# Capítulo 3

# Análisis de componentes principales

## 3.1. INTRODUCCIÓN

El Análisis de componentes principales trata de describir las características principales de un conjunto de datos multivariantes, en los que se han medido $p$ variables cuantitativas a $n$ objetos.

Esto se realiza construyendo un conjunto de variables incorreladas (componentes principales) que son combinación lineal de las variables originales.

Las componentes principales están ordenadas en órden decreciente de importancia: la primera componente recoge la mayor variabilidad posible de los datos; la segunda recoge la mayor variabilidad entre las variables incorreladas con la primera, etc.

Seleccionando las componentes más importantes se resume la información que proporcionan los datos con poca pérdida de información.

Los datos se podrán representar en un espacio de menor dimensión con ejes ortogonales.

**Ejemplo 3.1** Ventas de dos productos en 25 establecimientos nacionales.

| $X_1$ | $X_2$ | $X_1$ | $X_2$ | $X_1$ | $X_2$ |
|---|---|---|---|---|---|
| 191 | 155 | 179 | 158 | 192 | 154 |
| 195 | 149 | 183 | 147 | 174 | 143 |
| 181 | 148 | 174 | 150 | 176 | 139 |
| 183 | 153 | 190 | 159 | 197 | 167 |
| 176 | 144 | 188 | 151 | 190 | 163 |
| 208 | 157 | 163 | 137 | | |
| 189 | 150 | 195 | 155 | | |
| 197 | 159 | 186 | 153 | | |
| 188 | 152 | 181 | 145 | | |
| 192 | 150 | 175 | 140 | | |

Vector de medias y la matriz de covarianzas:

$$\overline{X} = \begin{pmatrix} 185{,}7 \\ 151{,}1 \end{pmatrix}, \quad S_X = \begin{pmatrix} 95{,}29 & 52{,}87 \\ 52{,}87 & 54{,}36 \end{pmatrix}$$

Variables centradas: $Y_1 = X_1 - \bar{x}_1$ e $Y_2 = X_2 - \bar{x}_2$.

Centramos los datos y representamos los puntos centrados

$$P_i = (y_{i1}, y_{i2}), \quad i = 1, 2, \ldots, 25.$$

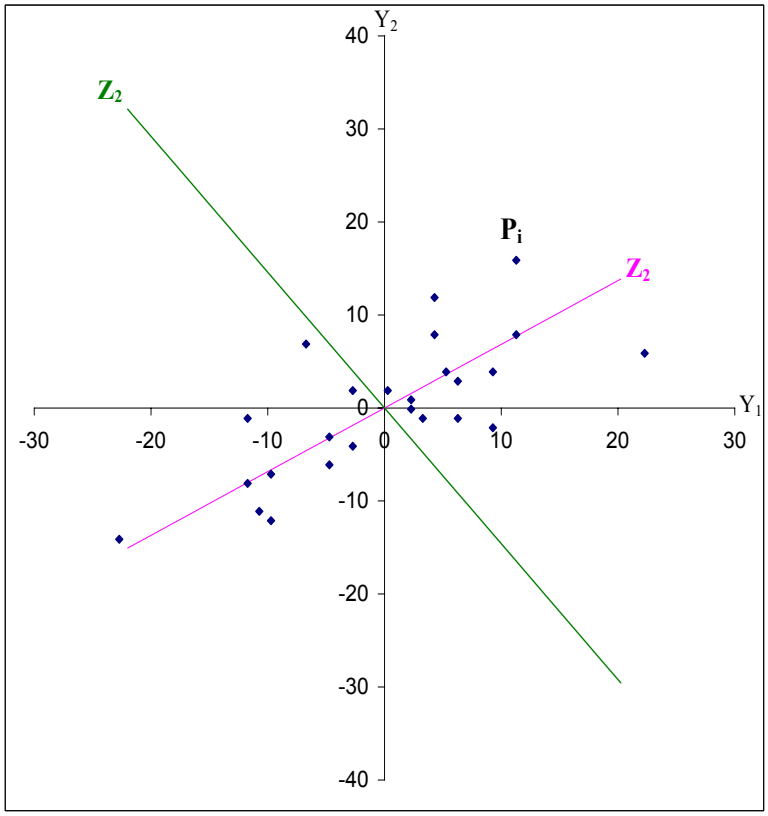


Figura 3.1: Nube de puntos $(y_{i1}, y_{i2})$, $i = 1, \ldots, 25$.

