Capítulo 7

Componentes Principales

El análisis de componentes principales está relacionado con la explicación de la estructura de covarianzas de una conjunto de p variables a través de pequeñas combinaciones lineales de estas variables. Estas combinaciones lineales representan la selección de un nuevo sistema de coordenadas obtenido al rotar el sistema original. Los nuevos ejes representan las direcciones con máxima variabilidad y proporcionan una forma más simple de describir las estructuras de covarianza.
Con las componentes principales se busca reducir la dimensionalidad del problema inicial e interpretar dichas combinaciones. Aunque p componentes son necesarias para reproducir la variabilidad total del sistema, algunas veces mucha de esta variabilidad podría ser explicada por un número pequeño k de componentes principales. Es decir, existe casi tanta información en las k componentes principales, como la que se obtiene de las p variables originales.
Así, el conjunto de n mediciones en p variables, puede ser reducido a un conjunto de n mediciones y k componentes principales, $k < p$. Este tipo de análisis puede develar comportamientos en los datos que previamente no se sospechaban, y permitir algunas interpretaciones que no eran obvias con los datos originales.

7.1. Componentes principales poblacionales

Algebráicamente, las componentes principales son combinaciones lineales particulares de p variables aleatorias $X_1, \cdots, X_p$. Geométricamente representan un nuevo sistema de coordenadas, el cual es obtenido al rotar el sistema original generado por $X_1, \cdots, X_p$. Los nuevos ejes representan las direcciones con máxima variabilidad y proporcionan una descripción más simple de la estructura de covarianza.
Las componentes principales dependen solamente de la matriz de covarianzas $\mathbf{\Sigma}$ (o de la matriz de correlaciones $\boldsymbol{\rho}$) del vector $\mathbf{X} = (X_1, \cdots, X_p)'$. Su desarrollo no requiere del supuesto de normalidad multivariada. En el caso en que la distribución del vector $\mathbf{X}$ sea normal multivariado, las componentes principales tendrán interpretaciones útiles relacionadas con densidades elipsoidales constantes. Este supuesto facilita la inferencia sobre las componentes principales muestrales.

Sea $\mathbf{X} = (X_1, \cdots, X_p)'$ un vector aleatorio con matriz de covarianzas $\mathbf{\Sigma}$ cuyos valores propios son

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Considere el siguiente conjunto de combinaciones lineales:

$$\begin{aligned} Y_1 &= \mathbf{a}_1' \mathbf{X} = a_{11} X_1 + \cdots + a_{1p} X_p \\ Y_2 &= \mathbf{a}_2' \mathbf{X} = a_{21} X_1 + \cdots + a_{2p} X_p \\ &\vdots = \vdots = \vdots \\ Y_p &= \mathbf{a}_p{}' \mathbf{X} = a_{p1} X_1 + \cdots + a_{pp} X_p \end{aligned}$$

Entonces,

$$\text{Var}\,[Y_i] = \mathbf{a}_i' \, \mathbf{\Sigma} \, \mathbf{a}_i \quad i = 1, \ldots, p$$

y

$$\text{Cov}\,[Y_i \, , \, Y_k] = \mathbf{a}_i' \, \mathbf{\Sigma} \, \mathbf{a}_k \quad i, k = 1, \cdots, p \; ; \; i \neq k \; .$$

Las componentes principales (poblacionales), son aquellas combinaciones lineales $Y_i, Y_2, \cdots, Y_p$ incorrelacionadas, tales que sus varianzas son tan grandes como sea posible. Para efectos prácticos, y procurando eliminar cualquier incremento en la varianza que no sea atribuible a la variabilidad propia de las combinaciones lineales, se enfocará la atención solo a vectores de coeficientes de norma 1.

El proceso para encontrar dichas componentes principales es como sigue:

- La primera componente principal corresponde a la combinación lineal $\mathbf{a}_1' \, \mathbf{X}$ que maximiza $\text{Var}\,[\mathbf{a}_1' \, \mathbf{X}]$, sujeto a que $\mathbf{a}_1' \, \mathbf{a}_1 = 1$.
- La segunda componente principal corresponde a la combinación lineal $\mathbf{a}_2' \, \mathbf{X}$ que maximiza $\text{Var}\,[\mathbf{a}_2' \, \mathbf{X}]$, sujeto a que $\mathbf{a}_2' \, \mathbf{a}_2 = 1$ y $\text{Cov}\,[\mathbf{a}_1' \, \mathbf{X}, \, \mathbf{a}_2' \, \mathbf{X}] = 0$.

$$\vdots$$

- La i-ésima componente principal corresponde a la combinación lineal $\mathbf{a}_i' \, \mathbf{X}$ que maximiza $\text{Var}\,[\mathbf{a}_i' \, \mathbf{X}]$, sujeto a que $\mathbf{a}_i' \, \mathbf{a}_i = 1$ y $\text{Cov}\,[\mathbf{a}_i' \, \mathbf{X}, \, \mathbf{a}_k' \, \mathbf{X}] = 0$, para $k = 1, 2, \cdots, i-1$.

Proposición 7.1.1. *Sea $B_{p\times p}$ una matriz definida positiva con vectores y valores propios $(\mathbf{e}_1, \, \lambda_1), \cdots, \, (\mathbf{e}_p, \, \lambda_p)$, con $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Entonces:*

$$\max_{\mathbf{X} \neq \mathbf{0}} \; \frac{\mathbf{X}' \, B \, \mathbf{X}}{\mathbf{X}' \, \mathbf{X}} \; = \; \lambda_1 \quad ; \quad \textit{el cual se alcanza cuando} \quad \mathbf{X} = \mathbf{e}_1$$

$$\min_{\mathbf{X} \neq \mathbf{0}} \; \frac{\mathbf{X}' \, B \, \mathbf{X}}{\mathbf{X}' \, \mathbf{X}} \; = \; \lambda_p \quad ; \quad \textit{el cual se alcanza cuando} \quad \mathbf{X} = \mathbf{e}_p \; .$$

Más aún,

$$\max_{\mathbf{X} \perp \mathbf{e}_1, \cdots, \mathbf{e}_k} \; \frac{\mathbf{X}' \, B \, \mathbf{X}}{\mathbf{X}' \, \mathbf{X}} \; = \; \lambda_{k+1} \quad ; \quad \textit{el cual se alcanza cuando} \quad \mathbf{X} = \mathbf{e}_{k+1} \; .$$

Teorema 7.1.1. *Sea* $\mathbf{\Sigma}$ *la matriz de covarianzas del vector aleatorio* $\mathbf{X} = (X_1, \cdots, X_p)'$ *y* $(\lambda_1, \mathbf{e}_1), \cdots, (\lambda_p, \mathbf{e}_p)$ *los valores y vectores propios de* $\mathbf{\Sigma}$*, donde* $\lambda_1 \geq \lambda_2, \ldots \geq \lambda_p \geq 0$*. Entonces, la i-ésima componente principal de* $\mathbf{\Sigma}$ *está dada por:*

$$Y_i = \mathbf{e}_i' X = e_{i1}\, X_1 + \cdots + e_{ip}\, X_p \quad , \quad donde \quad i = 1, 2, \cdots, p\,.$$

Además:

$$Var\,[Y_i] = \mathbf{e}_i'\, \mathbf{\Sigma}\, \mathbf{e}_i = \lambda_i \quad , \quad i = 1, \cdots, p$$

y

$$Cov\,[Y_i,\, Y_k] = \mathbf{e}_i'\, \mathbf{\Sigma}\, \mathbf{e}_k = 0 \quad , \quad i, k = 1, \cdots, p\,;\; i \neq k\,.$$

Prueba.
Haciendo $B = \Sigma$ en la proposición 4 se tiene que:

$$\max_{\mathbf{X} \neq \mathbf{0}} \frac{\mathbf{X}'\, \mathbf{\Sigma}\, \mathbf{X}}{\mathbf{X}'\, \mathbf{X}} = \lambda_1 = \mathbf{e}_1'\, \mathbf{\Sigma}\, \mathbf{e}_1 = \mathrm{Var}\,[Y_1]\,.$$

De igual manera se tiene que:

$$\max_{\mathbf{X} \perp \mathbf{e}_1, \cdots, \mathbf{e}_k} \frac{\mathbf{X}'\, \mathbf{\Sigma}\, \mathbf{X}}{\mathbf{X}'\, \mathbf{X}} = \lambda_{k+1} = \mathbf{e}_{k+1}'\, \mathbf{\Sigma}\, \mathbf{e}_{k+1} = \mathrm{Var}\,[Y_{k+1}]\,.$$

Como $\mathrm{Cov}\,[Y_i,\, Y_k] = \mathrm{Cov}\,[\mathbf{e}_i'\, \mathbf{X},\, \mathbf{e}_k'\, \mathbf{X}]$, entonces

$$\mathrm{Cov}\,[Y_i,\, Y_k] = \mathbf{e}_i'\, \mathbf{\Sigma}\, \mathbf{e}_k = \mathbf{e}_i'\, \mathbf{e}_k\, \lambda_k = 0 \quad , \quad \text{para} \quad i \neq k\,.$$

Se sabe que si los valores propios de $\mathbf{\Sigma}$ son todos diferentes, entonces los vectores propios son ortogonales. En caso contrario, se puede usar el proceso de ortogonalización de Grahm-Schimit, para lograr una base de $\mathbb{R}^p$, usando los vectores propios.

Proposición 7.1.2. *Sea* $\mathbf{X} = (X_1, \cdots, X_p)'$ *un vector aleatorio con matriz de covarianzas* $\mathbf{\Sigma}$ *y pares de valores y vectores propios* $(\lambda_1, \mathbf{e}_1), \cdots, (\lambda_p, \mathbf{e}_p)$*, tales que* $\lambda_1 \geq \lambda_2, \ldots \geq \lambda_p \geq 0$*. Sean* $Y_1 = \mathbf{e}_1'\, \mathbf{X}, \cdots, Y_p = \mathbf{e}_p'\, \mathbf{X}$*, las componentes principales asociadas a* $\mathbf{\Sigma}$*. Entonces:*

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} Var\,[X_i] = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} Var\,[Y_i]\,.$$

Prueba. Se sabe que $\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \mathrm{tr}\,(\mathbf{\Sigma})$. Ahora, como $\mathbf{\Sigma} = P\, \Delta\, P'$, donde P es una matriz ortogonal formadas por los vectores propios de $\mathbf{\Sigma}$ y Δ es una matriz diagonal que contiene los valores propios de $\mathbf{\Sigma}$. Asi, se tiene que:

$$\begin{aligned} \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} &= \mathrm{tr}\,(\mathbf{\Sigma}) \\ &= \mathrm{tr}\,(P\, \Delta\, P') \\ &= \mathrm{tr}\,(\Delta\, P\, P') \\ &= \mathrm{tr}\,(\Delta) \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_p\,. \end{aligned}$$

La cantidad $\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp}$ se conoce como *Varianza poblacional total*. La proporción de varianza total explicada por la i-ésima componente principal se calcula como $\dfrac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$, para $i = 1, \cdots, p$.

En relación con la i-ésima componente principal, suponga que $\mathbf{e}_i = (e_{i1}, \cdots, e_{ip})'$. La magnitud de e_{ik} mide la importancia de la k-ésima variable X_k sobre la i-ésima componente principal. Más aún, e_{ik} es proporcional a la correlación entre $\mathbf{X}_k$ y Y_i.

Proposición 7.1.3. *Sea $\mathbf{X} = (X_1, \cdots, X_p)'$ un vector aleatorio con matriz de covarianzas $\mathbf{\Sigma}$ y pares de valores y vectores propios $(\lambda_1, \mathbf{e}_1), \cdots, (\lambda_p, \mathbf{e}_p)$, tales que $\lambda_1 \geq \lambda_2, \cdots \geq \lambda_p \geq 0$. Sean $Y_1 = \mathbf{e}_1' \mathbf{X}, \cdots, Y_p = \mathbf{e}_p' \mathbf{X}$ las componentes principales asociadas a $\mathbf{\Sigma}$. Entonces:*

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \quad ; \quad i, k = 1, 2, \cdots, p \, .$$

Prueba. Sea $\mathbf{a}$ un vector en $\mathbb{R}^p$ con 1 en la k-ésima entrada y cero en las demás, es decir, $\mathbf{a} = (0, \cdots, 0, 1\, 0, \cdots, 0)'$. De esta manera se tiene que $X_k = \mathbf{a}' \mathbf{X}$. Como $Y_i = \mathbf{e}_i' \mathbf{X}$, se tiene que

$$\text{Cov}\,[Y_i, X_k] = \text{Cov}\,[\mathbf{a}' \mathbf{X}, \mathbf{e}_i' \mathbf{X}] = \mathbf{a}' \mathbf{\Sigma}\, \mathbf{e}_i = \mathbf{a}' \lambda_i \mathbf{e}_i = \lambda_i e_{ik} \, .$$

De esta manera se tiene que:

$$\rho_{Y_i, X_k} = \frac{\text{Cov}\,[Y_i, X_k]}{\sqrt{\text{Var}\,[Y_i]}\sqrt{\text{Var}\,[X_k]}} = \frac{\lambda_{iik}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \, .$$

Aunque las correlaciones entre las componentes principales y las variables sean útiles para interpretar las componentes principales, solo miden la contribución individual de las variables aleatorias sobre una componente particular. Esto es, no indican la importancia de una variable X sobre una componente principal Y en presencia de las demás variables $X's$. Por esta razón se recomienda que solo los coeficientes e_{ik}, y no las correlaciones, sean usados para interpretar las componentes principales.

