# Análisis de Componentes Principales

Ramón Gutiérrez Sánchez.
Universidad de Granada

# Índice general

# Tema 1

# El modelo de Análisis de Componentes Principales (ACP).

Las componentes principales (CP) asociadas a un vector de variables aleatorias $X = (X_1, \ldots, X_n)'$, son combinaciones lineales de dichas variables sometidas a ciertas propiedades. Es una técnica cuyo objetivo básico es la reducción de la dimensión de un problema con $p$ variables a otro con un número posiblemente menor de nuevas variables. Por otro lado, en el ACP paramétrico, que aquí abordamos, el vector aleatorio $X$ se supondrá modelizado, a la hora de realizar inferencia, por una distribución normal $p$-dimensional.

## 1.1. Construcción de las CP.

El ACP pretende explicar la estructura de covarianza de un vector aleatorio $X$ mediante la búsqueda de un nuevo sistema de ejes coordenados (las CP) que indican las direcciones de mayor variabilidad en una situación teórica dada (con $\Sigma$ matriz de covarianza de $X$ conocida) o posteriormente de una matriz $\Sigma$ estimada a partir de datos observados.

Analizamos en primer lugar el Modelo Teórico del ACP, supuesto que conocemos la matriz $\Sigma$ o la matriz de correlaciones $\mathbb{R}$ del vector $X$. Estudiaremos el método clásico de obtención de las CP.

Supongamos $X = (X_1, \ldots, X_p)'$ con $Cov(X) = \Sigma$ semidefinida positiva[1] y con $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, las raíces características correspondientes. Sean las combinaciones lineales

$$\left\{ \begin{array}{lcl} Y_1 & = & l_1'X = l_{11}'X_1 + \cdots + l_{1p}'X_p \\ \vdots & & \\ Y_p & = & l_p'X = l_{p1}'X_1 + \cdots + l_{pp}'X_p \end{array} \right.$$

Consideremos el vector aleatorio $Y = (Y_1, \ldots, Y_p)'$. Dadas dos cualesquiera de sus componentes, $i$ y $j$, es claro que

$$Var(Y_i) = l_i'\Sigma l_i \quad Cov(Y_i, Y_j) = l_i'\Sigma l_j$$

y esto es cierto para todo vector $X$, cualquiera que sea su distribución.

### 1.1.1. Definición.

Se llaman *componentes principales (CP)* las combinaciones lineales $Y_1, \ldots, Y_p$ que son incorreladas entre sí y tales que hacen máximas, en el sentido que luego se precisará, las varianzas $l_i'\Sigma l_i, i = 1, \ldots, p$.

[1] En general $\Sigma$ es definida no negativa

Para construir las CP así definidas realizaremos el siguiente proceso:

- Consideremos la combinación lineal de varianza máxima (llamémosla $Y_1$) de modo que esta varianza será $Var(Y_i) = l_i'\Sigma l_i$. Obviamente esto tiene una indeterminación ya que dicha varianza aumentará sin más que multiplicar $l$ por una constante positiva.
- Introducimos por tanto la restricción de que los vectores $l$ sean unitarios en todas las CP a obtener, por tanto $l_i'l_i = 1$.
- Llamamos primera componente principal a la CL $Y_1 = l_1'X$ tal que hace máxima $Var(Y_1)$ con la restricción $l_1'l_1 = 1$.
- Llamamos segunda componente principal a la CL $Y_2 = l_2'X$ tal que hace máxima $Var(Y_2)$ con la restricción $l_2'l_2 = 1$ y con la restricción adicional de ser incorrelada con $Y_1$, esto es
$$Cov(l_1'X, l_2'X) = 0.$$
- El procedimiento se continúa hasta construir las p combinaciones lineales $Y_1, \ldots, Y_p$. Tal que una $Y_i$ cualquiera, $i = 1, \ldots, p$, por definición, maximiza $Var(l_i'X)$ sujeta a $l_i'l_i = 1$ y a $Cov(l_i'X, l_k'X) = 0$ para $k < i$.

Estos sucesivos problemas de máximo (condicionados) se pueden resolver fácilmente, obteniéndose sucesivamente las p CP. Enfocando el cálculo mediante multiplicadores de Lagrange se obtienen sucesivamente las CP.

### 1.1.2. Cálculo de la primera componente principal.

La primera componente principal la definimos como

$$Y_1 = e_1'X,\ e_1'e_1 = 1$$

tal que

$$Var(Y_1) = Var(l'X),\ \max_l Var(l'X) = Var(e_1'X) = e_1'\Sigma e_1.$$

Estando ante el Problema de Lagrange de maximización condicionada:

$$\left\{ \begin{array}{l} \max_l\{l'\Sigma l\} \\ l'l = 1 \end{array} \right\} \Rightarrow \Phi_1(l) = l'\Sigma l - \lambda(l'l - 1) \Rightarrow \frac{\partial \Phi_1(l)}{\partial l} = 2\Sigma l - 2\lambda l = 0 \Rightarrow (\sigma - \lambda I)l = 0$$

Supuesto que $\Sigma_{p\times p}$ tiene autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$,[2] con autovectores asociados $e_1, e_2, \ldots, e_p$ y como $l'\Sigma l = \lambda l'l = 1$, $Var(l'\Sigma l) = \lambda$ y es claro que tomando $l = e_1$, correspondiente al mayor autovalor, se resuelve el problema planteado, de modo que la primera CP es $Y_1 = e_1'X$ y se tiene $Var(Y_1) = \lambda_1$.

### 1.1.3. Cálculo de la segunda componente principal.

Se trata de obtener, según la definición anterior, una combinación lineal $Y_2 = l'X$, incorrelada con $Y_1$ y de varianza máxima. Por tanto,

$$\max_l\{l'Xl\},\ con\ l'l = 1,\ l'\Sigma e_1 = 0$$

$$\Phi_2(l) = l'\Sigma l - \lambda(l'l - 1) - 2v(l'\Sigma e_1) \Rightarrow \frac{\partial \Phi_2(l)}{\partial l} = 2\Sigma l - 2\lambda l - 2v\Sigma e_1.$$

El problema se resuelve con la raiz $\lambda_2$, segunda en orden decreciente y con el correspondiente autovector $e_2$ de modo que $Y_2 = e_2'X$ y $Var(Y_2) = \lambda_2$.

[2] $\Sigma$, en general, como matriz de covarianza, es semidefinida positiva

### 1.1.4. Cálculo de la $(r+1)$-ésima componente principal $(1 \leq r+1 \leq p)$.

En este caso tenemos

$$Y_{r+1} = l'X;\ l'l = 1;\ l'\Sigma e_i = 0,\ i = 1, \dots r$$

$$\Phi_{r+1}(l) = l'\Sigma l - \lambda(l'l - 1) - 2\sum_{i=1}^{r} v_i l'\Sigma e_i.$$

Puede demostrarse que, siendo $\lambda_i \neq 0,\ i = 1, \dots, r$ el problema conduce a $v_i = 0,\ i = 1, \dots, r$ de modo que el sistema que resuelve el problema de maximización es

$$\{2\Sigma l - 2\lambda l = 0, \Sigma l - \lambda l = 0, (\Sigma - \lambda I)l = 0\}\cdot$$

Si $\lambda_{r+1} \neq 0$, basta tomar $\lambda = \lambda_{r+1}$, $l = e_{r+1}$ y se obtiene la $(r+1)$-ésima CP que es

$$Y_{r+1} = e'_{r+1}X,\ Var(Y_{r+1}) = \lambda_{r+1}.$$

En el caso en que $\lambda_{r+1} = 0,\ \lambda_i \neq 0, i \neq r+1$, se toma una CL de $\alpha_{r+1}$ y $\alpha_i$ para la cual $\alpha_i \neq 0$.

Una vez conseguidos $A = (e_1, \dots, e_p),\ \Lambda = diag(\lambda_1, \dots, \lambda_p),\ \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, como $A'A = I$ y $\Sigma A = A\Lambda$, se tiene que $A'\Sigma A = \Lambda$.

En caso que haya raíces características múltiples, es posible probar el siguiente teorema:

**Teorema 1** (Teorema de las raíces características múltiples)**.** *Si $\lambda_{r+1} = \lambda_{r+m} = \lambda$, entonces $\Sigma - \lambda I$ es de rango $p - m$. Los correspondientes vectores característicos $\alpha_{r+1}, \dots, \alpha_{r+m}$ están unívocamente determinados, salvo multiplicación por la derecha por una matriz ortogonal.*

### 1.1.5. Construcción conjunta de las $p$ componentes principales.

En lugar de ir obteniendo sucesivamente las CP resolviendo los sucesivos problemas de máximo condicionado y al final considerar globalmente todos, como antes se ha descrito, cabe, metodológicamente, actuar globalmente desde un comienzo. Por supuesto obtenemos los mismos resultados, pero en lugar de ir aplicando y resolviendo los sucesivos problemas de máximos condicionados de Lagrange, nos basaremos en un conocido resultado de maximización.

**Lema 1** (Lema de maximización)**.** *Sea $A$ una matriz $p \times p$ definida positiva, con autovalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ y autovalores normalizados $e_1, \dots, e_p$, y sea $x$ un vector $p \times 1$, arbitrario no nulo. Entonces se cumple que:*

$$\max_x \frac{x'Ax}{x'x} = \lambda_1,\ \textit{alcanzado en } x = e_1,$$

$$\min_{x'x \neq 0} \frac{x'Ax}{x'x} = \lambda_p,\ \textit{alcanzado en } x = e_p,$$

$$\max_{x \perp e_1, \dots, e_p} \frac{x'Ax}{x'x} = \lambda_{k+1},\ \textit{alcanzado en } x = e_{k+1}, k = 1, 2, \dots, p-1.$$

Es posible, entonces, demostrar el siguiente resultado.

**Teorema 2** (Teorema de componentes principales.)**.** *Sea $X = (X_1, \dots, X_p)'$ un vector aleatorio con matriz de covarianza conocida $\Sigma$ definida positiva y real y sean $(\lambda_i, e_i)$ los autovalores-autovectores de $\Sigma$ con $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. La CP $i$-ésima $Y_i$ antes definida viene dada por*

$$Y_i = e'_i X = e_{i1}X_1 + \cdots + e_{ip}X_p,\ i = 1, \dots, p,$$

*verificándose*

$$Var(Y_i) = e'_i \Sigma e_i = \lambda_i \qquad Cov(Y_i, Y_j) = e'_i \Sigma e_j = 0, i \neq j.$$

*Si hay autovalores iguales, pongamos $\lambda_k$, entonces los $e_k$ asociados no son únicos, por lo que, en este caso, las respectivas CP no son únicas.*

Dada una matriz B, $p \times p$, definida positiva con descomposición espectral $B = \sum_{i=1}^{p} \lambda_i e_i'$, sea la matriz $\mathbb{P} = (e_1, \ldots, e_p)$ formada por columnas, con los autovectores normalizados $e_i$. Entonces, $B = \mathbb{P}\Lambda\mathbb{P}' = \sum_{i=1}^{p} \lambda_i e_i e_i'$, siendo $\mathbb{P}\mathbb{P}' = I$ y $\lambda = diag(\lambda_1, \ldots, \lambda_p)$.

En el caso en que $\lambda_i > 0$ se puede utilizar esta descomposición para definir la matriz $B^{1/2}$, raíz cuadrada de B, ya que, al ser $B^{-1} = \mathbb{P}\Lambda^{-1}\mathbb{P}' = \sum_{i=1}^{p} \frac{1}{\lambda_i} e_i e_i'$, se define $B^{1/2} = \mathbb{P}\Lambda^{1/2}\mathbb{P}' = \sum_{i=1}^{p} \sqrt{\lambda_i} e_i e_i'$.

## 1.2. Estructura de la matriz de covarianza $\Sigma$ en el ACP.

Del teorema de las CP se deduce que $\Sigma = \mathbb{P}\Lambda\mathbb{P}'$ ($\Sigma > 0$), donde $\Lambda$ es la matriz diagonal de autovalores y $\mathbb{P}$ es la matriz de los autovectores, por columnas, que verifica $\mathbb{P}\mathbb{P}' = \mathbb{P}'\mathbb{P} = I$. Por tanto, el ACP induce una factorización estructural de la matriz de covarianzas $\Sigma$ del vector $X$.[3] Esta factorización tiene una importante propiedad, que conduce a la invarianza.