Se desea construir dos nuevos ejes que representen mejor los datos.

$P_i = (y_{i1}, y_{i2})$ punto genérico de la nube
$\mathbf{p}_i' = (y_{i1}, y_{i2})$ vector que va desde el origen hasta el punto $P_i$
$\mathbf{a}_1' = (a_{11}, a_{21})$ vector director (unitario) de un nuevo eje $Z_1$
$\mathbf{a}_2' = (a_{12}, a_{22})$ vector director (unitario) de un nuevo eje $Z_2$

Coordenadas de $P_i$ en los nuevos ejes $Z_1$ y $Z_2$:

$$z_{i1} = \ldots \quad ,$$
$$z_{i2} = \ldots \quad ,$$

Estos son los valores del establecimiento $i$ respecto a dos nuevas variables $Z_1$ y $Z_2$.

Por tanto, buscar dos nuevos ejes donde representar los datos es equivalente a buscar dos nuevas variables $Z_1$ y $Z_2$ que son combinación lineal de las variables originales.

El primer eje va a ser el que minimiza la distancia euclídea de cada punto a su proyección ortogonal (o coordenada) en dicho eje.

El vector director del primer eje, $\mathbf{a}_1' = (a_{11}, a_{21})$, se obtiene resolviendo el problema

$$\begin{array}{ll} \text{mín}_{\mathbf{a}_1} & \sum_{i=1}^{25} z_{i2}^2 \\ \text{s.a.} & \mathbf{a}_1'\mathbf{a}_1 = 1\,. \end{array}$$

Pero por el teorema de Pitágoras,

$$d(O, P_i)^2 = z_{i1}^2 + z_{i2}^2\,,$$

y despejando $z_{i2}^2$, tenemos que

$$z_{i2}^2 = d(O, P_i)^2 - z_{i1}^2\,.$$

Además, $d(O, P_i)^2$ es una cantidad que no depende de $\mathbf{a}_1' = (a_{11}, a_{21})$; es decir, para cualquier valor que tome $\mathbf{a}_1$, la distancia es la misma. Por tanto, encontrar el vector $\mathbf{a}_1$ que minimiza $\sum_{i=1}^{25} z_{i2}^2$ es equivalente a encontrar el vector $\mathbf{a}_1$ que maximiza $\sum_{i=1}^{25} z_{i1}^2$.

Por otro lado, al estar $Y_1$ e $Y_2$ centradas ($\bar{y}_1 = \bar{y}_2 = 0$), entonces $Z_1$ también está centrada, ya que

$$\bar{z}_1 = ... \qquad\qquad ,$$

y por tanto la varianza de las proyecciones en el eje $Z_1$ es

$$s^2_{Z_1} = \frac{1}{n}\sum_{i=1}^{n} z^2_{i1}\,.$$

Así, el eje principal que minimiza la distancia euclídea de los puntos a sus proyecciones ortogonales, es el que maximiza la varianza de las proyecciones. Es decir, el primer eje principal es el que produce la mayor separación de los puntos respecto de su origen.

El problema a resolver para encontrar dicho eje es

$$\begin{array}{ll} \text{máx}_{a_{11},a_{21}} & s^2_{Z_1} \\ \text{s.a.} & a^2_{11} + a^2_{21} = 1\,. \end{array}$$

La segunda componente es una recta ortogonal a $Z_1$.

Interpretación de los ejes:

- $Z_1$ separa a los establecimientos que venden muchas unidades de ambos productos, de los establecimientos que venden pocas unidades. Por tanto, se puede interpretar que $Z_1$ es la variable que mide la proporcionalidad en la venta de ambos productos.
- $Z_2$ da los establecimientos que venden muchas unidades de uno de los productos y pocas del otro. Sin embargo, dado que hay pocos establecimientos con estas características, podemos considerar que este eje no representa demasiado a los datos que tenemos. Si lo eliminásemos, nos quedaríamos con un espacio de dimensión uno.

