Varianza muestral generalizada y Normal multivariada

martes, 8 de febrero de 2022 9:06 a.m.

$$S_n = \frac{1}{n}\sum_j (x_{ji} - \bar{x}_i)^2$$

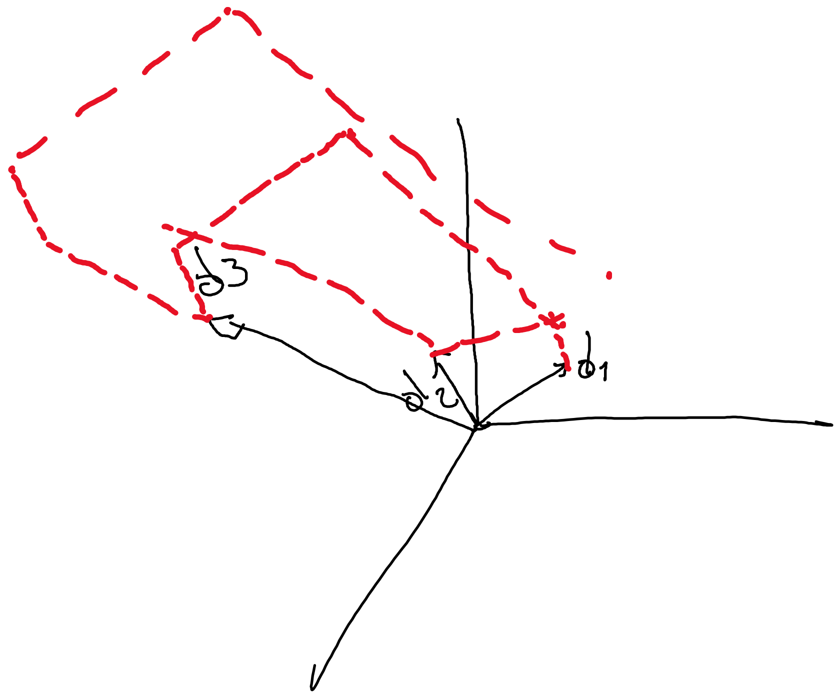
$$S = \frac{1}{n-1}\sum_j (x_{ij} - \bar{x}_i)^2$$

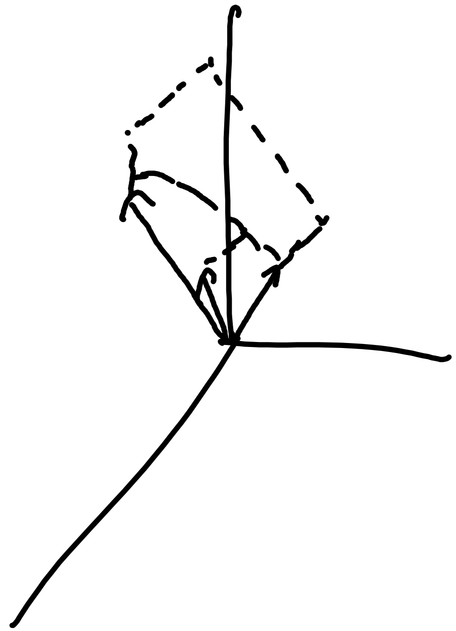
$$S = \begin{bmatrix} \ \end{bmatrix}$$

Matriz de varianzas y covarianzas muestral

Insesgada

Varianza muestral generalizada $= |S|$





La varianza generalizada es cero cuando y solo cuando al menos un vector desviación yace en el hiperplano formado por todas las combinaciones lineales de los restantes vectores. Es decir cuando al menos una columna de la matriz S es linealmente dependiente de otra.

$n \leq p$ $\qquad X = n\left\{ \left[\quad \right] \right. \qquad S = \left[\quad \right]$

Varianza generalizada determinada por $|R|$

- La varianza muestral generalizada se ve afectada indebidamente por la variabilidad de las mediciones en una sola variable.

