Capítulo 5

ANALISIS DE COMPONENTES PRINCIPALES

5.1 Definición y obtención de las componentes principales

Sea $\mathbf{X} = [X_1, \ldots, X_p]$ una matriz de datos multivariantes. Lo que sigue también vale si $\mathbf{X}$ es un vector formado por p variables observables.

Las componentes principales son unas variables compuestas incorrelacionadas tales que unas pocas explican la mayor parte de la variabilidad de $\mathbf{X}$.

Definició 5.1.1 *Las componentes principales son las variables compuestas*

$$Y_1 = \mathbf{X}\mathbf{t}_1, Y_2 = \mathbf{X}\mathbf{t}_2, \ldots, Y_p = \mathbf{X}\mathbf{t}_p$$

tales que:

1. *var*(Y_1) *es máxima condicionado a* $\mathbf{t}_1'\mathbf{t}_1 = 1$.
2. *Entre todas las variables compuestas* Y *tales que* $cov(Y_1, Y) = 0$, *la variable* Y_2 *es tal que var*(Y_2) *es máxima condicionado a* $\mathbf{t}_2'\mathbf{t}_2 = 1$.
3. Y_3 *es una variable incorrelacionada con* Y_1, Y_2 *con varianza máxima. Análogamente definimos las demás componentes principales.*

Si $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_p]$ es la matriz $p \times p$ cuyas columnas son los vectores que definen las componentes principales, entonces la transformación lineal $\mathbf{X} \to \mathbf{Y}$

$$\mathbf{Y} = \mathbf{XT} \tag{5.1}$$

se llama transformación por componentes principales.

Teorema 5.1.1 *Sean* $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_p$ *los* p *vectores propios normalizados de la matriz de covarianzas* $\mathbf{S}$, *es decir,*

$$\mathbf{St}_i = \lambda_i \mathbf{t}_i, \quad \mathbf{t}_i'\mathbf{t}_i = 1, \quad i = 1, \ldots, p.$$

Entonces:

1. *Las variables compuestas* $Y_i = \mathbf{Xt}_i$, $i = 1, \ldots, p$, *son las componentes principales.*

2. *Las varianzas son los valores propios de* $\mathbf{S}$

$$\mathrm{var}(Y_i) = \lambda_i, \quad i = 1, \ldots, p.$$

3. *Las componentes principales son variables incorrelacionadas:*

$$\mathrm{cov}(Y_i, Y_j) = 0, \quad i \neq j = 1, \ldots, p.$$

Demost.: Supongamos $\lambda_1 > \cdots > \lambda_p > 0$. Probemos que las variables $Y_i = \mathbf{Xt}_i$, $i = 1, \ldots, p$, son incorrelacionadas:

$$\begin{array}{lll} \mathrm{cov}(Y_i, Y_j) & = \mathbf{t}_i'\mathbf{St}_j & = \mathbf{t}_i'\lambda_j\mathbf{t}_j \quad = \lambda_j\mathbf{t}_i'\mathbf{t}_j, \\ \mathrm{cov}(Y_j, Y_i) & = \mathbf{t}_j'\mathbf{St}_i & = \mathbf{t}_j'\lambda_j\mathbf{t}_i \quad = \lambda_i\mathbf{t}_j'\mathbf{t}_i, \end{array}$$

$\Rightarrow (\lambda_j - \lambda_i)\mathbf{t}_i'\mathbf{t}_j = 0, \Rightarrow \mathbf{t}_i'\mathbf{t}_j = 0, \Rightarrow \mathrm{cov}(Y_i, Y_j) = \lambda_j\mathbf{t}_i'\mathbf{t}_j = 0$, si $i \neq j$.
Además:

$$\mathrm{var}(Y_i) = \lambda_i\mathbf{t}_i'\mathbf{t}_j = \lambda_i.$$

Sea ahora $Y = \sum_{i=1}^p a_i X_i = \sum_{i=1}^p \alpha_i Y_i$ una variable compuesta tal que $\sum_{i=1}^p \alpha_i^2 = 1$. Entonces

$$\mathrm{var}(Y) = \mathrm{var}(\sum_{i=1}^p \alpha_i Y_i) = \sum_{i=1}^p \alpha_i^2 \mathrm{var}(Y_i) = \sum_{i=1}^p \alpha_i^2\lambda_i \leq (\sum_{i=1}^p \alpha_i^2)\lambda_1 = \mathrm{var}(Y_1),$$

que prueba que Y_1 tiene varianza máxima.

