ANÁLISIS MULTIVARIADO

Algo de historia

Los métodos estadísticos multivariados, en su forma más simple, hacen referencia al análisis simultáneo de dos o más variables aleatorias. El primer método para medir la relación estadística entre dos variables se debe a Francis Galton (1822 – 1911), que introduce el concepto de *recta de regresión* y la idea de *correlación entre variables* en su libro *Natural Inheritance*, publicado en 1889 cuando Galton tenía 67 años. Estos descubrimientos surgen en sus investigaciones sobre la transmisión de los rasgos hereditarios, motivadas por su interés en contrastar empíricamente la teoría de la evolución de las especies, propuesta por su *primo Charles Darwin en 1859*. El concepto de correlación es aplicado en las ciencias sociales por Francis Edgeworth (1845 – 1926), que estudia la *normal multivariada* y la *matriz de correlación*. Karl Pearson (1857 – 1936), un distinguido estadístico británico creador del famosa χ^2 *de Pearson*, obtuvo el estimador del coeficiente de correlación muestral, y se enfrentó al problema de determinar si dos grupos de personas, de los que se conocen su medidas físicas, pertenen a la misma raza (problema simple de discriminación de poblaciones). Este problema intrigó a Harold Hotelling (1885 – 1973), un joven matemático y economista estadounidense, que, atraído por la Estadística, entonces una joven disciplina emergente, viaja en 1929 a la estación de investigación agrícola de Rothamsted en el Reino Unido para trabajar con el ya célebre científico y figura destacada de la estadística, R. A. Fisher (1890 – 1962). Hotelling se interesó por el problema de comparar tratamientos agrícolas en función de varias variables, y descubrió las semejanzas entre este problema y el planteado por Pearson. Debemos a Hotelling (1931) el contraste que lleva su nombre (T de Hotelling), que permite comparar si dos muestras multivariadas provienen de la misma población. A su regreso a la Universidad de Columbia en Nueva York, Truman Kelley, profesor de pedagogía en Harvard, planteó a Hotelling el problema de encontrar los factores capaces de explicar los resultados obtenidos por un grupo de personas en pruebas (test) de inteligencia. Hotelling (1933) inventó *los componentes principales*, que son indicadores capaces de resumir de forma óptima un conjunto amplio de variables y que dan lugar, posteriormente, al *análisis factorial*. El problema de obtener el mejor indicador resumen de un conjunto de variables había sido abordado y resuelto desde otro punto de vista por Karl Pearson en 1921, en su trabajo para

encontrar el plano de mejor ajuste a un conjunto de observaciones astronómicas. Posteriormente, Hotelling generaliza la idea de componentes principales introduciendo el *análisis de correlación canónica*, que permiten resumir simultáneamente dos conjuntos de variables.

El problema de encontrar factores que expliquen los datos fue planteado por primera vez por Charles Spearman ($1863 - 1945$), que observó que los niños que obtenían buenas puntuaciones en un test de habilidad mental también las obtenían en otros, lo que le llevó a postular que se debían a un factor general de inteligencia, el factor g (Spearman, 1904). L. Thurstone ($1887 - 1955$) estudió el modelo con varios factores y escribió uno de los primeros textos de análisis factorial (Thurstone, 1947). El análisis factorial fue considerado hasta los años 60 como una técnica psicométrica con poca base estadística, hasta que los trabajos de Lawley y Maxwell (1971) establecieron formalmente la estimación y el contraste del modelo factorial bajo la hipótesis de normalidad. Desde entonces, las aplicaciones del modelo factorial se han extendido a todas las ciencias sociales. La generalización del modelo factorial cuando tenemos dos conjuntos de variables y unas explican la evolución de las otras es el modelo *de ecuaciones estructurales*, que ha sido ampliamente estudiado por Joreskov (1973), entre otros.

La primera solución al problema de clasificación se debe a Fisher en 1933. Fisher inventa un método general, basado en el análisis de la varianza, para resolver un problema de discriminación de cráneos en antropología. El problema era clasificar un cráneo encontrado en una excavación arqueológica como perteneciente o no a un homínido (término que se utiliza para nombrar al ejemplar que pertenece al orden de los primates superiores, que tienen al ser humano (Homo sapiens) como la única especie que sobrevive). La idea de Fisher es encontrar una variable indicadora, combinación lineal de las variables originales de las medidas del cráneo, que consiga máxima separación entre las dos poblaciones en consideración. En 1937 Fisher visita la India invitado por P. C. Mahalanobis ($1893 - 1972$), que había inventado la medida de distancia que lleva su nombre, para investigar las diferentes razas en la India. Fisher percibe enseguida la relación entre la *medida (distancia) de Mahalanobis* y sus resultados en *análisis discriminante* y ambos consiguen unificar estas ideas y relacionarlas con los resultados de Hotelling sobre el contraste de medias de poblaciones multivariadas. Unos años después, un estudiante de Mahalanobis, C. R. Rao, va a extender el análisis de Fisher para clasificar un elemento en más de dos poblaciones.

Las ideas anteriores se desarrollan para variables cuantitativas (numéricas), pero se aplican

poco después a variables cualitativas o atributos (categóricas). Karl Pearson había introducido el estadístico que lleva su nombre para contrastar la independencia en una tabla de contingencia y Fisher, en 1940, aplica sus ideas de análisis discriminante a estas tablas. Paralelamente, Guttman ($1916-1987$), en Psicometría, presenta un procedimiento para asignar valores numéricos (construir escalas) a variables cualitativas que está muy relacionado con el método de Fisher. Como este último trabaja en Biometría, mientras Guttman lo hace en Psicometría, la conexión entre sus ideas tardó más de dos décadas en establecerse. En Ecología, Hill (1973) introduce un método para cuantificar variables cualitativas que está muy relacionado con los enfoques anteriores. En los años 60 en Francia un grupos de estadísticos y lingüistas estudian tablas de asociación entre textos literarios y J. P. Benzecri inventa el *análisis de correspondencias* con un enfoque geométrico que generaliza, y establece un marco común, para muchos de los resultados anteriores. Benzecri visita la Universidad de Princeton y los laboratorios Bell donde Carroll y Shepard están desarrollando los métodos de *escalamiento multidimensional* para analizar datos cualitativos, que habían sido iniciados en el campo de la Psicometría por Torgeson (1958). A su vuelta a Francia, Benzecri funda en 1965 el Departamento de Estadística de la Universidad de París y publica en 1972 sus métodos de análisis de datos cualitativos mediante análisis de correspondencias.

La aparición de la computadora transforma radicalmente los métodos de análisis multivariado que experimentan un gran crecimiento desde los años 70. En el campo descriptivo, las computadoras hacen posible la aplicación de métodos de clasificación de observaciones (*análisis de conglomerados o análisis de clusters*) que se basan cada vez más en un uso extensivo de la computadora. MacQueen (1967) introduce el a*lgoritmo de k-medias*. El primer ajuste de una *mezcla de distribuciones* fue realizado por el método de momentos por K. Pearson y el primer algortimo de estimación multivariada se debe a Wolfe (1970). Por otro lado, en el campo de la inferencia, la computadora permite la estimación de modelos sofisticados de mezclas de distribuciones para clasificación, tanto desde el punto de vista clásico, mediante nuevos algoritmos de estimación de variables latentes, como el algoritmo EM, debido a Dempster, Laird y Rubin (1977), como desde el punto de vista Bayesiano, con los métodos modernos de simulación de cadenas de Markov, o métodos MCMC (Markov Chain Monte Carlo).

En los últimos años, los métodos multivariados están sufriendo una transformación en dos direcciones: en primer lugar, las grandes masas de datos disponibles en algunas aplicaciones

están conduciendo al desarrollo de métodos de aproximación local, que no requieren hipótesis generales sobre el conjunto de observaciones. Este enfoque permite construir indicadores no lineales, que resumen la información por segmentos en lugar de intentar una aproximación general. En el análisis de grupos, este enfoque local está obteniendo también ventajas apreciables. La segunda dirección prescinde de las hipótesis sobre las distribuciones de los datos y cuantifica la incertidumbre mediente métodos de computación intensiva. Es de esperase que las crecientes posibilidades de cálculo proporcionadas por las computadoras actuales amplie el campo de aplicación de estos métodos a problemas más complejos y generales.

INTRODUCCIÓN

Los datos multivariados se presentan cuando el investigador recaba varias variables sobre cada "unidad" en su muestra. La mayoría de los conjuntos de datos que se colectan para una investigación son multivariados. Aunque algunas veces tiene sentido estudiar por separado cada una de las variables, en la mayoría de los casos no. En el común de las situaciones, las variables están relacionas de tal manera que si se analizan por separado, no se revela la estructura completa de los datos. En la gran mayoría de los conjuntos de datos multivariados, todas las variables necesitan analizarse de manera simultánea para descubrir patrones y características esenciales de la información que contienen. El análisis multivariado incluye métodos que son totalmente descriptivos y otros que son inferenciales. El objetivo principal es revelar la estructura de los datos, eliminando el "ruido" de los mismos.

Un aspecto muy importante a considerar en los datos multivariados, es que, por lo general, las variables que los componen tienen diferentes escalas de medición, hecho que se debe considerar al momento de realizar el análisis estadístico.

Estructura de los datos multivariados

Matriz de datos

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^{'} \\ \mathbf{x}_2^{'} \\ \vdots \\ \mathbf{x}_n^{'} \end{bmatrix}$$

Donde cada vector $\mathbf{x}_j^{'}$, es un vector columna, $p \times 1$, que representa los valores de las p variables sobre el individuo j. Y x_{jk} es el valor de la k-ésima variable (k=1,2,...,p) del j-ésimo individuo (j=1,2,...,n).

Resumen mediante descripciones numéricas

En una extensión simple de los procesos descriptivos que se realizan con una muestra, podemos hacer los correspondientes resúmenes numéricos para cada una de las variables involucradas en el análisis.

- Resúmenes univariados, respetando la escala de medición de cada variable

- **Vector de medias**

$$\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_p)'$$

con $\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}, \quad k = 1, 2, ..., p.$

- **Matriz de Varianza-Covarianza**

$$\mathbf{S}^2 = \begin{pmatrix} s_{11}^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22}^2 & \cdots & s_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp}^2 \end{pmatrix}$$

con las varianzas muestrales $s_{kk}^2 = \frac{1}{n-1}\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2, k = 1, 2, ..., p,$ y

las covarianzas muestrales $s_{ik} = \frac{1}{n-1}\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k),\ i \neq k = 1, 2, ..., p$

- **Matriz de correlación**

$$\mathbf{R} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix}$$

con las correlaciones muestrales $r_{ik} = \frac{s_{ik}}{s_{ii}s_{kk}}, i \neq k = 1, 2, ..., p$

Algunas características de las correlaciones

- $-1 \leq r_{ik} \leq 1$

- r_{ik} es una medida de la fuerza de la asociación lineal entre las variables involucradas

- r_{ik} es invariante ante cambios de escala

- r_{ik} usualmente se refiere a la correlación de *Pearson*. Para medidas generales de correlación (incluida la no lineal), se pueden utilizar la *tau de Kendall* o *rho de Spearman*.

Representación matricial

- **Media muestral**: $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_p)'$

- **Matriz de varianza-covarianza muestral**: $\mathbf{S} = [s_{ik}]$

- **Matriz de correlación muestral**: $\mathbf{R} = [r_{ij}]$, con $r_{ii} = 1$

ALGUNOS RESULTADOS IMPORTANTES DE ÁLGEBRA LINEAL

Como vimos, la forma de presentar la información propia para un análisis multivariado, es a través de vectores y matrices, por tal razón, en este apartado haremos una breve presentación de algunos de los conceptos de álgebra lineal que son de uso común en el análisis multivariado.

• **Producto interior de dos vectores**. $\mathbf{x}$ y $\mathbf{y} \in \mathbb{R}^p$ se define el producto interior de estos vectores como:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^t\mathbf{y} = \sum_{j=1}^{p} x_j y_j = \mathbf{y}^t\mathbf{x}$$

• **Norma**. $\mathbf{x} \in \mathbb{R}^p$. Se define la norma de un vector como:

$$\|\mathbf{x}\| = \langle \mathbf{x}, \mathbf{x} \rangle^{1/2} = \left(\sum_{j=1}^{p} x_j^2 \right)^{1/2}$$

• **Ortogonalidad**. $\mathbf{x}$ y $\mathbf{y} \in \mathbb{R}^p$, se dice que son ortogonales si su producto interior es cero, i.e., $\mathbf{x}^t\mathbf{y} = 0$. Y son **ortonormales**, si son ortogonales y ambos tienen norma *uno*.

• **Ángulo entre vectores**. $\mathbf{x} \in \mathbb{R}^p$. Se define el ángulo entre estos vectores como:

$$\cos(\theta) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \, \|\mathbf{y}\|}$$

• **Matriz transpuesta**. Se define la transpuesta de una matriz $\mathbf{A}$, como la matriz que tiene como renglones las columnas de $\mathbf{A}$, y la denotaremos por $\mathbf{A}^t$.

• **Matriz simétrica**. Se dice que una matriz $\mathbf{A}$, es simétrica si $a_{ij} = a_{ji} \ \forall i \neq j$.

• **Matriz diagonal**. Se dice que $\mathbf{A}$ es diagonal, si $a_{ij} = 0 \ \forall i \neq j$

• **Matriz ortogonal**. Si $\mathbf{A}$ es una matriz cuadrada, tal que $\mathbf{A}\mathbf{A}^t$=$\mathbf{I}$, se dice que $\mathbf{A}$ es una matriz *ortogonal*, y $\mathbf{A}^t$=$\mathbf{A}^{-1}$

• **Traza de una matriz**. La traza de una matriz es la suma de los elementos de su diagonal.

$$traza(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$$

Propiedades de la traza

i) $traza(\mathbf{AB}) = traza(\mathbf{BA})$

ii) $traza(\mathbf{ABC}) = traza(\mathbf{CAB}) = traza(\mathbf{BCA})$ (Cíclica)

• **Rango de una matriz**. El rango de una matriz $\mathbf{A}$, es el número de renglones o columnas linealmente independientes.

• **Inversa de una matriz**. Si $\mathbf{A}$ es una matriz no singular $p \times p$, existe una única matriz $\mathbf{B}$ tal que $\mathbf{AB}$=$\mathbf{BA}$=$\mathbf{I}$, donde $\mathbf{I}$ es la matriz identidad. Entonces, $\mathbf{B}$ es la inversa de $\mathbf{A}$, y la denotamos por $\mathbf{A}^{-1}$.

Eigenvalores y eigenvectores

Si $\mathbf{A}$ es una matriz cuadrada *pxp*, sus *eigenvalores* (valores característicos, valores propios) son las raíces de la ecuación

$$|\mathbf{A} - \lambda\mathbf{I}| = \mathbf{0}$$

Esta ecuación característica es un polinomio de grado p en λ. Sus raíces, los eigenvalores de $\mathbf{A}$, se denotan por $\lambda_1, \lambda_2, ..., \lambda_p$. Para cada eigenvalor λ_i, hay un correspondiente *eigenvector* $\mathbf{e_i}$, que se encuentra resolviendo la ecuación

$$|\mathbf{A} - \lambda_i\mathbf{I}|\,\mathbf{e_i} = \mathbf{0}$$

Existen muchas soluciones para $\mathbf{e_i}$. Para fines estadísticos, consideraremos un eigenvector con norma uno, i.e., $\|\mathbf{e_i}\| = 1$.

Dos resultados asociados a estos eigenvalores de mucha utilidad en análisis multivariado, son:

i) $traza(\mathbf{A}) = \sum_{\mathbf{i=1}}^{\mathbf{p}} \lambda_\mathbf{i}$

ii) $|\mathbf{A}| = \prod_{i=1}^{p} \lambda_i$ con $|\cdot|$ el determinante de la matriz

Si $\mathbf{A}$ es simétrica

iii) Los eigenvectores de norma uno, asociados a eigenvalores distintos son *ortonormales*

• **Matriz semi definida positiva**. Una matriz $\mathbf{A}$ $p \times p$ es una matriz semi definida positiva si $\mathbf{X}^t\mathbf{A}\mathbf{X} \geq \mathbf{0}$ para todo vector $\mathbf{X}$ de dimensión p.

• **Matriz definida positiva**. Una matriz $\mathbf{A}$ $p \times p$ es una matriz definida positiva si $\mathbf{X}^t\mathbf{A}\mathbf{X} > \mathbf{0}$ para todo vector $\mathbf{X} \neq \mathbf{0}$ de dimensión p.

Resultados importantes asociados a matrices semi y definidas positivas

i) $\mathbf{A}_{pxp}$ simétrica, entonces si $\mathbf{A}$ es simi definida positiva $\Rightarrow \lambda \geq 0$

ii) $\mathbf{A}_{pxp}$ simétrica, entonces si $\mathbf{A}$ es definida positiva $\Rightarrow \lambda > 0$

• **Descomposición espectral**. $\mathbf{A}_{pxp}$ simétrica, entonces su *descomposición espectral* es

$$\mathbf{A} = \lambda_1\mathbf{e_1}\mathbf{e_1}' + \lambda_2\mathbf{e_2}\mathbf{e_2}' + \cdots + \lambda_\mathbf{p}\mathbf{e_p}\mathbf{e_p}'$$

donde $\mathbf{e_i}'\mathbf{e_i} = \mathbf{1}$, $\mathbf{e_i}'\mathbf{e_j} = \mathbf{0}$ $\forall i \neq j$. Las λ_i son los eigenvalores de $\mathbf{A}$ y $\mathbf{e_i}$ son los correspondientes eigenvectores.

De esta descomposición se desprenden varios resultados muy importantes

i) $\mathbf{A} = \sum_{i=1}^{p} \lambda_i\mathbf{e_i}\mathbf{e_i}' = \mathbf{P}\mathbf{\Lambda}\mathbf{P}'$. Donde $\mathbf{P} = [\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_p}]$ es la matriz de eigenvectores y $\mathbf{\Lambda} = \mathbf{diag}\{\lambda_1, \lambda_2, ..., \lambda_\mathbf{p}\}$. Algunas veces se supone $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.

ii) $\mathbf{A}^{-1} = \mathbf{P}\mathbf{\Lambda}^{-\mathbf{1}}\mathbf{P}' = \sum_{\mathbf{i=1}}^{\mathbf{p}} \lambda_\mathbf{i}^{-\mathbf{1}}\mathbf{e_i}\mathbf{e_i}'$

iii) La raíz cuadrada de $\mathbf{A}$ es $\mathbf{A}^{1/2} = \sum_{i=1}^{p} \lambda_i^{1/2}\mathbf{e_i}\mathbf{e_i}' = \mathbf{P}\mathbf{\Lambda}^{\mathbf{1/2}}\mathbf{P}'$

Vectores y Matrices Aleatorias

Definición. $\mathbf{X} = [X_{ij}]$ es una *matriz aleatoria* si X_{ij} es una variable aleatoria

• **Esperanza**: $\mathbb{E}(\mathbf{X}) = [\mathbb{E}(X_{ij})]$

• Si $\mathbf{X}$ y $\mathbf{Y}$ son dos matrices aleatorias, entonces

1.- $\mathbb{E}(\mathbf{X} + \mathbf{Y}) = \mathbb{E}(\mathbf{X}) + \mathbb{E}(\mathbf{Y})$

2.- Si $\mathbf{A}$ y $\mathbf{B}$ son matrices no aleatorias, entonces $\mathbb{E}(\mathbf{A}\ \mathbf{X}\ \mathbf{B}) = \mathbf{A}\mathbb{E}(\mathbf{X})\mathbf{B}$

Vectores aleatorios

Para cada sujeto, podemos definir el vector aleatorio, $\mathbf{X}$, de dimensión p que tiene las mediciones de las p variables del sujeto.

Entonces, $\mathbb{E}(\mathbf{X}) = (\mathbb{E}(X_1), \mathbb{E}(X_2), ..., \mathbb{E}(X_p))' = \underline{\mu} = (\mu_1, \mu_2, ..., \mu_p)'$ y

$$\mathbb{C}ov(\mathbf{X}) = \Sigma = \begin{pmatrix} \mathbb{V}(X_1) & \mathbb{C}ov(X_1, X_2) & \cdots & \mathbb{C}ov(X_1, X_p) \\ \mathbb{C}ov(X_2, X_1) & \mathbb{V}(X_2) & \cdots & \mathbb{C}ov(X_2, X_p) \\ \vdots & \vdots & \cdots & \vdots \\ \mathbb{C}ov(X_p, X_1) & \mathbb{C}ov(X_p, X_2) & \cdots & \mathbb{V}(X_p) \end{pmatrix}$$

Entonces para cualquier vector no aleatorio, $\mathbf{c}$, de dimensión p, $\mathbb{V}(\mathbf{c}'\mathbf{X}) = \mathbf{c}'\underline{\mu}$ y

$\mathbb{V}(\mathbf{c}'\mathbf{X}) = \mathbf{c}'\mathbb{V}(\mathbf{X})\mathbf{c}$. Además $\mathbb{E}(\mathbf{X}\mathbf{X}') = \Sigma + \underline{\mu}\,\underline{\mu}'$

Si $\mathbf{X}$ es un vector de media $\underline{\mu}$. Entonces

$$\mathbb{C}ov(\mathbf{X}) = \mathbb{E}((\mathbf{X} - \underline{\mu})'(\mathbf{X} - \underline{\mu}))$$

Muestras aleatorias

Sea $\mathbf{X}_1, \mathbf{X}_2..., \mathbf{X}_n$ una muestra aleatoria de una distribución conjunta de dimensión p, que tiene media $\underline{\mu}$ y matriz de covarianza Σ. Ojo, aquí se toma una muestra de tamaño n de vectores de dimensión p.

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i, \quad \mathbf{S}_n = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)'$$

Entonces

$$\mathbb{E}(\bar{\mathbf{X}}) = \underline{\mu}, \quad \mathbb{C}ov(\bar{\mathbf{X}}) = \frac{1}{n}\Sigma \quad y \quad \mathbb{E}(\mathbf{S}_n) = \frac{n-1}{n}\Sigma$$

Demostración

$\mathbb{E}(\bar{\mathbf{X}}) = \underline{\mu}$ es trivial. Para $\mathbb{C}ov(\bar{\mathbf{X}})$, tenemos

$$\begin{aligned}(\bar{\mathbf{X}} - \underline{\mu})(\bar{\mathbf{X}} - \underline{\mu})' &= \left[\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \underline{\mu})\right]\left[\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \underline{\mu})\right]' \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{X}_i - \underline{\mu})(\mathbf{X}_j - \underline{\mu})'\end{aligned}$$

Entonces

$$\begin{aligned}\mathbb{C}ov(\bar{\mathbf{X}}) = \mathbb{E}\left[(\bar{\mathbf{X}} - \underline{\mu})(\bar{\mathbf{X}} - \underline{\mu})'\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[(\mathbf{X}_i - \underline{\mu})(\mathbf{X}_j - \underline{\mu})'\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(\mathbf{X}_i - \underline{\mu})(\mathbf{X}_i - \underline{\mu})'\right] \text{ (por independencia)} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\Sigma = \frac{1}{n}\Sigma\end{aligned}$$

Para $\mathbb{E}(\mathbf{S}_n)$, primero observemos que

$$\begin{aligned}\frac{1}{n}\Sigma = \mathbb{C}ov(\bar{\mathbf{X}}) &= \mathbb{E}\left[\left(\bar{\mathbf{X}}-\underline{\mu}\right)\left(\bar{\mathbf{X}}-\underline{\mu}\right)'\right] \\ &= \mathbb{E}\left(\bar{\mathbf{X}}\bar{\mathbf{X}}'\right) - \mathbb{E}\left(\underline{\mu\mu}'\right) \\ &= \mathbb{E}\left(\bar{\mathbf{X}}\bar{\mathbf{X}}'\right) - \underline{\mu\mu}'\end{aligned}$$

Entonces

$$\mathbb{E}\left(\bar{\mathbf{X}}\bar{\mathbf{X}}'\right) = \frac{1}{n}\Sigma + \underline{\mu\mu}'$$

Ahora sí, demostramos la proposición.

$$\begin{aligned}\mathbb{E}(\mathbf{S}_n) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\mathbf{X}_i-\bar{\mathbf{X}})(\mathbf{X}_i-\bar{\mathbf{X}})'\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i' - \mathbf{X}_i\bar{\mathbf{X}}' - \bar{\mathbf{X}}\mathbf{X}_i' + \bar{\mathbf{X}}\bar{\mathbf{X}}'\right] \\ &= \frac{1}{n}\left\{\sum_{i=1}^{n}\mathbb{E}(\mathbf{X}_i\mathbf{X}_i') + \mathbb{E}\left[-\sum_{i=1}^{n}\mathbf{X}_i\bar{\mathbf{X}}' - \sum_{i=1}^{n}\bar{\mathbf{X}}\mathbf{X}_i' + \sum_{i=1}^{n}\bar{\mathbf{X}}\bar{\mathbf{X}}'\right]\right\} \\ &= \frac{1}{n}\left\{\sum_{i=1}^{n}\mathbb{E}(\mathbf{X}_i\mathbf{X}_i') - n\mathbb{E}(\bar{\mathbf{X}}\bar{\mathbf{X}}') - n\mathbb{E}(\bar{\mathbf{X}}\bar{\mathbf{X}}') + n\mathbb{E}(\bar{\mathbf{X}}\bar{\mathbf{X}}')\right\} \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}\mathbb{E}(\mathbf{X}_i\mathbf{X}_i') - n\mathbb{E}(\bar{\mathbf{X}}\bar{\mathbf{X}}')\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\mathbf{X}_i\mathbf{X}_i') - \mathbb{E}(\bar{\mathbf{X}}\bar{\mathbf{X}}') \\ &= \frac{1}{n}\sum_{i=1}^{n}(\Sigma + \underline{\mu\mu}') - (\frac{1}{n}\Sigma + \underline{\mu\mu}') \\ &= \Sigma + \underline{\mu\mu}' - \frac{1}{n}\Sigma - \underline{\mu\mu}' \\ &= \frac{n-1}{n}\Sigma\end{aligned}$$

Similar al caso univariado, $\mathbf{S}_n$ es sesgado, pero $\mathbf{S} = \frac{n}{n-1}\mathbf{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)^{'}$ es un estimador insesgado de Σ.

Función generadora de momentos

La función generadora de momentos (fgm) de $\mathbf{X}$ es una función de $\mathbb{R}^p \to [0, \infty]$, dada por

$$M_{\mathbf{X}}(\mathbf{t}) = M_{\mathbf{X}}(t_1, t_2, ..., t_p) = \mathbb{E}\left[e^{t_1 X_1 + \cdots + t_p X_p}\right]$$

Normal multivariada

Definición: Sea $\mathbf{X} = (X_1, ..., X_p)$ un vector aleatorio de dimensión p. Diremos que $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$ si $\mathbf{X}$ tiene función de densidad de probabilidad

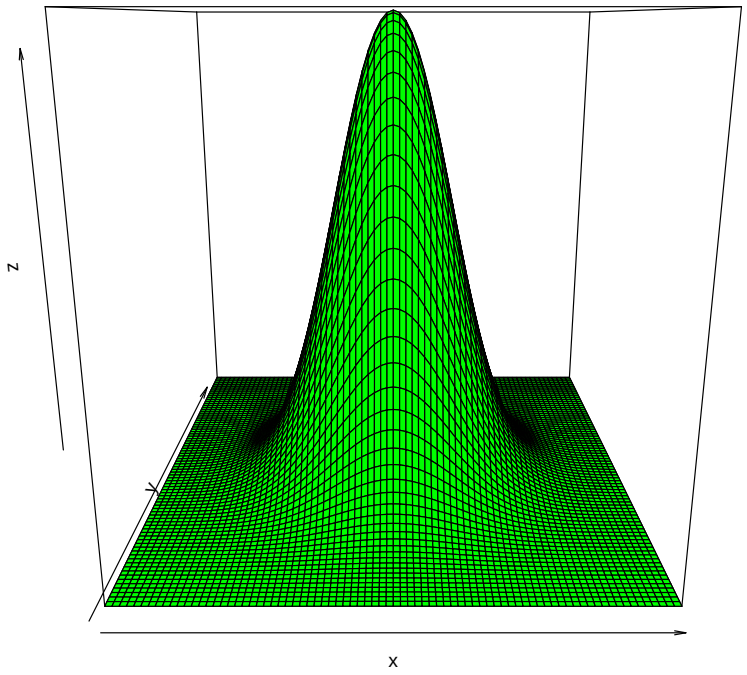
$$f(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} exp\left[-\frac{1}{2}(\mathbf{x} - \mu)^{'}\Sigma^{-1}(\mathbf{x} - \mu)\right]$$

Donde $\mu = (\mu_1, ..., \mu_p)^{'}$ y $\mathbf{\Sigma}$ es una matriz pxp definida positiva.

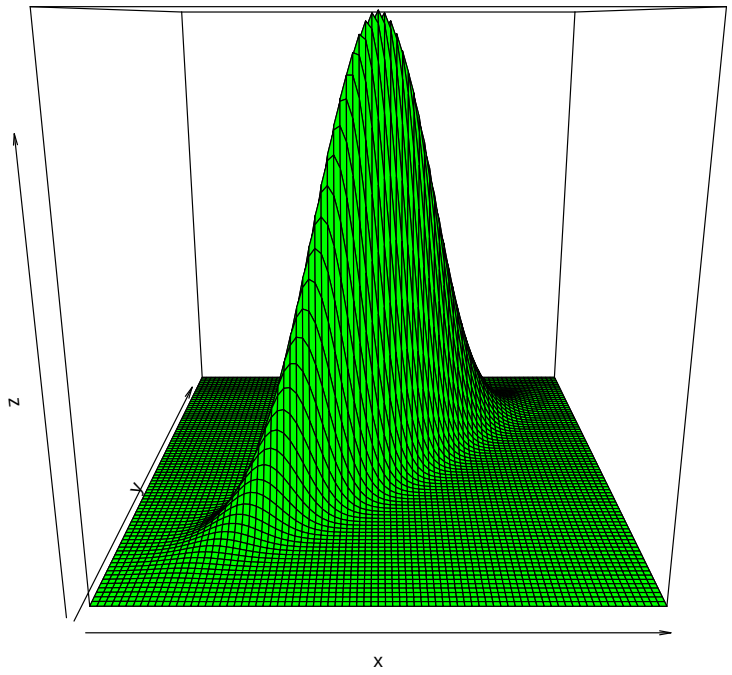
Resultados básicos

- $\mathbb{E}(\mathbf{X}) = \mu$
- $\mathbb{C}ov(\mathbf{X}) = \mathbf{\Sigma}$
- Función característica: $\phi(\mathbf{t}) = \mathbb{E}(e^{i\mathbf{t}^{'}\mathbf{X}}) = exp\left[i\mathbf{t}^{'}\mu - \frac{1}{2}\mathbf{t}^{'}\mathbf{\Sigma}\mathbf{t}\right]$, con $\mathbf{t} = (t_1, ..., t_p)$
- Función generadora de momentos: $\Phi(\mathbf{t}) = exp\left[\mathbf{t}^{'}\mu + \frac{1}{2}\mathbf{t}^{'}\mathbf{\Sigma}\mathbf{t}\right]$

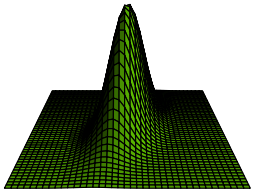
Normal bivariada estándar

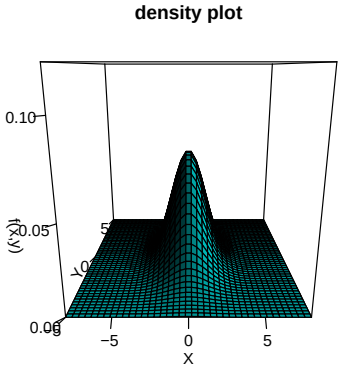


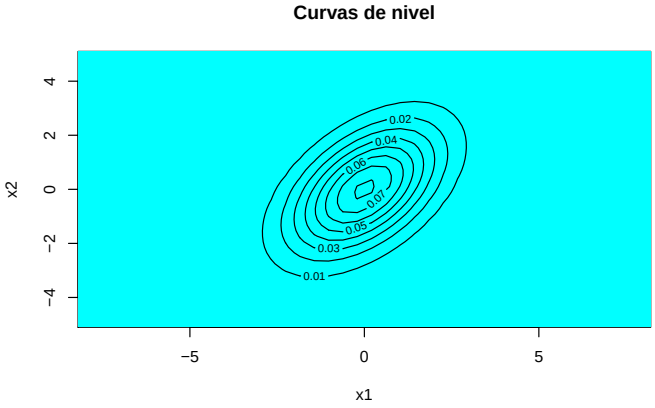
Normal bivariada con correlación=0.9

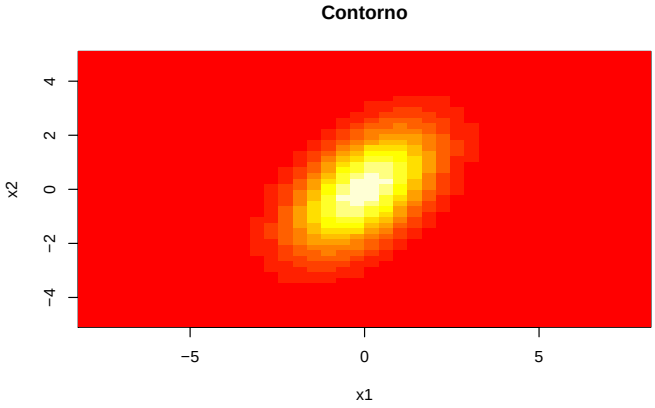


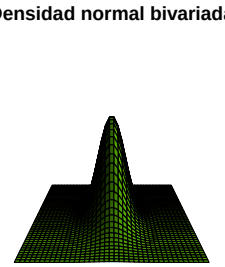
Aspectos de una normal bivariada

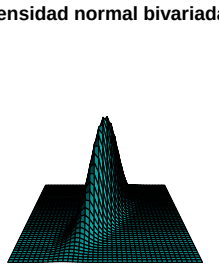


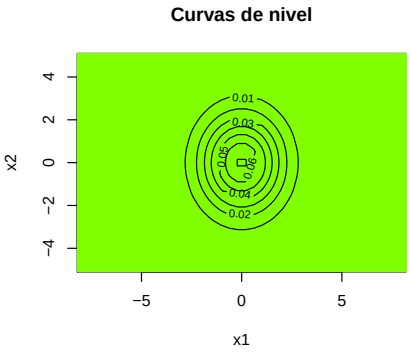


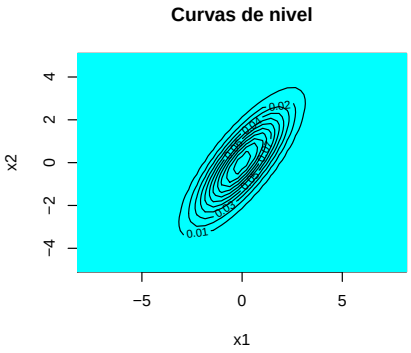





Densidad normal bivariada


Densidad normal bivariada


Curvas de nivel
x2
x1
0.01
0.02
0.03
0.04
0.05
0.06


Curvas de nivel
x2
x1
0.01
0.02

Propiedades importantes de la normal multivariada

Si $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$

• Sea $\mathbf{Y}$=$\mathbf{CX}$ con $\mathbf{C}$ una matriz de de $c \times p$ con $Rango(\mathbf{C}) = k \leq p$. Entonces,
$\mathbf{Y} \sim N_k(\mathbf{C}\mu, \mathbf{C\Sigma C}^{'})$

• Todos los subconjuntos de componentes de $\mathbf{X}$ se distribuyen normal (multivariada). Sea $\mathbf{X} = (\mathbf{X}_1^{'}, \mathbf{X}_2^{'})$, donde $\mathbf{X}_1^{'} = (X_1, ..., X_k)^{'}$ y $\mathbf{X}_2^{'} = (X_{k+1}, ..., X_p)^{'}$, $1 \leq k < p$. Particionando a μ y $\mathbf{\Sigma}$, como

$$\mu = (\mu_1^{'}, \mu_2^{'}), \qquad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}$$

entonces $\mathbf{X}_1 \sim N_k(\mu_1, \mathbf{\Sigma}_{11})$ y $\mathbf{X}_2 \sim N_{p-k}(\mu_2, \mathbf{\Sigma}_{22})$. En particular, cada componente, $\mathbf{X}_i \sim N(\mu_i, \sigma_{ii})$, con σ_{ii} el elemento (i, i) de $\mathbf{\Sigma}$.

• Si $\mathbf{X} = (\mathbf{X}_1^{'}, \mathbf{X}_2^{'})^{'} \sim N_p(\mu, \mathbf{\Sigma})$, entonces, $\mathbf{X}_1$ y $\mathbf{X}_2$ son independendientes si y sólo si $cov(\mathbf{X}_1, \mathbf{X}_2) = 0$.

• Las distribuciones condicionales de los componentes son normales (multivariadas). Nuevamente consideremos la partición anterior. Tenemos

$$\mathbf{X}_1 | \mathbf{X}_2 = \mathbf{x}_2 \sim N_k \left(\mu_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \mu_2), \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} \right)$$

• La forma cuadrática: $(\mathbf{X} - \mu)^{'} \mathbf{\Sigma}^{-1} (\mathbf{X} - \mu) \sim \chi_p^2$

Existen muchos más resultados importantes relacionados con la normal multivariada y también con las distribuciones muestrales de los estimadores de su media y su varianza, pero ya comentamos que difícilmente en *análisis multivariado* se tiene posibilidad de hacer un análisis a nivel inferencial. Esencialmente, el análisis multivariado es *descriptivo*.

Resumen mediante descripciones gráficas

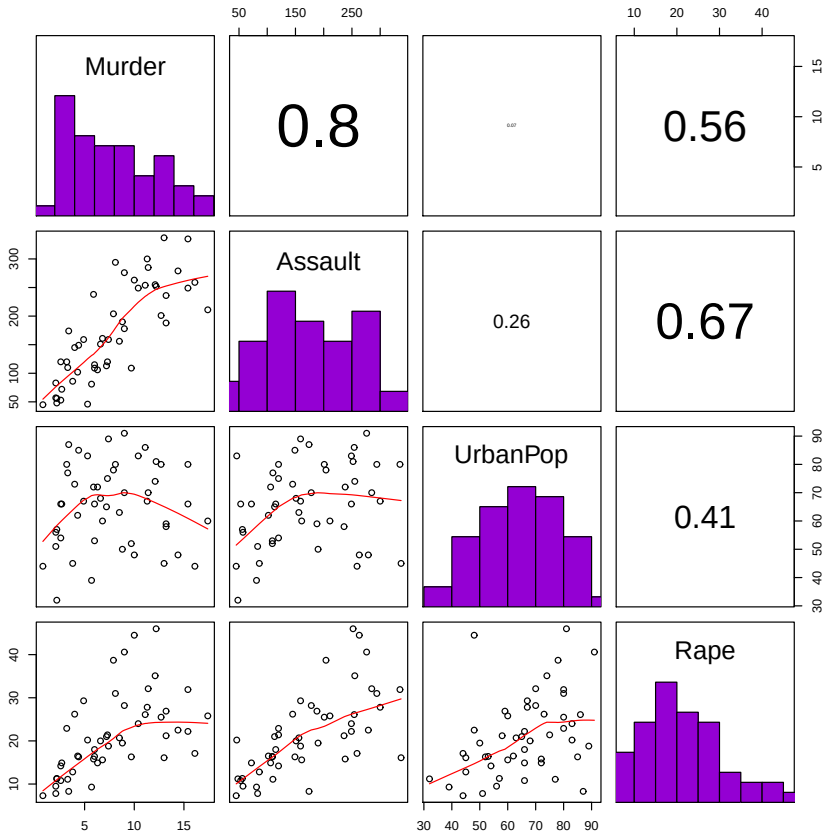
Una manera natural en estadística de mostrar la información contenida en un conjunto de datos, es a través de algunas representaciones gráficas de los mismos. Similar al análisis univariado estándar, se pueden hacer las representaciones gráficas que se considere necesarias, para cada variable. Pero, dada la naturaleza multivarida de nuestros datos, es más conveniente realizar estas representaciones tratando de involucrar a todas las variables de manera simultánea. El problema para graficar datos multivariados, es su dimensión.