## 3.2. DEFINICIÓN DE LAS COMPONENTES PRINCIPALES

Tabla de datos de $p$ variables $X_1, \ldots, X_p$ medidas a $n$ individuos:

| Indiv. | $X_1$ | $X_2$ | $\cdots$ | $X_p$ |
|---|---|---|---|---|
| 1 | $x_{11}$ | $x_{12}$ | $\cdots$ | $x_{1p}$ |
| 2 | $x_{21}$ | $x_{22}$ | $\cdots$ | $x_{2p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| n | $x_{n1}$ | $x_{n2}$ | $\cdots$ | $x_{np}$ |

Variables centradas:

$$Y_j = X_j - \bar{x}_j, \quad j = 1, \ldots, p$$

Datos centrados:

$$y_{ij} = x_{ij} - \bar{x}_j, \ j = 1, \ldots, p, \ i = 1, \ldots, n\,,$$

Matriz de datos centrados:

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix}.$$

Vector de variables aleatorias centradas: $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)'$
matriz de covarianzas de $Y$: $S_Y = \frac{1}{n} Y'Y$

**Primera componente principal ($Z_1$):** Combinación lineal de $Y_1, \ldots, Y_p$ con varianza máxima.

Sea $\mathbf{a}_1 = (a_{11}, a_{21}, \ldots, a_{p1})'$ el vector de coeficientes de la combinación lineal. La primera componente principal es

$$Z_1 = \mathbf{a}_1'\mathbf{Y} = a_{11}Y_1 + a_{21}Y_2 + \cdots + a_{p1}Y_p\,.$$

Su varianza es

$$s^2_{Z_1} = \ldots \qquad .$$

El vector $\mathbf{a}_1$ se obtiene resolviendo el problema

$$\begin{array}{ll} \text{máx}_{\mathbf{a}_1} & s^2_{Z_1} = \ldots \\ \text{s.a.} & \mathbf{a}_1'\mathbf{a}_1 = 1\,. \end{array}$$

Llamamos $\lambda_1$ al valor máximo solución de este problema.

**Segunda componente principal ($Z_2$):** Combinación lineal de $Y_1, \ldots, Y_p$, incorrelada con $Z_1$, con varianza máxima.

Sea $\mathbf{a}_2 = (a_{12}, a_{22}, \ldots, a_{p2})'$ el vector de coeficientes de la combinación lineal. La segunda componente principal es

$$Z_2 = \mathbf{a}_2'\mathbf{Y} = a_{12}Y_1 + a_{22}Y_2 + \cdots + a_{p2}Y_p\,.$$

Su varianza es

$$s^2_{Z_2} = \ldots \qquad .$$

La covarianza entre $Z_1$ y $Z_2$ es

$$s_{Z_1,Z_2} = \ldots \qquad .$$

El vector $\mathbf{a}_2$ se obtiene resolviendo el problema

$$\begin{array}{ll} \text{máx}_{\mathbf{a}_2} & s^2_{Z_2} = \ldots \\ \text{s.a.} & \mathbf{a}_2'\mathbf{a}_2 = 1\,, \\ & \mathbf{a}_1'S_Y\mathbf{a}_2 = 0\,. \end{array}$$

Si $\lambda_2$ es este máximo, se verifica $\lambda_1 \geq \lambda_2$.

**Componente principal $j$-ésima ($Z_j$):** Combinación lineal de $Y_1, \ldots, Y_p$, incorrelada con $Z_1$, $Z_2$, …, $Z_{j-1}$, con varianza máxima.

