TEMA 4
ANÁLISIS DE COMPONENTES PRINCIPALES
(ACP)

Curso 2017/18

1 Introducción

Objetivo del ACP: Determinar un espacio de dimensión "reducida" que represente adecuadamente un conjunto de n observaciones $p-$ dimensionales.

- El ACP pretende sustituir las variables originales por un número pequeño de combinaciones lineales de las variables originales, incorreladas y "perdiendo" poca información.
  - ¿Puede representarse los datos en un espacio bidimensional que recoja la mayor parte de la información?

- En ocasiones el ACP es un input para análisis posteriores: análisis de regresión, análisis de conglomerados, ANOVA,...

- Su construcción no requiere supuesto de normalidad. No obstante, en poblaciones normales se pueden realizar tests de hipótesis y proporcionan interpretaciones útiles de los elipsoides de densidad constante

2 Notación y resultados previos

- Sea $\underline{x}_1, ...\underline{x}_n$ una realización muestral de un vector $\underline{X} = (X_1, ..., X_p)' \sim (\mu, \mathbf{\Sigma})$

- Sea $\mathbf{X} = [\underline{x}_1, ..., \underline{x}_n]'$ la matriz de datos muestrales.

- Sean $\underline{x}_{(1)}, ..., \underline{x}_{(p)}$ las columnas de $\mathbf{X} = \left[\underline{x}_{(1)}, ..., \underline{x}_{(p)}\right]$

- Sea $\widehat{\mathbf{\Sigma}}$ la matriz de (cuasi) varianzas-covarianzas muestrales

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})'$$

- $\underline{x}_{(k)}$ es una muestra de la variable X_k
  - Varianza muestral de $\underline{x}_{(k)}$:

$$\widehat{\sigma}^2_{\underline{x}_{(k)}} = \widehat{\sigma}_{kk}$$

  - Covarianza muestral entre $\underline{x}_{(k)}$ y $\underline{x}_{(h)}$:

$$\widehat{\sigma}_{\underline{x}_{(k)},\underline{x}_{(h)}} = \widehat{\sigma}_{kh}$$

2.1 Transformaciones lineales de los datos

- Sea $\underline{t} \in \mathbb{R}^p$ y considérese la variable $y = \underline{t}'\underline{X} = \sum_{k=1}^{p} t_k X_k$ (suma ponderada de las variables originales).

- $\begin{pmatrix} \underline{t}'\underline{x}_1 \\ \vdots \\ \underline{t}'\underline{x}_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \mathbf{X}\underline{t} = [\underline{x}_1, ..., \underline{x}_n]' \, \underline{t}$ proporciona n observaciones de la nueva variable y

  - Media muestral: $\overline{y} = \underline{t}'\overline{\underline{x}}$

  - Varianza muestral: $\widehat{\sigma}_y^2 = \widehat{\sigma}_{\underline{t}'\underline{x}}^2 = \underline{t}'\widehat{\mathbf{\Sigma}}\underline{t}$

  - Dados $\underline{t}, \underline{d} \in \mathbb{R}^p$, la covarianza muestral:

$$\widehat{\sigma}_{\underline{t}'\underline{x}, \underline{d}'\underline{x}} = \underline{t}'\widehat{\mathbf{\Sigma}}\underline{d}$$

2.2 Propiedades de $\widehat{\mathbf{\Sigma}}$

- $\widehat{\mathbf{\Sigma}}$ es simétrica, s.d.p.

- Los autovalores son reales (simétrica) y positivos (s.d.p.)

$$\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq ... \geq \widehat{\lambda}_p \geq 0.$$

- Los autovectores son ortogonales (simétrica)

$$\underline{\widehat{e}}_k' \, \underline{\widehat{e}}_h = 0, \quad k \neq h$$

- Sean $\underline{\widehat{e}}_1, \ \underline{\widehat{e}}_2, ..., \underline{\widehat{e}}_p$ los **autovectores unitarios ortogonales** y

$$\widehat{\mathbf{E}} = [\underline{\widehat{e}}_1, \ \underline{\widehat{e}}_2, ..., \underline{\widehat{e}}_p]$$

- $\widehat{\mathbf{E}}$ es una matriz ortogonal

$$\widehat{\mathbf{E}}'\widehat{\mathbf{E}} = \widehat{\mathbf{E}}\widehat{\mathbf{E}}' = \mathbf{I}_p$$

