

$$\chi_s = H \chi \bar{P}^{-\frac{1}{2}}, \quad \text{donde}$$

$$\chi = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \text{matriz de datos}$$

$$H = I_n - \frac{1}{n} \mathbf{1}_{(n)} \mathbf{1}_{(n)}^T, \quad \text{matriz de centrado}$$

$$\bar{P} = \text{DIAG}(S_{X_1 X_1}, \dots, S_{X_p X_p}) = \begin{pmatrix} S_{X_1 X_1} & 0 & \cdots & 0 \\ 0 & S_{X_2 X_2} & \cdots & 0 \\ \vdots & & & \\ 0 & \cdots & \cdots & S_{X_p X_p} \end{pmatrix}$$

es una matriz diagonal, tal que el i -ésimo elemento en su diagonal es

$$S_{X_i X_i} = \sum_{j=1}^n (x_{ji} - \bar{x}_{\cdot i})^2 = \widehat{\text{VAR}}(X_i).$$

$$\chi_s = (\chi - \mathbf{1}_{(n)} \bar{x}_{(n)}^T) \bar{P}^{-\frac{1}{2}}$$

$$= \left(\chi - \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} (\bar{x}_{\cdot 1}, \dots, \bar{x}_{\cdot p}) \right) \bar{P}^{-\frac{1}{2}}$$

$$= \left[\chi - \begin{pmatrix} \bar{x}_{\cdot 1} & \cdots & \bar{x}_{\cdot p} \\ \bar{x}_{\cdot 1} & \cdots & \bar{x}_{\cdot p} \\ \vdots & & \vdots \\ \bar{x}_{\cdot 1} & & \bar{x}_{\cdot p} \end{pmatrix} \right] \bar{P}^{-\frac{1}{2}}$$

$$= \begin{pmatrix} x_{11} - \bar{x}_{\cdot 1} & x_{12} - \bar{x}_{\cdot 2} & \cdots & x_{1p} - \bar{x}_{\cdot p} \\ x_{21} - \bar{x}_{\cdot 1} & x_{22} - \bar{x}_{\cdot 2} & \cdots & x_{2p} - \bar{x}_{\cdot p} \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{x}_{\cdot 1} & x_{n2} - \bar{x}_{\cdot 2} & \cdots & x_{np} - \bar{x}_{\cdot p} \end{pmatrix} \bar{P}^{-\frac{1}{2}}$$

$$\chi_s = \begin{pmatrix} \frac{(x_{11} - \bar{x}_{\cdot 1})}{S_{X_1 X_1}^{1/2}} & \frac{(x_{12} - \bar{x}_{\cdot 2})}{S_{X_2 X_2}^{1/2}} & \cdots & \frac{x_{1p} - \bar{x}_{\cdot p}}{S_{X_p X_p}^{1/2}} \\ \vdots & \vdots & & \vdots \\ \frac{(x_{n1} - \bar{x}_{\cdot 1})}{S_{X_1 X_1}^{1/2}} & \frac{(x_{n2} - \bar{x}_{\cdot 2})}{S_{X_2 X_2}^{1/2}} & \cdots & \frac{x_{np} - \bar{x}_{\cdot p}}{S_{X_p X_p}^{1/2}} \end{pmatrix}$$

↑

"matriz de datos estandarizados"

$$\bar{\mathbf{x}}_{(n)}^s = \frac{1}{n} \chi_s' \mathbf{1}_n = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^n (x_{i1} - \bar{x}_{\cdot 1}) \\ \vdots \\ \sum_{i=1}^n (x_{ip} - \bar{x}_{\cdot p}) \end{pmatrix}$$

$$= \begin{pmatrix} \bar{x}_{\cdot 1} - \bar{x}_{\cdot 1} \\ \vdots \\ \bar{x}_{\cdot p} - \bar{x}_{\cdot p} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow p \times 1 \quad \dots \dots \dots \text{ (sm) }$$

$$\therefore \hat{\Sigma}^s = \frac{1}{n} \chi_s' \chi_s - \bar{\mathbf{x}}_{(n)}^s \bar{\mathbf{x}}_{(n)}^{s'} = \frac{1}{n} \chi_s' \chi_s$$

