

# Componentes Principales

**Graciela Boente**

# Motivación

Sea  $\mathbf{x} \in \mathbb{R}^p$  tal que

$$\mathbb{E}(\mathbf{x}) = \mathbf{0} \quad \text{VAR}(\mathbf{x}) = \mathbf{\Sigma}$$

El método de componentes principales busca elegir  $q$  combinaciones lineales

$$z_1 = \gamma_1^T \mathbf{x}, \quad z_2 = \gamma_2^T \mathbf{x}, \quad \dots \quad z_q = \gamma_q^T \mathbf{x}$$

de modo tal que si  $\mathbf{z} = (z_1, \dots, z_q)$  entonces,  $\mathbf{z}$  explica una porción razonable de la dispersión total medida a través de  $\text{traza}(\mathbf{\Sigma})$ .

Como ejemplo, tomemos las mediciones del tamaño de la cabeza del primer y segundo hijo que vieron en la práctica.

## Ejemplo 1

Teníamos como medidas

- $x_1$  = longitud de la cabeza,
- $x_2$  = ancho de la cabeza

medidas sobre el primer y segundo hijo de 25 familias, dando origen a observaciones  $\mathbf{x}_{ij}$ ,  $1 \leq j \leq 25$ ,  $i = 1, 2$  donde  $i = 1$  indica las observaciones correspondientes al hijo mayor e  $i = 2$  aquellas correspondientes al segundo hijo.

$$\bar{\mathbf{x}}_1 = \begin{pmatrix} 187.40 \\ 151.12 \end{pmatrix} \quad \bar{\mathbf{x}}_2 = \begin{pmatrix} 183.32 \\ 149.36 \end{pmatrix}$$

## Ejemplo 1

Tomemos

$$\hat{\gamma}_1 = (0.8346, 0.5509)^T$$

Definamos para cada hijo mayor los *scores*  $u_{1j}$  y las nuevas observaciones  $\mathbf{y}_{1j}$

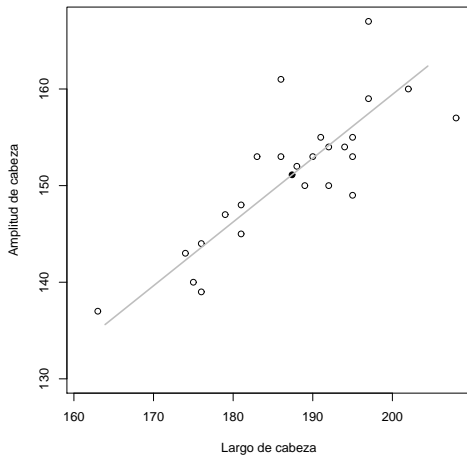
$$u_{1j} = \hat{\gamma}_1^T (\mathbf{x}_{1j} - \bar{\mathbf{x}}_1) \quad \mathbf{y}_{1j} = \bar{\mathbf{x}}_1 + \hat{\gamma}_1 u_{1j} = \bar{\mathbf{x}}_1 + \hat{\gamma}_1 \hat{\gamma}_1^T (\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)$$

Observemos que  $\mathbf{y}_{1j}$  están en la recta que pasa por  $\bar{\mathbf{x}}_1$  con dirección  $\hat{\gamma}_1$ .

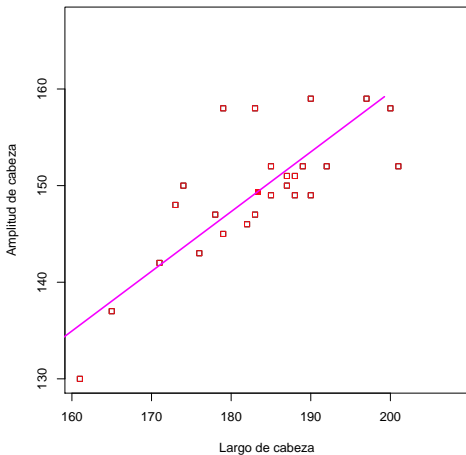
Hagamos lo mismo con el segundo hijo usando el vector

$$\hat{\gamma}_2 = (0.8507, 0.5257)^T$$

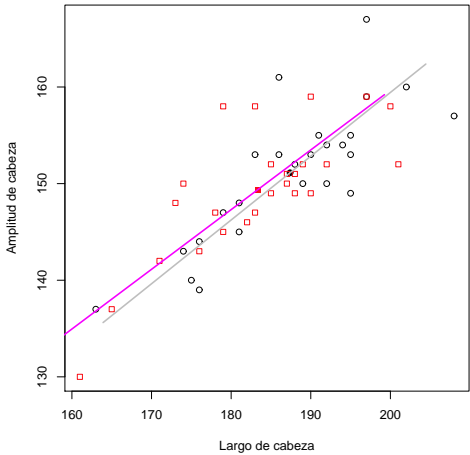
# Hijo Mayor



# Segundo Hijo



# Ambos Hijos

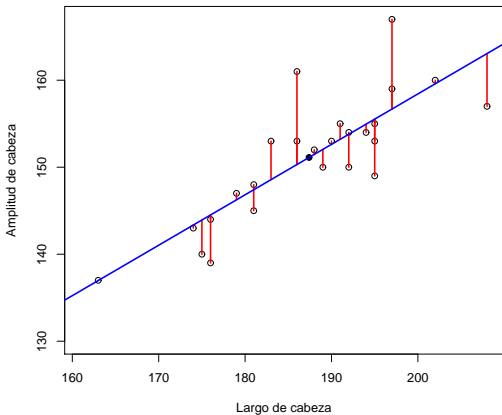


## Ejemplo 1

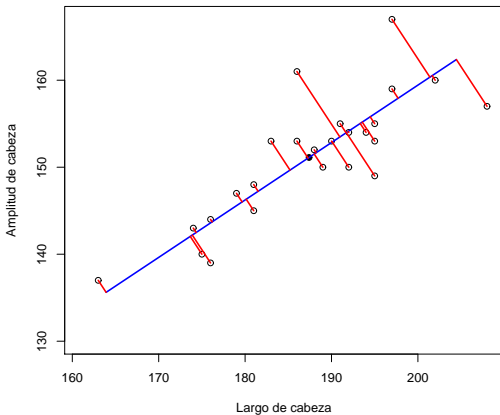
- En este ejemplo, ninguna de las dos variables  $x_1$  o  $x_2$  puede ser declarada como **independiente** o **dependiente**.
- Esto constituye la diferencia esencial con el análisis de regresión.
- La recta que obtuvimos **no es la recta de regresión** y se obtuvo minimizando la distancia de los puntos a la recta pero **en lugar de medir la distancia** verticalmente como en regresión, **la medimos en forma ortogonal a la recta**.
- Es el principio de mínimos cuadrados ortogonales de Pearson (1901).