- **Teorema de descomposición espectral**

$$\widehat{\mathbf{E}}' \; \widehat{\mathbf{\Sigma}} \; \widehat{\mathbf{E}} = \widehat{\mathbf{\Lambda}} = diag\{\widehat{\lambda}_1, \widehat{\lambda}_2, ..., \widehat{\lambda}_p\}$$

  –

$$\widehat{\mathbf{\Sigma}} \; = \widehat{\mathbf{E}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{E}}' = \sum_{k=1}^{p} \widehat{\lambda}_k \underline{\widehat{e}}_k \underline{\widehat{e}}_k'$$

-

$$\text{Traza}\left(\widehat{\mathbf{\Sigma}}\;\right) = \sum_{k=1}^{p} \widehat{\sigma}^2_{\underline{x}_{(k)}} = \text{Traza}\left(\widehat{\mathbf{\Lambda}}\right) = \sum_{k=1}^{p} \widehat{\lambda}_k$$

3 Componentes principales (muestrales)

Definición 1 *Se define la* $\mathbf{k}-$**ésima componente principal muestral** *como*

$$\underline{\widehat{e}}'_k \underline{X}, \quad k = 1, ..., p.$$

- Las variables originales $\underline{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}$ se han transformado en unas nuevas variables

  $$\underline{Y} = \begin{pmatrix} \underline{\widehat{e}}'_1 \underline{X} \\ \vdots \\ \underline{\widehat{e}}'_p \underline{X} \end{pmatrix}.$$

- Puntuaciones correspondientes a la $k - ésima$ componente principal

$$\underline{y}_{(k)} = \begin{pmatrix} y_{1k} \\ \vdots \\ y_{nk} \end{pmatrix} = \begin{pmatrix} \widehat{\underline{e}}'_k \underline{x}_1 \\ \vdots \\ \widehat{\underline{e}}'_k \underline{x}_n \end{pmatrix} = \mathbf{X}\widehat{\underline{e}}_k, \quad k = 1, ..., p$$

- Puntaciones de las p componentes principales

$$\mathbf{Y} = \begin{pmatrix} \underline{x}'_1 \widehat{e}_1 & \cdots & \underline{x}'_1 \widehat{e}_p \\ \vdots & & \vdots \\ \underline{x}'_n \widehat{e}_1 & \cdots & \underline{x}'_n \widehat{e}_p \end{pmatrix} = \mathbf{X}\widehat{\mathbf{E}} = (\underline{y}_{(1)}, ..., \underline{y}_{(p)})$$

- Transformación de los datos muestrales

$$\mathbf{Y} = \begin{pmatrix} \underline{x}_1'\widehat{e}_1 & \cdots & \underline{x}_1'\widehat{e}_p \\ \vdots & & \vdots \\ \underline{x}_n'\widehat{e}_1 & \cdots & \underline{x}_n'\widehat{e}_p \end{pmatrix} = \mathbf{X}\widehat{\mathbf{E}} = \begin{pmatrix} \underline{y}_1' \\ \vdots \\ \underline{y}_n' \end{pmatrix}$$

con

$$\underline{y}_i = \begin{pmatrix} \underline{x}_1'\widehat{e}_1 \\ \vdots \\ \underline{x}_1'\widehat{e}_p \end{pmatrix} = \begin{pmatrix} \widehat{e}_1' \\ \vdots \\ \widehat{e}_p' \end{pmatrix} \underline{x}_i = \widehat{\mathbf{E}}' \, \underline{x}_i, \; i = 1, ..., n$$

  - $\underline{y}_1, ..., \underline{y}_n$ representan los valores de los n elementos de la muestra respecto de las nuevas variables (CP).

  - Como $\widehat{\mathbf{E}}$ es una matriz ortogonal se mantienen las distancias entre los datos originales y los datos transformados

$$\begin{aligned} d^2\left(\underline{y}_i, \underline{y}_j\right) &= \left(\underline{y}_j - \underline{y}_i\right)'\left(\underline{y}_j - \underline{y}_i\right) = \left(\underline{x}_j - \underline{x}_i\right)' \widehat{\mathbf{E}}\widehat{\mathbf{E}}' \left(\underline{x}_j - \underline{x}_i\right) = \\ &= \left(\underline{x}_j - \underline{x}_i\right)'\left(\underline{x}_j - \underline{x}_i\right) = d^2\left(\underline{x}_i, \underline{x}_j\right) \end{aligned}$$