Ejemplo 7.1.1. *Suponga que X_1, X_2 y X_3 son variables aleatorias con matriz de covarianzas $\mathbf{\Sigma}$, dada por: $\mathbf{\Sigma} = \begin{pmatrix} 6 & -2 & 2 \\ -2 & 5 & 1 \\ 2 & 1 & 2 \end{pmatrix}$. Halle las componente principales, la varianza total, la proporción de varianza explicada por cada componente principal y las correlaciones ρ_{Y_i, X_k}. Verifique que las componentes principales son incorrelacionadas.*

Solución. Primero hallemos los vectores y valores propios de $\mathbf{\Sigma}$:

$$\det\,(\mathbf{\Sigma} - \lambda\, I_3) = |\mathbf{\Sigma} - \lambda\, I_3| = \left| \begin{pmatrix} 6-\lambda & -2 & 2 \\ -2 & 5-\lambda & 1 \\ 2 & 1 & 2-\lambda \end{pmatrix} \right|$$
$$= -\lambda^3 + 13\,\lambda^2 - 43\,\lambda + 26 = 0 \, .$$

La solución de este polinomio característico da como resultado tres valores propios:

$$\lambda_1 = 7.754 \quad , \quad \lambda_2 = 4.759 \quad , \quad \lambda_3 = 0.488 \; .$$

Los vectores propios asociados a estos valores propios son:

$$\mathbf{e}_1 = \begin{pmatrix} 0.826 \\ -0.529 \\ 0.195 \end{pmatrix} \quad , \quad \mathbf{e}_2 = \begin{pmatrix} 0.361 \\ 0.762 \\ 0.538 \end{pmatrix} \quad , \quad \mathbf{e}_3 = \begin{pmatrix} -0.433 \\ -0.374 \\ 0.820 \end{pmatrix} \; .$$

Las respectivas componentes principales son:

$$Y_1 = 0.826\, X_1 - 0.529\, X_2 + 0.195\, X_3 \; ,$$

$$Y_2 = 0.361\, X_1 + 0.762\, X_2 + 0.538\, X_3 \; ,$$

$$Y_3 = -0.433\, X_1 - 0.374\, X_2 + 0.820\, X_3 \; .$$

La varianza total está dada por:
$\lambda_1 + \lambda_2 + \lambda_3 = 7.754 + 4.759 + 0.488 = 13 = \mathrm{tr}\,(\mathbf{\Sigma})$.

La proporción de varianza explicada por cada componente es:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{7.754}{13} = 0.596 \; .$$

$$\frac{\lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{4.759}{13} = 0.366 \; .$$

$$\frac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{0.488}{13} = 0.0375 \; .$$

Note que las 2 primeras componentes principales juntas explican 96.2 % de la variación total o sea que las Y_1 y Y_2 pueden reemplazar las 3 variables originales sin perder mucha información (solo se

pierde aproximadamente el 3.8 %). Finalmente,

$$\rho_{Y_1,\,X_1} = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = \frac{0.826\,\sqrt{7.754}}{\sqrt{6}} = 0.939$$
$$\rho_{Y_1,\,X_2} = \frac{e_{12}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}} = \frac{-0.529\,\sqrt{7.754}}{\sqrt{5}} = -0.659$$
$$\rho_{Y_1,\,X_3} = \frac{e_{13}\sqrt{\lambda_1}}{\sqrt{\sigma_{33}}} = \frac{0.195\,\sqrt{7.754}}{\sqrt{2}} = 0.384$$
$$\rho_{Y_2,\,X_1} = \frac{e_{21}\sqrt{\lambda_2}}{\sqrt{\sigma_{11}}} = \frac{0.361\,\sqrt{4.759}}{\sqrt{6}} = 0.322$$
$$\rho_{Y_2,\dot{X}_2} = \frac{e_{22}\sqrt{\lambda_2}}{\sqrt{\sigma_{22}}} = \frac{0.762\,\sqrt{4.759}}{\sqrt{5}} = 0.743$$
$$\rho_{Y_2,\,X_3} = \frac{e_{23}\sqrt{\lambda_2}}{\sqrt{\sigma_{33}}} = \frac{0.538\,\sqrt{4.759}}{\sqrt{2}} = 0.830$$
$$\rho_{Y_3,\,X_1} = \frac{e_{31}\sqrt{\lambda_3}}{\sqrt{\sigma_{11}}} = \frac{-0.433\,\sqrt{0.488}}{\sqrt{6}} = -0.123$$
$$\rho_{Y_3,\dot{X}_2} = \frac{e_{32}\sqrt{\lambda_3}}{\sqrt{\sigma_{22}}} = \frac{-0.374\,\sqrt{0.488}}{\sqrt{5}} = -0.117$$
$$\rho_{Y_3,\,X_3} = \frac{e_{33}\sqrt{\lambda_3}}{\sqrt{\sigma_{33}}} = \frac{0.820\,\sqrt{0.488}}{\sqrt{2}} = 0.405$$

7.2. Componentes principales obtenidas de variables estandarizadas

Es posible también obtener componentes principales a partir de variables estandarizadas.

Sea $\mathbf{X} = (X_1, \cdots, X_p)'$ un vector aleatorio con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$. Sea $\mathbf{Z} = (Z_1, \cdots, Z_p)'$ otro vector aleatorio tal que $Z_i = \dfrac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}$, para $i = 1, \cdots, p$. Entonces:

$$\mathbf{Z} = (V^{\frac{1}{2}})^{-1}\,(\mathbf{X} - \boldsymbol{\mu}) \quad , \quad \text{donde} \quad V^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\sigma_{11}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\sigma_{pp}} \end{pmatrix} .$$

Claramente $E\,[\mathbf{Z}] = \mathbf{0}$ y $\mathrm{Cov}\,[\mathbf{Z}] = (V^{\frac{1}{2}})^{-1}\,\boldsymbol{\Sigma}\,(V^{\frac{1}{2}})^{-1} = \boldsymbol{\rho}$. Las componentes principales de $\mathbf{Z}$ pueden ser obtenidas de los vectores propios de $\boldsymbol{\rho}$, la matriz de correlaciones de $\mathbf{X}$. En general las componentes principales obtenidas con $\boldsymbol{\Sigma}$ son diferentes de las obtenidas con $\boldsymbol{\rho}$. Cuando las escalas de medición de las variables son muy diferentes, se deben estandarizar.

Proposición 7.2.1. *Sea* $\mathbf{X} = (X_1, \cdots, X_p)'$ *un vector aleatorio con vector de medias* $\boldsymbol{\mu}$ *y matriz de covarianzas* $\boldsymbol{\Sigma}$ *y sea* $\mathbf{Z} = (V^{\frac{1}{2}})^{-1}(\mathbf{X} - \boldsymbol{\mu})$, *donde* $V^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\sigma_{11}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\sigma_{pp}} \end{pmatrix}$. *La i-ésima componente principal de las variables estandarizadas está dada por:*

$$Y_i = \mathbf{e}_i' \mathbf{Z} = \mathbf{e}_i' (V^{\frac{1}{2}})^{-1} (\mathbf{X} - \boldsymbol{\mu}) \quad ; \quad i = 1, \cdots, p \, .$$

Más aún,

$$\sum_{i=1}^{p} Var[Y_i] = \sum_{i=1}^{p} Var[Z_i] = p \quad y \quad \rho_{Y_i, Z_k} = e_{ik} \sqrt{\lambda_i} \; ; \; i, k = 1, \cdots, p \, ,$$

donde $(\lambda_1, \mathbf{e}_1), \cdots, (\lambda_p, \mathbf{e}_p)$ *son los pares de valores y vectores propios de* $\boldsymbol{\rho}$, *con* $\lambda_1 \geq \lambda_2, \cdots \geq \lambda_p \geq 0$.

Ejemplo 7.2.1. *Determine las componentes principales poblacionales,* Y_1 *y* Y_2 *para la matriz* $\boldsymbol{\Sigma} = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$. *Con dichas componentes calcule la proporción de varianza total explicada por la primera componente principal. Repita el proceso usando la matriz de correlaciones. Comente. Calcule las correlaciones entre las componentes principales y las variables estandarizadas.*

Solución. Se hallan los valores propios de $\boldsymbol{\Sigma}$:

$$\det(\boldsymbol{\Sigma} - \lambda I) = \left| \begin{pmatrix} 5 - \lambda & 2 \\ 2 & 2 - \lambda \end{pmatrix} \right| = \lambda^2 - 7\lambda + 6 \, .$$

Igualando a cero la ecuación anterior se tiene que: $\lambda_1 = 6$ y $\lambda_2 = 1$. Los respectivos vectores propios asociados son: $\mathbf{e}_1 = \left(\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}} \right)'$ y
$\mathbf{e}_2 = \left(\frac{-1}{\sqrt{5}}, \frac{2}{\sqrt{5}} \right)'$. Así, las componentes principales son:

$$Y_1 = \frac{2}{\sqrt{5}} X_1 \; + \; \frac{1}{\sqrt{5}} X_2 = 0.894 \, X_1 + 0.447 \, X_2$$

$$Y_2 = \frac{2}{\sqrt{5}} X_1 \; + \; \frac{1}{\sqrt{5}} X_2 = -0.447 \, X_1 + 0.894 \, X_2 \, .$$

La proporción de varianza total explicada por Y_1 es $\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{6}{7} = 0.857$. Además

$$\rho_{Y_1, X_1} = \frac{e_{11} \sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = \frac{\frac{2}{\sqrt{5}} \sqrt{6}}{\sqrt{5}} = 0.9798$$

$$\rho_{Y_1, X_2} = \frac{e_{12} \sqrt{\lambda_2}}{\sqrt{\sigma_{22}}} = \frac{\frac{2}{\sqrt{5}} \sqrt{6}}{\sqrt{5}} = 0.7746 \, .$$

La matriz de correlación es $\boldsymbol{\rho} = \begin{pmatrix} 1 & \frac{2}{\sqrt{10}} \\ \frac{2}{\sqrt{10}} & 1 \end{pmatrix}$. Los valores propios de $\boldsymbol{\rho}$ se obtienen como:

$$\det(\boldsymbol{\rho} - \lambda I) = \left| \begin{pmatrix} 1 - \lambda & \frac{2}{\sqrt{10}} \\ \frac{2}{\sqrt{10}} & 1 - \lambda \end{pmatrix} \right| = (1 - \lambda)^2 - \frac{4}{10} = 0 \, .$$

Al resolver esta ecuación se obtiene $\lambda_1 = 1 + \frac{2}{\sqrt{10}}$ y $\lambda_2 = 1 - \frac{2}{\sqrt{10}}$. Los respectivos vectores propios son: $\mathbf{e}_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)'$ y $\mathbf{e}_1 = \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)'$.

Las componentes principales para las variables estandarizadas son:

$$Y_1 = \frac{1}{\sqrt{2}} Z_1 + \frac{1}{\sqrt{2}} Z_2 = \frac{1}{\sqrt{2}} \left(\frac{X_1 - \mu_1}{5}\right) + \frac{1}{\sqrt{2}} \left(\frac{X_2 - \mu_2}{2}\right)$$

$$Y_2 = -\frac{1}{\sqrt{2}} Z_1 + \frac{1}{\sqrt{2}} Z_2 = -\frac{1}{\sqrt{2}} \left(\frac{X_1 - \mu_1}{5}\right) + \frac{1}{\sqrt{2}} \left(\frac{X_2 - \mu_2}{2}\right)$$

Lo cual es equivalente a escribir:

$$Y_1 = 0.1414\,(X_1 - \mu_1) + 0.3536\,(X_2 - \mu_2)$$

$$Y_2 = -0.1414\,(X_1 - \mu_1) + 0.3536\,(X_2 - \mu_2)\ .$$

Para $\boldsymbol{\rho}$, la proporción de varianza total explicada por la primera componente es $\dfrac{\lambda_1}{2} = \dfrac{1 + \frac{2}{\sqrt{10}}}{2} = 0.8162$. Finalmente:

$$\rho_{Y_1, Z_1} = e_{11}\sqrt{\lambda_1} = \frac{1}{\sqrt{2}}\sqrt{1 + \frac{2}{\sqrt{10}}} = 0.9035$$

$$\rho_{Y_1, Z_2} = e_{12}\sqrt{\lambda_1} = -\frac{1}{\sqrt{2}}\sqrt{1 + \frac{2}{\sqrt{10}}} = -0.9035\ .$$

La primera componente principal es sensiblemente afectada por la estandarización. Observe que cuando las componentes principales de $\boldsymbol{\rho}$ se expresan en términos de las variables originales, las magnitudes relativas de los pesos están en concordancia.

7.3. Componentes principales muestrales

Sea $\mathbf{X}_1, \cdots, \mathbf{X}_n$ una muestra aleatoria de una distribución p-dimensional, con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$ desconocida. De esta muestra se obtiene el vector de medias muetrales $\bar{\mathbf{X}}$, la matriz de covarianzas muestral S y la matriz de correlaciones muestrales R. El objetivo aqui es poder encontrara combinaciones lineales de las características medidas que contengan la mayor cantidad de variabilidad muestral posible. Tales combinaciones lineales serán llamadas *Componentes principales muestrales*. El siguiente resultado indica como hallar las componentes principales muestrales.

