Capítulos 1 y 2: resumen

- Algebra básica de Matrices y vectores (lectura individual) -
- Los elementos más importantes a recordar son:

1) Concepto de distancia estadística:

La distancia Euclidiana entre los puntos $P = (x_1, x_2, \cdots x_p)$ y $Q = (y_1, y_2, \cdots y_p)$

$$d(P, Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

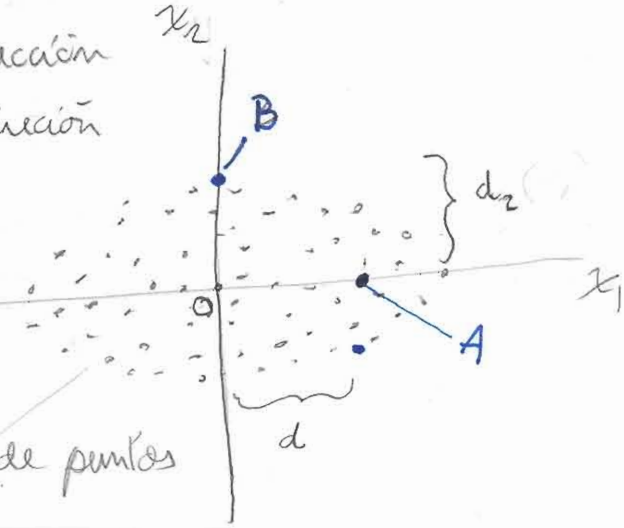
es insatisfactoria en análisis estadístico porque:

i) cada coordenada contribuye igualmente al cálculo de la distancia

ii) cuando las coordenadas representan mediciones provenientes de fluctuaciones aleatorias es deseable darle menor "peso" a las coordenadas que están sujetas a variaciones mayores. Para entender porqué, consideren este ejemplo en 2 dimensiones:

- $d(O,B)$ es más "inusual" en la dirección del eje x_2 que $d(O,A)$ en la dirección x_1, aún cuando

$d(O,B) = d(O,A)$



Dicho de otro modo, $d(O,B)$ es más inusual que $d(O,A)$ porque $\sqrt{s_{11}} > \sqrt{s_{22}}$, dónde s_{ii} es la varianza de la muestra en dirección del eje 'i'.

La distancia estadística pesa entonces cada coordenada por la varianza de la muestra en cada eje. Si $P = (x_1, x_2)$ es un punto, entonces $d(O,P) = \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}}}$ (*) ec.

iii) Sólo si la variabilidad en cada eje es igual, la dist. euclídiana es adecuada.

iv) Si las mediciones en el eje x_1 no son indep. de las med. en x_2 entonces podemos rotar los ejes así

entonces $d(O,P) =$

$$\sqrt{\frac{\tilde{x}_1^2}{\tilde{s}_{11}} + \frac{\tilde{x}_2^2}{\tilde{s}_{22}}}, \quad \text{donde} \begin{cases} \tilde{x}_1 = x_1 \cos(\theta) + x_2 \sin(\theta) \\ \tilde{x}_2 = -x_1 \sin(\theta) + x_2 \cos(\theta) \end{cases}$$

Después de sustituir y un poco de algebra; la distancia entre $P = (\tilde{x}_1, \tilde{x}_2)$ y $O = (0, 0)$ se re-escribe así:

$$d(O,P) = \sqrt{\underbrace{a_{11}x_1^2}_{1} + \underbrace{2a_{12}x_1x_2}_{2} + \underbrace{a_{22}x_2^2}_{3}} \quad \text{ec.} (**)$$

dónde
$$\left.\begin{aligned} a_{11} &= f_1(s_{11}, s_{12}, s_{22}, \theta) \\ a_{12} &= f_2(s_{11}, s_{12}, s_{22}, \theta) \\ a_{22} &= f_3(s_{11}, s_{12}, s_{22}, \theta) \end{aligned}\right\} \text{ ver p. 35 cap. 1}$$

Lo importante es que el término 2 que da cuenta de la correlación no-nula r_{12}

- $(**)$ es una generalización de $(*)$
- $P = (x_1, x_2)$, $Q = (y_1, y_2) \Rightarrow d(P,Q) = \sqrt{\begin{pmatrix} a_{11}(x_1-y_1)^2 + 2a_{12}(x_1-y_1)(x_2-y_2) \\ + a_{22}(x_2-y_2)^2 \end{pmatrix}}$
- Todos los puntos $P = (x_1, x_2)$ que están a una dist. al cuadrado cste. c^2 de Q satisfacen:

ec: $(***)$ $\quad a_{11}(x_1-y_1)^2 + 2a_{12}(x_1-y_1)(x_2-y_2) + a_{22}(x_2-y_2)^2 = c^2$

esta es la ec. de una elipse centrada en Q. En p dimensiones

$$d(O,P) = \sqrt{a_{11}x_1^2 + a_{22}x_2^2 + \dots + a_{pp}x_p^2 + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + \dots + 2a_{p-1,p}x_{p-1}x_p}$$

ec. $(\boxtimes)$

y además

$$d^2(O,P) = a_{11}x_1^2 + a_{22}x_2^2 + \dots + a_{pp}x_p^2 + 2(a_{12}x_1x_2 + a_{13}x_1x_3 + \dots + a_{p-1,p}x_{p-1}x_p)$$

si $d^2(O,P) > 0 \;\; \forall\, [x_1, x_2, \dots, x_p] \neq [0, 0, \dots 0]$.