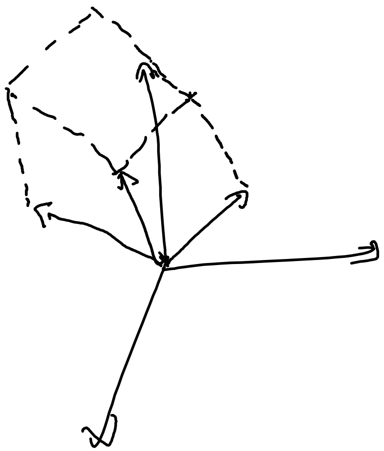
- Por ejemplo, si s_{ii} es muy grande (o pequeño). Geométricamente el vector de desviación correspondiente $d_i = (y_i - \bar{x}_i \mathbf{1})$ será muy largo (o muy corto) y afectará el volumen.

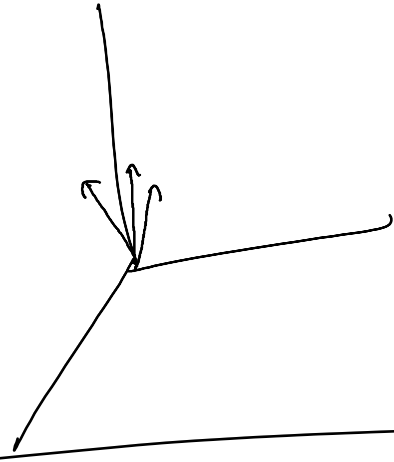
- Entonces, es útil escalar todos los vectores de desviación para que tengan la misma longitud.

- Equivale a reemplazar la observación original x_{ji} por $(x_{ji} - \bar{X}_i)/\sqrt{S_{ii}}$

Entonces, la matriz de varianza y covarianz es $R \to$ La matriz de correlación

$|R|$ = Varianza muestral generalizada de las variables estandarizadas





Varianza total de la muestra $= S_{11} + S_{22} + \dots + S_{pp}$

Normal multivariada

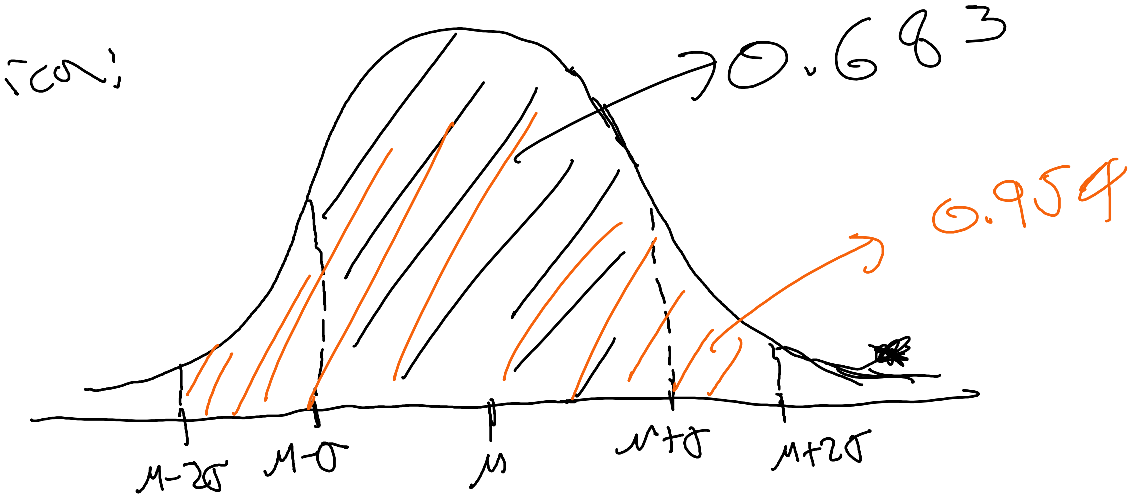
Recordemos:

La PDF de una v.a. normal con media μ y var σ^2 es:

$(x-\mu)/_{\sigma^2}$ $\qquad x \in \mathbb{R}$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \qquad x \in \mathbb{R}$$

Ley empírica:



Denotamos la distr. normal con media μ y var $\sigma^2 \longrightarrow N(\mu, \sigma^2)$

X es normal con media μ y var σ^2

$$X \sim N(\mu, \sigma^2)$$

$$\frac{(X-\mu)^2}{\sigma^2} \longrightarrow$$ Distancia estadística entre X y μ

$$= (X-\mu)(\sigma^2)^{-1}(X-\mu)$$

Generalizando $\vec{x}$ vector de observaciones de $\mathbf{X}$ $\rightarrow$ una fila de la matriz X

distancia estadística $\mathbf{x}$ a μ

$$(\mathbf{x}-\boldsymbol{\mu})' \underline{\underline{\Sigma}}^{-1} (\mathbf{x}-\boldsymbol{\mu})$$

Entonces podemos generalizar y obtener la PDF de la normal multivariada.

$$f(\mathbf{x}) = \frac{1}{\quad} \cdot e^{-(\mathbf{x}-\boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})/2}$$

$$\underbrace{(2\pi)^{p/2} |\Sigma|^{1/2}}$$

$$\mathbb{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$$

de un vector aleatorio $\mathbb{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$ → v.a.

Denotamos la normal multivariada como

vector → matriz.

$$\mathbb{X} \sim N_p(\mu, \Sigma)$$

Obs:

Si $p=2$ y $\rho_{12}=0$ $\quad (cov(X_1, X_2)=0)$

$$\Downarrow$$

X_1, X_2 indep.

Teorema:

Si Σ def positiva (Σ^{-1} existe) entonces si e es un vector propio de Σ con valor propio asociado λ, e es un vect propio de Σ^{-1} con val propio asociado $\frac{1}{\lambda}$. Además

Σ^{-1} def positiva.

chi-cuadrada

Obs:

1) $(X-\mu)' \Sigma^{-1} (X-\mu) \leq \chi^2_p(\alpha)$

g.l

con probabilidad $1-\alpha$ $y^{\cup}$.

Proposición 1)

Si $\mathbb{X}$ vector aleatorio normal multivariado.

$\mathbb{X} \sim N_p(\mu, \Sigma)$ Entonces

Si $a \in \mathbb{R}^p$, $a'\mathbb{X} = a_1 X_1 + a_2 X_2 + \dots + a_p X_p$

$\sim N(a'\mu, a'\Sigma a)$

Prop 2) Si $a'\mathbb{X} \sim N(a'\mu, a'\Sigma a)$, $\forall a \in \mathbb{R}^p$

entonces $\mathbb{X} \sim N_p(\mu, \Sigma)$

Prop 3) $\mathbb{X} \sim N_p(\mu, \Sigma)$

Sea $\underset{q\times p}{A} \in \mathbb{R}^{q\times p}$

$$A\mathbb{X} \sim N_q(A\mu, A\Sigma A')$$

Prop 4) $\mathbb{X} \sim N_p(\mu, \Sigma)$ $\quad d \in \mathbb{R}^p$

$$\mathbb{X} + d \sim N_p(\mu + d, \Sigma)$$

Prop 5) Todas las particiones de vect. aleatorios normales resultan en vect. aleat normales.

$$\mathbb{X} \sim N_p(\mu, \Sigma) \qquad \mathbb{X} = \begin{pmatrix} \mathbb{X}_{(1)} \\ \hline \mathbb{X}_{(2)} \end{pmatrix} \begin{matrix} \}\, q \\ \}\, p-q \end{matrix} \qquad \overset{d}{\mu} = \begin{pmatrix} \mu_{(1)} \\ \hline \mu_{(2)} \end{pmatrix}$$

$$\sim (\mu, \Sigma) \qquad \Sigma = (\Sigma \;\; \Sigma)$$

$$X_{(1)} \sim N_q\left(\mu_{(1)}, \Sigma_{11}\right) \qquad = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right)$$

$$X_{(2)} \sim N_{p-q}\left(\mu_{(2)}, \Sigma_{22}\right)$$

Prop 6) 1) Si $\underset{q_1\times 1}{X_{(1)}}$, $\underset{q_2\times 1}{X_{(2)}}$ normales multivariados e independ.