Proposición 7.3.1. *Sea $\mathbf{X}_1, \cdots, \mathbf{X}_n$ una muestra aleatoria de una distribución p-dimensional, con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$ desconocida. Sea S la matriz de covarianzas muestral y*
$(\widehat{\lambda}_1, \widehat{\mathbf{e}}_1), \cdots, (\widehat{\lambda}_p, \widehat{\mathbf{e}}_p)$ los pares de valores y vectores propios de S, donde $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \cdots \geq \widehat{\lambda}_p \geq 0$. Entonces, la i-ésima componente principal asociada a S está dada por:

$$\widehat{Y}_i = \widehat{\mathbf{e}}_i' \mathbf{X} = \widehat{e}_{i1} X_1 + \cdots + \widehat{e}_{ip} X_p \quad ; \quad i = 1, \cdots, p\ ,$$

donde $\mathbf{X}$ *es cualquier observación p-varida de las variables* $(X_1, \cdots, X_p)$. *Además*

$$\textit{Varianza muestral de } \widehat{Y}_k = \widehat{\lambda}_k \ , \quad k = 1, \cdots, p$$

$$\textit{Covarianza muestral de } \widehat{Y}_i \textit{ y } \widehat{Y}_k = 0 \ , \quad i \neq k$$

$$s_{11} + \cdots + s_{pp} = \widehat{\lambda}_1 + \cdots + \widehat{\lambda}_p$$

$$r_{\widehat{Y}_i, X_k} = \frac{\widehat{e}_{ik}\sqrt{\widehat{\lambda}_i}}{\sqrt{s_{kk}}} \ , \quad i, k = 1, \cdots, p \ .$$

Si las observaciones $\mathbf{X}_i$ son centradas alrededor de $\bar{\mathbf{X}}$, esta operación no tiene efecto sobre la matriz de covarianzas muestral. Es ese caso la i-ésima componente principal está dada por: $\widehat{Y}_i = \widehat{\mathbf{e}}_i'(\mathbf{X} - \bar{\mathbf{X}})$, para $i = 1, \cdots, p$ y para cualquier observación p-variada $\mathbf{X}$.

Haciendo $\widehat{Y}_{ij} = \widehat{\mathbf{e}}_i'(\mathbf{X}_j - \bar{\mathbf{X}})$, para $i = j, \cdots, p$, se tiene que:

$$\bar{\widehat{Y}}_i = \frac{1}{n}\sum_{j=1}^{n} \widehat{\mathbf{e}}_i'(\mathbf{X}_j - \bar{\mathbf{X}}) = \frac{1}{n}\widehat{\mathbf{e}}_i \left(\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})\right) = \frac{1}{n}\widehat{\mathbf{e}}_i'\,\mathbf{0} = 0 \ .$$

Ejemplo 7.3.1. *Un censo proporciona información por región de cinco variables socioeconómicas para las áres de Madison y Wisconsin. Las variables de interés fueron* X_1 : *Población total (en miles),* X_2 : *Promedio de años escolares,* X_3 : *Total de empleados (en miles),* X_4 : *Empleados del servicio de salud (en cientos) y* X_5 : *Ingresos familiares (en miles). Los datos de* 14 *regiones se muestran en la tabla* 7.

Tabla 7: Datos de Censo

X_1	X_2	X_3	X_4	X_5
5.935	14.2	2.265	2.27	2.91
1.523	13.1	0.597	0.75	2.62
2.599	12.7	1.237	1.11	1.72
4.009	15.2	1.649	0.81	3.02
4.687	14.7	2.312	2.50	2.22
8.044	15.6	3.641	4.51	2.36
2.766	13.3	1.244	1.03	1.97
6.538	17.0	2.618	2.39	1.85
6.451	12.9	3.147	5.52	2.01
3.314	12.2	1.606	2.18	1.82
3.777	13.0	2.119	2.83	1.80
1.530	13.8	0.798	0.84	4.25
2.768	13.6	1.336	1.75	2.64
6.585	14.9	2.763	1.91	3.17

Obtenga las componentes principales usando la matriz de covarianzas. Modifique la escala de la variable Ingreso (X_5) y usando la matriz de covarianzas halle nuevamente las componentes principales. Comente.

Solución. Usando la matriz de covarianzas para las variables X_1 a X_5, en R.

```
censo <- read.table(file.choose(), header=T)
round(cov(censo),2)
      X1   X2    X3    X4    X5
X1  4.31 1.68  1.80  2.16 -0.25
X2  1.68 1.77  0.59  0.18  0.18
X3  1.80 0.59  0.80  1.06 -0.16
X4  2.16 0.18  1.06  1.97 -0.36
X5 -0.25 0.18 -0.16 -0.36  0.50
au <- prcomp(censo)
au
Standard deviations (1, .., p=5):
[1] 2.6326932 1.3360929 0.6242194 0.4790918 0.1189747

Los valores propios se obtienen al elevar estos valores al cuadrado

round(au\$sdev^2,2)
[1] 6.93 1.79 0.39 0.23 0.01

Rotation (n x k) = (5 x 5):
           PC1         PC2          PC3         PC4          PC5
X1 -0.78120807  0.07087183 -0.003656607  0.54171007  0.302039670
X2 -0.30564856  0.76387277  0.161817438 -0.54479937  0.009279632
X3 -0.33444840 -0.08290788 -0.014841008  0.05101636 -0.937255367
X4 -0.42600795 -0.57945799 -0.220453468 -0.63601254  0.172145212
X5  0.05435431  0.26235528 -0.961759720  0.05127599 -0.024583093

summary(au)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     2.6327 1.3361 0.62422 0.47909 0.11897
Proportion of Variance 0.7413 0.1909 0.04168 0.02455 0.00151
Cumulative Proportion  0.7413 0.9323 0.97394 0.99849 1.00000

# Paquete factoextra
install.packages("factoextra")
library(factoextra)
library(ggplot2)

fviz_eig(au)
fviz_pca_ind(au, col.ind=1:14, gradient.cols=c("blue", "red","black","blue", "orange"))

aux <- svd(cov(censo))
plot(aux$u[,1],aux$u[,2], type="n", xlab="Comp1", ylab="Comp2")
text(aux$u[,1],aux$u[,2], labels=colnames(censo))
abline(h=0)
abline(v=0)
arrows(0,0,aux$u[,1]+0.03,aux$u[,2]-0.03,col="red")
```

La primera componente principal explica el 74.13 % de la variabilidad total y las primeras dos componentes principales explican el 93.23 % de dicha variabilidad.

Un gráfico muy útil que muestra la variabilidad asociada a cada componente es el *Scree Plot*. Adicionalmente, para identificar que las componentes principales muestrales si sean incorrelacionadas, se grafican ambas y se observan posibles anomalías. El Scree plot y el grafico de dispersión para las dos primeras componentes principales muestrales se muestran en las figuras 7.1 y 7.2.
Usando las variables, se obtiene el gráfico de la figura 7.3.

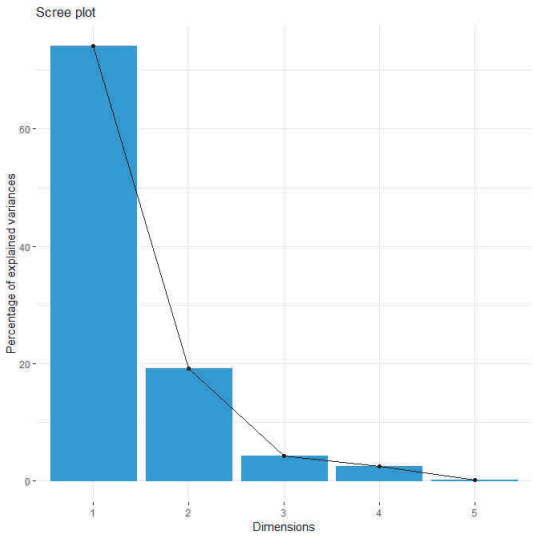


Figura 7.1: Scree Plot para datos del Censo

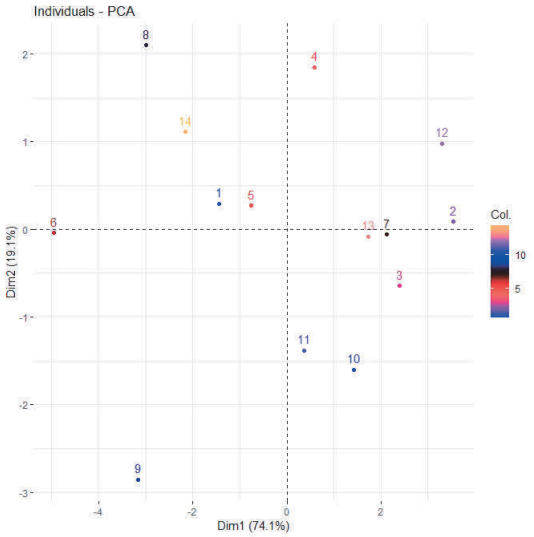


Figura 7.2: Gráficos de las dos primeras componentes principales datos del Censo

En el gráfico 7.2 se pueden evidenciar las observaciones que tienen mayor peso en las componentes principales y en el gráfico 7.3 las variables con mayor peso en las dos componentes principales.

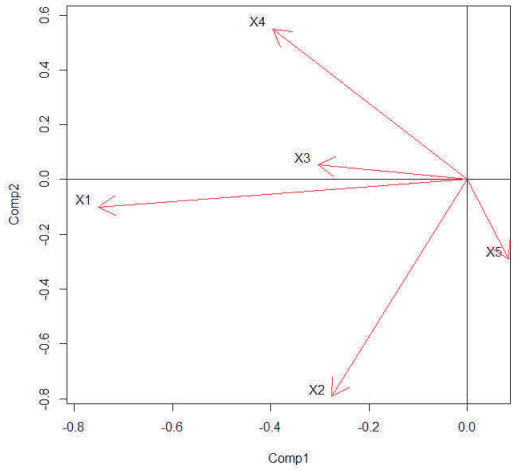


Figura 7.3: Gráficos de las dos primeras componentes principales y Variables del Censo

Estandarizando la variable X_5 y se denotandola Ing, se obtienen los siguientes resultados:

```
censo1 <- censo
censo1[,5] <- (censo[,5]-mean(censo[,5]))/sd(censo[,5])
colnames(censo1)=c("X1", "X2", "X3", "X4", "Ing")
au1 <- prcomp(censo1)
round(au1\$sdev^2,2)
[1] 2.64 1.39 0.84 0.48 0.12
> summary(au1)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     2.6368 1.3931 0.84117 0.47928 0.11899
Proportion of Variance 0.7062 0.1971 0.07187 0.02333 0.00144
Cumulative Proportion  0.7062 0.9034 0.97523 0.99856 1.00000

fviz_eig(au1)
fviz_pca_ind(au1, col.ind=1:14, gradient.cols=c("blue", "red","black","blue", "orange"))

aux <- svd(cov(censo1))
plot(aux$u[,1],aux$u[,2], type="n", xlab="Comp1", ylab="Comp2")
text(aux$u[,1],aux$u[,2], labels=colnames(censo1))
abline(h=0)
abline(v=0)
arrows(0,0,aux$u[,1]+0.03,aux$u[,2]-0.03,col="red")
```

El Scree plot y el grafico de dispersión para las dos primeras componentes principales muestrales con observaciones y variables se muestran en las figuras 7.4, 7.5 Y 7.6.

La primera componente principal explica el 70.62 % de la variabilidad total y las primeras dos componentes principales explican el 90.34 % de dicha variabilidad.

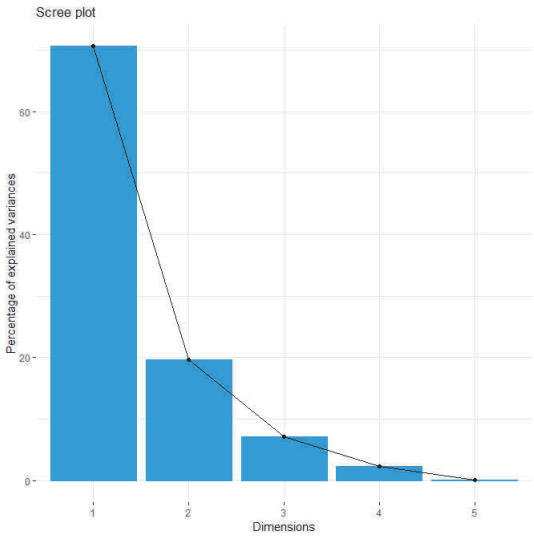


Figura 7.4: Scree Plot para datos del Censo, Ingresos estandarizados

Ejemplo 7.3.2. *Se tienen los resultados de los juegos olímpicos de los Ángeles* 1984. *Se muestran los registros de competencia en atletismo para cada país participante en las pruebas de* 100, 200 *y* 400 *metros en segundos y* 800, 1500, 5000, 10000 *y maratón, en minutos, para los hombres. Similarmente las pruebas de* 100, 200 *y* 400 *metros en segundos y* 800, 1500, 3000 *y maratón para mujeres medidos en minutos. Los datos aparecen en los archivos olihom y olimuje. Realice un análisis de componentes principales para las variables discriminando por género.*

Solución. Debido a que las variables medidas están en direrentes escalas, para todos los efectos se usará la matriz de correlaciones muestrales. Los comandos en R para los resultados de los Hombres son:

```
oli_hom <- read.table(file.choose(), header=T)

au1 <- prcomp(oli_hom[,1:8], scale=T)
summary(au1)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5    PC6    PC7     PC8
Standard deviation     2.5734 0.9368 0.39915 0.35221 0.28263 0.2607 0.2155 0.15033
Proportion of Variance 0.8278 0.1097 0.01992 0.01551 0.00999 0.0085 0.0058 0.00283
Cumulative Proportion  0.8278 0.9375 0.95739 0.97289 0.98288 0.9914 0.9972 1.00000

round(au1$rotation[,1],3)
   P100    P200    P400    P800   P1500   P5000  P10000 MARATON
  0.318   0.337   0.356   0.369   0.373   0.364   0.367   0.342

round(au1$rotation[,2],3)
  P100    P200    P400    P800   P1500   P5000  P10000  MARATON
  0.567   0.462   0.248   0.012  -0.140  -0.312  -0.307  -0.439

plot(au1$rotation[,1],au1$rotation[,2], type="n", xlab="Comp1", ylab="Comp2", xlim=c(-0.2,0.4))
text(au1$rotation[,1],au1$rotation[,2], labels=colnames(oli_hom))
abline(h=0)
abline(v=0)
arrows(0,0,au1$rotation[,1]+0.03,au1$rotation[,2]-0.03,col="blue")
```