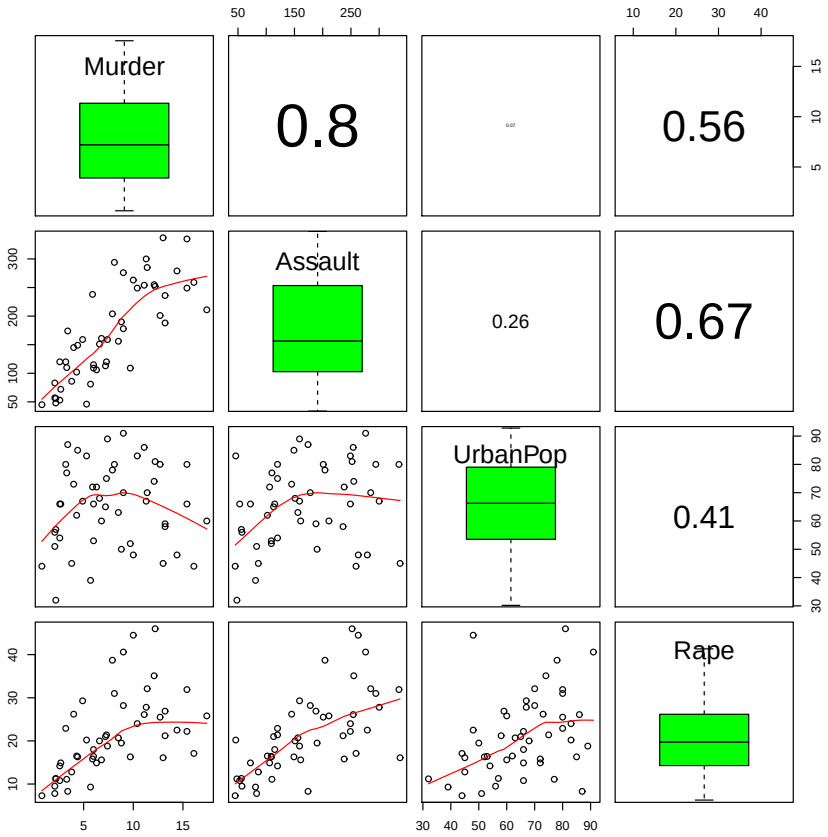
Existen diversas técnicas gráficas para desplegar datos multivariados. La finalidad esencial de éstas es tratar de identificar grupos similiares de sujetos, observaciones atípicas, dispersión de las variables, correlación entre ellas, etc.

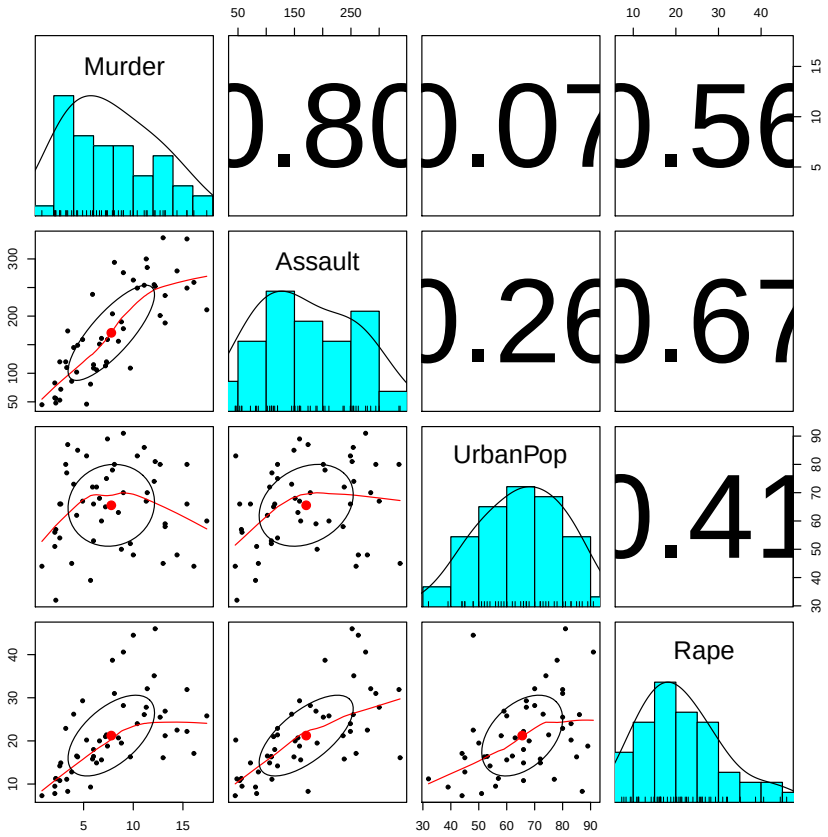
El uso de diagramas y gráficas ahorra tiempo, ya que las características esenciales de grandes volúmenes de datos estadísticos puede apreciarse de un solo vistazo.

Gráfica de la matriz de datos

Una procedimiento útil para iniciar una exploración de las variables en datos multivariados, es desplegar gráficas de dispersión entre pares de variables contenidas en la matriz de datos. Dijimos que para que un análisis multivariado tenga sentido, debemos tener una *fuerte* correlación entre las variables involucradas. Una gráfica que es útil para estos propósitos y que proporciona información adicional, se obtiene con el comando *pairs* de **R**. Los datos pertenecen a la base en **R**, *USArrests* que reporta el número de arrestos por asesinatos (Murder), asaltos (Assault), y violaciones (Rape), además del porcentaje de población urbana (Urban Pop) de los 50 estados que constituyen los Estados Unidos de América


Murder
0.8
0.56
Assault
0.26
0.67
UrbanPop
0.41
Rape

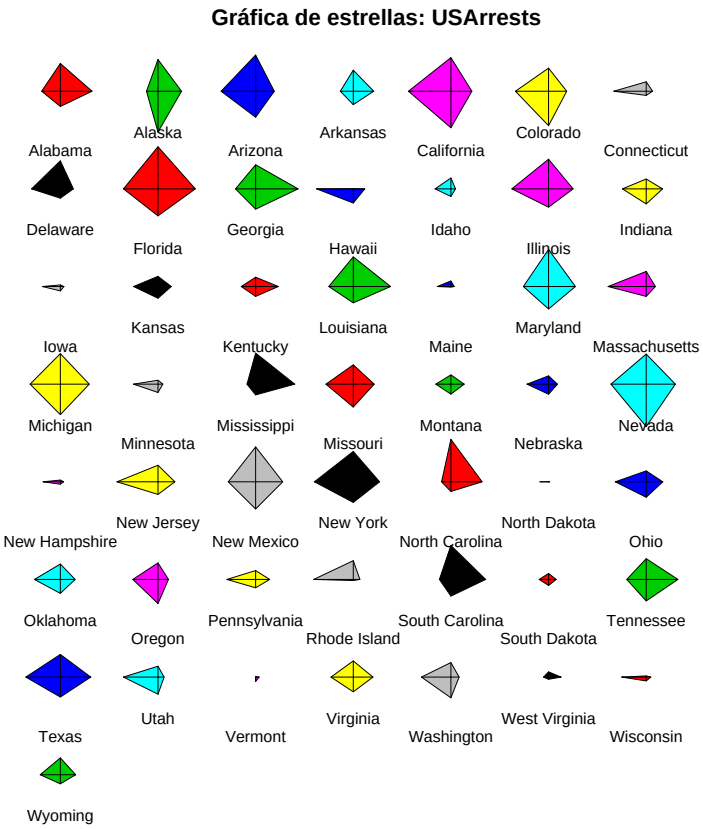

Murder
0.8
0.56
Assault
0.26
0.67
UrbanPop
0.41
Rape


Murder
Assault
UrbanPop
Rape
0.80
0.07
0.56
0.26
0.67
0.41

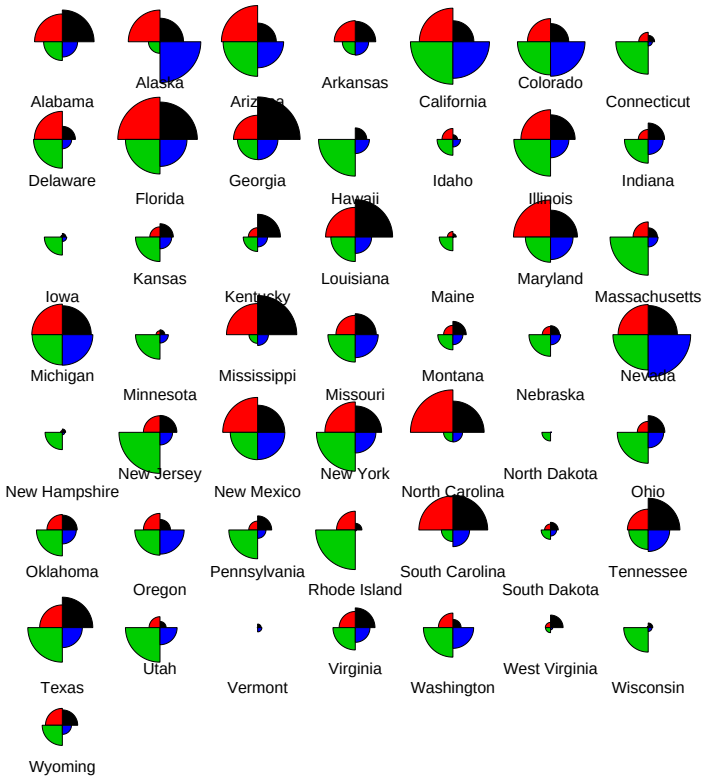
La gráfica anterior presenta características de la forma de la densidad de la variable (histograma y densidad tipo kernel) y de la correlación entre el grupo de variables. Pero no sería útil para descubrir qué estados son similares de acuerdo a este grupo de variables medidas. Para ello, recurriremos a algunas técnicas que intentan resumir todas las variables en una sola gráfica.

Diagramas de estrellas

Cada individuo se representa en una estrella, con tantos rayos o ejes como variables posea su vector de observaciones. Cada eje representa el valor de la variable re-escalada de manera independiente entre variables. Para re-escalar se utilizan todos los datos. En todas las estrellas se usa siempre el mismo eje para representar la misma variable. El eje j en la estrella del individuo i depende de x_{ij} (en valor absoluto o relativo)



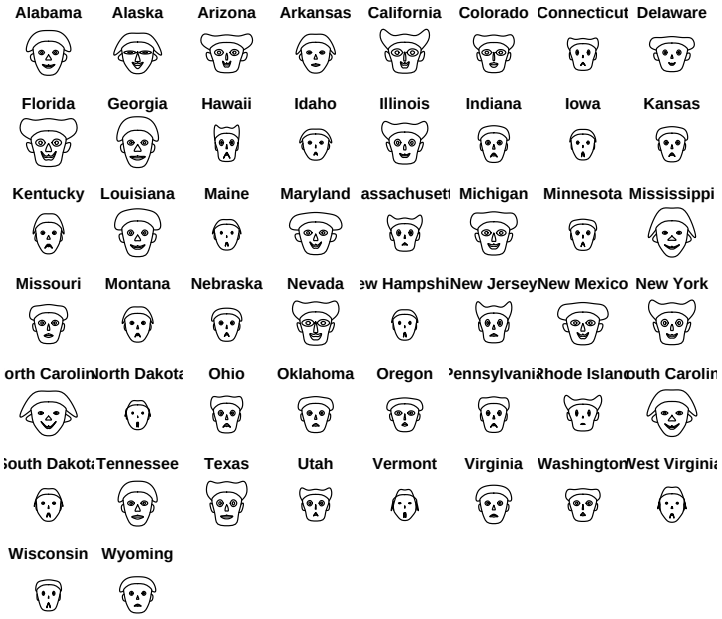
Gráfica de estrellas: USArrests

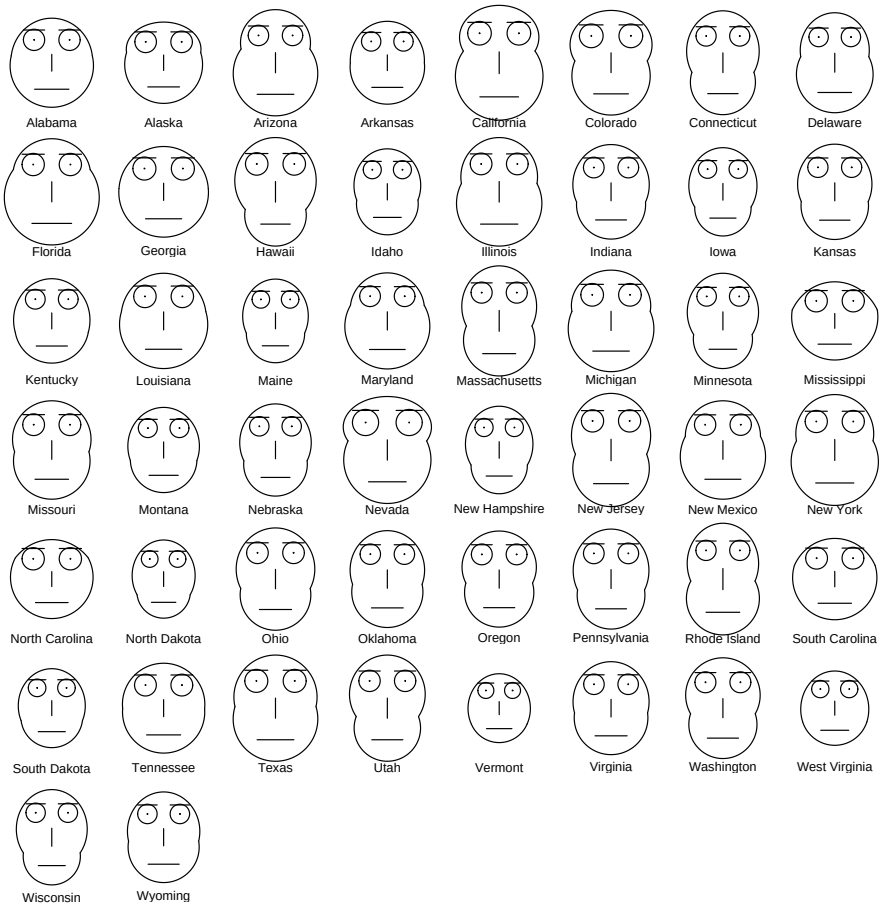


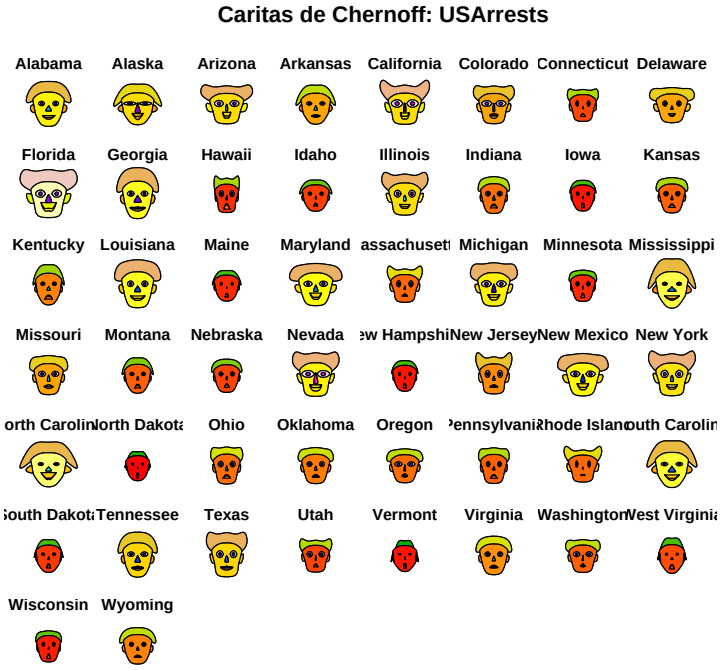
Caritas de Chernoff

El objetivo en esta técnica es asociar el valor de cada variable con alguna característica de una cara humana. Las variables están asociadas con seis aspectos básicos de la carita: *forma de la cara, la boca, la nariz, los ojos, las cejas y las orejas.* Cuando el número de variables es grande, algunas de ellas estarán asociadas con varios aspectos relacionados con los anteriores: *Amplitud de la cara, longitud de las cejas, altura de la cara, separación de los ojos, posición de las pupilas, longitud de la nariz, ancho de la nariz, diámetro de las orejas, nivel de las orejas, longitud de la boca, inclinación de los ojos, altura de las cejas,* etc. Bernard Flury ideó, con base al trabajo de Chernoff, duplicar la cantidad de variables para representar la carita, dejando de lado la simetría, i.e., del lado izquierdo del rostro es posible graficar 18 variables y otras tantas del lado derecho.

Caritas de Chernoff: USArrests




Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming


Caritas de Chernoff: USArrests
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
assachusett
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
ew Hampshi
New Jersey
New Mexico
New York
orth Carolin
orth Dakota
Ohio
Oklahoma
Oregon
ennsylvania
hode Island
outh Carolin
outh Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming

Curvas de Andrew

Supongamos que cada individuo tiene p variables medidas $(X_{i1}, X_{i2}, ..., X_{ip})$. Se define la función

$$f_{X_i} = \frac{X_{i1}}{\sqrt{2}} + X_{i2}sen(t) + X_{i3}cos(t) + X_{i4}sen(2t) + X_{i5}cos(2t) + ... \quad -\pi < t < \pi$$

Algunas propiedades interesantes de estas curvas

i) Preserva medias, i.e.

$$f_{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} f_{X_i}(t)$$

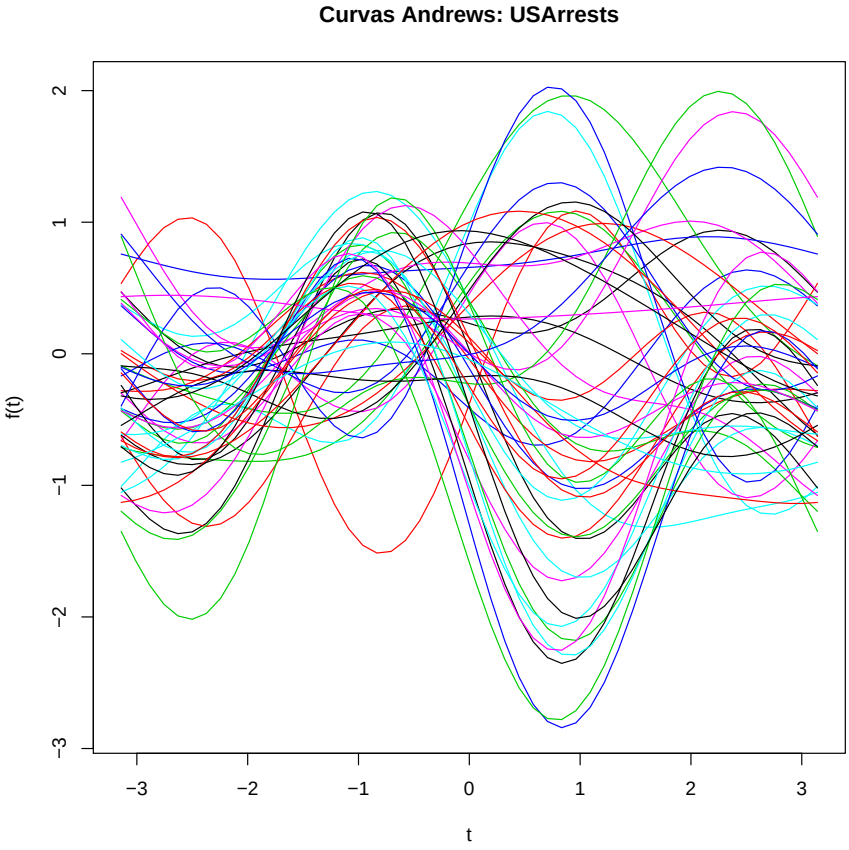
ii) Preserva distancias

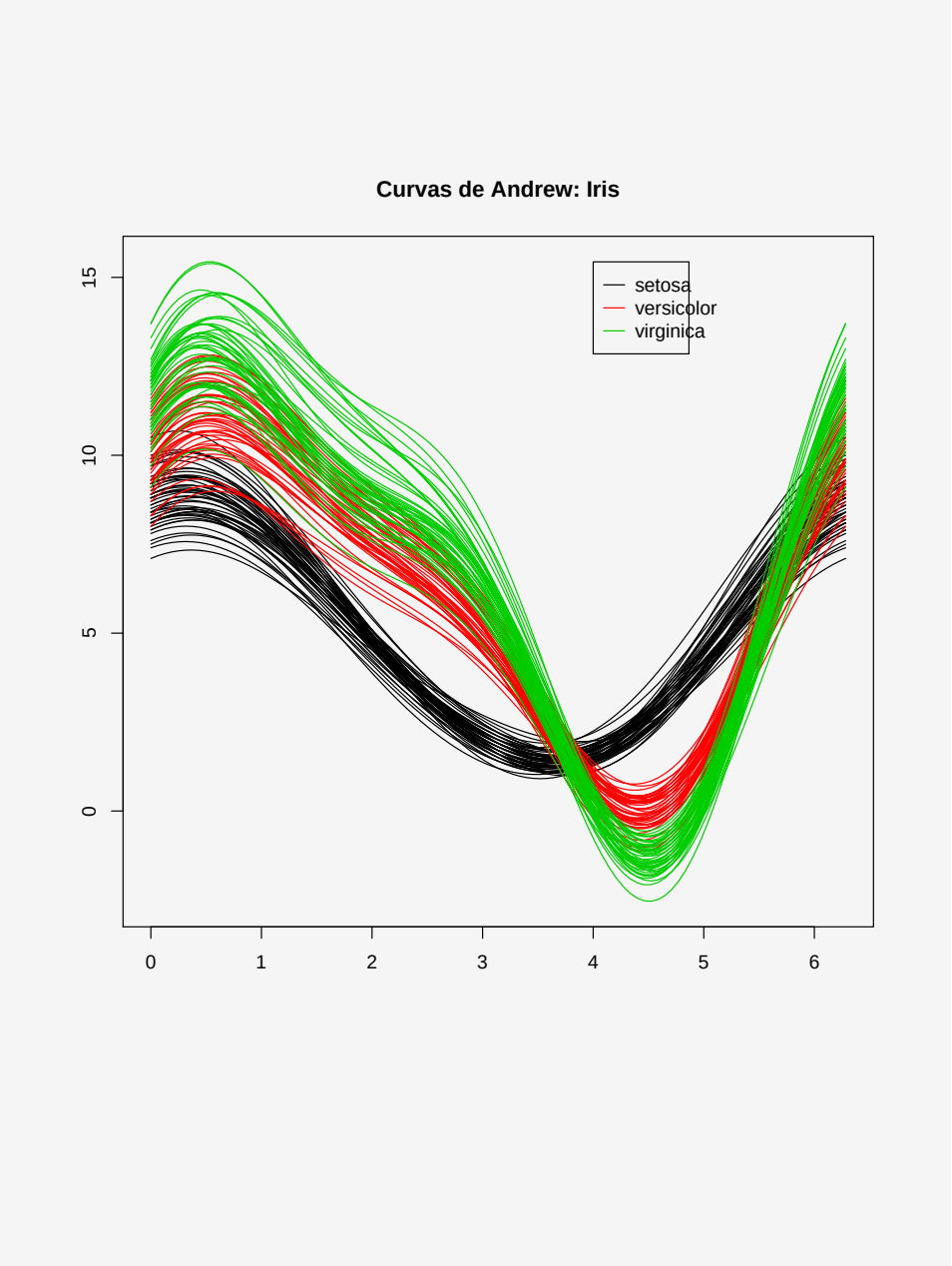
$$\left\| f_{X_i}(t) - f_{X_j}(t) \right\|^2 = \int_{-\pi}^{\pi} \left(f_{X_i}(t) - f_{X_j}(t) \right)^2 dt = \pi \sum_{k=1}^{p} \left(X_{ik} - X_{jk} \right)^2$$

Por lo tanto, si los sujetos X_i, X_j , están cerca, las respectivas curvas lo estarán también.

En esta representación gráfica, el orden de las variables juega un papel importante. Si la dimensión de $\mathbf{X}$ es muy alta, las últimas variables tendrán una contribución pequeña. Por lo que se recomienda ordenar las variables de manera que las variables "más importantes" aparezcan al principio (por ejemplo, aquéllas que discriminan mejor los posibles subgrupos presentes en los datos). También es recomendable no incluir demasiadas observaciones (curvas) en una sola gráfica.

En este tipo de gráficas, las observaciones atípicas aparecen como curvas aisladas que se distinguen claramente de las demás.


Curvas Andrews: USArrests
f(t)
t
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3



Nota: Cada una de estas técnicas se vuelve inadecuada si el número de sujetos es muy grande.

Estas no son las únicas técnicas de represantación gráfica de datos multivariados, existen otras como

- Gráficas de perfiles

- Parallel coordinates plot

TÉCNICAS DE REDUCCIÓN DE DIMENSIÓN

Comentamos al final de la sección anterior que si es muy grande el número de observaciones en nuestro estudio, el despliegue gráfico de estas observaciones, con el fin de encontrar grupos de observaciones semejantes entre ellas, resulta poco útil. Por lo tanto, requerimos de *técnicas esencialmente numéricas* para representar, de preferencia gráficamente, nuestras observaciones y que nos permitan visualizar los grupos que subyacen en ellas.

ANÁLISIS DE COMPONENTES PRINCIPALES

INTRODUCCIÓN

El objetivo principal de la mayoría de las técnicas numéricas de análisis multivariado, es reducir la dimensión de nuestros datos. Por supuesto, si esta reducción se puede hacer a 2 ó 3 dimensiones, se tiene la posibilidad de una visión gráfica de los mismos. Obvio, *siempre* es posible hacer la reducción a este número de dimensiones, pero es importante juzgar si estas pocas dimensiones son suficientes para resumir la información contenida en todas las variables.

El *análisis de componentes principales* tiene este objetivo: dadas n observaciones de p variables, se analiza si es posible representar adecuadamente esta información con un número menor ($q << p$) de *variables construidas como combinaciones lineales de las originales*, llamadas *componentes principales*. Esta técnica se debe a Hotelling (1933), aunque sus orígenes se encuentran en los ajustes ortogonales por mínimos cuadrados introducidos por K. Pearson (1901).
En concreto, los objetivos del análisis de componentes principales son:

- Reducir la dimensión de los datos ($q << p$)

- Generar nuevas variables: Componentes principales

¿Para qué?

- Explorar datos multivariados

• Encontrar agrupaciones

• Encontrar datos atípicos

• Como auxiliar para combatir la multicolinealidad en los modelos de regresión

¿Qué hace?

Forma nuevas variables llamadas *Componentes Principales* (c.p.) con las siguientes características:

1) No están correlacionadas (bajo el supuesto de distribución normal, son independientes)

2) La primera c.p. explica la mayor cantidad de varianza de los datos, que sea posible

3) Cada componente subsecuente explica la mayor cantidad de la variabilidad restante de los datos, que sea posible.

Las componentes son de la forma:

$$\begin{aligned} Z_i &= a_i^{'}X = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p \quad i = 1, 2, ..., p \quad \text{ó} \\ Z_i &= a_i^{'}(X - \mu)\ (centradas) \end{aligned}$$

Es decir, son combinaciones lineales de las p variables.

Para la primer componente, el objetivo es construir esta combinación lineal, de tal manera que la varianza de ella sea máxima. Por supuesto, suena a resolver un problema de maximización. Entonces, el problema consiste en encontrar el vector a_1, que haga máxima la varianza de esta primer componente. Para garantizar la unicidad de la solución, forzaremos el procedimiento a que a_1 sea de *norma uno* ($\|a_1\| = 1$).
En concreto, debe elegirse a_1, un vector de norma uno, $\|a_1\| = a_1^{'}a_1 = 1$, de tal manera que:

$$\mathbb{V}(Z_1) = \mathbb{V}(a_1{'}X) = a_1^{'}\mathbb{V}(X)a_1 = a_1'\Sigma a_1 \quad \text{sea máxima}$$

Bajo esta restricción, el problema se transforma a encontrar un máximo con restricciones, para lo que utilizaremos la técnica de los *multiplicadores de Lagrange*.

Deducción de la construcción de la primer componente

El problema se plantea de la siguiente manera. Maximizar

$$F(a) = \mathbb{V}(Z) = \mathbb{V}\left(a'X\right) = a'\mathbb{V}(X)\,a = a'\Sigma a$$
$$\text{s.a } \lambda \|a\| = \lambda a'a = 1$$

Que genera la función

$$F(a) = a'\Sigma a - \left(\lambda a'a - 1\right)$$

Derivando respecto al vector a, obtenemos

$$\frac{\partial F(a)}{\partial a} = 2\Sigma a - 2\lambda a = 0$$

cuya solución está dada por la igualdad

$$\Sigma a = \lambda a$$

que, como vimos en el repaso de los conceptos de álgebra lineal, implica que a es un eigenvector de la matriz Σ y λ el eigenvalor correspondiente a este eigenvector.

Para determinar cuál valor propio de Σ es el que corresponde a la solución de la ecuación anterior, multipliquemos por la izquierda por a', dicha ecuación

$$a'\Sigma a = \lambda a'a \Rightarrow a'\Sigma a = \lambda$$

y observamos, entonces, que $\mathbb{V}(Z) = \lambda$, y como esta cantidad es la que deseamos maximizar, entonces λ es el eigenvalor más grande de la matriz Σ con a el eigenvector asociado a este eigenvalor, llamémoslos λ_1 y a_1, respectivamente.

La siguiente componente debe cumplir con las condiciones de tener la mayor varianza del remanente, una vez calculada la primera, y no estar correlacionada con ésta. Obsérvese que esta última condición se obtiene si los correspondientes vectores, digamos a_1 y a_2 son ortogonales, y como pediremos que a_2 sea también de norma uno, entonces serán ortonormales.

Una manera de garantizar que esta segunda componente es la de mayor varianza posible, después de la primera, es que la suma de estas dos varianzas sea máxima. Entonces el problema se puede plantear de la siguiente manera. Maximizar

$$F(a_1, a_2) = a_1^{'}\Sigma a_1 + a_2^{'}\Sigma a_2$$
$$\text{s.a } \lambda_1 a_1^{'} a_2 = 1 \ , \ \lambda_2 a_2^{'} a_2 = 1 \ y \ \mu a_1^{'} a_2 = 0$$

Derivando esta función respecto a los vectores a_1 y a_2, tenemos

$$\begin{aligned} \frac{\partial F(a_1, a_2)}{\partial a_1} &= 2\Sigma a_1 - 2\lambda_1 a_1 + \mu a_2 = 0 \\ \frac{\partial F(a_1, a_2)}{\partial a_2} &= 2\Sigma a_2 - 2\lambda_2 a_2 + \mu a_1 = 0 \end{aligned}$$

Multiplicando la parcial respecto a a_1 por $a_1^{'}$ por la izquierda y recordando que $a_1^{'} a_2 = 0$, porque son ortonormales, tenemos

$$a_1^{'}\Sigma a_1 = \lambda_1 \Rightarrow a_1 a_1^{'}\Sigma a_1 = \lambda_1 a_1 \Rightarrow \Sigma a_1 = \lambda_1 a_1$$

De manera similar, multiplicando la parcial respecto a a_2 por $a_2^{'}$ por la izquierda y recordando que $a_2^{'} a_1 = 0$, porque son ortonormales, tenemos

$$a_2^{'}\Sigma a_2 = \lambda_2 \Rightarrow a_2 a_2^{'}\Sigma a_2 = \lambda_2 a_2 \Rightarrow \Sigma a_2 = \lambda_2 a_2$$

que implica que a_1 y a_2 deben ser eigenvectores de Σ. Tomando estos vectores propios de norma uno y sustituyendo en la función objetivo, obtenemos

$$\lambda_1 a_1^{'} a_1 + \lambda_2 a_2^{'} a_2 - \lambda_1 \left(a_1^{'} a_1 - 1\right) - \lambda_2 \left(a_2^{'} a_2 - 1\right) - \mu a_1^{'} a_2 = \lambda_1 + \lambda_2$$

Por lo que es claro que λ_1 y λ_2 deben ser los dos eigenvalores más grandes de la matriz Σ y a_1 y a_2 sus correspondientes eigenvectores.

De manera general, la *j-ésima* componente principal será

$Z_j = a_j' X \;\; j = 1, 2, ..., p$ con a_j el eigenvector de la matriz Σ asociado al eingevalor λ_j

y $\lambda_1 > \lambda_2 > \cdots > \lambda_p$.

Propiedades de los componentes principales

Los componentes principales como variables derivadas de las originales, tienen las siguientes propiedades:

• *Conservan la variabilidad original de los datos*: En el sentido de que la suma de las varianzas de los componentes es igual a la suma de las varianzas de las variables originales.

Por construcción tenemos que

$$\mathbb{V}(Z_1) = \lambda_1, \; \mathbb{V}(Z_2) = \lambda_2, \text{ etc.}$$

y además se tiene también que $\mathbb{C}ov(Z_1, Z_2) = 0$. En general $\mathbb{C}ov(Z_i, Z_j) = 0$ para toda $i \neq j \;\; i, j = 1, 2, ..., p$. Entonces

$$traza(\Sigma) = \sum_{i=1}^{p} \mathbb{V}(X_i) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} Var(Z_i)$$

Las nuevas variables Z_i tienen conjuntamente la misma variabilidad que las variables originales, la suma de varianzas es la misma, pero su estructura o constitución es muy diferente.

• La proporción de la varianza total explicada por una componente, es el cociente entre su varianza (el valor propio asociado al vector propio que la define), y la suma de los valores propios de la matriz. Por esta razón se dice que el i-ésimo componente principal explica una proporción de varianza igual a:

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

y los primeros r de ellos

$$\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \quad r \leq p$$

• Las covarianza entre el vector de variables originales X y la i-ésima componente principal Z_i, es:

$$\mathbb{C}ov(X, Z_i) = Cov\left(X, a_i'X\right) = a_i'\mathbb{C}ov(X, X) = a_i'\Sigma = a_i'\lambda_i = \Sigma a_i = \lambda_i a_i \quad i = 1, 2, ..., p$$

Es decir

$$\mathbb{C}ov(X, Z_i) = \mathbb{C}ov\left(X_1, X_2, ..., X_p, Z_i\right) = \lambda_i a_i = \lambda_i\left(a_{i1}, a_{i2}, ..., a_{ip}\right)$$

Entonces, la covarianza entre la i-ésima componente y la j-ésima variable es:

$$\mathbb{C}ov\left(X_j, Z_i\right) = \lambda_i a_{ij}$$

Como $\mathbb{V}(X_j) = \sigma_{jj}^2$ y $\mathbb{V}(Z_i) = \lambda_i$, entonces tenemos que:

$$\mathbb{C}or(X_j, Z_i) = \frac{\mathbb{C}ov\left(X_j, Z_i\right)}{\sqrt{\mathbb{V}\left(X_j\right)\mathbb{V}\left(Z_i\right)}} = \frac{\lambda_i a_{ij}}{\sqrt{\sigma_{jj}^2\lambda_i}} = \frac{\sqrt{\lambda_i}a_{ij}}{\sigma_{jj}}$$

El peso que tiene la variable j en la componente i, está dado por a_{ij}. El tamaño relativo de las $a_{ij}'s$ reflejan la contribución relativa de cada variable en la componente. Para interpretar, en el contexto de los datos, una componente, debemos analizar el patrón de las a_{ij} de cada componente.

Si utilizamos la matriz de correlación para realizar el análisis de *c.p.*, como $\sigma_{jj}^2 = 1$, entonces

$$a_{ij}^* = \sqrt{\lambda_j}a_{ij}$$

se interpreta como el coeficiente de correlación entre la variable j y el componente i. Esta es una de las interpretaciones particularmente más usuales.

Análisis de la matriz de componentes principales

Denotemos por $\mathbf{Z}$ a la matriz de componentes principales, entonces

$$\mathbf{Z} = \mathbf{XA}$$

con

$$\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_p) = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{p1} \\ a_{12} & a_{22} & \cdots & a_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{pp} \end{pmatrix}$$

Propiedades de $\mathbf{A}$.

- En la matriz $\mathbf{A}$, cada columna es un vector propio de Σ.

- $\mathbf{A}'\mathbf{A} = \mathbf{A}\mathbf{A}' = \mathbf{I}_p \Rightarrow \mathbf{A}' = \mathbf{A}^{-1} \Rightarrow \mathbf{A}$ es ortogonal

- $\Sigma\mathbf{A} = \mathbf{A}\Lambda$ con $\Lambda = diag\,(\lambda_1, ..., \lambda_p)\,,\ \lambda_1 > \lambda_2 > \cdots > \lambda_p$ (resultado análogo a $\Sigma\mathbf{a}_i = \lambda_i\mathbf{a}_i$)

Estructura de correlación

- $\mathbb{V}(\mathbf{Z}) = \mathbb{V}(\mathbf{XA}) = \mathbf{A}'\mathbb{V}(X)\mathbf{A} = \mathbf{A}'\Sigma\mathbf{A} = \mathbf{A}'\mathbf{A}\Lambda = \Lambda = diag\,(\lambda_1, ..., \lambda_p)$. Entonces $Cov(\mathbf{Z}_i, \mathbf{Z}_j) = 0,$ si $i \neq j$ y $Var(\mathbf{Z}_i) = \lambda_i \geq Var(\mathbf{Z}_j) = \lambda_j$ si $i \leq j$

Además

$$traza(\Sigma) = traza(\Sigma\mathbf{A}\mathbf{A}') = traza(\mathbf{A}\Lambda\mathbf{A}') = traza(\mathbf{A}'\mathbf{A}\Lambda) = traza(\Lambda) = \sum_{j=1}^{p} \lambda_j.$$

Ya que $traza(\Sigma) = \sum_{j=1}^{p} \sigma_{jj}^2$. Entonces

$\sum_{j=1}^{p} \lambda_j$ es una medida de la variación total de los datos (variación total de $\mathbf{X}$)

Componentes muestrales

Como sabemos, Σ es desconocida, pero podemos estimarla con $\mathbf{S}$ la matriz de varianza-covarianza muestral, que es un estimador con muy buenas propiedades estadísticas. Entonces, con datos reales, el análisis de componentes principales se realiza con esta matriz y se obtienen los estimadores

$$\hat{\lambda}_i \quad y \quad \hat{a}_i$$

¿Matriz de varianza-covarianza o de correlación?