Consideremos ahora las variables Y incorrelacionadas con Y_1. Las podemos expresar como:

$$Y = \sum_{i=1}^{p} b_i X_i = \sum_{i=2}^{p} \beta_i Y_i \text{ condicionado a } \sum_{i=2}^{p} \beta_i^2 = 1.$$

Entonces:

$$\text{var}(Y) = \text{var}(\sum_{i=2}^{p} \beta_i Y_i) = \sum_{i=2}^{p} \beta_i^2 \text{var}(Y_i) = \sum_{i=2}^{p} \beta_i^2 \lambda_i \leq (\sum_{i=2}^{p} \beta_i^2)\lambda_2 = \text{var}(Y_2),$$

y por lo tanto Y_2 está incorrelacionada con Y_1 y tiene varianza máxima. Si $p \geq 3$, la demostración de que $Y_3, \ldots, Y_p$ son también componentes principales es análoga.□

5.2 Variabilidad explicada por las componentes principales

La varianza de la componente principal Y_i es $\text{var}(Y_i) = \lambda_i$ y la variación total es $\text{tr}(\mathbf{S}) = \sum_{i=1}^{p} \lambda_i$. Por lo tanto:

1. Y_i contribuye con la cantidad λ_i a la variación total $\text{tr}(\mathbf{S})$.

2. Si $q < p$, $Y_1, \ldots, Y_q$ contribuyen con la cantidad $\sum_{i=1}^{q} \lambda_i$ a la variación total $\text{tr}(\mathbf{S})$.

3. El porcentaje de variabilidad explicada por las m primeras componentes principales es
$$P_m = 100 \frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_p}. \tag{5.2}$$

En las aplicaciones cabe esperar que las primeras componentes expliquen un elevado porcentaje de la variabilidad total. Por ejemplo, si $m = 2 < p$, y $P_2 = 90\%$, las dos primeras componentes explican una gran parte de la variabilidad de las variables. Entonces podremos sustituir $X_1, X_2, \ldots, X_p$ por las componentes principales Y_1, Y_2. En muchas aplicaciones, tales componentes tienen interpretación experimental.

5.3 Representación de una matriz de datos

Sea $\mathbf{X} = [X_1, \ldots, X_p]$ una matriz $n \times p$ de datos multivariantes. Queremos representar, en un espacio de dimensión reducida m (por ejemplo, $m = 2$), las filas $\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n'$ de $\mathbf{X}$. Necesitamos introducir una distancia (ver Sección 1.9).

Definició 5.3.1 *La distancia euclídea (al cuadrado) entre dos filas de* $\mathbf{X}$

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{ip}), \quad \mathbf{x}_j = (x_{j1}, \ldots, x_{jp}),$$

es

$$\delta_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j) = \sum_{h=1}^{p} (x_{ih} - x_{jh})^2.$$

La matriz $\Delta = (\delta_{ij})$ *es la matriz* $n \times n$ *de distancias entre las filas.*

Podemos representar las n filas de $\mathbf{X}$ como n puntos en el espacio R^p distanciados de acuerdo con la métrica δ_{ij}. Pero si p es grande, esta representación no se puede visualizar. Necesitamos reducir la dimensión.