es la matriz con entradas $K_{1,l}$

$$\hat{\Sigma}_{kl}^s = \frac{1}{n} \sum_{i=1}^n \frac{(x_{ik} - \bar{x}_{\cdot k})(x_{il} - \bar{x}_{\cdot l})}{\sqrt{S_{X_k X_k}} \sqrt{S_{X_l X_l}}} \quad (= \hat{\rho}_{X_k X_l})$$

$$\hat{\Sigma}^s = \hat{\mathcal{R}} \leftarrow \text{matriz de correlaciones de } \chi$$

$$\hat{\mathcal{R}} = \mathbf{G}_R \mathbf{L}_R \mathbf{G}_R' \quad \begin{array}{l} \text{descomposición} \\ \text{de Jordan de } \hat{\mathcal{R}} \end{array}$$

$\mathbb{I}_R = \text{DIAG}(l_1^R, \dots, l_p^R)$, $l_1^R \geq \dots \geq l_p^R$ son los valores propios de $\hat{\mathcal{R}}$, con correspondientes vectores propios g_1^R, \dots, g_p^R . Como $\text{tr}(\hat{\mathcal{R}}) = \sum_{j=1}^p l_j^R$ y la matriz $\hat{\mathcal{R}}$ tiene diagonal dada por $\mathbb{1}_P = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$, entonces

$$\sum_{j=1}^p l_j^R = P.$$

Las componentes principales normalizadas son entonces (dado que $\bar{x}_{(n)}^S = 0$)

$$\tilde{z} = \chi_s g_R \quad \dots \quad (\text{CPN})$$

(comárese con la ecuación (CP), página 35).

$$\chi_s = \begin{pmatrix} \frac{(x_{11}-\bar{x}_{\cdot 1})}{S_{X_1 X_1}^{1/2}} & \frac{(x_{12}-\bar{x}_{\cdot 2})}{S_{X_2 X_2}^{1/2}} & \cdots & \frac{x_{1p}-\bar{x}_{\cdot p}}{S_{X_p X_p}^{1/2}} \\ \vdots & \vdots & & \vdots \\ \frac{(x_{n1}-\bar{x}_{\cdot 1})}{S_{X_1 X_1}^{1/2}} & \frac{(x_{n2}-\bar{x}_{\cdot 2})}{S_{X_2 X_2}^{1/2}} & \cdots & \frac{x_{np}-\bar{x}_{\cdot p}}{S_{X_p X_p}^{1/2}} \end{pmatrix}$$

↑
"matriz de datos estandarizados"

$$\bar{x}_{(n)}^s = \frac{1}{n} \chi_s' \mathbf{1}_n = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^n (x_{i1}-\bar{x}_{\cdot 1}) \\ \vdots \\ \sum_{i=1}^n (x_{ip}-\bar{x}_{\cdot p}) \end{pmatrix}$$

$$= \begin{pmatrix} \bar{x}_{\cdot 1} - \bar{x}_{\cdot 1} \\ \vdots \\ \bar{x}_{\cdot p} - \bar{x}_{\cdot p} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow p \times 1 \quad \dots \dots \dots \text{ (sm)}$$

$$\therefore \hat{\Sigma}^s = \frac{1}{n} \chi_s' \chi_s - \bar{x}_{(n)}^s \bar{x}_{(n)}'^s = \frac{1}{n} \chi_s' \chi_s$$

es la matriz con entradas $\kappa_{i,l}$

$$\hat{\Sigma}_{kl}^s = \frac{1}{n} \sum_{i=1}^n \frac{(x_{ik}-\bar{x}_{\cdot k})(x_{il}-\bar{x}_{\cdot l})}{\sqrt{S_{X_k X_k}}} \sqrt{S_{X_l X_l}} \quad (= \hat{\rho}_{X_k X_l})$$

$$\hat{\Sigma}^s = \hat{\mathcal{R}} \leftarrow \text{matriz de correlaciones de } \chi$$

$$\hat{\mathcal{R}} = \mathbf{G}_{\mathcal{R}} \mathbf{D}_{\mathcal{R}} \mathbf{G}_{\mathcal{R}}'$$