### 1.2.1. Propiedad de invariancia

En las condiciones del teorema de las CP, $tr(\Sigma) = tr(\Lambda)$. En efecto,

$$tr(\Sigma) = tr(\mathbb{P}\Lambda\mathbb{P}') = tr(\Lambda\mathbb{P}'\mathbb{P}) = tr(\Lambda I) = tr(\Lambda)$$

es decir, $\sigma_{11} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \cdots + \lambda_p = \sum_{i=1}^{p} Var(Y_i)$.

Esta invariancia es en realidad la base de la aplicación práctica del ACP, pues la proporción de la varianza total del vector $X$, es decir $\sum_{i=1}^{p} Var(X_i)$, que es debida a $j$-ésima CP, $Y_j$, es $\frac{\lambda_j}{\lambda_1 + \cdots + \lambda_p}, j = 1, \ldots, p$.

*Si las CP son tales que unas pocas explican un alto porcentaje de la varianza total, merece la pena sustituir el vector X original por esas CP.*

Por otra parte, también es un invariante la varianza generalizada (Wilks) respecto de las variables originales y respecto de las CP, ya que de la estructura $\Sigma = \mathbb{P}\Lambda\mathbb{P}'$ se deduce que $|\Sigma| = |\mathbb{P}\Lambda\mathbb{P}'| = |\Lambda|$.

### 1.2.2. Correlación entre las CP y las variables originales.

Sean las $p$ CP, $Y_j$ asociadas al vector aleatorio $X$ de matriz de covarianzas $\Sigma$ conocida y sean $(\lambda_i, e_i)$ sus autovalores-autovectores. Vamos a calcular $\rho_{Y_i,X_k}$ para ello consideremos $h_k' = (0, \ldots, 0, 1, 0, \ldots, 0)$, definido por $h_{ki} = \delta_{ki}$. Entonces,

$$Cov(Y_i, X_k) = Cov(e_i'X, h_k'X) = e_i'\Sigma h_k = h_k'\Sigma e_i = h_k'\lambda_i e_i = \lambda_i h_k' e_i = \lambda_i e_{ki}.$$

Por otro lado, $Var(Y_i) = \lambda_i, Var(X_k) = \sigma_{kk}$, luego tenemos

$$\rho_{Y_i,X_k} = \frac{\lambda_i e_{ki}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} = \frac{e_{ki}\sqrt{\lambda_i}}{\sigma_k} \quad i, k = 1, \ldots, p.$$

¿Qué significado tiene este resultado? A la vista de la expresión obtenida para $\rho_{Y_i,X_k}$ queda claro que la componente $k$-ésima del autovector $e_i$ que proporciona la CP $Y_i$, mide la importancia que la variable original $k$-ésima, $X_k$, tiene en dicha CP, de modo que cuanto mayor sea $|e_{ki}|$ mayor es la correlación entre $X_k$ y la $Y_i$ considerada.

[3] En el caso en que haya autovalores iguales, no es única la matriz $\mathbb{P}$ de autovectores.

**Ejemplo 1.** *Supongamos* $X = (X_1, X_2, X_3)$ *con* $\Sigma = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$. *Los autovalores son* $\lambda_1 = 5{,}83, \lambda_2 = 2{,}00, \lambda_3 = 0{,}17$ *y los autovectores son* $e_1' = (0{,}383, -0{,}924, 0)$, $e_2' = (0, 0, 1)$ *y* $e_3' = (0{,}924, 0{,}383, 0)$. *Las componentes principales son* [4]

$$\begin{aligned} Y_1 &= 0{,}383X_1 - 0{,}924X_2 \\ Y_2 &= X_3 \\ Y_3 &= 0{,}924X_1 + 0{,}383X_2 \end{aligned}$$

*Sabemos que* $Var(Y_i) = \lambda_i$. *Comprobémoslo para* $Y_1$.

$$Var(Y_1) = Var(0{,}383X_1 - 0{,}924X_2) = E\left[0{,}383(X_1 - \mu_1) - 0{,}924(X_2 - \mu_2)\right]^2 =$$

$$(0{,}383)^2 E(X_1 - \mu_1)^2 + (-0{,}924)^2 E(X_2 - \mu_2)^2 + 2(0{,}383)(-0{,}924)E\left[(X_1 - \mu_1)(X_2 - \mu_2)\right] =$$

$$(0{,}383)^2 Var(X_1) + (-0{,}924)^2 Var(X_2) + 2(0{,}383)(-0{,}924)Cov(X_1, X_2) =$$

$$(0{,}383)^2 \cdot 1 + (-0{,}924)^2 \cdot 5 + 2(0{,}383)(-0{,}924) \cdot (-2) = 0{,}147 + 0{,}854 \cdot 5 + 0{,}708 \cdot 2 = 5{,}83.$$

*Sabemos que las CP son incorreladas. Comprobémoslo, por ejemplo, con*

$$Cov(Y_1, Y_2) = Cov(0{,}383X_1 - 0{,}924X_2, X_3) = E\left[0{,}383(X_1 - \mu_1)(X_2 - \mu_2)\right] =$$

$$0{,}383E\left[(X_1 - \mu_1), (X_3 - \mu_3)\right] - 0{,}924E\left[(X_2 - \mu_2), (X_3 - \mu_3)\right] =$$

$$0{,}383Cov(X_1, X_3) - 0{,}924Cov(X_2, X_3) = 0{,}383 \cdot 0 - 0{,}924 \cdot 0 = 0.$$

*Obsérvese también que la traza es invariante, ya que*

$$1 + 5 + 2 = 5{,}83 + 2{,}00 + 0{,}17.$$

*El porcentaje de varianza explicado por la primera componente* $Y_1$ *es del* $73\,\%$ $\left(\dfrac{5{,}83}{8} \times 100\right)$. *Análogamente, entre* $Y_1$ *e* $Y_2$ *explican el* $98\,\%$, *por lo que, a efectos prácticos, podemos sustituir el vector* $(X_1, X_2, X_3)$ *por el vector* $(Y_1, Y_2)$.

*Podemos también calcular las correlaciones entre* $Y_i$ *y* $X_j$. *Así, por ejemplo,* $\rho_{Y_1,X_1} = \dfrac{0{,}383\sqrt{5{,}83}}{\sqrt{1}} = 0{,}925$ *y* $\rho_{Y_1,X_2} = \dfrac{-0{,}924\sqrt{5{,}83}}{\sqrt{5}} = -0{,}998$,[5] *de donde se deduce que* $X_1$ *y* $X_2$ *son prácticamente igual de importantes para la primera CP. Del mismo modo,* $\rho_{Y_2,X_1} = 0$, $\rho_{Y_2,X_2} = 0$ *y* $\rho_{Y_2,X_3} = 1$. *Pueden calcularse,* $\rho_{Y_3,X_1}$, $\rho_{Y_3,X_2}$ *y* $\rho_{Y_3,X_3}$.

### 1.2.3. Componentes principales extraídas sobre $\Sigma$ y sobre $\mathbb{R}$ (estandarización)

Sea la variable $X = (X_1, \ldots, X_p)'$ con $E(X) = \mu$ y $Cov(X) = \Sigma$. Realicemos la transformación de estandarización $Z = \mathbb{D}^{-1/2}(X - \mu)$, siendo $\mathbb{D} = diag(\sigma_{11}, \ldots, \sigma_{pp})$, Esto es:

$$\begin{pmatrix} Z_{11} \\ \vdots \\ Z_{pp} \end{pmatrix} = \begin{pmatrix} \sigma_1^{-1} & & \\ & \ddots & \\ & & \sigma_p^{-1} \end{pmatrix} \begin{pmatrix} X_1 - \mu_1 \\ \vdots \\ X_p - \mu_p \end{pmatrix}.$$

El vector $Z$ tiene unas CP basadas en su propia matriz de covarianzas. Ahora bien,

$$Cov(Z) = Cov[\mathbb{D}^{-1/2}(X - \mu)(X - \mu)'\mathbb{D}^{-1/2}] = \mathbb{D}^{-1/2}\Sigma\mathbb{D}^{-1/2} = \mathbb{R},$$

[4] Obsérvese que $X_3$ es una CP por que es incorrelada con las otras dos.
[5] Obsérvese que puede ser engañoso ver los coeficientes, sólo en $X_1$

siendo $\mathbb{R}$ la matriz de correlación de $X$.

La pregunta que se plantea inmediatamente es la siguiente: ¿Son invariantes las CP por un cambio como el que hemos realizado[6]? ¿Son homogéneas frente al cambio?. La contestación a ambas preguntas es negativa en general, pudiéndose enunciar el siguiente resultado.

**Lema 2.** *La i-ésima componente principal del vector tipificado $Z$ con matriz de covarianzas $\mathbb{R}$, viene dada por $Y_i = \alpha_i' Z = \alpha_i' \mathbb{D}^{-1/2}(X-\mu), i = 1,\ldots,p$ siendo $\alpha_i$ los autovectores asociados a los autovalores $\lambda_i$ de $\mathbb{R}$,* [7] *cumpliendo la propiedad de que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, verificándose además que $\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = p$.*

**Ejemplo 2.** *Sea $X = (X_1, X_2)$ con $\Sigma = \begin{pmatrix} 1 & 4 \\ 4 & 100 \end{pmatrix}$. En este caso será $\mathbb{D}^{-1/2} = \begin{pmatrix} 1 & 0 \\ 0 & 0{,}1 \end{pmatrix}$ y por tanto $\mathbb{R} = \mathbb{D}^{-1/2}\Sigma\mathbb{D}^{-1/2} = \begin{pmatrix} 1 & 0{,}4 \\ 0{,}4 & 1 \end{pmatrix}$.*

*Si trabajamos con la matriz $\Sigma$ tenemos:*

$$|\Sigma - \lambda I| = 0 \Rightarrow \lambda_1 = 100{,}16, \lambda_2 = 0{,}84, \lambda_1 + \lambda_2 = 101.$$

$$\alpha_1 = (0{,}040, 0{,}999)' \quad \alpha_2 = (0{,}999, -0{,}040)'.$$

$$Y_1 = 0{,}040X_1 + 0{,}999X_2 \quad Y_2 = 0{,}999X_1 - 0{,}040X_2.$$

$Y_1$ *explica el* $100 \times \dfrac{100{,}16}{101}\,\% = 99{,}2\,\%$ $Y_2$ *explica el* $100 \times \dfrac{0{,}84}{101}\,\% = 0{,}8\,\%$

$$\rho_{Y_1,X_1} = 0{,}400 \quad \rho_{Y_1,X_2} = 0{,}100$$

*Si trabajamos con la matriz $\mathbb{R}$ tenemos:*

$$|\mathbb{R} - \lambda I| = 0 \Rightarrow \lambda_1 = 1{,}4, \lambda_2 = 0{,}6, \lambda_1 + \lambda_2 = 2.$$

$$\alpha_1 = (0{,}707, 0{,}707)' \quad \alpha_2 = (0{,}707, 0{,}707)'.$$

$$Y_1 = 0{,}707Z_1 + 0{,}707Z_2 \quad Y_2 = 0{,}707Z_1 - 0{,}707Z_2.$$

$Y_1$ *explica el* $100 \times \dfrac{1{,}4}{2}\,\% = 70\,\%$ $Y_2$ *explica el* $100 \times \dfrac{0{,}6}{2}\,\% = 30\,\%$

$$\rho_{Y_1,Z_1} = 0{,}837 \quad \rho_{Y_1,Z_2} = 0{,}837$$

Observamos, por tanto, que cuando las variables se estandarizan, tanto $Z_1$ como $Z_2$ contribuyen por igual a la primera CP de $\mathbb{R}$, explicando el 70 % de la varianza total.

Así pues, la estructura de las CP cambia según nos basemos en $\Sigma$ o en $\mathbb{R}$. A menudo es práctica habitual la tipificación, especialmente si el rango de medición es muy diferente.

### 1.2.4. Estructuras especiales de $\Sigma$

A veces nos encontramos con estructuras especiales, como es el caso de un problema en Biología en que la matriz de covarianzas es de la forma

$$\Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{pmatrix}$$

[6] Una afinidad.