# Hijo Mayor: Recta de regresión



# Hijo Mayor: Componente principal



## Definición

Sea  $\mathbf{x} \in \mathbb{R}^p$  tal que

$$\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu} \quad \text{VAR}(\mathbf{x}) = \boldsymbol{\Sigma}$$

Sean

- $\lambda_1 \geq \dots \geq \lambda_p$  los autovalores de  $\boldsymbol{\Sigma}$
- $\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_p$  los autovectores de  $\boldsymbol{\Sigma}$  asociados a  $\lambda_1 \geq \dots \geq \lambda_p$
- $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_p)$ ,  $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \mathbf{I}_p$
- $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p)$

$$\boldsymbol{\Gamma}^T \boldsymbol{\Sigma} \boldsymbol{\Gamma} = \boldsymbol{\Lambda}, \quad \boldsymbol{\Sigma} = \sum_{j=1}^p \lambda_j \boldsymbol{\gamma}_j \boldsymbol{\gamma}_j^T = \boldsymbol{\Gamma} \boldsymbol{\Lambda} \boldsymbol{\Gamma}^T$$

Luego, podemos escribir a  $\mathbf{x}$  como

$$\mathbf{x} = \boldsymbol{\mu} + \sum_{j=1}^p \boldsymbol{\gamma}_j^T (\mathbf{x} - \boldsymbol{\mu}) \boldsymbol{\gamma}_j$$

## Definición

Sea el vector  $\mathbf{v} = \mathbf{\Gamma}^T(\mathbf{x} - \boldsymbol{\mu})$ . Las coordenadas  $v_1, \dots, v_p$  de  $\mathbf{v}$  se llaman **las componentes principales de  $\mathbf{x}$** .

La  $j$ -ésima componente principal es, por lo tanto,

$$v_j = \gamma_j^T(\mathbf{x} - \boldsymbol{\mu}),$$

corresponde a la proyección ortogonal de  $(\mathbf{x} - \boldsymbol{\mu})$  en la dirección  $\gamma_j$ .

Se llama  $j$ -ésima **componente principal estandarizada** a la variable

$$z_j = \lambda_j^{-\frac{1}{2}} v_j = \lambda_j^{-\frac{1}{2}} \gamma_j^T(\mathbf{x} - \boldsymbol{\mu})$$

**Propiedad 1.** La componentes principales  $v_1, \dots, v_p$  son no correlacionadas y  $\text{VAR}(v_j) = \lambda_j$ , o sea,

$$\text{VAR}(\mathbf{v}) = \mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p)$$

## Correlaciones

Supongamos que  $\mathbb{E}\mathbf{x} = \boldsymbol{\mu}$  y  $\text{VAR}(\mathbf{x}) = \boldsymbol{\Sigma}$ .

En la práctica van a calcular la correlación entre  $x_j$ , la coordenada  $j$ -ésima de  $\mathbf{x}$ , y  $v_\ell$  y van a ver que, si  $\boldsymbol{\gamma}_\ell = (\gamma_{\ell,1}, \dots, \gamma_{\ell,p})^T$ ,

$$\text{Corr}(x_j, v_\ell) = \rho_{x_j, v_\ell} = \gamma_{\ell,j} \sqrt{\frac{\lambda_\ell}{\sigma_{jj}}} \quad (1)$$

Supongamos que predecimos a  $\mathbf{x}$  usando un predictor lineal basado en  $\mathbf{v}_q = (v_1, \dots, v_q)^T$ . El mejor predictor lineal de  $\mathbf{x}$  basado en  $\mathbf{v}_q$  es

$$\mathbf{x}^* = \boldsymbol{\mu} + \text{COV}(\mathbf{x}, \mathbf{v}_q) \{ \text{VAR}(\mathbf{v}_q) \}^{-1} \mathbf{v}_q = \boldsymbol{\mu} + \boldsymbol{\Gamma}_q \mathbf{v}_q$$

donde  $\boldsymbol{\Gamma}_q = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_q)$  y el residuo es  $\mathbf{u} = \mathbf{x} - \mathbf{x}^*$ .

## Correlaciones

Luego, si  $\mathbf{\Lambda}_q = \text{diag}(\lambda_1, \dots, \lambda_q)$

$$\text{VAR}(\mathbf{u}) = \mathbf{\Sigma} - \mathbf{\Gamma}_q \mathbf{\Lambda}_q \mathbf{\Gamma}_q^T = \mathbf{\Sigma} - \sum_{\ell=1}^q \lambda_\ell \boldsymbol{\gamma}_\ell \boldsymbol{\gamma}_\ell^T$$

o, sea,

$$\text{VAR}(x_j - x_j^*) = \sigma_{jj} - \sum_{\ell=1}^q \lambda_\ell \gamma_{\ell,j}^2$$

El término  $\lambda_\ell \gamma_{\ell,j}^2$  es la parte de la varianza de  $x_j$  explicada por  $v_\ell$  y por (1) es igual a  $\sigma_{jj} \rho_{x_j, v_\ell}^2$  de donde

$$\text{VAR}(x_j - x_j^*) = \sigma_{jj} \left( 1 - \sum_{\ell=1}^q \rho_{x_j, v_\ell}^2 \right)$$