descomposición
de Jordan de $\hat{\mathcal{R}}$

$\mathbb{I}_{\hat{\mathcal{R}}} = \text{DIAG}(\hat{l}_1^{\hat{\mathcal{R}}}, \dots, \hat{l}_p^{\hat{\mathcal{R}}})$, $\hat{l}_1^{\hat{\mathcal{R}}} \geq \dots \geq \hat{l}_p^{\hat{\mathcal{R}}}$ son los valores propios de $\hat{\mathcal{R}}$, con correspondientes vectores propios $\hat{g}_1^{\hat{\mathcal{R}}}, \dots, \hat{g}_p^{\hat{\mathcal{R}}}$. Como $\text{tr}(\hat{\mathcal{R}}) = \sum_{j=1}^p \hat{l}_j^{\hat{\mathcal{R}}}$ y la matriz $\hat{\mathcal{R}}$ tiene diagonal dada por $\mathbb{1}_p = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$, entonces

$$\sum_{j=1}^p \hat{l}_j^{\hat{\mathcal{R}}} = p.$$

Las componentes principales normalizadas son entonces (dado que $\bar{x}_{(n)}^S = \mathbf{0}$)

$$\tilde{Z} = Z_S \hat{g}_{\hat{\mathcal{R}}} \quad \dots \quad (\text{CPN})$$

(comárese con la ecuación (CP), página 35).

$\tilde{Z}_{n \times p} = (z_1, \dots, z_p)$ & z_j tiene n observaciones de la columna j de la j -ésima componente principal normalizada

comárese con y_j en la ecuación (CP) página 35.

ejercicio: verifique que

$$(i) \bar{\Sigma}_{(n)}^{-1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}_{p \times 1}$$

$$\begin{aligned} (ii) \hat{\Sigma}_x &= \mathbf{G}_R' \hat{\Sigma}^s \mathbf{G}_R \\ &= \mathbf{G}_R' \hat{\mathcal{R}} \mathbf{G}_R \end{aligned}$$

A nivel poblacional las componentes principales normalizadas estarían dadas por

$$\Sigma = \mathbf{G}_R' \mathcal{D}^{1/2} (\mathbf{x} - \mathbf{\mu}), \quad \left(\leftarrow \begin{matrix} \text{dims} \\ p \times 1 \end{matrix} \right)$$

en donde $\mathcal{R} = \mathbf{G}_R \Lambda_R \mathbf{G}_R'$ es la matriz de correlaciones de \mathbf{x} con entrada i,j

$P_{ij} = \text{corr}(x_i, x_j)$; $i, j = 1, 2, \dots, P$, \mathcal{D} es una matriz diagonal y el elemento i en la diagonal es $\text{VAR}(x_i)$; $i = 1, 2, \dots, P$.

$\mathbf{x}_s \equiv \mathcal{D}^{-1/2} (\mathbf{x} - \mathbf{\mu})$ es el vector aleatorio estandarizado con componente i dada

por $\frac{x_i - \mu_i}{\sqrt{\text{VAR}(x_i)}}$; $i = 1, 2, \dots, P$.

Para estudiar la correlación entre x_i y z_j

$$\begin{aligned}
 \text{cov}(x_s, z_i) &= \text{cov} [\bar{\sigma}^{-1/2}(x - \mu), \Gamma_R^{-1} \bar{\sigma}^{1/2}(x - \mu)] \\
 &= \bar{\sigma}^{-1/2} \text{cov}[(x - \mu), (x - \mu)] \bar{\sigma}^{1/2} \Gamma_R \\
 &= \bar{\sigma}^{-1/2} \sum \bar{\sigma}^{-1/2} \Gamma_R \\
 &= \hat{\mathbf{R}} \Gamma_R = \Gamma_R \Lambda_R \Gamma_R^{-1} \Gamma_R = \Gamma_R \Lambda_R.
 \end{aligned}$$

A nivel muestral el análogo estaría dado por

$$\begin{aligned}
 (\text{A}) \quad \widehat{\text{cov}}(x_s, z_i) &\stackrel{(a)}{=} \frac{1}{n} \mathbf{x}_s' \mathbf{z} = \frac{1}{n} \mathbf{x}_s' \mathbf{x}_s \mathbf{g}_R^* \\
 &= \hat{\mathbf{R}} \mathbf{g}_R^* = \mathbf{g}_R^* \mathbf{I}_R \mathbf{g}_R^* \\
 &= \mathbf{g}_R^* \mathbf{I}_R. \quad (\text{dim } s \times p)
 \end{aligned}$$

La ecuación (A) tiene una matriz cuya entrada i,j vale $\widehat{\text{cov}}(x_i^*, z_j)$, si de ahí quisieramos obtener $\widehat{\text{corr}}(x_i^*, z_j)$ tenemos que dividir cada entrada por