Propiedades 1

1. Matriz de varianzas-covarianzas de $\underline{y}_i = \widehat{\mathbf{E}}' \, \underline{x}_i$, $i = 1, ..., n$:

$$\widehat{\mathbf{\Sigma}}_y = \widehat{\mathbf{E}}' \, \widehat{\mathbf{\Sigma}} \, \widehat{\mathbf{E}} = \widehat{\mathbf{\Lambda}} = diag\{\widehat{\lambda}_1, \widehat{\lambda}_2, ..., \widehat{\lambda}_p\}$$

2. Varianza muestral de la kth CP: $\quad \widehat{\sigma}^2_{\underline{y}_{(k)}} = \widehat{\sigma}^2_{\underline{\widehat{e}}'_k \underline{x}} = \widehat{\lambda}_k \, , \;\; k = 1, ..., p$

3. Covarianza muestral entre dos CPs: $\quad \widehat{\sigma}_{\underline{y}_{(k)}, \underline{y}_{(h)}} = \widehat{\sigma}_{\underline{\widehat{e}}'_k \underline{x}, \underline{\widehat{e}}'_h \underline{x}} = 0, \;\; \forall k \neq h$

4. $\sum_{k=1}^{p} \widehat{\sigma}^2_{\underline{x}_{(k)}} = \mathrm{tr}(\widehat{\mathbf{\Sigma}}) = \sum_{k=1}^{p} \widehat{\lambda}_k = \sum_{k=1}^{p} \widehat{\sigma}^2_{\underline{y}_{(k)}}$

5. El coeficiente de correlación muestral entre $\underline{y}_{(k)}$ y $\underline{x}_{(h)}$ viene dado por

$$r(\underline{y}_{(k)}; \underline{x}_{(h)}) = \frac{\widehat{e}_{kh} \sqrt{\widehat{\lambda}_k}}{\widehat{\sigma}_{\underline{x}_{(h)}}}$$

Teorema

1. La 1^a CP muestral es la combinación lineal (normalizada) de máxima varianza:

$$\widehat{\sigma}^2_{\underline{y}_{(1)}} = \widehat{\sigma}^2_{\underline{\widehat{e}}'_1 \underline{x}} = \sup_{\{\underline{t}\in\mathbb{R}^p:\ \underline{t}'\underline{t}=\mathbf{1}\}} \widehat{\sigma}^2_{\underline{t}'\underline{x}}$$

2. La 2^a CP muestral es la c.l. de máxima varianza incorrelada con la 1^a CP:

$$\widehat{\sigma}^2_{\underline{y}_{(2)}} = \widehat{\sigma}^2_{\underline{\widehat{e}}'_2 \underline{x}} = \sup_{\left\{\underline{t}\in\mathbb{R}^p:\ \underline{t}'\underline{t}=\mathbf{1}\ ,\ \underline{t}'\widehat{\mathbf{\Sigma}}\underline{\widehat{e}}_1=0\right\}} \widehat{\sigma}^2_{\underline{t}'\underline{x}}$$

3. La $k-$ésima CP muestral verifica:

$$\widehat{\sigma}^2_{\underline{y}_{(k)}} = \widehat{\sigma}^2_{\underline{\widehat{e}}'_k \underline{x}} = \sup_{\left\{\underline{t}\in\mathbb{R}^p:\ \underline{t}'\underline{t}=\mathbf{1}\ ,\ \underline{t}'\widehat{\mathbf{\Sigma}}\underline{\widehat{e}}_h=0\ \ h=1,\ldots,k-1\right\}} \widehat{\sigma}^2_{\underline{t}'\underline{x}}, \quad k = 1, \ldots, p$$

Demostración

Sea $\underline{t} \in \mathbb{R}^p$ con $\underline{t}'\underline{t}$=1

Como $rg(\widehat{\mathbf{E}})$ =p, los autovectores constituyen una base ortonormal de $\mathbb{R}^p$,

- $\exists \underline{c} = (c_1, ..., c_p)' \in \mathbb{R}^p$ tal que $\underline{t} = \sum_{k=1}^{p} c_k \widehat{\underline{e}}_k = \widehat{\mathbf{E}}\underline{c} \Rightarrow \underline{c} = \widehat{\mathbf{E}}'\underline{t}$

