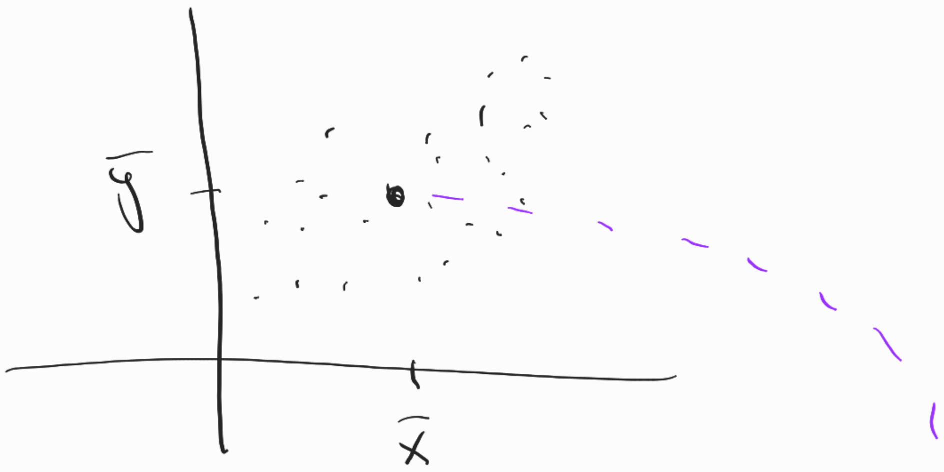
Si ponemos 2 características y m individuos, cada $a_i = (x_i, y_i)$

$$A = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \\ x_m & y_m \end{pmatrix} \rightarrow \bar{a} = \frac{1}{m}\sum_{i=1}^{m}(x_i, y_i)$$

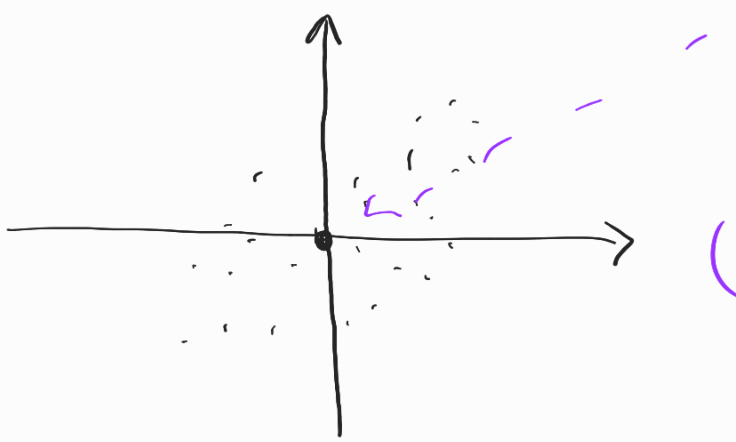
$$= \Bigg(\underbrace{\frac{1}{m}\sum_{i=1}^{m} x_i}, \underbrace{\frac{1}{m}\sum_{i=1}^{m} y_i}\Bigg)$$

$$\bar{x} = (\bar{a})_1 \qquad \bar{y} = (\bar{a})_2$$

Gráficamente



Si restamos cada $(x_i, y_i) - (\bar{x}, \bar{y})$ tendremos los datos "centrados"



se corre todo dejando $(\bar{x}, \bar{y})$ en el origen $(0,0)$

En general vemos que ahora

$$\frac{1}{m}\sum_{i=1}^{m}(a_i - \bar{a}) = \frac{1}{m}\left(\sum_{i=1}^{m} a_i - m\bar{a}\right)$$

$$= \frac{1}{m}\sum_{i=1}^{m} a_i - \bar{a} = 0.$$

Llamamos $b_i = a_i - \bar{a}$ y armamos

$$B = \begin{pmatrix} \underline{\quad b_1 \quad} \\ \vdots \\ \overline{\quad b_m \quad} \end{pmatrix} = \begin{pmatrix} \underline{\quad a_1 - \bar{a} \quad} \\ \vdots \\ \overline{\quad a_m - \bar{a} \quad} \end{pmatrix}$$

la matriz de los datos centrados.