Observe que la primera componente responde por el 82.8 % de la variabilidad en los datos. Esta primera componente puede verse como un indicador de los países con mejores tiempos, esto debido a

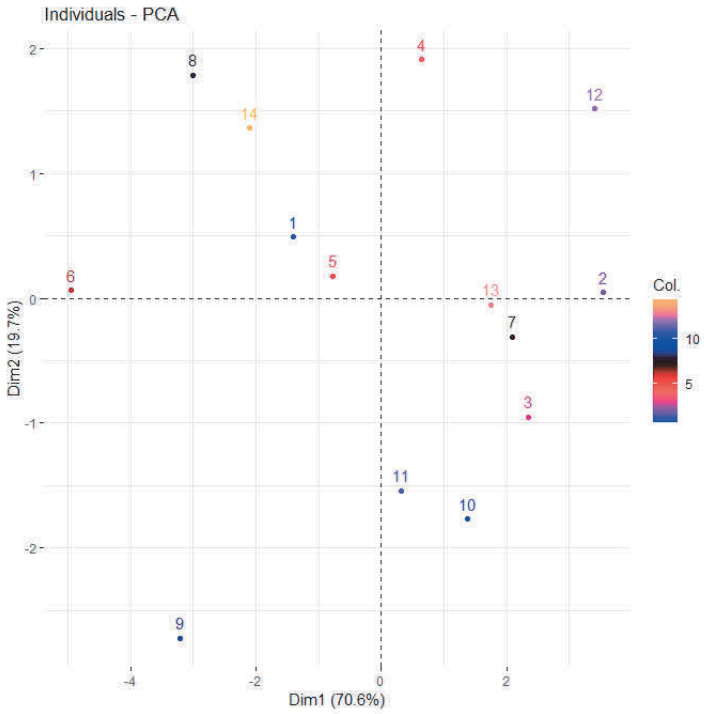


Figura 7.5: Gráficos de las dos primeras componentes principales, Ingreso estandarizado

que todas las constantes son positivas. La siguiente tabla muestra el ranking generado por la primera componente principal.

En la tabla anterior se observa que el país mejor rankeado es USA. La segunda componente principal parece un contraste entre los países más fuertes en velocidad y los más fuertes en pruebas de fondo. Así, si el valor de la segunda componente es negativo, es indicador de un país fuerte en pruebas de fondo. Si es positivo, es un país fuerte en pruebas de velocidad.

Se puede repetir el mismo análisis para los records de las Mujeres. De estos se obtienen los siguientes resultados:

```
oli_muj <- read.table(file.choose(), header=T)
au2 <- prcomp(oli_muj[,1:7], scale=T)
summary(au2)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     2.4095 0.80848 0.54762 0.35423 0.23198 0.19761 0.14981
Proportion of Variance 0.8294 0.09338 0.04284 0.01793 0.00769 0.00558 0.00321
Cumulative Proportion  0.8294 0.92276 0.96560 0.98353 0.99122 0.99679 1.00000

au2$rotation
round(au2$rotation[,1],3)
   P100    P200    P400    P800   P1500   P3000 MARATON
  0.368   0.365   0.382   0.385   0.389   0.389   0.367

round(au2$rotation[,2],3)
   P100    P200    P400    P800   P1500   P3000 MARATON
  0.490   0.537   0.247  -0.155  -0.360  -0.348  -0.369
$
```

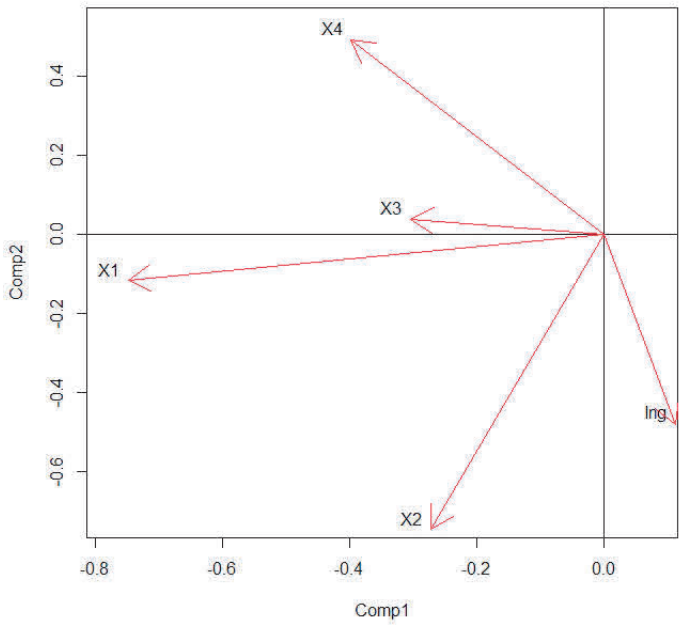


Figura 7.6: Gráficos de las dos primeras componentes principales y Variables del Censo

7.4. Inferencia para muestras grandes

Se ha mostrado que los vectores y valores propios de la matriz de covarianzas (o de correlaciones), son la esencia de las componentes principales; los primeros indican las direcciones de máxima variabilidad y los segundos la magnitud de dicha variabilidad. Cuando los primeros (pocos) valores propios son muy grandes en comparación con los demás, mucha de la variabilidad total de las variables originales, puede ser explicada por pocas componentes principales (menos de p). En la práctica las decisiones sobre la calidad de las componentes principales aproximadas, debe hacerse sobre la base de los pares de valores y vectores propios $(\widehat{\lambda}_i, \widehat{\mathbf{e}}_i)$ obtenidos de la matriz S o R.

Dado que estos pares dependen de una muestra aleatoria, es natural pensar en cuales serían sus distribuciones muestrales, y por esa misma razón, pueden difererir de sus contrapartes poblacionales $(\lambda_i, \mathbf{e}_i)$. Las distribuciones muestrales de $\widehat{\lambda}_i$ y $\widehat{\mathbf{e}}_i$ son difíciles de derivar. Si el tamaño de la muestra es grande, es posible encontrar las distribuciones asintóticas de $\widehat{\lambda}_i$ y $\widehat{\mathbf{e}}_i$.

Sea $\mathbf{X}_1, \cdots, \mathbf{X}_n$ una muestra aleatoria de una distribución $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Asuma que los valores propios de $\boldsymbol{\Sigma}$, $\lambda_1, \cdots, \lambda_p$, son distintos y positivos, esto es, $\lambda_1 > \lambda_2, > \cdots > \lambda_p > 0$, los cuales son desconocidos. Se supone que el número de valores propios es conocido. Cuando el supuesto de normalidad es violado, *Anderson* y *Girshick*, establecen las siguientes distribuciones muestrales para $\widehat{\boldsymbol{\lambda}}' = (\widehat{\lambda}_1, \cdots, \widehat{\lambda}_p)$ y para los vectores propios estimados $\widehat{\mathbf{e}}_1, \cdots, \widehat{\mathbf{e}}_p$ de la matriz S.

1. Sea Δ una matriz diagonal que contiene los valores propios $\lambda_1, \cdots, \lambda_p$ de la matriz $\boldsymbol{\Sigma}$. Entonces $\sqrt{n}\left(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}\right)$ tiene una distribución aproximadamente $N_p\left(\mathbf{0}, 2\,\Delta^2\right)$.

ind_vel	PAIS
7905,2	USA
7912,9	AUSTRALI
7929,7	PORTUGAL
7933,6	JAPAN
7952,9	NZ
7955,1	NETHERLA
7958,7	GBNI
7969,9	MEXICO
7994,2	KENYA
8005,0	GDR
8007,3	BELGIUM
8027,7	CANADA
8043,4	USSR
8053,0	SWEDEN

ind_vel	PAIS
8062,4	FINLAND
8064,7	DENMARK
8074,6	ITALY
8085,8	SWITZERL
8090,2	DPRKOREA
8095,8	COLUMBIA
8100,5	NORWAY
8106,0	SPAIN
8108,3	POLAND
8112,5	TURKEY
8143,0	FRG
8144,3	INDIA
8151,7	FRANCE
8152,8	IRELAND

ind_vel	PAIS
8159,9	RUMANIA
8174,1	HUNGARY
8209,0	BRAZIL
8239,6	CHINA
8269,4	CHILE
8273,3	CZECH
8298,3	GREECE
8360,3	AUSTRIA
8400,9	KOREA
8416,1	COSTA
8472,6	ISRAEL
8489,1	ARGENTIN
8588,2	TAIPEI
8592,6	GUATEMAL

ind_vel	PAIS
8632,3	BURMA
8692,8	LUXEMBOU
8950,4	PHILIPPI
9028,5	BERMUDA
9131,8	PNG
9161,4	INDONESI
9246,0	THAILAND
9373,5	MAURITIU
9473,1	MALAYSIA
9480,4	DOMREP
9694,7	SINGAPOR
9976,3	WSAMOA
10154,4	COOKIS

Figura 7.7: Ranking de los paises según mejores tiempos

2. Sea
$$\mathbf{E}_i = \lambda_i \sum_{k=1,\, k\neq i}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2}\, \mathbf{e}_k\, \mathbf{e}_k' \,,$$
entonces $\sqrt{n}\,(\widehat{\mathbf{e}}_i - \mathbf{e}_i)$ es aproximadamente $N_p\,(\mathbf{0},\, \mathbf{E}_i)$, con $i = 1, \cdots, p$.

3. Cada $\widehat{\lambda}_i$ tiene una distribución que es independiente de los elementos asociados a $\widehat{\mathbf{e}}_i$.

El numeral 1 implica que, para n grande, $\widehat{\lambda}_1, \cdots, \widehat{\lambda}_p$ están distribuídos de manera independiente. Más aún, $\widehat{\lambda}_i$ tiene una distribución aproximadamente $N\left(\lambda_i,\, \frac{2\lambda_i^2}{n}\right)$, para $i = 1, \cdots, p$.
Así, para α dado, un intervalo de confianza aproximado al $100\,(1-\alpha)\,\%$ para λ_i está dado por:
$$\left(\frac{\widehat{\lambda}_i}{\left(1 + z_{\frac{\alpha}{2}}\sqrt{\frac{2}{n}}\right)} \;\; \frac{\widehat{\lambda}_i}{\left(1 - z_{\frac{\alpha}{2}}\sqrt{\frac{2}{n}}\right)} \right) ,$$
donde $z_{\frac{\alpha}{2}}$ corresponde al percentil $100\,\frac{\alpha}{2}\,\%$ superior de una normal estándar.

El numeral 2 implica que los vectores aleatorios $\widehat{\mathbf{e}}_i$ tiene una distribución aproximadamente normal, para $i = 1, \cdots, p$, y sus componentes están correlacionadas.

Ejemplo 7.4.1. *Considere nuevamente los datos de los registros obtenidos en las pruebas de atletismo para hombre y mujeres en las Olimpiadas de los Ángeles,* 1984*).*

De los cálculos allí derivados usando la matriz de correlaciones muestrales, se encuentra que para los hombres el primer valor propio era $\widehat{\lambda}_1 = 6.622$ y la primera componente principal explica el $82.78\,\%$ de la variabilidad total. Asumiendo qure $n = 55$ es grande, un intervalo de confianza aproximado al $95\,\%$ para λ_1, está dado por:
$$\left(\frac{\widehat{\lambda}_1}{\left(1 + z_{0.025}\sqrt{\frac{2}{55}}\right)} \;\; \frac{\widehat{\lambda}_1}{\left(1 - z_{0.025}\sqrt{\frac{2}{55}}\right)} \right) = \left(\frac{6.622}{\left(1 + 1.96\sqrt{\frac{2}{55}}\right)} \;\; \frac{6.622}{\left(1 - 1.96\sqrt{\frac{2}{55}}\right)} \right)$$
$$= (4.8204\,,\; 10.574)\,.$$

Analogamente, para las mujeres, el primer valor propio era $\widehat{\lambda}_1 = 5.806$ y la primera componente principal explica el $82.94\,\%$ de la variabilidad total. Asumiendo qure $n = 55$ es grande, un intervalo de confianza al $95\,\%$ para λ_1, está dado por:

$$\left(\frac{\widehat{\lambda}_1}{\left(1 + z_{0.025}\sqrt{\frac{2}{55}}\right)} \;\; \frac{\widehat{\lambda}_1}{\left(1 - z_{0.025}\sqrt{\frac{2}{55}}\right)} \right) = \left(\frac{5.806}{\left(1 + 1.96\sqrt{\frac{2}{55}}\right)} \;\; \frac{5.806}{\left(1 - 1.96\sqrt{\frac{2}{55}}\right)} \right)$$
$$= (4.226\,,\; 9.271)\,.$$

Ambos intervalos de confianza tiene límites muy similares, para hombres y mujeres.

Capítulo 8

Análisis de Cluster o Agrupamiento

Algunos veces los procedimientos exploratorios pueden ser poco útiles a la hora de permitir entender la naturaleza compleja de las interrelaciones multivaridas. Una técnica exploratorias importante se basa en la búsqueda de una estrutucra natural de agrupamiento en los datos multivariados. Estos agrupamientos pueden dar mejor información acerca de posibles outliers, sugerir hipótesis interesantas o permitir una mejor estimación de los parámetros poblacionales asociados con la distribución multivariada de los datos. El agrupamiento o cluster, puede ser considerada como una forma diferente de clasificar observaciones multivariadas. Estos procedimientos se conocen como análisis de Cluster. El objetivo básico en este análisis es descubrir agrupamientos naturales en los items (o variables). Primero se propone o elabora una escala sobre la cual se mide la asociación (o similaridad) entre objetos o variables y luego a través de dicha escala, se construyen los posibles agrupamientos.