(a) Ya que $\bar{\mathbf{z}}_{(n)}^T = \mathbf{0}$ (tarea i) y además de la ecuación (sm) $\bar{\mathbf{x}}_{(n)}^T = \mathbf{0}$
página 69

$\sqrt{\widehat{\text{VAR}}(x_i^s)} \cdot \sqrt{\widehat{\text{VAR}}(z_j)}$, pero como
 $\widehat{\Sigma}^s = \widehat{R}$ y esta última matriz tiene
 1's sobre su diagonal

$$\widehat{\text{VAR}}(x_i^s) = 1.$$

Para calcular $\widehat{\text{VAR}}(z_j)$, dada la note (a)
 al pie de la página 72

$$\begin{aligned}\widehat{\Sigma}_Z &= \frac{1}{n} \widehat{Z}^T \widehat{Z} = \frac{1}{n} \mathbf{G}_{\widehat{R}}^T \mathbf{X}_s^T \mathbf{X}_s \mathbf{G}_{\widehat{R}} \\ &= \mathbf{G}_{\widehat{R}}^T \left(\frac{1}{n} \mathbf{X}_s^T \mathbf{X}_s \right) \mathbf{G}_{\widehat{R}} \\ &= \mathbf{G}_{\widehat{R}}^T \widehat{R} \mathbf{G}_{\widehat{R}} \\ &= \mathbf{I}_{\widehat{R}}^T \mathbf{I}_{\widehat{R}} \widehat{R} \mathbf{I}_{\widehat{R}}^T \mathbf{I}_{\widehat{R}} \mathbf{G}_{\widehat{R}} \\ &= \mathbf{I}_{\widehat{R}},\end{aligned}$$

es decir $\widehat{\text{cov}}(z_i, z_j) = 0 \quad \forall i \neq j$ y

$$\widehat{\text{VAR}}(z_j) = \lambda_j^{\widehat{R}}$$

Por tanto la matriz de correlación entre \mathbf{x}_s y \mathbf{z} , es

$$\hat{\mathcal{R}}_{\mathbf{x}_s, \mathbf{z}} = \mathbf{g}_s^T \mathbf{I}_s^{-1/2} \mathbf{g}_s^{1/2} = \mathbf{g}_s^T \mathbf{I}_s^{1/2}$$

Se puede probar que las correlaciones entre las variables originales x_i y las componentes principales normalizadas z_j , están dadas por

$$r_{x_i z_j} = \sqrt{l_j^{\hat{\mathcal{R}}}} g_{ij}^{\hat{\mathcal{R}}} \quad \dots \text{ (cor n.p.),}$$

(comprérese con $r_{x_i y_j}$ en la página 45).

$$\sum_{j=1}^p l_j^{\hat{\mathcal{R}}} (g_{ij}^{\hat{\mathcal{R}}})^2 = (g_i^{\hat{\mathcal{R}}})^T \mathbf{I}_{\hat{\mathcal{R}}} g_i^{\hat{\mathcal{R}}} = \text{elemento } (i,i)$$

de la matriz $\mathbf{g}_s^T \mathbf{I}_s^{-1/2} \mathbf{g}_s^{1/2} = \hat{\mathcal{R}}$, la cual tiene 1's en la diagonal, por tanto

$$\sum_{j=1}^p r_{x_i z_j}^2 = \sum_{j=1}^p l_j^{\hat{\mathcal{R}}} (g_{ij}^{\hat{\mathcal{R}}})^2 = 1.$$

Nuevamente podemos interpretar $r_{x_i z_j}^2$ como

75

"la proporción de la varianza de X_i explicada por la j-ésima componente principal Z_j ".

ejercicio: pruebe la ecuación (cor npc).

"la proporción de la varianza de X_i explicada por la j-ésima componente principal Z_j ".

Ejercicio: pruebe la ecuación (cor NPC).

La figura F muestra gráficos de las componentes principales calculadas a partir de los datos estandarizados. Para esta figura se usa el carácter "o" para los billetes verdaderos y el carácter "+" para los billetes falsos.

El vector de valores propios de \hat{R} es

$$\lambda^{\hat{R}} = (2.946, 1.278, 0.869, 0.450, 0.269, 0.189)$$

El panel inferior izquierdo en la figura F muestra una gráfica de las parejas ordenadas $(i, \lambda_i^{\hat{R}}) ; i=1, 2, \dots, 6$.