¿Cuándo una, cuándo otra?

Varianza-covarianza

- Variables medidas en las mismas unidades o, por lo menos, en unidades comparables
- Varianzas de tamaño semejante.

Si las variables no están medidas en las mismas unidades, entonces cualquier cambio en la escala de medición en una o más variables tendrá un efecto sobre las *c.p.* Por ejemplo, supongamos que una variable que se midió originalmente en pies, se cambió a pulgadas. Esto significa que la varianza de la variable se incrementará en $12^2 = 144$. Ya que *c.p.* se basa en la varianza, esta variable tendría una mayor influencia sobre los *c.p.* cuando se mide en pulgadas que en pies.

Si una variable tiene una varianza mucho mayor que las demás, dominará el primer componente principal, sin importar la estructura de covarianza de las variables.

Si no se tienen las condiciones para realizar un análisis de *c.p.* con la matriz de varianza-covarianza, se recomienda hacerlo con la matriz de correlación.

Aplicar análisis de *c.p.* a la matriz de correlación, es equivalente a aplicarlo a datos estandarizados ("puntajes z"), en lugar de los datos crudos. Realizar el análisis de *c.p.* con la matriz de correlación, implica intrínsecamente asumir que todas las variables tienen igual importancia dentro del análisis, supuesto que no siempre puede ser cierto.

Pueden presentarse situaciones en donde las variables no estén en unidades comparables y en las que el investigador considere que tienen una importancia distinta. Algunos paquetes

estadísticos permiten asignar pesos a las variables. Entonces se procedería a estandarizar las variables y posteriormente asignar pesos mayores a aquéllas que el investigador considere más importantes.

Análisis de *c.p.* con la matriz de correlación

Estandarizar los datos, hacer análisis de *c.p.* utilizando la matriz de correlación en lugar de la de varianza-covarianza.

Importante: El análisis de *c.p.* transforma un conjunto de variables correlacionadas en un nuevo conjunto de variables no correlacionadas. Si las variables originales no están correlacionadas o están muy poco correlacionadas esta técnica no tiene ninguna utilidad y la dimensión real de los datos es la misma que el número de variables medidas.

¿Cómo decidir cuántas componentes es apropiado considerar?

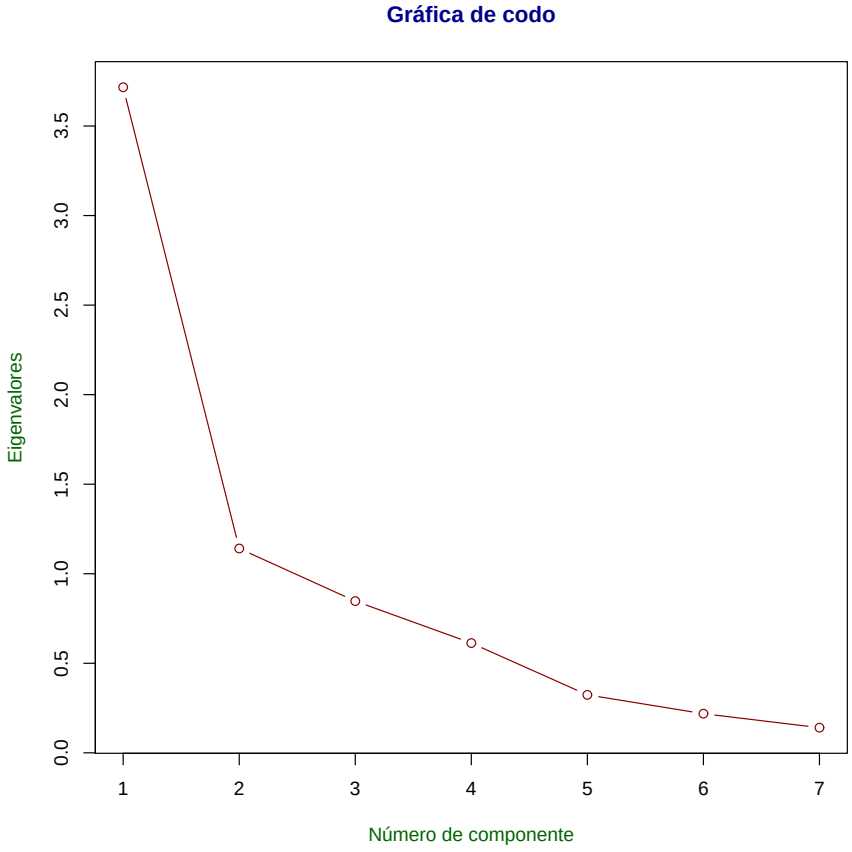
- Porcentaje de varianza explicada requerido (Matriz de varianza-covarianza)

- Porcentaje requerido $\gamma * 100\%$ de la variabilidad total.

Encontrar el número de componentes que cubra este requerimiento. Este criterio depende de la población bajo estudio y del investigador.

Gráfica de codo (SCREE). Cuando los puntos en la gráfica tienden a nivelarse (horizontalmente), los eigenvalores están lo suficientemente cercanos a cero y pueden ignorarse. Entonces, elegir el número de componentes igual al número de eigenvalores antes de que la gráfica se nivele.

Desafortunadamente, mientras más componentes se requiere, menos útiles resultan cada una.


Gráfica de codo
Eigenvalores
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1
2
3
4
5
6
7
Número de componente

Matriz de correlación

- Los criterios mostrados para la matriz de varianza-covarianza.
- Uno más. Considerar el número de componentes cuyo eigenvalor sea mayor que uno.

Puntajes factoriales

Dado que se han generado p componentes principales a partir de las p variables originales, es claro que cada uno de los individuos en nuestra matriz de información, tiene asociados *un valor por cada componente principal*, mismo que se calcula de la siguiente manera

$$\mathbf{Z}_i = \mathbf{A}^{'}\mathbf{X}_i, \quad i = 1, 2, ..., p$$

que proporcionan las coordenadas de la observación $\mathbf{X}_i$ en el nuevo sistema de ejes generado por las *c.p.*

$$z_{ij} = \mathbf{a}_j^{'}\mathbf{X}_i = \sum_{k=1}^{p} a_{jk}x_{ik}$$

es el valor de la j-ésima componente para el i-ésimo individuo.

Entonces, podemos representar un individuo en el plano, mediante la pareja (z_{i1}, z_{i2}).

Ya que uno de los usos comunes de esta técnica es identificar individuos similares, es importante tener en cuenta que *las c.p. preservan la distancia entre las observaciones*, como mostraremos en seguida.

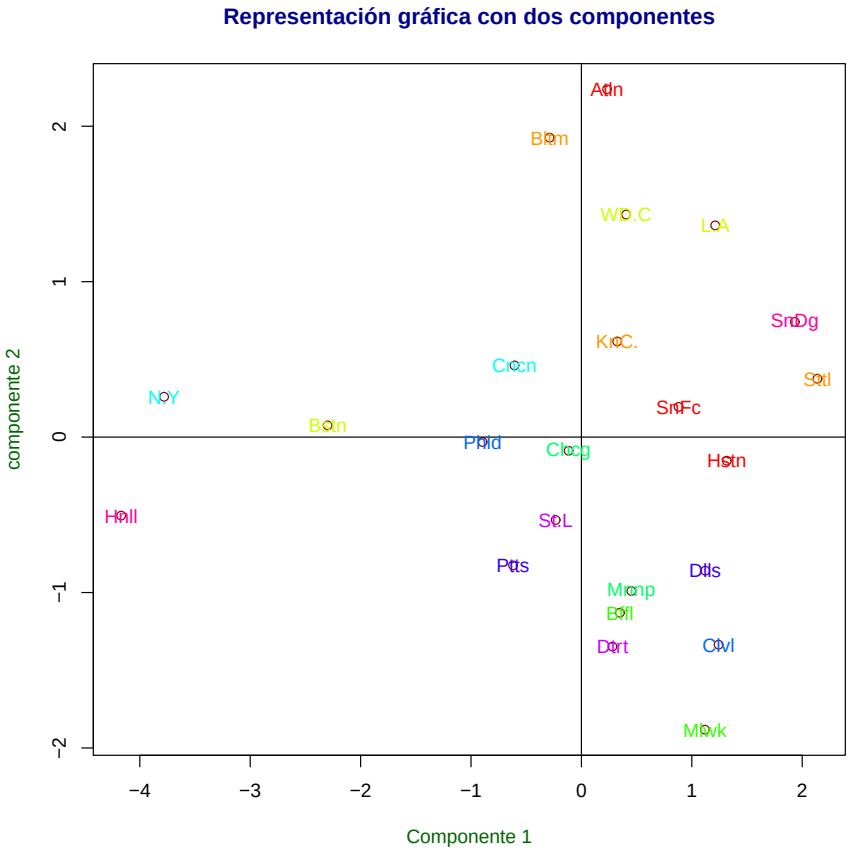
Denotemos por $\mathbf{Z}_i$: Vector de c.p. del individuo $\mathbf{X}_i$ y por $\mathbf{Z}_j$: Vector de c.p. del individuo $\mathbf{X}_j$. Entonces, se trata de mostrar que la distancia entre estas componentes es igual a la distancia entre los vectores originales de los sujetos.

$$
\begin{aligned}
\|\mathbf{Z}_i - \mathbf{Z}_j\|^2 &= (\mathbf{Z}_i - \mathbf{Z}_j)' (\mathbf{Z}_i - \mathbf{Z}_j) \\
&= \left(\mathbf{A}'\mathbf{X}_i - \mathbf{A}'\mathbf{X}_j\right)' \left(\mathbf{A}'\mathbf{X}_i - \mathbf{A}'\mathbf{X}_j\right) \\
&= \left(\mathbf{A}' (\mathbf{X}_i - \mathbf{X}_j)\right)' \left(\mathbf{A}' (\mathbf{X}_i - \mathbf{X}_j)\right) \\
&= (\mathbf{X}_i - \mathbf{X}_j)' \mathbf{A}\mathbf{A}' (\mathbf{X}_i - \mathbf{X}_j) \\
&= (\mathbf{X}_i - \mathbf{X}_j)' \mathbf{A}\mathbf{A}^{-1} (\mathbf{X}_i - \mathbf{X}_j) \quad (\mathbf{A} \text{ es ortogonal}) \\
&= (\mathbf{X}_i - \mathbf{X}_j)' \mathbf{I}_p (\mathbf{X}_i - \mathbf{X}_j) \\
&= (\mathbf{X}_i - \mathbf{X}_j)' (\mathbf{X}_i - \mathbf{X}_j) \\
&= \|\mathbf{X}_i - \mathbf{X}_j\|^2
\end{aligned}
$$

Observación. Esta distancia se conserva en el espacio original de los vectrores, que es de diminsión p. Si sólo tomamos pocas componentes (2 ó 3) para representar las observaciones, entonces

$$\|\mathbf{X}_i - \mathbf{X}_j\|^2 \approx \left\|\mathbf{Z}_i^* - \mathbf{Z}_j^*\right\|^2$$

con $\mathbf{Z}^*$ un vector de dimensión 2 ó 3, únicamente. Esta aproximación será adecuada si estas pocas dimensiones explican un alto porcentaje de la varianza total de los datos.


Representación gráfica con dos componentes
componente 2
Componente 1
2
1
0
−1
−2
−4
−3
−2
−1
0
1
2

Aplicación de c.p. con variables medidas en diversas escalas

El análisis de c.p. se realiza, generalmente, utilizando variables continuas; no obstante, existen aplicaciones donde se presentan diversas escalas de medición en las variables. Una manera generaliza de abordar esta situación, es realizar el análisis ignorando la escala de medición, i.e., suponiendo que todas provienen de una escala de intervalo. En este caso, la correlación entre cualquier par de variables, es la de Pearson. El hecho de no respetar la escala de cada variable, propicia que las correlaciones sean más pequeñas de lo debido, lo que, para una técnica basada en la asociación entre las variables, resulta poco deseable. Otra alternativa es construir variables *dummy's* con las variables medidas en escalas nominal y ordinal. Este procedimiento tiene la desventaja de incrementar el número de variables dentro del análisis (hay que recordar que si una variable nominal u ordinal tiene k categorías, entonces genera un número igual de variables dummy's). Este incremento de dimensión repercutirá en el hecho de que tendremos menos posibilidades de poder representar nuestros datos en pocas dimensiones, i.e., tendremos poca varianza explicada por unas cuantas dimensiones.

Una forma alternativa de enfrentar este problema, es utilizando la *matriz de correlaciones policóricas*. En esta matriz se utiliza un tipo de correlación de acuerdo a la escala de medición de las dos variables en cuestión. La siguiente tabla muestra las correlaciones que se sugiere calcular.

Escala de medición	Continua	Ordinal	Dicotómica
Continua	Pearson	Policórica	Punto biserial
Ordinal		Policórica	Policórica
Dicotómica			Tetracórica

Una vez calculada esta matriz, el análisis de c.p. se lleva a cabo utilizándola para realizar todos los procesos de cálculo.

BIPLOTS

Podemos dividir el análisis de datos multivariados en un análisis que se centre en la estructura de asociación entre las variables, y uno basado en las relaciones entre las observaciones (los sujetos). Es deseable tener una técnica que nos permita mostrar las relaciones entre las variables, entre los sujetos y entre ambos. El *biplot* es una representación bidimensional de la matriz de datos $\mathbf{X}$ en la que tanto los renglones (sujetos) como las columnas (variables) se representan a través de puntos. La representación se basa en la *descomposición en valor singular* de la matriz de datos.

Descomposición en valor singular

Sea $\mathbf{X}_{n\times p}$ una matriz. Mostraremos que se puede escribir como el producto de una matriz de columnas ortogonales $(n \times n)$, una matriz diagonal $(n \times p)$ con elementos no negativos y una matriz ortogonal $(p \times p)$. En concreto, la descomposición en valor singular es

$$\mathbf{X}_{n\times p} = \mathbf{U}_{n\times n}\Sigma_{n\times p}\mathbf{V}'_{p\times p}$$

Además

- $\mathbf{U}$ es ortogonal, i.e., $\mathbf{U}'\mathbf{U} = \mathbf{I}$
- $\mathbf{V}$ es ortogonal, i.e., $\mathbf{V}'\mathbf{V} = \mathbf{I}$ y
- Σ es diagonal.

Demostración

La matriz $\mathbf{X}\mathbf{X}'$ es una matriz cuadrada de $p \times p$ de rango p. La matriz $\mathbf{X}'\mathbf{X}$ es una matriz cuadrada de $n \times n$ de rango p (ya que $\mathbf{X}$ es de rango p). Como las matrices son simétricas y positivas definidas, deben tener p eigenvalores positivos y p eigenvectores ortonormales, asociados a estos eigenvalores.

Sean $\mathbf{v}_i$, $i = 1, 2, ..., p$ los vectores propios de $\mathbf{X}'\mathbf{X}$. Estos vectores pertenecen al espacio de los renglones de $\mathbf{X}$. Llamemos $\mathbf{u}_i$, $i = 1, 2, ..., p$ a los correspondientes vectores propios, asociados a los valores propios no nulos, de $\mathbf{X}\mathbf{X}'$. Estos vectores pertenecen al espacio de las

columnas de $\mathbf{X}$.

Estos vectores propios tienen una notable relación

$$\mathbf{X}\mathbf{v}_1 = \sigma_1\mathbf{u}_1;\ \mathbf{X}\mathbf{v}_2 = \sigma_2\mathbf{u}_2;\ \ldots;\mathbf{X}\mathbf{v}_p = \sigma_p\mathbf{u}_p \qquad \cdots(1)$$

con $\sigma_1, \sigma_2, ..., \sigma_p$ valores positivos llamados *valores singulares* de la matriz $\mathbf{X}$.

Esta relación se puede escribir a nivel matricial como

$$\mathbf{X}\left(\mathbf{v}_1\ \mathbf{v}_1 \cdots\ \mathbf{v}_p\right) = \left(\mathbf{u}_1\ \mathbf{u}_2 \cdots\ \mathbf{u}_p\right)\begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_p \end{pmatrix}$$

de donde se obtiene la descomposición

$$\mathbf{X}\mathbf{V} = \mathbf{U}\Sigma$$

y como $\mathbf{V}\mathbf{V}^{'} = \mathbf{I}$, multiplicando por la derecha por $\mathbf{V}^{'}$ la igualdad anterior, tenemos la *descomposición en valor singular*

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^{'}$$

Esta representación en valor singular, tiene una especialmente atractiva representación

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^{'} = \mathbf{u}_1\sigma_1\mathbf{v}_1^{'} + \mathbf{u}_2\sigma_2\mathbf{v}_2^{'} + \cdots + \mathbf{u}_p\sigma_p\mathbf{v}_p^{'}$$

donde cada elemento de la suma *tiene rango 1.* Si ordenamos los valores singulares $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$, esta decomposición en valor singular representa a la matriz $\mathbf{X}$ en elementos de rango uno, *en orden de importancia.*

Para hacer propiamente la demostración de esta descomposición, debemos mostrar que la relación mencionada en *(1)* es cierta. Así que comencemos dicha demostración.

Si λ_i es un eigenvalor no nulo de $\mathbf{X}^{'}\mathbf{X}$ con eigenvector asociado, $\mathbf{v}_i$, entonces, podemos escribir

$$\mathbf{X}'\mathbf{X}\mathbf{v}_i = \sigma_i^2\mathbf{v}_i, \quad \text{con } \sigma_i = \sqrt{\lambda_i} \text{ la raíz positiva de } \lambda_i$$

Entonces

$$\mathbf{v}_i'\mathbf{X}'\mathbf{X}\mathbf{v}_i = \sigma_i^2\mathbf{v}_i'\mathbf{v}_i = \sigma_i^2$$

y por lo tanto

$$\mathbf{v}_i'\mathbf{X}'\mathbf{X}\mathbf{v}_i = (\mathbf{X}\mathbf{v}_i)'(\mathbf{X}\mathbf{v}_i) = \|\mathbf{X}\mathbf{v}_i\| = \sigma_i^2$$

Además, de la misma igualdad, pero multiplicando por $\mathbf{X}$ por la izquierda, obtenemos

$$\mathbf{X}\mathbf{X}'\mathbf{X}\mathbf{v}_i = \sigma_i^2\mathbf{X}\mathbf{v}_i$$

lo que implica que $\mathbf{X}\mathbf{v}_i$ es un eigenvector de $\mathbf{X}\mathbf{X}'$ con eigenvalor asociado σ_i^2. Pero los eigenvectores de esta matriz eran $\mathbf{u}_i$, entonces

$$\mathbf{u}_i = \frac{\mathbf{X}\mathbf{v}_i}{\sigma_i} \Rightarrow \mathbf{X}\mathbf{v}_i = \sigma_i\mathbf{u}_i$$

que demuestra la relación que mencionamos entre estos eigenvalores.

BIPLOTS

Ahora, hagamos uso de esta descomposición para representar a los individuos y las variables de nuestros datos. Es claro que para lograr una buena representación de los individuos y de las variables en pocas dimensiones, debemos suponer que podemos reconstruir la matriz de datos considerando sólo unas cuantas dimensiones. En concreto, debemos suponer que

$$\mathbf{X} \approx \sum_{j=1}^{q} \lambda_j^{1/2} \mathbf{u}_j \mathbf{v}_j^{'} = \mathbf{U}_q \mathbf{\Sigma}_q \mathbf{V}_q^{'}$$

para la representación bidimensional, pediríamos $q = 2$. Ya que Σ_q es una matriz diagonal, la podemos asociar a la matriz $\mathbf{U}$ a $\mathbf{V}$ o a ambas a la vez. Por ejemplo, podemos definir

$$\mathbf{G}_q = \mathbf{U}_q \Sigma_q^{1-c} \quad y \quad \mathbf{H}_q^{'} = \Sigma_q^c \mathbf{V}_q^{'}$$

$0 \leq c \leq 1$. Para cada valor de c que elijamos, tenemos

$$\mathbf{X} = \mathbf{G}_q \mathbf{H}_q = \mathbf{U}_q \Sigma_q^{1-c} \Sigma_q^c \mathbf{V}_q^{'}$$

El exponente c se puede elegir de varias maneras. Las eleccioners habituales son $c = 0$, $c = \frac{1}{2}$ y $c = 1$

Sea $\mathbf{g}_i$ el i-ésimo renglón de $\mathbf{G}$ y $\mathbf{h}_j$ el j-ésimo renglón de $\mathbf{H}$ (por tanto, la j-ésima columna de $\mathbf{H}^{'}$). Si q=2, los $\mathbf{n+p}$ vectores $\mathbf{g}_i$ y $\mathbf{h}_j$ pueden representarse en el plano, dando lugar a la representación conocida como *biplot*. Los puntos $\mathbf{g}_i$ representan observaciones, y los puntos $\mathbf{h}_j$ representan variables.

Interpretación

Antes de interpretar el biplot, debemos relacionarlo con nuestra matriz de datos. Primero, denotemos como $\mathbf{S}$ (el estimador de Σ) a la matriz de varianza-covarianza muestral de $\mathbf{X}$ centrada sobre la media de cada variable, entonces tenemos que

$$\mathbf{S} = \frac{\mathbf{X}^{'}\mathbf{X}}{n-1} \quad \Rightarrow \quad \mathbf{X}^{'}\mathbf{X} = (n-1)\mathbf{S}$$

Por otro lado, escribimos la matriz de componentes principales como $\mathbf{Z} = \mathbf{XA}$, entonces

$$\mathbf{Z}'\mathbf{Z} = (\mathbf{XA})'(\mathbf{XA}) = \mathbf{A}'\mathbf{X}'\mathbf{XA} = (n-1)\mathbf{A}'\mathbf{SA} = (n-1)\mathbf{L}$$

$\mathbf{L}$ es la correspondiente matriz Λ, sólo que de eigenvalores estimados, ℓ_i.

Suponiendo, como es usual, que $\ell_i \neq 0 \ \forall i$, podemos definir la matriz diagonal $\mathbf{L}^{-1/2}$, cuyos elementos son $\ell_i^{-1/2}$.

Ya sabemos que $\mathbf{X}$ se puede representar mediante la descomposición en valor singular de una matriz. Entonces, definamos las siguientes matrices

$\mathbf{U} = (n-1)^{-1/2}\mathbf{ZL}^{-1/2} = (n-1)^{-1/2}\mathbf{XAL}^{-1/2}$ $\Big($cuya k-ésima columna es $(n-1)^{-1/2}\ell_k^{-1/2}\mathbf{Xa}_k$, k=1,2,...,p$\Big)$

$\mathbf{L} = (n-1)^{1/2}\mathbf{L}^{1/2}$ (abuso de notación. Matriz diagonal cuyo k-ésimo elemento es $(n-1)^{1/2}\lambda_k^{1/2}$), y

$\mathbf{A} = \mathbf{A}$ (cuyas columnas son los eigenvectores $\mathbf{a}_k$, k=1,2,...,p)

Obsérvese que

$$\mathbf{ULA}' = (n-1)^{-1/2}\left[\ell_1^{-1/2}\mathbf{Xa}_1, \ell_2^{-1/2}\mathbf{Xa}_2, ..., \ell_p^{-1/2}\mathbf{Xa}_p\right](n-1)^{1/2}\left[\ell_1^{1/2}\mathbf{a}_1, \ell_2^{1/2}\mathbf{a}_2, ..., \ell_1^{1/2}\mathbf{a}_p\right]'$$

$$= \sum_{k=1}^{p} \ell_k^{-1/2}\mathbf{Xa}_k\ell_k^{1/2}\mathbf{a}_k' = \sum_{k=1}^{p}\mathbf{Xa}_k\mathbf{a}_k' = \mathbf{X}$$

Entonces, hemos escrito $\mathbf{X}$ en términos de la descomposición dada por estas tres matrices, i.e.

$$\mathbf{X}_{n\times p} = \mathbf{U}_{n\times p}\mathbf{L}_{p\times p}\mathbf{A}'_{p\times p}$$

La identificación con las matrices que resultaron del desarrollo del proceso de descomposición en valor singular es

$$\mathbf{U} = \mathbf{U}, \quad \Sigma = \mathbf{L} \quad y \quad \mathbf{A}' = \mathbf{V}'$$

Ahora sí, para construir el biplot, definimos los elementos de la descomposición de $\mathbf{X}$ como

$$\mathbf{X} = \mathbf{G}\mathbf{H}^{'}, \quad con \quad \mathbf{G} = \mathbf{U} \quad y \quad \mathbf{H}^{'} = \mathbf{L}\mathbf{A}^{'}$$

Esta definición implica tomar *c=1* en la representación general de los biplots. Si denotamos por $\mathbf{g}_i^{'}, i = 1, 2, ..., n$ y $\mathbf{h}_j^{'}, j = 1, 2, ..., p$ los renglones de $\mathbf{G}$ y $\mathbf{H}$, respectivamente. Entonces, el elemento *(i,j)* de $\mathbf{X}$ se puede escribir como

$$x_{ij} = \mathbf{g}_i^{'}\mathbf{h}_j$$

Varios resultados

1.- $\mathbf{U}'\mathbf{U} = \left((n-1)^{-1/2}\mathbf{Z}\mathbf{L}^{-1/2}\right)' \left((n-1)^{-1/2}\mathbf{Z}\mathbf{L}^{-1/2}\right) = (n-1)^{-1}\mathbf{L}^{-1/2}\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A}$

$= (n-1)^{-1}\mathbf{L}^{-1/2}(n-1)\mathbf{L}\mathbf{L}^{-1/2} = \mathbf{I}_p$

2.-$\mathbf{X}'\mathbf{X} = \mathbf{H}\mathbf{H}' = (n-1)\mathbf{S}$

Demostración

$$(n-1)\mathbf{S} = \mathbf{X}'\mathbf{X} = \left(\mathbf{G}\mathbf{H}'\right)' \left(\mathbf{G}\mathbf{H}'\right) = \mathbf{H}\mathbf{U}'\mathbf{U}\mathbf{H}' = \mathbf{H}\mathbf{H}'$$

3.- $\mathbf{h}_j'\mathbf{h}_j = \|\mathbf{h}_j\|^2 = \ell_j^{1/2}\mathbf{a}_j'\ell_j^{1/2}\mathbf{a}_j = \ell_j\mathbf{a}_j'\mathbf{a}_j = \ell_j = Var(X_j),\ j = 1, 2, ..., p$

4.- $Cov\left(X_i, X_j\right) = \mathbf{h}_i'\mathbf{h}_j$

5.- $Corr\left(X_i, X_j\right) = \dfrac{\mathbf{h}_i'\mathbf{h}_j}{\|\mathbf{h}_i\|\,\|\mathbf{h}_j\|}$, es el coseno del ángulo entre los vectores $\mathbf{h}_i$ y $\mathbf{h}_j$

Obsérvese que los elementos de $\mathbf{H}$ representan a las variables y algunas de las características de ellas se obtienen a través de estos elementos.

Y los individuos?.

Observemos que $x_{ij} = \mathbf{g}_i'\mathbf{h}_j$ es un escalar que corresponde al valor que tiene el individuo i en la variable j. Si queremos escribir de esta forma al vector completo de observaciones del individuo i, lo debemos reescribir como $\mathbf{X}_i = \mathbf{g}_i'\mathbf{H}' = \left(\mathbf{g}_i'\mathbf{h}_1, \mathbf{g}_i'\mathbf{h}_2, ..., \mathbf{g}_i'\mathbf{h}_p\right), i = 1, 2, ..., n$ (que denota que estamos proyectando al vector g_i' sobre cada columna de $\mathbf{H}$). Recordar que $\mathbf{h}_j'$ son los renglones de $\mathbf{H}$, por lo tanto, $\mathbf{h}_j$ son las columnas de $\mathbf{H}'$. Y además, nuevamente abusando de la notación, escribimos el vector $\mathbf{X}_i$, como vector columna

$$\mathbf{X}_i = \mathbf{X}_i' = \left(\mathbf{g}_i'\mathbf{H}'\right)' = \mathbf{H}\mathbf{g}_i$$

Demostremos que la distancia entre dos elementos de $\mathbf{G}$; $\mathbf{g}_i, \mathbf{g}_j$, es proporcional a la distancia de *Mahalanobis* entre las observaciones $\mathbf{X}_i$. Antes necesitamos el siguiente resultado. Par-

tiendo nuevamente de la descomposición en valor singular, tenemos

$\mathbf{X} = \mathbf{ULA}^{'} \rightarrow \mathbf{X}^{'}\mathbf{X} = \mathbf{ALU}^{'}\mathbf{ULA}^{'} = \mathbf{AL}^{2}\mathbf{A}^{'}$. Por otro lado

$\mathbf{X}^{'}\mathbf{XA} = \mathbf{AL}^{2}\mathbf{A}^{'}\mathbf{A} = \mathbf{AL}^{2} \rightarrow \left(\mathbf{X}^{'}\mathbf{XA}\right)^{-1} = \mathbf{L}^{-2}\mathbf{A}^{-1}$ de donde

$\mathbf{A}^{-1}\left(\mathbf{X}^{'}\mathbf{X}\right)^{-1} = \mathbf{A}^{'}\left(\mathbf{X}^{'}\mathbf{X}\right) = \mathbf{L}^{-2}\mathbf{A}^{'}$

La distancia de Mahalanobis entre dos vectores es

$$\delta_{ij}^{2} = \left(\mathbf{X}_i - \mathbf{X}_j\right)^{'}\mathbf{S}^{-1}\left(\mathbf{X}_i - \mathbf{X}_j\right)$$

Entonces

$$\begin{aligned}
\delta_{ij}^{2} &= \left(\mathbf{Hg}_i - \mathbf{Hg}_j\right)^{'}\mathbf{S}^{-1}\left(\mathbf{Hg}_i - \mathbf{Hg}_j\right) \\
&= \left(\mathbf{g}_i - \mathbf{g}_j\right)^{'}\mathbf{H}^{'}\mathbf{S}^{-1}\mathbf{H}\left(\mathbf{g}_i - \mathbf{g}_j\right) \\
&= (n-1)\left(\mathbf{g}_i - \mathbf{g}_j\right)^{'}\mathbf{LA}^{'}\left(\mathbf{X}^{'}\mathbf{X}\right)^{-1}\mathbf{AL}\left(\mathbf{g}_i - \mathbf{g}_j\right) \\
&= (n-1)\left(\mathbf{g}_i - \mathbf{g}_j\right)^{'}\mathbf{LL}^{-2}\mathbf{A}^{'}\mathbf{AL}\left(\mathbf{g}_i - \mathbf{g}_j\right) \\
&= (n-1)\left(\mathbf{g}_i - \mathbf{g}_j\right)^{'}\left(\mathbf{g}_i - \mathbf{g}_j\right) \\
&\propto \left\|\mathbf{g}_i - \mathbf{g}_j\right\|^{2}
\end{aligned}$$

En resumen. Dada la descomposición en valor singular de $\mathbf{X}$

$$\mathbf{X} = \mathbf{GH}^{'}, \quad \text{con} \quad \mathbf{G} = \mathbf{U} \text{ y } \mathbf{H}^{'} = \mathbf{LA}^{'}$$

los elementos de $\mathbf{G}$ representan a los individuos con

$$\left\|\mathbf{g}_i - \mathbf{g}_j\right\|^{2} \propto \delta_{ij}^{2} = \left(\mathbf{X}_i - \mathbf{X}_j\right)^{'}\mathbf{S}^{-1}\left(\mathbf{X}_i - \mathbf{X}_j\right)$$

Los elementos de $\mathbf{H}$ representan a las variables, con las siguientes características

• $Var(\mathbf{X}_j) = \mathbf{h}_j^{'}\mathbf{h}_j = \|h_j\|^2$, j=1,2,...,p

• $Cov(\mathbf{X}_i, \mathbf{X}_j) = \mathbf{h}_i^{'}\mathbf{h}_j$

• $Corr(\mathbf{X}_i, \mathbf{X}_j) = \dfrac{\mathbf{h}_i^{'}\mathbf{h}_j}{\|\mathbf{h}_i\| \, \|\mathbf{h}_j\|}$

Entonces el **Biplot** es una representación gráfica bidimensional de los individuos y las variables, a través de los vectores **g** y **h**, suponiendo que esta representación en dos dimensiones es una buena aproximación. Es decir que

$$x_{ij} \approx g_i^{*'} h_j^*$$

Con g^* y h^* vectores en $\mathbb{R}^2$. Entonces, el biplot se construye graficando a los individuos como puntos $\mathbf{g}_i^{*'} = (\ell_1^{1/2} u_{1i}, \ell_2^{1/2} u_{2i})$ y los p vectores, cuyo punto final se encuentra en $\mathbf{h}_j^{'} = (\ell_1^{1/2} a_{1i}, \ell_2^{1/2} a_{2i})$.

Ahora sí estamos en posibilidad de hacer la interpretación del biplot.

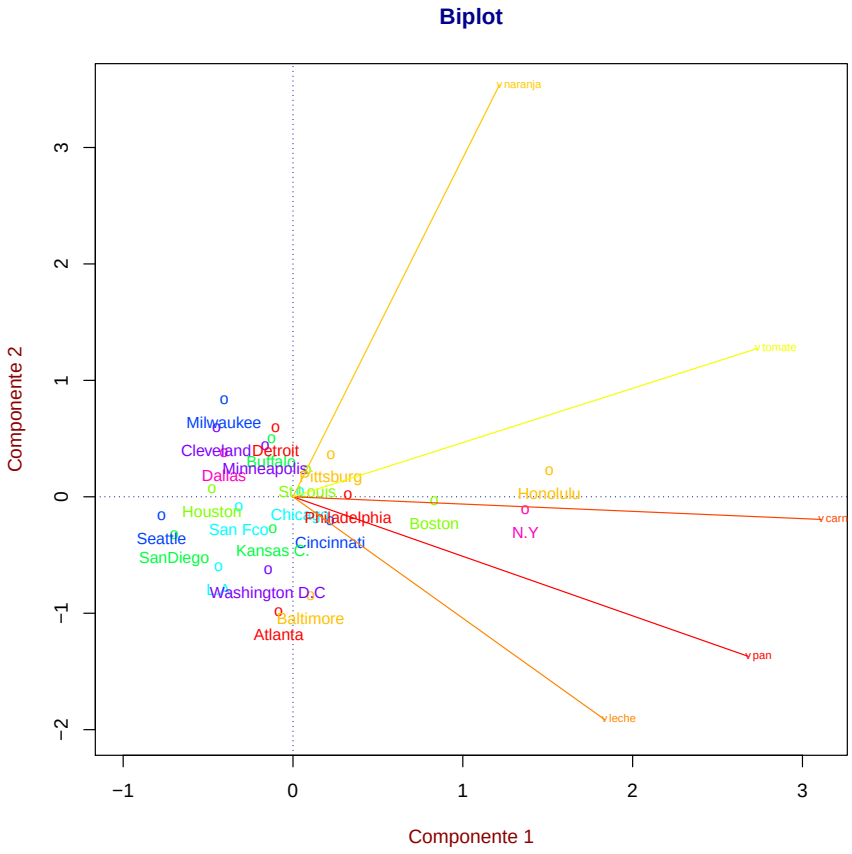
• Individuos semejantes representarán puntos cercanos en la gráfica

• Variables cuyo ángulo entre los vectores que las representan sea pequeño, serán variables con una fuerte correlación, ya que $cos(\theta)$ es una función decreciente de 0^0 a 90^0 y $cos(0^0) = 1$ (los vectores son colineales) y $cos(90^0) = 0$ (los vectores son ortogonales).
Colineales $\Rightarrow$ corr=1, ortogonales $\Rightarrow$ corr=0.

• Finalmente, ya que escribimos a los elementos de la matriz **X** como

$$x_{ij} \approx \mathbf{g}_i^{'}\mathbf{h}_j = \|\mathbf{g}\| \, \|\mathbf{h}\| \, cos(\theta_{ij})$$

que es la proyección de la observación i en la variable j. Para apreciar la magnitud del registro de un individuo en una variable, hay que proyectar el punto que representa al individuo sobre el vector que representa la variable, mientras más pequeña sea esta proyección, más grande será la magnitud del registro del individuo en la variable.


Biplot
v naranja
v tomate
v carn
v pan
v leche
Milwaukee
Cleveland
Detroit
Buffalo
Minneapolis
Dallas
Pittsburg
St.Louis
Honolulu
Houston
Chic
Philadelphia
Boston
N.Y
San Fco
Seattle
Cincinnati
SanDiego
Kansas C.
Washington D.C
Baltimore
Atlanta
Componente 1
Componente 2
−1
0
1
2
3
−2

ANÁLISIS DE FACTORES

Introducción

El análisis factorial es una técnica estadística multivariada que se incorpora a la metodología cuantitativa que involucra *variables latentes*. De uso común en diversas área del conocimiento relacionadas con las ciencias sociales. Por ejemplo, el análisis factorial se ha utilizado en psicología en estudios de habilidades, motivación, aprendizaje, etc.; en pedagogía, en estudios relacionados con el aprovechamiento escolar, la tipología de profesores, etc.; en sociología, en dimensiones de grupo, actitudes políticas, afinidad política, etc., y en muchas otras disciplinas como: ecología, economía, medicina, metrología, educación, evaluación, sólo por mencionar algunas.

Concepto de factor

Un factor, también conocido como *variable latente o constructo* (psicología), se puede definir como una variable que no puede medirse de manera directa, pero que está asociada con un conjunto de variables observadas correlacionadas entre sí. Más aún, se supone que la correlación de estas variables observadas se debe precisamente a que tienen en común a este factor.

Ejemplos clásicos de factores

- Inteligencia
- Nivel socieconómico
- Salud
- Bienestar
- Satisfacción
- Desarrollo
- Personalidad, etc.

El análisis factorial tiene por objeto explicar la estructura de correlación entre un conjunto de variables observadas, a través de un pequeño número (reducción de dimensión) de *variables latentes, no observadas y no observables, llamadas factores*. Por ejemplo, supongamos que hemos tomado varias medidas físicas del cuerpo de una persona: estatura, longitud del tronco y de las extremidades, anchura de hombros, peso, etc. Es intuitivamente claro que todas estas medidas no son independientes entre sí, y podrían contener factores relacionados con *la talla* y *la masa corporal* de los sujetos. Como segundo ejemplo, supongamos que estamos interesados en estudiar el desarrollo humano (*factor*) en los países del mundo, y que disponemos de variables económicas, sociales y demográficas, en general dependientes entre sí, que están relacionadas con este factor de desarrollo. Como tercer ejemplo, supongamos que medimos, con distintas pruebas, la capacidad mental de un individuo para procesar información y resolver problemas. Podemos preguntarnos si existen factores, no observables, que expliquen el conjunto de resultados observados. El conjunto de estos factores será lo que llamamos inteligencia y es importante conocer cuántas dimensiones distintas tiene este concepto y cómo caracterizarlas y medirlas. El análisis factorial surge impulsado por el interés de Charles Spearman (1904) en comprender las dimensiones de la inteligencia humana, y muchos de sus avances se han producido en el área de la Psicometría.