8.1. Medidas de similaridad

Recuerde que:

- Si $\mathbf{x} = (x_1, \cdots, x_p)'$ y $\mathbf{y} = (y_1, \cdots, y_p)'$, la distancia euclidiana entre $\mathbf{x}$ e $\mathbf{y}$ se define como:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T (\mathbf{x}-\mathbf{y})} .$$

- Si $\mathbf{X}$ e $\mathbf{Y}$ son vectores aleatorios p-variados, la distancia estadística se obtiene como:

$$d(\mathbf{X}, \mathbf{Y}) = \sqrt{(\mathbf{X}-\mathbf{Y})' A (\mathbf{X}-\mathbf{Y})} .$$

  Usualmente $A = S^{-1}$. Como en la práctica no se tiene conocimiento de los diferentes grupos, S^{-1} puede no ser calculable.

Por lo anterior, la distancia euclidiana es preferida para realizar el análisis de cluster. Otras distancias usualmente utilizadas son:

- Distancia de Minkowski:

$$d(\mathbf{X}, \mathbf{Y}) = \left[\sum_{i=1}^{p} |X_i - Y_i|^m \right]^{\frac{1}{m}} .$$

- Distancia o métrica de Canberra:

$$d(\mathbf{X},\, \mathbf{Y}) = \sum_{i=1}^{p} \frac{|X_i - Y_i|}{(X_i + Y_i)} \,.$$

- Distancia o métrica de Czekanowski:

$$d(\mathbf{X},\, \mathbf{Y}) = 1 - \frac{2 \sum\limits_{i=1}^{p} \min(X_i,\, Y_i)}{\sum\limits_{i=1}^{p} X_i + Y_i} \,.$$

Cuando los items no pueden ser representados por medidas p dimensionales, pares de ítems son a veces comparados sobre la presencia o ausencia de una característica de interés. Los ítems similares tendrían más características comunes que los que no lo son. La presencia o ausencia de una característica puede ser medida a través de variables "Dummy"(binarias o bernoullis): 1: si la característica está presente y 0: si no lo está.

Para ilustrar esta situación, suponga que se tienen $p = 5$ variables y que para cada variable se define una variable binaria. Suponga que para dos observaciones o ítems i y k, se evalúan dichas variables binarias:

	1	2	3	4	5
ítem i	1	0	0	1	1
ítem k	1	1	0	1	0

Sea X_{ij} el puntaje (1 o 0) para la j-ésima variable binaria sobre el i-ésimo ítem, y sea X_{kj} el puntaje para la j-ésima variable sobre el k-ésimo ítem, $j = 1, \cdots, p\,;\ i, k = 1, \cdots, n$. Consecuentemente

$$(X_{ij} - X_{kj})^2 = \begin{cases} 0 & , \text{ si } X_{ij} = X_{kj} = 1 \text{ o } X_{ij} = X_{kj} = 0 \\ 1 & , \qquad \text{si } X_{ij} \neq X_{kj} \end{cases} \,.$$

La distancia euclidiana proporciona el conteo de el número de "desigualdades"(no correspondencias).

Para analizar las correspondencias, y posibles combinaciones de los aciertos y no aciertos para los ítems i y k, se suele utilizar una tabla de contiengencia:

Retomando el ejemplo, se tiene $a = 2$, $b = 1$, $c = 1$ y $d = 1$.

A partir de este tipo de tablas, se pueden generar diferentes maneras de medir similaridad entre ítems, que permitan posteriormente agruparlos de manera adecuada. Algunas propuestas de coeficientes de similaridad para agrupamientos de cluster se muestran a continuación:

Ítem i \ Ítem k	1	0	totales
1	a	b	$a+b$
0	c	d	$c+d$
totales	$a+c$	$b+d$	$a+b+c+d=p$

Figura 8.1: Tabla de contingencia para aciertos y no aciertos, ítems i y k

1. $\dfrac{a+d}{p}$ iguales pesos para aciertos 1-1 *y* 0-0

2. $\dfrac{2\,(a+d)}{2\,(a+d)+(b+c)}$ doble peso para aciertos 1-1 *y* 0-0

3. $\dfrac{a+d}{(a+d)+2\,(b+c)}$ doble peso para no aciertos

4. $\dfrac{a}{p}$ ningún acierto 0-0

5. $\dfrac{a}{a+b+c}$ ningún acierto 0-0 en el numerador o el denominador

6. $\dfrac{2\,a}{2\,a+b+c}$ ningún acierto 0-0 en el numerador o el denominador
doble peso para aciertos 1-1

7. $\dfrac{a}{a+2\,(b+c)}$ ningún acierto 0-0 en el numerador o el denominador
doble peso para aciertos 0-0

8. $\dfrac{a}{b+c}$ razón de aciertos sobre no aciertos
pero excluyendo los aciertos $0-0$

Ejemplo 8.1.1. *Ciertas características asociadas con los más recientes presidentes de USA, se muestran en la tabla 8. Calcule los coeficientes de similaridad Tipo* 1.

Tabla 8: Datos de Expresidentes U.S.

Presidente	L. Nacimiento	Elegido Primer Término	Partido	Exp. Congreso	Fue Vice-presidente
1. R. Reagan	Medio oeste	SI	Republicano	NO	NO
2. J. Carter	Sur	SI	Demócrata	NO	NO
3. G. Ford	Medio oeste	NO	Republicano	SI	SI
4. R. Nixon	Oeste	SI	Republicano	SI	SI
5. L. Johnson	Sur	NO	Demócrata	SI	SI
6. J. Kennedy	Este	SI	Demócrata	SI	NO

Solución. Defina las siguientes variables:

$$X_1 = \begin{cases} 1 & , \text{ Lugar de nacimiento es el Sur} \\ 0 & , \text{ otro caso} \end{cases} ,$$

$$X_2 = \begin{cases} 1 & , \text{ Elegido en primer término} \\ 0 & , \text{ otro caso} \end{cases},$$

$$X_3 = \begin{cases} 1 & , \text{ Republicano} \\ 0 & , \text{ Demócrata} \end{cases},$$

$$X_4 = \begin{cases} 1 & , \text{ Tiene experiencia en el congreso} \\ 0 & , \text{ otro caso} \end{cases},$$

$$X_5 = \begin{cases} 1 & , \text{ Sirvio como Vicepresidente} \\ 0 & , \text{ otro caso} \end{cases}.$$

Los puntajes para los expresidentes 1 y 2 usando las $p = 5$ variables y la respectiva tabla de contingencia se muestran a continuación.

Ind 1		X_1	X_2	X_3	X_4	X_5
	1	0	1	1	0	0
	2	1	1	0	0	0

		Individuo 1		
		1	0	totales
Ind. 2	1	1	1	2
	0	1	2	3
	totales	2	3	5

Figura 8.2: Puntajes y Tabla de contingencia para los ítems 1 y 2

Usando el coeficiente 1, se tiene $\frac{a+d}{p} = \frac{3}{5}$. Realizando el mismo proceso para todas las combinaciones posibles y calculando los coeficientes respectivos de similaridad, se obtiene la siguiente matriz de similaridades:

$$\begin{pmatrix} 1 & & & & & \\ \frac{3}{5} & 1 & & & & \\ \frac{2}{5} & 0 & 1 & & & \\ \frac{3}{5} & \frac{1}{5} & \frac{4}{5} & 1 & & \\ 0 & \frac{2}{5} & \frac{3}{5} & \frac{2}{5} & 1 & \\ \frac{3}{5} & \frac{3}{5} & \frac{2}{5} & \frac{3}{5} & \frac{2}{5} & 1 \end{pmatrix}$$

Se observan dos grupos relativamente homogéneos: $(1, 2, 5, 6)$ y $(3, 4)$.

8.2. Medidas de similaridad y asociación para pares de variables

Cuando las variables son binarias, se pueden organizar los datos usando una tabla de contingencia. Sinembargo, las variables, más que los ítems, definen las categorías. Para cada par de variables, existen n ítems categorizados en la tabla. Por ejemplo

		Variable k		
		1	0	totales
Variable i	1	a	b	$a+b$
	0	c	d	$c+d$
	totales	$a+c$	$b+d$	$n=a+b+c+d$

Figura 8.3: Tabla de Contingencia para variables i y k

Observe que el par $1-1$ se dá en a de los n ítems de la muestra. Similarmente, el par $0-0$ se dá d veces. En este caso puede calcularse un coeficiente de correlación, el cual se obtiene como:

$$r = \frac{a\,d - b\,c}{[(a+b)\,(c+d)\,(a+c)\,(b+d)]^{\frac{1}{2}}}\ .$$

Este número es tomado como una medida de similaridad entre dos variables; además, es usado como estadístico de prueba χ^2 para probar la independencia de dos variables categóricas: $r^2 = \frac{\chi^2}{n}$.

8.3. Métodos jerárquicos de agrupamientos

Los métodos jerárquicos de agrupamiento pueden ser catalogados de dos maneras: Por una serie de uniones o de divisiones sucesivas. En el primer caso se habla de métodos **Aglomerativos**, en el segundo de métodos **Divisorios**. En el primer caso se tienen tantos clusters como objetos. En el segundo se parte de un solo cluster que contiene todos los objetos.
Los resultados de ambos procedimientos son usualmente mostrados en forma de un diagrama bidimensional, llamado **Dendograma**. Este diagrama muestra las márgenes o divisiones que han sido hechas a cada nivel de agrupamiento.

Los métodos que se considerarán, son jerárquicos aglomerativos y en pareticular, los llamados de enlace o *Linkeo*. Estos métodos son deseables para agrupar ítems o variables. Entre estos métodos están: *Linkeo simple* (mínima distancia entre clusters), *Linkeo completo* (máxima distancia entre clusters) y *Linkeo promedio* (distancias pormedio entre clusters). En la figura 8.4 se ilustra el proceso en los tres métodos:

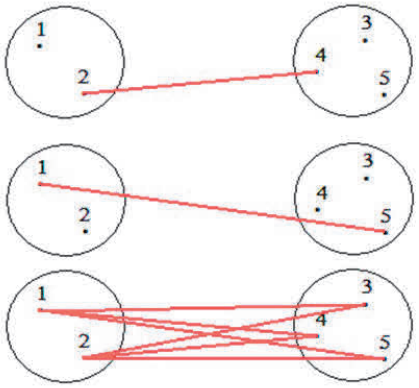
Los siguientes son los pasos a seguir en el algorítmo de agrupamiento jerárquico aglomerativo para agrupar N objetos (ítems o variables):

1. Se empieza con N clusters, cada uno contiene una sola entidad y una matriz simétrica $D_{N\times N}$ de distancias o similaridades, con $D = [(d_{ij})]$.

2. Busque dentro de la matriz D el par de clusters más similares. Suponga que dichos clusters son U y V y sea d_{UV}la distancia entre estos clusters.

3. Una los clusters U y V para formar un nuevo cluster y denótelo (UV). Actualice las entradas de la matriz D haciendo:

   - Borrando las filas y columnas de D correspondientes a los clusters U y V



Distancia entre Clusters Disimilaridad	Tipo de Linkeo
d_{24}	Simple
d_{15}	Completo
$\frac{d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25}}{6}$	Promedio

Figura 8.4: Métodos de Linkeo

- Adiccione una nueva fila y columna que contiene las distancias o similaridades entre el cluster (UV) y los demás clusters.

4. Repita los pasos 2 y 3 un total de $N-1$ veces. Registre la identidad de los clusters agrupados y los niveles (distancias o similaridades) para los cuales dichas agrupaciones son hechas.

8.3.1. Linkeo Simple

Las entradas en el algorítmo de linkeo simple pueden ser distancias o similaridades entre pares de objetos. Los grupos son formados de entidades individuales juntando los vecinos más cercanos (menor distancia o mayor similaridad). La distancia entre un cluster (UV) y otro cluster W se denotará $d_{(UV)W}$ y se calcula como: $d_{(UV)W} = \min\{d_{UW}, d_{VW}\}$, donde d_{UW} es la distancia entre los miembros más cercanos de los clusters U y W y d_{VW} es la distancia entre los miembros más cercanos de los clusters V y W. Estos resultados se presentan graficamente en un Dendograma, donde los clusters se representan en paréntesis, y estos a su vez se muestran junto con unos nodos cuya posición a lo largo del eje Y (de distancias), indican el nivel al cual el respectivo cluster es formado.

8.3.2. Linkeo completo

En este procedimiento, muy similar al linkeo simple, hay una diferencia: en cada paso la distancia entre clusters es determinada por la distancia entre los dos elementos de ambos clusters más distantes. Todos los ítems de un cluster son los que distan más unos de otros. El algorítmo empieza encontrando la mínima entrada en la matriz D y juntando los correspondientes objetos, digamos U y V, formando el cluster (UV). La distancia entre el Cluster (UV) y otro cluster W, se obtiene como $d_{(UV)W} = \max\{d_{UW}, d_{VW}\}$, donde d_{UW} es la distancia entre los miembros más lejanos de los clusters U y W; y d_{VW} es la distancia entre miembros más lejanos de los clusters V y W.

8.3.3. Linkeo promedio

Se empieza buscando en la matriz D los dos objetos más cercanos, digamos U y V. Estos se unen para formar el cluster (UV). Luego la distancia entre el cluster (UV) y W se obtiene como:

$$d_{(UV)W} = \frac{\sum_{i=1} \sum d_{ik}}{N_{(UV)}\, N_W} \,,$$

donde d_{ik} es la distancia entre el objeto i del cluster (UV) y el objeto k del cluster W; $N_{(UV)}$ y N_W son los tamaños de los cluster (UV) y W respectivamente.