## Predicción

Si  $\mathbb{E}\mathbf{x} = \boldsymbol{\mu}$  tenemos que la mejor predicción lineal está dada por

$$\mathbf{x}_q^* = \boldsymbol{\mu} + \boldsymbol{\Gamma}_q \mathbf{v}_q$$

donde

- $\boldsymbol{\Gamma}_q = (\gamma_1, \dots, \gamma_q)$
- $\mathbf{v}_q = \boldsymbol{\Gamma}_q^T (\mathbf{x} - \boldsymbol{\mu}) = (v_1, \dots, v_q)^T = (\gamma_1^T (\mathbf{x} - \boldsymbol{\mu}), \dots, \gamma_q^T (\mathbf{x} - \boldsymbol{\mu}))^T$

Una medida de la bondad del ajuste es

$$Q_1 = \|\mathbf{x} - \mathbf{x}_q^*\|^2 = (\mathbf{x} - \boldsymbol{\mu} - \boldsymbol{\Gamma}_q \mathbf{v}_q)^T (\mathbf{x} - \boldsymbol{\mu} - \boldsymbol{\Gamma}_q \mathbf{v}_q) = \sum_{j=q+1}^p v_j^2$$

y es utilizada en control de calidad.

## Predicción

Sea

$$\theta_m = \sum_{j=q+1}^p \lambda_j^m$$

entonces si  $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\left( \frac{Q_1}{\theta_1} \right)^\nu \approx N \left( 1 + \frac{\theta_2 \nu (\nu - 1)}{\theta_1^2}, 2 \frac{\theta_2 \nu^2}{\theta_1^2} \right)$$

donde

$$\nu = 1 - \left( \frac{2\theta_1\theta_3}{3\theta_2^2} \right)$$

Un valor grande de  $Q_1$  sugiere que el ajuste no es adecuado y que el punto  $\mathbf{x}$  está fuera de control.



## Lemas previos

**Lema 1.** Sea  $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$  una matriz simétrica definida no-negativa. Sean  $\lambda_1 \geq \dots \geq \lambda_p$  los autovalores de  $\mathbf{\Sigma}$  y  $\gamma_1, \dots, \gamma_p$  los autovectores de  $\mathbf{\Sigma}$  asociados a  $\lambda_1 \geq \dots \geq \lambda_p$ . Entonces

a)  $\sup_{\mathbf{u} \neq \mathbf{0}} \frac{\mathbf{u}^T \mathbf{\Sigma} \mathbf{u}}{\mathbf{u}^T \mathbf{u}} = \lambda_1$  y el supremo se alcanza en  $\gamma_1$ .

b)  $\inf_{\mathbf{u} \neq \mathbf{0}} \frac{\mathbf{u}^T \mathbf{\Sigma} \mathbf{u}}{\mathbf{u}^T \mathbf{u}} = \lambda_p$  y el infimo se alcanza en  $\gamma_p$ .

c)  $\sup_{\substack{\mathbf{u} \neq \mathbf{0} \\ \mathbf{u}^T \gamma_i = 0 \quad 1 \leq i \leq k}} \frac{\mathbf{u}^T \mathbf{\Sigma} \mathbf{u}}{\mathbf{u}^T \mathbf{u}} = \lambda_{k+1}$  y el supremo se alcanza en  $\gamma_{k+1}$ .

## Lemas previos

**Teorema de Courant–Fisher.** Sea  $\Sigma \in \mathbb{R}^{p \times p}$  una matriz simétrica definida no-negativa. Sean

- $\lambda_1 \geq \dots \geq \lambda_p$  los autovalores de  $\Sigma$  y
- $\gamma_1, \dots, \gamma_p$  los autovectores de  $\Sigma$  asociados a  $\lambda_1 \geq \dots \geq \lambda_p$ .

Entonces

$$\inf_{\mathbf{B} \in \mathbb{R}^{p \times k}} \sup_{\mathbf{B}^T \mathbf{u} = \mathbf{0}} \frac{\mathbf{u}^T \Sigma \mathbf{u}}{\mathbf{u}^T \mathbf{u}} = \lambda_{k+1}$$

y se alcanza en  $\mathbf{B}_0 = (\gamma_1, \dots, \gamma_k)$ .

## Lemas previos

**Teorema de separación de Poincaré.** Sea  $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$  una matriz simétrica definida no-negativa. Sean

- $\lambda_1 \geq \dots \geq \lambda_p$  los autovalores de  $\mathbf{\Sigma}$  y
- $\gamma_1, \dots, \gamma_p$  los autovectores de  $\mathbf{\Sigma}$  asociados a  $\lambda_1 \geq \dots \geq \lambda_p$ .

Entonces, si  $\mathbf{B} \in \mathbb{R}^{p \times k}$  es tal que  $\mathbf{B}^T \mathbf{B} = \mathbf{I}_k$ , se tiene que

$$\begin{aligned}\lambda_j(\mathbf{B}^T \mathbf{\Sigma} \mathbf{B}) &\leq \lambda_j = \lambda_j(\mathbf{\Sigma}) & 1 \leq j \leq k \\ \lambda_{k-j}(\mathbf{B}^T \mathbf{\Sigma} \mathbf{B}) &\geq \lambda_{p-j} = \lambda_{p-j}(\mathbf{\Sigma}) & 0 \leq j \leq k-1 \\ \lambda_s(\mathbf{B}^T \mathbf{\Sigma} \mathbf{B}) &\geq \lambda_{p-k+s} = \lambda_{p-k+s}(\mathbf{\Sigma}) & 1 \leq s \leq k\end{aligned}$$

donde  $\lambda_j(\mathbf{A})$  indica el  $j$ -ésimo autovalor de  $\mathbf{A}$ .