Sea $\mathbf{a}_j = (a_{1j}, a_{2j}, \ldots, a_{pj})'$ el vector de coeficientes de la combinación lineal. La $j$-ésima componente principal es

$$Z_j = \mathbf{a}_j'\mathbf{Y} = a_{1j}Y_1 + a_{2j}Y_2 + \cdots + a_{pj}Y_p\,.$$

Su varianza es

$$s^2_{Z_j} = \ldots \qquad .$$

La covarianza entre $Z_i$ y $Z_j$ es

$$s_{Z_i,Z_j} = \ldots \qquad .$$

Por tanto, $\mathbf{a}_j$ se obtiene resolviendo el problema

$$\begin{array}{ll} \text{máx}_{\mathbf{a}_j} & s^2_{Z_j} = \mathbf{a}_j' S_Y \mathbf{a}_j \\ \text{s.a.} & \mathbf{a}_j'\mathbf{a}_j = 1\,, \\ & \mathbf{a}_i' S_Y \mathbf{a}_j = 0,\ i = 1, \ldots, j-1\,. \end{array}$$

Si $\lambda_j$ es este máximo, se verifica $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{j-1} \geq \lambda_j$. Esto ocurre para $j = 1, \ldots, p$.

**Solución - Descomposición espectral de $S_Y$:**

Denotamos por $\mathbf{Z} = (Z_1, \ldots, Z_p)'$ al vector de componentes principales, y $A = [\mathbf{a}_1\ \mathbf{a}_2\ \ldots\ \mathbf{a}_p]$. Expresamos la el vector de componentes principales en función del vector de variables originales:

$$\mathbf{Z} = A'\mathbf{Y}\,.$$

La matriz de varianzas-covarianzas de $\mathbf{Z}$ es

$$S_Z = A' S_Y A = \text{diag}\{S^2_{Z_1}, \ldots, S^2_{Z_p}\}\,.$$

Por otro lado, al ser $S_Y$ simétrica, el Teorema de Descomposición Espectral nos dice que existe una matriz ortogonal $P$ y otra matriz diagonal $\Lambda$, tales que

$$S_Y = P\Lambda P' \Leftrightarrow \Lambda = P' S_Y P = \text{diag}\{\lambda_1, \ldots, \lambda_p\}\,,$$

donde las columnas de $P$ son los vectores propios de $S_Y$ normalizados, y los elementos de $\Lambda$ son los valores propios de $S_Y$.

El siguiente resultado nos dice que las dos igualdades anteriores coinciden, es decir, $P = A$ y $\Lambda = S_Z$, con lo cual, se verifica:

- ✓ $\mathbf{a}_j$ es el vector propio normalizado de $S_Y$ asociado a $\lambda_j$.
- ✓ $s^2_{Z_j} = \lambda_j$, donde $\lambda_j$ es el $j$-ésimo mayor valor propio de $S_Y$.

**Teorema 3.1** Sea $S_Y$ la matriz de covarianzas asociada al vector $\mathbf{Y} = (Y_1, \ldots, Y_p)'$. Sean $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ los valores propios de $S_Y$, con vectores propios asociados $\mathbf{a}_1$, $\mathbf{a}_2$, ..., $\mathbf{a}_p$. Se verifica:

(i) La componente principal $i$-ésima es

$$Z_i = \mathbf{a}_i' \mathbf{Y} = a_{1i} Y_1 + a_{2i} Y_2 + \cdots + a_{pi} Y_p,\ i = 1, \ldots, p\,.$$

(ii) La varianza de la componente principal $i$-ésima es

$$s^2_{Z_i} = \mathbf{a}_i' S_Y \mathbf{a}_i = \lambda_i,\ i = 1, \ldots, p\,.$$

(iii) La covarianza entre dos componentes $Z_i$ y $Z_j$ es

$$s_{Z_i Z_j} = \mathbf{a}_i' S_Y \mathbf{a}_j = 0,\ \forall i \neq j\,.$$

**Ejemplo 3.2** En el Ejemplo 3.1, los valores y vectores propios de $S_Y = S_X$ son

$$\lambda_1 = 131{,}52, \;\; \mathbf{a}_1' = (a_{11}, a_{21}) = (0{,}825, 0{,}565)\,,$$
$$\lambda_2 = 18{,}146, \;\; \mathbf{a}_2' = (a_{12}, a_{22}) = (-0{,}565, 0{,}825)\,.$$