Definició 5.3.2 *La variabilidad geométrica de la matriz de distancias* Δ *es la media de sus elementos al cuadrado*

$$V_\delta(\mathbf{X}) = \frac{1}{2n^2} \sum_{i,j=1}^{n} \delta_{ij}^2.$$

Si $\mathbf{Y} = \mathbf{XT}$ *es una transformación lineal de* $\mathbf{X}$*, donde* $\mathbf{T}$ *es una matriz* $p \times q$ *de constantes,*

$$\delta_{ij}^2(q) = (\mathbf{y}_i - \mathbf{y}_j)'(\mathbf{y}_i - \mathbf{y}_j) = \sum_{h=1}^{q} (y_{ih} - y_{jh})^2$$

es la distancia euclídea entre dos filas de $\mathbf{Y}$*. La variabilidad geométrica en dimensión* $q \leq p$ *es*

$$V_\delta(\mathbf{Y})_q = \frac{1}{2n^2} \sum_{i,j=1}^{n} \delta_{ij}^2(q).$$

Teorema 5.3.1 *La variabilidad geométrica de la distancia euclídea es la traza de la matriz de covarianzas*

$$V_\delta(\mathbf{X}) = tr(\mathbf{S}) = \sum_{h=1}^{p} \lambda_h.$$

Demost.: Si $x_1, \ldots, x_n$ es una muestra univariante con varianza s^2, entonces

$$\frac{1}{2n^2} \sum_{i,j=1}^{n} (x_i - x_j)^2 = s^2. \tag{5.3}$$

En efecto, si $\overline{x}$ es la media

$$\begin{aligned} \tfrac{1}{n^2}\textstyle\sum_{i,j=1}^{n}(x_i - x_j)^2 &= \tfrac{1}{n^2}\textstyle\sum_{i,j=1}^{n}(x_i - \overline{x} - (x_j - \overline{x}))^2 \\ &= \tfrac{1}{n^2}\textstyle\sum_{i,j=1}^{n}(x_i - \overline{x})^2 + \tfrac{1}{n^2}\textstyle\sum_{i,j=1}^{n}(x_j - \overline{x})^2 \\ &\quad + \tfrac{2}{n^2}\textstyle\sum_{i,j=1}^{n}(x_i - \overline{x})(x_j - \overline{x}))^2 \\ &= \tfrac{1}{n} n s^2 + \tfrac{1}{n} n s^2 + 0 = 2s^2. \end{aligned}$$

Aplicando (5.3) a cada columna de $\mathbf{X}$ y sumando obtenemos

$$V_\delta(\mathbf{X}) = \sum_{j=1}^{p} s_{jj} = \mathrm{tr}(\mathbf{S}).\square$$

Una buena representación en dimensión reducida q (por ejemplo, $q = 2$) será aquella que tenga máxima variabilidad geométrica, a fin de que los puntos estén lo más separados posible.

Teorema 5.3.2 *La transformación lineal* $\mathbf{T}$ *que maximiza la variabilidad geométrica en dimensión* q *es la transformación por componentes principales (5.1), es decir,* $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_q]$ *contiene los* q *primeros vectores propios normalizados de* $\mathbf{S}$.

Demost.: Aplicando (5.3), la variabilidad geométrica de $\mathbf{Y} = \mathbf{XT}$, donde $\mathbf{T}$ es cualquiera, es

$$V_\delta(\mathbf{Y})_q = \sum_{j=1}^{p} s^2(Y_j) = \sum_{j=1}^{p} \mathbf{t}_j' \mathbf{S} \mathbf{t}_j,$$

siendo $s^2(Y_j) = \mathbf{t}_j' \mathbf{S} \mathbf{t}_j$ la varianza de la variable compuesta Y_j. Alcanzamos la máxima varianza cuando Y_j es una componente principal: $s^2(Y_j) \leq \lambda_j$. Así:

$$\max V_\delta(\mathbf{Y})_q = \sum_{j=1}^{p} \lambda_j . \square$$

El porcentaje de variabilidad geométrica explicada por $\mathbf{Y}$ es

$$P_q = 100 \frac{V_\delta(\mathbf{Y})_q}{V_\delta(\mathbf{X})_p} = 100 \frac{\lambda_1 + \cdots + \lambda_q}{\lambda_1 + \cdots + \lambda_p}.$$

Supongamos ahora $q = 2$. Si aplicamos la transformación (5.1), la matriz de datos $\mathbf{X}$ se reduce a

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} \\ \vdots & \vdots \\ y_{i1} & y_{i2} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{pmatrix}.$$

Entonces, representando los puntos de coordenadas $(y_{i1}, y_{i2}), i = 1, \ldots, n$, obtenemos una representación óptima en dimensión 2 de las filas de $\mathbf{X}$.