Objetivo del análisis de factores

• Explicar la estructura de correlación entre un conjunto de variables medidas

• Determinar si el conjunto de variables exhiben patrones de relación entre sí, de tal manera que se puedan dividir en subgrupos (factores) en los que las variables que integran cada subgrupo, estén más fuertemente correlacionadas entre ellas, que con el resto de los subconjuntos.

• Entonces, lo que se tiene es un subconjunto de variables medidas $X_1, X_2, ..., X_p$ y se supone que a este conjunto de variables subyacen k factores con $k << p$.

El modelo de factores

$$
\begin{aligned}
X_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1k} f_{1k} + u_1 \\
X_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2k} f_{1k} + u_2 \\
&\vdots \\
X_p &= \lambda_{p1} f_1 + \lambda_{p2} f_2 + \cdots + \lambda_{2k} f_{pk} + u_p
\end{aligned}
$$

!Como un modelo de regresión lineal múltiple, en el que ahora "la respuesta" es cada una de las $X's$ y donde los factores $f_1, f_2, ..., f_k$ son las variables explicativas! Y los errores son las $u's$, llamados factores específicos.

En notación matricial

$$\mathbf{X} = \Lambda \mathbf{F} + \mathbf{U}$$

con

$$
\mathbf{X}_{n\times p} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \quad \Lambda_{p\times k} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1k} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pk} \end{pmatrix} \quad \mathbf{F}_{k\times 1} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_k \end{pmatrix} \quad \mathbf{U}_{p\times 1} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{pmatrix}
$$

A $f_1, f_2, ..., f_k$ se les denomina factores comunes (comunalidad) y $u_1, u_2, ..., u_p$ factores específicos (especificidad).

El modelo tiene algunos supuestos sobre los que se construye, que son:

• Los factores comunes f_j j=1,2,...,k no están correlacionados y tienen media cero y varianza uno

• Los factores específicos u_i no están correlacionados y tienen media cero y varianza ψ_i i=1,2,...,p

• Los factores comunes no están correlacionados con los factores específicos

Bajo estos supuestos tenemos que

$$\mathbb{V}(X_i) = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_i = h_i^2 + \psi_i, \;\; i = 1, 2, ..., p$$

con h_i^2 conocida como la comunalidad de la variable (la varianza de la variable X_i explicada por los k factores comunes) y ψ_i conocida como la especificidad (la correspondiente varianza no explicada por los factores comunes). Totalmente análogo a regresión.

Además se tiene que

$$\mathbb{C}ov(X_i, X_l) = \mathbb{C}ov\left(\sum_{j=1}^{k} \lambda_{ij} f_j + u_i, \sum_{j=1}^{k} \lambda_{lj} f_j + u_l\right) = \sum_{j=1}^{k} \lambda_{ij}\lambda_{lj}, \;\; \forall\, i \neq l, \; i, l = 1, 2, ..., p$$

Podemos observar que los factores comunes explican las relaciones existentes entre las variables del problema (relaciones que se establecieron a través de la matriz de correlación). Es por esta razón que los factores que tienen interés y son susceptibles de interpretación son los factores comunes. Los factores únicos o factores específicos se incluyen en el modelo dada la imposibilidad de expresar, en general, p variables en función de un número más reducido, k, de factores. Entonces, los factores comunes y sus características asociadas (comunalidades, especificidades, número, etcétera) representan el objeto de interés en el análisis factorial.

El hecho de que la varianza y covarianza de las variables medidas se pueda expresar en términos del modelo factorial, implica que la matriz de correlación de las variables se puede escribir como

$$\Sigma = \Lambda\Lambda^{'} + \Psi$$

Entonces, el objetivo del análisis factorial es determinar *k: número de factores*, $\hat{\Lambda}$, $\hat{\Psi}$ utilizando la matriz de correlación muestral $\hat{\Sigma} = \mathbf{R}$. Con lo que se obtiene

$$\mathbf{R} = \hat{\Lambda}\hat{\Lambda}^{'} + \hat{\Psi}$$

Soluciones múltiples al modelo

Un aspecto muy importante es que la solución del modelo de factores no es única, en el sentido de que si tenemos una matriz ortogonal $\mathbf{M}$ (la condición de ortogonalidad $\Rightarrow \mathbf{M}\mathbf{M}^{'} = \mathbf{I}$), podemos escribir:

$$\begin{aligned}
\mathbf{R} &= \Lambda\Lambda^{'} + \Psi \\
\mathbf{R} &= \Lambda\mathbf{I}\Lambda^{'} + \Psi \\
\mathbf{R} &= \Lambda\mathbf{M}\mathbf{M}^{'}\Lambda^{'} + \Psi \\
\mathbf{R} &= (\Lambda\mathbf{M})(\Lambda\mathbf{M})^{'} + \Psi
\end{aligned}$$

Entonces, si Λ es una matriz de cargas factoriales, $\Lambda\mathbf{M}$ también lo es, para toda matriz ortogonal, $\mathbf{M}$. Por lo tanto, la matriz de cargas factoriales no es única, y esto implica que los factores tampoco son únicos.

Para garantizar una solución única en este modelo debemos anexar alguna restricción. La forma usual de este tipo de restricciones es alguna de las siguientes:

$$\Lambda^{'}\Lambda = \Gamma, \quad \Lambda^{'}\Psi^{-1}\Lambda = \Gamma, \quad \text{ó} \quad \Lambda^{'}\mathbf{D}^{-1}\Lambda = \Gamma$$

con Λ y $\mathbf{D}$ matrices no singulares.

Obsérvese que el producto de $\Lambda^{'}\Lambda$ no genera una matriz diagonal, aunque las restricciones del modelo exigen que lo sea, es decir, que los elementos fuera de la diagonal de este producto sean cero. Por ello, y ya que fuera de la diagonal tenemos *k(k-1)* elementos, entonces es necesario este número de restricciones para garantizar una solución única del modelo.

Número máximo de factores

De acuerdo con la discusión anterior, conviene saber cuál es el máximo número de factores que podemos extraer de un conjunto de p variables medidas.

En el análisis factorial ¿quién o qué constituye nuestra información?

Como la idea es descomponer la matriz de correlación, entonces los elementos no redundantes de ésta, representan nuestra información. En el caso de que tengamos p variables medidas, el número de elementos no redundantes es $p(p+1)/2$. Ahora bien, necesitamos estimar $p * k$ cargas factoriales totales y p especificidades, entonces necesitamos estimar $p(k+1)$ parámetros de nuestro modelo. Y necesitamos imponer a este número de parámetros por estimar, $k(k-1)$ restricciones para obtener una solución única. Es lógico suponer que esta diferencia entre los parámetros por estimar y las restricciones no debe exceder el número de elementos no redundantes de la matriz de correlación (nuestra información observada). Entonces, se debe cumplir que:

$$\frac{p(p+1)}{2} \geq p(k+1) - \frac{k(k-1)}{2} \Rightarrow (p-k)^2 \geq p+k$$

A partir de esta desigualdad podemos observar que el mínimo de variables requeridas para extraer un factor es 3 (véase que en este caso se cumple la igualdad). Con cinco variables observadas podemos tener a lo más dos factores; con 20 el número máximo de factores puede ser hasta de 14; sin embargo, en la práctica no se busca encontrar este número máximo, sino aquél que nos permita explicar, de la mejor manera posible, las correlaciones entre estas variables medidas. Entonces, en la situación donde el número de parámetros por estimar sobrepase al número de elementos no redundantes de la matriz de correlación, simplemente afirmaremos que el modelo de factores *no existe*. En el caso de que existan tantos parámetros como elementos no redudantes, es posible que el modelo de factores exista, pero también es posible que no exista. Finalmente, cuando los elementos no redundantes de la matriz son más que el número de parámetros por estimar, el modelo de factores existe y es posible que proporcione una explicación más simple de las relaciones entre las variables observadas, que la que proporciona la matriz de correlación, $\mathbf{R}$.

Un ejemplo del caso de igualdad

Como acotamos en el párrafo anterior, cuando se tienen tres variables manifiestas y un solo factor, se cumple la igualdad en este criterio para el número máximo de factores. Al respecto, Everitt (2001) proporciona el siguiente ejemplo, que, además de tratar con detalle esta situación, nos proporcionará una visión clara de los procesos inmersos en la solución de estos modelos. Se tienen las calificaciones de exámenes de un grupo de estudiantes, en las asignaturas de X_1: Literatura clásica, X_2: Francés y X_3: Inglés, de las que se obtiene la siguiente matriz de correlaciones:

$$\mathbf{R} = \begin{pmatrix} 1 & & \\ 0.83 & 1 & \\ 0.78 & 0.67 & 1 \end{pmatrix}$$

Ya que no puede ser de otra forma, supongamos que se tiene un solo factor subyacente a los datos, que podrías llamar como *habilidad linguística*. Entonces, el proceso para estimar los parámetros es el siguiente:

El modelo de factores subyacente es:

$$X_1 = \lambda_{11} f_1 + u_1$$
$$X_2 = \lambda_{21} f_1 + u_2$$
$$X_3 = \lambda_{31} f_1 + u_3$$

Obsérvese que:

$\frac{p(p+1)}{2} = \frac{3*4}{2} = 6$ y $p(k+1) = 3*(1+1) = 6$ con número de restricciones $k(k-1) = 0$. Entonces, el número de parámetros por estimar coincide con el número de elementos no redundantes de la matriz de correlación. Como comentamos líneas arriba, el objetivo es encontrar, a partir de la matriz de correlación, $\mathbf{R}$, las matrices $\hat{\Lambda}$ y $\hat{\Psi}$. Recordando cómo se escriben las varianzas y covarianzas de las variables, en términos de los elementos del modelo de factores, en este caso tenemos:

$\mathbf{R} = \Lambda\Lambda' + \Psi \Rightarrow$

$$\begin{pmatrix} 1 & & \\ 0.83 & 1 & \\ 0.78 & 0.67 & 1 \end{pmatrix} = \begin{pmatrix} \lambda_{11} \\ \lambda_{21} \\ \lambda_{31} \end{pmatrix} (\lambda_{11}, \lambda_{21}, \lambda_{31}) + \begin{pmatrix} \psi_1 & & \\ & \psi_2 & \\ & & \psi_3 \end{pmatrix}$$

$$= \begin{pmatrix} \lambda_{11}^2 + \psi_1 & & \\ \lambda_{21}\lambda_{11} & \lambda_{21}^2 + \psi_2 & \\ \lambda_{31}\lambda_{11} & \lambda_{31}\lambda_{21} & \lambda_{31}^2 + \psi_3 \end{pmatrix}$$

De este sistema se desprenden las ecuaciones:

$$\lambda_{11} * \lambda_{21} = 0.83$$
$$\lambda_{11} * \lambda_{31} = 0.78$$
$$\lambda_{21} * \lambda_{31} = 0.67$$

que puede resolverse de diversas manera para obtener

$$\hat{\lambda}_{11} = 0.98 \quad \hat{\lambda}_{21} = 0.84 \quad \hat{\lambda}_{31} = 0.79$$

De las relaciones

$$\lambda_{11}^2 + \psi_1 = \lambda_{21}^2 + \psi_2 = \lambda_{31}^2 + \psi_3 = 1$$

obtenemos

$$\hat{\psi}_1 = 0.04 \quad \hat{\psi}_2 = 0.29 \quad \hat{\psi}_3 = 0.39$$

Por lo que

$$\hat{\Lambda} = \begin{pmatrix} \hat{\lambda}_{11} \\ \hat{\lambda}_{21} \\ \hat{\lambda}_{31} \end{pmatrix} = \begin{pmatrix} 0.98 \\ 0.84 \\ 0.79 \end{pmatrix} \qquad \hat{\Psi} = \begin{pmatrix} \hat{\psi}_1 & & \\ & \hat{\psi}_2 & \\ & & \hat{\psi}_3 \end{pmatrix} = \begin{pmatrix} 0.04 & & \\ & 0.29 & \\ & & 0.39 \end{pmatrix}$$

podemos observar que todos los parámetros estimados tienen valores admisibles.

Supongamos ahora que tomamos una nueva muestra sobre estos exámenes, que arroja la siguiente matriz de correlación:

$$\mathbf{R} = \begin{pmatrix} 1 & & \\ 0.84 & 1 & \\ 0.60 & 0.35 & 1 \end{pmatrix}$$

Entonces, realizando el procedimiento anterior llegamos a:

$$\hat{\Lambda} = \begin{pmatrix} \hat{\lambda}_{11} \\ \hat{\lambda}_{21} \\ \hat{\lambda}_{31} \end{pmatrix} = \begin{pmatrix} 1.12 \\ 0.70 \\ 0.50 \end{pmatrix} \qquad \hat{\Psi} = \begin{pmatrix} \hat{\psi}_1 & & \\ & \hat{\psi}_2 & \\ & & \hat{\psi}_3 \end{pmatrix} = \begin{pmatrix} -0.44 & & \\ & 0.51 & \\ & & 0.75 \end{pmatrix}$$

que tiene dos parámetros estimados inadmisibles, $\mathbb{V}(X_1) = \hat{\psi}_1 = -0.44$ y $\hat{\lambda}_{11} = 1.2$. Este último debido a que estima la correlación entre X_1 f_1^*, por lo que no puede ser mayor que uno. El ejemplo muestra que la igualdad en el criterio del número máximo de factores que se pueden extraer, puede generar resultados inapropiados, por lo que es preferible considerar la desigualdad estricta. También ilustra el principio sobre el que se basa el proceso de estimación: igualar la matriz de correlaciones generada por el modelo, que involucra a los parámetros que lo componen, con la matriz de correlación estimada con la información.

Tareita

Demuestre $*$. Es decir, demuestre que λ_{ij} es la correlación entre X_i y f_j

Estimación de los parámetros

Antes de presentar los distintos métodos para estimar los parámetros involucrados en este modelo, es importante remarcar que el análisis de factores se basa precisamente *en un mo delo*, es decir, se asume que a los datos por analizar *subyace un modelo*; esta condición lo hace diferente al análisis de componentes principales que no asume la existencia de ningún modelo y se basa simplemente en la descomposición de la matriz de varianza-covarianza o de correlación, en sus eigenvalores y eigenvectores. En este sentido, es claro que en el análisis de factores es necesario hacer alguna(s) prueba(s) de *bondad de ajuste* para verificar si los datos se ajustan al modelo propuesto. Pero, en qué momento se propuso un modelo?. Aunque generalmente no se hace de *manera totalmente explícita*, al decidir retener k factores en el análisis, intrínsecamente se asume que el *modelo propuesto es un modelo factorial con k factores*. Entonces, en esencia, estaríamos afirmando que *la estructura de correlación de las variables o la matriz de correlación de ellas, se puede explicar a través de estos k factores retenidos*. Por lo tanto, deberíamos probar que este modelo con *k factores ajusta adecuadamente a nuestros datos*.

Habíamos comentado que el hecho de que la varianza y covarianza de las variables medidas se pueda expresar en términos del modelo factorial, implicaba que la matriz de correlación de las variables se podía escribir como:

$$\Sigma = \Lambda\Lambda^{'} + \Psi$$

entonces, es claro que Σ, la matriz de correlaciones que se desprende del modelo, depende de los parámetros del mismo modelo; entonces $\Sigma = \Sigma\left(\underline{\theta}\right)$. Y si $\mathbf{R}$ representa la respectiva matriz de correlación de los datos, entonces el objetivo de los métodos de estimación es minimizar alguna función de distancia entre estas dos matrices, es decir, la función por minimizar es de la forma:

$$\mathbb{F} = G\left(\left|\Sigma\left(\underline{\theta}\right) - \mathbf{R}\right|\right)$$

con G alguna función específica. Los valores en $\Sigma\left(\underline{\theta}\right)$ que minimicen esta función de distancia serán los estimadores de sus parámetros. Tomando en cuenta que Σ se puede descomponer como:

$$\Sigma = \Lambda\Lambda' + \Psi$$

los procesos que minimizan esta función de distancia entre estas dos matrices son equivalentes a encontrar los estimadores de Λ y Ψ tales que:

$$\mathbf{R} = \hat{\Sigma} \approx \hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}$$

Máxima Verosimilitud

En este caso, la función de distancia se desprende de la verosimilitud del modelo, y tiene la forma

$$\mathbb{F}\left(\Sigma\left(\underline{\theta}\right), \mathbf{R}\right) \propto -\frac{1}{2} log\left(\Sigma\left(\underline{\theta}\right)\mathbf{R}^{-1}\right) - traza\left(\Sigma\left(\underline{\theta}\right)\mathbf{R}^{-1}\right)$$

Aunque en este método el objetivo es maximizar la verosimilitud, cabe recordar que maximizar es equivalente a minimizar el negativo de esta verosimilitud.

Este método de estimación demanda que $\mathbf{X}$ tenga una distribución normal multivariada, hecho que en la práctica es muy difícil que se cumpla. No obstante, se ha encontrado que el método es robusto ante desviaciones de la normalidad. Sin embargo, es inadecuado su uso con variables nominales u ordinales.

Mínimos Cuadrados

En este caso, la función que se minimiza es:

$$\mathbb{F}\left(\Sigma\left(\underline{\theta}\right), \mathbf{R}\right) = traza\left[\left(\mathbf{R} - \Sigma\left(\underline{\theta}\right)\right)^2\right]$$

que también puede considerarse una medida de distancia entre la matriz observada, $\mathbf{R}$ y la matriz generada por el modelo, $\Sigma\left(\underline{\theta}\right)$. Se minimiza la suma de cuadrados de las diferencias entre estas dos matrices. Nuevamente, los valores de los parámetros que minimicen esta función serán los estimadores.

Mínimos Cuadrados Generalizados

Este método es una generalización del de mínimos cuadrados; la función por minimizar es:

$$\mathbb{F}\left(\Sigma\left(\underline{\theta}\right), \mathbf{R}\right) = traza\left[\left(\left(\mathbf{R} - \Sigma\left(\underline{\theta}\right)\right)\mathbf{R}^{-1}\right)^2\right]$$

la intención es minimizar la suma de cuadrados de todos los elementos en este producto de matrices.

Mínimos Cuadrados Ponderados

En este método el objetivo es minimizar la diferencia entre la matriz generada por el modelo y la estimada por nuestros datos, ponderando estas diferencias por una matriz de pesos. Concretamente, la función que debemos minimizar tiene la forma:

$$\mathbb{F}\left(\Sigma\left(\underline{\theta}\right), \mathbf{R}\right) = traza\left[\left(\left(\mathbf{R} - \Sigma\left(\underline{\theta}\right)\right)\Psi^{-1}\right)^2\right]$$

con Ψ la matriz definida anteriormente.

Método de Ejes Principales (Principal axis Factor Analysis)

Este método de estimación no requiere ningún supuesto sobre la distribución de la matriz de datos, $\mathbf{X}$, por lo que es preferible a cualquiera de los anteriores. En este caso se utiliza la llamada matriz reducida $\mathbf{R}^*$ definida como

$$\mathbf{R}^* = \mathbf{R} - \hat{\Psi} = \hat{\Lambda}'\hat{\Lambda}$$

por lo que los elementos en la diagonal de $\mathbf{R}^*$ son las comunalidades estimadas. Este proceso requiere de una estimación inicial de estas comunalidades. Los métodos más frecuentes para estas estimaciones iniciales son:

• El coeficiente de correlación múltiple entre cada X_i y el resto de las variables, y

• El mayor coeficiente de correlación, en valor absoluto, entre X_i y cualquiera de las otras variables, es decir:

$$\tilde{h}_i^2 = \max_{i \neq j} |r_{ij}|$$

con r_{ij} la correlación entre las variables X_i y X_j. A partir de las estimaciones iniciales de las comunalidades se hace un proceso de componentes principales sobre R^* para encontrar

las cargas factoriales. Posteriormente se actualizan los estimadores de las comunalidades. El proceso continúa de forma iterativa, hasta que el cambio en las estimaciones entre dos iteraciones consecutivas es prácticamente nulo.

Bondad de Ajuste

Dado que supusimos que subyace un modelo a nuestros datos, entonces es necesario verificar lo adecuado del ajuste de este modelo a nuestra información, a través de alguna(s) prueba(s) de bondad de ajuste.

Residuos

Un elemento fundamental en todos los procesos de bondad de ajuste sobre un modelo, lo constituye los llamados *residuos* que, como sabemos, corresponden a la diferencia entre los valores observados y los valores ajustados por el modelo propuesto. En este caso, como la intención de los métodos de estimación, fue encontrar el valor de los parámetros que mejor aproximara la matriz de correlación generada por el modelo de factores y la generada por los datos, estos residuos son

$$\mathbf{R} - \Sigma(\hat{\theta}) = \mathbf{R} - \left(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}\right)$$

Si el ajuste del modelo de factores a los datos es adecuado, entonces esta matriz debe tener valores pequeños en todas las entradas. Este "buen ajuste" lo que significa es que *efectivamente subyacen a los datos los k factores propuestos*. Obsérvese que las entradas de esta matriz son correlaciones que están entre *-1 y 1*, así que se esperan valores realmente pequeños para que se considere un buen ajuste.

Prueba sobre el número de factores en el modelo

En esta prueba el objetivo es contrastar si el modelo con *k factores que hemos propuesto* ajusta bien a los datos. En otras palabras: si *k factores son suficientes para explicar la estructura de correlación subyacente a las variables medidas*. Esta prueba supone que la matriz de datos $\mathbf{X}$ tiene una distribución normal multivariada. Entonces, se trata de realizar la prueba

$$\mathbb{H}_0 : \Sigma = \Sigma\left(\underline{\theta}\right) = \Lambda\Lambda' + \Psi \;\; vs. \;\; \mathbb{H}_a : \Sigma \neq \Sigma\left(\underline{\theta}\right) = \Lambda\Lambda' + \Psi$$

Bajo el supuesto de normalidad multivariada, el estadístico de prueba

$$\left(n - \frac{2(p+2k)+11}{6}\right) ln \left(\frac{\left|\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}\right|}{|\mathbf{R}|}\right)$$

que se distribuye como una $\chi^2_{(\nu)}$ con $\nu = \frac{1}{2}\left[(p-k)^2 - (p+k)\right]$. Entonces, rechazar $\mathbb{H}_0$ implica que el número de factores elegido no es suficiente para la descripción adecuada de la estructura de correlación, y hay necesidad de agregar más factores. Ya comentamos que esta prueba se basa en la normalidad multivariada de $\mathbf{X}$, que es difícil de cumplir, por lo que, en la mayoría de los casos, sólo se podrá usar como una referencia.

Puntajes Factoriales

Una vez que se ha estimado el modelo de factores propuesto, es necesario calcular los *puntajes factoriales* que le corresponden a cada individuo en cada uno de los factores. A este respecto existen principalmente dos métodos:

• **Método de Bartlett o mínimos cuadados ponderados**. El desarrollo de este método de construcción de puntajes es como sigue:

Generamos $\mathbf{Z}$ la matriz de datos estandarizados. Entonces, el modelo de factores se puede escribir en función de esta matriz, como

$$\mathbf{Z} = \Lambda\mathbf{F} + \mathbf{U}, \quad \text{con } \mathbb{E}(\mathbf{U}) = \mathbf{0}\, y\, \mathbb{V}(\mathbf{U}) = \Psi$$

De donde obtenemos

$$\mathbf{U}'\mathbf{U} = (\mathbf{Z} - \Lambda\mathbf{F})'(\mathbf{Z} - \Lambda\mathbf{F}) \quad \text{Mínimos cuadados o}$$
$$\mathbf{U}'\Psi^{-1}\mathbf{U} = (\mathbf{Z} - \Lambda\mathbf{F})'\Psi^{-1}(\mathbf{Z} - \Lambda\mathbf{F}) \quad \text{Mínimos cuadados ponderados}$$

con Ψ una matriz de pesos.

Bartlett sugiere encontrar f que minimice

$$\left(\mathbf{Z}_i - \hat{\Lambda} f\right)' \hat{\Psi}^{-1} \left(\mathbf{Z}_i - \hat{\Lambda} f\right), \quad i = 1, 2, ..., n$$

el valor f_i que minimiza esta expresión es

$$f_i = \left(\hat{\Lambda}' \hat{\Psi}^{-1} \hat{\Lambda}\right) \hat{\Lambda}' \hat{\Psi}^{-1} \mathbf{Z}_i$$

Entonces, se toma a f_i como el puntaje factorial del individuo *i, i=1,2,...,n.*

• **Método de Thompson o de regresión**.

Este método supone que tanto la matriz de datos $\mathbf{X}$ como los factores f son normales. Bajo estos supuestos, los puntajes factoriales se calculan como

$$\hat{f}_i = \hat{\Lambda}' \left(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}^{-1}\right)^{-1} \mathbf{Z}_i, \quad i = 1, 2, ..., n$$

Un concepto muy controversial: rotación de factores

Cuando el modelo en cuestión está determinado por un solo factor, su solución es única; sin embargo, las soluciones de los modelos multifactoriales, no son únicas y, como vimos, para lograr esta unicidad, se introduce una restricción esencialmente arbitraria, es decir, no es inherente al modelo. Entonces, diferentes tipos de "restricciones" pueden proporcionar soluciones diversas a este modelo de factores. Este aspecto ha suscitado críticas sobre el análisis factorial, ya que se piensa que depende de cuestiones subjetivas, que pudieran encaminar las soluciones a resultados preconcebidos por el investigador. Estas críticas son erróneas en dos aspectos: primero, el investigador *no obtiene la solución que él desea*; segundo, es más adecuado decir que la misma solución puede expresarse de diferentes maneras; de hecho, varias características de las soluciones, por ejemplo las comunalidades, permanecen inalteradas. *Rotación* es el nombre que se le da al proceso de cambiar de una solución a otra, y proviene de la representación geométrica de este procedimiento.

La razón principal para rotar una solución es clarificar la estructura de las cargas factoriales. Los factores deben tener un significado claro para el investigador, a partir del contexto de aplicación. Si la estructura que muestran las cargas factoriales de la solución inicial son confusas o difíciles de interpretar, una rotación puede proporcionar una estructura más fácil de interpretar.

Rotaciones ortogonales

Uno de los patrones de cargas factoriales más usuales y de hecho más deseables es la llamada *estructura simple de cargas factoriales*. Se dice que las cargas factoriales presentan una estructura simple si cada variable tiene una gran carga en un solo factor, con cargas cercanas a cero en el resto de los factores. Una de las rotaciones que procura generar una estructura de cargas simple son las *rotaciones ortogonales* (los nuevos ejes después de la rotación siguen siendo ortogonales). Existen varios métodos para realizar una rotación ortogonal, pero el más popular es la llamada *varimax*, implementada en la mayoría de los paquetes estadísticos. *Importante*: No hay garantía de que una rotación produzca necesariamente una estructura de cargas simple, pero, de hacerlo, puede ayudar a una interpretación mucho más fácil de los factores. Existen otras rotaciones ortogonales (como quartimax y equimax), pero ninguna tiene la popularidad de varimax.

Rotaciones oblicuas

Contrario a las rotaciones ortogonales, las rotaciones oblicuas permiten relajar la restricción de ortogonalidad con el fin de ganar simplicidad en la interpretación de los factores. Con este método los factores resultan correlacionados, aunque generalmente esta correlación es pequeña. El uso de rotaciones oblicuas se justifica porque en muchos contextos es lógico suponer que los factores están correlacionados. Pese a que pueden ser de utilidad en algunas situaciones, estas rotaciones raramente se usan, a diferencia de las ortogonales. Entre las rotaciones oblicuas, *promax* es conceptualmente simple; sin embargo, la más popular es *oblimin*.

Tipos de análisis factorial

Análisis factorial exploratorio

En muchas ocasiones no se tiene certeza sobre el número de factores, k, que subyacen en la estructura de datos; por ende, se puede realizar la extracción de factores de manera secuencial, se inicia con $k = 1$ y se llega hasta un número de factores que permita lograr un buen ajuste del modelo a los datos. Este procedimiento de incorporar factores hasta lograr un buen ajuste da lugar al llamado análisis factorial exploratorio, en el que el investigador no conoce de antemano el número de factores que subyacen en las variables observadas. Una desventaja de este tipo de análisis: puede ocurrir que los factores encontrados no tengan ninguna interpretación para el investigador, es decir, la estructura de cargas factoriales no sea interpretable por el investigador, para reconocer el constructo subyacente a este factor.

Análisis factorial confirmatorio

Por el contrario, cuando en una investigación se determina de forma precisa el número de factores, se está ante un *análisis factorial confirmatorio*. La forma usual de proponer este número de factores es en atención a alguna teoría propuesta en el área de aplicación. En este caso, los objetivos de la investigación se centran en la confirmación del número de factores y, consecuentemente, en la validación de esta teoría mediante la evidencia empírica proporcionada por los datos. Si el ajuste estadístico de los datos al modelo teórico es satisfactorio, se podrá concluir que el modelo es adecuado.

Entonces, cuando el análisis factorial es de tipo exploratorio, se tiene la necesidad de decidir cuántos factores se deben retener en el análisis. En seguida se enuncian algunos criterios establecidos para decidir este número.

Se pueden utilizar los mismos criterios que para componentes principales: *porcentaje de varianza explicada* y *gráfica de codo*, con uno más que es

El criterio del eigenvalor> 1

La lógica que sigue este criterio se basa en la idea de que cada uno de los factores extraídos debería justificar, al menos, la varianza de una variable individual (de lo contrario no se cumpliría con el objetivo de reducir la dimensión de los datos originales). En el contexto del

análisis factorial, los eigenvalores representan la cantidad de varianza de todas las variables medidas que puede ser explicada por un factor determinado. Cada una de las variables contribuye con un valor de 1 en el eigenvalor (varianza) total. Por lo tanto, de acuerdo con este criterio, deberían elegirse los factores con *eigenvalores mayores a 1* para garantizar que explican la varianza de al menos una variable.

Cómo determinar a priori si es conveniente llevar a cabo un análisis de factores

Al igual que en componentes principales, un elemento fundamental en análisis factorial es la fuerza de asociación de las variables medidas, que se manifiesta en la matriz de correlación. Entonces, veamos algunos estadísticos que nos pueden auxiliar en la determinación de si es conveniente o no llevar a cabo este análisis.

• **Determinante de la matriz de correlación**. Una medida global de la correlación entre todas las variables la proporciona el *determinante de la matriz de correlación*. Si este determinante está cercano a cero, será indicativo de que existe una estructura de correlación importante entre las variables, y el análisis factorial puede ser pertinente.

• **KMO (Medida de adecuación muestral)**. La llamada *medida de adecuación muestral* (Measure of Sampling Adequacy) está definida por:

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} r_{ij \bullet m}^2}$$

Esta prueba es un índice que compara los coeficientes de correlación r_{ij}^2 con los coeficientes de correlación parcial $r_{ij \bullet m}^2$. Esta última correlación es la correlación entre dos variables, eliminando el efecto de las restantes variables incluídas en el análisis. Entonces, si un par de variables está fuertemente correlacionada con el resto, la correlación parcial debe ser pequeña, ya que implica que buena parte de la correlación entre estas variables puede ser explicada por las otras variables en el análisis. Esto significa que está presente una fuerte estructura de correlación entre ellas y, por lo tanto, tiene sentido realizar el análisis de factores. En este caso, el denominador de KMO será cercano en magnitud al numerador, puesto que la contribución de las correlaciones parciales es prácticamente nula, y el índice KMO estará cercano a uno. Por el contrario, si esta correlación parcial es grande, implica que estas variables tienen poca correlación con el resto, lo que significa una estructura de correlación débil entre el conjunto, y hace cuestionable el análisis factorial. En este escenario, la contribución de las correlaciones parciales es importante, y el denominador será mucho mayor que el numerador, con KMO próximo a cero. Como regla empírica se considera que

si $KMO < 0.6$, es inadecuado realizar un análisis factorial a los datos.

La prueba de esfericidad de Bartlett

Si no hubiera estructura de correlación entre las variables involucradas en el análisis factorial, la matriz de correlación sería la matriz identidad, es decir, tendría ceros fuera de la diagonal (no habría correlación entre cualesquiera dos variables) y unos en la diagonal. Entonces, debemos probar, como parte fundamental para iniciar nuestro análisis factorial, que la matriz de correlaciones de nuestros datos es distinta de la identidad. A este respecto, la *prueba de esfericidad de Bartlett* contrasta la hipótesis nula de que la matriz de correlación es la identidad contra la hipótesis alternativa de que es distinta de la identidad. Desafortunadamente, esta prueba asume que las variables tienen una distribución normal multivariada, por lo que en muchas aplicaciones debe usarse únicamente como una referencia.

Análisis factorial con variables medidas en diversas escalas

Esta técnica, al igual que componentes principales, se presenta para datos medidos en escala continua; cuando las variables involucradas tengan otras escalas de medición, utilizaremos la matriz policórica para hacer este análisis.

Interpretación de la matriz de cargas factoriales

Una vez que se han estimado las cargas factoriales es importante establecer criterios que permitan interpretar los resultados obtenidos. Esta interpretación hará posible establecer una conexión entre los resultados vertidos por el análisis factorial y los constructos teóricos relacionados con los datos. En este sentido, la extracción de un determinado número de factores por los criterios estadísticos ya mencionados, carecerá de sentido si no podemos darle un significado lógico a cada uno de ellos, que además esté justificado teóricamente.

¿Cómo podemos determinar si una carga factorial es lo suficientemente "grande" para concluir que la correlación entre la variable y el factor es significativa? Hair *et al.* (1998-1999) proponen ciertas directrices para determinar si una carga factorial es o no significativa, dependiendo del tamaño de la muestra utilizada para el análisis (esta tabla se basa en estudios de potencia estadística).

Directrices

Carga Factorial	Tamaño de muestra necesario para la significancia
0.30	325
0.35	250
0.40	200
0.45	150
0.50	120
0.55	100
0.60	85
0.65	70
0.70	60
0.75	50

Estos cálculos están basados en un nivel de significancia de 0.05, una potencia de 80% y los errores estándar dos veces mayores que los coeficientes convencionales de correlación.

Análisis de Correspondencias

Introducción

El *análisis de correspondencias* es una técnica multivariada *exploratoria* para analizar tablas de frecuencias multidimensionales, esto es, tablas de clasificación cruzada de dos o más variables categóricas. El desarrollo del método de correspondencias se centra, generalmente, en las tablas de dos dimensiones; no obstante, el análisis de tablas multidimensionales depende en mucho de las mismas ideas que se desarrollan para las tablas bidimensionales. Esta técnica es el equivalente de componentes principales para variables cualitativas.

Entonces, por su similitud con el análisis de componentes principales, podemos decir que el análisis de correspondencias es una técnica para desplegar de forma gráfica datos categóricos multivariados (generalmente bidimensionales), derivando coordenadas para representar las categorías de las variables que constituyen los *renglones y columnas de una tabla de contingencia*, para plasmar gráficamente la asociación entre estas variables. Entonces, el análisis de correspondencias **(AC)** es:

• Técnica exploratoria para analizar tablas multidimensionales de contingencia o clasificación cruzada, entre dos o más variables categóricas.

• El objetivo es desplegar en una gráfica las asociaciones entre las categorías de una tabla de contingencia. Asociaciones tanto entre renglones, columnas y renglones y columnas. Para descubrir qué categorías están asociadas.

• Es una técnica de reducción de dimensión. Idealmente esperaríamos representar estas asociaciones entre columnas y renglones, en gráficas en dos o tres dimensiones, siempre que con estas pocas dimensiones, se logre una buena representación de ellas.

Análisis de tablas bidimensionales

En este caso, la información está constituída por una matriz de dimensiones $\mathbf{I} \times \mathbf{J}$, que representa las frecuencias absolutas observadas de dos variables cualitativas en una muestra de n elementos. La primera variable representa los renglones de esta tabla, que toma $\mathbf{I}$ valores posibles distintos, y la segunda representa las columnas, y toma $\mathbf{J}$ valores posibles distintos.

Ejemplo. Esta tabla presenta la clasificación de $n = 5387$ escolares escoceses por el color de sus ojos, con cuatro categorías posibles: $\mathbf{I} = 4$, y el color de su cabello, con cinco categorías: $\mathbf{J} = 5$. Esta tabla tiene interés histórico ya que fue utilizada por Fisher en *1940* para ilustrar un método de análisis de tablas de contingencia que está muy relacionado con el que aquí presentamos.

	Color de cabello					
Color de ojos	Rubio	Pelirrojo	Castaño	Obscuro	Negro	Total
claros	688	116	584	188	4	1580
azules	326	38	241	110	3	718
castaños	343	84	909	412	26	1774
oscuros	98	48	403	681	85	1315
Total	1455	286	2137	1391	118	5387

Conceptos asociados al análisis de correspondencias

Antes de presentar varios conceptos asociados al análisis de correspondencias, presentemos la forma general de una tabla de contingencia similar a la del ejemplo. Esta tabla tiene la forma

	Y: Variable columna				
X: Variable renglón	y_1	y_2	$\cdots$	y_J	Total
x_1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	$n_{I\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet J}$	$n_{\bullet\bullet} = n$

con n_{ij} el número de observaciones en el cruce de la categoría i de la variable $\mathbf{X}$ y la categoría j de la variable $\mathbf{Y}$. Además

$$n_{i\bullet} = \sum_{j=1}^{J} n_{ij}, \qquad n_{\bullet j} = \sum_{i=1}^{I} n_{ij} \quad y \quad n_{\bullet\bullet} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$$

Perfiles

El concepto de *perfil* (un vector de frecuencias relativas) es de suma importancia en *AC*. Estos vectores de frecuencias relativas tienen características especiales, debido al hecho de que la suma de sus elementos es *uno o el 100%*. En el análisis de estas tablas de frecuencias, consideraremos los vectores de frecuencias relativas *por renglón* y los vectores de frecuencias relativas *por columna*, que llamaremos *perfil renglón* y *perfil columna*, respectivamente. Cuyas definiciones, basándonos en la forma de la tabla general, son

Perfil Renglón

$$\left(\frac{n_{i1}}{n_{i\bullet}}, \frac{n_{i2}}{n_{i\bullet}}, ..., \frac{n_{iJ}}{n_{i\bullet}}\right) \text{ i=1,2,...,I.}$$

que corresponde a un vector de frecuencias relativas de las columnas, por categoría de renglón.