[7] Nótese que serán las raices de $|\mathbb{R} - \lambda I| = 0$

a la que corresponde una matriz de correlaciones

$$\mathbb{R} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

que es la matriz de covarianzas de las variables primitivas tipificadas.[8]

Se puede demostrar que las raíces de la ecuación $|\Sigma - \lambda I| = 0$ son, cuando $\rho$ es positivo, las siguientes:

$$\lambda_1 = 1 + (p-1)\rho \quad \lambda_2 = \cdots = \lambda_p = 1 - \rho$$

por tanto, una raíz mayor $\lambda_1$ y una de orden de multiplicidad $p-1$. A la primera corresponde un autovector $\alpha_1 = (1, \ldots, 1)' \times p^{-1/2}$ en tanto que a la raíz múltiple corresponde un subespacio de dimensión $p-1$ en el que podemos definir por ejemplo:

$$\begin{array}{rcl} \alpha_2 & = & (1, -1, 0, \ldots, 0, \ldots, 0)' \sqrt{1 \times 2}. \\ \alpha_3 & = & (1, 1, -2, \ldots, 0, \ldots, 0)' \sqrt{2 \times 3}. \\ \vdots & & \vdots \\ \alpha_i & = & (1, 1, \ldots, -(i-1), 0, \ldots, 0)' \sqrt{(i-1) \times i}. \\ \vdots & & \vdots \\ \alpha_p & = & (1, 1, 1, \ldots\ldots\ldots, -(p-1))' \sqrt{(p-1) \times p}. \end{array}$$

La primera CP es $Y_1 = \alpha_1' X = p^{-1/2} \times \sum_{i=1}^{p} X_1$ que explica $\dfrac{\lambda}{p} = \rho + \dfrac{1-\rho}{\rho}$.

### 1.2.5. Muestras de combinaciones lineales de variables aleatorias.

Sea $X = (X_1, X_2, \ldots, X_p)'$ y una combinación lineal (CL) definida $c'X$. Si tomamos una muestra de tamaño $N$, las CL muestrales serán

$$c'x_j = c_1 x_{1j} + \cdots + c_p x_{pj}, j = 1, \ldots, N$$

siendo $x_j = (x_{1j}, \ldots, x_{pj})$ el $j$-ésimo individuo de la muestra.

La varianza muestral de las CL muestrales será:

$$\frac{1}{N-1}\left[(c'x_1 - c'\overline{x})^2 + (c'x_2 - c'\overline{x})^2 + \cdots + (c'x_N - c'\overline{x})^2\right] =$$

$$\frac{1}{N-1}\left[c'(x_1 - \overline{x})(x_1 - \overline{x})'c + c'(x_2 - \overline{x})(x_2 - \overline{x})'c + \cdots + c'(x_N - \overline{x})(x_N - \overline{x})'c\right] =$$

$$\frac{1}{N-1} c' \left[(x_1 - \overline{x})(x_1 - \overline{x})' + (x_2 - \overline{x})(x_2 - \overline{x})' + \cdots + (x_N - \overline{x})(x_N - \overline{x})'\right] c =$$

$$c' \frac{\sum_{i=1}^{N} (x_i - \overline{x})(x_i - \overline{x})'}{N-1} c = c'Sc.$$

Supongamos otra CL distinta $b'X$ para la misma muestra. Es fácil ver que su media muestral sería $b'\overline{x}$ y su varianza muestral $b'Sb$ y que la covarianza muestral entre las dos CL consideradas será $b'Sc = c'Sb$.

En efecto, la covarianza sería:

$$\frac{1}{N-1}\left[(b'x_1 - b'\overline{x})(c'x_1 - c'\overline{x})' + (b'x_2 - b'\overline{x})(c'x_2 - c'\overline{x})' + \cdots + (b'x_N - b'\overline{x})(c'x_N - c'\overline{x})'\right] =$$

[8] La estructura especial, indica que todas las variables estén igualmente correlacionadas.

$$\frac{1}{N-1}\left[b'(x_1-\overline{x})(x_1-\overline{x})'c+b'(x_2-\overline{x})(x_2-\overline{x})'c+\cdots+b'(x_N-\overline{x})(x_N-\overline{x})'c\right]=$$

$$b'\frac{(x_1-\overline{x})(x_1-\overline{x})'+(x_2-\overline{x})(x_2-\overline{x})'+\cdots+(x_N-\overline{x})(x_N-\overline{x})'}{N-1}c=b'Sc.$$

## 1.3. Análisis de Componentes Principales Muestral (ACPM).

Consideremos la siguiente situación: Se dispone de una muestra aleatoria de tamaño $N$, $x_1, x_2, \ldots, x_N$, de una población $X=(X_1,\ldots,X_p)'$, de vector de medias $E(X)=\mu$ y matriz de covarianzas $Cov(X)=\Sigma$ (desconocida). Sean $\overline{x}$ y $S$ los correspondientes valores muestrales. El objetivo del ACPM es conseguir explicar el mayor porcentaje posible de variación de la muestra con unas CL incorreladas de las variables que hagan máximas las varianzas.

Así pues, dada la muestra $x_1,\ldots,x_N$ tendremos una CL definida por

$$l_i'x_j=l_{1i}x_{1j}+l_{2i}x_{2j}+\cdots+l_{pi}x_{pj}\quad j=1,\ldots,N.$$

Se tendrá, por tanto, para cada CL, $l_i'x_j$, una media muestral $l_i'\overline{x}$ y una varianza muestral $l_i'Sl_i$, y para cada par $l_i'x_j$ y $l_k'x_j$ una covarianza muestral $l_i'Sl_k$.

Llamamos *primera componente principal muestral* a una CL $l_1'X$ tal que al considerar sus $N$ valores sobre la muestra, $\{l_1'x_1, l_1'x_2,\ldots,l_1'x_N\}$, éstos hacen máxima la varianza $Var[\{l_1'x_1,l_1'x_2,\ldots,l_1'x_N\}]=l_1'Sl_1$ sujeto a la restricción de $l_1'l_1=1$.

Llamamos *segunda componente principal muestral* a una CL $l_2'X$ tal que al considerar sus $N$ valores sobre la muestra, $\{l_2'x_1, l_2'x_2,\ldots,l_2'x_N\}$, éstos hacen máxima la varianza $Var[\{l_2'x_1,l_2'x_2,\ldots,l_2'x_N\}]=l_2'Sl_2$ sujeto a las restricciones de que $l_2'l_2=1$ y además

$$Cov[\{l_1'x_1,l_1'x_2,\ldots,l_1'x_N\},\{l_2'x_1,l_2'x_2,\ldots,l_2'x_N\}]=0$$

esto es, $l_1'Sl_2=l_2'Sl_1=0$.

Llamamos *$i$-ésima componente principal muestral* a una CL $l_i'X$ tal que al considerar sus $N$ valores sobre la muestra, $\{l_i'x_1, l_i'x_2,\ldots,l_i'x_N\}$, éstos hacen máxima la varianza $Var[\{l_i'x_1,l_i'x_2,\ldots,l_i'x_N\}]=l_i'Sl_i$ sujeto a las restricciones de que $l_i'l_i=1$ y además

$$Cov[\{l_k'x_1,l_k'x_2,\ldots,l_k'x_N\},\{l_i'x_1,l_i'x_2,\ldots,l_i'x_N\}]=0$$

esto es, $l_k'Sl_i=l_i'Sl_k=0$, para $k<i$.

**Teorema 3.** *Sea $X=(X_1,X_2,\ldots,X_p)'$ una variable aleatoria con $E(X)=\mu$ y $Cov(X)=\Sigma$ desconocida. Sea una muestra $x_1,\ldots,x_N$ de $X$, con $x_j=(x_{1j},\ldots,x_{pj}), j=1,...N$. Sea $\bar{x}=\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}x_i$ la media muestral y $S=(s_{ij})=\dfrac{1}{N-1}\displaystyle\sum_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x})'$ la covarianza muestral. Sean $\hat{\lambda}_1\geq\cdots\geq\hat{\lambda}_p\geq 0$ los $p$ autovalores de $S$,[9] solución de la ecuación $|S-\lambda I|=0$ y sean $\hat{e}_1,\ldots,\hat{e}_p$ los respectivos autovectores.*

*Sean $\hat{y}_i=\hat{e}_ix$, donde $x$ es cualquier observación de la variable $X$, $\hat{y}_i$ son las CP muestrales.*

*Se cumple que*

$$\textit{Varianza Muestral}(\hat{y}_i)=\hat{\lambda}_i.$$

[9] Suponemos que S es definida no negativa

$$Covarianza\ Muestral(\hat{y}_i, \hat{y}_k) = 0 \quad si\ i \neq k.$$

$$Varianza\ Total\ Muestral = \sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \cdots + \hat{\lambda}_p \quad y \quad \hat{\rho}_{\hat{y}_i, x_k} = \frac{\hat{e}_{ki}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}.$$

También en el caso de ACP muestral, es frecuente tipificar las observaciones, con un comportamiento análogo al caso del modelo teórico. Así, tipificando la muestra $\{x_1, \ldots, x_N\}$, siendo $x_j = (x_{1j}, \ldots, x_{pj})$ se obtiene

$$z_j = \hat{\mathbb{D}}^{-1/2}(x_j - \bar{x}),\ esto\ es\ z_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_{ii}}}, i = 1, \ldots, p; j = 1, \ldots, N.$$

Es fácil comprobar que $\bar{z} = \frac{1}{N}\sum_1^N z_j = 0$, $S_z = \frac{1}{N-1}\sum_1^N (z_i - \bar{z})(z_i - \bar{z})' = \hat{\mathbb{R}}$, matriz de correlación muestral, de tal modo que $\mathbb{R} = (r_{ij}) = \mathbb{D}^{-1/2} S \mathbb{D}^{-1/2}$. En nomenclatura matricial, llamando $\mathbf{1} = (1, \ldots, 1)'_{N\times 1}$ y $Z = (z_1, \ldots, z_N)_{p\times N}$ se tiene

$$\bar{z} = \frac{1}{N} Z\mathbf{1} \quad S_z = \frac{1}{N-1}(Z - \bar{z}\mathbf{1}')(Z - \bar{z}\mathbf{1}')'.$$

## 1.4. Análisis de Componentes Principales en poblaciones normales.

Hasta ahora no hemos supuesto que $X = (X_1, \ldots, X_p)'$ sea normal $p$-variante, sino sólo que $E(X) = \mu$ y $Cov(X) = \Sigma$. A su vez hemos analizado dos casos:

1. $\Sigma$ conocida, con lo que $\lambda_i$ y $e_i$ son conocidos determinísticamente.
2. $\Sigma$ desconocida, en cuyo caso hemos basado el ACP en una muestra de la población y, al no conocer $\Sigma$, nos hemos basado en una matriz de cuasivarianzas muestrales $S$ de dicha muestra, desarrollando el ACP muestral.

Para poder conocer el comportamiento de $\hat{\lambda}_i$ y $\hat{e}_i$ y, en definitiva, de $\hat{y}_i$, obtenidos en el ACP muestral, es preciso basarse en la distribución en el muestreo de $\hat{\lambda}_i$, raices características de la matriz muestral $S$ y, en consecuencia, hay que modelizar la distribución de $S$ o, de manera análoga, de $\hat{\mathbb{R}}$ y de sus raices $\hat{\rho}_i$.

Para ello hay que empezar modelizando $X$. El caso bien conocido del Análisis Multivariante teórico corresponde a la distribución $N_p(\mu, \Sigma)$. El esquema es el siguiente:

$$\begin{array}{ccccccc} X & \longrightarrow & \left\{\begin{array}{l} \mu \to \bar{x} \\ \Sigma \to \hat{\Sigma} \to S \\ \lambda_i\ (teóricos) \end{array}\right\} & \longrightarrow & S & \longrightarrow & \hat{\lambda}_i \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ N_p(\mu, \Sigma) & & T.\ Fisher\ y\ Zehna & & Wishart & & \begin{array}{l} Distribución\ de\ las \\ r.c.\ muestrales\ Wishart \end{array} \end{array}$$

Nótese que si $X \rightsquigarrow N_p(\mu, \Sigma)$, $\Sigma > 0$ desconocida, y $X_{p\times N}$ es la matriz de una muestra, se puede dar una interpretación de la matriz $S$ muestral. En efecto, en este caso $S = \frac{A}{N-1}$ $y$ $\hat{\Sigma} = \frac{A}{N}$, o lo que es igual $\Sigma = \frac{N-1}{N} S$. Si $\Sigma > 0$ (según el Teorema de Dykstra) $S$ (o $A$) son definidas positivas y todos sus autovalores son distintos (c.s.). Si $\Sigma$ no es definida positiva puede utilizarse la modelización normal con $rang(\Sigma) < p$.

En el primer caso es claro que las CP muestrales son los estimadores de máxima verosimilitud de su contrapartida teórica: las CP teóricas asociadas a $\Sigma$ que no serán conocidas nunca (según el Teorema Zhenna).