Ejemplo 8.3.1. *Las distancias entre* 5 *pares de ítems se muestran en la figura 8.5. Agrupe los* 5 *ítems, usando los tres tipos de linkeos: Simple, Compuesto y Promedio.*

	1	2	3	4	5
1	0				
2	4	0			
3	6	9	0		
4	1	7	10	0	
5	6	3	5	8	0

Figura 8.5: Matriz de Distancias

Solución

- Usando *Linkeo Simple*. De la figura anterior se observa que la mínima distancia es $d_{14} = 1$. Así, formamos el cluster (14). Calculamos las distancias del cluster (14) a los demás clusters:

$$d_{(14)2} = \min\{d_{12},\, d_{42}\} = \min\{4,\, 7\} = 4$$

$$d_{(14)3} = \min\{d_{13},\, d_{43}\} = \min\{6,\, 10\} = 6$$

$$d_{(14)5} = \min\{d_{15},\, d_{45}\} = \min\,,\, 8\} = 6\,.$$

  La nueva matriz de distancias se muestra en la figura 8.6.

	(14)	2	3	5
(14)	0			
2	4	0		
3	6	9	0	
5	6	3	5	0

Figura 8.6: Matriz de distancias con Cluster (14)

La menor distancia entre pares de clusters es $d_{25} = 3$ y así, un nuevo cluster puede ser formado: (25). Se construye una nueva matriz de distancias considerando solo los cluster (14), (25) y 3: $d_{(14)3} = 6$,

$$d_{(14)(25} = \min\{d_{(14)2},\, d_{(14)5}\} = \min\{4,\, 6\} = 4$$

$$d_{(25)3} = \min\{d_{23},\, d_{53}\} = \min\{9,\, 5\} = 5\ .$$

La nueva matriz de distancias se muestra en la figura 8.7.

	(14)	(25)	3
(14)	0		
(25)	4	0	
3	6	5	0

Figura 8.7: Matriz de distancias con Clusters (14) y (25)

La menor distancia entre pares de clusters es $d_{(14)(25)} = 4$. Así, el siguiente cluster es (25 14) y $d_{(25\,14)3} = \min\{d_{(14)3},\, d_{(25)3}\} = \min\{6,\, 5\} = 5$. La nueva matriz se muestra en al figura 8.8. El

	(14 25)	3
(14 25)	0	
3	5	0

Figura 8.8: Matriz de distancias con Clusters (14 25) y 3

último cluster es (14 25 3). El respectivo dendograma se muestra en la figura 8.9.

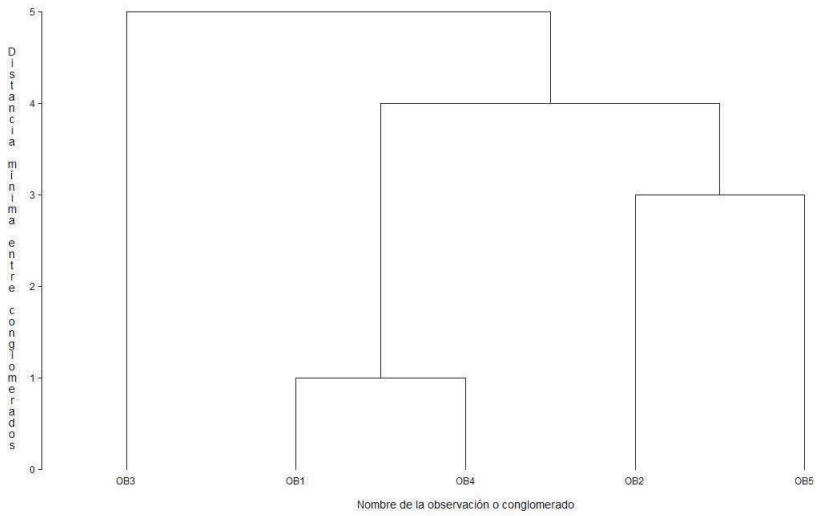


Figura 8.9: Dendograma Linkeo Simple

- Usando *Linkeo Completo*. De la figura 8.5 se observa que la mínima distancia es $d_{14} = 1$. Así, formamos el cluster (14). Calculamos las distancias del cluster (14) a los demás clusters:

$$d_{(14)2} = \max\{d_{12},\, d_{42}\} = \max\{4,\, 7\} = 7$$

$$d_{(14)3} = \max\{d_{13},\, d_{43}\} = \max\{6,\, 10\} = 10$$

$$d_{(14)5} = \max\{d_{15},\, d_{45}\} = \max,\, 8\} = 8\ .$$

La nueva matriz de distancias se muestra en la figura 8.10. La menor distancia entre pares de

	(14)	2	3	5
(14)	0			
2	7	0		
3	10	9	0	
5	8	3	5	0

Figura 8.10: Matriz de distancias con Cluster (14)

clusters es $d_{25} = 3$ y así, un nuevo cluster puede ser formado: (25). Se construye una nueva matriz de distancias considerando solo los cluster (14), (25) y 3: $d_{(14)3} = 10$,

$$d_{(14)(25} = \max\{d_{(14)2},\, d_{(14)5}\} = \max\{7,\, 5\} = 7$$

$$d_{(25)3} = \max\{d_{23},\, d_{53}\} = \max\{9,\, 5\} = 9\ .$$

La nueva matriz de distancias se muestra en la figura 8.11.

	(14)	(25)	3
(14)	0		
(25)	8	0	
3	10	9	0

Figura 8.11: Matriz de distancias con Clusters (14) y (25)

La menor distancia entre pares de clusters es $d_{(14)(25)} = 4$. Así, el siguiente cluster es (25 14) y

$$d_{(25\,14)3} = \max\{d_{(14)3},\, d_{(25)3}\} = \max\{10,\, 9\} = 10\ .$$

La nueva matriz se muestra en al figura 8.12. El último cluster es (14 25 3). El respectivo dendograma se muestra en la figura 8.13. Observe que se obtienen los mismos cluster en el mismo orden que usando linkeo simple, la diferencia está en la escala del eje vertical.

	(14 25)	3
(14 25)	0	
3	10	0

Figura 8.12: Matriz de distancias con Clusters (14 25) y 3

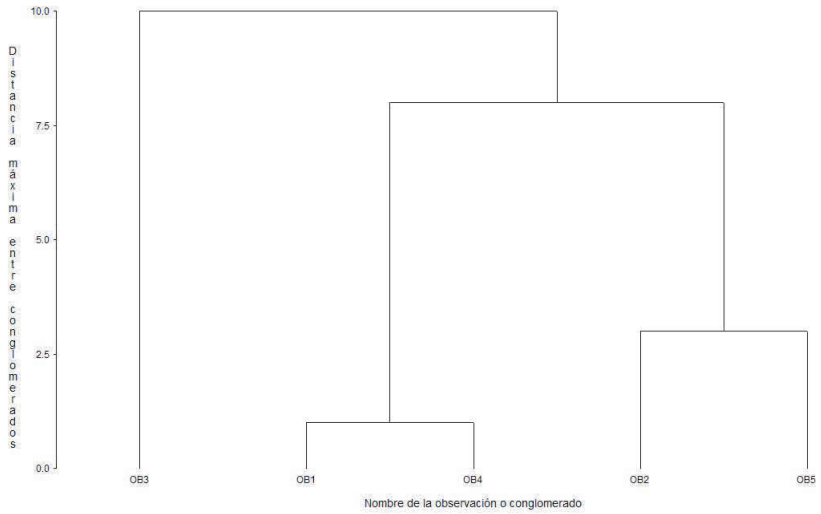


Figura 8.13: Dendograma Linkeo Completo

- Usando *Linkeo Promedio*.

  De la figura 8.5 se observa que la mínima distancia es $d_{14} = 1$. Así, formamos el cluster (14). Calculamos las distancias del cluster (14) a los demás clusters:

$$d_{(14)2} = \frac{d_{12} + d_{42}}{(2)(1)} = \frac{4+7}{2} = 5.5$$

$$d_{(14)3} = \frac{d_{13} + d_{43}}{(2)(1)} = \frac{6+10}{2} = 8$$

$$d_{(14)5} = \frac{d_{15} + d_{45}}{(2)(1)} = \frac{6+8}{2} = 7\,.$$

  La nueva matriz de distancias se muestra en la figura 8.14. La menor distancia entre pares de

	(14)	2	3	5
(14)	0			
2	5.5	0		
3	8	9	0	
5	7	3	5	0

Figura 8.14: Matriz de distancias con Cluster (14)

clusters es $d_{25} = 3$ y así, un nuevo cluster puede ser formado: (25). Se construye una nueva matriz de distancias considerando solo los cluster (14), (25) y 3: $d_{(14)3} = 8$,

$$d_{(14)(25} = \frac{d_{12} + d_{15} + d_{42} + d_{45}}{(2)(2)} = \frac{4 + 6 + 7 + 8}{4} = 6.25$$

$$d_{(25)3} = \frac{d_{23} + d_{53}}{(2)(1)} = \frac{9 + 5}{2} = 7 \, .$$

La nueva matriz de distancias se muestra en la figura 8.15.

	(14)	(25)	3
(14)	0		
(25)	6.25	0	
3	8	7	0

Figura 8.15: Matriz de distancias con Clusters (14) y (25)

La menor distancia entre pares de clusters es $d_{(14)(25)} = 4$. Así, el siguiente cluster es (14 25) y

$$d_{(14\,25)3} = \frac{d_{23} + d_{53} + d_{13} + d_{43}}{(4)(1)} = \frac{9 + 5 + 6 + 10}{4} = 7.5 \, .$$

La nueva matriz se muestra en al figura 8.16.

	(14 25)	3
(14 25)	0	
3	7.5	0

Figura 8.16: Matriz de distancias con Clusters (25 14) y 3

El último cluster es (3 25 14). El respectivo dendograma se muestra en la figura 8.17.

Ejemplo 8.3.2. *Se tiene información sobre las utilidades de* 22 *compañías públicas en estados Unidos en 1975. Nueve variables son registradas:* X_1*: ingreso/deduda,* X_2*: Tasa de retorno de capital,* X_3*: Costo per-cápita en el lugar,* X_4*: Factor de carga anual,* X_5*: Crecimiento per-cápita de la demanda de 1974 a 1975,* X_6*: Ventas (uso per-cápita de KWH anual),* X_7*: Porcentaje nuclear,* X_8*: Costo total de combustible (centavos por KWH) y Lugar. Los datos se muestran a continuación.*

Datos sobre utilidades

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	*Lugar*
1.06	*9.2*	*151*	*54.4*	*1.6*	*9077*	*0.0*	*0.628*	*Arizona*
0.89	*10.3*	*202*	*57.9*	*2.2*	*5088*	*25.3*	*1.555*	*Boston*
1.43	*15.4*	*113*	*53.0*	*3.4*	*9212*	*0.0*	*1.058*	*Central*
1.02	*11.2*	*168*	*56.0*	*0.3*	*6423*	*34.3*	*0.700*	*Common*
1.49	*8.8*	*192*	*51.2*	*1.0*	*3300*	*15.6*	*2.044*	*Consolid*
1.32	*13.5*	*111*	*60.0*	*-2.2*	*11127*	*22.5*	*1.241*	*Florida*
1.22	*12.2*	*175*	*67.6*	*2.2*	*7642*	*0.0*	*1.652*	*Hawaiian*
1.10	*9.2*	*245*	*57.0*	*3.3*	*13082*	*0.0*	*0.309*	*Idaho*
1.34	*13.0*	*168*	*60.4*	*7.2*	*8406*	*0.0*	*0.862*	*Kentucky*
1.12	*12.4*	*197*	*53.0*	*2.7*	*6455*	*39.2*	*0.623*	*Madison*
0.75	*7.5*	*173*	*51.5*	*6.5*	*17441*	*0.0*	*0.768*	*Nevada*
1.13	*10.9*	*178*	*62.0*	*3.7*	*6154*	*0.0*	*1.897*	*NewEngla*
1.15	*12.7*	*199*	*53.7*	*6.4*	*7179*	*50.2*	*0.527*	*Northern*
1.09	*12.0*	*96*	*49.8*	*1.4*	*9673*	*0.0*	*0.588*	*Oklahoma*
0.96	*7.6*	*164*	*62.2*	*-0.1*	*6468*	*0.9*	*1.400*	*Pacific*
1.16	*9.9*	*252*	*56.0*	*9.2*	*15991*	*0.0*	*0.620*	*Puget*
0.76	*6.4*	*136*	*61.9*	*9.0*	*5714*	*8.3*	*1.920*	*SanDiego*
1.05	*12.6*	*150*	*56.7*	*2.7*	*10140*	*0.0*	*1.108*	*Southern*
1.16	*11.7*	*104*	*54.0*	*-2.1*	*13507*	*0.0*	*0.636*	*Texas*
1.20	*11.8*	*148*	*59.9*	*3.5*	*7287*	*41.1*	*0.702*	*Wisconsi*
1.04	*8.6*	*204*	*61.0*	*3.5*	*6650*	*0.0*	*2.116*	*United*
1.07	*9.3*	*174*	*54.3*	*5.9*	*10093*	*26.6*	*1.306*	*Virginia*

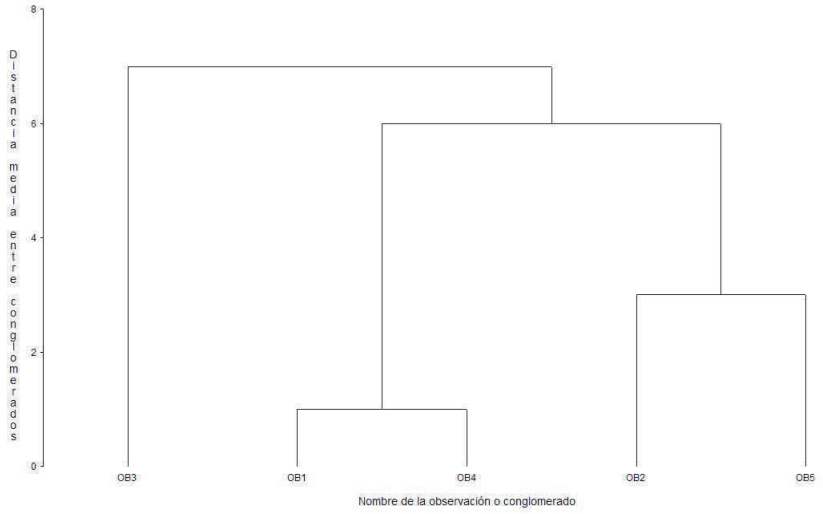


Figura 8.17: Dendograma Linkeo Promedio

Realice un análisis de cluster agrupando primero ítems y luego variables. Para el primer caso utilice el linkeo simple; para el segundo el linkeo completo.