Podemos calcular la proporción de la varianza explicada por la componente i , al dividir

$$\lambda_i^{\hat{R}} \text{ por } \sum_{j=1}^P \lambda_j^{\hat{R}} = P$$

$$\hat{P^R} = \left(\frac{2.946}{6}, \frac{1.278}{6}, \frac{0.869}{6}, \frac{0.450}{6}, \frac{0.269}{6}, \frac{0.189}{6} \right)^T$$

$$= (0.491, 0.213, 0.145, 0.075, 0.045, 0.032).$$

Si $\hat{\Psi}_q^R = \frac{\sum_{j=1}^q l_j \hat{x}}{\sum_{j=1}^6 l_j \hat{x}}$ es la proporción

de la variabilidad explicada por las primeras q componentes (normalizadas), entonces el vector $\hat{\Psi}^R = (\hat{\Psi}_1^R, \hat{\Psi}_2^R, \dots, \hat{\Psi}_6^R)$

$$\text{es } \hat{\Psi}^R = (0.491, 0.704, 0.849, 0.924, 0.969, 1.0)$$

Entonces, la primera componente principal (normalizada) explica 49% de la variabilidad y las primeras dos componentes explican 70.4% de la variabilidad en los datos.

La matriz \hat{W}^R de vectores propios está dada por

$$\mathbf{g}_R = \begin{pmatrix} -0.007 & 0.815 & -0.018 & 0.575 & -0.059 & -0.031 \\ 0.468 & 0.342 & 0.103 & -0.395 & 0.639 & 0.298 \\ 0.487 & 0.252 & 0.123 & -0.430 & -0.614 & -0.349 \\ 0.407 & -0.266 & 0.584 & 0.404 & -0.215 & 0.462 \\ 0.368 & -0.091 & -0.788 & 0.110 & -0.220 & 0.419 \\ -0.493 & 0.274 & 0.114 & -0.392 & -0.340 & 0.632 \end{pmatrix}$$

$$Z_1 = -0.007x_1 + 0.468x_2 + 0.487x_3 + 0.407x_4 \\ + 0.368x_5 - 0.493x_6$$

$$Z_2 = 0.815x_1 + 0.342x_2 + 0.252x_3 - 0.266x_4 \\ - 0.091x_5 + 0.274x_6$$

Notemos que los pesos asignados a cada variable en estas combinaciones lineales son ahora de "magnitud similar".

La figura G muestra los perfiles ordenados $(r_{x_i z_1}, r_{x_i z_2})$; $i=1, 2, \dots, P$, los valores de estas correlaciones están en la siguiente tabla

variable	$r_{X_iz_1}$	$r_{X_iz_2}$	$r_{X_iz_1}^2 + r_{X_iz_2}^2$
X_1 longitud	-0.012	0.922	0.850
X_2 ancho (izq.)	0.803	0.387	0.794
X_3 ancho (der.)	0.835	0.285	0.78
X_4 long. borde inferior	0.698	-0.301	0.58
X_5 long. borde superior	0.631	-0.104	0.41
X_6 long. diag	-0.847	0.309	0.81

Observando la ecuación para Z_1 (los magnitudes y signos de los pesos), así como la columna $r_{X_iz_1}$ de la tabla, notamos que Z_1 queda descrita por la diferencia entre: la suma de los anchos (izquierdo y derecho) y la longitud de la diagonal de los billetes. Usando la ecuación para Z_2 y la columna $r_{X_iz_2}$, notamos que Z_2 está dominada por la longitud de los billetes.

Ahora nos concentraremos en la figura F, tomando las anteriores observaciones en cuenta. Vemos que para los billetes verdaderos (con carácter "o") se tienen longitudes de su diagonal que son más grandes y valores de longitud del ancho (izquierdo y derecho) más pequeños⁽¹⁾. Para los billetes falsos, se tienen valores de la longitud de la diagonal más pequeños y valores de longitud del ancho (izquierdo y derecho) más grandes⁽²⁾

Con respecto a la longitud de los billetes X_1 , esta tiende a ser más grande para los billetes verdaderos⁽³⁾

(1) $Z_1 < 0$ (2) $Z_1 > 0$ (3) $Z_2 > 0$