5.4 Inferencia

Hemos planteado el ACP sobre la matriz $\mathbf{S}$, pero lo podemos también plantear sobre la matriz de covarianzas poblacionales Σ. Las componentes principales obtenidas sobre $\mathbf{S}$ son, en realidad, estimaciones de las componentes principales sobre Σ.

Sea $\mathbf{X}$ matriz de datos $n \times p$ donde las filas son independientes con distribución $N_p(\mu, \Sigma)$. Recordemos que:

1. $\overline{\mathbf{x}}$ es $N_p(\mu, \Sigma/n)$.

2. $\mathbf{U} = n\mathbf{S}$ es Wishart $W_p(\Sigma, n-1)$.

3. $\overline{\mathbf{x}}$ y $\mathbf{S}$ son estocásticamente independientes.

Sea $\Sigma = \Gamma\Lambda\Gamma'$ la diagonalización de Σ. Indiquemos

$$\Gamma = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p], \quad \boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_p], \quad \Lambda = diag(\lambda_1, \ldots, \lambda_p),$$

los vectores propios y valores propios de Σ. Por otra parte, sea $\mathbf{S} = \mathbf{GLG}'$ la diagonalización de $\mathbf{S}$. Indiquemos:

$$\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_p], \quad \mathbf{l} = [l_1, \ldots, l_p], \quad \mathbf{L} = \text{diag}(l_1, \ldots, l_p)$$

los vectores propios y valores propios de $\mathbf{S}$. A partir de ahora supondremos

$$\lambda_1 \geq \ldots \geq \lambda_p.$$

5.4.1 Estimación y distribución asintótica

Teorema 5.4.1 *Se verifica:*

1. *Si los valores propios son diferentes, los valores y vectores propios obtenidos a partir de* $\mathbf{S}$ *son estimadores máximo-verosímiles de los obtenidos a partir de* Σ

$$\widehat{\lambda}_i = l_i, \quad \widehat{\boldsymbol{\gamma}}_i = \mathbf{g}_i \quad , i = 1, \ldots, p.$$

2. *Cuando* $k > 1$ *valores propios son iguales a* λ

$$\lambda_1 > \ldots > \lambda_{p-k} = \lambda_{p-k+1} = \ldots = \lambda_p = \lambda,$$

el estimador máximo verosímil de λ *es la media de los correspondientes valores propios de* $\mathbf{S}$

$$\widehat{\lambda} = (l_{p-k+1} + \ldots + l_p)/k$$

Demost.: Los valores y vectores propios están biunívocamente relacionados con Σ y por lo tanto 1) es consecuencia de la propiedad de invariancia de la estimación máximo verosímil. La demostración de 2) se encuentra en Anderson (1959).□

Teorema 5.4.2 *Los vectores propios* $[\mathbf{g}_1, \ldots, \mathbf{g}_p]$ *y valores propios* $\mathbf{l} = [l_1, \ldots, l_p]$ *verifican asintóticamente:*

1. $\mathbf{l}$ *es* $N_p(\boldsymbol{\lambda}, 2\Lambda^2/n)$. *En particular:*

$$l_i \quad \text{es} \quad N(\lambda_i, 2\lambda_i^2/n), \qquad cov(l_i, l_j) = 0, \quad i \neq j,$$

es decir, l_i, l_j *son normales e independientes.*

2. $\mathbf{g}_i$ *es* $N_p(\boldsymbol{\gamma}_i, \mathbf{V}_i/n)$ *donde*

$$\mathbf{V}_i = \lambda_i \sum_{j \neq i} \frac{\lambda_i}{(\lambda_i - \lambda_j)^2} \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i'$$

3. $\mathbf{l}$ *es independiente de* $\mathbf{G}$.