Prescindiendo del desarrollo teórico del estudio, en el caso normal, del comportamiento de $\hat{\lambda}_i$, nos limitamos a dar algunos contrastes básicos de carácter asintótico.

Suponemos que todos los autovalores de $\Sigma$ son distintos y positivos ($\Sigma$ definida positiva), esto es $\lambda_1 > \lambda_2 > \cdots \lambda_p > 0$.

### 1.4.1. Resultados de Anderson-Girschick.

En las condiciones enunciadas sea $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_p)'$ y $(\hat{e}_1, \ldots, \hat{e}_p)$ los autovalores y autovectores de $S$ y análogamente $\lambda$ y $(e_1, \ldots, e_p)$ de $\Sigma$. Sea $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$ y $E_i = \lambda_i \sum_{\substack{k=1 \\ k \neq i}}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} e_k e_k'$. Entonces:

$$\sqrt{N}(\hat{\lambda} - \lambda) \rightsquigarrow N_p(0, 2\lambda^2) \tag{1.1}$$

$$\sqrt{N}(\hat{e}_i - e_i) \rightsquigarrow N_p(0, E_i) \tag{1.2}$$

y además cada $\hat{\lambda}_i$ se distribuye independientemente de los elementos del respectivo $\hat{e}_i$.

**Nota 1.** *El resultado 1.1 implica que si N tiende a $\infty$, los $\hat{\lambda}_i$ se distribuyen independientemente (ya que la matriz de covarianza de la $N_p$ asintótica es diagonal). Además, aproximadamente, $\hat{\lambda}_i$ es $N(\lambda_i, 2\lambda_i^2/N)$. Ello permite establecer intervalos de confianza al $100(1-\alpha)\,\%$ del siguiente modo:*

$$P\left(|\hat{\lambda}_i - \lambda_i| \leq z_{\alpha/2}\lambda_i\sqrt{2/N}\right) = 1 - \alpha \Rightarrow$$

$$\Rightarrow \frac{\hat{\lambda}_i}{1 + z_{\alpha/2}\sqrt{2/N}} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{1 - z_{\alpha/2}\sqrt{2/N}}$$

*Un intervalo simultáneo (Bonferroni), para un $\lambda_i$, será poniendo $z_{\alpha/2m}$.*

**Nota 2.** *Hay que tener cuidado con estos intervalos cuando un $\lambda_i$ es muy grande, aun cuando N no lo sea. Conviene actuar, siempre que se pueda, sobre $\mathbb{R}$. La razón es que en estos casos se producen intervalos muy amplios.*

**Nota 3.** *Del resultado 1.2 se deduce que los $\hat{e}_i$ se distribuyen normalmente alrededor de los $e_i$ respectivos para $N \to \infty$. Pero los elementos de $\hat{e}_i$ están correlacionados, no son independientes, y el grado de correlación depende de la separación de los autovalores $\lambda_1, \ldots, \lambda_p$, que no se conoce, y del tamaño N.*

Los errores típicos aproximados de los coeficientes $\hat{e}_{ki}$, componentes de $\hat{e}_i$, vienen dados por la diagonal de $\frac{1}{N}\hat{E}_i$, donde $\hat{E}_i$ coincide con $E_i$ sustituyendo $\lambda_k$ por $\hat{\lambda}_k$.

Obsérvese finalmente que los anteriores resultados asintóticos de Anderson (1963) y Girschick (1939) suponen que las raices características teóricas de $\Sigma$ en la distribución base son todas distintas y no nulas. Si esta hipótesis no es cierta no pueden aplicarse, aparte el hecho de que aún cuando sí se apliquen, en el resultado $(ii)$ la matriz $E_i$ depende, en elementos fuera de la diagonal principal, de los valores teóricos que no son conocidos.

**Ejemplo 3.** *Un ejemplo concreto que se puede abordar y resolver asintóticamente es el caso en que la matriz $\Sigma$ sea de la forma $\sigma_{ij} = \sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}\rho$ o lo que es lo mismo, que la matriz de correlaciones sea*

$$\mathbb{R}_0 = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}.$$

*Si se supone el test* $H_0 : \mathbb{R} = \mathbb{R}_0$ *frente a* $H_1 : \mathbb{R} \neq \mathbb{R}_0$*, se puede abordar por el método del cociente de verosimilitudes o bien por el método de Lawley (1963). Este resultado asintótico es el siguiente: Se rechaza* $H_0$ *si*

$$\frac{N-1}{(1-\bar{r})^2}\left[\sum_{1<k}(r_{ik}-\bar{r})^2 - \hat{\gamma}\sum_{j=1}^{p}(\hat{r}_j-\hat{r})\right] > \chi^2_{(p+1)(p+2)/2}(\alpha)$$

*siendo*

$$\begin{aligned}\bar{r}_k &= \frac{1}{p-1}\sum_{\substack{i=1\\ i\neq k}}^{p} r_{ik},\ k=1,\dots p.\\ \bar{r} &= \frac{2}{p(p+1)}\sum_{i<k} r_{ik}.\\ \hat{\gamma} &= \frac{(p-1)^2[1-(1-r)^2]}{p-(p-2)(1-\bar{r})^2}.\end{aligned}$$

## 1.5. Cálculo de las Componentes Principales poblacionales.

Sea $X = (X_1, \dots, X_p)'$ un vector aleatorio $p$-variante con $E[X] = \mu$ y matriz de covarianza conocida $\Sigma$. Consideremos casos en los cuales $\Sigma$ es una matriz semidefinida positiva y admitimos que pueda tener raices múltiples. Como sólo nos interesan varianzas y covarianzas de $X$, supondremos que $\mu = 0$.

La primera componente principal de $X$ es la combinación lineal normalizada de $X$: $Y_1 = e'X$, $e = (e_1, \dots, e_p) \in \mathbb{R}^p$ con $e'e = 1$ tal que

$$Var(e'X) = \max_{l} Var(l'X) \quad \forall l \in \mathbb{R}^p \text{ satisfaciendo } l'l = 1.$$

Sabemos que $Var(l'X) = l'\Sigma l$. Entonces, para encontrar la primera componente principal $e'X$ necesitamos encontrar el $e$ que maximiza $l'\Sigma l$ para todas las elecciones de $l \in \mathbb{R}^p$ sujeto a la restricción $l'l = 1$. Usando multiplicadores de Lagrange, $\lambda$, buscamos el $e$ que maximiza:

$$\Phi_1(l) = l'\Sigma l - \lambda(l'l - 1) \ \ \forall l \in \mathbb{R}^p \ \text{ tal que } \ l'l = 1$$

Como $l'\Sigma l$ y $l'l$ tienen derivada, podemos derivar $\Phi_1$ con respecto a $l$, e igualando a 0 obtenemos la ecuación que debe verificar:

$$2\Sigma e - 2\lambda e = 0 \tag{1.3}$$

ó equivalentemente

$$(\Sigma - \lambda I)e = 0.$$

Como $e \neq 0$, por ser $e'e = 1$, la Eq(1.3) tiene solución si $det(\Sigma - \lambda I) = 0$; esto es, $\lambda$ es una raíz característica de $\Sigma$ y $e$ es el vector característico correspondiente. Como $\Sigma$ es de dimensión $p \times p$, hay $p$ valores de $\lambda$ que satisfacen Eq(1.3). Sean $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ las raíces características ordenadas de $\Sigma$ y sean:

$$e_1 = (e_{11}, \dots, e_{1p})', \cdots, e_p = (e_{p1}, \dots, e_{pp})'$$

los vectores característicos correspondientes de $\Sigma$. Como $\Sigma$ es semidefinida positiva, algunas de las raices características pueden ser cero, es más, algunas de las raices pueden tener multiplicidad mayor que la unidad. De Eq(1.3) se tiene:

$$e'\Sigma e = \lambda e'e = \lambda$$

entonces, si $e$ es tal que $e'e = 1$ y satisface , verifica:

$$Var(e'X) = e'\Sigma e = \lambda$$

donde $\lambda$ es la raíz característica de $\Sigma$ correspondiente a $e$.

Para maximizar $Var(e'X)$ necesitamos que $\lambda = \lambda_1$, la raíz característica mayor de $\Sigma$, y $e = e_1$, el vector característico de $\Sigma$ correspondiente a $\lambda_1$.

Se define la *primera componente principal* como la función lineal normalizada $Y_1 = e_1'X = \sum_{i=1}^{p} e_{1i}X_i$ donde $e_1$ es el vector característico normalizado de $\Sigma$ correspondiente a su raíz característica mayor $\lambda_1$.

**Nota 4.** *Hasta ahora no hemos supuesto que $X$ siga una distribución especial. Si $X$ se distribuye según una normal p-variante con matriz de covarianza $\Sigma$, definida positiva, entonces, las superficies de densidad de probabilidad constante son los elipsoides de concentración y, $Y_1 = e_1'X$ representa el eje mayor principal de este elipsoide. En general, bajo la suposición de normalidad de $X$, las componentes principales implicaron una rotación de los ejes coordenados a los ejes principales de estos elipsoides. Si hay raíces múltiples, estos ejes no están únicamente determinados.*

La *segunda componente principal* es la función lineal normalizada $e'X = Y_2$ que tiene máxima varianza entre todas las funciones normalizadas lineales $l'X$ que están incorreladas con $Y_1$.

Si toda función lineal normalizada $l'X$ está incorrelada con $Y_1$, entonces

$$0 = Cov(l'X; Y_1) = (conE[X] = \mu = 0)E[l'XY_1'] = E[l'XX'e_1] =$$

$$E[XX'] = \Sigma l'\Sigma e_1 = l'\lambda_1 e_1 = \lambda_1 l'e_1 = 0 \tag{1.4}$$

Esto implica que los vectores $l$ y $e_1$ son ortoganales [10]. Queremos encontrar una combinación lineal $e'X$ que tenga varianza máxima entre todas las combinaciones lineales normalizadas $l'X$ incorreladas con $Y_1$. Usando multiplicadores de Lagrange $\lambda$, $\nu$, buscamos el $e$ que maximiza:

$$\Phi_2(l) = l'\Sigma l - \lambda(l'l - 1) - 2\nu(l'\Sigma e_1).$$

Derivando con respecto a $l$

$$\frac{\partial \Phi_2}{\partial l} = 2\Sigma l - 2\lambda l - 2\nu\Sigma e_1. \tag{1.5}$$

Por tanto, $e$ debe satisfacer

$$e_1'\Sigma e - \lambda e_1'e - \nu e_1'\Sigma e_1 = 0$$

De la Eq.(1.4) tenemos que: $e_1'\Sigma e = 0$ [11] y $e_1'\Sigma e_1 = \lambda_1 \Rightarrow \nu\lambda_1 = 0$. Como $\lambda_1 \neq 0 \Rightarrow \nu = 0 \Rightarrow$ (usando Eq.(1.5)) $2\Sigma e - 2\lambda e = 0 \Rightarrow (\Sigma - \lambda I)e = 0$ y por tanto, los coeficientes de la segunda componente principal de $X$ son los elementos del vector característico $e_2$ de $\Sigma$ normalizado, correspondiente a su segunda raíz característica mayor $\lambda_2$ (recordemos que $\Sigma$ es simétrica y si es definida positiva, todas sus raices características son reales y positivas). La segunda componente principal de $X$ es $Y_2 = e_2'X$.

Se puede continuar así hasta $r$ $(r < p)$ componentes principales obteniendo $Y_r$. Para la $r+1$-ésima componente principal queremos encontrar una combinación lineal $e'X$ que tenga varianza máxima entre todas las combinaciones lineales normalizadas $l'X$, incorreladas con $Y_1, \ldots, Y_r$.

Así, si $Y_i = e_i'X \quad i = 1, \ldots, r$

$$Cov(l'X; Y_i) = l'\Sigma e_i = l'\lambda_i e_i = \lambda_i l'e_i = 0; \quad i = 1, \ldots, r$$

Para encontrar $e$ se necesita maximizar:

$$\Phi_{r+1}(l) = l'\Sigma l - \lambda(l'l - 1) - 2\sum_{i=1}^{r} \nu_i l'\Sigma e_i. \tag{1.6}$$

donde $\lambda, \nu_1, \ldots, \nu_r$ son los multiplicadores de Lagrange.

[10] Nótese que $\lambda_1 l'e_1 = 0 \Rightarrow l'e_1 = 0$ cuando $\lambda_1 \neq 0$, y $\lambda_1 \neq 0$ si $\Sigma \neq 0$. El caso $\Sigma = 0$ es trivial y no se considera.
[11] Dado que $l'\Sigma e_1 = 0$, $e'\Sigma e_1 = 0 \Rightarrow (e'\Sigma e_1)' = 0 \Rightarrow e_1\Sigma e = 0$.