## Propiedades de optimalidad

**Propiedad 1.** (*Pearson, 1901*) Sea  $\mathbf{x} \in \mathbb{R}^p$  tal que

$$\mathbb{E}(\mathbf{x}) = \mathbf{0} \quad \text{VAR}(\mathbf{x}) = \mathbf{\Sigma}$$

- $\lambda_1 \geq \dots \geq \lambda_p$  los autovalores de  $\mathbf{\Sigma}$  y
- $\gamma_1, \dots, \gamma_p$  los autovectores de  $\mathbf{\Sigma}$  asociados a  $\lambda_1 \geq \dots \geq \lambda_p$ .
- $\mathcal{H}_0$  el subespacio generado por  $\gamma_1, \dots, \gamma_q$  donde  $\lambda_q > \lambda_{q+1}$ .

Indiquemos por  $\pi(\mathbf{x}, \mathcal{H})$  a la proyección ortogonal de  $\mathbf{x}$  sobre el subespacio  $\mathcal{H}$ . Entonces, se tiene que para todo subespacio  $\mathcal{H}$  de dimensión  $q$

$$\mathbb{E}\|\mathbf{x} - \pi(\mathbf{x}, \mathcal{H}_0)\|^2 \leq \mathbb{E}\|\mathbf{x} - \pi(\mathbf{x}, \mathcal{H})\|^2$$

o sea, las componentes principales dan el mejor ajuste lineal de dimensión  $q$ .

## Propiedades de optimalidad

**Propiedad 2.** Sea  $\mathbf{x} \in \mathbb{R}^p$  tal que

$$\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu} \quad \text{VAR}(\mathbf{x}) = \boldsymbol{\Sigma} > 0$$

Sean  $\lambda_1 \geq \dots \geq \lambda_p$  los autovalores de  $\boldsymbol{\Sigma}$  y  $\gamma_1, \dots, \gamma_p$  los autovectores de  $\boldsymbol{\Sigma}$  asociados a  $\lambda_1 \geq \dots \geq \lambda_p$ . Entonces,

a)  $\max_{\|\mathbf{a}\|=1} \text{VAR}(\mathbf{a}^T \mathbf{x}) = \text{VAR}(v_1)$ , o sea, el máximo se alcanza en  $\gamma_1$ .

b)  $\max_{\substack{\|\mathbf{a}\|=1 \\ \text{Cov}(\mathbf{a}^T \mathbf{x}, v_j) = 0 \quad 1 \leq j \leq k}} \text{VAR}(\mathbf{a}^T \mathbf{x}) = \text{VAR}(v_{k+1})$ ,

es decir, el máximo se alcanza en  $\gamma_{k+1}$ .

La condición  $\text{COV}(\mathbf{a}^T \mathbf{x}, v_j) = 0$  asegura que no se repite información.

c)  $\sum_{j=1}^p \text{VAR}(v_j) = \text{traza}(\boldsymbol{\Sigma})$ .

## Propiedades de optimalidad

**Propiedad 3.** Sea  $\mathbf{x} \in \mathbb{R}^p$  tal que

$$\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu} \quad \text{VAR}(\mathbf{x}) = \boldsymbol{\Sigma}$$

- $\lambda_1 \geq \dots \geq \lambda_p$  los autovalores de  $\boldsymbol{\Sigma}$  y
- $\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_p$  los autovectores de  $\boldsymbol{\Sigma}$  asociados a  $\lambda_1 \geq \dots \geq \lambda_p$ .

Queremos reemplazar a  $\mathbf{x}$  por  $q < p$  combinaciones lineales elegidas de modo a perder lo menos posible.

Tomemos  $y_j = \mathbf{a}_j^T \mathbf{x}$ ,  $1 \leq j \leq q$  y supongamos que  $\|\mathbf{a}_j\| = 1$  y  $\mathbf{a}_j^T \mathbf{a}_\ell = 0$  si  $j \neq \ell$ .

## Propiedades de optimalidad

**Propiedad 3.** Luego,

$$\text{VAR}(\mathbf{a}_j^T \mathbf{x}) = \mathbf{a}_j^T \boldsymbol{\Sigma} \mathbf{a}_j$$

por lo que las  $q$  combinaciones lineales  $(y_1, \dots, y_q)$  aportan

$$\sum_{j=1}^q \mathbf{a}_j^T \boldsymbol{\Sigma} \mathbf{a}_j$$

de la variación total de  $\mathbf{x}$  medida a través de la traza( $\boldsymbol{\Sigma}$ ).

Entonces, se cumple que

$$\max_{\substack{\|\mathbf{a}_j\|=1 \\ \mathbf{a}_j^T \mathbf{a}_\ell = 0 \quad j \neq \ell}} \sum_{j=1}^q \mathbf{a}_j^T \boldsymbol{\Sigma} \mathbf{a}_j = \sum_{j=1}^q \boldsymbol{\gamma}_j^T \boldsymbol{\Sigma} \boldsymbol{\gamma}_j = \sum_{j=1}^q \lambda_j$$

## Propiedades de optimalidad

Si  $\lambda_{q+1} = \dots = \lambda_p = 0$ , entonces  $v_{q+1}, \dots, v_p$  tienen varianza 0, o sea,

$$\mathbb{P}(\gamma_j^T(\mathbf{x} - \boldsymbol{\mu}) = 0 \quad \text{para todo } q + 1 \leq j \leq p) = 1$$

es decir,  $\mathbf{x} - \boldsymbol{\mu}$  yace en un subespacio de dimensión  $q$ .

Si esto no ocurre, deberíamos elegir  $q$  tal que  $\sum_{j=1}^q \lambda_j$  sea un porcentaje alto de la variación total de  $\mathbf{x}$ , o sea, de modo que por ejemplo

$$\frac{\sum_{j=1}^q \lambda_j}{\sum_{j=1}^p \lambda_j} = 0.95$$

Daremos test para verificar esta hipótesis basados en una muestra  $\mathbf{x}_1, \dots, \mathbf{x}_n$ .