Así, las dos componentes principales serían

$$Z_1 = ...$$
$$Z_2 = ...$$

El primer eje principal es la recta que pasa por el origen y con vector director $\mathbf{a}_1 = (0{,}825, 0{,}565)'$, cuya ecuación es

$$Y_2 = ... \qquad Y_1\,.$$

**Corolario 3.1** Sea el vector de variables $\mathbf{Y} = (Y_1, \ldots, Y_p)'$, con matriz de covarianzas $S_Y$. Sean $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ los valores propios de $S_Y$, y $\mathbf{a}_1$, $\mathbf{a}_2$, ..., $\mathbf{a}_p$ los vectores propios asociados. Sean las componentes principales

$$\begin{aligned} Z_1 &= \mathbf{a}_1'\mathbf{Y}, \\ Z_2 &= \mathbf{a}_2'\mathbf{Y}, \\ &\vdots \\ Z_p &= \mathbf{a}_p'\mathbf{Y} \end{aligned}$$

Las componentes principales conservan la varianza total, es decir

$$\mathrm{tr}(S_Y) = \sum_{i=1}^{p} s_{Y_i}^2 = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} s_{Z_i}^2\,.$$

**Ejemplo 3.3** Para el Ejemplo 3.1, sabemos que $S_Y = S_X$, con lo cual, podemos calcular la traza de $S_Y$

$$\text{tr}(S_X) = ... \qquad .$$

Por otro lado, sumando los valores propios obtenemos

$$\lambda_1 + \lambda_2 = ... \qquad .$$

**Nota 3.1** En la Figura 3.1 se puede observar cómo las observaciones se pueden envolver en una elipsoide. Esto ocurre cuando el vector $(Y_1, Y_2)$ tiene una distribución normal bivariante. Entonces, el primer eje principal tiene la dirección del eje principal de la elipsoide, y el segundo eje tiene la dirección del eje secundario de la elipsoide. Además, $\sqrt{\lambda_1}$ es la longitud del semieje principal de la elipse, y $\sqrt{\lambda_2}$ es la del semieje secundario.

## 3.3. SELECCIÓN E INTERPRETACIÓN DE LAS COMPONENTES

### 3.3.1. ELECCIÓN DEL NÚMERO DE COMPONENTES

En la literatura se proponen tres criterios:

- Retener un número de componentes tales que en conjunto recojan un porcentaje de variabilidad de al menos un 75 %.

  Proporción de la variabilidad total de la $i$-ésima componente:

$$\frac{\lambda_i}{\text{tr}(S_Y)} = \frac{\lambda_i}{\sum_{k=1}^{p} \lambda_k}, \ i = 1, \ldots, p\,.$$

  Proporción de variabilidad de las primeras $p_1 < p$ componentes:

$$\frac{\sum_{k=1}^{p_1} \lambda_i}{\text{tr}(S_Y)} = \frac{\sum_{k=1}^{p_1} \lambda_k}{\sum_{k=1}^{p} \lambda_k}\,.$$

- Seleccionar las componentes con valores propios (varianzas) mayores que el valor propio (varianza) promedio:

$$\overline{\lambda} = \frac{1}{p} \sum_{k=1}^{p} \lambda_k \, .$$

- Usar el "test scree" o gráfico de sedimentación, en el que se representa el número de órden del valor propio $i$ frente a su magnitud $\lambda_i$. Se descartan las componentes que ya no aportan significativamente a la varianza.

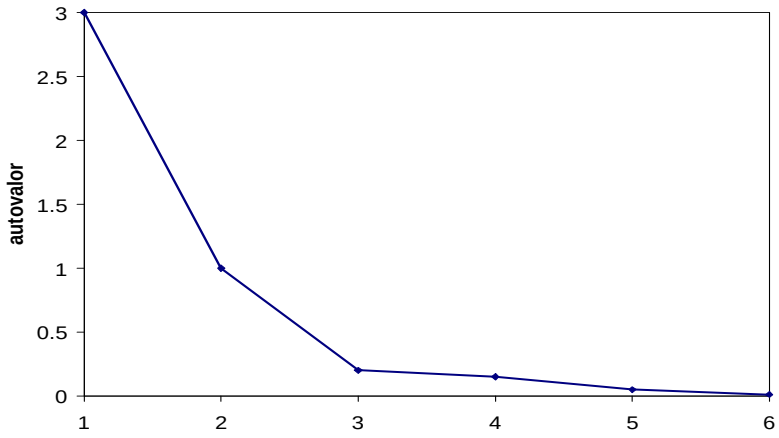


Figura 3.2: Gráfico de sedimentación

**Ejemplo 3.4** En el Ejemplo 3.1, el porcentaje de variabilidad explicado por la primera componente principal $Z_1$ es

...