- $\underline{t}'\underline{t}=1 \Longrightarrow \underline{c}'\underline{c}=\underline{t}'\mathbf{E}'\mathbf{E}\underline{t}=\underline{t}'\underline{t}=1.$

- Varianza de la combinación lineal $\underline{t}'\underline{x}$

$$\widehat{\sigma}^2_{\underline{t}'\underline{x}} = \underline{t}'\widehat{\mathbf{\Sigma}}\underline{t} = \underline{c}'\widehat{\mathbf{E}}'\widehat{\mathbf{\Sigma}}\widehat{\mathbf{E}}\underline{c} = \underline{c}'\widehat{\mathbf{\Lambda}}\underline{c} = \sum_{k=1}^{p} \widehat{\lambda}_k \, c_k^2 \tag{1}$$

- Covarianza entre $\underline{t}'\underline{x}$ y $\underline{e}_k'\underline{x}$

$$\widehat{\sigma}_{\underline{t}'\underline{x}, \underline{e}_k'\underline{x}} = \underline{t}'\widehat{\mathbf{\Sigma}}\widehat{\underline{e}}_k = \underline{c}'\widehat{\mathbf{E}}'\widehat{\mathbf{\Sigma}}\widehat{\underline{e}}_k = \underline{c}'\widehat{\mathbf{E}}'\widehat{\boldsymbol{\lambda}}_k\widehat{\underline{e}}_k = c_k\widehat{\lambda}_k \tag{2}$$

1. **Primera CP**

Sea $\underline{t} \in \mathbb{R}^p$ con $\underline{t}'\underline{t}$=1.

De (1)

$$\widehat{\sigma}^2_{\underline{t}'\underline{x}} = \sum_{k=1}^{p} \widehat{\lambda}_k \, c_k^2 \leq \widehat{\lambda}_1 \underline{c}'\underline{c} = \widehat{\lambda}_1 = \widehat{\sigma}^2_{\underline{y}_{(1)}} = \widehat{\sigma}^2_{\underline{\widehat{e}}_1'\underline{x}}$$

2. **Segunda CP**

Sea $\underline{t} \in \mathbb{R}^p$, con $\underline{t}'\underline{t}$=1 y $\widehat{\sigma}_{\underline{t}'\underline{x},\widehat{\underline{e}}_1'\underline{x}} = 0$

De (2)

$$\widehat{\sigma}_{\underline{t}'\underline{x},\widehat{\underline{e}}_1'\underline{x}} = \underline{t}'\widehat{\mathbf{\Sigma}}\widehat{\underline{e}}_1 = c_1\widehat{\lambda}_1 = 0$$

- Podemos suponer que $\widehat{\lambda}_1 > 0$, ya que si $\widehat{\lambda}_1 = 0 \Rightarrow \widehat{\lambda}_1 = \cdots = \widehat{\lambda}_p = 0 \Rightarrow$

$$\widehat{\sigma}^2_{\underline{t}'\underline{x}} = \sum_{k=1}^{p} \widehat{\lambda}_k\, c_k^2 = 0, \quad \forall\ \underline{t} \in R^p$$

  - Trivialmente $\widehat{\underline{e}}_2'\underline{x}$ es la 2^a CP

- Si $\widehat{\lambda}_1 > 0 \Rightarrow c_1 = 0$ y por (1)

$$\widehat{\sigma}^2_{\underline{t}'\underline{x}} = \sum_{k=1}^{p} \widehat{\lambda}_k\, c_k^2 = \sum_{k=2}^{p} \widehat{\lambda}_k\, c_k^2 \leq \widehat{\lambda}_2\underline{c}'\underline{c} = \widehat{\lambda}_2 = \widehat{\sigma}^2_{\underline{y}_{(2)}} = \widehat{\sigma}^2_{\widehat{\underline{e}}_2'\underline{x}}$$

3. **k-ésima CP**

Sea $\underline{t} \in R^p$, con $\underline{t}'\underline{t}$=1 y $\widehat{\sigma}_{\underline{t}'\underline{x},\widehat{\underline{e}}'_h\underline{x}} = 0$, $h = 1, ..., k-1$.