## Normal Multivariada Singular

**Definición.** Sea  $\Sigma$  simétrica definida no-negativa. Se dice que  $\mathbf{x} \sim N(\mathbf{0}, \Sigma)$  si su función característica es

$$\varphi(\mathbf{u}) = \exp\left\{-\frac{1}{2}\mathbf{u}^T \Sigma \mathbf{u}\right\}$$

**Propiedad.** Si  $\mathbf{x} \sim N(\mathbf{0}, \Sigma)$  y  $\text{rango}(\Sigma) = q < p$  entonces

- $v_1, \dots, v_q$  son independientes
- $v_j \sim N(0, \lambda_j), 1 \leq j \leq q$
- para  $j \geq q + 1, \mathbb{P}(v_j = 0) = 1.$

## Normal Multivariada Singular: Ejemplo

Sea  $\mathbf{x} \sim N(\mathbf{0}, \Sigma)$  con

$$\Sigma = (1 - \rho)\mathbf{I}_p + \rho\mathbf{1}_p\mathbf{1}_p^T = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

- Los autovalores de  $\Sigma$  son  $(1 - \rho)$  con multiplicidad  $p - 1$  y  $1 + (p - 1)\rho$  con multiplicidad 1.

Tomemos

$$\rho = -\frac{1}{p-1}$$

de forma que  $\Sigma$  sea singular.

## Normal Multivariada Singular: Ejemplo

$$\Sigma = (1 - \rho)\mathbf{I}_p + \rho\mathbf{1}_p\mathbf{1}_p^T \quad \rho = -\frac{1}{p-1}$$

- Los autovalores de  $\Sigma$  son  $(1 - \rho)$  con multiplicidad  $p - 1$  y  $1 + (p - 1)\rho = 0$  con multiplicidad 1.
- El autovector asociado a  $\lambda_p = 0$  es  $\gamma_p = (1/\sqrt{p})\mathbf{1}_p$
- Cualquier conjunto de  $p - 1$  vectores ortogonales a  $\mathbf{1}_p$  se pueden tomar como los autovectores asociados a  $(1 - \rho)$ .

## Normal Multivariada Singular: Ejemplo

Podemos tomar entonces como componentes principales

$$\begin{aligned}v_1 &= \frac{x_1 - x_2}{\sqrt{2}} \\v_2 &= \frac{x_1 + x_2 - 2x_3}{\sqrt{2 \times 3}} \\&\vdots \\v_{p-1} &= \frac{x_1 + \cdots + x_{p-1} - (p-1)x_p}{\sqrt{p(p-1)}} \\v_p &= \frac{x_1 + \cdots + x_p}{\sqrt{p}}\end{aligned}$$

## Normal Multivariada Singular: Ejemplo

- $v_1, \dots, v_{p-1}$  son i.i.d. tales que  $v_j \sim N\left(0, 1 - \rho = \frac{p}{p-1}\right)$
- $\mathbb{P}(v_p = 0) = 1$

De esta forma, se obtiene por ejemplo, que

$$\begin{aligned}\mathbb{P}(3x_1 + x_2 - x_3 + x_4 + \dots + x_p > 0) &= \mathbb{P}(\sqrt{p}v_p + \sqrt{6}v_2 + \sqrt{2}v_1 > 0) \\ &= \mathbb{P}(\sqrt{3}v_2 + v_1 > 0) = \frac{1}{2}\end{aligned}$$

pues

$$\sqrt{3}v_2 + v_1 \sim N\left(0, \frac{4(p-2)}{p-1}\right)$$

## Componentes principales muestrales

En la práctica,  $\boldsymbol{\mu}$  y  $\boldsymbol{\Sigma}$  son desconocidos y deben ser estimados a partir de una muestra aleatoria  $\mathbf{x}_1, \dots, \mathbf{x}_n$ .

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}, \quad \mathbf{Q} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \quad \text{y} \quad \hat{\boldsymbol{\Sigma}} = \frac{\mathbf{Q}}{n}$$

Cuando  $\mathbf{x}$  tiene densidad, si  $n > p$ ,

$$\mathbb{P}(\mathbf{Q} > 0) = 1$$

y además,

$$\mathbb{P}(\lambda_1(\mathbf{Q}) > \lambda_2(\mathbf{Q}) > \dots > \lambda_p(\mathbf{Q})) = 1$$

## Componentes principales muestrales: $\widehat{\Sigma} = \widehat{\Gamma} \widehat{\Lambda} \widehat{\Gamma}^T$

- $\widehat{\lambda}_1 > \dots > \widehat{\lambda}_p$  los autovalores de  $\widehat{\Sigma}$  y
- $\widehat{\gamma}_1, \dots, \widehat{\gamma}_p$  los autovectores de  $\widehat{\Sigma}$  asociados a  $\widehat{\lambda}_1 > \dots > \widehat{\lambda}_p$ .

**Definición.** Para cada observación  $\mathbf{x}_i$  definimos el vector de componentes principales muestrales como

$$\widehat{\mathbf{v}}_i = \widehat{\Gamma}^T (\mathbf{x}_i - \bar{\mathbf{x}})$$

La coordenada  $j$ -ésima de  $\widehat{\mathbf{v}}_i$ ,  $\widehat{v}_{i,j}$ , se llama **la  $j$ -ésima componente principal de  $\mathbf{x}_i$** .

La  $j$ -ésima componente principal es, por lo tanto,

$$\widehat{v}_j = \widehat{\gamma}_j^T (\mathbf{x}_i - \bar{\mathbf{x}}),$$

corresponde a la proyección ortogonal de  $(\mathbf{x}_i - \bar{\mathbf{x}})$  en la dirección  $\widehat{\gamma}_j$ .

Las propiedades que vimos anteriormente se cumplen en términos de la distribución empírica.

## Ejemplo 2

Los *Microtus multiplex* son una familia de roedores presentes en Europa. En este ejemplo se tomaron 43 especímenes y para cada uno se midieron 8 variables

- Ancho del molar superior izquierdo # 1 (0.001mm)
- Ancho del molar superior izquierdo # 2 (0.001mm)
- Ancho del molar superior izquierdo # 3 (0.001mm)
- Largo de la fosa incisiva (0.001mm)
- Largo del hueso palatal (0.001mm)
- Largo del cráneo (0.01mm)
- Altura del cráneo sobre bullae (0.01mm)
- Ancho del cráneo a través del rostro (0.01mm)

obteniendo entonces vectores  $\mathbf{y}_i \in \mathbb{R}^8$ . Por conveniencia numérica, se presentan los resultados obtenidos con  $\mathbf{x}_i = \mathbf{y}_i/10$ .