Si miramos por características (en las columnas) podemos también escribir

$$B = \left(c_1 \,\middle|\, \cdots \,\middle|\, c_n \right)$$

Estudiamos

$$B^t \cdot B = \begin{pmatrix} \overline{\quad c_1 \quad} \\ \vdots \\ \overline{\quad c_n \quad} \end{pmatrix} \left(c_1 \middle| \ \cdots \ \middle| c_n \right)$$

$$= \begin{pmatrix} \langle c_1, c_1 \rangle & \langle c_1, c_2 \rangle & \cdots & \langle c_1, c_n \rangle \\ \langle c_2, c_1 \rangle & \langle c_2, c_2 \rangle & \cdots & \langle c_2, c_n \rangle \\ \vdots & & & \\ \langle c_n, c_1 \rangle & & \cdots & \langle c_n, c_n \rangle \end{pmatrix}$$

Volvamos al ej de 2 caract.

$$B^t B = \begin{pmatrix} \langle c_1, c_1 \rangle & \langle c_1, c_2 \rangle \\ \langle c_2, c_1 \rangle & \langle c_2, c_2 \rangle \end{pmatrix}$$

donde $c_1 = \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix}$ $c_2 = \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}$

$$\Rightarrow \langle c_1, c_1 \rangle = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\langle c_2, c_2 \rangle = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$\langle c_1, c_2 \rangle = \langle c_2, c_1 \rangle = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$B^t B$ se la llama matriz de covarianza.

también se usa $\frac{1}{n} B^t B$ $\left(ó \ \frac{1}{n-1} B^t B \right)$

porque de esta manera se tiene efectivamente varianza y covarianza en sus entradas.

Notar que en este caso se reescalan todos los valores singulares con $\frac{1}{n}$ y los autovectores de $B^t B$ son los mismos.

A partir de aquí se pueden calcular las componentes ppales como hicimos la clase pasada que nos dará U, V, Σ:

$B\,V = U\,\Sigma$ de donde

$V = (v_1 | \dots | v_n)$ con $\{v_1, \dots, v_n\}$ BON
y autovectores de $B^t B$

$\Rightarrow$

$B \cdot v_1 = \sigma_1 u_1$ de manera que si

$v_1 = (v_1^1, \dots, v_1^n)$ son sus coordenadas

obtenemos que

$$\sigma_1\, u_1 = v_1^1 C_1(B) + \dots + v_1^n C_n(B)$$

$$v_1^1 \cdot \underbrace{C_1(B)}_{\substack{\text{col 1 de A} \\ \text{menos } (\bar{a})_1}} + \dots + v_1^n \underbrace{C_n(B)}_{\substack{\text{col n de A} \\ \text{menos } (\bar{a})_n}}$$

En el ej de dim 2 , $B^t B \in \mathbb{R}^{2\times 2}$

$$\Rightarrow \sigma_1 u_1 = \left[v_1^1 \begin{pmatrix} x_1-\bar{x} \\ \vdots \\ x_n-\bar{x} \end{pmatrix} + v_1^2 \begin{pmatrix} y_1-\bar{y} \\ \vdots \\ y_n-\bar{y} \end{pmatrix} \right]$$

De aquí podemos ver qué jerarquía tiene cada característica centrada (columnas de A). Si $v_1^1 >> v_1^2$ vemos que la componente principal depende más de la coord x es decir de la primera característica que sería la variable que mejor explica la variabilidad conjunta de los datos.

En algunos conjuntos de datos puede haber variables que tengan ordenes de magnitud muy diferentes que otras. Por ej. si tengo datos de PBI de países junto con datos de porcentaje de mujeres (número entre 0 y 1).