Demost.: Anderson (1959), Mardia, Kent y Bibby (1979).□

Como consecuencia de que l_i es $N(\lambda_i, 2\lambda_i^2/n)$, obtenemos el intervalo de confianza asintótico con coeficiente de confianza $1 - \alpha$

$$\frac{l_i}{(1 + az_{\alpha/2})^{1/2}} < \lambda_i < \frac{l_i}{(1 - az_{\alpha/2})^{1/2}}$$

siendo $a^2 = 2/(n-1)$ y $P(|Z| > z_{\alpha/2}) = \alpha/2$, donde Z es $N(0,1)$.

Se obtiene otro intervalo de confianza como consecuencia de que $\log l_i$ es $N(\log \lambda_i, 2/(n-1))$

$$l_i e^{-az_{\alpha/2}} < \lambda_i < l_i e^{+az_{\alpha/2}}.$$

5.4.2 Tests de hipótesis

Determinados tests de hipótesis relativos a las componentes principales son casos particulares de un test sobre la estructura de la matriz Σ.

A. Supongamos que queremos decidir si la matriz Σ es igual a una matriz determinada Σ_0. Sea $\mathbf{X}$ un matriz $n \times p$ con filas independientes $N_p(\mu, \Sigma)$. El test es:

$$H_0 : \Sigma = \Sigma_0 \quad (\mu \quad \text{desconocida})$$

Si L es la verosimilitud de la muestra, el máximo de $\log L$ bajo H_o es

$$\log L_0 = -\frac{n}{2} \log |2\pi\Sigma_0| - \frac{n}{2} tr(\Sigma_0^{-1}\mathbf{S}).$$

El máximo no restringido es

$$\log L = -\frac{n}{2} \log |2\pi\mathbf{S}| - \frac{n}{2} p.$$

El estadístico basado en la razón de verosimilitud λ_R es

$$\begin{aligned} -2\log\lambda_R &= 2(\log L - \log L_0) \\ &= n\text{tra}(\Sigma_0^{-1}\mathbf{S}) - n\log|\Sigma_0^{-1}\mathbf{S}| - np. \end{aligned} \tag{5.4}$$

Si $L_1, \ldots, L_p$ son los valores propios de $\Sigma_0^{-1}\mathbf{S}$ y a, g son las medias aritmética y geométrica

$$a = (L_1 + \ldots + L_p)/p, \quad g = (L_1 \times \ldots \times L_p)^{1/p}, \tag{5.5}$$

entonces, asintóticamente

$$-2\log\lambda_R = np(a - \log g - 1) \sim \chi^2_q, \tag{5.6}$$

siendo $q = p(p+1)/2 - \text{par}(\Sigma_0)$ el número de parámetros libres de Σ menos el número de parámetros libres de Σ_0.

B. Test de independencia completa.

Si la hipótesis nula afirma que las p variables son estocásticamente independientes, el test se formula como

$$H_0 : \Sigma = \Sigma_d = \text{diag}(\sigma_{11}, \cdots, \sigma_{pp}) \quad (\mu \quad \text{desconocida}).$$

Bajo H_0 la estimación de Σ_d es $\mathbf{S}_d = \text{diag}(s_{11}, \cdots, s_{pp})$ y $\mathbf{S}_d^{-1}\mathbf{S} = \mathbf{R}$ es la matriz de correlaciones. De (5.4) y de $\log|2\pi\mathbf{S}_d| - \log|2\pi\mathbf{S}| = \log|\mathbf{R}|$, $\text{tra}(\mathbf{R}) = p$, obtenemos

$$-2\log\lambda_R = -n\log|\mathbf{R}| \sim \chi^2_q$$

siendo $q = p(p+1)/2 - p = p(p-1)/2$. Si el estadístico $-n\log|\mathbf{R}|$ no es significativo, entonces podemos aceptar que las variables son incorrelacionadas y por lo tanto, como hay normalidad multivariante, independientes.