Según el primer criterio, nos quedaríamos con una componente principal. La media de los valores propios es

...

Por tanto, ambos criterios indican lo mismo. El criterio del test-scree no es necesario cuando solo se dispone de dos componentes principales.

### 3.3.2. COMUNALIDADES

**Comunalidad de una variable original:** Cantidad de su varianza que recogen las componentes principales seleccionadas:

Por el Teorema de Descomposición espectral,

$$S_Y = A\Lambda A' = \lambda_1 \mathbf{a}_1 \mathbf{a}_1' + \cdots + \lambda_p \mathbf{a}_p \mathbf{a}_p'$$

La varianza de cada variable original $Y_i$ se descompone en una suma de aportaciones de cada componente principal $Z_j$,

$$s_{Y_i}^2 = \lambda_1 a_{i1}^2 + \cdots + \lambda_1 a_{ip}^2,\ i = 1, \ldots, p\,.$$

La *comunalidad* de una variable original $Y_i$ es la cantidad de variabilidad que permanece en el sumatorio anterior al seleccionar solo las $p_1 < p$ primeras componentes principales,

$$h_{Y_i}^2 = \lambda_1 a_{i1}^2 + \cdots + \lambda_{p_1} a_{ip_1}^2,\ i = 1, \ldots, p\,.$$

Las comunalidades nos indican lo bien representadas que están las variables originales por el conjunto de componentes seleccionadas.

**Ejemplo 3.5** Si en el Ejemplo 3.1 nos quedamos con una componente principal

$$Z_1 = a_{11} Y_1 + a_{21} Y_2 = ... \qquad ,$$

con varianza $\lambda_1 = ...$ , entonces las comunalidades de la variables originales son

$$h^2_{Y_1} = ... \quad ,$$
$$h^2_{Y_2} = ... \quad .$$

La proporción de la varianza de las variables originales recogida (explicada) por la primera componente principal es

$$\frac{h^2_{Y_1}}{S^2_{Y_1}} = \text{——} = ... \,,$$
$$\frac{h^2_{Y_2}}{S^2_{Y_2}} = \text{——} = ... \,,$$

En SPSS, estas proporciones se denominan *comunalidades reescaladas*. Una variable se considerará bien representada por las componentes principales seleccionadas si su comunalidad reescalada es al menos 0.6.

### 3.3.3. COMPONENTES PRINCIPALES REESCALADAS

La matriz de covarianzas $S_Y$ se puede descomponer en el producto de una matriz por su traspuesta, de la forma

$$S_Y = A\Lambda A' = A\Lambda^{1/2}\Lambda^{1/2}A' = A\Lambda^{1/2}\left(A\Lambda^{1/2}\right)' = A^*\left(A^*\right)' \,,$$

donde $A^* = A\Lambda^{1/2}$. Es decir,

$$A^* = [\mathbf{a}^*_1\ \mathbf{a}^*_2\ \cdots\ \mathbf{a}^*_p], \text{ con } \mathbf{a}^*_j = \lambda_j^{1/2}\mathbf{a}_j, j = 1, \ldots, p\,.$$

Es decir, las columnas $\mathbf{a}^*_j$ de la matriz $A^*$ son los coeficientes de las componentes principales, reescalados de manera que se le da un

peso mayor a las componentes que aportan mayor variabilidad. De esta manera, las *componentes principales reescaladas* son

$$Z_i^* = (\mathbf{a}_i^*)'\mathbf{Y}, \quad i = 1, \ldots, n\,.$$

La matriz $A^*$ se denomina *matriz factorial*, o *matriz de saturaciones*.