Con **perfil renglón promedio**, dado por

$$\frac{n_{\bullet j}}{n_{\bullet\bullet}} \text{ j=1,2,...,J}$$

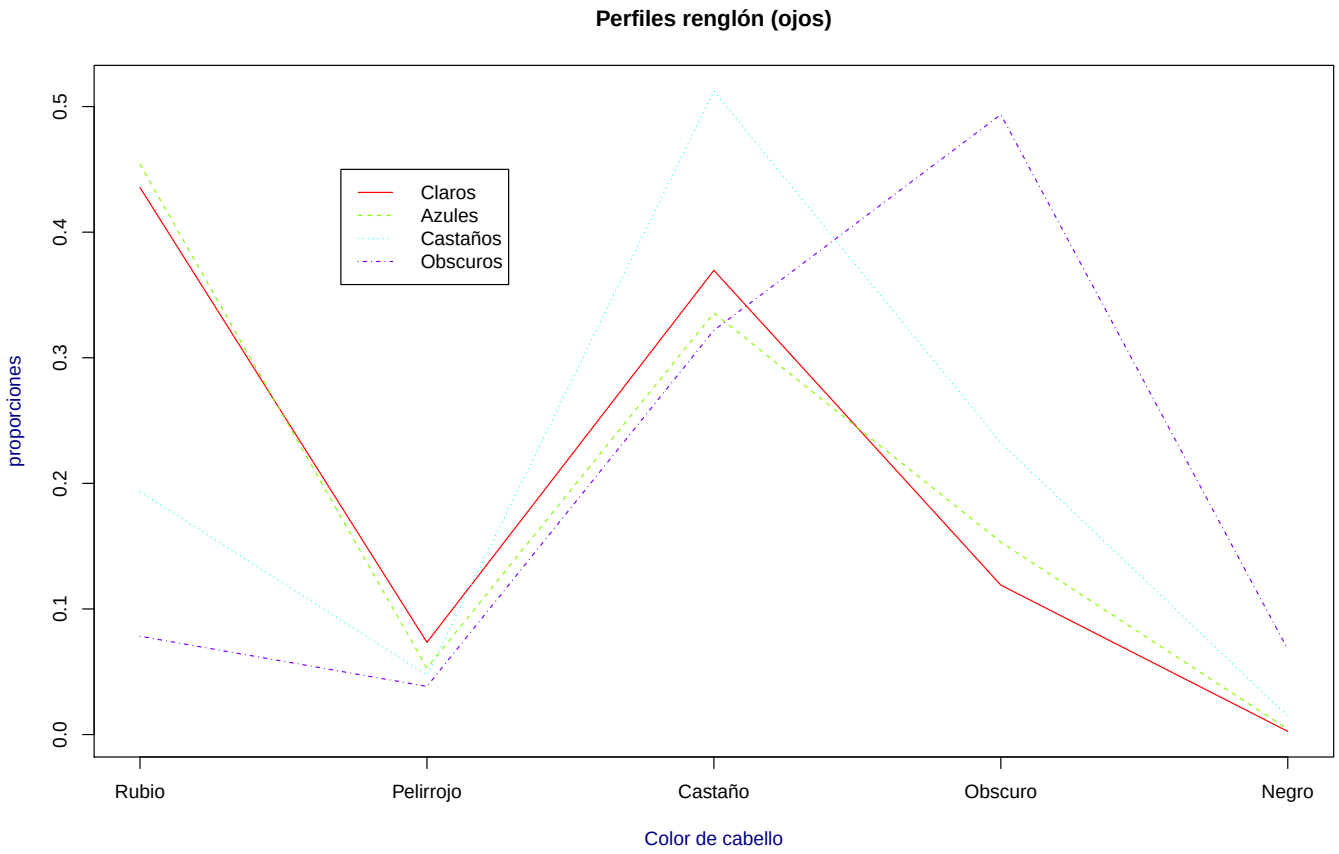
que corresponde al vector de frecuencias relativas del total por columna, entre el total de los datos.

La tabla de perfiles renglón para nuestro ejemplo es:

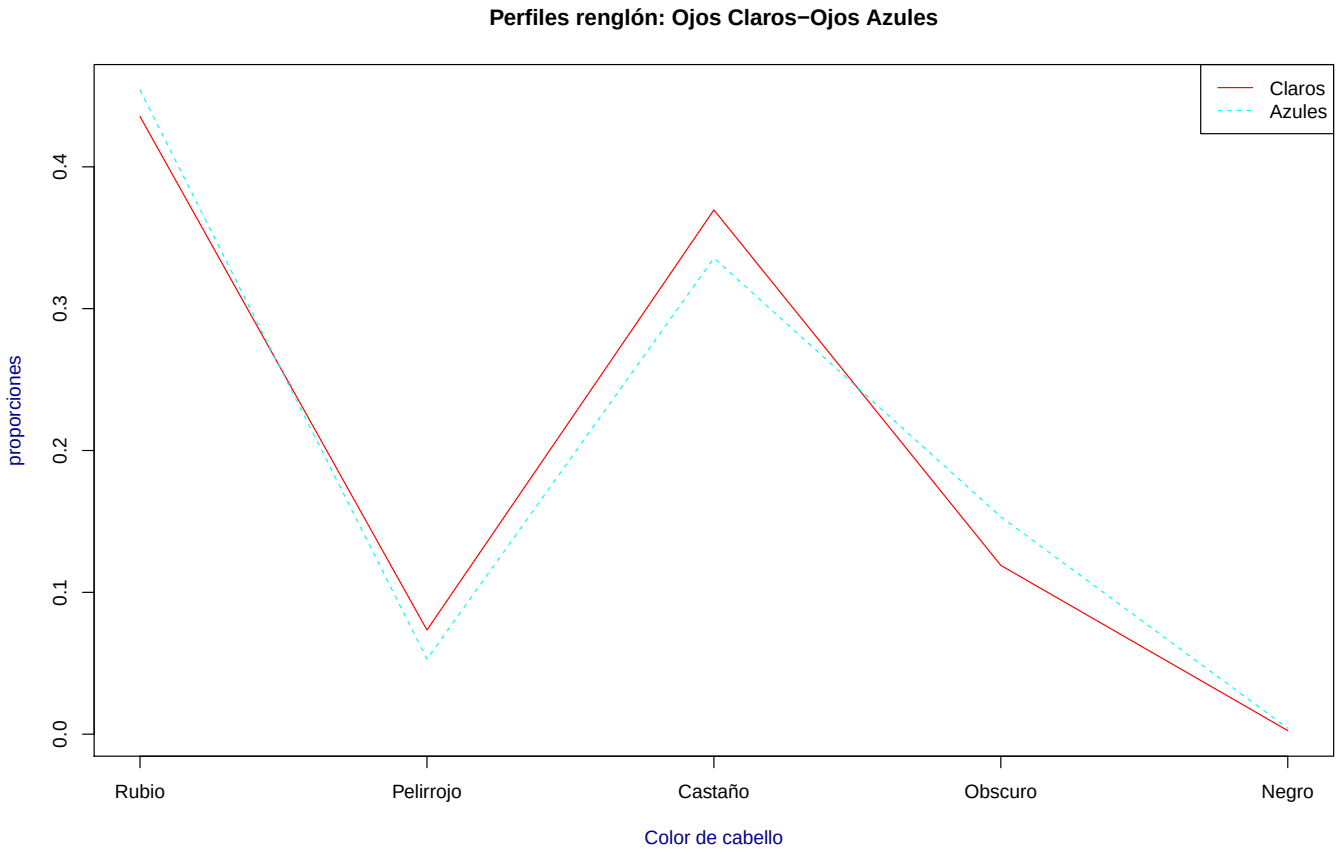
Tabla de Perfiles Renglón

	Color de cabello					
Color de ojos	Rubio	Pelirrojo	Castaño	Obscuro	Negro	Total
claros	0.435	0.073	0.370	0.119	0.003	1
azules	0.454	0.053	0.336	0.153	0.004	1
castaños	0.193	0.047	0.512	0.232	0.015	1
oscuros	0.078	0.038	0.322	0.494	0.068	1
Perfil renglón promedio	0.273	0.054	0.401	0.261	0.022	1
Total	1.161	0.212	1.539	0.998	0.0892	4

Observemos, por ejemplo, que el perfil del renglón: Ojos claros $(0.435, 0.073, 0.370, 0.119, 0.003)$, corresponde a $(688/1580, 116/1580, 584/1580, 188/1580, 4/1580)$, que son los valores observados en ese renglón, entre el total del mismo.



En esta gráfica de los perfiles renglón, podemos observar una gran similitud entre los perfiles de los sujetos de ojos claros y los de ojos azules.



Perfil columna

$$\left(\frac{n_{1j}}{n_{\bullet j}}, \frac{n_{2j}}{n_{\bullet j}}, ..., \frac{n_{2I}}{n_{\bullet j}}\right) \text{ j=1,2,...,J.}$$

que corresponde al vector de frecuencias relativas de los renglones, por categoría de columna.

Con **perfil columna promedio**

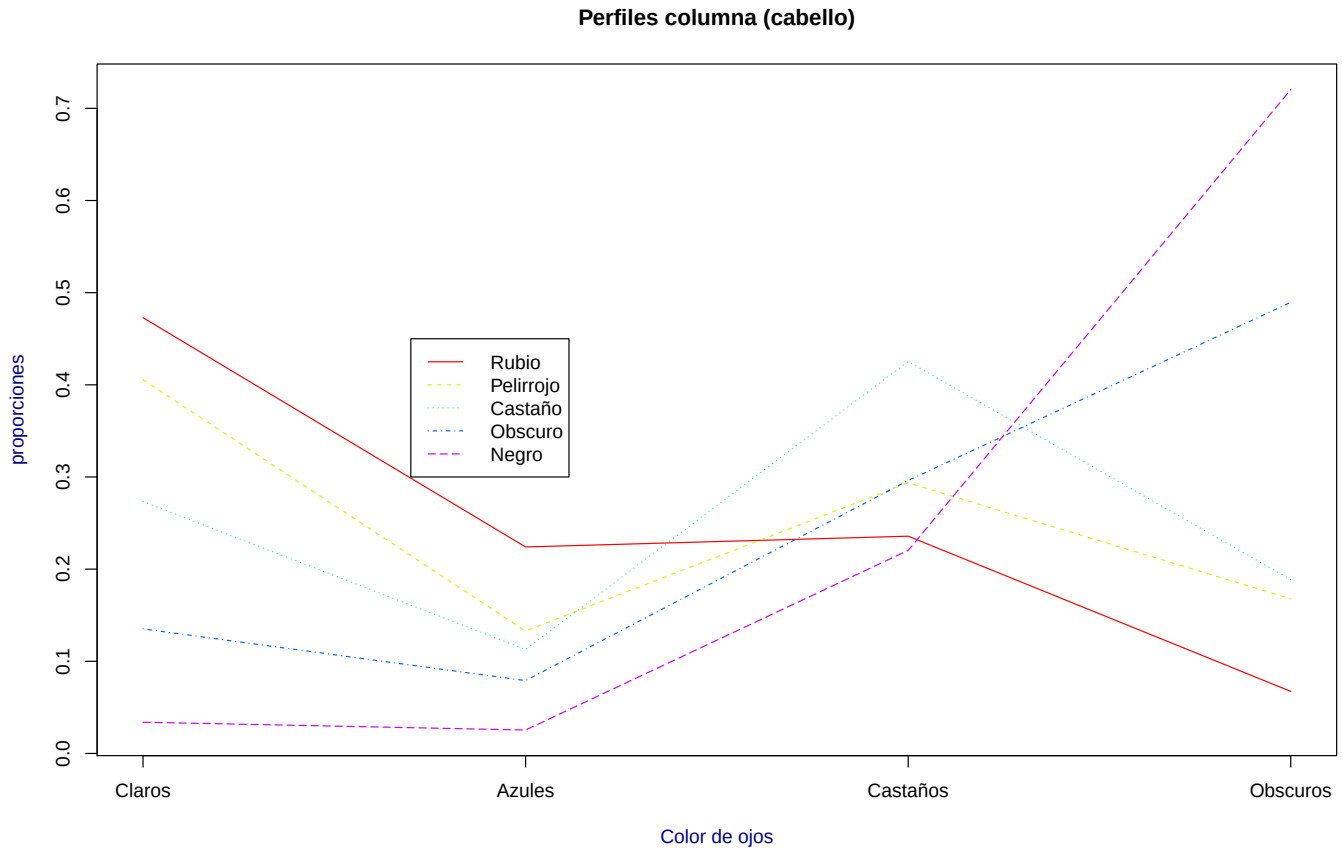
$$\frac{n_{i\bullet}}{n_{\bullet\bullet}} \text{ i=1,2,...,I}$$

que corresponde al vector de frecuencias relativas del total por renglón, entre el total de los datos.

La tabla de perfiles columna para nuestro ejemplo es:

Tabla de Perfiles Columna

	Color de cabello						
Color de ojos	Rubio	Pelirrojo	Castaño	Obscuro	Negro	Perfil columna promedio	Total
Claros	0.473	0.406	0.273	0.142	0.034	1.327	0.2967693
Azules	0.224	0.133	0.113	0.083	0.025	0.578	0.1348610
Castaños	0.236	0.294	0.425	0.310	0.220	1.485	0.3332081
Obscuros	0.067	0.168	0.189	0.465	0.720	1.609	0.2351615
Total	1	1	1	1	1	4	



Masa

En el cálculo usual de la media (no ponderada), todos los puntos tienen la misma masa (o peso). Sin embargo, una media ponderada permite asociar diferentes masas a los diferentes valores (puntos) que la conforman. Cuando ponderamos estos valores de distinta forma, el centroide no se sitúa exactamente en el centro "geográfico" de la nube de puntos, sino que tiende a situarse cerca de los puntos con mayor masa.

Por ejemplo, supongamos que en una clase de 30 estudiantes la media de calificaciones, calculada sumando sus calificaciones y dividiendo entre 30, es 7.43. Se sabe que tres estudiantes

obtuvieron 9 de calificación, siete obtuvieron 8 y 20 obtuvieron 7, entonces, podemos calcular la media de manera equivalente, asignando un peso de 3/30 a la calificación de 9, 7/30 a la de 8 y 20/30 a la de 7. Dado que la calificación de 7 tiene mayor peso que las otras, el valor de la media ponderada, 7.43, se encuentra "más cerca" de esta calificación. La media aritmética usual de los valores 7, 8 y 9 es 8.

En el AC, los pesos asignados a los perfiles reciben el nombre de *masas*. Los totales de las columnas, con relación al total de la tabla, son las masas de las columnas que asignaremos a los perfiles columna. En símbolos

$$m_j = \frac{n_{\bullet j}}{n_{\bullet\bullet}} \quad j = 1, 2, ..., J$$

El perfil columna promedio, está constituido por los totales de las columnas, divididos entre el total de la tabla.

De manera semejante, los totales de los renglones, con relación al total de la tabla, son las masas de los renglones que asignaremos a los perfiles renglón.

$$m_i = \frac{n_{i\bullet}}{n_{\bullet\bullet}} \quad i = 1, 2, ..., I$$

El perfil renglón promedio, está constituido por los totales de los renglones, divididos entre el total de la tabla.

Entonces, un perfil con mayor masa se hubicará más cercano a su correspondiente perfil columna promedio o perfil renglón promedio, según corresponda a un perfil columna o renglón.

Distancias entre perfiles

Una forma de saber qué tan parecido es un renglón con respecto a otro, o una columna con respecto a otra, es a través de la distancia entre sus correspondientes perfiles.

La distancia χ^2

En AC la forma de calcular la distancia entre perfiles es un poco complicada, y se realiza a través de la llamada *distancia* χ^2. Existen varias maneras de justificar el uso de esta

distancia, algunas más técnicas que no viene al caso considerar, y otras más intuitivas que utilizaremos aquí.

Recordemos que la estadística χ^2 asociada a estas tablas de contingencia tiene la forma

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \mathbb{E}_{ij})^2}{\mathbb{E}_{ij}}$$

Reflexionemos un poco sobre esta estadística

Mencionamos que el objetivo en este análisis de correspondencias es determinar las asociaciones entre los distintos elementos de una tabla de contingencia, a saber, entre renglones, entre columnas y entre renglones y columnas. Recordemos que uno de los usos de la estadística χ^2 es probar la asociación *a nivel global* entre dos variables categóricas. Al comparar los valores observados en la tabla contra los esperados (estos últimos se calculan bajo el supuesto de que las variables involucradas *son independientes*), entonces, es claro que la discrepancia entre los valores observados y esperados

$$n_{ij} - \mathbb{E}_{ij}, \;\; i = 1, 2, ..., I \;\; j = 1, 2, ..., J$$

es una cantidad que evidencia el grado de *no inpendencia* entre las celdas correspondientes, i.e., es, en algún sentido, una medida del grado de asoción entre estas celdas, lo que conduce a que el valor de esta χ^2 sea una medida del nivel de asociacion global de las variables. Recordemos también que, bajo el supuesto de indpendencia entre las variables de la tabla, los valores esperados se calculan como

$$\mathbb{E}_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n_{\bullet\bullet}}, \;\; i = 1, 2, ..., I \;\; j = 1, 2, ..., J$$

Retomando la expresión de esta estadística, observemos que la podemos escribir como

$$
\begin{aligned}
\chi^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - \dfrac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}\right)^2}{\dfrac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}} \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{i\bullet}\left(\dfrac{n_{ij}}{n_{i\bullet}} - \dfrac{n_{\bullet j}}{n_{\bullet\bullet}}\right)\right)^2}{\dfrac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}} \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{i\bullet}^2\left(\dfrac{n_{ij}}{n_{i\bullet}} - \dfrac{n_{\bullet j}}{n_{\bullet\bullet}}\right)^2}{\dfrac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}} \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{i\bullet}\left(\dfrac{n_{ij}}{n_{i\bullet}} - \dfrac{n_{\bullet j}}{n_{\bullet\bullet}}\right)^2}{\dfrac{n_{\bullet j}}{n_{\bullet\bullet}}}
\end{aligned}
$$

Ahora, obsérvese que

$\dfrac{n_{ij}}{n_{i\bullet}}$ es el perfil renglón para $i = 1, 2, ..., I$ y

$\dfrac{n_{\bullet j}}{n_{\bullet\bullet}}$ el perfil renglón promedio o perfil esperado.

Por lo tanto, esta χ^2 la podemos reescribir, como

$$
\sum_i \text{total renglón } i \times \frac{(\text{perfil observado renglón } i\text{-perfil esperado renglón } i)^2}{\text{perfil esperado renglón } i}
$$

justamente como una distancia entre los perfiles renglón y su perfil promedio.

De manera similar, si en el segundo paso del desarrollo anterior, factorizamos $n_{\bullet j}$, en lugar de $n_{i\bullet}$, obtenemos

$$
\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{\bullet j} \frac{\left(\dfrac{n_{ij}}{n_{\bullet j}} - \dfrac{n_{i\bullet}}{n_{\bullet\bullet}}\right)^2}{\dfrac{n_{i\bullet}}{n_{\bullet\bullet}}}
$$

Que podemos reescribir, como

$$\sum_j \text{total columna } j \times \frac{(\text{perfil observado columna } j\text{-perfil esperado columna } j)^2}{\text{perfil esperado columna } j}$$

justamente como una distancia entre los perfiles columna y su perfil promedio.

Inercia

Otro de los conceptos importantes, de hecho, muy importante, en AC es el de la *inercia*. Desde el punto de vista de la Física, en particular de la mecánica, que es de donde se traslada este concepto, se tiene que cualquier objeto tiene un centro de gravedad (centroide); cualquier partícula en el objeto tiene cierta masa y cierta distancia al centroide; entonces, el momento de inercia está dado por $I = md^2$ sumado sobre todas las partículas que constituyen el objeto. Es decir:

$$\mathbf{I} = \sum \mathbf{md^2}$$

El concepto análogo en AC consiste en considerar a los puntos perfiles, cuya masa suma uno. Estos puntos tienen un centroide (su perfil promedio) y una distancia (distancia ji-cuadrada) entre puntos perfiles. Cada punto en un perfil contribuye a la inercia en la nube total de puntos.

Ahora, hagamos la deducción analítica de este concepto partiendo de la expresión de la distancia χ^2 escrita como distancias entre perfiles. En esta expresión, dividamos ambos lados de la igualdad entre el total de la muestra, $n_{\bullet\bullet} = n$, con lo que obtenemos

$$\frac{\chi^2}{n} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{i\bullet}}{n} \frac{\left(\frac{n_{ij}}{n_{i\bullet}} - \frac{n_{\bullet j}}{n_{\bullet\bullet}}\right)^2}{\frac{n_{\bullet j}}{n_{\bullet\bullet}}}$$

obsérvese que $\frac{n_{i\bullet}}{n}$ corresponde a la masa de cada perfil renglón; entonces tenemos una expresión semejante a la que define la inercia. En este caso *la masa es la masa de cada perfil*, *el centroide es el perfil promedio o medio* y *la distancia del perfil a este centroide está dada*

por la distancia χ^2. En AC a esta cantidad se le conoce con el nombre de *inercia* o *inercia total*.

Dada esta relación

$$\frac{\chi^2}{n} = \mathbf{I}$$

es claro que la inercia es una medida de la varianza o variabilidad de nuestros datos. De hecho, sabemos que es una medida de la asociación entre los categorías en la tabla. Además, al escribirla como una medida de la discrepancia entre un perfil y su perfil medio, también es una medida de qué tan "lejos" se hallan los perfiles renglón o columna de su perfil medio. Podemos considerar que este pefil medio representa la hipótesis de homogeneidad, en este caso, de homogeneidad entre los perfiles. Entonces, debe ser claro que si los perfiles difieren poco de sus perfiles medios, el valor de la inercia sería bajo, e implicaría una pobre asociación entre las variables, así como entre los renglones, columnas, y renglones columnas de la tabla de contingencia.

Reducción de dimensión

La dimensión natural de una tabla de contingencia de $\mathbf{I} \times \mathbf{J}$ es $min\,(\mathbf{I}-1, \mathbf{J}-1)$, así que si la menor de estas dimensiones es grande, entonces debemos hacer una reducción de dimensión, idealmente a *dos* o *tres* dimensiones para poder representarlas gráficamente, de manera que la varianza explicada, en este caso la *inercia explicada*, por esas pocas dimensiones sea cercana al 100%.

Descomposición en valor singular

En muchas de las técnicas multivariadas, la información se concentra en la matriz de varianza-covarianza o de correlación, ¿existe un concepto semejante en AC?

Matriz asociada a la χ^2. Ya vimos que la información relevante para las asociaciones de las categorías de la tabla, la proporcionan las diferencias entre los perfiles observados y los perfiles medios, estos últimos asumiendo que los perfiles medios representan la homogeneidad entre los perfiles correspondientes. Equivalentemente, esta matriz también representa las asociaciones de las categorías de la tabla, medidas a través de la diferencia entre los valores esperados, que se obtienen bajo el supuesto de independencia de esta tabla, y los valores observados. Entonces, consideremos la matriz

$$\mathbf{S} = (s_{ij}) \qquad \text{con } s_{ij} = \frac{n_{ij} - \mathbb{E}_{ij}}{\sqrt{\mathbb{E}_{ij}}} = \frac{n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}}{\sqrt{\frac{n_{i\bullet} n_{\bullet j}}{n}}} \quad i = 1, 2, ..., I \;\; J = 1, 2, ..., J$$

que constituye nuestra "matriz de correlación".

Por lo que, nuestro objetivo de representar gráficamente las asociaciones, se convierte en el de encontrar un número reducido de dimensiones (idealmente 2 ó 3) donde se puedan representar estas desviaciones expresadas en "nuestra matriz de correlación". En este sentido, este objetivo es similar al de componentes principales, factores, etc. Para lograrlo, debemos hacer la *descomposición en valor singular*, de esta matriz.

$$\mathbf{S} = \mathbf{ULA}^{'} \text{ con } \mathbf{G} = \mathbf{UL} \;\; \mathbf{y} \;\; \mathbf{H} = \mathbf{A}'$$

entonces, los elementos de $\mathbf{S}$ se pueden escribir como

$$s_{ij} = \sum_{k=1}^{R} \lambda_k^{1/2} g_{ik} h_{ik}, \quad i = 1, 2, ..., I \;\; j = 1, 2, ..., J$$

con $R = min(I-1, J-1)$, el número máximo de dimensiones que se pueden tener, y que corresponde al rango de $\mathbf{S}$. Los valores g_{ik} y h_{ik} son los elementos de la k-ésima columna de $\mathbf{G}$ y $\mathbf{H}$, respectivamente. $\lambda_1, \lambda_2, ..., \lambda_R$ son los eigenvalores de $\mathbf{S}$, que constituyen la matriz $\mathbf{L}$.

Obsérvese que

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} s_{ij}^2 = \mathbf{SS}' = \mathbf{ULA}' \left(\mathbf{ULA}'\right)' = \mathbf{ULA}'\mathbf{AL}'\mathbf{U}' = \mathbf{ULL}'\mathbf{U}' = \mathbf{L^2UU}' = \mathbf{L^2}$$

$$\Rightarrow \; traza\left(\mathbf{SS}'\right) = traza(\mathbf{LL}') = traza(\mathbf{L^2}) = \sum_{k=1}^{R} \lambda_k$$

por lo tanto, la variabilidad de la matriz asociada a la tabla de contingencia, es igual a la suma de sus eigenvalores. Con lo que tenemos una analogía completa con el proceso de componentes principales.

Entonces, lo que deseamos es poder representar estos elementos de $\mathbf{S}$ en pocas dimensiones, pero asegurándonos de que esta representación es buena, en algún sentido. Si queremos una representación bidimensional, entonces

$$s_{ij} \approx \sum_{k=1}^{2} \lambda_k^{1/2} g_{ik} h_{ik} \quad i = 1, 2, ..., I \;\; j = 1, 2, ..., J$$

¿Y cómo determinamos si, en este caso, dos dimensiones proporcionan una buena aproximación de los elementos de esta matriz?

Similar al criterio de componentes principales, consideramos el porcentaje de inercia que explican estas dos dimensiones

$$\frac{\sum_{k=1}^{2} \lambda_k}{\sum_{k=1}^{R} \lambda_k} \times 100\%$$

debe ser grande, idealmente cercano al 100%.

Representación bidimensional

En la práctica, es común representar el AC en una gráfica bidimensional. Aunque se pueden proyectar los datos en cualquier subespacio de dimensión menor, las proyecciones bidimensionales son particularmente atractivas, ya que representan nuestra forma habitual de representar una gráfica.
En este caso, cada categoría de un renglón estará representada por un par coordenado (g_{i1}, g_{i2}) $i = 1, 2, ..., I$. Y cada categoría de una columna, por el par coordenado (h_{j1}, h_{j2}) $j = 1, 2, ..., J$. Por lo que, para representarlos hay que desplegarlos en una gráfica bidimensional.

Representaciones bidimensionales

- Representación bidimensional de los renglones

- Representación bidimensional de las columnas

- Representación bidimensional de renglones y columnas (generalmente la más usual. *Biplot*).

Interpretación del AC

Dimensiones: Algunas veces es posible interpretar o "dar nombre" a las dimensiones que se obtienen a través del **AC**. Podemos examinar la posición de las categorías renglones/columnas en cada dimensión y analizar qué tienen en común las categorías de estos renglones/columnas que aparecen juntas, y qué distingue a aquéllas que aparecen separadas. Sin embargo, cuando se interpreta una dimensión, es importante prestar particular atención a aquellos puntos que contribuyen más a la inercia de cada dimensión.

Podemos particionar la contribución de cada punto a la inercia total, en su contribución a la inercia de cada dimensión. La cantidad de inercia de la k-ésima dimensión explicada por el renglón i es

$$\frac{(masa\ del\ \text{renglón}\ i) * g_{ik}^2}{\sqrt{\lambda_k}} \quad \text{equivalentemente} \quad \frac{(masa\ de\ la\ columna\ j) * h_{ik}^2}{\sqrt{\lambda_k}}$$

entonces, puntos correspondientes a renglones con una gran masa y grandes coordenadas en

la k-ésima dimensión, contribuirán más a la inercia de esta dimensión. Puntos con una relativa mayor contribución a la inercia de una dimensión, son más importantes para la misma y proporcionan la clave para su correspondiente interpretación.

Puntos

Como debe ser obvio, categorías de renglones que tienen un perfil similar deben aparecer cercanos en la representación bidimensional. Misma situación para categorías de columnas.

Asociación entre renglones y columnas: En este caso, la distancia entre una categoría renglón y otra columna, no representan ninguna similitud de los mismos. Para interpretar puntos de distintas naturalezas, se recurre a su llamada representación en *biplot*.

Como sabemos, el *biplot* se basa en el producto escalar entre los vectores columna y renglón, por lo que depende más de las longitudes y ángulos formados por estos vectores que de la distancia entre los puntos.

Geométricamente, el producto escalar entre vectores es igual al producto de las longitudes de los vectores multiplicado por el coseno del ángulo formado entre ellos, es decir

$$\mathbf{x}'\mathbf{y} = \sum_i x_i y_i = \|\mathbf{x}\| \, \|\mathbf{y}\| \, cos\,(\theta)$$

Recordemos también que la proyección perpendicular de un vector $\mathbf{x}$ sobre la dirección definida por otro vector $\mathbf{y}$, tiene una longitud igual al producto de la longitud de $\mathbf{x}$, multiplicada por el ángulo que forman estos vectores. En concreto

$$Proy\; x_y = \|x\| \, cos\,(\theta)$$

Entonces, en el AC, la idea es que a través de los productos escalares del biplot, podamos recuperar de manera aproximada, los elementos de la matriz dada por la tabla de contingencia o clasificación cruzada.

En la palabra biplot, el prefijo *bi* indica que en el mapa se representan conjuntamente renglones y columnas, pero no indica que el mapa sea bidimensional, ya que los biplots pueden tener cualquier dimensionalidad. No obstante, lo más frecuente es una representación en dos

dimensiones.

Recordemos que $g_{i2} = (g^*_{i1}, g^*_{i2})$ es un punto en dimensión dos que representa el i-ésimo renglón (i=1,2,...,I), y $h_{i2} = (h^*_{i1}, h^*_{i2})$ es un punto bidimensional que representa a la j-ésima columna (j=1,2,...,J). Utilizando los conceptos asociados al biplot, el producto $g^*_{ik} \times h^*_{ik}$ representa la contribución conjunta del renglón i y la columna j al residuo (que dijimos que era una medida de asociación) en la dimensión k, es decir, la "asociación" entre el renglón i y la columna j. O, de forma más precisa, la contribución del renglón i y la columna j, a la asociación global medida por la χ^2.

En este sentido, un valor grande y positivo de $g^*_{ik} \times h^*_{ik}$ indica una asociación positiva entre el renglón i y la columna j en la dimensión k, misma que se obtiene si ambos son grandes y positivos, o grandes y negativos.

Si $g^*_{ik} \times h^*_{ik}$ proporciona un valor grande y negativo, implica una asociación negativa entre el rengón i y la columna j en la dimensión k, misma que se obtiene si ambos son grandes y uno es positivo y el otro negativo.

Un valor cercano a cero de $g^*_{ik} \times h^*_{ik}$ indica que no hay asociación entre el renglón i y la columna j en la dimensión k, que se obtiene si alguno o ambos están cercanos a cero en esa dimensión.

Calidad de la representación de un punto. Dado que se han elegido un número reducido de dimensiones para representar un punto, una pregunta de interés es saber qué calidad de representación tiene cada punto en estas pocas dimensiones. La calidad de esta representación se mide a través del cociente entre la distancia al origen del punto en las dimensiones elegidas, y su distancia al origen en el máximo de dimensiones posibles ($min(I-1, J-1)$). Si un punto tiene baja calidad, implica que su representación en este espacio reducido, no es adecuada.

Correspondencias múltiples

El *análisis de correspondencias múltiples (MCA)*, es una extensión del *AC* que permite analizar las asociaciones entre más de dos variables categóricas.

Un punto importante en *MCA* es la forma en que se debe manejar la información de una tabla de frecuencias multidimensional. La manera de hacerlo es a través de la llamada *matriz indicadora* o *matriz disjunta* que no es mas que una matriz cuyos valores son sólo "0" ó "1". Entonces, el *MCA* consiste en analizar una serie de observaciones descritas por un conjunto de *variables nominales* o *variables dummy's*.

Ejemplo: Enfermedad de Hodgkin. En este caso sólo tenemos una tabla bidimensional, pero servirá para ilustrar la construcción de la matriz indicadora o disjunta.

Enfermedad de Hodgkin

	Tipo de respuesta			
Tipo histológico	Positiva	Parcial	Nula	Total
LP	74	18	12	104
NS	68	16	12	96
MC	154	54	58	266
LD	18	10	44	72
Total	314	98	126	538

La matriz indicadora correspondiente

Matriz indicadora: Enfermedad de Hodgkin

	Tipo histológico				Respuesta		
Sujeto	LP	NS	MC	LD	Positiva	Parcial	Nula
1	1	0	0	0	1	0	0
2	1	0	0	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
74	1	0	0	0	1	0	0
75	1	0	0	0	0	1	0
76	1	0	0	0	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
92	1	0	0	0	0	1	0
93	1	0	0	0	0	0	1
94	1	0	0	0	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Extensión a tres variables categóricas. Si aumentamos un criterio de clasificación referente, por ejemplo, al género, la tabla se ampliaría de la siguiente manera

Matriz indicadora: Enfermedad de Hodgkin con tres variables categóricas

	Tipo histológico				Respuesta			Género	
Sujeto	LP	NS	MC	LD	Positiva	Parcial	Nula	F	M
1	1	0	0	0	1	0	0	1	0
2	1	0	0	0	1	0	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
74	1	0	0	0	1	0	0	0	1
75	1	0	0	0	0	1	0	1	0
76	1	0	0	0	0	1	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
92	1	0	0	0	0	1	0	0	1
93	1	0	0	0	0	0	1	1	0
94	1	0	0	0	0	0	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Entonces, el punto inicial del análisis de correspondencias múltiples es construir la matriz indicadora, $\mathbf{Z}$. Cada renglón de esta matriz tiene k *"unos"* y *C-k "ceros"*, donde k es el número de variables categóricas en cuestión, y C es el total de categorías de estas k variables, esto es

$$C = \sum_{i=1}^{k} c_i$$

con c_i el número de categorías de la *i-ésima variable*. Por lo tanto, la matriz indicadora, $\mathbf{Z}$, es de tamaño *(n,k)*, donde k es el número total de variables. Las suma de cada una de sus filas es igual a k, el número de variables, y la suma de cada columna es el número de individuos que que tiene la característica en cuestión.

Para una tabla de contingencia con k variables, la matriz indicadora puede escribirse como:

$$\mathbf{Z} = [\mathbf{Z}_1 \ \mathbf{Z}_2 \ \cdots \ \mathbf{Z}_k]$$

con $\mathbf{Z}_i$ es la matriz de $n_i \times c_i$ de la i-ésima tabla de contingencia.

La matriz de Burt

La matriz

$$\mathbf{B} = \mathbf{Z}'\mathbf{Z}$$

se conoce como la *matriz de Burt*, y contiene las submatrices $\mathbf{Z}_i'\mathbf{Z}_j$ de la tabla de contingencia bidimensional, basada en las variables i y j. Esto es

$$\mathbf{B} = \begin{bmatrix} \mathbf{Z}_1'\mathbf{Z}_1 & \mathbf{Z}_1'\mathbf{Z}_2 & \cdots & \mathbf{Z}_1'\mathbf{Z}_k \\ \mathbf{Z}_2'\mathbf{Z}_1 & \mathbf{Z}_2'\mathbf{Z}_2 & \cdots & \mathbf{Z}_2'\mathbf{Z}_k \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{Z}_k'\mathbf{Z}_1 & \mathbf{Z}_k'\mathbf{Z}_2 & \cdots & \mathbf{Z}_k'\mathbf{Z}_k \end{bmatrix}$$

Entonces, el análisis de correspondencias múltiple consiste, esencialmente, en aplicar todos los procesos de correspondencias simples, ya sea a la matriz indicadora o a la matriz de Burt.

La matriz o tabla **B** es simétrica y está conformada por $k \times k$ subtablas. Las k subtablas diagonales son a su vez diagonales y contienen las frecuencias marginales de cada una de las variables. Las subtablas fuera de esta diagonal, son las tablas de contingencia entre parejas de variables.

Ejemplo de una matriz de Burt con tres variables categóricas

Matriz de Burt

	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	C1	C2	C3	C4	C5
A1	119	0	0	0	0	27	28	30	22	12	49	40	18	7	5
A2	0	322	0	0	0	38	74	84	96	30	67	142	60	41	12
A3	0	0	204	0	0	3	48	63	73	17	18	75	70	34	7
A4	0	0	0	178	0	3	21	23	79	52	16	50	40	56	16
A5	0	0	0	0	48	0	3	5	11	29	2	9	9	16	12
B1	27	38	3	3	0	71	0	0	0	0	43	19	4	3	2
B2	28	74	48	21	3	0	174	0	0	0	36	88	34	15	1
B3	30	84	63	23	5	0	0	205	0	0	37	90	57	19	2
B4	22	96	73	79	11	0	0	0	281	0	27	88	75	74	17
B5	12	30	17	52	29	0	0	0	0	140	9	31	27	43	30
C1	49	67	18	16	2	43	36	37	27	9	152	0	0	0	0
C2	40	142	75	50	9	19	88	90	88	31	0	316	0	0	0
C3	18	60	70	40	9	4	34	57	75	27	0	0	197	0	0
C4	7	41	34	56	16	3	15	19	74	43	0	0	0	154	0
C5	5	12	7	16	12	2	1	2	17	30	0	0	0	0	52

Algunas características de estas matrices

Entonces, las características de estas matrices son:

- La matriz es $\mathbf{Z} = (z_{ij})$ con $z_{ij} = \begin{cases} 1 \\ 0 \end{cases}$
- $\sum_i \sum_j z_{ij} = nk.$
- n: número de individuos.
- k: número de variables categóricas.

- c_i: número de categorías de la variable i, i=1,2,...,k.

- C=$\sum_{i=1}^{k} c_i$: total de categorías.

- *Marginales*: $z_{i\bullet} = k$ (puesto que hay k variables, y por lo tanto, k *uno's* por renglón).

- *Marginales*: $z_{\bullet j} =$ individuos que tienen la característica j.

- *Perfil renglón*: $\frac{z_{ij}}{z_{i\bullet}} = \frac{z_{ij}}{k}$.

- *Masa del perfil renglón*: $\frac{z_{i\bullet}}{nk} = \frac{k}{nk} = \frac{1}{n}$

- *Perfil columna*:$\frac{z_{ij}}{z_{\bullet j}}$

- *Masa del perfil columna*: $\frac{z_{\bullet j}}{nk}$

Una manera de definir la distancia χ^2 entre los perfiles es a través de las frecuencias relativas de la tabla, de la siguiente manera

$$d^2\left(i, i^{'}\right) = \sum_j \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i^{'}j}}{f_{i^{'}\bullet}}\right)^2$$

$$d^2\left(j, j^{'}\right) = \sum_i \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - \frac{f_{ij^{'}}}{f_{\bullet j^{'}}}\right)^2$$

en este caso, con $f_{ij} = \frac{z_{ij}}{nk}$, $f_{i\bullet} = \frac{z_{i\bullet}}{nk}$ y $f_{j\bullet} = \frac{z_{\bullet j}}{nk}$

Observemos que se pondera las diferencias cuadradas de los perfiles renglón o columna, por el inverso de su frecuencia, lo que hace que perfiles con poca frecuencia, contribuyan de manera similar a los que tienen mayor frecuencia en la tabla. De hecho, lo único que se está

haciendo es dotar de una métrica distinta a estas distancias.