C. Test de igualdad de valores propios.

Este es un test importante en ACP. La hipótesis nula es

$$H_0 : \lambda_1 > \ldots > \lambda_{p-k} = \lambda_{p-k+1} = \ldots = \lambda_p = \lambda.$$

Indicamos los valores propios de $\mathbf{S}$ y de $\mathbf{S}_0$ (estimación de Σ si H_0 es cierta)

$$\mathbf{S} \sim (l_1, \ldots, l_k, l_{k+1}, \ldots, l_p), \quad \mathbf{S}_0 \sim (l_1, \ldots, l_k, a_0, \ldots, a_0),$$

donde $a_0 = (l_{k+1} + \ldots + l_p)/(p-k)$ (Teorema 5.4.1). Entonces

$$\mathbf{S}_0^{-1}\mathbf{S} \sim (1, \ldots, 1, l_{k+1}/a_0, \ldots, l_p/a_0),$$

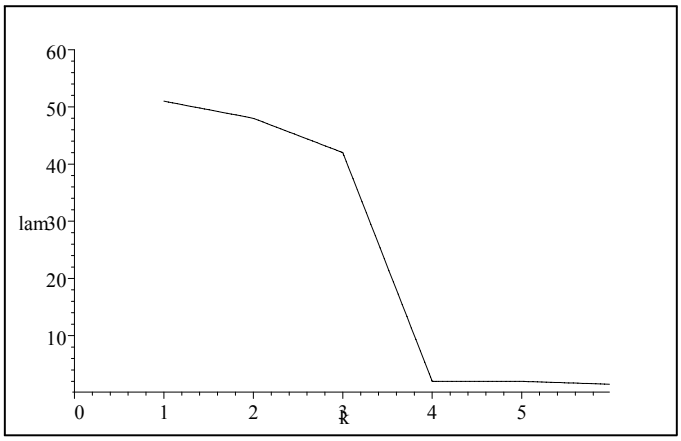

Figura 5.1: Ejemplo de representación de los valores propios, que indicaría 3 componentes principales.

las medias (5.5) son $a = 1$ y $g = (l_{k+1} \times \ldots \times l_p)^{1/p} a_0^{(k-p)/p}$ y aplicando (5.6)

$$-2 \log \lambda_R = n(p-k) \log(l_{k+1} + \ldots + l_p)/(p-k) - n(\sum_{i=k+1}^{p} \log l_i) \sim \chi_q^2, \quad (5.7)$$

donde $q = (p-k)(p-k+1)/2 - 1$.

5.5 Número de componentes principales

En esta sección presentamos algunos criterios para determinar el número $m < p$ de componentes principales.

5.5.1 Criterio del porcentaje

El número m de componentes principales se toma de modo que P_m sea próximo a un valor especificado por el usuario, por ejemplo el 80%. Por otra parte, si la representación de $P_1, P_2, \ldots, P_k, \ldots$ con respecto de k prácticamente se estabiliza a partir de un cierto m, entonces aumentar la dimensión apenas aporta más variabilidad explicada.

5.5.2 Criterio de Kaiser

Obtener las componentes principales a partir de la matriz de correlaciones $\mathbf{R}$ equivale a suponer que las variables observables tengan varianza 1. Por lo tanto una componente principal con varianza inferior a 1 explica menos variabilidad que una variable observable. El criterio, llamado de Kaiser, es entonces:

Retenemos las m primeras componentes tales que $\lambda_m \geq 1$,
donde $\lambda_1 \geq \ldots \geq \lambda_p$ son los valores propios de $\mathbf{R}$, que también son las varianzas de las componentes. Estudios de Montecarlo prueban que es más correcto el punto de corte $\lambda^* = 0.7$, que es más pequeño que 1.