Se calcula $\dfrac{\partial \Phi_{r+1}}{\partial l}$ y se iguala a 0, obteniéndose que el vector buscado $e$ ha de satisfacer:

$$2\Sigma e - 2\lambda e - 2\sum_{i=1}^{r} \nu_i \Sigma e_i = 0$$

o equivalentemente

$$e_i'\Sigma e - \lambda e_i' e - e\sum_{i=1}^{r} \nu_i \lambda_i = 0$$

dado que $e_i'\Sigma e_i = \lambda_i$.

Se concluye que si $\lambda_i \neq 0 \Rightarrow \nu_i\lambda_i = 0 \Rightarrow \nu_i = 0$. Si $\lambda_i = 0 \Rightarrow \Sigma e_i = \lambda_i e_i = 0 \Rightarrow l'\Sigma e_i = 0 \Rightarrow$ el factor $l'\Sigma e_i$ desaparece en Eq.(1.6).

Así, el $e$ que maximiza la expresión considerada es el vector característico de $\Sigma$, ortogonal a $e_i,\ i = 1, \ldots, r$, correspondiente a su raíz característica $\lambda$. Si $\lambda_{r+1} \neq 0$, tomando $\lambda = \lambda_{r+1}$ y $e$ como el vector característico normalizado $e_{r+1}$ correspondiente a la $(r+1)$-ésima raíz característica mayor $\lambda_{r+1}$, obtenemos la $(r+1)$-ésima componente principal, que es $Y_{r+1} = e_{r+1}'X$.

Sin embargo, si $\lambda_{r+1} = 0$ y $\lambda_i = 0$ para $i \neq r+1$ entonces $e_i'\Sigma e_{r+1} = 0$ no implica que $e_i'e_{r+1} = 0$. En tales casos, reemplazando $e_{r+1}$ por una combinación lineal de $e_{r+1}$ y el $e_i$ para el cual $\lambda_i = 0$, podemos construir el nuevo $e_{r+1}$ ortogonal a todos los $e_i,\ i = 1, \ldots, r$.

Continuamos de esa forma hasta el $m$-ésimo paso, de tal manera que en el $(m+1)$-ésimo paso no podamos encontrar un vector normalizado $e$ tal que $e'X$ sea incorrelado con todas las componentes principales $Y_1, \ldots, Y_m$. Como $\Sigma$ es $p \times p$, obviamente $m < p$ ó $m = p$. Veamos que $m = p$ es la única solución. Supongamos que fuera $m < p$, entonces existirían $p - m$ vectores ortogonales normalizados:

$$B_{m+1}, \ldots, B_p \ \text{ tales que } \ e_i'B_j = 0 \quad i = 1, \ldots, m \ j = m+1, \ldots, p.$$

Sea $B = (B_{m+1}, \ldots, B_p)$. Consideremos una raíz de $|B'\Sigma B - \lambda I| = 0$ y el correspondiente vector $B^0 = (B^0_{m+1}, \ldots, B^0_p)$ satisfaciendo

$$(B'\Sigma B - \lambda I)B^0 = 0 \tag{1.7}$$

Como

$$e_i'\Sigma B B^0 = \lambda_i e_i' \sum_{j=m+1}^{p} B_j B_j^0 = \lambda_i \sum_{j=m+1}^{p} B_j^0 e_i' B_j = 0$$

el vector $e_i$ es ortogonal a $\Sigma B B^0 = BC$, donde $C$ es un vector de $p - m$ componentes. Ahora

$$B'\Sigma B B^0 = B'BC = C$$

de Eq.(1.7) tenemos que

$$\lambda B^0 = C \quad \Sigma(BB^0) = \lambda B B^0.$$

Entonces $(BB^0)X$ es incorrelada con $e_j'X,\ j = 1, \ldots, m$, y conduce a un nuevo $e_{m+1}$. Esto contradice la suposición de que $m < p$ y se tiene entonces que $m = p$.

## 1.6. Manejo simultáneo de todas las componentes principales.

Sea $P = (e_1, \ldots, e_p)$ y $\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}$ donde $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ son todas las raices características ordenadas de $\Sigma$ y $e_1, \ldots, e_p$ son los vectores característicos normalizados correspondientes. Como $P'P = I$ y $\Sigma P = P\Lambda$, se tiene $P'\Sigma P = \Lambda$. Entonces para $Y = (Y_1, \ldots, Y_p)'$ tenemos el siguiente teorema:

**Teorema 4.** *Existe una transformación ortogonal $Y = P'X$ tal que $Cov(P) = \Lambda$ donde $\Lambda$ es una matriz diagonal de elementos $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ que son las raíces ordenadas de $|\Sigma - \lambda I| = 0$. La i-ésima columna de $P$, $e_i$, satisface $(\Sigma - \lambda I)e_i = 0$. Las componentes de $Y$ son incorreladas y $Y_i$ tiene varianza máxima entre todas las combinaciones lineales normalizadas incorreladas con $Y_1, \ldots, Y_{i-1}$.*

El vector $Y$ es llamado el vector de componentes principales de $X$. En el caso de raices múltiples, supongamos que: $\lambda_{r+1} = \cdots = \lambda_{r+m} = \lambda$ entonces $(\Sigma - \lambda I)\alpha_i = 0$, $i = r+1, \ldots, r+m$. Esto es, $\alpha_i$ $(i = r+1, \ldots, r+m)$ son $m$ soluciones linealmente independientes. Para mostrar que no puede haber otra solución linealmente independiente de

$$(\Sigma - \lambda I)\alpha = 0, \tag{1.8}$$

tomamos $\sum_{i=1}^{p} a_i\alpha_i$ ($a_i$ escalares) solución de Eq.(1.8), con lo que se tendrá:

$$\lambda \sum_{i=1}^{p} a_i\alpha_i = \Sigma\left(\sum_{i=1}^{p} a_i\alpha_i\right) = \sum_{i=1}^{p} a_i\Sigma\alpha_i = \sum_{i=1}^{p} a_i\lambda_i\alpha_i$$

Como $\lambda a_i = \lambda_i a_i$ implica que $a_i = 0$, a menos que $i = r+1, \ldots, r+m$. Esto es, el rango de $(\Sigma - \lambda I)\alpha$ es $p-m$.

Obviamente, si $(\alpha_{r+1}, \ldots, \alpha_{r+m})$ es una solución de Eq.(1.8), entonces para cualquier matriz no singular $C$, $(\alpha_{r+1}, \ldots, \alpha_{r+m})C$ es también solución de Eq.(1.8). Pero de la condición de ortonormalidad de $\alpha_{r+1}, \ldots, \alpha_{r+m}$ se concluye fácilmente que $C$ es una matriz ortogonal. Así se tiene el siguiente teorema:

**Teorema 5.** *Si $\lambda_{r+1} = \cdots = \lambda_{r+m} = \lambda$, entonces $(\Sigma - \lambda I)$ es una matriz de rango $p-m$. Además, los correspondientes vectores característicos $e_{r+1}, \ldots, e_{r+m}$ están únicamente determinados salvo multiplicación por la derecha de una matriz ortogonal.*

## 1.7. Test en el ACP basados en la matriz $S$ de covarianzas muestrales.

A continuación damos algunos tests basados en $S$, útiles en ACP. Recordemos que $(= \hat{\Sigma} = A/N$.

### 1.7.1. Test de Bartlett (1947).

Sirve para contrastar que los $p-k$ autovalores más pequeños son todos iguales. Es decir

$$H_0 : \ \lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_p$$

Se efectúa mediante el estadístico

$$\left(N - k - 1 - \frac{2q + 1 + \frac{2}{q}}{6}\right)\left(-\ln|S| \sum_{j=1}^{k} \ln l_{(j)} + q \ln l\right)$$

en donde: $q = p-k$, $l_{(j)} = j$-ésimo autovalor de $S$ y

$$l = \frac{1}{q}\left(tr(S) - \sum_{j=1}^{k} l_{(j)}\right).$$

Bajo la hipótesis nula, sigue una distribución $\chi^2$ con $\frac{1}{2}(p-k-1)(p-k+2)$ grados de libertad, asintóticamente.

### 1.7.2. Test de Bartlett-Lawley (1956).

El anterior estadístico se corrige añadiendo: $l^2 \sum_{j=1}^{k} \frac{1}{(l_{(j)} - l)^2}$, obteniéndose una $\chi^2$ con $\frac{1}{2}(q+2)(q-1)$ grados de libertad asintóticamente. Esta corrección depende del valor $l$ antes indicado.

### 1.7.3. Test de Anderson (1963).

En el conjunto de autovalores de $\Sigma$:

$$\lambda_1 > \lambda_2 > \cdots > \lambda_k > \lambda_{k+1} \geq \cdots \geq \lambda_{q+r} > \cdots > \lambda_p$$

se contrasta:

$$H_0 : \ \lambda_{k+1} = \cdots = \lambda_{q+r} = \lambda.$$

El método del cociente de verosimilitudes y su comportamiento asintótico conducen al siguiente estadístico:

$$-(N-1)\sum_{i=q+1}^{q+r} \ln l_i + (N-1)v \ln \left( \frac{1}{v} \sum_{i=q+1}^{q+r} l_i \right) \rightsquigarrow \chi^2_{\frac{1}{2}[r(r+1)]-1}$$

el cual, cuando $q + r = p$ (igualdad de las últimas $p - q$ raíces características) coincide con el de Bartlett. Asimismo, si $q = 0$, igualdad de todas la raíces características, proporciona el test de Bartlett para dicho caso (caso de esfericidad):

$$-\left( (N-1) - \frac{1}{6}(2p+1+\frac{2}{p}) \right) \left( \ln |S| + p \ln (1/p) \sum_{i=1}^{p} l_i \right) p \rightsquigarrow \chi^2_{(p-1)(p+2)/2}$$

($k = 0$ en la expresión de Bartlett).

Este test de Anderson es dado sin la corrección de Lawley.

**Nota 5.** *El test de Bartlett-Lawley antes dado, basado en la matriz de covarianzas muestrales $S$, es decir en:*

$$S = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})'$$

*que en el caso de una población $N_p(\mu; \Sigma)$, con $\Sigma > 0$, es tal que:*

$$S = \hat{\Sigma},$$

*depende de un valor desconocido ($\lambda$ = valor común en la $H_0$), por tanto no es estrictamente hablando un "estadístico".*

*El resultado original de Bartlett-Lawley dice: Para contrastar la hipótesis nula, $H_0 : \ \lambda_{k+1} = \lambda_{k+2} = \ldots = \lambda_p = \lambda$ sobre la base de $S = \hat{\Sigma}$, se construye la variable:*

$$L = \left\{ N - 1 - k - \frac{1}{6}(2q + \frac{2}{q} + 1) + \lambda^2 \sum_{i=1}^{k} \frac{1}{(\lambda_i - \lambda)^2} \right\} \cdot \left\{ -\ln \left( \frac{|\hat{\Sigma}|}{\prod_{i=1}^{k} \hat{\lambda}_i} \right) + q \ln \left( \frac{tr\hat{\Sigma} - \sum_{i=1}^{k} \hat{\lambda}_i}{q} \right) \right\}; \quad q = p - k.$$

*Esta variable se comporta, con un orden de aproximación de $(1/N^2)$, según una $\chi^2$ con $1/2(q+2)(q-1)$ grados de libertad.*

*Ahora bien, esta variable no es, como antes decíamos, un estadístico al no ser $\lambda$ conocido, ni tampoco los $\lambda_i$, raices características de $\Sigma$. En tal caso se sustituyen los $\lambda_i$ por los $\hat{\lambda}_i$, sus estimadores de máxima verosimilitud (en una $N_p(\mu; \sigma)$ y por el Teorema de Zehna sobre $\hat{\Sigma} = S$), y $\lambda$ por*

$$\frac{1}{p-k} \left\{ tr\hat{\Sigma} - \sum_{i=1}^{k} \hat{\lambda}_i \right\} = \frac{1}{q} \left\{ tr\hat{\Sigma} - \sum_{i=1}^{k} \hat{\lambda}_i \right\}$$

*obsérvese que* $\dfrac{1}{p-k} \left\{ tr\hat{\Sigma} - \sum_{i=1}^{k} \hat{\lambda}_i \right\} = \dfrac{1}{p-k} \left\{ \sum_{i=1}^{p} \hat{\lambda}_i \right\} = \hat{\lambda}.$

Es interesante también dar una variante del test de Bartlett-Lawley preparada para cuando $\lambda = \lambda_0$ por hipótesis. En este caso puede probarse que

$$\left\{ N-1-k-\frac{1}{6}(2q+1-\frac{2}{q+1}-\frac{1}{q+1}\left(\sum_{i=1}^{k}\frac{\lambda_i}{\lambda_i-\lambda}\right)^2+\lambda^2\sum_{i=1}^{k}\frac{1}{(\lambda_i-\lambda)^2}\right\}\cdot$$

$$\cdot\left\{ p\ln\lambda-\ln\frac{|\hat{\Sigma}|}{\prod_{i=1}^{k}\hat{\lambda}_i}+\frac{tr\hat{\Sigma}-\sum_{i=1}^{k}\hat{\lambda}_i}{\lambda}-q\right\};\quad q=p-k$$

se comporta según una $\chi^2$ con $\frac{1}{2}q(q+1)$ grados de libertad, con un orden de aproximación de $1/N^2$.