Entonces, en este caso de correspondencias múltiples, estas distancias son

- $d^2\left(i,i^{'}\right)=\sum_j \frac{nk}{z_{\bullet j}}\left(\frac{z_{ij}}{z_{i\bullet}}-\frac{z_{i^{'}j}}{z_{i^{'}\bullet}}\right)^2=\sum_j \frac{nk}{z_{\bullet j}}\left(\frac{z_{ij}}{k}-\frac{z_{i^{'}j}}{k}\right)^2=\frac{n}{k}\sum_j \frac{1}{z_{\bullet j}}\left(z_{ij}-z_{i^{'}j}\right)^2$

- $d^2\left(j,j^{'}\right)=\sum_i \frac{nk}{z_{i\bullet}}\left(\frac{z_{ij}}{z_{\bullet j}}-\frac{z_{ij^{'}}}{z_{\bullet j}}\right)^2=\sum_i \frac{nk}{k}\left(\frac{z_{ij}}{z_{\bullet j}}-\frac{z_{ij^{'}}}{z_{\bullet j}}\right)^2=n\sum_i \left(\frac{z_{ij}}{z_{\bullet j}}-\frac{z_{ij^{'}}}{z_{\bullet j}}\right)^2$

Análisis de estas distancias

Podemos reescribir la distancia entre los perfiles renglón de la siguiente manera:

$$\frac{n}{k}\sum_j \frac{1}{z_{\bullet j}}\left(z_{ij}-z_{i^{'}j}\right)^2=\frac{n}{k}\sum_{j\in M_{ii^{'}}}\frac{1}{z_{\bullet j}}$$

con $M_{ii^{'}}$ modalidades que poseen sólo un individuo i ó $i^{'}$. Entonces, los perfiles serán más parecidos (distancia más pequeña), conforme poseean más modalidades en común.

Perfiles columna

$$n\sum_i \left(\frac{z_{ij}}{z_{\bullet j}}-\frac{z_{ij^{'}}}{z_{\bullet j^{'}}}\right)^2=n\frac{\#\left(ind[j,no\ j^{'}]\right)\#\left(ind[j^{'},no\ j]\right)}{z_{\bullet j}z_{\bullet j^{'}}}$$

Entonces, entre más objetos tengan sólo una de j o $j^{'}$ mayor es la distancia.

Interpretación

- Dos modalidades escogidas por los mismos individuos coinciden
- Dos individuos son cercanos si escogen las mismas modalidades
- Modalidades con poca frecuencia están alejadas del centro de gravedad

Inercia en ACM

- *Centroide*: $G = \left(\frac{1}{n}, ..., \frac{1}{n}\right)$

- *Distancia del perfil columna al centroide*

$$d^2(j, G) = n\sum_i^n \left(\frac{z_{ij}}{z_{\bullet j}} - \frac{1}{n}\right)^2 = n\sum_i^n \left(\frac{z_{ij}}{z_{\bullet j}^2} - 2\frac{z_{ij}}{z_{\bullet j}} + \frac{1}{n^2}\right) = \frac{n}{z_{\bullet j}} - 1 \quad \left(\text{recordando que } \sum_i^n z_{ij} = z_{\bullet j}\right)$$

  Cuya distancia es más grande si $z_{\bullet j}$ es pequeña.

- *Inercia de un perfil columna*

$$\mathbf{I}(j) = \frac{z_{\bullet j}}{nk} d^2(j, G) = \frac{z_{\bullet j}}{nk}\left(\frac{n}{z_{\bullet j}} - 1\right) = \frac{1}{k}\left(1 - \frac{z_{\bullet j}}{n}\right)$$

  Mayor inercia si $z_{\bullet j}$ es pequeña.

- *Inercia de la k-ésima variable*

$$\mathbf{I}_k = \sum_{j=1}^{c_k} I(j) = \sum_{j=1}^{c_k} \frac{1}{k}\left(1 - \frac{z_{\bullet j}}{n}\right) = \frac{1}{k}(c_k - 1)$$

  que crece con el número de categorías.

- *Inercia total*

$$\mathbf{I} = \sum_k \mathbf{I}_k = \sum_k \frac{1}{k}(c_k - 1) = \frac{1}{k}(C - k) = \frac{C}{k} - 1$$

  que no tiene ningún significado estadístico.

Como mencionamos, el *ACM* es una extension del análisis de correspondencias simples (ACS), y se basa en un *ACS* de la matriz indicadora, $\mathbf{Z}$, o de la matriz de Brurt, $\mathbf{B} = \mathbf{Z}'\mathbf{Z}$.

En el *ACM* la identificación de la verdadera dimensión de los datos (de la tabla) es particularmente difícil, pese a que un *MCA* es un *ACS* de una tabla particular, porque la prueba ji-cuadrada no tiene sentido. Es decir, para la matriz de Burt $\mathbf{B}$, se puede calcular la estadística χ^2 como si fuera una tabla de contingencia usual, y esta estadística puede simplificarse a

$$\chi^2_{\mathbf{B}} = 2\sum_{k}^{i=1}\sum_{i-1}^{j=1}\chi^2_{ij} + n(C-k)$$

con χ^2_{ij} la estadística ji-cuadrada de la subtabla $\mathbf{Z}_i^{'}\mathbf{Z}_j$, $i \neq j$. Pero, desafortunadamente, la correspondiente estadística χ^2 de independencia, calculada con la matriz $\mathbf{Z}$ *no se distribuye como una ji-cuadrada.*

Por otro lado, ya que el *ACM* codifica cada variable en varias variables binarias, entonces, este esquema de codificación, crea, *de manera artificial, dimensiones adicionales a la tabla*, ya que una variable categórica se codifica en múltiples columnas. Como consecuencia de esto, la inercia (i.e. la varianza) se infla de manera artificial y, por lo tanto, el porcentaje de inercia explicado por la primer dimensión (de hecho, por pocas dimensiones) es severamente subestimado.

El término *inflación* que se aplica al alto número de eigenvalores del *MCA*, fue derivado por Benzécri (1979) que lo explica en términos del arbitrario número de niveles en que una característica continua puede discretizarse al hacerla cualitativa (discreta o categórica), y el hecho que, si comparamos el *ASC* y el *ACM* aplicado a la misma tabla de contingencia bidimensional, es posible encontrar una relación entre los eigenvalores. Es decir, al particionar una tabla de Burt, $\mathbf{B} = \mathbf{Z}^{'}\mathbf{Z}$, de dos variables discretas en submatrices, se puede demostrar la relación

$$\mu_\alpha = \frac{1 \pm \sqrt{\lambda_\alpha}}{2}$$

que se cumple entre los eigenvalores de $\mathbf{Z}$ (μ_α) y los del *ACS* de la tabla de contingencia entre estas dos variables (λ_α). En este caso, es evidente que los eigenvalores, $\lambda_\alpha = 0$, del

ACS corresponden a eigenvalores, $\mu_\alpha = \frac{1}{2}$, del *ACM* de $\mathbf{Z}$, y $\nu_\alpha = \frac{1}{4}$ de $\mathbf{B}$[1], mientras que los otros dos eigenvalores son uno mayor y otro menor que $\frac{1}{2}$ y $\frac{1}{4}$, respectivamente. Realizando la generalización de este argumento a $k > 2$ variables categóricas, resulta que hay que limitar la atención en *ACM* únicamente a aquellos eigenvalores mayores que su media, esto es

$$\mu_\alpha \geq \bar{\mu}_\alpha = \frac{1}{k}$$

Este argumento es discutido por Benzécri (1979) y Greenacre (1988, 2006). Ambos autores sugieren, para dar una medida de importancia a cada dimensión, reevaluar los vectores propios más grandes que su media, de la siguiente manera

$$\mu_\alpha \begin{cases} \left[\left(\frac{k}{k-1}\right)\left(\lambda_\alpha - \frac{1}{k}\right)\right]^2, & si\ \ \mu_\alpha \geq \bar{\mu}_\alpha = \frac{1}{k} \\ 0 \qquad\qquad\qquad\qquad\qquad , & si\ \ \mu_\alpha < \frac{1}{k} \end{cases}$$

Entonces, utilizando esta fórmula se tiene una mejor estimación de la inercia extraida por cada eigenvalor. Benzécri sugiere considerar el total de inercia, como la suma de los eigenvalores reevaluados, y tomar como porcentaje de inercia explicado por un de estos eigenvalores, al cociente

$$\frac{\mu_\alpha}{\sum_\alpha \mu_\alpha}$$

Esto resulta en una *dramática* reevaluación de la importancia relativa del primer eigenvalor.

Análisis de correspondencias conjunto

Greenacre (1988) critica el enfoque *ACM* ya que en su opinión "no es una natural generalización de la geometría [...] de la aproximación de mínimos cuadrados del [SCA] " y propone el *análisis de correspondencias conjunto* (JCA) como su generalización natural en el caso de

[1]Ya que el análisis se puede hacer sobre la matriz $\frac{\mathbf{Z}}{k}$, donde k es el número de variables categóricas en la tabla

los datos nominales, considerándolo como un conjunto de tablas de contingencia obtenidas cruzándolas sobre los mismos individuos. Según él, "en el *ACM* no parece haber justificación para el ajuste de las subtablas en la diagonal de la matriz de Burt, **B**, que contribuyen el término *n(C-k)* en la variación total ", un término que "infla artificialmente la variabilidad total, puede ocasionar que el porcentaje de varianza explicada por los ejes principales pueda ser muy baja, especialmente si *J-Q* es grande. Una medida más natural del total de variación es la suma

$$\sum_{i}\sum_{i\neq j}\chi^2_{ij}$$

Esto sugiere una alternativa en la generalización del análisis de correspondencias, que ajusta sólo las tablas de contingencia *fuera de la diagonal*, análogo a análisis de factores donde los valores de la diagonal de la matriz de varianza-covarianza o de correlación, no tienen un interés obvio.

En efecto, la redefinición propuesta de la variación total, mediante la eliminación de las matrices *diagonales por bloque*, en la diagonal de la matriz **B**, produciría un sesgo importante debido a la manera como se realiza la aplicación en la Tabla de Burt de las métricas de Ji-cuadrada, ya que la estructura de estas matrices *diagonales por bloque* de la diagonal, representa una gran desviación de los valores esperados, que el *ACM* analiza como si se tratara de una verdadera desviación. Dada esta situación, el uso del *ACM* no es muy adecuado, por lo que (JCA) parece ser una mejor propuesta.

Interpretación de ACM

Al igual que con *ACS*, la interpretación en el *ACM* se basa en las proximidades entre los puntos en el mapa de pocas dimensiones (es decir, dos o tres dimensiones). Así como para *ACS*, las proximidades sólo son significativos entre los puntos del mismo conjunto (es decir, renglones con renglones, columnas con columnas). Específicamente, cuando dos perfiles renglón están cerca uno de otro, implica que tienden a presentar los mismos niveles de las variables nominales.

Para interpretar la proximidad entre los perfiles columna es necesario distinguir dos casos. En primer lugar, la proximidad entre los niveles de diferentes perfiles columna, significa que

estos niveles tienden a aparecer juntos en las observaciones. En segundo lugar, debido a que los niveles de la misma variable nominal no pueden ocurrir al mismo tiempo, necesitamos un tipo diferente de la interpretación para este caso. Aquí la proximidad entre los niveles significa que el grupos de observaciones asociados con estos dos niveles son en sí mismos similares.

ANÁLISIS DISCRIMINANTE

INTRODUCCIÓN

Un problema muy importante en estadística lo constituye el llamado problema de clasificación. En esta sección y en la siguiente discutiremos el problema de clasificación desde dos perspectivas diferentes. Al considerar grupos de objetos en un conjunto de datos multivariados pueden surgir dos situaciones: en algunos casos es de interés determinar si de manera natural las observaciones forman grupos o clases, mientras que en otras ocasiones nos interesa clasificar a los objetos de acuerdo a un conjunto de categorías definidas previamente. En este último caso se trata de un problema de *clasificación supervisada* y será discutido en esta sección.

El problema de discriminación o clasificación es habitual en muchas áreas de la actividad humana, que van desde un diagnóstico médico hasta los sistemas que posibilitan la concesión de un crédito bancario o de reconocimiento de falsas obras de arte (pinturas o escritos).

El problema de *discriminar* aparece en muchas situaciones en que es necesario clasificar elementos con información incompleta. Por ejemplo, los sistemas automáticos de concesión de créditos (*credit scoring*) implementados en muchas instituciones financieras o bancarias, deben utilizar algunas variables de los individuos sujetos al crédito, tales como : nivel de ingresos, historial crediticio, antigüedad en el trabajo, patrimonio, edo. civil, etc., para decidir si el sujeto es o no confiable para otorgarle dicho crédito. En ingeniería este problema se conoce con el nombre de reconocimiento de patrones (pattern recognition), para diseñar máquinas capaces de realizar clasificaciones de manera automática. Por ejemplo, reconocer voces y sonidos, clasificar billetes o menedas, reconocer caracteres escritos en una pantalla de una computadora o clasificar cartas según el distrito postal. Otros ejemplos de aplicaciones del análisis discriminante son: asignar la autoría de un texto escrito de procedencia desconocida a uno de entre varios autores por las frecuencias de uso de palabras; asignar una partitura musical o un cuadro a un artista; determinar una declaración de impuestos como potencialmente fraudulenta o no; determinar una empresa como en riesgo de quiebra o no; un paciente como enfermo de cáncer o no; en Biología se presenta en la llamada taxonomía de especies, que consiste en asignar diversos individuos en *taxones*, etc.

El nombre del análisis discriminante como *técnica de clasificación supervisada*, proviene del

hecho que conocemos una muestra de elementos bien clasificados (nuestra muestra) que sirve de pauta o modelo para la clasificación de futuras observaciones.

Planteamiento estadístico del problema

Desde el punto de vista estadístico, el anális discriminante tiene los siguientes elementos.

• Se dispone de un conjunto de elementos que pueden provenir de dos o más poblaciones distintas.

• En cada elemento se ha observado un vector aleatorio de dimensión p: $\mathbf{X} = (x_1, x_2, ..., x_p)$ de características de los individuos que, suponemos, son potencialmente distintas en las poblaciones, i.e., pueden ayudar a discriminar entre estas poblaciones.

Objetivos del análisis discriminante

• **Disciminación**: Describir las características que diferencian a los distintos grupos conocidos de una población. Para encontrar factores discriminantes cuyos valores numéricos sean tales que separen a los grupos lo más posible.

• **Clasificación**: Asignar nuevos sujetos a un grupo, de entre dos o más. Derivar una regla que pueda usarse para asignar de forma *óptima* un individuo a un grupo de los ya conocidos.

Nota hitórica: La primera aplicación del análisis discriminante consitió en clasificar los restos de un cráneo descubierto en una excavación, como humano, utilizando la distribución de medidas físicas para los cráneos humanos y los de antropoides (Def. dicccionario: Que se parece al ser humano en sus características externas | antropomorfo).

En resumen

¿Qué es el análisis discriminante?

• Es una técnica estadística *de reducción de dimensión*, cuyo objetivo es maximizar la separación entre los datos de $p >> 2$ ó 3 dimensiones, cuando se realice esta reducción de dimensión a 2 ó 3.

¿Para qué?

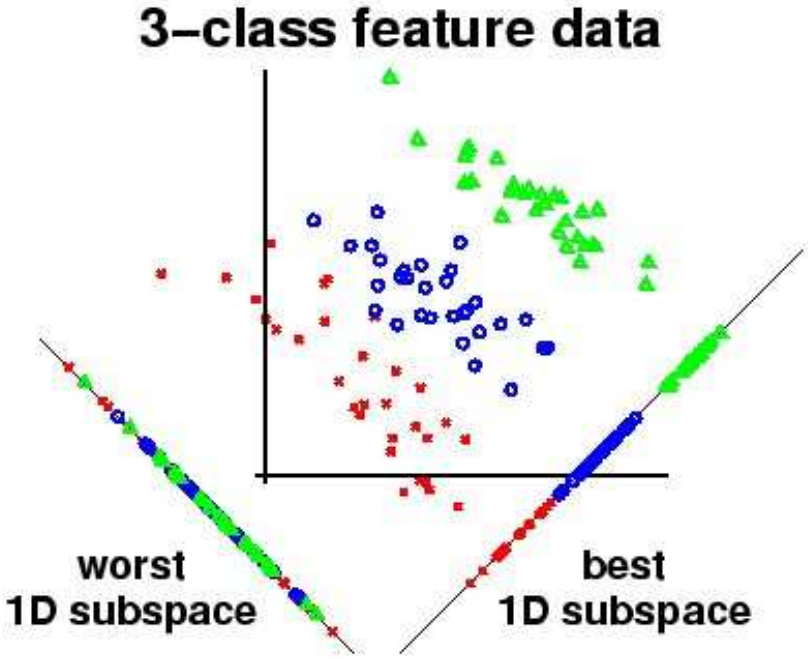
- Identificación de las características de los grupos
- Variables que discriminan entre los grupos
- Selección de las variables discriminantes
- Clasificación de nuevos individuos en los grupos ya existentes

IMPORTANTE: Para implementar esta técnica, *los grupos deben estar definidos de antemano*. Esta agrupación podría ser producto de algunos de los métodos multivariados para este fin, como *cluster o componentes principales*, producto de una agrupación natural o de la experiencia del usuario.

Variables canónicas discriminantes: Dos grupos

Mencionamos que el análisis discriminante es una técnica de reducción de dimension, reducción que sabemos se logra proyectando nuetras observaciones, originalmente en dimensión p, a un espacio de dimensión menor, idealmente *2* ó *3* para poderlas visualizar gráficamente. En este caso, esta proyección debiera ser tal que logre la mayor separación posible de los grupos en el espacio donde se proyectan. El ejemplo de la gráfica siguiente muestra que la elección del plano de proyección (en este caso una recta) no es trivial, por lo que se requieren de elementos técnicos para su determinación.

Podemos observar que la proyección sobre el plano en $\mathbb{R}^1$: la línea recta del lado izquierdo de la gráfica, no posibilita la separación de los tres grupos de observaciones. Por el contrario, una proyección de estos datos sobre el plano en $\mathbb{R}^1$, representado por la línea recta del lado derecho, logra una muy buena separación de los grupos en este espacio reducido. En este caso de dos grupos, el problema se transforma en elegir, de todas las posibles líneas rectas, aquella que maximice la separación de estas proyecciones, que son valores escalares.


3–class feature data
worst
1D subspace
best
1D subspace

Las funciones lineales discriminantes de Fisher

La función lineal discriminante para dos grupos fue deducida por primera vez por Fisher, a través de un razonamiento intuitivo. El criterio propuesto por Fisher es encontrar una variable escalar, que sea tal que maximice la distancia entre los datos proyectados.

$$Y = \mathbf{a}'\mathbf{X}$$

Como tenemos sólo dos poblaciones, entonces necesitamos una única función lineal discriminante

$$Y = \mathbf{a}'\mathbf{X} = a_1 X_1 + ... + a_p X_p$$

Entonces, de manera general, tenemos el siguiente planteamiento.

Dos poblaciones: π_1 y π_2, donde cada uno de los individuos que las componenen tiene un vector de p variables medidas $\mathbf{X}' = (X_1, ..., X_p)$, con $\mathbf{X}_1$ y $\mathbf{X}_2$, las matrices de datos de los sujetos en cada uno de los dos grupos, respectivamente.

Entonces, una vez proyectados los datos originales $\mathbf{X}' = (X_1, ..., X_p)$ a través de las funciones lineales (combinaciones lineales). Tenemos

• Todos los puntos (sujetos) $(\mathbf{X}_1, \mathbf{X}_2)$ son proyectados (mapeados) sobre el plano, Y.

• Hay que elegir a Y de tal manera que logremos la mayor separación entre los grupos proyectados.

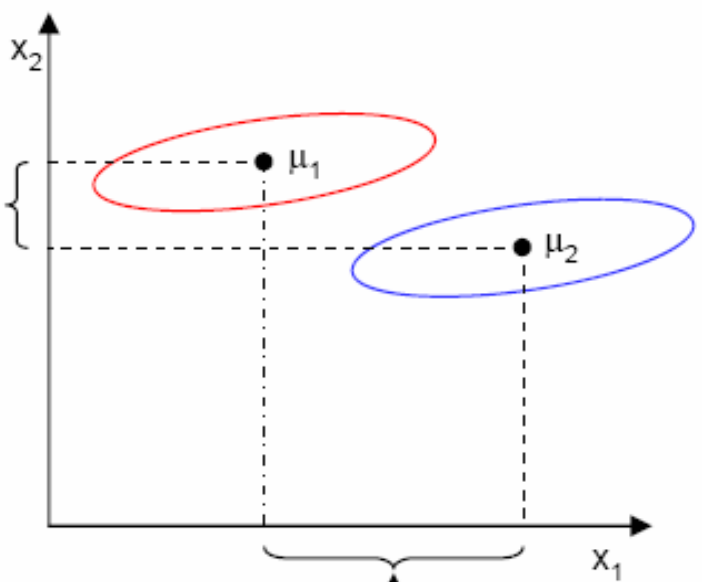
Pero, para encontrar un vector que proporciones una "buena proyección", en el sentido que dijimos, necesitamos definir una medida de separación entre estas proyecciones. Una buena alternativa podría ser elegir la distancia entre las medias proyectadas por estas funciones lineales, como nuestra función objetivo, es decir

$$J(\mathbf{a}) = |\tilde{\mu}_1 - \tilde{\mu}_2| = \left|\mathbf{a}'\mu_1 - \mathbf{a}'\mu_2\right| = \left|\mathbf{a}'\left(\mu_1 - \mu_2\right)\right|$$

con

$$\begin{aligned} \mu_1 &= \mathbf{E}(\mathbf{X}|\pi_1) \text{ : media de } \mathbf{X} \text{ en la población 1} \\ \mu_2 &= \mathbf{E}(\mathbf{X}|\pi_2) \text{ : media de } \mathbf{X} \text{ en la población 2} \end{aligned}$$

Sin embargo, la distancia entre las medias proyectadas de cada grupo, no es una muy buena medida, ya que no toma en cuenta la variabilidad (varianza o desviación estándar) dentro de estos grupos. En la gráfica siguiente se muestra que aunque existe mayor separación de las medias proyectando sobre el eje horizontal, se logra una mejor separación de los grupos, proyectando sobre el eje vertical.


x2
μ1
μ2
x1

La solución propuesta por Fisher para salvar esta dificultad, fue maximizar una función que represente esta diferencia de medias, pero normalizada (escalada) por una medida de la variabilidad dentro de los grupos.

Para cada grupo, esta variabilidad es equivalente a la varianza del grupo proyectado

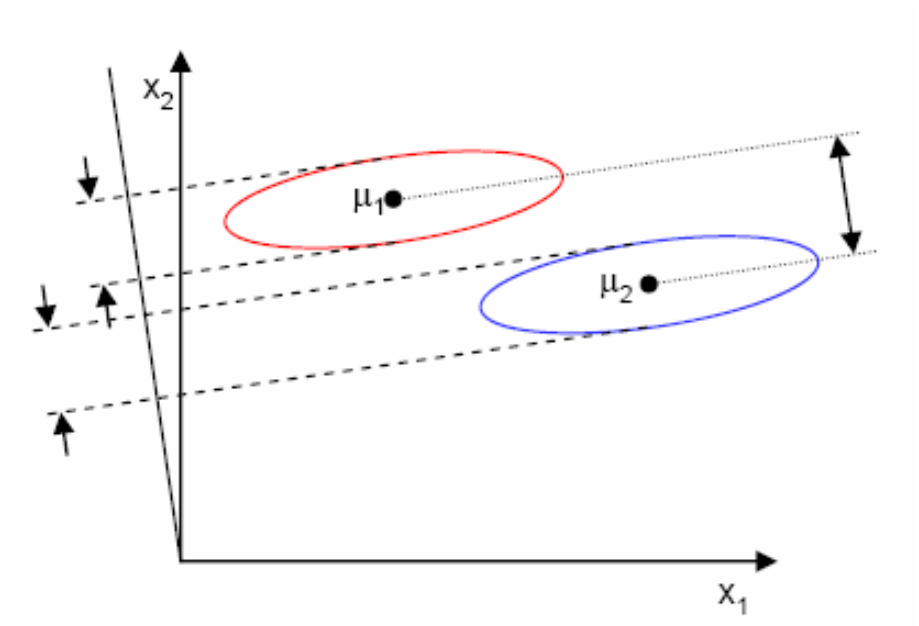
$$\tilde{S}_i^2 = \sum_{Y \in G_i} \left(Y - \tilde{\mu}_i\right)^2, \;\; i = 1, 2$$

Entonces, $\tilde{S}_i$ mide la *variabilidad dentro del grupo* i después de que ha sido proyectado en el plano Y.

Por lo tanto, $\tilde{S}_1^2 + \tilde{S}_2^2$ mide la variabilidad dentro de los dos grupos, una vez realizada la proyección, denominada *variabilidad dentro de grupos* de las muestras proyectadas.

Entonces, la función lineal discriminante de Fisher, se define como la función lineal: $Y = \mathbf{a}'\mathbf{X}$ que maximiza la función objetivo

$$J\left(\mathbf{a}\right) = \frac{\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|^2}{\tilde{S}_1^2 + \tilde{S}_2^2}$$


x_2
μ_1
μ_2
x_1

Para encontrar el vector, $\mathbf{a}^*$, que maximice esta expresión, es necesario escribirla de forma explícita como función de $\mathbf{a}$.

Definamos la variabilidad en y dentro de los grupos en el espacio original $\mathbf{X}$, como

$$S_i = \sum_{x \in G_i} (\mathbf{X} - \mu_i)(\mathbf{X} - \mu_i)^{'}, \;\; i = 1, 2 \text{ y}$$

$$S_w = S_1 + S_2$$

Donde S_i es la matriz de varianza-covarianza del grupo i, y S_w la matriz de dispersión dentro de grupos.

Ahora, regresemos a estas mismas definiciones, pero con las observaciones proyectadas en el plano Y. Y tenemos

$$\begin{aligned} \tilde{S}_i^2 = \sum_{Y \in G_i} (Y - \tilde{\mu}_i)^2 &= \sum_{Y \in G_i} \left(\mathbf{a}^{'}\mathbf{X} - \mathbf{a}^{'}\mu_i \right)^2 \\ &= \sum_{x \in G_i} \mathbf{a}^{'} (\mathbf{X} - \mu_i)(\mathbf{X} - \mu_i)^{'} \mathbf{a} \\ &= \mathbf{a}^{'} S_i \mathbf{a} \end{aligned}$$

y

$$\begin{aligned} \tilde{S}_1^2 + \tilde{S}_2^2 &= \mathbf{a}^{'} S_1 \mathbf{a} + \mathbf{a}^{'} S_2 \mathbf{a} \\ &= \mathbf{a}^{'} (S_1 + S_2) \mathbf{a} \\ &= \mathbf{a}^{'} S_w \mathbf{a} \\ &= \tilde{S}_w \end{aligned}$$

Con $\tilde{S}_w$ la matriz de dispersión dentro de grupos proyectados.

De modo similar, las medias proyectadas en el espacio Y, pueden escribirse en términos de las medias en el espacio original, de la siguiente manera

$$
\begin{aligned}
\left(\tilde{\mu}_1 - \tilde{\mu}_2\right)^2 &= \left(\mathbf{a}^{'}\mu_1 - \mathbf{a}^{'}\mu_2\right)^2 \\
&= \mathbf{a}^{'}\left(\mu_1 - \mu_2\right)\left(\mu_1 - \mu_2\right)^{'}\mathbf{a} \\
&= \mathbf{a}^{'}S_B\mathbf{a} \\
&= \tilde{S}_B
\end{aligned}
$$

La matriz S_B se conoce como *la matriz de dispersión entre los grupos*, mientras que $\tilde{S}_B$ es la matriz de dispersión entre grupos de las muestras proyectadas.

Ya que $\tilde{S}_B$ es el producto interno entre dos vectores, es de rango a lo más *uno*.

Finalmente, podemos expresar el criterio de Fisher en términos de las dos matrices de dispersion, S_w y S_B, como

$$
J\left(\mathbf{a}\right) = \frac{\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|^2}{\tilde{S}_1^2 + \tilde{S}_2^2} = \frac{\mathbf{a}^{'}S_B\mathbf{a}}{\mathbf{a}^{'}S_w\mathbf{a}}
$$

Una vez planteada la función objetivo, lo que resta es derivarla respecto a $\mathbf{a}$, para encontrar el máximo de ella. Este procedimiento se realiza, por supesto, utilizando técnicas de cálculo vectorial. El desarrollo es el siguiente

Entonces, queremos encontrar el valor de $\mathbf{a}$ que hace máxima la función $J\left(\mathbf{a}\right)$. Diferenciemos esta expresión e igualémosla a cero. Estos es

$$
\begin{aligned}
\frac{d}{d\mathbf{a}}J(\mathbf{a}) &= \frac{d}{d\mathbf{a}}\left(\frac{\mathbf{a}^{'}S_B\mathbf{a}}{\mathbf{a}^{'}S_w\mathbf{a}}\right)=0 \\
&\Rightarrow \mathbf{a}^{'}S_w\mathbf{a}\frac{d}{d\mathbf{a}}\left(\mathbf{a}^{'}S_B\mathbf{a}\right)-\mathbf{a}^{'}S_B\mathbf{a}\frac{d}{d\mathbf{a}}\left(\mathbf{a}^{'}S_w\mathbf{a}\right)=0 \\
&\Rightarrow \left(\mathbf{a}^{'}S_w\mathbf{a}\right)2S_B\mathbf{a}-\left(\mathbf{a}^{'}S_B\mathbf{a}\right)2S_w\mathbf{a}=0, \qquad \cdots(1) \\
&\quad \text{Dividiendo por } 2\mathbf{a}^{'}S_w\mathbf{a} \text{ (que es un escalar)} \\
&\Rightarrow \left(\frac{\mathbf{a}^{'}S_w\mathbf{a}}{\mathbf{a}^{'}S_w\mathbf{a}}\right)S_B\mathbf{a}-\left(\frac{\mathbf{a}^{'}S_B\mathbf{a}}{\mathbf{a}^{'}S_w\mathbf{a}}\right)S_w\mathbf{a}=0 \\
&\Rightarrow S_B\mathbf{a}-J(\mathbf{a})S_w\mathbf{a}=0 \\
&\Rightarrow S_w^{-1}S_B\mathbf{a}-J(\mathbf{a})\mathbf{a}=0 \\
&\Rightarrow S_w^{-1}S_B\mathbf{a}=J(\mathbf{a})\mathbf{a}
\end{aligned}
$$

Obsérvese que $J(\mathbf{a})$ es un escalar, digamos, λ. Entonces tenemos que resolver el problema generalizado de eigenvalores. En concreto, tenemos que encontrar el eigenvalor del sistema

$$S_w^{-1}S_B\mathbf{a}=\lambda\mathbf{a}, \qquad \text{con } \lambda=J(\mathbf{a}) \text{ un escalar}$$

cuya solución es

$$\mathbf{a}^{*}=S_w^{-1}(\mu_1-\mu_2)$$

Una forma alternativa de deducir esta solución es considerar la igualdad (1) de este desarrollo y continuar como sigue

$$\Rightarrow \quad \left(\mathbf{a}^{'} S_w \mathbf{a}\right) 2S_B\mathbf{a} - \left(\mathbf{a}^{'} S_B \mathbf{a}\right) 2S_w\mathbf{a} = 0$$

$$\Rightarrow \quad \left(\mathbf{a}^{'} S_w \mathbf{a}\right) S_B\mathbf{a} = \left(\mathbf{a}^{'} S_B \mathbf{a}\right) S_w\mathbf{a}$$

$$\Rightarrow \quad \left(\mathbf{a}^{'} S_w \mathbf{a}\right) (\mu_1 - \mu_2)(\mu_1 - \mu_2)^{'} \mathbf{a} = \mathbf{a}^{'} (\mu_1 - \mu_2)(\mu_1 - \mu_2)^{'} \mathbf{a} S_w\mathbf{a}$$

$$\Rightarrow \quad (\mu_1 - \mu_2)\left(\mathbf{a}^{'} S_w \mathbf{a}\right) = \mathbf{a}^{'} (\mu_1 - \mu_2) S_w\mathbf{a}$$

$$\Rightarrow \quad (\mu_1 - \mu_2) = \frac{S_w\mathbf{a}\left(\mathbf{a}^{'} (\mu_1 - \mu_2)\right)}{\mathbf{a}^{'} S_w\mathbf{a}}$$

$$\Rightarrow \quad \mathbf{a} = \frac{(\mu_1 - \mu_2)\,\mathbf{a}^{'} S_w\mathbf{a}}{S_w\left(\mathbf{a}^{'} (\mu_1 - \mu_2)\right)}$$

$$\Rightarrow \quad \mathbf{a} = \lambda S_w^{-1} (\mu_1 - \mu_2)$$

con $\lambda = \dfrac{\mathbf{a}^{'} S_w\mathbf{a}}{\mathbf{a}^{'} (\mu_1 - \mu_2)}$ un escalar. Ahora bien, como la función a maximizar es invariante ante multiplicaciones por constantes, y λ lo es, entonces, podemos normalizar $\mathbf{a}$, de tal manera que $\lambda = 1$, de donde obtenemos

$$\mathbf{a}^* = S_w^{-1} (\mu_1 - \mu_2)$$

Función lineal discriminante estimada

Para utilizar el discriminante con datos muestrales, es necesario estimar esta función lineal discriminante a través de los datos observados, recordando que estos datos son los que se generan una vez proyectados a través de la función discriminante. Entonces, las matrices que necesitamos son

$$\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 = \sum_{i=1}^{n}\left(\mathbf{a}'\left(y_i-\overline{y}\right)\right)^2 = \mathbf{a}'\mathbf{T}\mathbf{a}: \text{ Suma de cuadrados totales o varianza total, y}$$

$$\sum_{g=1}^{G} n_g\left(\overline{X}_g-\overline{X}\right)^2 = \sum_{g=1}^{G} n_g\left(\mathbf{a}'\left(\overline{y}_g-\overline{y}\right)\right)^2 = \mathbf{a}'\mathbf{E}\mathbf{a}: \text{ Suma de cuadrados entre grupos o varianza entre g}$$

$$\widehat{\mu}_1 = \overline{\mathbf{X}}_1,\ \widehat{\mu}_2 = \overline{\mathbf{X}}_2 \text{ Las medias en los grupos 1 y 2, respectivamente, y}$$

$$\mathbf{S}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j}-\overline{\mathbf{X}}_1\right)\left(X_{1j}-\overline{\mathbf{X}}_1\right)',\quad \mathbf{S}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2}\left(X_{2j}-\overline{\mathbf{X}}_2\right)\left(X_{2j}-\overline{\mathbf{X}}_2\right), \text{ las respectivas varianza}$$

$$\mathbf{S} = \mathbf{S}_{pool} = \frac{\left(n_1-1\right)\mathbf{S}_1+\left(n_2-1\right)\mathbf{S}_2}{n_1+n_2-2}: \text{ La varianza conjunta de los grupos}$$

El supuesto de matrices de varianza-covarianza iguales dentro de los dos grupos, es fundamental y hace que la matriz de varianza-covarianza total se estime como un *pool* de las correspondientes matrices de cada grupo.

Entonces, la función lineal estimada queda como

$$\widehat{y} = \widehat{\mathbf{a}}'\mathbf{X} = \left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)'\mathbf{S}_{pool}^{-1}\mathbf{X}$$

Urgente: Un ejemplo

49 adultos mayores del sexo masculino participaron en un estudio interdisciplinario sobre su condición humana, y fueron clasificados en dos grupos: "factor senil presente" y "factor

senil ausente", basados en una intesiva evaluación psicológica. Los siguientes resultados son el resultado de cuatro pruebas realizadas a estos sujetos.

	No Senil (n=37)		*Senil* ($n = 12$)	
Prueba	$\overline{\mathbf{X}}$	**S.D.**	$\overline{\mathbf{X}}$	**S.D.**
Información	12.566	3.387	8.750	3.251
Similaridades	9.486	3.380	5.333	4.271
Aritmética	11.514	3.363	8.500	3.631
Pintura	7.973	1.922	4.750	3.571

$$\mathbf{S}_{Senil} = \begin{pmatrix} 11.47 & 8.55 & 6.39 & 2.07 \\ 8.55 & 11.42 & 5.49 & 0.29 \\ 6.39 & 5.49 & 11.31 & 1.82 \\ 2.07 & 0.29 & 1.82 & 3.69 \end{pmatrix}, \quad \mathbf{S}_{No\text{Senil}} = \begin{pmatrix} 10.57 & 10.45 & 9.68 & 7.66 \\ 10.45 & 18.24 & 12.09 & 8.91 \\ 9.68 & 12.09 & 13.18 & 5.32 \\ 7.66 & 8.91 & 5.32 & 12.75 \end{pmatrix},$$

$$\mathbf{S}_{poool} = \begin{pmatrix} 11.26 & 9.00 & 7.16 & 3.38 \\ 9.00 & 13.02 & 7.04 & 2.31 \\ 7.16 & 7.04 & 11.75 & 2.64 \\ 3.38 & 2.31 & 2.64 & 5.81 \end{pmatrix}$$

$$\begin{aligned} \left(\overline{X}_{No\text{Senil}} - \overline{X}_{\text{Senil}}\right)' &= (3.82, 4.15, 3.01, 3.23) \\ \mathbf{a}' &= \left(\overline{X}_{No\text{Senil}} - \overline{X}_{\text{Senil}}\right)' \mathbf{S}_{poool}^{-1} \\ &= (3.82, 4.15, 3.01, 3.23) \begin{pmatrix} 0.249 & -0.127 & -0.060 & 0.066 \\ -0.127 & 0.180 & -0.034 & 0.0182 \\ -0.060 & -0.034 & 0.146 & -0.017 \\ 0.066 & 0.0182 & -0.017 & 0.211 \end{pmatrix} \\ &= (0.02453159, 0.2162928, 0.01043125, 0.4510016) \end{aligned}$$

Las medias de los grupos proyectados son

$$\overline{y}_{No\text{Senil}} = (0.02453159, 0.2162928, 0.01043125, 0.4510016) \begin{pmatrix} 12.566 \\ 9.486 \\ 11.514 \\ 7.973 \end{pmatrix} = 6.07$$

$$\overline{y}_{\text{Senil}} = (0.02453159, 0.2162928, 0.01043125, 0.4510016) \begin{pmatrix} 8.750 \\ 5.333 \\ 8.500 \\ 4.750 \end{pmatrix} = 3.59$$

La función lineal disciminante para cada sujeto es

$$\begin{aligned} y_j = \mathbf{a}^{'} X_j &= (0.02453159, 0.2162928, 0.01043125, 0.4510016) \left(X_{1j}, X_{2j}, X_{3j}, X_{4j}\right)^{'} \\ &= 0.02X_{1j} + 0.22X_{2j} + 0.01X_{3j} + 0.45X_{4j} \end{aligned}$$

Clasificación

Un manera muy simple de utilizar esta función lineal, Y, para clasificar una nueva observación, X_0, a alguno de los grupos es

1.- Calcular la proyección en el plano Y, de esta observación

$$y_0 = (\mu_2 - \mu_1)^{'} S_w^{-1} X_0$$

2.- Encontrar el punto medio de las medias de los grupos proyectadas $\tilde{\mu}_1$ y $\tilde{\mu}_2$.

$$\begin{aligned} m &= \frac{1}{2}(\tilde{\mu}_1 + \tilde{\mu}_2) \\ &= \frac{1}{2}\left(\mathbf{a}^{'}\mu_1 + \mathbf{a}^{'}\mu_2\right) \\ &= \frac{1}{2}(\mu_2 - \mu_1)^{'} S_w^{-1}(\mu_2 + \mu_1) \end{aligned}$$

3.- Regla de clasificación

$$\begin{aligned}\text{Asignar X}_0 \text{ al grupo 1 } (\pi_1) \text{ si } y_0 &= \left(\mu_2 - \mu_1\right)' S_w^{-1} X_0 \geq m, \text{ y} \\ \text{Asignar X}_0 \text{ al grupo 2 } (\pi_2) \text{ si } y_0 &= \left(\mu_2 - \mu_1\right)' S_w^{-1} X_0 < m\end{aligned}$$

o bien si

$$y_0 - m = \left(\mu_2 - \mu_1\right)' S_w^{-1} X_0 \geq 0 \text{ ó } < 0$$

Estimación de la regla de clasificación

La regla de clasificación estimada queda como

$$\begin{aligned}\text{Asignar X}_0 \text{ al grupo 1 } (\pi_1) \text{ si } y_0 &= \left(\bar{X}_2 - \bar{X}_1\right)' S_{pool}^{-1} X_0 \geq m, \text{ y} \\ \text{Asignar X}_0 \text{ al grupo 2 } (\pi_2) \text{ si } y_0 &= \left(\bar{X}_2 - \bar{X}_1\right)' S_{pool}^{-1} X_0 < m\end{aligned}$$

En nuestro caso

$$m = \frac{1}{2}\left(6.07 + 3.59\right) = 4.83$$

Entonces, un nuevo individuo se asignaría al grupo: *No senil* si y_0, su puntaje dado por la proyección de sus valores en el plano Y, es mayor que 4.83, y se asignaría al grupo *Senil* si es menor a 4.83.