Además, si $a_{ij} = a_{ji} \;\; \forall\, i,j = 1,2,\dots p$ entonces

$$0 < d^2(O,P) = [x_1 \;\; x_2 \;\; \cdots \;\; x_p] \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

$$= \underline{x}' A \underline{x} \quad \text{para } \underline{X} \neq 0,$$

donde A es una matriz positiva definida. Entonces una forma cuadrática positiva definida puede ser interpretada como una distancia (estadíst.) al cuadrado. !

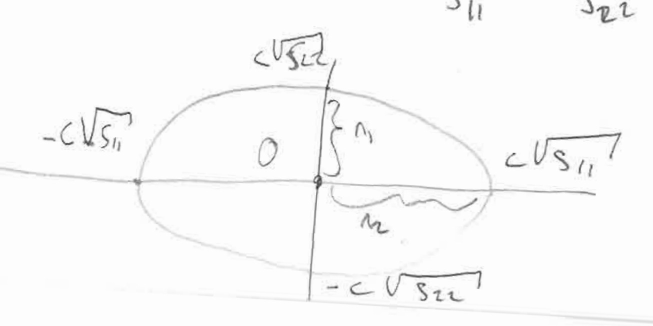
Interpretación Geométrica para $p=2$:

los puntos $\underline{x}' = [x_1 \;\; x_2]$ a distancia c est. del origen satisfacen:

$$\underline{X}' A \underline{X} = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 = c^2.$$ Cual es el tamaño de los ejes r_1 y r_2?

comentario:

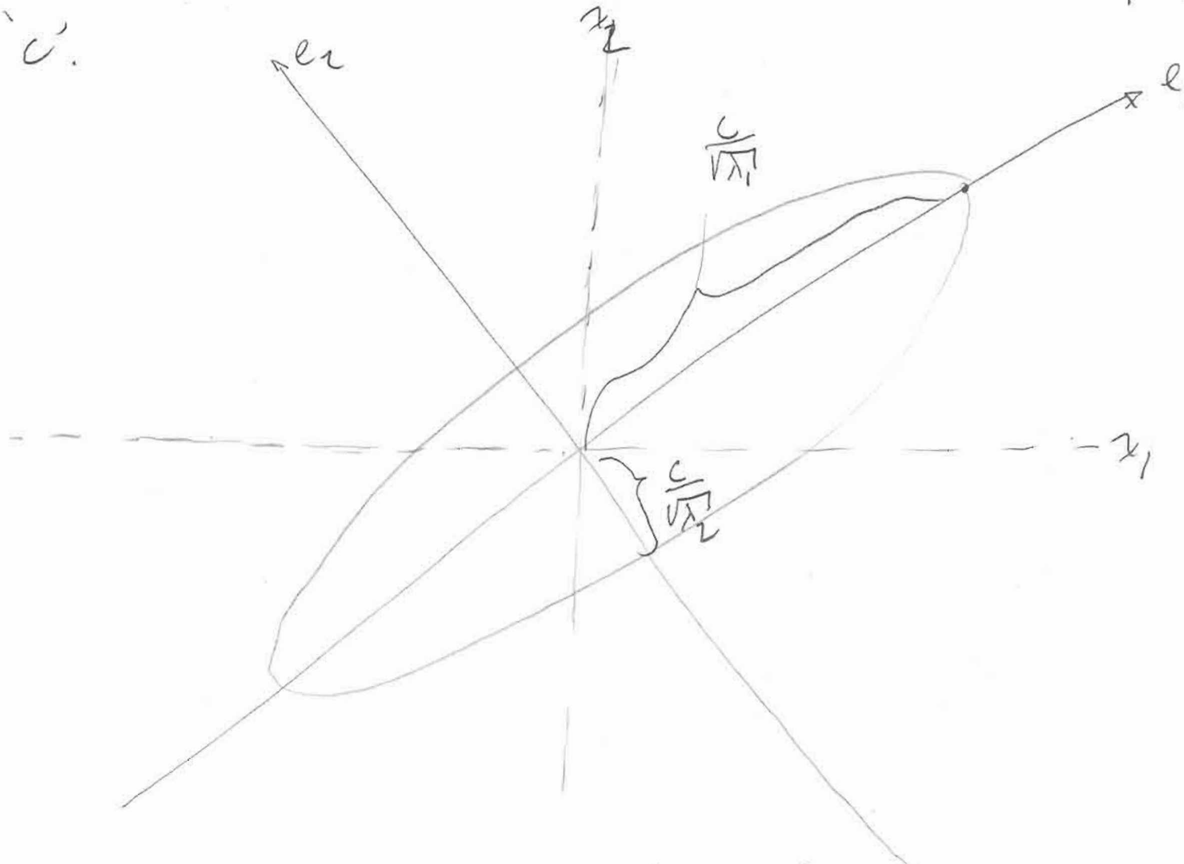
$$\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} = c$$ es la ec. de una elipse centrada en el origen