Señalemos finalmente la filosofía del test Bartlett-Lawley respecto de la práctica del ACP Hemos visto que versa sobre la hipótesis nula

$$H_0: \quad \lambda_{k+1}=\cdots=\lambda_p=\lambda$$

Supongamos que se han "extraído" $k$ componentes principales, correspondientes a las $k$ primeras raices características $\hat{\lambda}_i$; $i=1,\ldots,k$, de la matriz $S=\hat{\Sigma}$. ¿Cómo decidir que las restantes $(p-k)$ componentes principales no son significativas? Es claro que si es verdad la hipótesis nula, con un $\lambda$ pequeño, podemos prescindir de esas $p-k$ componentes principales restantes. En este sentido hay que aplicar el test de Bartlett-Lawley.

## 1.8. Test en ACP sobre $\mathbb{R}$.

Los tests dados antes (en especial el de Bartlett-Lawley) se basan en $S$. Pero ya se vio que en la práctica del ACP muestral, es preciso en muchos casos tipificar los valores observados y por tanto hay que basar el análisis en la matriz de correlaciones $\mathbb{R}$, estimado por la matriz de correlaciones muestrales $\hat{\mathbb{R}}$ (estimadora de máxima verosimilitud de aquella, en el caso de población $N_p(\mu;\Sigma)$).

Los tests en este caso son considerablemente más complicados que los basados en $S$. El problema fue estudiado por Lawley y recogido y aplicado por Dhrymes, entre otros. El test se plantea con la hipótesis nula:

$$H_0: \quad \rho_{k+1}=\cdots=\rho_p=\rho;\quad k<p$$

en donde $\rho_i$ son las raices características de $\mathbb{R}$, estimados por $\hat{\rho}_i$, raíces características de $\hat{\mathbb{R}}$. Si se considera:

$$(N-1)\left\{-\ln\frac{|\hat{\mathbb{R}}|}{\prod_{I=1}^{k}\hat{\rho}_i}+q\ln\frac{tr\hat{\mathbb{R}}-\sum_{i=1}^{k}\hat{\rho}_i}{q}\right\};\quad q=p-k$$

este estadístico, bajo $H_0$, se comporta asintóticamente, con un orden de aproximación de $1/N$, según una $\chi^2$ cuyos grados de libertad vienen dados por la expresión

$$p^*=\frac{1}{2}(q-1)(q+2)-\frac{1}{q}\left\{(q-1)\rho\sum_{i=1}^{p}\sum_{j=1}^{p}c_{ij}^2\rho_{ij}^2\sum_{i=1}^{p}\sum_{j=1}^{p}c_{ii}c_{jj}\rho_{ij}^2\right\}$$

con $c_{ij}\in C=I-\theta_1\theta_1'$; $\theta_1$ matriz de vectores característicos de las $k$ primeras raices de $\Sigma$. Obsérvese que $p^*$ depende de muchos parámetros desconocidos. Por ejemplo $\rho$ y todos los $\rho_{ij}$ (toda la matriz teórica $\mathbb{R}$). En consecuencia, no disponemos de un verdadero estadístico. En todo caso se hará práctica esa variable y $p^*$, sustituyendo todos los parámetros por sus estimadores de máxima verosimilitud, y en cualquier caso, calculado $p^*$, se aproximará por el entero más próximo. Obviamente un cálculo en ordenador para $p^*$ se hace imprescindible.

A veces uno se contenta con un contrate de hipótesis muy particular, que obviamente interesará rechazar en la mayoría de los casos. Se trata de

$$H_0: \quad \mathbb{R}=I$$

En principio este *test de esfericidad* sobre $\mathbb{R}$, equivale al antes utilizado sobre $S$ (test de esfericidad de Bartlett). En efecto, bajo la hipótesis nula

$$-\left\{N-1-\frac{1}{6}(2p+5)\right\}\ln|\hat{\mathbb{R}}| \;\rightsquigarrow \chi^2_{p(p-1)/2}.$$

Otro test útil es el correspondiente a un caso de estructura teórica conocida de $\Sigma$. En efecto, un ejemplo concreto que se puede abordar y resolver asintóticamente es el caso en que la matriz $\Sigma$ sea así:

$$\Sigma = \left(\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}\rho\right)$$

o lo que es lo mismo, que la matriz de correlaciones sea

$$\mathbb{R}_0 = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

Si se supone el test:

$$\begin{cases} H_0: & \mathbb{R} = \mathbb{R}_0 \\ H_1: & \mathbb{R} \neq \mathbb{R}_0 \end{cases}$$

se puede abordar por el método del cociente de verosimilitudes o bien por el resultado de Lawley (1963); este resultado asintótico es el siguiente: Se rechaza $H_0$ si

$$\frac{N-1}{(1-\bar{r})^2}\left\{\sum_{i<k}(r_{ik}-\bar{r})^2 - \hat{\gamma}\sum_{k=1}^{p}(\bar{r}_k-\bar{r})^2\right\} > \chi^2_{(p+1)(p+2)/2}(\alpha)$$

siendo

$$\bar{r}_k = \frac{1}{p-1}\sum_{\substack{i=1\\ i\neq k}}^{p} r_{ik}; \quad k=1,\ldots,p,$$

$$\bar{r} = \frac{2}{p(p-1)}\sum_{i<k} r_{ik},$$

$$\hat{\gamma} = \frac{(p-1)^2\{1-(1-\bar{r})^2\}}{p-(p-2)(1-\bar{r})^2}.$$

## 1.9. Sobre la selección del número de componentes principales a retener.

Cuando el ACP tiene como objetivo prioritario la reducción de la dimensión de un problema multivariable, es preciso decidir con cuantas componentes principales nos quedamos. Este problema no es fácil pues aunque teóricamente disponemos de algunos tests estadísticos que permiten fundamentar objetivamente tal decisión, la dificultad de éstos junto a las fuertes hipótesis bajo las que se obtienen, hacen inviable o poco útil muchas veces esta vía. Esto es particularmente verdad cuando se actúa con matriz de correlaciones $\mathbb{R}$. Por ello, en la práctica del ACP, existen una serie de criterios prácticos que sientan una metodología aceptada en general.

### 1.9.1. Actuación con matriz de covarianzas muestrales.

En este caso la vía de los tests antes dada es plausible, en general. De todas formas, un análisis de la proporción:

$$\frac{\left(\sum_{i=1}^{k}\hat{l}_i\right)}{\left(\sum_{i=1}^{p}\hat{l}_i\right)}; \quad k<p$$

es la base de elección del número de componentes principales, siendo, desde luego, subjetiva la ley de parada.

### 1.9.2. Actuación con matriz de correlaciones muestrales.

Este caso, al que posiblemente nos vemos avocados en gran parte de los problemas prácticos, por razones ya expuestas, es prácticamente poco abordable por tests estadísticos. Los criterios más utilizados, alternativamente a aquellos, son estos:

1. Criterio de Kaiser (1958), o criterio de raíz característica mayor de 1. Seleccionamos aquellas componentes principales cuyo autovalor es mayor que 1. Tiene su base en que una componente principal cualquiera deberá explicar más varianza que una de las variables originales.

2. Criterio de Catell (1966). ("Screen test"). Consiste en representar, los autovalores en el orden de extracción y analizar el "punto de ruptura" respecto de la recta determinada por los autovalores más pequeños. Catell-Jaspers (1967) sugieren tomar hasta el inmediato antes de comienzo de la recta. (El que aparezcan en una recta indica su trivialidad respecto de los dados antes del punto de ruptura). Esto puede tener complicaciones como varios puntos de ruptura o no haber un punto de ruptura claro.

3. Criterio de Horn (1965). Se representan igual que en el criterio de Catell, los autovalores de las componentes principales. Por otra parte, se consideran $K$ conjuntos de una Normal $p$-variante, de tamaño $N$ todos, y conocemos la estructura de correlación de esa población[12]. Se generan entonces estas $K$ muestras. Se factorizan en CP cada muestra, se calculan los "autovalores-medios" (media aritmética de los autovalores en los $K$ casos) y se representan, el "primer autovalor medio", el "segundo autovalor medio", etc. Algunos pueden ser mayores que 1. Cabe esperar que la ordenada 1 se alcanza en $p/2$. Obsérvese que para estos datos simulados, las CP representan el caso en que todos los autovalores son 1, bajo la hipótesis nula. El criterio de Horn consiste en quedarse con las componentes principales anteriores al punto de cruce.

## 1.10. Análisis de componentes principales y observaciones anómalas.

La explicación técnica de las representaciones gráficas utilizadas la interpretación del Análisis de Componentes Principales, se basa en la idea antes expuesta de considerar el comportamiento de

$$x_j - (\text{ valor asignado por las CP o valor "predicho" para } x_j)$$

es decir, en definitiva, en medir el error cometido al ajustar el dato mediante las componentes principales.

De manera global, esta idea se condensa en un resultado, que sigue, basado en medidas centralizadas:

$$(x_j - \bar{x};\ \ j = 1, \ldots, N).$$

El error de ajuste a $(x_j - \bar{x};\ \ j = 1, \ldots, N)$ mediante una matriz $A = (a_1 \cdots a_N)$, vendrá dado por:

$$error = \sum_{j=1}^{N} (x_j - \bar{x} - a_j)'(x_j - \bar{x} - a_j) = \sum_{i=1}^{p} \sum_{j=1}^{N} (x_{ij} - \bar{x}_i - a_{ij})^2.$$

En definitiva, suponemos que la matriz

$$(x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_N - \bar{x})_{p \times N}$$

es ajustada por la matriz

$$A = (a_1, a_2, \ldots, a_N)_{p \times N}.$$

En general podemos suponer que $rg(A) \leq r < \min(p, N)$. Desde luego si el ACP se efectúa bajo hipótesis de normalidad se podrá precisar mejor esta condición y ver sus implicaciones en $\Sigma$.