## Ejemplo 2

$$\bar{x} = (205.4535, 163.6465, 181.9930, 396.6488, 526.0209, 238.5977, 80.9442, 46.8698)^T$$

$$S = \begin{pmatrix} 171.5130 & 97.4108 & 121.2151 & 158.7597 & 213.4108 & 88.4330 & 27.0469 & 23.2574 \\ 97.4108 & 102.3087 & 110.3706 & 161.7584 & 142.5469 & 73.9892 & 21.7843 & 17.8412 \\ 121.2151 & 110.3706 & 232.5688 & 250.9282 & 225.8311 & 110.3502 & 26.2622 & 24.0643 \\ 158.7597 & 161.7584 & 250.9282 & 737.7635 & 148.4182 & 187.5194 & 32.9356 & 42.2246 \\ 213.4108 & 142.5469 & 225.8311 & 148.4182 & 855.6855 & 159.8781 & 45.5893 & 36.5392 \\ 88.4330 & 73.9892 & 110.3502 & 187.5194 & 159.8781 & 87.0845 & 19.2189 & 19.3642 \\ 27.0469 & 21.7843 & 26.2622 & 32.9356 & 45.5893 & 19.2189 & 11.2949 & 5.2852 \\ 23.2574 & 17.8412 & 24.0643 & 42.2246 & 36.5392 & 19.3642 & 5.2852 & 5.7445 \end{pmatrix}$$

Los autovalores y autovectores de  $S$  son  $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_p)$  y  $\hat{\Gamma} = (\hat{\gamma}_1, \dots, \hat{\gamma}_p)$  donde

$$\hat{\Lambda} = \text{diag}(1305.4337, 651.5147, 123.2253, 75.9081, 27.8237, 13.2150, 5.7182, 1.1248)$$

$$\hat{\Gamma} = \begin{pmatrix} 0.2719 & -0.0219 & -0.5571 & 0.6380 & 0.4369 & 0.1191 & -0.0428 & -0.0344 \\ 0.2179 & 0.0559 & -0.3577 & 0.1295 & -0.8556 & 0.2432 & -0.1161 & 0.0019 \\ 0.3409 & 0.0863 & -0.5097 & -0.7495 & 0.2152 & 0.0895 & 0.0174 & 0.0141 \\ 0.5404 & 0.7174 & 0.4063 & 0.0853 & 0.0603 & 0.1285 & 0.0277 & -0.0067 \\ 0.6404 & -0.6854 & 0.3389 & -0.0108 & -0.0046 & 0.0716 & -0.0002 & 0.0015 \\ 0.2328 & 0.0652 & -0.1129 & 0.0380 & -0.1206 & -0.9226 & -0.1878 & -0.1623 \\ 0.0563 & -0.0053 & -0.0875 & 0.0571 & -0.1100 & -0.1322 & 0.9686 & -0.1347 \\ 0.0534 & 0.0140 & -0.0402 & 0.0478 & -0.0209 & -0.1684 & 0.1010 & 0.9768 \end{pmatrix}$$

## Ejemplo 2

$$\frac{\hat{\lambda}_1}{\sum_{j=1}^p \hat{\lambda}_j} = 0.5923$$

$$\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{\sum_{j=1}^p \hat{\lambda}_j} = 0.8879$$

$$\frac{\sum_{j=1}^3 \hat{\lambda}_j}{\sum_{j=1}^p \hat{\lambda}_j} = 0.9438$$

Además, si suponemos que los datos son normales, un estimador del desvío estandar de  $\hat{\lambda}_j$  es

$$\sqrt{\frac{2}{n}} \lambda_j$$

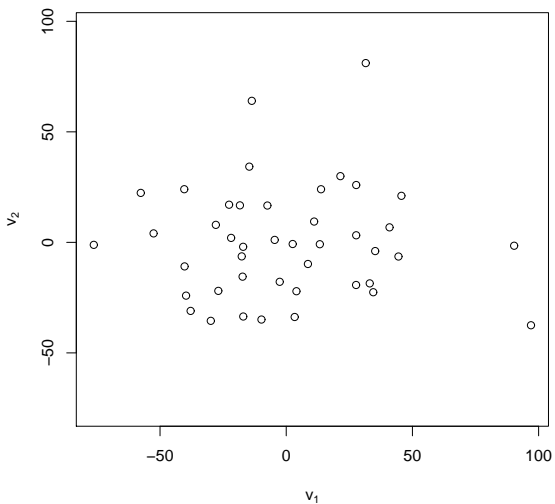
Luego, los desvíos estandar estimados de los autovalores  $s_{\hat{\lambda}_j}$  dan

$j$	1	2	3	4	5	6	7	8
$\hat{\lambda}_j$	1305.434	651.515	123.225	75.908	27.824	13.215	5.718	1.125
$s_{\hat{\lambda}_j}$	281.537	140.509	26.575	16.371	6.001	2.850	1.233	0.243

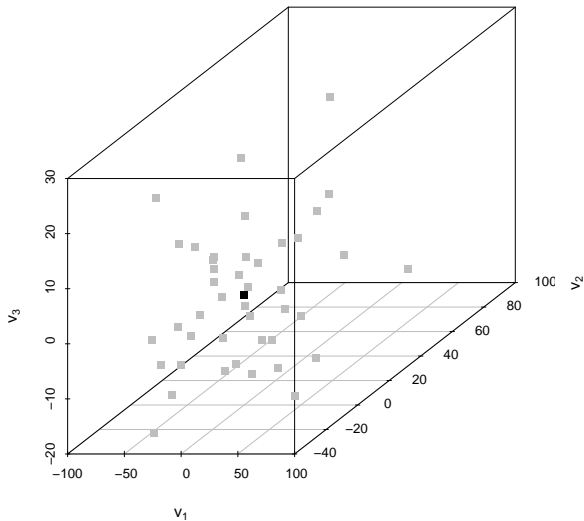
Es decir,

- podemos pensar que la segunda componente está bien determinada, o sea, que  $\lambda_3 \neq \lambda_2$  y
- quizás dudemos sobre la tercera o sea, no podemos asegurar todavía que  $\lambda_3 \neq \lambda_4$ .