**Ejemplo 3.6** Para el Ejemplo 3.1, las componentes principales reescaladas son

$$\begin{aligned} \mathbf{a}_1^* &= \lambda_1^{1/2}\mathbf{a}_1 = ... \quad , \\ \mathbf{a}_2^* &= \lambda_2^{1/2}\mathbf{a}_2 = ... \quad . \end{aligned}$$

La matriz de saturaciones es

$$A^* = [\mathbf{a}_1^*\ \mathbf{a}_2^*] = \begin{pmatrix} \cdots & \\ & \end{pmatrix}\,.$$

### 3.3.4. CORRELACIONES ENTRE VARIABLES Y COMPONENTES PRINCIPALES

**Corolario 3.2** Sea la componente principal $j$-ésima

$$Z_j = a_{1j}Y_1 + \cdots + a_{pj}Y_p\,.$$

Las correlaciones de $Y_1, Y_2, \ldots, Y_p$ con $Z_j$ son

$$r_{Y_1,Z_j} = \frac{\sqrt{\lambda_j}a_{1j}}{s_{Y_1}}\,,\ r_{Y_2,Z_j} = \frac{\sqrt{\lambda_j}a_{2j}}{s_{Y_2}}\,,\ \cdots\,,\ r_{Y_p,Z_j} = \frac{\sqrt{\lambda_j}a_{pj}}{s_{Y_p}}\,.$$

**Ejemplo 3.7** Para el Ejemplo 3.1, las correlaciones entre las variables $Y_1$ e $Y_2$ con las componentes principales

$$\begin{aligned} Z_1 &= ...Y_1 + ...Y_2\,, \\ Z_2 &= ...Y_1 + ...Y_2\,. \end{aligned}$$

son

$r_{Y_1,Z_1} = \text{————————} = \ldots \quad , \quad r_{Y_2,Z_1} = \text{————————} = \ldots \quad .$

$r_{Y_1,Z_2} = \text{————————} = \ldots \quad , \quad r_{Y_2,Z_2} = \text{————————} = \ldots \quad .$

## 3.4. ANÁLISIS DE LA MATRIZ DE CORRELACIONES

Si las variables originales $X_j$, $j = 1, \ldots, p$, tienen rangos muy distintos, o están medidas en unidades no comparables, es necesario tipificarlas (estandarizarlas) de la forma

$$Y_j = \frac{X_j - \bar{x}_j}{\sqrt{s_j^2}}, \quad s_j^2 = \frac{1}{n-1}\sum_{j=1}^{p}(x_{ij} - \bar{x}_j)^2, \ j = 1, \ldots, p\,.$$

Así, la matriz de covarianzas $S_Y$ del vector $(Y_1, \ldots, Y_p)$ coincidirá con la matriz de correlaciones $R_Y$. Entonces se verifica:

- $\text{tr}(S_Y) = \text{tr}(R_Y) = \ldots = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} s_{Z_i}^2$.
- La media de valores propios es $\overline{\lambda} = \sum_{i=1}^{p} \lambda_i / p = \ldots \quad .$
- La correlación entre la variable $Y_i$ y la componente principal $Z_j$ es

$$r_{Y_i,Z_j} = \text{————————} = \ldots \quad .$$

- Las comunalidades son

$$h_{Y_i}^2 = \sum_{j=1}^{p_1} \lambda_j a_{ij}^2 = \ldots \quad .$$

- La matriz factorial es $A^* = [\mathbf{a}_1^* \ \ldots \ \mathbf{a}_p^*]$, donde

$$\mathbf{a}_j^* = \sqrt{\lambda_j}\mathbf{a}_j = (r_{Y_i,Z_j}, \ldots, r_{Y_p,Z_j})', \ j = 1, \ldots, p\,.$$

## 3.5. REPRESENTACIÓN GRÁFICA

Normalmente se realizan dos tipos de gráficos:

- **Gráfico de componentes:** Representa las correlaciones de las variables en el espacio de las componentes principales. De este gráfico se puede obtener una interpretación de cada componente en función de las variables originales.
- **Representación de los individuos en el espacio de las componentes:** Es un gráfico de dispersión en el que se representan las coordenadas de los $n$ individuos en cada par de componentes seleccionadas. Permite caracterizar o describir a los individuos.

  Este gráfico se interpreta con ayuda de las correlaciones de las variables originales con las componentes principales (Gráfico anterior).