[12] Por ejemplo de $\mathbb{R} = I$

Por otro lado, recuérdese que en el ACP muestral, las componentes principales son

$$\hat{y}_i = \hat{e}_i' Z = \hat{e}_{1i} z_1 + \hat{e}_{2i} z_2 + \cdots + \hat{e}_{pi} z_p; \quad i = 1, \ldots, p$$

con variables tipificadas; o bien

$$\hat{y}_i = (\hat{e}_i')_{(1\times p)} X_{(p\times 1)} = \hat{e}_{1i} x_1 + \cdots + \hat{e}_{pi} x_p; \quad i = 1, \ldots, p.$$

En conjunto:

$$\hat{y}_{(p\times 1)} = (\,\hat{e}_1 \quad \cdots \quad \hat{e}_p\,)'_{(p\times p)} \cdot X_{(p\times 1)} = (\,\hat{e}_1 \quad \cdots \quad \hat{e}_p\,)' \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}.$$

Si se consideran los valores de las componentes principales sobre toda la muestra $(x_j; \mathrm{J} = 1, \ldots, N)$, tendremos:

$$\hat{Y}_{(p\times N)} = \begin{pmatrix} \hat{y}_{11} & \cdots & \hat{y}_{1N} \\ \vdots & & \vdots \\ \hat{y}_{N1} & \cdots & \hat{y}_{NN} \end{pmatrix} = (\,\hat{e}_1 \quad \cdots \quad \hat{e}_p\,)'_{(p\times p)} \cdot X_{(p\times N)}.$$

En efecto:

$$\hat{y}_i = \hat{e}_i' X = (\,\hat{e}_{1i} \quad \cdots \quad \hat{e}_{pi}\,) \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} = \hat{e}_{1i} x_1 + \cdots + \hat{e}_{pi} x_p,$$

$$\hat{y} = \begin{pmatrix} \hat{e}_{11} x_1 + \cdots + \hat{e}_{p1} x_p \\ \vdots \\ \hat{e}_{1i} x_1 + \cdots + \hat{e}_{pi} x_p \\ \vdots \\ \hat{e}_{1p} x_1 + \cdots + \hat{e}_{pp} x_p \end{pmatrix} = \begin{pmatrix} \hat{e}_{11} & \cdots & \hat{e}_{p1} \\ \vdots & & \vdots \\ \hat{e}_{1i} & \cdots & \hat{e}_{pi} \\ \vdots & & \vdots \\ \hat{e}_{1p} & \cdots & \hat{e}_{pp} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} =$$

$$= \begin{pmatrix} \hat{e}_1' \\ \vdots \\ \hat{e}_i' \\ \vdots \\ \hat{e}_p' \end{pmatrix} \cdot X = (\,\hat{e}_1 \quad \cdots \quad \hat{e}_i \quad \cdots \quad \hat{e}_p\,)' \cdot X$$

considerando toda la muestra $(x_j;\ j = 1, \ldots, N)$

$$\hat{Y} = \begin{pmatrix} \hat{y}_{11} & \cdots & \hat{y}_{1N} \\ \vdots & & \vdots \\ \hat{y}_{i1} & \cdots & \hat{y}_{iN} \\ \vdots & & \vdots \\ \hat{y}_{p1} & \cdots & \hat{y}_{pN} \end{pmatrix}_{p\times N} = (\,\hat{e}_1 \quad \cdots \quad \hat{e}_i \quad \cdots \quad \hat{e}_p\,)' \begin{pmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & & \vdots \\ x_{i1} & \cdots & x_{iN} \\ \vdots & & \vdots \\ x_{p1} & \cdots & x_{pN} \end{pmatrix} =$$

$$\begin{pmatrix} \hat{e}_1 \\ \vdots \\ \hat{e}_i \\ \vdots \\ \hat{e}_p \end{pmatrix} (\,x_1 \quad \cdots \quad x_N\,) = (\,\hat{e}_1 \quad \cdots \quad \hat{e}_p\,)'_{p\times p} (\,x_1 \quad \cdots \quad x_N\,)_{p\times N}$$

Entonces, despejando, será

$$(\,\hat{e}_1 \quad \cdots \quad \hat{e}_i \quad \cdots \quad \hat{e}_p\,) \cdot \hat{Y} = (\,x_1 \quad \cdots \quad x_N\,)$$

y desarrollándolo nos queda

$$\begin{pmatrix} \hat{e}_{11} & \cdots & \hat{e}_{1i} & \cdots & \hat{e}_{1p} \\ \vdots & & \vdots & & \vdots \\ \hat{e}_{p1} & \cdots & \hat{e}_{pi} & \cdots & \hat{e}_{pp} \end{pmatrix} \cdot \begin{pmatrix} \hat{y}_{11} & \cdots & \hat{y}_{1N} \\ \vdots & & \vdots \\ \hat{y}_{p1} & \cdots & \hat{y}_{pN} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1N} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{iN} \\ \vdots & & \vdots & & \vdots \\ x_{p1} & \cdots & x_{pj} & \cdots & x_{pN} \end{pmatrix}$$

$$( \hat{e}_1 \quad \cdots \quad \hat{e}_i \quad \cdots \quad \hat{e}_p ) \cdot \hat{Y} = ( x_1 \quad \cdots \quad x_j \quad \cdots \quad x_N )$$

de donde

$$( \hat{e}_1 \quad \cdots \quad \hat{e}_i \quad \cdots \quad \hat{e}_p ) \cdot \begin{pmatrix} \hat{y}_{1j} \\ \vdots \\ \hat{y}_{pj} \end{pmatrix} = x_j; \quad j = 1, \ldots, N$$

Es decir

$$\hat{y}_{1j}\hat{e}_1 + \hat{y}_{2j}\hat{e}_2 + \cdots + \hat{y}_{ij}\hat{e}_i + \cdots + \hat{y}_{pj}\hat{e}_p = x_j; \quad j = 1, \ldots, N.$$

donde $(\hat{y}_{1j}; \hat{y}_{2j}; \ldots; \hat{y}_{pj})$ son los valores sobre el elemento $x_j$ de la muestra, de las $p$ componentes principales. Por ejemplo, $\hat{y}_{ij}$ es el valor de la $i$-ésima componente principal sobre $x_j$, que sabemos por otro lado que vale $\hat{y}_{ij} = \hat{e}_i' x_j$.

De modo que se puede escribir:

$$x_j = \hat{y}_{1j}\hat{e}_1 + \hat{y}_{2j}\hat{e}_2 + \cdots + \hat{y}_{ij}\hat{e}_i + \cdots + \hat{y}_{pj}\hat{e}_p =$$

$$(\hat{e}_1' x_j)\hat{e}_1 + (\hat{e}_2' x_j)\hat{e}_2 + \cdots + (\hat{e}_i' x_j)\hat{e}_i + \cdots + (\hat{e}_p' x_j)\hat{e}_p.$$

De las anteriores expresiones se deduce lo siguiente: Si tomamos un conjunto formado por las primeras $q$ componentes principales y el conjunto de las $(p - q)$ últimas, y la parte de $x_j$ que es explicada por ambos conjuntos de componentes principales, es decir:

$$\hat{y}_{1j}\hat{e}_1 + \cdots + \hat{y}_{qj}\hat{e}_q$$

$$\hat{y}_{q+1,j}\hat{e}_{q+1} + \cdots + \hat{y}_{pj}\hat{e}_p$$

y consideramos el ajuste de $x_j$ mediante la primera, la segunda será

$$x_j - (\hat{y}_{1j}\hat{e}_1 + \cdots + \hat{y}_{qj}\hat{e}_q)$$

una medida del "error" cometido en la aproximación. La expresión

$$\hat{y}_{q+1,j}^2 + \cdots + \hat{y}_{pj}^2$$

nos da la longitud al cuadrado, como error cometido. Este será grande en la medida en que sobre alguno de los "ejes principales" $(\hat{e}_{q+1}; \cdots; \hat{e}_p)$ la coordenada respectiva sea grande.

Es claro que esta medida del error será menor, por otra parte, cuanto mejor sea el ajuste del dato $x_j$ por las $q$ primeras CP, y es claro también que si una observación es *estructuralmente anómala* frente a las demás, provocará que el error sea grande, al ser grande, por ejemplo, una coordenada.

$$\hat{y}_{i,j}^2; \quad i = q + 1, \ldots, p.$$

Pero este análisis se ha realizado sobre una componente $x_j$, individualmente considerada. ¿Se podrá expresar el error cometido al ajustar *todos* los datos $x_j;\ j = 1, \ldots, N$ por el grupo de las primeras $q$ componentes principales?

Porque lo anterior, con $x_j$ individuales, puede servir como un método de detección de observaciones anómalas, supuesto que estructuralmente las $q$ componentes principales primeras ajustan bien al conjunto de las

observaciones y se buscan entonces las que estructuralmente son erróneas ("outliers"). Es preciso entonces conocer el error global sobre toda la muestra (j=1,...,N) que se comete al aproximar estructuralmente por las primeras $q$ componentes principales todos los elementos de la muestra. Para ello es preciso analizar la 'geometría'" del ACP muestral.

De entrada, volviendo al planteamiento dado en al principio, puede probarse este resultado.

**Teorema 6.** *Al aproximar* $(x_j - \bar{x};\ j = 1, \ldots, N)$ *por una matriz* $A = (\, a_1 \quad \cdots \quad a_N \,)$, *con rango* $rg(A) \leq r < \text{mín}\,(p, N)$, *el error global*

$$\sum_{j=1}^{N} (x_j - \bar{x} - a_j)'(x_j - \bar{x} - a_j)$$

*se minimiza cuando se toma por* $A$ *la matriz* $\hat{A} = \hat{E}\,(\, \hat{y}_1 \quad \cdots \quad \hat{y}_r \,)'$ *donde*

$$\hat{E} = (\, \hat{e}_1 \quad \cdot \quad \hat{e}_r \,)$$

*formada con los primeros* $r$ *autovectores. De modo que*

$$\hat{A}_{p\times N} = (\, \hat{e}_1 \quad \cdots \quad \hat{e}_r \,)_{p\times r} \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_r \end{pmatrix}_{r\times N} = (\, \hat{a}_1 \quad \cdots \quad \hat{a}_N \,)$$

*con* $\hat{a}_j = \hat{y}_{1j}\hat{e}_1 + \cdots + \hat{y}_{rj}\hat{e}_r$ *y siendo*

$$(\, \hat{y}_{1j} \quad \cdots \quad \hat{y}_{rj} \,) = (\, \hat{e}_1'(x_j - \bar{x}) \quad \cdots \quad \hat{e}_r'(x_j - \bar{x}) \,)$$

*los valores de las primeras* $r$ *componentes principales muestrales sobre el elemento* $j$*-ésimo de la muestra, centrado en* $\bar{x}$.

*El mínimo alcanzado ("Error Cuadrático") vale:*

$$\sum_{j=1}^{N} (x_j - \bar{x} - a_j)'(x_j - \bar{x} - a_j) = (N-1)(\hat{\lambda}_{r+1} + \cdots + \hat{\lambda}_p).$$

**Nota 6.** *Este teorema nos da pues el error cometido al aproximar toda la muestra por las primeras* $r$ *componentes principales y, además, nos lo expresa en términos de los autovalores muestrales. Pero también nos interpreta el significado de las componentes principales obtenidas mediante la minimización de un error cuadrático cometido al aproximar la muestra centrada por los* $a_j$*: se minimiza el error cuando la aproximación* $A$ *se construye precisamente con las* $r$ *primeras componentes principales, con* $rg(A) \leq r < \text{mín}\,(p, N)$.

## 1.11. Representaciones gráficas en el ACP.

Basándonos en los resultados de la interpretación geométrica del ACP, pueden establecerse unas útiles prácticas gráficas, que recogen estas ideas.

En primer lugar, es de interés comprobar la *normalidad* de las componentes principales primeras, lo cual se realiza efectuando las representaciones gráficas de los pares $(\hat{y}_i, \hat{y}_l)$ de componentes principales. Una normalidad conjunta puede aceptarse si el contorno de los valores de $(\hat{y}_i, \hat{y}_l)$ sobre $(x_j;\ j = 1, \ldots, N)$ es sensiblemente elíptico, para valores no anómalos.

En segundo lugar se representan, vía un "Q-Q-plot", los valores de cada componentes principales sobre la muestra $(y_j,\ j = 1, \ldots, N)$, en la idea de detectar también valores anómalos.

Ambas cosas conviene hacerlas también con las últimas componentes principales.

## 1.12. Aplicaciones del ACP: ACP sobre k-grupos.

En la práctica podemos encontrarnos con que la muestra $(x_j;\ j=1,\ldots,N)$ proviene de varias poblaciones distintas, con lo que esa muestra global no es aleatoria independiente. Cuando esto ocurre, como para el problema con $K$ muestras respecto del vector de medias, podemos optar por uno de estos dos caminos:

1. Aplicar ACP a cada grupo, por separado, y comparar las componentes principales deducidas en cada caso.
2. Plantear un tratamiento global de la situación, como es el ANOVA respecto de un test de diferencia de medias dos a dos.

Desde luego, queda fijado que en esta situación de varios grupos, el objetivo es contrastar si son homogéneos respecto de su estructura de componentes principales, si es conocida la estructura de los grupos. Si esta estructura no se conoce, se hará el ACP sobre toda la muestra, y puede éste utilizarse para obtener posibles cluster o grupos entre ellos.

Vamos a suponer a continuación algunos modelos que se han propuesto para abordar esta situación.