## Ejemplo 2: Dos Primeras CP



## Ejemplo 2: Tres Primeras CP



## Ejemplo 2

Correlaciones absolutas entre las variables y las 3 primeras componentes principales

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$
$\hat{\gamma}_1$	0.272	0.218	0.341	0.540	0.640	0.233	0.056	0.053
$ \hat{\rho}_{x_j, v_1} $	0.750	0.779	0.808	0.719	0.791	0.901	0.605	0.805
$\hat{\gamma}_2$	-0.022	0.056	0.086	0.717	-0.685	0.065	-0.005	0.014
$ \hat{\rho}_{x_j, v_2} $	0.043	0.141	0.144	0.674	0.598	0.178	0.04	0.149
$\hat{\gamma}_3$	-0.557	-0.358	-0.510	0.406	0.339	-0.113	-0.088	-0.040
$ \hat{\rho}_{x_j, v_3} $	0.472	0.393	0.371	0.166	0.129	0.134	0.289	0.186

Observemos que las coordenadas de  $\hat{\gamma}_1$  son todas positivas, esto ocurre porque  $\mathbf{S}$  tiene todos sus elementos positivos.

Por lo tanto, todas las correlaciones  $\hat{\rho}_{x_j, v_1}$  son positivas.

## Inferencia en el caso normal

**Teorema.** Sean  $\mathbf{x}_1, \dots, \mathbf{x}_n$  i.i.d  $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , con  $\lambda_1 > \lambda_2 > \dots > \lambda_p > 0$  entonces

- $\hat{\boldsymbol{\Lambda}} = \text{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_p)$  es el EMV de  $\boldsymbol{\Lambda}$
- $\hat{\boldsymbol{\Gamma}} = (\hat{\gamma}_1, \dots, \hat{\gamma}_p)$  es el EMV de  $\boldsymbol{\Gamma}$

Además,

$$\hat{L}_j = \sqrt{n}(\hat{\lambda}_j - \lambda_j) \xrightarrow{D} N(0, 2\lambda_j^2)$$

asintóticamente independientes entre sí.

Si las observaciones no son normales se puede probar que  $\hat{L}_j$  es asintóticamente normal con varianza  $c\lambda_j^2$  pero no son necesariamente independientes

## Inferencia en el caso normal

Por otra parte, si  $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , con

$$\boldsymbol{\Sigma} = \text{diag}(\lambda_1, \dots, \lambda_p)$$

donde  $\lambda_1 > \lambda_2 > \dots > \lambda_p > 0$  y  $\hat{\boldsymbol{\gamma}}_k = (\hat{\gamma}_{k,1}, \dots, \hat{\gamma}_{k,p})^T$  entonces,

- $\star \hat{L}_j = \sqrt{n}(\hat{\lambda}_j - \lambda_j),$
- $\star F_{kj} = \sqrt{n}\hat{\gamma}_{k,j},$
- $\star E_{kk} = \sqrt{n}(\hat{\gamma}_{k,k} - 1)$

son independientes

- $F_{kj} \xrightarrow{D} N\left(0, \frac{\lambda_k \lambda_j}{(\lambda_k - \lambda_j)^2}\right)$
- $E_{kk} \xrightarrow{P} 0$

$H_{0,(r,h)} : \lambda_{r+1} = \lambda_{r+2} = \dots = \lambda_{r+h}$  versus  $H_{1,(r,h)} : \lambda_{r+1} > \lambda_{r+2} > \dots > \lambda_{r+h}$

El test del cociente de máxima verosimilitud se basa en

$$M_{r,h} = \frac{\prod_{j=r+1}^{r+h} \hat{\lambda}_j}{\left( \frac{1}{h} \sum_{j=r+1}^{r+h} \hat{\lambda}_j \right)^h}$$

Rechazando para valores chicos de  $M_{r,h}$  y se tiene que **cuando**  
 $H_{0,(r,h)}$  **es cierta**

$$-n \log(M_{r,h}) \xrightarrow{D} \chi_{\frac{h(h+1)}{2}-1}^2$$



$$H_{0,k} = H_{0,(p-k,k)} : \lambda_{p-k+1} = \lambda_{p-k+2} = \cdots = \lambda_p \text{ versus}$$
$$H_{1,k} : \lambda_{p-k+1} > \lambda_{p-k+2} > \cdots > \lambda_p$$

El test del cociente de máxima verosimilitud se basa en

$$M_k = \frac{\prod_{j=p-k+1}^p \hat{\lambda}_j}{\left( \frac{1}{k} \sum_{j=p-k+1}^p \hat{\lambda}_j \right)^k}$$

Rechazando para valores chicos de  $M_k$  y se tiene que **cuando**  $H_{0,k}$  es cierta

$$-n \log(M_k) \xrightarrow{D} \chi_{\frac{k(k+1)}{2}-1}^2$$

## Ejemplo 2

Veamos si  $H_{0,(1,2)} : \lambda_2 = \lambda_3$  y  $H_{0,(2,2)} : \lambda_3 = \lambda_4$  son ciertas.  
Tenemos que,

$$\begin{aligned}M_{1,2} &= 0.5350 & -n \log(M_{1,2}) &= 26.8942 \\M_{2,2} &= 0.9435 & -n \log(M_{2,2}) &= 2.4991 \\ \chi_{2,0.05}^2 &= 5.9915 & \chi_{2,0.01}^2 &= 9.2103\end{aligned}$$

Luego, rechazamos  $H_{0,(1,2)}$  pero no rechazamos  $H_{0,(2,2)}$ .  
Los  $p$ -valores son respectivamente,  $1.44 * 10^{-6}$  y 0.2866.