Por ejemplo, a qué grupo asignaríamos a un individuo que tiene el siguiente vector de observaciones: $X_0 = (8.150, 6.001, 9.050, 4.510)$?

Notemos primeramente que este vector está cercano a las medias del grupo *senil*: (8.750, 5.333, 8.500, 4.750), entonces, debería de clasificarse en ese grupo. Calculemos su proyección al plano Y, i.e., calculemos

$$\begin{aligned}y_0 = \mathbf{a}' X_0 &= \left(0.02453159, 0.2162928, 0.01043125, 0.4510016\right)\left(X_{10}, X_{20}, X_{30}, X_{40}\right)' \\ &= 0.02X_{10} + 0.22X_{20} + 0.01X_{30} + 0.45X_{40} \\ &= \left(0.02453159, 0.2162928, 0.01043125, 0.4510016\right) * \left(8.150, 6.001, 9.050, 4.510\right) = 3.626326\end{aligned}$$

Si consideramos ahora un sujeto con valores más cercanos a las medias del grupo *No senil*: (12.566 , 9.486, 11.514, 7.973), digamos, $X_0 = (11.950, 10.00, 10.73, 8.103)$, debería clasificarse como No senil. Su proyección es: 6.222474; que corrobora nuestra especulación.

Discriminante clásico

Se denomina discriminante clásico al discriminante que asume poblaciones normales multivariadas para cada uno de los grupos. Es decir, se supone que cada población tiene función de densidad de probabilidad, dada por

$$f_i(\mathbf{X}) = \frac{(2\pi)^{-p/2}}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_i)' \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_i)\right\}, \quad \text{i=1,2,...,G}$$

Discriminante clásico con dos grupos

En el caso de que tengamos dos grupos con probabilidades a priori de pertencia a cada uno de ellos π_1 y π_2, respectivamente ($\pi_1 + \pi_2 = 1$). Entonces, para clasificar a un nuevo individuo, $\mathbf{x}_0$, por ejemplo, al grupo *2*, sólo debemos comparar sus densidades, y lo asignamos al grupo *2* síi

$$\pi_2 f_2(\mathbf{x}_0) > \pi_1 f_1(\mathbf{x}_0)$$

si las probabilidades iniciales son iguales, entonces lo asignamos a dicho grupo sii

$$f_2(\mathbf{x}_0) > f_1(\mathbf{x}_0)$$

Bajo el supuesto de que las densidades sean normales de dimensión *p,* tenemos

$$\begin{aligned}
\frac{\pi_2}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_2)' \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_2)\right\} &> \frac{\pi_1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_1)' \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_1)\right\} \Longrightarrow \\
\log(\pi_2) - \frac{1}{2}(\mathbf{x}-\mu_2)' \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_2) &> \log(\pi_1) - \frac{1}{2}(\mathbf{x}-\mu_1)' \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_1) \\
(\mathbf{x}-\mu_1)' \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_1) &> (\mathbf{x}-\mu_2)' \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_2) - 2\log\left(\frac{\pi_2}{\pi 1}\right)
\end{aligned}$$

si denotamos como D_i^2 el cuadrado de la distancia de Mahalanobis entre el punto observado, $\mathbf{x}$, y la media de la población i=1,2, tenemos

$$D_i^2 = (\mathbf{x}-\mu_i)' \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_i)$$

si suponemos probabilidades iniciales iguales, entonces la regla que se obtiene para clasificar $\mathbf{x}$ en el grupo *2* es: Clasificar esta observación en el grupo *2* sii

$$D_1^2 > D_2^2$$

es decir, clasificar la observación en el grupo cuyas medias estén más próximas, según la distancia de Mahalanobis cuadrada.

Interpretación de la regla anterior

Desarrollemos las siguientes expresiones

$$\begin{aligned}
\left(\mathbf{x}-\mu_1\right)^{'}\mathbf{\Sigma}^{-1}\left(\mathbf{x}-\mu_1\right) &= \mathbf{x}^{'}\mathbf{\Sigma}^{-1}\mathbf{x}-2\mu_1^{'}\mathbf{\Sigma}^{-1}x+\mu_1^{'}\mathbf{\Sigma}^{-1}\mu_1, \text{ y} \\
\left(\mathbf{x}-\mu_2\right)^{'}\mathbf{\Sigma}^{-1}\left(\mathbf{x}-\mu_2\right) &= \mathbf{x}^{'}\mathbf{\Sigma}^{-1}\mathbf{x}-2\mu_2^{'}\mathbf{\Sigma}^{-1}x+\mu_2^{'}\mathbf{\Sigma}^{-1}\mu_2
\end{aligned}$$

entonces, la regla divide al conjunto de valores posibles de $\mathbf{x}$, en dos regiones cuya frontera es

$$-2\mu_1^{'}\mathbf{\Sigma}^{-1}x+\mu_1^{'}\mathbf{\Sigma}^{-1}\mu_1=-2\mu_2^{'}\mathbf{\Sigma}^{-1}x+\mu_2^{'}\mathbf{\Sigma}^{-1}\mu_2$$

que es equivalente, como función de $\mathbf{x}$ a

$$\left(\mu_2-\mu_1\right)^{'}\mathbf{\Sigma}^{-1}\mathbf{x}=\left(\mu_2-\mu_1\right)^{'}\mathbf{\Sigma}^{-1}\left(\frac{\mu_2+\mu_1}{2}\right)$$

Observemos que el hecho de suponer matriz de varianzas covarianzas iguales entre los grupos, permite el agrupamiento de los términos de esta manera. Si denotamos por

$$\mathbf{a}^{'}=\left(\mu_2-\mu_1\right)^{'}\mathbf{\Sigma}^{-1}$$

entonces, la frontera entre las dos regiones de clasificación para π_1 y π_2 puede escribirse como

$$\mathbf{a}^{'}\mathbf{x}=\mathbf{a}^{'}\left(\frac{\mu_2+\mu_1}{2}\right)$$

que es la ecuación de un hiperplano. También equivalente a

$$2\mathbf{a}^{'}\mathbf{x} = \mathbf{a}^{'}(\mu_1 + \mu_2)$$

$$\mathbf{a}^{'}\mathbf{x} - \mathbf{a}^{'}\mu_1 = \mathbf{a}^{'}\mu_2 - \mathbf{a}^{'}\mathbf{x}\,(*)$$

Se puede demostrar que esta regla equivale a proyectar el punto $\mathbf{x}$ que queremos clasificar y las medias de ambas poblaciones sobre la función lineal discriminante, y después asignar el punto a aquella población de cuya media se encuentre más próxima en la proyección. Situación que habíamos visto anteriormente.

Esta última ecuación indica que el procedimiento para clasificar un elemento X_0 puede resumirse como sigue:

• Calcular el vector $\mathbf{a}^{'}$, mediante la expresión correspondiente

• Construir la función lineal discriminante

$$Y = \mathbf{a}^{'}\mathbf{X} = a_1X_1 + ... + a_pX_p$$

• Calcular la proyección en el plano Y, $Y_0 = \mathbf{a}^{'}X_0$, del individuo $X_0 = (X_{10}, ..., X_{p0})$, y el valor de las medias proyectadas de las poblaciones, $\tilde{\mu}_i = \mathbf{a}^{'}\mu_i$. Clasificar esta observación en aquella población donde la distancia, $|Y_0 - \tilde{\mu}_i|$, sea mínima.

Obsérvese que

$$\mathbb{E}(Y|\pi_i) = \tilde{\mu}_i = \mathbf{a}^{'}\mu_i,\ \ i = 1, 2$$

Entonces, la regla de decisión que se desprende de $(*)$, equivale a clasificar la observación en el grupo π_2, síi

$$|Y - \tilde{\mu}_1| > |Y - \tilde{\mu}_2|$$

Esta variable aleatoria Y tiene varianza dada por

$$\mathbb{V}(Y) = \mathbb{V}\left(\mathbf{a}'\mathbf{X}\right) = \mathbf{a}'\mathbb{V}(\mathbf{X})\,\mathbf{a} = \mathbf{a}'\Sigma\mathbf{a} = (\mu_2 - \mu_1)' \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}(\mu_2 - \mu_1)$$
$$= (\mu_2 - \mu_1)' \boldsymbol{\Sigma}^{-1}(\mu_2 - \mu_1) = D^2$$

y el cuadrado de la distancia que es un escalar, entre las medias proyectadas es la distancia de Mahalanobis entre los vectores de medias originales:

$$(\tilde{\mu}_2 - \tilde{\mu}_1)^2 = \left(\mathbf{a}'(\mu_2 - \mu_1)\right)^2 = (\mu_2 - \mu_1)' \boldsymbol{\Sigma}^{-1}(\mu_2 - \mu_1) = D^2$$

Cálculo de probabilidades de clasificación errónea

Una vez que hemos obtenido la regla de clasificación, en este caso, para dos poblaciones, debemos estimar o calcular las probabilidades de clasificación errónea o estimar el error de clasificación. Bajo el supuesto de que las poblaciones son normales multivariadas con igual matriz de varianza-covarianza, obtuvimos que

$$Y = \mathbf{a}'\mathbf{X} \sim N\left(\mathbf{a}'\mu_i, D^2\right)$$

Bajo esta situación, podemos calcular la probabilidad de clasificar de manera errónea una observación. En concreto, la probabilidad de clasificar erróneamente una observación $\mathbf{X}$ cuando $\mathbf{X} \in \pi_1$, es

$$\mathbb{P}(\pi_2|\pi_1) = \mathbb{P}\left\{y \geq \frac{\tilde{\mu}_1 + \tilde{\mu}_2}{2} | y \sim N\left(\tilde{\mu}_1, D^2\right)\right\}$$

Si construimos la variable estandarizada $z = \dfrac{y - \tilde{\mu}_1}{D} \sim N(0,1)$, entonces esta probabilidad es

$$\mathbb{P}(\pi_2|\pi_1) = \mathbb{P}\left\{z \geq \frac{\frac{\tilde{\mu}_1 + \tilde{\mu}_2}{2} - \tilde{\mu}_1}{D}\right\} = 1 - \Phi\left(\frac{D}{2}\right)$$

De forma análoga, la probabilidad de clasificar de manera errónea una observación $\mathbf{X}$ cuando $\mathbf{X} \in \pi_2$, es

$$\mathbb{P}\left(\pi_1|\pi_2\right) = \mathbb{P}\left\{z < \frac{\frac{\tilde{\mu}_1 + \tilde{\mu}_2}{2} - \tilde{\mu}_2}{D}\right\} = \Phi\left(-\frac{D}{2}\right)$$

Por la simetría de la distribución normal, estas probabilidades son iguales, además de que la regla de clasificación obtenida hace mínimas estas probabilidades de error, y los errores de clasificación sólo dependen de las distancias de Mahalanobis entre las medias.

Probabilidades a posteriori

El grado de certeza de la regla de clasificación, depende de la probabilidad de acertar (clasificar correctamente) mediante la misma. La *probabilidad a posteriori* de que la observación, $\mathbf{X}$, sea clasificada o asignada a la población, π_1, se calcula como

$$\begin{aligned}
\mathbb{P}\left(\pi_1|\mathbf{X}\right) &= \frac{\pi_1 f_1\left(\mathbf{X}\right)}{\pi_1 f_1\left(\mathbf{X}\right) + \pi_2 f_2\left(\mathbf{X}\right)} \\
&= \frac{\pi_1 exp\left\{-\frac{1}{2}\left(\mathbf{X}-\mu_1\right)^{'}\mathbf{\Sigma}^{-1}\left(\mathbf{X}-\mu_1\right)\right\}}{\pi_1 exp\left\{-\frac{1}{2}\left(\mathbf{X}-\mu_1\right)^{'}\mathbf{\Sigma}^{-1}\left(\mathbf{X}-\mu_1\right)\right\} + \pi_2 exp\left\{-\frac{1}{2}\left(\mathbf{X}-\mu_2\right)^{'}\mathbf{\Sigma}^{-1}\left(\mathbf{X}-\mu_2\right)\right\}}
\end{aligned}$$

Que puede escribirse en términos de las distancias cuadradas de Mahalanobis entre la observación y cada una de las dos medias, D_1^2 y D_2^2, como:

$$\mathbb{P}\left(\pi_1|\mathbf{X}\right) = \frac{1}{1 + \frac{\pi_1}{\pi_2} exp\left\{-\frac{1}{2}\left(D_1^2 - D_2^2\right)\right\}}$$

Discriminante logístico

En el problema de clasificación a través del análisis discriminante que hemos tratado en esta sección, vimos que si la distribución conjunta de las observaciones es normal multivariada, utilizar las distancias de Mahalanobis estimadas suele dar buenos resultados y resulta óptimo con muestras grandes. Sin embargo, frecuentemente los datos recabados para realizar esta clasificación no son normales. Por ejemplo, en muchos problemas de clasificación se utilizan variables discretas, lo que haría cuestionable la distribución normal multivariada de los datos, y la condición óptima de los resultados basados en ésta.

El modelo Logit

Consideremos el problema de la discriminación únicamente entre dos poblaciones. Una forma de abordar el problema es definir una variable de clasificación, y, que tome el *valor cero* cuando la observación pertenezca a la primera población, π_1, *y uno* cuando pertenece a la segunda, π_2. Entonces, la muestra consistirá en n elementos del tipo $(y_i, \mathbf{X}_i)$, donde y_i determina el valor de la variable de clasificación, y, y $\mathbf{X}_i$ es un vector de variables explicativas o predictoras. A continuación, construiremos un modelo para pronosticar el valor de la variable de respuesta o de clasificación, y, de una nueva observación, cuando se conocen las variables predictoras, $\mathbf{X}$. El primer enfoque simple es formular el modelo de regresión lineal correspondentiene. Basta analizar este enfoque a través del modelo de regresión simple, aunque el vector de variables explicativas puede ser de dimensión mayor a uno. El modelo es

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, ..., n$$

Entonces, sabemos que

$$\mathbb{E}\left(y_i | x_i\right) = \beta_0 + \beta_1 x_i$$

Ya que nuestra variable de respuesta, y, es una indicadora de pertenencia a los grupos, es claro que esta esperanza es la probabilidad de que un sujeto sea clasificado en la población, π_2. Llamenos p_i a la probabilidad de que esta variable tome el *valor uno*, i.e., que la observación pertenezca al grupo, *dos*. Entonces

$$p_i = \mathbb{P}(y = 1|x_i)$$

Entonces, la variable de clasificación, y, es binomial y toma los valores posibles *uno* y *cero* con probabilidades p_i y $1 - p_i$. Por lo que su esperanza es:

$$\mathbb{E}(y|x_i) = p_i \times 1 + (1 - p_i) \times 0 = p_i$$

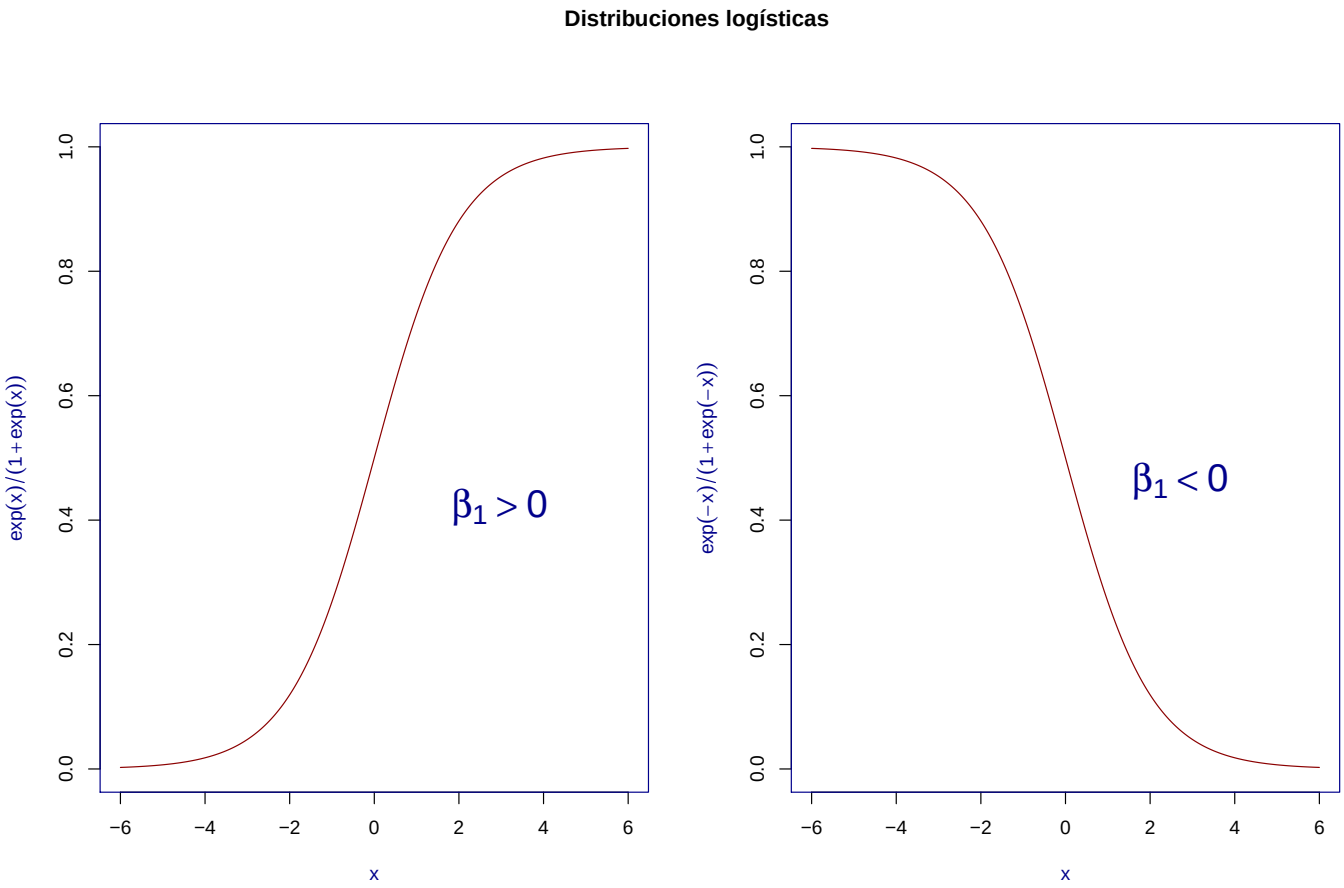
por lo tanto, concluimos que

$$p_i = \beta_0 + \beta_1 x_i$$

Si, aparentemente, este modelo funciona bien, entonces...

Porqué no usar el modelo lineal?

Reflexionemos un poco sobre el modelo propuesto. El rango de variación de p_i está entre *cero y uno*, porque es una probabilidad; sin embargo, el rango de variación de $\beta_0 + \beta_1 x_i$ no necesariamente está dentro de este rango, de hecho, **¡podría ser negativo!**, lo que es, por supuesto, una incongruencia. Un inconveniente más es que nuestra variable de respuesta se distribuye $Bernoulli(p_i)$ y no tiene varianza constante, ya que su varianza es $p_i(1 - p_i)$. Por lo tanto, esta propuesta de modelo no parece ser adecuada para el tipo de respuesta que queremos modelar. Necesitamos un modelo cuyos valores para la respuesta estén contenidos en el intervalo (0,1) y que si $\beta_1 > 0$ y x es "grande" entonces $p_i \to 1$ o si $x \to -\infty$ $p_i \to 0$. Por otro lado, si $\beta_1 < 0$ y $x \to \infty$, entonces $p_i \to 0$, o si $x \to -\infty$, entonces $p_i \to 1$ de hecho, necesitamos una función cuya gráfica sea de la forma



La función que permite ajustar este tipo de curvas es *la función logística* y el modelo de regresión asociado es el *modelo de regresión logística*. Este modelo se escribe de la siguiente manera

$$\pi(x) = Pr(Y = 1|x) = \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}$$

este es un modelo *no lineal* para nuestra respuesta, $\pi(x)$. Para lograr un modelo lineal a partir del modelo anterior, primero, construyamos el momio correspondiente

$$\frac{\pi(x)}{1 - \pi(x)} = \frac{Pr(Y = 1|x)}{1 - Pr(Y = 1|x)} = exp\left(\beta_0 + \beta_1 x\right)$$

si aplicamos la función logaritmo en ambos lados de la igualdad tenemos

$$logit(\pi(x)) = log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x$$

que es un modelo lineal, para el logaritmo del momio de la respuesta, conocido como *Logit*. Obviamente, este modelo se puede generalizar para más de una variable explicativa. El modelo con *p regresores* es

$$logit(\pi(x)) = log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

la escala de medición de los regresores puede ser cualquiera.

Los parámetros de este modelo se estiman por *máxima verosimilitud*. El modelo predice la probabilidad de cada individuo de presentar la respuesta *y=1*, por lo que, generalmente, se asigna a una observación a la población asociada a este valor de respuesta, si esta probabilidad es *mayor que 0.5* y se asigna a la otra población, en caso contrario.

Funciones lineales discriminantes para varios grupos

El enfoque de Fisher puede generalizarse para encontrar las funciones lineales que tengan máximo poder discriminante para clasificar nuevos elementos entre $G > 2$ poblaciones. La manera de hacerlo es semejante al caso de dos grupos, sólo que ahora se tienen *k=min(G-1,p)* funciones discriminantes. Es decir

$$\begin{aligned} Y_1 &= \mathbf{a}_1^{'} X \\ Y_2 &= \mathbf{a}_2^{'} X \\ &\vdots \\ Y_k &= \mathbf{a}_k^{'} X \end{aligned}$$

Entonces, en este caso, el proceso de clasificación es como sigue:

• Proyectamos las medias de cada grupo. Esto es, obtenemos

$$\tilde{\mu}_i = \mathbf{E}\left(Y|\pi_i\right) = \mathbf{E}\left(\mathbf{a}^{'}\mathbf{X}|\pi_i\right) = \mathbf{a}^{'}\mathbf{E}\left(\mathbf{X}|\pi_i\right) = \mathbf{a}^{'}\mu_i \quad \text{i=1,2,...,G}$$

• Proyectamos el vector de covariables del sujeto a clasificar, X_0, y obtenemos y_0 su proyección sobre el espacio Y.

• Clasificamos el punto en aquella población de cuya media se encuentre más cercana.

Las distancias se miden con la distancia euclideana en el espacio de las variables canónicas, y. Es decir, clasificaremos al sujeto , X_0, en la población i si:

$$\left(y_0 - \tilde{\mu}_i\right)^{'}\left(y_0 - \tilde{\mu}_i\right) = \min_g \left(y_0 - \tilde{\mu}_g\right)^{'}\left(y_0 - \tilde{\mu}_g\right)$$

Como tenemos varios grupos, la separación entre las medias la mediremos por el cociente entre la variabilidad entre grupos, y la variabilidad dentro de los grupos. Este es el criterio habitual para comparar varias medias en el análisis de la varianza y genera el estadístico *F de Fisher*. De hecho, lo que estamos haciendo es plantear un análisis de varianza en el

espacio de proyección, Y.

Nuevamente, para obtener las variables lineales discriminantes, comenzamos buscando un vector de proyección, $\mathbf{a}$, de norma uno, tal que los grupos de observaciones proyectados sobre él tengan separación relativa máxima. La proyección de la media de las observaciones del grupo g en esta dirección corresponde al escalar:

$$\bar{\tilde{\mu}}_g = \mathbf{a}'\bar{\mathbf{X}}_g$$

Con la correspondiente proyección para la media de todos los datos, dada por

$$\bar{\tilde{\mu}} = \mathbf{a}'\bar{\mathbf{X}}$$

ambas medias proyectadas son vectores de dimensión $p \times 1$, sólo que la primera es para los individuos en el grupo *g, g=1,2,...,k*, y la segunda es para todos los datos, sin importar la pertenencia a algún grupo.

Entonces, tomando como medida de la distancia entre las medias de los grupos proyectadas: $\tilde{\mu}_1, \tilde{\mu}_2, ..., \tilde{\mu}_k$, la varianza total dentro de grupos es

$$\sum_{g=1}^{k} n_g \left(\bar{\tilde{\mu}}_g - \bar{\tilde{\mu}}\right)^2$$

que debemos comparar contra la varianza dentro de grupos o variabilidad total, dada por

$$\sum_i \sum_g \left(y_{ig} - \bar{\tilde{\mu}}_g\right)^2$$

El proceso para encontrar las funciones lineales se realiza mediante el cociente de las varianzas entre grupos y total (idéntico al procedimiento ANOVA, sólo que aquí estas varianzas se obtienen con los elementos proyectados).

$$\frac{\sum_{g=1}^{k} n_g \left(\bar{\tilde{\mu}}_g - \bar{\tilde{\mu}}\right)^2}{\sum_i \sum_g \left(y_{ig} - \bar{\tilde{\mu}}_g\right)^2}$$

Ahora, expresemos este criterio en función de los datos originales. La suma de cuadrados *dentro de grupos*, para los puntos proyectados, es:

$$\sum_{i=1}^{n_g}\sum_{g=1}^{k}\left(y_{ig}-\bar{\tilde{\mu}}_g\right)^2=\sum_{i=1}^{n_g}\sum_{g=1}^{k}\mathbf{a}^{'}\left(\mathbf{X}_{ig}-\bar{X}_g\right)\left(\mathbf{X}_{ig}-\bar{X}_g\right)^{'}\mathbf{a}=\mathbf{a}^{'}\mathbf{W}\mathbf{a}$$

con $\mathbf{W}$, dada por

$$\sum_{i=1}^{n_g}\sum_{g=1}^{k}\left(\mathbf{X}_{ig}-\bar{X}_g\right)\left(\mathbf{X}_{ig}-\bar{X}_g\right)^{'}$$

Esta matriz tiene dimensiones $p\times p$ y, en general, es de rango p, asumiendo que $n-k\geq p$. Estima la variabilidad de los datos respecto a las medias de su grupo.

Por otro lado, la suma de cuadrados *entre grupos*, para los puntos proyectados está dada por

$$\sum_{g=1}^{k}n_g\left(\bar{\tilde{\mu}}_g-\bar{\tilde{\mu}}\right)^2=\sum_{g=1}^{k}n_g\mathbf{a}^{'}\left(\bar{X}_g-\bar{X}\right)\left(\bar{X}_g-\bar{X}\right)^{'}\mathbf{a}=\mathbf{a}^{'}\mathbf{B}\mathbf{a}$$

Es decir, la matriz $\mathbf{W}$ corresponde a las diferencias dentro de grupos (withing) y la matriz $\mathbf{B}$ las diferencias entre grupos (between).

Entonces, la cantidad a maximizar para encontrar las funciones lineales discriminantes es

$$\mathbf{J}=\frac{\mathbf{a}^{'}\mathbf{B}\mathbf{a}}{\mathbf{a}^{'}\mathbf{W}\mathbf{a}}$$

Realizando el proceso usual, tenemos

$$\begin{aligned}\frac{2\mathbf{B}\mathbf{a}\left(\mathbf{a}^{'}\mathbf{W}\mathbf{a}\right)-\left(\mathbf{a}^{'}\mathbf{B}\mathbf{a}\right)\mathbf{W}\mathbf{a}}{\left(\mathbf{a}^{'}\mathbf{W}\mathbf{a}\right)^2} &= 0\\ \mathbf{B}\mathbf{a} &= \mathbf{W}\mathbf{a}\frac{\left(\mathbf{a}^{'}\mathbf{B}\mathbf{a}\right)}{\left(\mathbf{a}^{'}\mathbf{W}\mathbf{a}\right)}\\ \mathbf{B}\mathbf{a} &= \mathbf{J}\mathbf{W}\mathbf{a}\end{aligned}$$

Suponiendo que $\mathbf{W}$ tiene inversa, i.e., es no singular, y observando que $\mathbf{J}$ es un escalar, que denotaremos como λ, entonces, obtenemos el sistema

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{a} = \lambda\mathbf{a}$$

lo que implica que $\mathbf{a}$ debe ser un vector propio de la matriz $\mathbf{W}^{-1}\mathbf{B}$ y λ su valor propio asociado. Como el objetivo es maximizar $\lambda = \mathbf{J}$, que corresponde a la versión de la ANOVA en el espacio de proyección, Y, entonces $\mathbf{a}$ debe ser el vector propio asociado al valor propio más grande de la matriz $\mathbf{W}^{-1}\mathbf{B}$, que llameros $\mathbf{a}_1$. Con este vector construiríamos la primer función lineal discriminante

$$Y_1 = \mathbf{a}_1^{'}\mathbf{X}$$

Por construcción, esta función disciminante debe tener el mayor poder para discriminar entre los grupos. La segunda de estas funciones debe tener el mayor poder de disciminación restante, una vez construida la primer función discriminante, y debe ser ortogonal a la primera

$$Y_2 = \mathbf{a}_2^{'}\mathbf{X}, \quad Y_1 \perp Y_2 \Rightarrow \mathbf{a}_1 \perp \mathbf{a}_2$$

de forma análoga a la construcción de la primer función discriminante, se puede demostrar que el poder de disciminación de esta segunda función se maximiza si $\mathbf{a}_2$ es el correspondiente vector propoio asociado al segundo valor propio más grande de la matriz $\mathbf{W}^{-1}\mathbf{B}$. En general, se tiene que si $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_k$ son los vectores propios de $\mathbf{W}^{-1}\mathbf{B}$, asociados a los valores propios $\lambda_1, \lambda_2, ..., \lambda_k$, con $\lambda_1 > \lambda_2 > ... > \lambda_k$, entonces las funciones lineales

$$Y_i = \mathbf{a}_i^{'}\mathbf{X}, \quad i = 1, 2, ..., k$$

proporcionan máxima separación entre los G grupos proyectados. Además son ortogonales entre ellas.

Estimación

Supongamos que

- $\mathbf{X}_i$ es una matriz de datos $n_i \times p$ del grupo i=1,...,G

- $\overline{\mathbf{X}}_i = \dfrac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{X}_{ij}$ que es un estimador de μ_i

- $\mathbf{S}_i = \dfrac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i\right)\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i\right)'$ y

- $\overline{\mathbf{X}} = \left(\dfrac{1}{\sum_i n_i}\right)\sum_{i=1}^{G} n_i\overline{\mathbf{X}}_i = \left(\dfrac{1}{\sum_i n_i}\right)\sum_{i=1}^{G}\sum_{j=1}^{n_i}\mathbf{X}_{ij}$ que estima a $\overline{\mu}$

- $\mathbf{S}_{pool} = \sum_{g=1}^{G}\dfrac{n_g - 1}{n - G}\mathbf{S}_g$ es la matriz de covarianza común a los G grupos.

La estimación de $\mathbf{B}$, la correspondiente versión muestral de la suma de cuadrados entre grupos, es

$$\widehat{\mathbf{B}} = \sum_{i=1}^{G}\left(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}\right)\left(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}\right)'$$

con la correspondiente estimación de $\mathbf{W}$, la suma de cuadrados dentro de grupos, dada por

$$\widehat{\mathbf{W}} = \sum_{i=1}^{G}\sum_{j=1}^{n_i}\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i\right)\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i\right)'$$

Discriminante clásico para $G > 2$ grupos

Nuevamente, la idea para generalizar el procedimiento a G poblaciones normales es similar al anterior con dos poblaciones. En este caso, asignaremos el sujeto con covariables $\mathbf{X}$ al grupo $g = 1, 2, ..., G$ sii

$$\pi_g f_g\left(\mathbf{x}\right) > \pi_j f_j\left(\mathbf{x}\right) \quad \forall g \neq j \quad g, j = 1, 2, ..., G$$

Si las probabilidades apriori de pertenencia a cada grupo son iguales, y las matrices de varianza y covarianza son iguales entre los grupos, la condición anterior es equivalente a calcular la distancia de Mahalanobis del punto observado, $\mathbf{X}$, al centriode (vector de medias) de cada

población y clasificarlo en la población que haga mínima esta distancia. Al realizar el proceso que es semejante al caso de dos grupos obtenemos

Las funciones lineales discriminantes tienen las siguientes características:

• Y_1 es la combinación lineal que proporciona el mayor poder de discriminación entre los grupos, y está asociada al valor característico más grande de $\mathbf{W^{-1}E}$.

• Y_2 es la combinación lineal que proporciona el mayor poder discriminador entre los grupos, después de Y_1, y es *ortogonal* a Y_1. Esta función está asociada con el segundo valor característico más grande de $\mathbf{W^{-1}E}$.

Y así sucesivamente. El número máximo de funciones que se puede construir es *k=min(G-1,p)*.

En un proceso de análisis multivariado que lleva inmersa una reducción de dimensión, es muy importante determinar qué tan bien se reproducen los datos en las pocas dimensiones que se consideren para realizar su análisis. Una de las medidas más comunes para determinar lo adecuado de esta reducción de dimension, es el total de varianza explicada por las funciones lineales discriminantes. Una buena representación se logra si la varianza retenida por estas pocas dimensiones está cercana al 100%. El total de la variaza explicada por las primeras $m \leq k$ funciones lineales discriminantes es:

$$\sum_{i=1}^{m} \lambda_i \quad \text{y}$$

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{k} \lambda_i} \times 100\%$$ Porcentaje que explican las primeras m funciones discriminantes

Análisis de las funciones lineales discriminantes

Centroides de los grupos. La media de los puntajes que arrojen las evaluaciones de cada individuo en estas funciones lineales discriminantes. Debería ser una medida inicial de qué tan separados están los grupos *proyectados*. Si la discriminación es buena, deberíamos observar centroides muy alejados uno de otro.

Λ de Wilks. Esta estadística sirve para determinar el poder discriminante de cada una de las funciones discriminantes. Se determina de forma secuencial el número de estas funciones que debemos considerar, a través de la estadística

$$\mathbf{\Lambda} = \frac{\text{Suma de cuadrados dentro de grupos}}{\text{Suma de cuadrados totales}} = \frac{|\mathbf{B}|}{|\mathbf{T}|} = \frac{|\mathbf{B}|}{|\mathbf{W}+\mathbf{B}|}$$

Si la discriminación lograda es buena, entonces la varianza dentro de los grupos será pequeña y la varianza entre grupos será grande. Por lo tanto, $\mathbf{\Lambda}$ estará cercana a cero.

Para este fin, es preferible utilizar el estadístico $\mathbf{V}$ de Barlett, que es una función de $\mathbf{\Lambda}$ y tiene distribución asintótica χ^2. El procedimiento es, inicialmente, considerar sólo una función discriminante, realizar la prueba y, si ésta es significativa, querrá decir que es pertinente la incorporación de otra función discriminante, de lo contrario, será indicativo de que con el número actual de funciones se tiene el máximo poder discriminante; equivalentemente, que la inclusión de otra función no aporta nada a la discriminación entre los grupos.

Las hipótesis a probar mediante este procedimiento son

$$\mathbf{H}_0 : k \text{ funciones lineales son suficientes para discriminar vs.}$$
$$\mathbf{H}_a : \text{ son necesarias más de } k \text{ funciones } \; k = 1, 2, ..., min(G-1, p)$$

La manera de determinar la *importancia relativa de las variables dentro de las funciones discriminantes*, es a través de sus coeficientes estandarizados. La razón es que éstos ya están libres de unidades y son comparables. *La variable que posea el coeficiente estandarizado más grande en valor absoluto*, será la que tiene un *poder discriminante mayor*.

Coeficientes de correlación o de estructura, $\mathbb{C}orr(X_i, \mathbf{Y}_g)$: Correlación lineal entre cada una de las variables y cada una de las funciones lineales. Si esta correlación es grande (cercana a uno en valor absoluto) indica una relación lineal fuerte entre la variable y la función, por tanto, la variable tiene una contribución importante para discriminar entre los grupos. Si está cercana a cero, no tiene poder discriminatorio entre los grupos.

Tasa de error de clasificación: Un elemento muy importante, que determina qué tan bien clasifica nuestro discriminante a las observaciones en la población, es la *tasa de error de clasificación*. Si las covariables utilizadas realmente discriminan a los grupos en la población, esta tasa debe ser pequeña, de lo contrario, será grande y concluiremos que las variables utilizadas, no tienen poder de discriminación entre los grupos en la población.

Cuando se hace esta clasificación con la misma muestra que se utilizó para construir el discriminante, generalmente se logra una tasa de error de clasificación"artificialmente" baja. Una forma más honesta de calcular esta tasa, es a través de la llamada *clasificación cruzada*, que no es más que eliminar uno por uno a las observbaciones en la muestra, y utilizar el discriminante para asignarlas a algunos de los grupos; por lo regular, este procedimiento genera tasas de error más elevadas, pero más realistas.

Regresión multinomial

En el caso de que existan más de dos grupos en el proceso de análisis discriminante, el discriminante logístico se generaliza a una variable de respuesta con más de dos categorías nominales, dando origen al llamado *modelo de regresión multinomial*. En este caso tenemos

$\eta_{ij} = log\left(\frac{\pi_{ij}}{\pi_{iG}}\right) = \alpha_j + \mathbf{X}_i'\beta_j$ entonces

$$\pi_{ij} = P(G_j|\mathbf{X}_i) = \frac{exp\,(\eta_{ij})}{\sum_{i=1}^{p} exp\,(\eta_{ij})}, \quad j = 1, 2, ..., G$$

denota la probabilidad de que el sujeto i pertenezca al grupo j.

Discriminante cuadrático

Supongamos que las poblaciones son normales, pero que, como ocurre regularmente, no existe igualdad de varianzas; en el caso de dos grupos, $\Sigma_1 \neq \Sigma_2$. Entonces, la regla de clasificación, bajo el supuesto de probabilidades a priori iguales, es:

$$\mathbb{Q}(\mathbf{X}) = \frac{1}{2}\left(\Sigma_2^{-1} - \Sigma_1^{-1}\right)\mathbf{X} + \mathbf{X}'\left(\Sigma_2^{-1}\mu_1 - \Sigma_1^{-1}\mu_2\right) + \frac{1}{2}\mu_2'\Sigma_2^{-1}\mu_2 - \frac{1}{2}\mu_1'\Sigma_1^{-1}\mu_1 + \frac{1}{2}log\,|\Sigma_2| - \frac{1}{2}log\,|\Sigma_1|$$

Observemos que el término $\mu_i'\Sigma_i^{-1}\mu_i,\ i = 1, 2$, no puede cancelarse y origina términos de grado 2, ya sean cuadráticos o cruzados, lo que justifica el nombre de discriminante cuadrático.