### 1.12.1. Modelo de Okamoto (1976) o "modelo de efectos fijos"

Supongamos definidas las componentes principales escritas de manera centrada:

$$\hat{y}_i = \hat{e}_i'(x-\bar{x});\quad \hat{Y} = (\,\hat{e}_1 \quad \cdots \quad \hat{e}_p\,)'\,(X_1-\bar{x});$$

$$X^*_{p\times N} = (\,\hat{e}_1 \quad \cdots \quad \hat{e}_p\,)_{p\times p}\,\hat{Y}_{p\times N}$$

[13] que al recorrer la muestra $(x_j;\ j=1,\ldots,N)$ dará los valores:

$$\hat{y}_i;\quad \hat{y}_{ij} = \hat{e}_i'(x_j-\bar{x})\quad i=1,\ldots,p\ (CP);\ j=1,\ldots,N.$$

Ya vimos que $x_j = \hat{y}_{1j}\hat{e}_1+\cdots+\hat{y}_{pj}\hat{e}_p$, de donde, si tomamos $q$ componentes principales

$$x_j = \hat{y}_{1j}\hat{e}_1+\cdots+\hat{y}_{qj}\hat{e}_q+\hat{y}_{q+1,j}\hat{e}_{q+1}+\cdots+\hat{y}_{pj}\hat{e}_p$$

de donde:

$$x_{jl} = \hat{y}_{1j}\hat{e}_{1l}+\cdots+\hat{y}_{qj}\hat{e}_{ql}+\sum_{r=q+1}^{p}\hat{y}_{rj}\hat{e}_{rl}$$

($l$-ésima componente de $x_j;\ j=1,\ldots,p$).

$$x_{jl} = \bar{x}_l+\hat{y}_{1j}\hat{e}_{1l}+\cdots+\hat{y}_{ql}\hat{e}_{ql}+\sum_{r=q+1}^{p}\hat{y}_{rj}\hat{e}_{rl}$$

lo que sugiere el modelo teórico:

$$x_{jl} = \beta_l+\sum_{r=1}^{q}p_{rj}m_{rl}+\varepsilon_{jl};\quad j=1,\ldots,N;\ i=1,\ldots,p$$

en el que se supone impuesta un estructura: Los $m_{rl}$ y $p_{rj}$ son tales que verifican propiedades análogas a las verificadas por los $\hat{e}_i$ (ortogonalidad) y por la covarianzas entre $\hat{y}_{lj}$ (covarianzas nulas).

Bajo esta estructura, puede efectuarse un "análisis de varianza", como puede verse en Okamoto ("Random models and fixed model of PCA" ed. Ikeda: Essays in Prob. and Stat. (dedicated to J. Ogawa) pgs. 339-351. Tokyo).

[13] $X_{p\times N}-\bar{x}_{p\times 1}\cdot 1'_{1\times N} = X-(\bar{x}\ \overbrace{\cdots}^{N}\ \bar{x})_{p\times N}$ ya que $\bar{x}_{p\times 1}\cdot 1'_{1\times N} = \bar{x}(1\ \overbrace{\cdots}^{N}\ 1) = (\bar{x};\bar{x}\ \overbrace{\cdots}^{N}\ \bar{x})$

### 1.12.2. El ACP y la Regresión Lineal (Latenet root regression)

Consideremos un modelo de Regresión Lineal Múltiple:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i; \quad i = 1, \ldots, N$$

o bien:

$$y_i = \alpha + \beta_1 (x_{i1} - \bar{x}_{\cdot 1}) + \cdots + \beta_{p-1}(x_{i,p-1} - \bar{x}_{\cdot p-1}) + \varepsilon$$

que matricialmente, como Modelo Lineal, se expresa:

$$y_{N\times 1} = \alpha \cdot 1_N + \bar{\bar{X}}_{N\times p}\beta_{p\times 1} + \varepsilon_{N\times 1}; \quad \beta = (\alpha; \beta_1; \cdots; \beta_{p-1})'$$

$$\bar{\bar{X}} = (\, x_1 - \bar{x} \quad \cdots \quad x_N - \bar{x}\,)'_{N\times p}\,.$$

Sea entonces la matriz de cuadrados

$$(\bar{\bar{X}}'\bar{\bar{X}})_{p\times p} = \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})'$$

tal que $\dfrac{1}{N-1}(\bar{\bar{X}}'\bar{\bar{X}})$ es la matriz de covarianzas muestrales sobre la que se efectúa el ACP muestral.

Supongamos que un autovalor muestral $\hat{\lambda}$ es próximo a cero y su correspondiente vector es $\hat{e}$. Entonces:

$$(\bar{\bar{X}}'\bar{\bar{X}})\hat{e} - \hat{\lambda}\hat{e} = 0 \Rightarrow \bar{\bar{X}}'\bar{\bar{X}}\hat{e} \approx 0 \Rightarrow \hat{e}'\bar{\bar{X}}'\bar{\bar{X}}\hat{e} \approx 0 \Rightarrow \bar{\bar{X}}\hat{e} \approx 0$$

y ello significa que hay multicolinealidad.

Si hay un cierto número de restricciones lineales, por ejemplo $p - k$, entonces:

$$\bar{\bar{X}} \cdot \hat{\mathbb{E}}_2 \approx 0; \quad \hat{\mathbb{E}} = (\hat{\mathbb{E}}_1 | \hat{\mathbb{E}}_2)$$

(siendo $\mathbb{E}_2$ $p \times (p-k)$). En este caso general, el Modelo Lineal de Regresión se puede volver a escribir en términos de as componentes principales de $\mathbb{E}_1$, es decir, de $R$ componentes principales no nulas. En efecto:

$$\bar{\bar{X}}\beta = \bar{\bar{X}}\hat{\mathbb{E}}(\hat{\mathbb{E}}'\beta) = (\bar{\bar{X}}\hat{\mathbb{E}}_1|0)(\hat{\mathbb{E}}'\beta) = (\bar{\bar{X}}\hat{\mathbb{E}}_1)(\hat{\mathbb{E}}'\beta).$$

## 1.13. Resultados previos: Elipsoides equiprobables en una $N_p(\mu; \Sigma)$ y combinaciones lineales de un vector aleatorio multidimensional.

Supongamos un vector aleatorio $X \rightsquigarrow N_p(\mu; \Sigma)$, con $\Sigma$ definida positiva. Si en esta densidad se considera la familia de elipsoides

$$(X - \mu)'\Sigma^{-1}(X - \mu) = c; \quad c > 0$$

es claro que tal densidad es constante en cada elipsoide, correspondiente a un $c$ concreto. Por otra parte, dicha familia tiene como centro al vector $\mu$, mientras que las características de $\Sigma$ determinan la forma y orientación de los elipsoides. Por ejemplo, en el caso $p = 2$, dicha familia es de elipses.

Como es sabido, existe un elipsoide especial obtenido cuando $c = p+2$, llamado "elipsoide de concentración" (Cramer (1946)), caracterizado por verificar la siguiente propiedad:

$$f(x) = \begin{cases} \dfrac{\Gamma\left(\frac{1}{2}+1\right)}{|\Sigma|^{-1}(p+2)^{p/2}\pi^{p/2}} & ;\ si\ (x-\mu)'\Sigma^{-1}(x-\mu) \le p+2 \\ 0 & ;\ fuera \end{cases}$$

tiene la misma media y matriz de covarianzas que la ley $N_p(\mu;\Sigma)$.

Volviendo a la familia de elipsoides de equiprobabilidad nos planteamos el cálculo de los "ejes principales", y ello lo hacemos por un método analítico (multiplicadores de Lagrange) en vista de la metodología que luego se utilizará en el Análisis de Componentes Principales. En efecto; supongamos una recta desde el centro $\mu$ a la superficie del elipsoide, dada por sus coordenadas sobre dicha superficie. Es claro que el eje principal (primero) del elipsoide ha de cumplir:

$$\begin{cases} \max\limits_{x} [(x-\mu)'(x-\mu)] \\ sujeto\ a\ (x-\mu)'\Sigma^{-1}(x-\mu) = c \end{cases}$$

Obsérvese que $(x-\mu)'(x-\mu)$ es el cuadrado de la semilongitud de tal eje principal cuando, en efecto, $x$ está en la superficie considerada, a un punto para el que se verifique el máximo indicado.

Resolviendo este problema de máximo por multiplicadores de Lagrange, se tendrá

$$\Phi(x,\lambda) = (x-\mu)'(x-\mu) - \lambda(x-\mu)'\Sigma^{-1}(x-\mu)$$

de donde el $x$ buscado verificará

$$\frac{\partial\Phi(x,\lambda)}{\partial x} = (x-\mu) - \lambda\Sigma^{-1}(x-\mu) = 0$$

es decir $(I-\lambda\Sigma^{-1})(x-\mu)=0$; $(\Sigma-\lambda I)(x-\mu)=0$.

Si suponemos, como se dijo al principio, que $\Sigma > 0$, entonces todas sus raices características (soluciones de $|\Sigma-\lambda I|=0$) son reales y no nulas: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Por tanto, si tomamos la mayor de ellas, $\lambda_1$, es claro que:

1. El eje principal mayor está en la dirección determinada por vector característico $e_1$ asociado a dicha raíz.
2. El cuadrado de la longitud del eje principal considerado valdrá $4\lambda_1 c$, ya que

$$4(x-\mu)'(x-\mu) = 4\lambda_1(x-\mu)'\Sigma^{-1}(x-\mu) = 4\lambda_1 c$$

¿Cómo calcular los restantes ejes principales del elipsoide? Pues volviendo a reiterar el cálculo, tomando las sucesivas raices características de $\Sigma$, en orden decreciente, y los respectivos autovectores se van obteniendo por los mismos argumentos utilizados para el eje principal mayor.

En el caso en que haya una raíz característica multiple, con un orden $r$ de multiplicidad, el elipsoide es hiperesférico en el subespacio $r$-dimensional correspondiente. Y, obviamente, si todas las raices características son diferentes, sus autovectores correspondientes (y por tanto los ejes principales asociados) son ortogonales, ya que a dos raices distintas corresponden autovectores ortogonales.

Es interesante, en este punto, con vistas al posterior Análisis de Componentes Principales, utilizar los ejes principales calculados en la familia de elipsoides para definir una "transformación a los ejes principales". Recordemos que estamos en el caso de una $N_p(\mu;\Sigma)$, gracias a lo cual podemos hablar de ejes principales en su sentido geométrico. En efecto, sea la transformación

$$X \rightsquigarrow N_p(\mu;\Sigma) \longrightarrow Y = (Y_1,\ldots,Y_p)' = A(X-\mu)$$

en donde $A = (e_1,\ldots,e_p)$ con $e_1,\ldots,e_p$ autovectores normalizados de $\Sigma > 0$. Obviamente, al ser $X \rightsquigarrow N_p(\mu;\Sigma)$, se tiene

$$Y \rightsquigarrow N_p(0; A'\Sigma A)$$

Supongamos que todas las raices $\lambda_i$ de $\Sigma$ son distintas, entonces $A$ es ortogonal, es decir: $A'A = I \Leftrightarrow A' = A^{-1}$. Por tanto, tenemos una transformación:

$$X \rightsquigarrow Y = A(X-\mu).$$

tal que $A'\Sigma A$ es diagonal; es decir, que las componentes $Y_i$ de $Y$ son incorreladas. Los elementos de la diagonal principal no nula de $A'\Sigma A$ son las varianzas de las $Y_i$.

En definitiva, es posible definir una transformación ortogonal (giro) llevando el sistema de referencia al origen $\mu$ y girando los ejes hasta coincidir con los ejes principales, de tal forma que se transforma el vector $X$ en uno $Y$ que, respecto de dicho sistema nuevo, tiene sus componentes incorreladas, de tal forma además, que la longitud de los ejes de cualquier elipsoide dado ($c > 0$) es proporcional a la varianza de las variables $Y_i$.

Finalmente cabe preguntarse lo siguiente: ¿Se podrá definir esta transformación ortogonal así cuando $X$ no sea $N_p(\mu; \Sigma)$? Después se verá la respuesta.

### 1.13.1. Combinaciones lineales de un vector aleatorio $X$.

Dado un vector aleatorio $X = (X_1, \ldots, X_p)'$, no necesariamente normal, con media $E[X] = \mu$ y matriz de covarianzas $Cov(X) = \Sigma$, es claro que si tomamos una combinación lineal

$$\alpha' X; \text{ con } \alpha = (\alpha_1, \ldots, \alpha_p)' \in \mathbb{R}^p$$

se verifica

$$E[\alpha' X] = \alpha' \mu; \quad Cov(\alpha' X] = \alpha' \Sigma \alpha$$

Obviamente, como ya se ve en el estudio general de la normal multivariante, si $X \rightsquigarrow N_p(\mu; \Sigma)$ y tomamos combinaciones lineales $\alpha' X$, entonces

$$\alpha' X \rightsquigarrow N_p(\alpha' \mu; \alpha' \Sigma \alpha)$$

**Nota 7.** *El "caracter cerrado" por combinaciones lineales, de un cierto tipo o clase de distribuciones, es siempre de interés en Cálculo de Probabilidades y no es, naturalmente, privativo de la familia Normal.*