### Conclusión:

- No debemos dar ninguna interpretación relativa a  $v_3$  y  $v_4$  pues ese espacio no está bien determinado.
- Este resultado y el hecho que  $(\hat{\lambda}_1 + \hat{\lambda}_2) / \sum_{j=1}^p \hat{\lambda}_j = 0.8879$  sugeriría que la variabilidad en las mandíbulas de los roedores estudiados podría ser adecuadamente descrita por las dos primeras componentes principales.

## Test para porcentajes

Queremos testear

$$H_{0,p_0} : \frac{\sum_{j=1}^q \lambda_j}{\sum_{j=1}^p \lambda_j} = p_0 \quad \text{versus} \quad H_{1,p_0} : \frac{\sum_{j=1}^q \lambda_j}{\sum_{j=1}^p \lambda_j} \neq p_0$$

donde  $p_0$  es un porcentaje prefijado. También podría interesarnos

$$H_{0,p_0}^* : \frac{\sum_{j=1}^q \lambda_j}{\sum_{j=1}^p \lambda_j} \leq p_0 \quad \text{versus} \quad H_{1,p_0}^* : \frac{\sum_{j=1}^q \lambda_j}{\sum_{j=1}^p \lambda_j} > p_0$$

Para obtener un test para estas hipótesis nos basaremos en la distribución asintótica de los autovalores

$$\sqrt{n}(\hat{\lambda}_j - \lambda_j) \xrightarrow{D} N(0, 2\lambda_j^2)$$

y usaremos que son asintóticamente independientes entre sí.

## Test para porcentajes

Observemos que

$$\frac{\sum_{j=1}^q \lambda_j}{\sum_{j=1}^p \lambda_j} = p_0$$

es equivalente a  $\theta = 0$  donde

$$\theta_{p_0} = (1 - p_0) \sum_{j=1}^q \lambda_j - p_0 \sum_{j=q+1}^p \lambda_j.$$

Definamos el estimador de  $\theta$

$$\hat{\theta}_{p_0} = (1 - p_0) \sum_{j=1}^q \hat{\lambda}_j - p_0 \sum_{j=q+1}^p \hat{\lambda}_j.$$

Sea

$$\sigma_{p_0}^2 = 2(1 - p_0)^2 \sum_{j=1}^q \lambda_j^2 + 2p_0^2 \sum_{j=q+1}^p \lambda_j^2$$

## Test para porcentajes

Entonces,

$$\sqrt{n}(\hat{\theta}_{p_0} - \theta_{p_0}) \xrightarrow{D} N(0, \sigma_{p_0}^2)$$

Por lo tanto, definiendo

$$\hat{\sigma}_{p_0}^2 = 2(1 - p_0)^2 \sum_{j=1}^q \hat{\lambda}_j^2 + 2p_0^2 \sum_{j=q+1}^p \hat{\lambda}_j^2$$

resulta que

$$\sqrt{n}(\hat{\theta}_{p_0} - \theta_{p_0}) / \hat{\sigma}_{p_0} \xrightarrow{D} N(0, 1)$$

## Test para porcentajes

Un intervalo de confianza asintótico de nivel  $1 - \alpha$  para  $\theta$  es

$$\mathcal{I}_{p_0} = \left[ \hat{\theta}_{p_0} - \frac{\hat{\sigma}_{p_0}}{\sqrt{n}} z_{\frac{\alpha}{2}}, \hat{\theta}_{p_0} + \frac{\hat{\sigma}_{p_0}}{\sqrt{n}} z_{\frac{\alpha}{2}} \right]$$

Luego, rechazo  $H_{0,p_0}$  si  $0 \notin \mathcal{I}_{p_0}$  o equivalentemente si

$$\sqrt{n} \frac{|\hat{\theta}_{p_0}|}{\hat{\sigma}_{p_0}} \geq z_{\frac{\alpha}{2}}$$

Por otra parte, un test asintótico para  $H_{0,p_0}^*$  o sea, para  $H_{0,p_0}^* : \theta_{p_0} \leq 0$  versus  $\theta_{p_0} > 0$  rechaza si

$$\sqrt{n} \frac{\hat{\theta}_{p_0}}{\hat{\sigma}_{p_0}} \geq z_{\alpha}$$

## Ejemplo 2

$$\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{\sum_{j=1}^p \hat{\lambda}_j} = 0.8879$$

Estudiamos si vale  $H_{0,p_0}^*$

$$H_{0,p_0}^* : \frac{\lambda_1 + \lambda_2}{\sum_{j=1}^p \lambda_j} \leq p_0 \quad \text{versus} \quad H_{1,p_0}^* : \frac{\lambda_1 + \lambda_2}{\sum_{j=1}^p \lambda_j} > p_0$$

con  $p_0 = 0.80$  y  $p_0 = 0.85$ .

$p_0$	$\hat{\theta}_{p_0}$	$\hat{\sigma}_{p_0}$	$\sqrt{n} \hat{\theta}_{p_0} / \hat{\sigma}_{p_0}$	p-valor
0.80	193.778	445.38	2.853	0.0022
0.85	83.579	357.04	1.535	0.0624

## Ejemplo 2

$$\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{\sum_{j=1}^p \hat{\lambda}_j} = 0.8879$$

Estudiamos si vale  $H_{0,p_0}$

$$H_{0,p_0} : \frac{\lambda_1 + \lambda_2}{\sum_{j=1}^p \lambda_j} = p_0 \quad \text{versus} \quad H_{1,p_0} : \frac{\lambda_1 + \lambda_2}{\sum_{j=1}^p \lambda_j} \neq p_0$$

con  $p_0 = 0.90$ .

$p_0$	$\hat{\theta}_{p_0}$	$\hat{\sigma}_{p_0}$	$\sqrt{n}  \hat{\theta}_{p_0}  / \hat{\sigma}_{p_0}$	$p$ -valor	$\mathcal{I}_{p_0}$
0.90	-26.619	279.46	0.6246	0.532	(-62.213, 8.976)