Solución.

- *Agrupando observaciones (compañías).* El código en R usado es:

```
uti <- read.table(file.choose(), header=T)
clu_comp <- hclust(dist(uti[,1:8]),"single")
plot(clu_comp, xlab="Compañias", ylab="Distancia")
```

El dendograma obtenido se muestra en la figura 8.18.

En la figura 8.19 se muestra el dendograma con la propuesta de 5 grupos.

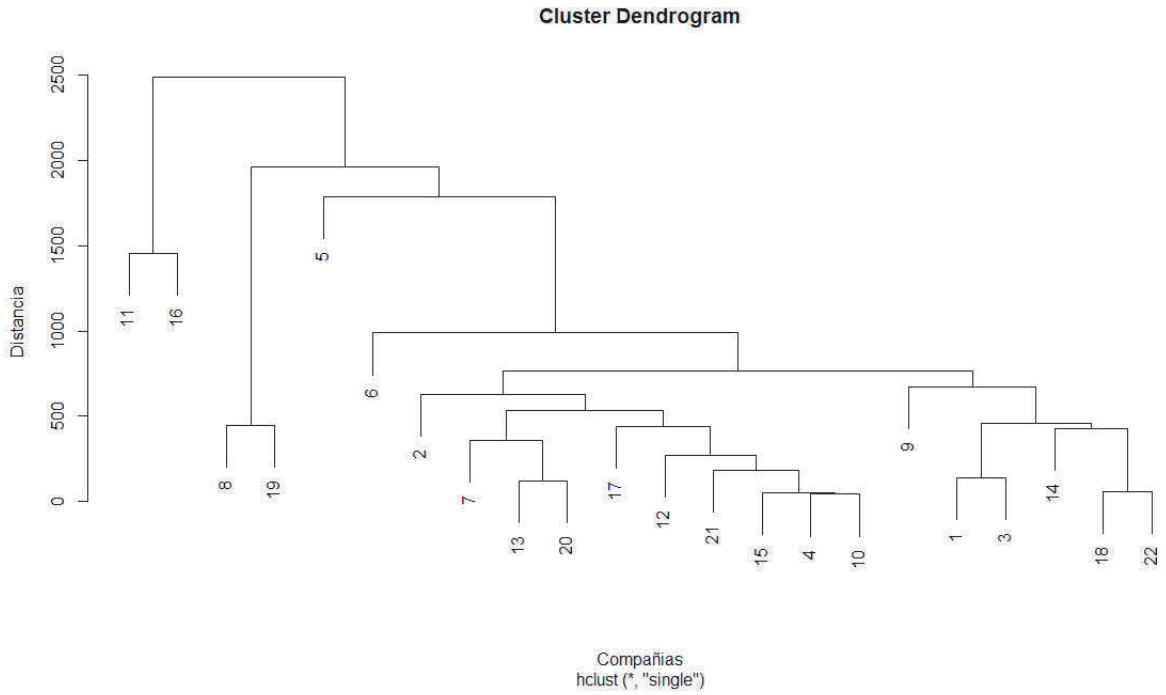


Figura 8.18: Dendograma Linkeo Simple

- *Agrupando Variables* El código en R y el respectivo dendograma, 8.20, con tres grupos propuestos es:

```
library(ClustOfVar)

tree <- hclustvar(uti[,1:8])
plot(tree)
rect.hclust(tree, k=3, border="red")
```

Otro método jerárquico de cluster es conocido como método *Ward's*.

El criterio de varianza mínima de Ward minimiza la varianza total dentro del cluster (ESS). Para implementar este método, en cada paso se encuetra el par de grupos que conducen a un aumento mínimo en la varianza total dentro del cluster fusionado. Este aumento es una distancia al cuadrado ponderada entre los centroides de los cluster. En el paso inicial, todos los grupos son singletons (grupos que contienen un solo punto).

Para aplicar un algoritmo recursivo bajo esta función objetivo, la distancia inicial entre objetos individuales debe ser proporcional a la distancia euclidiana al cuadrado. Las distancias iniciales en este método, son las distancias ecuclidianas al cuadrado.
Para un par de observaciones p-variadas $\mathbf{X}_i$ y $\mathbf{X}_j$, se tiene que

$$d_{ij} = d\left(\mathbf{X}_i\,,\,\mathbf{X}_j\right) = \|\mathbf{X}_i - \mathbf{X}_j\|^2\,.$$

Es importante verificar en el software que se use para implementar el método Ward's, si se usan las distacias euclidianas o las euclidianas al cuadrado.

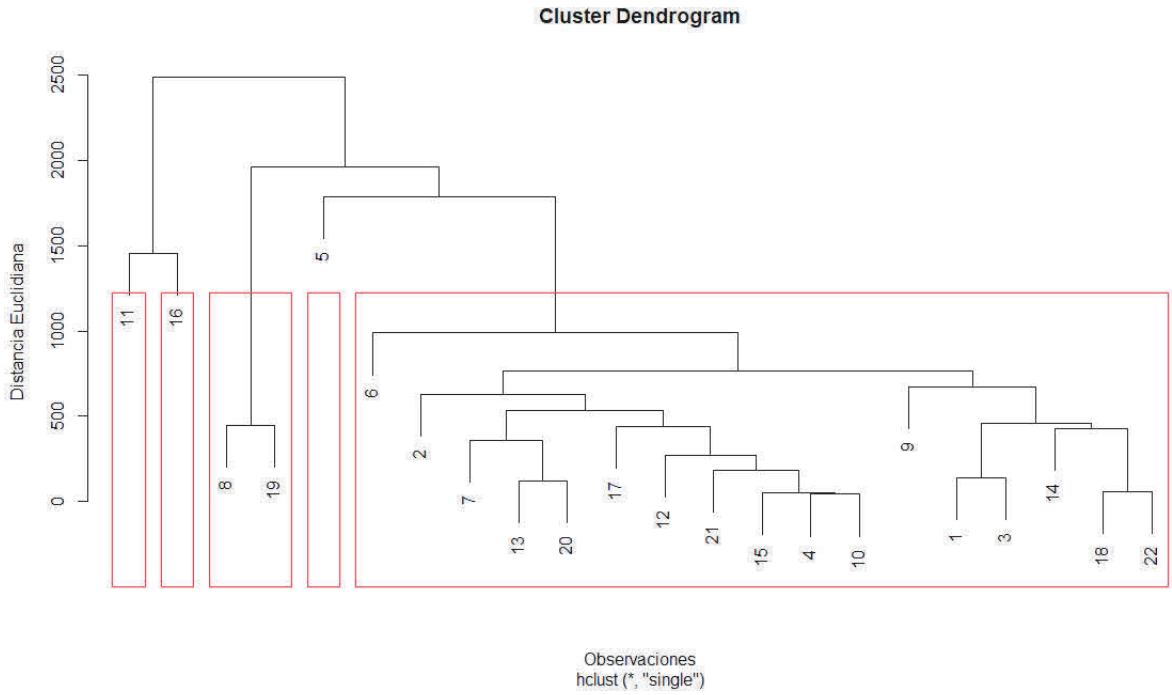


Figura 8.19: Dendograma Linkeo Simple con grupos

Inicialmente cada cluster consiste de un solo ítem, y si se tienen N ítems, entonces $ESS_k = 0$, para $k = 1, \cdots, N$. Con esto $ESS = 0$. Cuando todos los ítems se agrupan en un solo cluster, se obtiene el máximo valor para ESS, el cual está dado por:

$$ESS = \sum_{j=1}^{N} (\mathbf{X}_j - \bar{\mathbf{X}})' (\mathbf{X}_j - \bar{\mathbf{X}}) ,$$

donde $\mathbf{x}_j$ es la j-ésima medida multivariada asociada al ítem j y $\bar{\mathbf{X}}$ es el vector de medias muestral multivariado. Los resultados de este método se pueden presentar en un dendograma, donde el eje vertical corresponde a los valores de ESS.

Ejemplo 8.3.3. *Retomando los datos sobre las utilidades de las compañias, el respectivo código en R se muestra a continuación:*

```
# Metodo de Ward's Minima Varianza entre Clusters
# package:FactoClass
# ward.cluster
# Cargar paquete ggplot2

hw_uti <- ward.cluster(dist(uti[,1:8]), h.clust = 1)
plot(hw_uti)
rect.hclust(hw_uti, k=3, border=2:10)
```

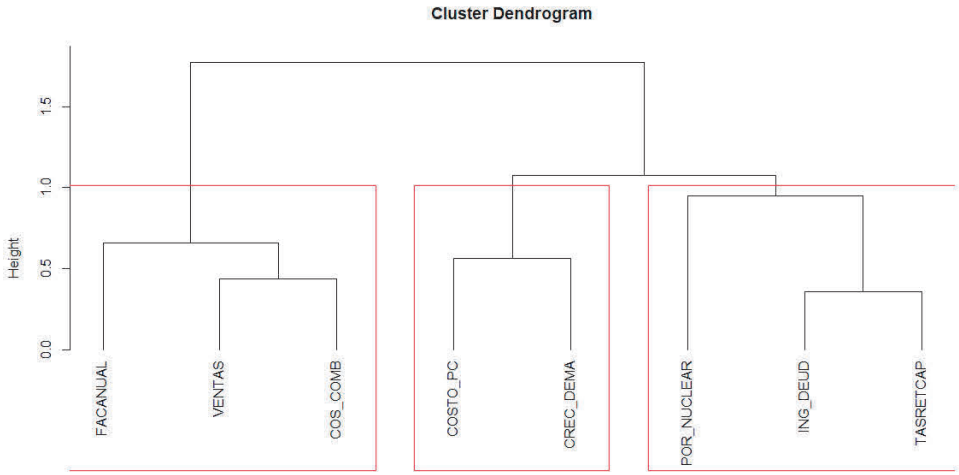


Figura 8.20: Dendograma Para Variables

```
clu_uti <- cutree(hw_uti, k = 3)
fviz_cluster(list(data = uti[,1:8], cluster = clu_uti))
```

Los graficos se muestran en las figuras 8.21 y 8.22.

Clusters usando K-Means.

Dentro de los métodos no jerárquicos, se destaca el K-means, o los métodos de agrupamiento basados en modelos. Para estos últimos los criterios de Información de Akaike y el criterio de información Bayesiano BIC, suelen se empleados para definir un número adecuado de clusters. Dado un conjunto de observaciones p-variadas, el método permite encontrar o particionar estas n observaciones en k conjuntos $\mathbf{S} = \{S_1, \cdots, S_k\}$*, con* $k \leq n$*, de manera que se minimice la suma de cuadrados entre clusters, es decir, la varianza.*

Formalmente, el objetivo es encontrar el conjunto $\mathbf{S}$ *que minimice:*

$$\overset{argmin}{\mathbf{S}} \sum_{1=1}^{k} \sum_{\mathbf{X} \in S_i} \|\mathbf{X} - \boldsymbol{\mu}_i\|^2 \; = \overset{argmin}{\mathbf{S}} \sum_{1=1}^{k} |S_i| \, Var[S_i] \; ,$$

donde $\boldsymbol{\mu}_i$ *es la media de los puntos en* S_i*. Esto es equivalente a minimizar los pares de desviaciones al cuadrado de los puntos dentro del mismo cluster:*

$$\overset{argmin}{\mathbf{S}} \sum_{1=1}^{k} \frac{1}{2\,|S_i| \, \sum_{\mathbf{X}, \mathbf{Y} \in S_i} \|\mathbf{X} - \mathbf{Y}\|^2} \; .$$

Ejemplo 8.3.4. *Nuevamente usando los datos sobre las utilidades de las compañias, el respectivo código en R se muestra a continuación:*

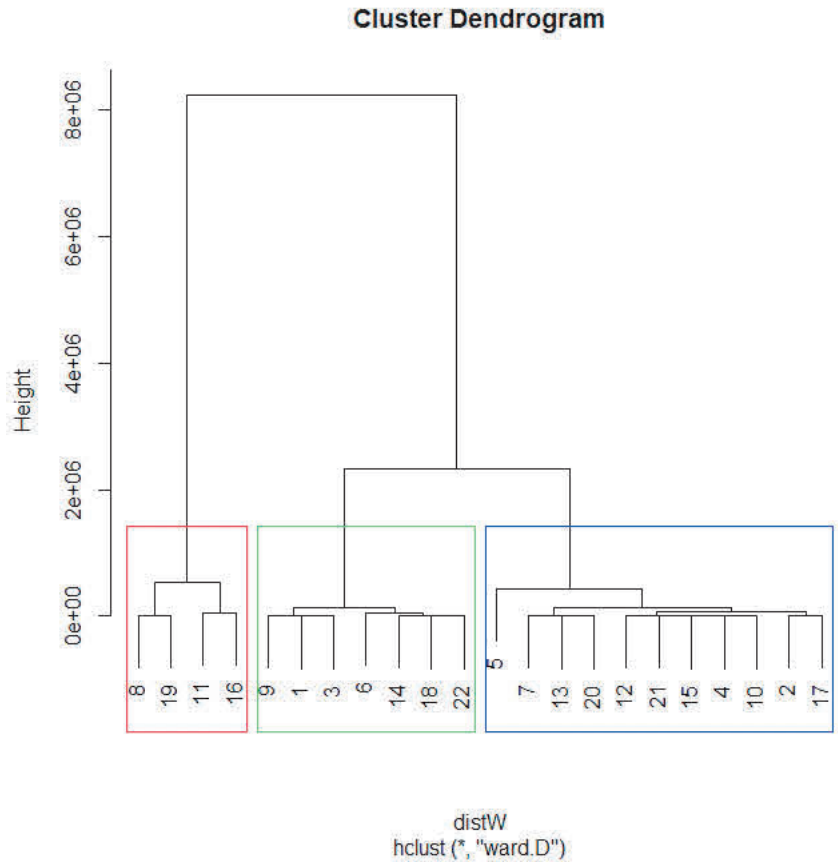


Figura 8.21: Dendograma para Cias Método Ward's

```
# Distancia entre Clusters Mínima usando Kmeans
# El algoritmo de Hartigan y Wong (1979) se usa por defecto.

cia_km <- kmeans(uti[,1:8], centers = 3, nstart = 10)
fviz_cluster(cia_km, data = uti[,1:8])
fviz_nbclust(x = uti[,1:8],FUNcluster = kmeans, method = 'wss' )
```

El gráfico resultante se muestra en la figura 8.23.