## 3.6. RESUMEN

Las componentes principales $Z_1, \ldots, Z_p$ son nuevas variables incorreladas, que se obtienen como combinación lineal de las variables originales $Y_1, \ldots, Y_p$:

$$\begin{aligned} Z_1 &= a_{11}Y_1 + a_{21}Y_2 + \ldots + a_{p1}Y_p\,, \\ Z_2 &= a_{12}Y_1 + a_{22}Y_2 + \ldots + a_{p2}Y_p\,, \\ &\vdots \\ Z_p &= a_{1p}Y_1 + a_{2p}Y_2 + \ldots + a_{pp}Y_p\,. \end{aligned}$$

En principio, hay tantas componentes principales como variables originales. Geométricamente, los ejes principales son los ejes ortogonales que producen la mayor separación de las proyecciones.

¿Cómo se calculan las componentes?, es decir, ¿cómo se calculan las constantes $a_{ij}$?

Se calculan a partir de la matriz de covarianzas de $\mathbf{Y} = (Y_1, \ldots, Y_p)$, llamada $S_Y$. Se calculan sus valores propios y se ordenan de mayor a menor, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Se calculan los vectores propios (normalizados) asociados a los valores propios, $\mathbf{a}_1, \ldots, \mathbf{a}_p$. Pues éstos son los vectores cuyas coordenadas me dan las constantes $a_{ij}$, es decir,

$$\begin{aligned} \mathbf{a}_1 &= (a_{11}, \ldots, a_{p1})', \\ \mathbf{a}_2 &= (a_{12}, \ldots, a_{p2})', \\ &\vdots \\ \mathbf{a}_p &= (a_{1p}, \ldots, a_{pp})'. \end{aligned}$$

¿Con cuántas componentes me quedo?

Hay tres criterios:

- Porcentaje de varianza: sabemos que los valores propios son las varianzas de las componentes principales, y que la suma de varianzas de las componentes es igual a la suma de las varianzas de las variables originales; por tanto,

  $$\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i} \cdot 100$$

  es el porcentaje de varianza que recoge la componente $Z_j$.

Ordenadas de mayor a menor porcentaje de varianza, nos quedamos con el número de componentes que en total recojan al menos un 75 % de la varianza.

- Nos quedamos con las componentes cuyo valor propio supere la media de los valores propios.
- Criterio del test scree o gráfico de sedimentación: Nos fijamos en los "codos" más pronunciados.

Una vez se han seleccionado un número $p_1 < p$ de componentes principales, se observan las comunalidades de las variables originales. Las comunalidad de una variable $Y_i$ es la cantidad de su varianza $S_{Y_i}$ que permanece al seleccionar las $p_1$ componentes principales.

Una variable se considera mínimamente representada por las componentes si éstas recogen al menos un 60 % de su varianza. Si no es así, esta variable no es interpretable en función de las componentes.

¿Qué miden las componentes?

Para interpretar el sentido de las componentes principales, el siguiente paso es observar las correlaciones entre las variables originales y dichas componentes seleccionadas.

Para una componente

$$Z_j = a_{1j}Y_1 + \cdots + a_{pj}Y_p\,,$$

las correlaciones de $Y_1, Y_2, \ldots, Y_p$ con $Z_j$ son

$$r_{Y_1,Z_j} = \frac{\sqrt{\lambda_j} a_{1j}}{s_{Y_1}}\,,\ r_{Y_2,Z_j} = \frac{\sqrt{\lambda_j} a_{2j}}{s_{Y_2}}\,,\ \cdots\,,\ r_{Y_p,Z_j} = \frac{\sqrt{\lambda_j} a_{pj}}{s_{Y_p}}\,.$$

Para una mejor interpretación, interesa que las variables tengan una correlación alta solamente con una de las componentes. Si esto no se verifica, se pueden efectuar rotaciones, e inspeccionar si éstas mejoran los resultados.

Estas correlaciones se pueden representar en gráficos.

Por último, representado a los individuos en el espacio de las componentes principales y teniendo en cuenta las correlaciones anteriores, se pueden describir las características de los individuos y extraer conclusiones prácticas.