En estos casos la varianza en la variable con valores altos será también mayor y por lo tanto se espera que la 1ra componente ppal sea cercana a esta dirección.

Para evitar este fenómeno, en estos casos se "normalizan" los datos dividiendo los datos centrados por el desvío standard de cada característica.

En el ej de dim 2 tenemos

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \underset{\text{centramos}}{\rightsquigarrow} \begin{pmatrix} x_1-\bar{x} \\ \vdots \\ x_n-\bar{x} \end{pmatrix} \underset{\substack{\text{Dividimos} \\ \text{x desvío}}}{\rightsquigarrow} \begin{pmatrix} \frac{x_1-\bar{x}}{s_1} \\ \vdots \\ \frac{x_n-\bar{x}}{s_1} \end{pmatrix}$$

$$\bar{x} = \frac{1}{n}\sum x_i \qquad s_1 = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{1/2}$$

Y lo mismo para la 2da característica $\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

Luego consideramos

$$B = \begin{pmatrix} \frac{x_1-\bar{x}}{s_1} & \frac{y_1-\bar{y}}{s_2} \\ \vdots & \vdots \\ \frac{x_n-\bar{x}}{s_1} & \frac{y_n-\bar{y}}{s_2} \end{pmatrix}$$

Luego se trabaja a partir de acá con esta B.

En este caso, notar que

$$\frac{1}{n} B^t \cdot B = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{s_1^2} & \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{s_1}\frac{(y_i-\bar{y})}{s_2} \\ \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{s_1}\frac{(y_i-\bar{y})}{s_2} & \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{s_2^2} \end{pmatrix}$$

y como $\frac{1}{n}\sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{s_1^2} = \frac{1}{s_1^2}\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}_{s_1^2} = 1$

y lo mismo para

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{s_2^2} = 1$$

Los elementos de la diagonal de

$\frac{1}{n}B^tB$ son 1's.

Más generalmente, si llamamos

$$B = \left(\, w_1 \;\middle|\; w_2 \,\right) \quad \text{con}$$

$$w_1 = \begin{pmatrix} \frac{x_1-\bar{x}}{s_1} \\ \vdots \\ \frac{x_n-\bar{x}}{s_1} \end{pmatrix} \quad y \quad w_2 = \begin{pmatrix} \frac{y_1-\bar{y}}{s_2} \\ \vdots \\ \frac{y_n-\bar{y}}{s_2} \end{pmatrix}$$

$$\frac{1}{n}B^tB = \begin{pmatrix} \frac{1}{n}\langle w_1, w_1\rangle & \frac{1}{n}\langle w_1, w_2\rangle \\ \frac{1}{n}\langle w_2, w_1\rangle & \frac{1}{n}\langle w_2, w_2\rangle \end{pmatrix}$$

donde $\frac{1}{n}\langle w_1, w_1\rangle = \frac{1}{n} \left\| \frac{x-\bar{x}}{s_1} \right\|_2^2$

$$= \frac{1}{n} \cdot \frac{1}{s_1^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = 1$$ como ya

vimos. De la misma manera

$$\frac{1}{n} \langle w_2, w_2 \rangle = 1.$$

Esto nos dice que $\|w_1\|^2 = \langle w_1, w_1\rangle = n$

$$\Rightarrow \|w_1\|_2 = \sqrt{n}$$

De la misma manera $\|w_2\|_2 = \sqrt{n}$.

Además

$$\left|\frac{1}{n}\langle w_1, w_2\rangle\right| = \frac{1}{n} \left|\langle w_1, w_2\rangle\right| \leq \frac{1}{n} \|w_1\|_2 \|w_2\|_2 = 1$$

Todo esto nos dice que en este caso:

$$\frac{1}{n} B^t B = \begin{pmatrix} 1 & \text{nro entre } -1 \text{ y } 1 \\ \text{nro entre } -1 \text{ y } 1 & 1 \end{pmatrix}.$$