Metodos para seleccionar número óptimo de clusters.

- **Método de Elbow**

  El método Elbow analiza el WSS (suma de cuadrados total entre clusters) en función del número de clústers. Se debe elegir un número de clusters de manera que agregar otro clúster no mejore mucho el WSS total. El número óptimo de clusters se puede definir de la siguiente manera:

  - *Calcule el algoritmo de agrupación (por ejemplo kmeans, Ward's, etc) para diferentes valores de k (núm,erod e grupos). Por ejemplo, variando k de 1 a 10 grupos.*

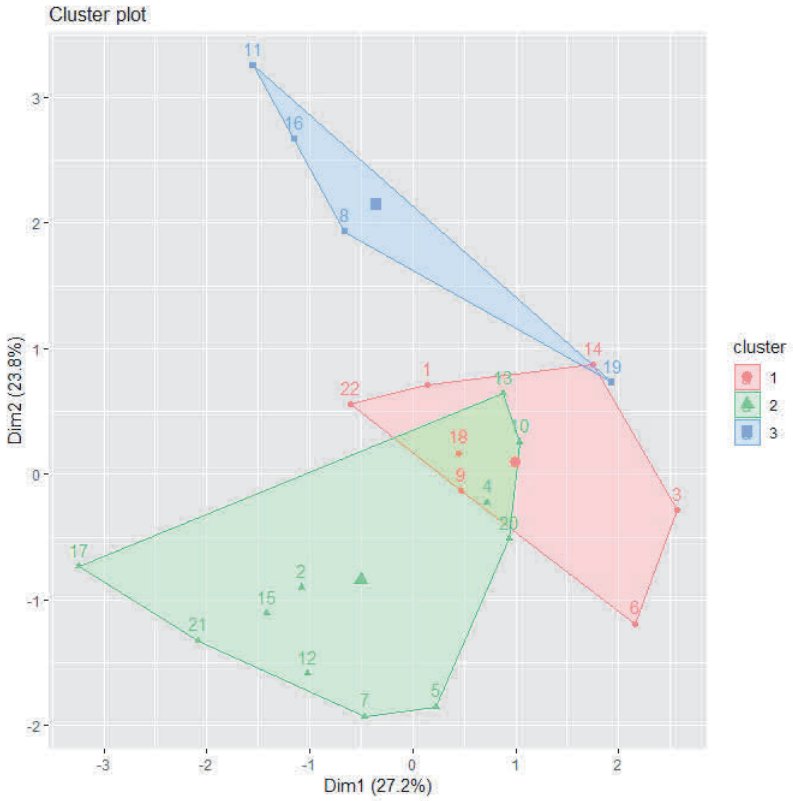


Figura 8.22: Clusters 2D para Compañias Método Ward's

- *Para cada k, calcule la suma total del cuadrado dentro del clúster (WSS).*
- *Trace la curva de WSS en función de k.*
- *La ubicación de una curva en forma de rodilla en la gráfica, generalmente se considera como un indicador del número apropiado de grupos.*

- **Método Silhouette Promedio**
  Este método mide la calidad de una agrupación, es decir, determina qué tan bien se encuentra cada objeto dentro de su grupo. Un alto valor de este indicador indica un buen agrupamiento.

  Este método calcula la silueta promedio de las observaciones para diferentes valores de k. El número óptimo de grupos k es el que maximiza la silueta promedio en un rango de valores posibles para k (Kaufman y Rousseeuw 1990).

  El algoritmo es similar al método de la rodilla y se puede calcular de la siguiente manera:
  - *Calcule el algoritmo de agrupación (por ejemplo kmeans, Ward's, etc) para diferentes valores de k (núm,erod e grupos). Por ejemplo, variando k de 1 a 10 grupos.*
  - *Para cada k, calcule la silueta promedio de las observaciones (avg.sil).*
  - *Trace la curva de avg.sil en función del número de grupos k.*
  - *La ubicación del máximo se considera como el número apropiado de clústeres.*

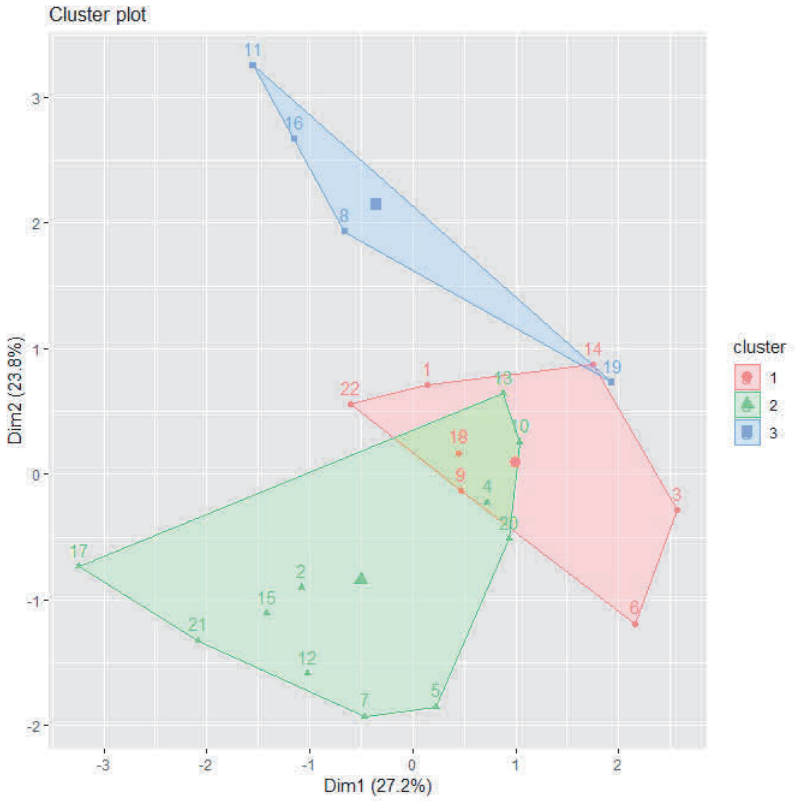


Figura 8.23: Clusters 2D para Compañias Método Kmeans

- **Método Gap** *El estadístico Gap fué publicado por R. Tibshirani, G. Walther y T. Hastie (Universidad de Standford, 2001). El enfoque se puede aplicar a cualquier método de agrupación.*

  Este estadístico compara la variación total dentro del cluster para diferentes valores de k, con sus valores esperados bajo una distribución de referencia nula de los datos. La estimación de los grupos óptimos será un valor que maximice este estadístico (es decir, que produzca el estadístic Gap más grande). Esto significa que la estructura de agrupamiento está muy lejos de la distribución aleatoria uniforme de puntos.

Algunos comandos en r, para realizar los gráficos antes mencionados, usando los datos de las utilidades de las compañias son:

```
# Determining Optimal clusters (k) Using WSS, Average Silhouette and Gap
fviz_nbclust(x = uti[,1:8], FUNcluster = kmeans, method = c("silhouette", "wss", "gap_stat") )
fviz_nbclust(x = uti[,1:8], FUNcluster = kmeans, method = 'wss' )
fviz_nbclust(x = uti[,1:8], FUNcluster = kmeans, method ="silhouette")
fviz_nbclust(x = uti[,1:8], FUNcluster = kmeans, method ="gap_stat")
```

Los respectivos gráficos se muestran en las figuras 8.24, 8.25 y 8.26.

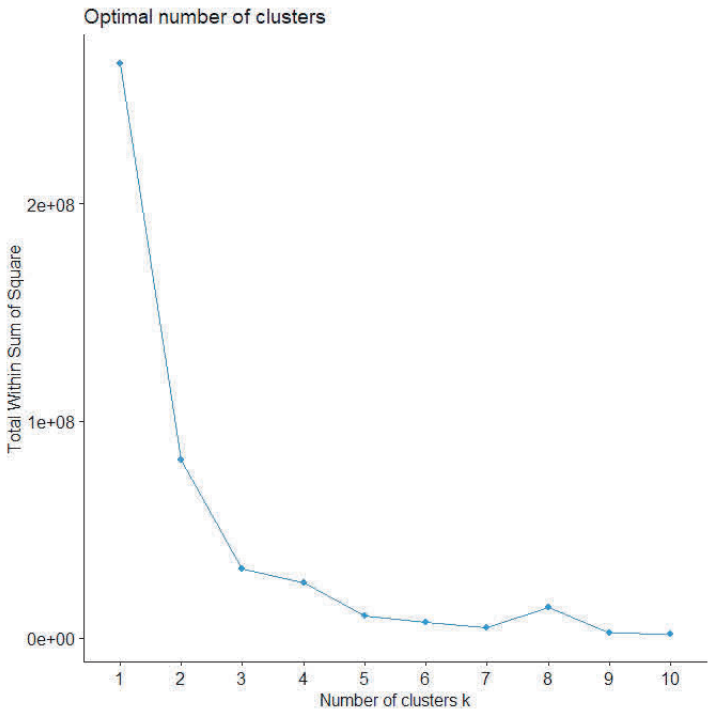


Figura 8.24: No de clusters óptimos usando método WSS Elbow

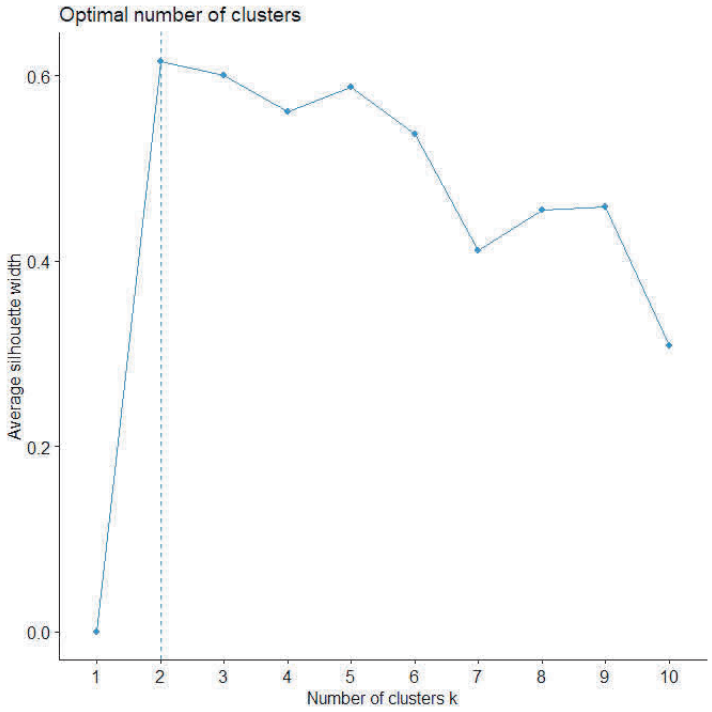


Figura 8.25: No de clusters óptimos usando método Silhouette

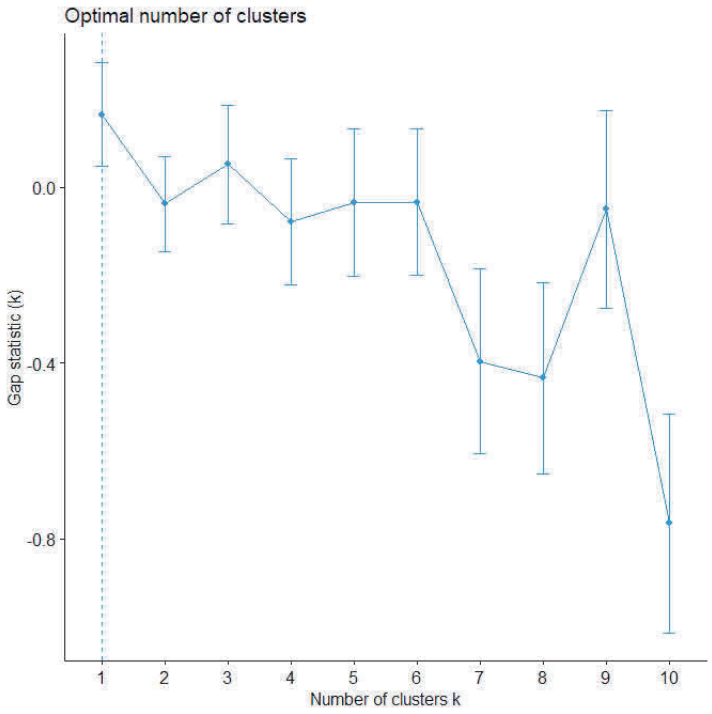


Figura 8.26: No de clusters óptimos usando método Gap