De (2)

$$\widehat{\sigma}_{\underline{t}'\underline{x},\widehat{\underline{e}}'_h\underline{x}} = c_h\widehat{\lambda}_h = 0$$

- Podemos suponer que $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq ... \geq \widehat{\lambda}_{k-1} > 0 \Rightarrow c_h = 0, \qquad h = 1, ..., k-1$

Por tanto,

$$\widehat{\sigma}^2_{\underline{t}'\underline{x}} = \sum_{h=k}^{p} \widehat{\lambda}_h c_h^2 \leq \widehat{\lambda}_k \underline{c}'\underline{c} = \widehat{\lambda}_k = \widehat{\sigma}^2_{\underline{y}_{(k)}} = \widehat{\sigma}^2_{\widehat{\underline{e}}'_k\underline{x}}$$

3.1 Componentes principales muestrales para datos estandarizados

- Las CPs no son en general invariantes respecto a cambios de escala.

- Se puede realizar el estudio de las CP a partir de la matriz de datos estandarizados

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \longrightarrow \mathbf{X}_{(s)} = \begin{pmatrix} \frac{x_{11}-\overline{x}_1}{\widehat{\sigma}_1} & \ldots & \frac{x_{1p}-\overline{x}_p}{\widehat{\sigma}_p} \\ \vdots & & \vdots \\ \frac{x_{n1}-\overline{x}_1}{\widehat{\sigma}_1} & \ldots & \frac{x_{1p}-\overline{x}_p}{\widehat{\sigma}_p} \end{pmatrix}$$

- La matriz de covarianzas asociada a $\mathbf{X}_{(s)}$ es la matriz de correlaciones de $\mathbf{X}$

$$\widehat{\mathbf{\Sigma}}_{\mathbf{X}_{(s)}} = \widehat{\mathbf{R}}_X$$

- Sean $\widehat{e}_{1(s)}, ..., \widehat{e}_{p(s)}$ los autovectores asociados a los autovalores de $\widehat{\mathbf{\Sigma}}_{\mathbf{X}_{(s)}}$, $\widehat{\lambda}_{1(s)} \geq ... \geq \widehat{\lambda}_{p(s)}$

- Las puntuaciones de las CPs a partir de los datos estandarizados son

$$\mathbf{Y}_{(s)} = \mathbf{X}_{(s)} \left[\widehat{e}_{1(s)}, ..., \widehat{e}_{p(s)}\right]$$

Nota 1 *En genera las CPs sobre los datos estandarizados y sin estandarizar no coinciden.*

- *Cuando las escalas de medida de las variables son muy distintas, las variables con valores más grandes tendrán más peso en el análisis. Se estandarizan las variables para que las magnitudes de las variables sean similares.*

- *La estandarización resuelve otro posible problema. Si las variabilidades de los datos originales son muy distintas, las variables con mayor varianza van a influir más en la determinación de la primera componente principal.*

3.2 Selección del número de componentes

- Variabilidad explicada por la k-th C.P. muestral

$$\frac{\widehat{\lambda}_k}{\sum_{h=1}^{p} \widehat{\lambda}_h}$$

- Proporción de la variabilidad total explicada/ no explicada por las k primeras C.P. muestrales:

$$\text{Explicada:} \quad \frac{\sum_{h=1}^{k} \widehat{\lambda}_h}{\sum_{h=1}^{p} \widehat{\lambda}_h} \qquad \text{No explicada:} \quad \frac{\sum_{h=k+1}^{p} \widehat{\lambda}_h}{\sum_{h=1}^{p} \widehat{\lambda}_h}$$

- El número de CPs a seleccionar dependerá de las magnitudes relativas de los autovalores.

  - Elegir las k primeras si

$$\frac{\widehat{\lambda}_{k+1}}{\sum_{h=1}^{p} \widehat{\lambda}_h} < \delta_1 \quad \text{o} \quad \frac{\sum_{h=k+1}^{p} \widehat{\lambda}_h}{\sum_{h=1}^{p} \widehat{\lambda}_h} < \delta_2$$

    donde δ_1 y δ_2 serán elegidos convenientemente.

  - Retener componentes suficientes hasta que expliquen un determinado porcentaje de la variabilidad total (80%)

  - Retener todas las componentes cuyos autovalores sean mayores que la media $\frac{\sum_{h=1}^{p} \widehat{\lambda}_h}{p}$.

    La media es 1 si las CP se calculan sobre las variables estandarizadas.

  - Hacer un gráfico de k frente a $\widehat{\lambda}_k$, y ver si se observa una división entre "autovalores grandes" y "pequeños".

  - Hacer un test de hipótesis (requiere normalidad multivariante)