Este criterio se puede extender a la matriz de covarianzas. Por ejemplo, m podría ser tal que $\lambda_m \geq v$, donde $v = \text{tra}(\mathbf{S})/p$ es la media de las varianzas. También es aconsejable considerar el punto de corte $0.7 \times v$.

5.5.3 Test de esfericidad

Supongamos que la matriz de datos proviene de una población normal multivariante $N_p(\mu, \Sigma)$. Si la hipótesis

$$H_0^{(m)} : \lambda_1 > \ldots > \lambda_m > \lambda_{m+1} = \ldots = \lambda_p$$

es cierta, no tiene sentido considerar más de m componentes principales. En efecto, no hay direcciones de máxima variabilidad a partir de m, es decir, la distribución de los datos es esférica. El test para decidir sobre $H_0^{(m)}$ está basado en el estadístico ji-cuadrado (5.7) y se aplica secuencialmente: Si aceptamos $H_0^{(0)}$ no hay direcciones principales, pero si rechazamos $H_0^{(0)}$, entonces repetimos el test con $H_0^{(1)}$. Si aceptamos $H_0^{(1)}$ entonces $m = 1$, pero si rechazamos $H_0^{(1)}$ repetimos el test con $H_0^{(2)}$, y así sucesivamente. Por ejemplo, si $p = 4$, tendríamos que $m = 2$ si rechazamos $H_0^{(0)}, H_0^{(1)}$ y aceptamos $H_0^{(2)} : \lambda_1 > \lambda_2 > \lambda_3 = \lambda_4$.

5.5.4 Criterio del bastón roto

Los valores propios suman $V_t = \text{tr}(\mathbf{S})$, que es la variabilidad total. Imaginemos un bastón de longitud V_t, que rompemos en p trozos al azar (asignando $p-1$ puntos uniformemente sobre el intervalo $(0, V_t)$) y que los trozos ordenados

son los valores propios $l_1 > l_2 > \ldots > l_p$. Si normalizamos a $V_t = 100$, entonces el valor esperado de l_j es

$$E(L_j) = 100 \times \frac{1}{p} \sum_{i=1}^{p-j} \frac{1}{j+i}.$$

Las m primeras componentes son significativas si el porcentaje de varianza explicada supera claramente el valor de $E(L_1) + \ldots + E(L_m)$. Por ejemplo, si $p = 4$, los valores son:

Porcentaje	$E(L_1)$	$E(L_2)$	$E(L_3)$	$E(L_4)$
Esperado	52.08	27.08	14.58	6.25
Acumulado	52.08	79.16	93.74	100

Si $V_2 = 93.92$ pero $V_3 = 97.15$, entonces tomaremos sólo dos componentes.

5.5.5 Un ejemplo

Exemple 5.5.1

Sobre una muestra de $n = 100$ estudiantes de Bioestadística, se midieron las variables

X_1 = peso (kg), X_2 =talla (cm.), X_3 =ancho hombros (cm.), X_4 = ancho caderas (cm.),

con los siguientes resultados:

1. medias: $\overline{x}_1 = 54.25, \overline{x}_2 = 161.73, \overline{x}_3 = 36.53, \overline{x}_4 = 30.1$.

2. matriz de covarianzas:

$$\mathbf{S} = \begin{pmatrix} 44.7 & 17.79 & 5.99 & 9.19 \\ 17.79 & 26.15 & 4.52 & 4.44 \\ 5.99 & 4.52 & 3.33 & 1.34 \\ 9.19 & 4.44 & 1.34 & 4.56 \end{pmatrix}$$

3. vectores y valores propios (columnas):

	$\mathbf{t}_1$	$\mathbf{t}_2$	$\mathbf{t}_3$	$\mathbf{t}_4$
	.8328	.5095	.1882	.1063
	.5029	−.8552	.0202	.1232
	.1362	−.0588	.1114	−.9826
	.1867	.0738	−.9755	−.0892
Val. prop.	58.49	15.47	2.54	2.24
Porc. acum.	74.27	93.92	97.15	100