Usando una descomposición espectral:

$$A = \lambda_1 \underset{\sim}{e}_1 \underset{\sim}{e}_1' + \lambda_2 \underset{\sim}{e}_2 \underset{\sim}{e}_2' \quad y \quad \underset{\sim}{x}' A \underset{\sim}{x} = \lambda_1 \underbrace{(x' \underset{\sim}{e}_1)^2}_{=y_1^2} + \lambda_2 \underbrace{(x' e_2)^2}_{=y_2^2}$$

entonces $R_1 = c \cdot \lambda_1^{-1/2}$, $R_2 = c \cdot \lambda_2^{-1/2}$ y los puntos situados a distancia 'c' del origen yacen en el perímetro de una elipse cuyos ejes están dados por los vectores propios de A con longitud proporcional a $\lambda_1^{-1/2}$ y $\lambda_2^{-1/2}$. La constante de proporcionalidad es 'c'.



para $p > 2$, $\quad c^2 = \lambda_1 (x' e_1)^2 + \dots + \lambda_p (x' e_p)^2$ dibuja una hiper elipsoide //

2) Vectores Media y Matrices de Varianzas y Co-varianzas.

Defs.

i) Vector aleatorio: un vector cuyos elementos son variables aleatorias

ii) Matriz aleatoria: una matriz cuyos elementos " " "

- Si $X = \{X_{ij}\}$ es una matriz $n \times p$, entonces $E(X) = \{E(X_{ij})\}_{n \times p}$
- El comportamiento colectivo de las variables aleatorias $X_1, X_2, \dots X_p$ en $\underline{X}' = [X_1, X_2, \dots, X_p]$ es descrito por una distribución conjunta $f(x_1, x_2, \dots, x_p) = f(\underline{X})$
- Indep. Si $f_{1,2,\dots,p}(x_1, x_2, \dots, x_p) = f_1(x_1) \cdot f_2(x_2) \cdots f_p(x_p) \quad \forall [x_1, x_2 \dots, x_p]$, entonces las variables aleatorias $X_1, X_2, \dots, X_p$ son mutuamente independientes.
- Si X_i y X_k son indep., entonces $Cov(X_i, X_k) = 0$

Sean $\mu_i = \mathbb{E}(X_i)$, $i = 1, 2, \dots, p$, y

- $\mathbb{E}(\underline{X}) = \begin{bmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} = \underline{\mu}$

- $\Sigma = \mathbb{E}\left[(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})'\right] = \mathbb{E}\left\{ \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_p - \mu_p \end{bmatrix} \left[X_1 - \mu_1, X_2 - \mu_2, \dots, X_p - \mu_p \right] \right\}$

$$= \mathbb{E}\left\{ \begin{bmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \cdots & (X_1-\mu_1)(X_p-\mu_p) \\ (X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & \cdots & (X_2-\mu_2)(X_p-\mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ (X_p-\mu_p)(X_1-\mu_1) & (X_p-\mu_p)(X_2-\mu_2) & \cdots & (X_p-\mu_p)^2 \end{bmatrix} \right\}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \sigma_{31} & \vdots & \ddots & \vdots \\ \vdots & & & \sigma_{pp} \\ \sigma_{p1} & \sigma_{p2} & \cdots & \end{bmatrix} = \mathrm{Cov}(X) = \Sigma$$

Nos referimos a $\underline{\mu}$ y a Σ como <u>la media</u> poblacional y la matriz de varianza - covarianza de la población.

La matriz de correlaciones en la población se define como:

$$\rho = \begin{bmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{22}}} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} & \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{bmatrix}$$

Si

$$\mathbb{V}^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & & 0 \\ & \sqrt{\sigma_{22}} & \\ 0 & & \ddots \\ & & & \sqrt{\sigma_{pp}} \end{bmatrix}$$

entonces (Chequear)

$$\mathbb{V}^{1/2} \rho \mathbb{V}^{1/2} = \Sigma \quad \text{y} \quad \rho = (\mathbb{V}^{1/2})^{-1} \Sigma (\mathbb{V}^{1/2})^{-1}$$

Estos resultados son válidos también cuando las cantidades poblacionales son reemplazadas por sus equivalentes en una muestra: Si $\underline{\bar{x}}' = [\bar{x}_1, \bar{x}_2, \ldots \bar{x}_p]$ es el vector de medias de una muestra de 'n' observaciones en 'p' variables aleatorias $X_1, X_2, \ldots X_p$, entonces la matriz de var-cov. de la muestra es:

$$S_n = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{1p} & \cdots & s_{pp} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{n}\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)^2 & \cdots & \frac{1}{n}\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)(x_{jp}-\bar{x}_p) \\ \vdots & \ddots & \vdots \\ \frac{1}{n}\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)(x_{jp}-\bar{x}_p) & \cdots & \frac{1}{n}\sum_{j=1}^{n}(x_{jp}-\bar{x}_p)^2 \end{bmatrix}$$