entonces $\underbrace{\operatorname{cov}\left(X_{(1)}, X_{(2)}\right)}_{\substack{\text{matriz} \\ q_1 \times q_2}} = 0 \longrightarrow$ Matriz 0

2) Si $X = \left(\begin{array}{c} X_{(1)} \\ \hline X_{(2)} \end{array}\right) \sim N_{q_1+q_2}\left(\left(\begin{array}{c} \mu_{(1)} \\ \hline \mu_{(2)} \end{array}\right), \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right)\right)$

$X_{(1)}$ y $X_{(2)}$ indep si y solo si $\Sigma_{12} = 0$

Covarianza 0 SÍ implica independencia

Si HAY normalidad,

3) $\underset{\sim}{X}_{(1)}^{q_1 \times 1}$, $\underset{\sim}{X}_{(2)}^{q_2 \times 1}$ normales indep si y solo si

$$\underset{\sim}{X} = \begin{pmatrix} \underset{\sim}{X}_{(1)} \\ \hline \underset{\sim}{X}_{(2)} \end{pmatrix} \sim N_{q_1+q_2}\left(\begin{pmatrix} \mu_{(1)} \\ \hline \mu_{(2)} \end{pmatrix}, \left(\begin{array}{c|c} \Sigma_{11} & 0 \\ \hline 0 & \Sigma_{22} \end{array} \right) \right)$$

Prop 7) Sea $\underset{\sim}{X} \sim N_p(\mu, \Sigma)$, $|\Sigma| > 0$

Entonces

1) $(\underset{\sim}{X} - \mu)' (\Sigma^{-1}) (\underset{\sim}{X} - \mu) \sim \chi_p^2$

la distancia de los valores a la media tiene distrib χ^2

La probabilidad del evento.

$$\left\{ (X-\mu)'(\Sigma^{-1})(X-\mu) \leq \chi^2_p(\alpha) \right\} \text{ es } 1-\alpha.$$

Muestreo de la normal multivariada.

Suponga que $X_1, \ldots X_n$ son una muestra aleatoria de una población normal multivariada con media μ y cov Σ

Densidad conjunta de $X_1 \ldots X_n$

$$\prod_{j=1}^{n} \left(\frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}\left[(X_j-\mu)' \Sigma^{-1} (X_j-\mu)\right]} \right)$$

función de verosimilitud (función de μ y Σ)

Dadas las observaciones $X_1 \ldots X_n$: $L(\mu, \Sigma)$

Método de máximo-verosimilitud:

Utilizar como estimaciones de parámetros poblacionales desconocidos los valores que maximizan $L(\mu, \Sigma)$

Los valores que "mejor" explican los datos

Teo: Sean $X_1 \ldots X_n$ una muestra aleatoria de una población normal con media μ y cov. Σ.

Entonces los MLE son:

$$\hat{\mu} = \bar{X} \qquad \hat{\Sigma} = \frac{1}{n}\left(\sum_{j=1}^{n}(X_j - \bar{X})(X_j - \bar{X})'\right)$$

$$= S_n = \frac{n-1}{n} S$$

Obs: 1) $\hat{\mu}$, $\hat{\Sigma}$ son VECTORES ALEATORIOS antes de tomar los datos

2) $$L(\hat{\mu}, \hat{\Sigma}) = \frac{1}{(2\pi)^{np/2}} e^{-np/2} \times \frac{1}{\underbrace{|\hat{\Sigma}|}_{\frac{n-1}{n}S}^{n/2}}$$

$$cte\,(\underbrace{|S|})^{-n/2}$$

la varianza generalizada determina el

máximo de la función de verosimilitud.