4. Número de componentes:

   a. Criterio de Kaiser: la media de las varianzas es $v = \text{tr}(\mathbf{S})/p = 19.68$. Los dos primeros valores propios son 58.49 y 15.47, que son mayores que $0.7 \times v$. Aceptamos $m = 2$.

   b. Test de esfericidad.

m	χ^2	g.l.
0	333.9	9
1	123.8	5
2	0.39	2

   Rechazamos $m = 0, m = 1$ y aceptamos $m = 2$.

   c. Test del bastón roto: Puesto que $P_2 = 93.92$ supera claramente el valor esperado 79.16 y que no ocurre lo mismo con P_3, aceptamos $m = 2$.

5. Componentes principales:

$$Y_1 = .8328X_1 + .5029X_2 + .1362X_3 + .1867X_4,$$
$$Y_2 = .5095X_1 - .8552X_2 - .0588X_3 + .0738X_4.$$

6. Interpretación: la primera componente es la variable con máxima varianza y tiene todos sus coeficientes positivos. La interpretamos como una componente de *tamaño*. La segunda componente tiene coeficientes positivos en la primera y cuarta variable y negativos en las otras dos. La interpretamos como una componente de *forma*. La primera componente ordena las estudiantes según su tamaño, de la más pequeña a la más grande, y la segunda según la forma, el tipo pícnico en contraste con el tipo atlético. Las dimensiones de tamaño y forma están incorrelacionadas.

5.6 Complementos

El Análisis de Componentes Principales (ACP) fué iniciado por K. Pearson en 1901 y desarrollado por H. Hotelling en 1933. Es un método referente a una población, pero W. Krzanowski y B. Flury han investigado las componentes principales comunes a varias poblaciones.

El ACP tiene muchas aplicaciones. Una aplicación clásica es el estudio de P. Jolicoeur y J. E. Mosimann sobre tamaño y forma de animales, en términos de la primera, segunda y siguientes componentes principales. La primera componente permite ordenar los animales de más pequeños a más grandes, y la segunda permite estudiar su variabilidad en cuanto a la forma. Nótese que tamaño y forma son conceptos "independientes".

El ACP puede servir para estudiar la capacidad. Supongamos que la caparazón de una tortuga tiene longitud L, ancho A, y alto H. La capacidad sería $C = L^\alpha A^\beta H^\gamma$, donde α, β, γ son parámetros. Aplicando logaritmos, obtenemos

$$\log C = \alpha \log L + \beta \log A + \gamma \log H = \log(L^\alpha A^\beta H^\gamma),$$

que podemos interpretar como la primera componente principal Y_1 de las variables $\log L, \log A, \log H$, y por tanto α, β, γ serían los coeficientes de Y_1.

Por medio del ACP es posible efectuar una regresión múltiple de Y sobre $X_1, \ldots, X_p$, considerando las primeras componentes principales $Y_1, Y_2, \ldots$ como variables explicativas, y realizar regresión de Y sobre $Y_1, Y_2, \ldots$, evitando así efectos de colinealidad, aunque las últimas componentes principales también pueden influir (Cuadras, 1993). La regresión ortogonal es una variante interesante. Supongamos que se quieren relacionar las variables $X_1, \ldots, X_p$ (todas con media 0), en el sentido de encontrar los coeficientes $\beta_1, \ldots, \beta_p$ tales que $\beta_1 X_1 + \ldots + \beta_p X_p \cong 0$. Se puede plantear el problema como $\mathrm{var}(\beta_1 X_1 + \ldots + \beta_p X_p) =$mínima, condicionado a $\beta_1^2 + \ldots + \beta_p^2 = 1$. Es fácil ver que la solución es la última componente principal Y_p.

Se pueden definir las componentes principales de un proceso estocástico y de una variable aleatoria. Cuadras y Fortiana (1995), Cuadras y Lahlou (2000) han estudiado las componentes principales de las variables uniforme, exponencial y logística.