Esta regla es equivalente a asignar a un individuo $\mathbf{X}_0$ al grupo donde se minimice la función

$$\min_{j\in(1,2)}\left[\frac{1}{2}log\,|\Sigma_j| + \frac{1}{2}\left(\mathbf{X}_0 - \mu_j\right)'\Sigma_j^{-1}\left(\mathbf{X}_0 - \mu_j\right)\right]$$

Para el caso de $G > 2$ grupos, y suponiendo que las matrices de varianza-covarianza no son iguales, la regla se extiene trivialmente como: asignar a un individuo $\mathbf{X}_0$ al grupo donde se minimice la función

$$\min_{j\in(1,\ldots,G)}\left[\frac{1}{2}log\,|\Sigma_j| + \frac{1}{2}\left(\mathbf{X}_0 - \mu_j\right)'\Sigma_j^{-1}\left(\mathbf{X}_0 - \mu_j\right)\right]$$

Análisis de conglomerados (clusters)

El análisis de conglomerados (clusters) es una técnica multivariada, cuyo objetivo es identificar los grupos que subyacen a un conjunto de observaciones. La idea es "descubrir" grupos de observaciones homogéneas y que estén separados de otros grupos. En mercadotecnia, por ejemplo, puede ocurrir que una muestra de consumidores con distintas características, esté formada por un pequeño número de grupos dentro de cada uno de los cuales dichas características sean similares. Esto podría tener implicaciones importantes para determinar una estrategia de mercado apropiada o para investigar la tipología del consumidor. En un contexto educativo, los grupos pueden ser conjuntos de individuos con distintas capacidades (grupos de excelencia, estándar o de bajo rendimiento) o con diversos intereses, que los pueden ubicar en distintas áreas de estudio (orientación vocacional). En Biología podría tratarse de diversos tipos de individuos que pertenecen a una misma especie. En ecología podrían referirse a distintos tipos de plantas. En seguros, podemos agrupar a los sujetos que representan riesgos diferentes en alguna cobertura sobre, por ejemplo, automóviles. En fin, existe un sinnúmero de situaciones reales donde, en algún sentido, se tiene que trabajar con grupos de observaciones o individuos.

Estos métodos se conocen también con el nombre de métodos de *clasificación automática o no supervisada, o de reconocimiento de patrones sin supervisión.* El nombre de no supervisados se aplica para distinguirlos del análisis discriminante, que estudiamos en la sección anterior. Este nombre se debe a que, a diferencia del análisis discriminante, aquí no conocemos la naturaleza de los grupos, de hecho, ni siquiera sabemos el número de grupos, antes de clasificar las observaciones dentro de los clusters.

Objetivos:

• Identificar los grupos que de manera natural se forman con los datos

Estos grupos se forman con base a las similaridades o disimilaridades entre los sujetos, no entre las variables. En este sentido, esta es una técnica de análisis multivariado determinada por los sujetos (casos) y no por las variables como, por ejemplo, en el análisis de componentes principales o en el análisis de factores.

• Podemos decir que esta técnica tiene más fundamento computacional que estadístico

• Es una técnica descriptiva

• Aunque el objetivo común es agrupar a los sujetos, el análisis de conglomerados también se puede utilizar para agrupar variables, de una forma similar al análisis de factores.

Consideraciones antes de realizar el análisis de conglomerados

Al hacer un análisis de conglomerados con un conjunto de datos, nos enfrentamos a una serie de cuestionamientos que debemos dar respuesta para llevar a cabo nuestro objetivo. A saber

• Una primer pregunta es ¿qué variables debemos elegir para realizar los clusters?. Aunque esta es una elección muy importante, raras veces es considerada como tal, y, en la práctica, involucra una mezcla de intuición y disponibilidad de los datos.

• ¿Qué medida de distancia utilizar entre los casos?

• ¿Qué tipo de liga utilizar para los grupos?

• ¿Qué tipo de técnica de construcción de los conglomerados usar?

Pasos en el análisis de conglomerados

• Si las variables no están medidas en la misma escala, es conveniente hacer el análisis con las variables estandarizadas. El objetivo es que las variables con mayores magnitudes no dominen el análisis (similar a Componentes Principales)

• Selección de variables. Como este proceso no proporciona ninguna medida acerca de la importancia de una variable en el análisis, ésta es una decisión que el usuario debe hacer *CUIDADOSAMENTE*.

• Construir y evaluar el modelo de conglomerados

• Identificar la pertenencia (membresía) de los casos a su correspondiente cluster.

Tipos de distancias para los casos, de acuerdo a su escala de medición

La primera decisión importante que se debe tomar es sobre cómo calcular la distancia entre dos observaciones. Es claro que esta elección dependerá de la escala de medición de las variables involucradas en la misma.

En realidad, es bastante subjetivo el hecho de elegir una medida de similitud ya que depende de las escalas de medida. Para variables nominales, generalmente se utilizan medidas de similaridad, mientras que para variable medidas en escala de intervalo o de razón usualmente se consideran matrices de distancias. Se pueden agrupar observaciones (sujetos) según la similitud expresada en términos de una distancia. Si el objetivo es agrupan variables (tipo análisis de factores), es habitual utilizar como medida de similitud los coeficientes de correlación en valor absoluto. Para variables categóricas existen también criterios basados en la posesión o no de los atributos (tablas de presencia-ausencia).

Distancia

Dados dos vectores $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^p$, la distancia entre ellos es una función d con las siguientes propiedades:

i) d: $\mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+ \cup 0$, i.e., $d(\mathbf{x}_i, \mathbf{x}_j) \geq 0$

ii) $d(\mathbf{x}_i, \mathbf{x}_i) = 0, \forall\ \mathbf{x}_i$

iii) $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$ (simetría)

iv) $d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_k) + d(\mathbf{x}_k, \mathbf{x}_j)$ (desigualdad del triángulo)

Variables continuas

• *Distancia euclidiana (la más común)* Supongamos que tenemos dos sujetos con p variables, i.e., $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^p$. Entonces, su distancia euclideana es

$$\begin{aligned} d(\mathbf{x}_1, \mathbf{x}_2) &= \left[\sum_{i=1}^{p} (x_{1i} - x_{2i})^2\right]^{1/2} \\ d^2(\mathbf{x}_1, \mathbf{x}_2) &= (\mathbf{x}_1 - \mathbf{x}_2)^t (\mathbf{x}_1 - \mathbf{x}_2) \text{ (forma vectorial)} \end{aligned}$$

• *Distancia euclideana al cuadrado*

$$d^2(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{p} (x_{1i} - x_{2i})^2$$

• *Distancia de Mahalanobis*

$$d^2(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1 - \mathbf{x}_2)^t \, \mathbf{S}^{-1} \, (\mathbf{x}_1 - \mathbf{x}_2)$$

donde $\mathbf{S}$ es la matriz de covarianzas entre las variables. De este modo, las distancias se ponderan según el grado de relación que exista entre las variables, es decir, si están más o menos correlacionadas. Si la correlación es nula, se obtiene la distancia euclideana.

• *Distancia City block*

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{p} |x_{1i} - x_{2i}|$$

• *Distancia de Minkowski*

$$d(\mathbf{x}_1, \mathbf{x}_2) = \left[\sum_{i=1}^{p} (x_{1i} - x_{2i})^p \right]^{1/p}$$

Si *p=1* tenemos la distancia City block y si *p=2* la distancia euclideana. Si $p = \infty$ se tiene la distancia de Chebychev, dada por

$$D_\infty = \max_{1 \leq i \leq p} |x_{1i} - x_{2i}|$$

• Distancia de Camberra

$$d_{Can}\,(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{p} \frac{|x_{1i} - x_{2i}|}{|x_{1i} + x_{2i}|}$$

Definida como *cero* si $x_{1i} = x_{2i}$.

Variables de conteo (numéricas discretas)

- Ji-cuadrada

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\frac{\sum_i \left(x_{1i} - \mathbb{E}\left(x_{1i}\right)\right)^2}{\mathbb{E}\left(x_{1i}\right)} + \frac{\sum_i \left(x_{2i} - \mathbb{E}\left(x_{2i}\right)\right)^2}{\mathbb{E}\left(x_{2i}\right)}}$$

- phi-cuadrada

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\frac{\frac{\sum_i \left(x_{1i} - \mathbb{E}\left(x_{1i}\right)\right)^2}{\mathbb{E}\left(x_{1i}\right)} + \frac{\sum_i \left(x_{2i} - \mathbb{E}\left(x_{2i}\right)\right)^2}{\mathbb{E}\left(x_{2i}\right)}}{n}}$$

Variables dicotómicas

- Distancia euclidiana

- Distancia euclidiana al cuadrado

Datos binarios (medidas de similaridad)

En este caso se desea medir la similaridad entre los vectores $x_i = \left(x_{i1}, ..., x_{ip}\right)^{'}$ y $x_j = \left(x_{j1}, ..., x_{jp}\right)^{'}$, con la característica particular de que $x_{ik},\ x_{jk} \in \{0, 1\}$, $\forall k = 1, 2, ..., p$. Que generan los siguientes casos

$$\begin{aligned} x_{ik} &= x_{jk} = 1, \\ x_{ik} &= x_{jk} = 0, \\ x_{ik} &= 1, x_{jk} = 0, \\ x_{ik} &= 0, x_{jk} = 0 \end{aligned}$$

Si definimos ahora

$$
\begin{aligned}
a_1 &= \sum_{k=1}^{p} \mathbf{I}\left(x_{ik} = x_{jk} = 1\right) \\
a_2 &= \sum_{k=1}^{p} \mathbf{I}\left(x_{ik} = 0, x_{jk} = 1\right) \\
a_3 &= \sum_{k=1}^{p} \mathbf{I}\left(x_{ik} = 1, x_{jk} = 0\right) \\
a_1 &= \sum_{k=1}^{p} \mathbf{I}\left(x_{ik} = x_{jk} = 0\right)
\end{aligned}
$$

en la práctica, es frecuente el uso de la siguiente medida de similiradidad para este tipo de datos

$$
s_{ij} = \frac{a_1 + \delta a_4}{a_1 + \delta a_4 + \lambda\left(a_2 + a_3\right)}
$$

Donde los valores de δ y λ representan pesos que definen distintas medidas de similaridad. Obsérvese que $a+b+c+d=p$. Las más comunes de estas medidas, se presentan en la tabla siguiente.

Nombre de la medida de similaridad	δ	λ	Definición
Jaccard	0	1	$\frac{a_1}{a_1 + a_2 + a_3}$
Tanimoto	1	2	$\frac{a_1 + a_4}{a_1 + 2\left(a_2 + a_3\right) + a_4}$
Simple Matching (M)	1	1	$\frac{a_1 + a_4}{p}$
Russel and Rao (RR)	–	–	$\frac{a_1}{p}$
Dice	0	1/2	$\frac{2a_1}{2a_1 + \left(a_2 + a_3\right)}$
Kulczynski	–	–	$\frac{a_1}{a_2 + a_3}$

El rango de estas medidas de similaridad está entre cero (mínima similaridad) y uno (máxima similaridad). La idea es ponderar de manera distinta el número de *acuerdos* y *desacuerdos* entre los valores de los vectores binarios observados.

Existen muchísimas medidas de similaridad definidas para datos binarios, estás son las principales, pero, por ejemplo, en los manuales del *innombrable* aparecen las siguientes, además de las ya mencionadas: Sokal and Sneath similarity measure 1, Sokal and Sneath similarity measure 2, Sokal and Sneath similarity measure 3, Ochiai similarity measure, Sokal and Sneath similarity measure 5, Fourfold point correlation (similarity), Binary Euclidean distance, Binary squared Euclidean distance, etc.

Variables continuas (medidas de similaridad)

- Correlación de Pearson

$$r(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{i=1}^{p} (x_{1i} - \bar{x}_{1i})(x_{2i} - \bar{x}_{2i})}{\sqrt{\sum_{i=1}^{p} (x_{1i} - \bar{x}_{1i})^2 \sum_{i=1}^{p} (x_{2i} - \bar{x}_{2i})^2}}$$

- Coseno

$$cos(\theta) = \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{\|\mathbf{x}_1\| \, \|\mathbf{x}_2\|} = \frac{\sum_{i=1}^{p} x_{i1} x_{2i}}{\sqrt{\sum_{i=1}^{p} x_{1i}^2} \sqrt{\sum_{i=1}^{p} x_{2i}^2}}$$

Coeficiente de similaridad de Gower

Es común que las variables involucradas en la construcción de los clusters sean de diversas escalas: continuas, nominales, ordinales, numéricas de conteo y binarias. Del mismo modo que como se calcula la *matriz polycórica* para componentes principales y análisis de factores, requeriríamos construir una matriz de distancias que tomara en cuenta la escala de medición de cada variable. El *Coeficiente general de similaridad de Gower* (1971) es la medida de

similaridad más popular para datos mixtos. Entonces, este coeficiente, S_{ij}, compara dos sujetos (casos) i y j de la siguiente manera:

$$S_{ij} = \frac{\sum_{k=1}^{p} w_{ijk} S_{ijk}}{\sum_{k=1}^{p} w_{ijk}}$$

Con

- S_{ijk} denota contribución a la distancia entre estos sujetos, proporcionada por la *k-ésima* variable, $k = 1, 2, ..., p$.

- Algunas veces la comparación entre estos individuos no es posible, debido a pérdida de información (datos faltantes) o, en el caso de variables dicotómicas, a que la característica no está presente en ninguno de los dos sujetos i y j. Por lo tanto, w_{ijk} sólo puede tomar los valores 1 ó 0.

- La medida de similitud, S_{ijk}, en el índice de Gower para variables ordinales y continuas, se calcula de la siguiente manera:

$$S_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{r_k}$$

  Con r_k el rango de la *k-ésima* variable.

- En caso de que las variables sean *nominales*, S_{ijk} es *1* si $x_{ik} = x_{jk}$ ó *0* si $x_{ik} \neq x_{jk}$. Entonces, $S_{ijk} = 1$ si los sujetos i y j, tienen la misma categoría en la variable k, ó *0* si tienen diferente categoría. Y $w_{ijk} = 1$ si ambos casos tienen observaciones (alguna categoría) de la variable k.

- Para variables binarias (o dicotómicas), la medida de similitud, S_{ijk}, de este índice y los pesos w_{ijk}, se definen de acuerdo a la siguiente tabla:

	valor del atributo			
Sujeto i	+	+	−	−
Sujeto j	+	−	+	−
S_{ijk}	1	0	0	0
w_{ijk}	1	1	1	0

Donde +, denota que el atributo está presente y **-** que está ausente.

Nota: Si todas las variables son binarias, el índice de Gower es equivalente al coeficiente de similaridad de Jaccard.

Entonces, estas funciones de distancia transforman nuestra matriz de datos $\mathbf{X}_{n\times p}$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

en una matriz de distancias o similaridades, $\mathbf{D}_{n\times n}$, entre los n sujetos

$$\mathbf{D} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

Distancias entre clusters

Una vez que se ha definido qué distancia conviene usar para los casos, debemos decidir cómo se habrá de calcular la distancia de un individuo a un conglomerado y la distancia entre los conglomerados. Para este fin, se tiene las siguientes medidas, conocidas en la literatura de análisis de conglomerados, como *ligas*. Ilustraremos cada una de estas ligas con la matriz de distancia

Distancias

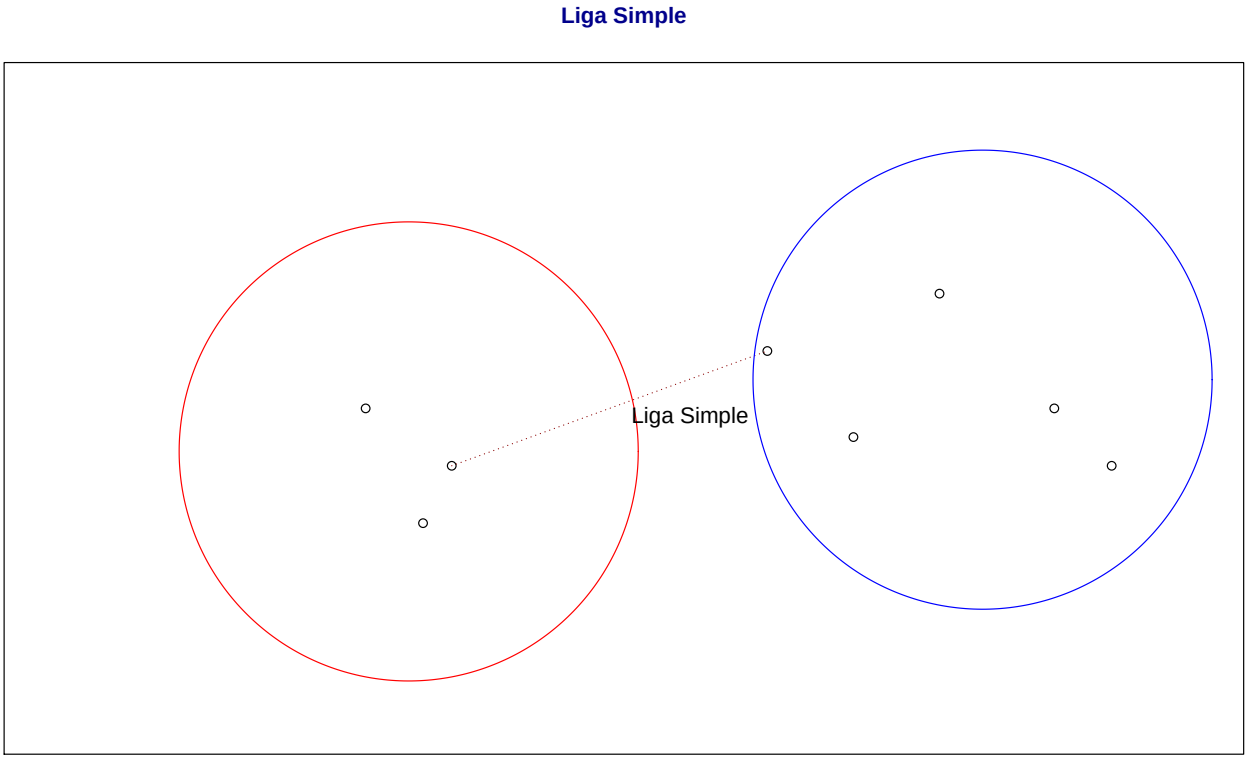
	1	**2**	**3**	**4**	**5**
1	0	9	3	6	11
2	9	0	7	5	10
3	3	7	0	9	2
4	6	5	9	0	8
5	11	10	2	8	0

Vecinos cercanos o liga simple

Aquí la distancia entre dos conglomerados es la distancia entre sus sujetos más cercanos.

En términos matemáticos, si tenemos un cluster **R** y otro **S**, entonces la distancia es:

$$d(R, S) = min(d_{ij},\ i \in \mathbf{R}, j \in \mathbf{S})$$



Observando las distancias entre todos los sujetos en nuestra matriz de distancias, $\mathbf{D}$, observamos que la mínima de éstas es *3* y corresponde a los sujetos $(3,5)$, por lo que son los primeros en unirse y lo hacen a una *altura=2*. Ahora bien, esta unión de sujetos ya constituye un grupo, por lo que hay que calcular las distancias de este grupo al resto de los elementos utilizando la liga simple, es decir, hay que calcular

$$\begin{aligned} d\{(3,5),1\} &= min\{d(3,1),d(5,1)\} = min\{3,11\} = 3 \\ d\{(3,5),2\} &= min\{d(3,2),d(5,2)\} = min\{7,10\} = 7 \\ d\{(3,5),4\} &= min\{d(3,4),d(5,4)\} = min\{9,8\} = 8 \end{aligned}$$

La nueva matriz de distancias $\mathbf{D}_1$, se obtiene considerando estas distancias calculadas con este primer grupo formado. En concreto tenemos

$\mathbf{D}_1$

	(3,5)	**1**	**2**	**4**
(3,5)	0	3	7	8
1	3	0	9	6
2	7	9	0	5
4	8	6	5	0

Realizando el mismo proceso inicial, la distancia mínima entre estos grupos es *3* y corresponde a los grupos $(3,5)$ y *1*. Entonces, el siguiente agrupamiento genera al grupo $(1,3,5)$. Nuevamente, debemos calcular la distancia de este grupo a cada uno de los otros elementos, a través de la liga simple

$$\begin{aligned} d\{(1,3,5),2\} &= min\{d(1,2),d(3,2),d(5,2)\} = min\{9,7,10\} = 7 \\ d\{(1,3,5),4\} &= min\{d(1,4),d(3,4),d(5,4)\} = min\{6,9,8\} = 6 \end{aligned}$$

Por lo que nuestra nueva matriz de distancias es

$\mathbf{D}_2$

	(1,3,5)	**2**	**4**
(1,3,5)	0	7	6
2	7	0	5
4	6	5	0

La distancia mínima en esta matriz es *5* y corresponde a la de los sujetos $(2,4)$, que constituyen el siguiente grupo formado y que, como vemos, es diferente al que ya habíamos constituido. Ahora debemos encontrar la distancia entre estos dos grupos

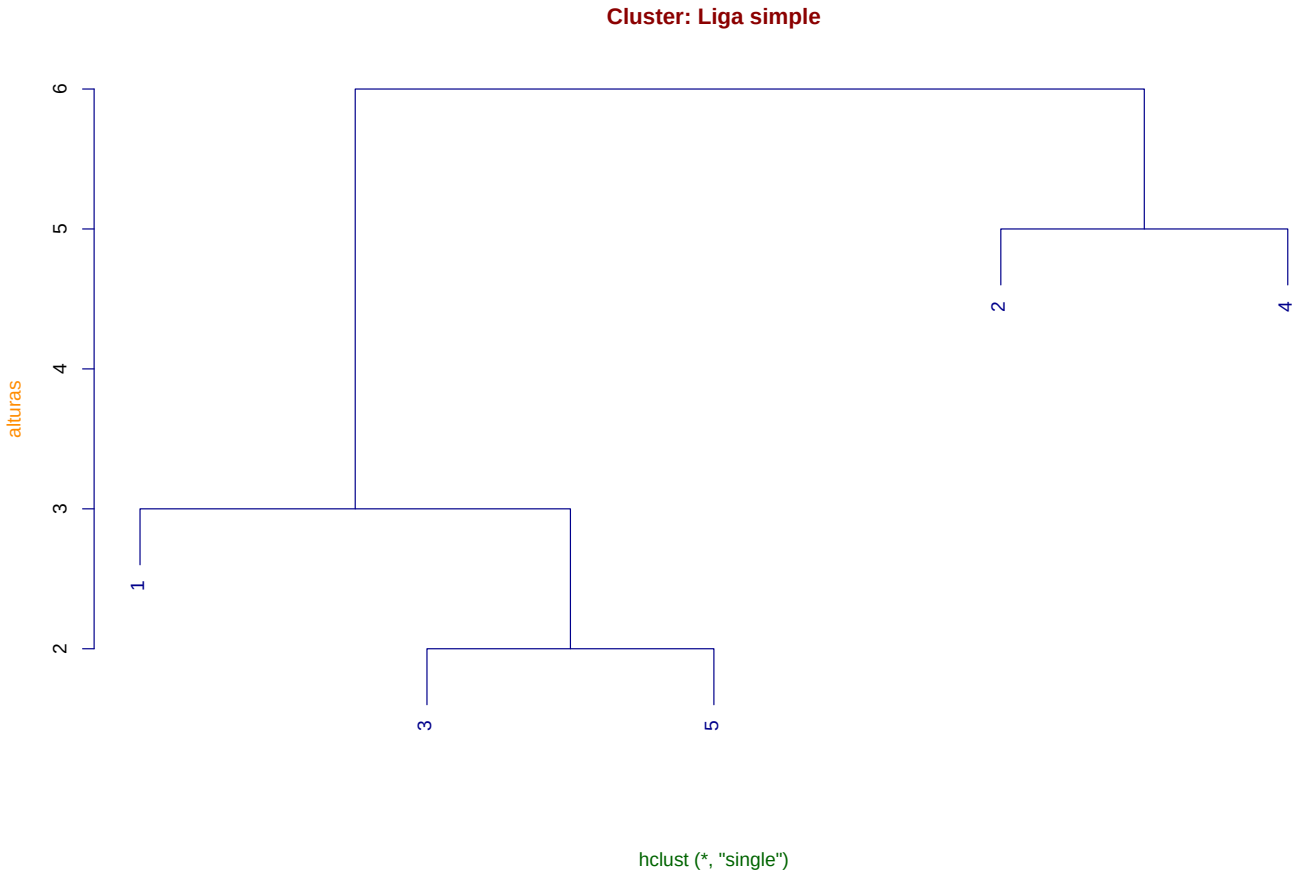
$$d\{(1,3,5),(2,4)\} = min\{d(1,2),d(3,2),d(5,2),d(1,4),d(3,4),d(5,4)\} = min\{9,7,10,6,9,8\} = 6$$

Y la última de nuestras matrices es

$\mathbf{D}_3$

	(1,3,5)	**(2,4)**
(1,3,5)	0	6
(2,4)	6	0

Esta distancia a la que se unen los dos grupos formados, constituye la distancia máxima a la que se unen *todas las observaciones* y genera un único grupo, como ya sabemos que debe suceder con un algoritmo *aglomerativo*. Observemos que esta agrupación, así como las distancias (alturas) de unión entre los grupos, coincide totalmente con el *dendrograma* mostrado en la gráfica siguiente.

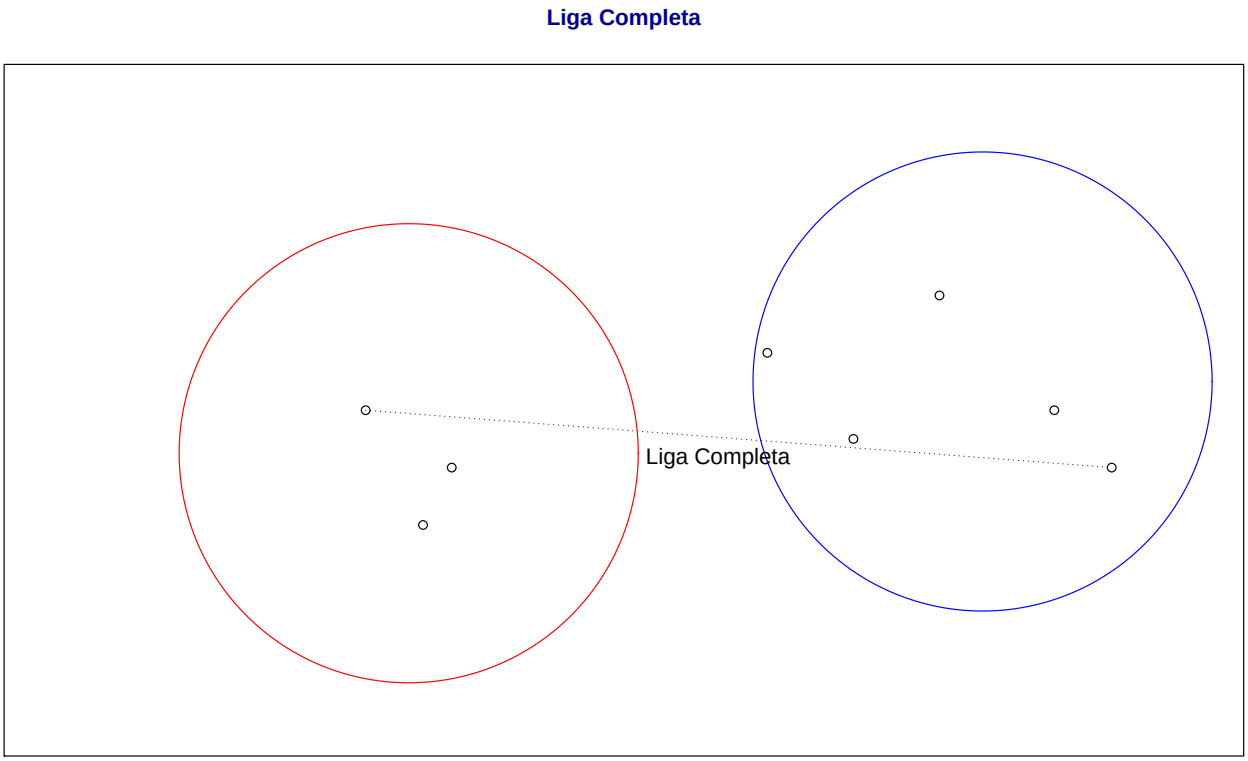


Vecinos lejanos o liga compuesta

Aquí La distancia entre dos conglomerados es la distancia entre sus dos sujetos más lejanos

En términos matemáticos, si tenemos un cluster **R** y otro **S**, entonces la distancia es:

$$d(R,S) = max(d_{ij},\ i \in \mathbf{R}, j \in \mathbf{S})$$



Para esta liga, el primer conglomerado se hace igual que en la liga simple, y está constituido por $(3,5)$, y se unen a altura 2. Una vez establecido este grupo, procedemos a encontrar su distancia al resto de los elementos, utilizando la liga completa, de la siguiente forma

$$\begin{aligned} d\{(3,5),1\} &= max\{d(3,1),d(5,1)\} = max\{3,11\} = 11 \\ d\{(3,5),2\} &= max\{d(3,2),d(5,2)\} = max\{7,10\} = 10 \\ d\{(3,5),4\} &= max\{d(3,4),d(5,4)\} = max\{9,8\} = 9 \end{aligned}$$

La nueva matriz de distancias $\mathbf{D}_1$, se obtiene considerando estas distancias calculadas con

este primer grupo formado. En concreto tenemos

$\mathbf{D}_1$

	(3,5)	**1**	**2**	**4**
(3,5)	0	11	10	9
1	11	0	9	6
2	10	9	0	5
4	9	6	5	0

La distancia mínima en esta matriz corresponde a los individuos $(2, 4)$, con altura=5, que forman el siguiente grupo. El siguiente paso es calcular la distancia entre estos nuevos grupos, con la liga completa.

$$\begin{aligned} d\{(3,5),(2,4)\} &= max\{d(3,2), d(3,4), d(5,2), d(5,4)\} = max\{7,9,10,8\} = 10 \\ d\{(3,5),1\} &= max\{d(3,1), d(5,1)\} = max\{3,11\} = 11 \\ d\{(2,4),1\} &= max\{d(2,1), d(4,1)\} = max\{9,6\} = 9 \end{aligned}$$

con lo que generamos la siguiente matriz

$\mathbf{D}_2$

	(3,5)	**1**	**(2,4)**
(3,5)	0	11	10
1	11	0	9
(2,4)	10	9	0

observamos que la distancia más pequeña es *9*, y corresponde a la unión de los grupos $(2, 4)$ y *1*, que generan el grupo $(1, 2, 4)$. Volvemos a calcular la distancia entre los grupos $(3, 5)$ y $(1, 2, 4)$ a través de la liga completa.

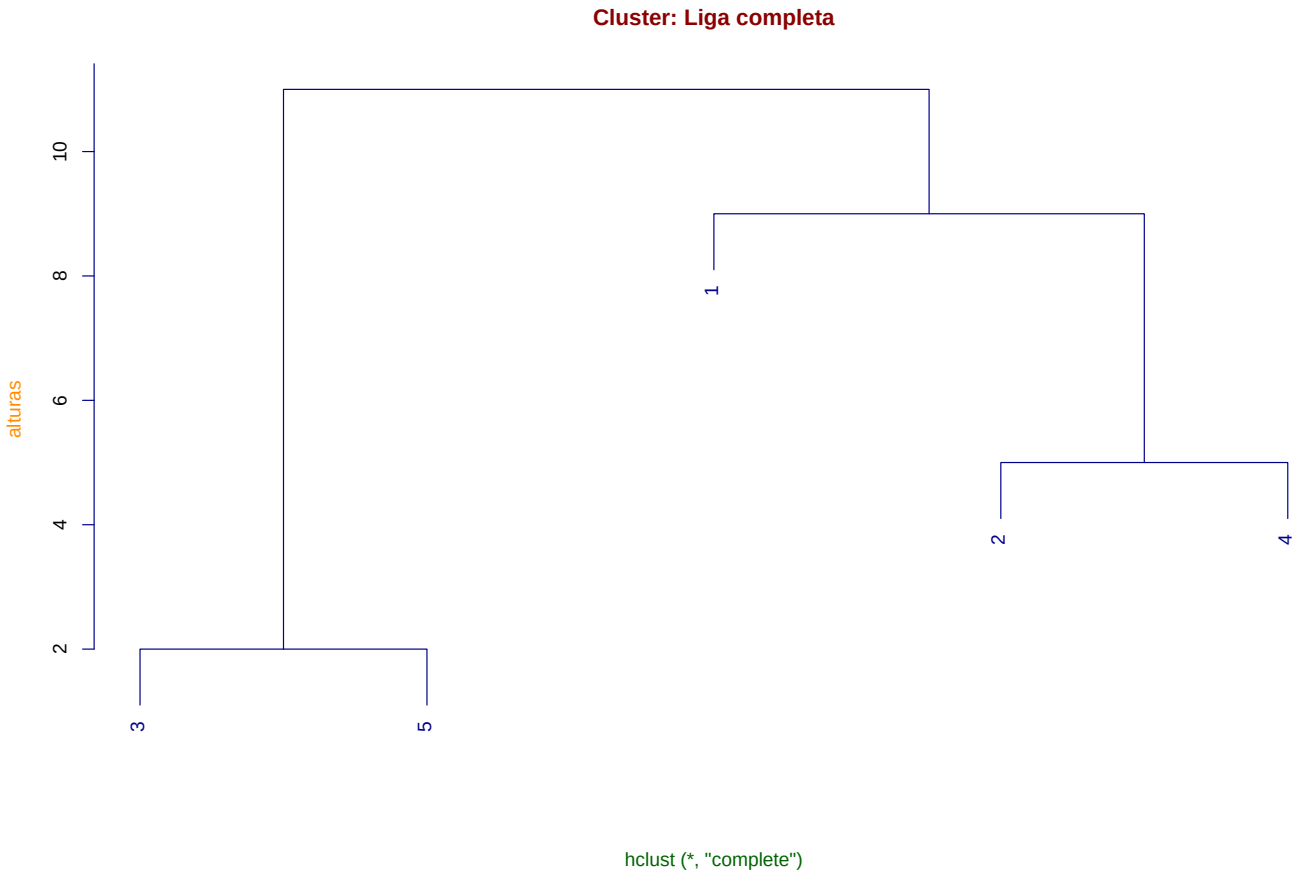
$$\begin{aligned} d\{(3,5),(1,2,4)\} &= max\{d(3,1), d(3,2), d(3,4), d(5,1), d(5,2), d(5,4)\} \\ &= max\{3,7,9,11,10,8\} = 11 \end{aligned}$$

y genera la matriz

$\mathbf{D_3}$

	(3,5)	**(1,2,4)**
(3,5)	0	11
(1,2,4)	11	0

esta es la distancia a la que todas las observaciones se unen en un solo grupo. Nuevamente observamos que los grupos formados y las distancias (alturas) que calculamos, coinciden con la gráfica de esta liga.

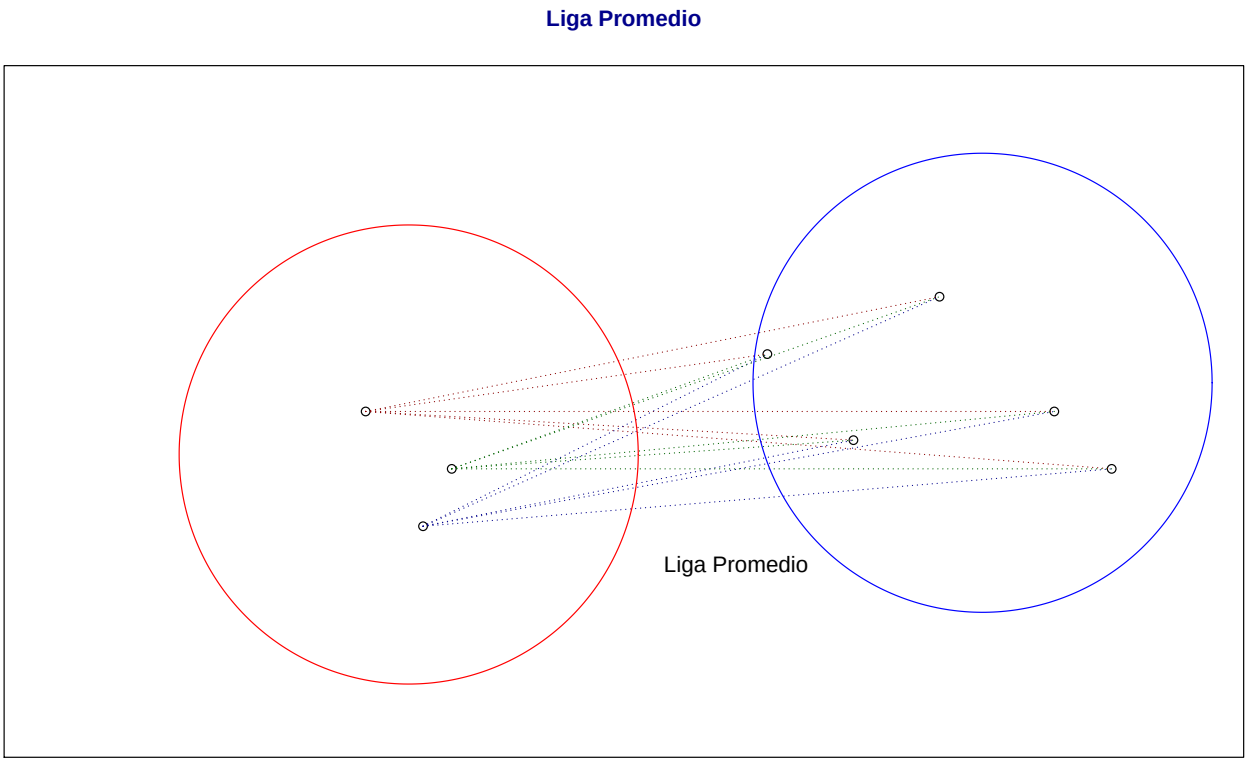


Liga promedio

Es la distancia promedio entre todas las posibles distancias intra o inter clusters. Apropiada, cuando el investigador asume que los grupos son homogéneos.

En símbolos, si tenemos un cluster $\mathbf{R}$ con n_R elementos, y otro $\mathbf{S}$, con n_S elementos, entonces la distancia es:

$$d(R,S) = \frac{1}{n_R}\frac{1}{n_S}\sum_{i\in\mathbf{R}}\sum_{j\in\mathbf{S}} d_{ij}$$



Igual que para los dos casos anteriores, las observaciones que se unen inicialmente, son $(3,5)$, que se unen a altura *2*. Una vez que se obtiene este cluster, hay que calcular sus distancia al resto de los elementos, utilizando la liga promedio. Es decir

$$
\begin{aligned}
d\{(3,5),1\} &= \frac{1}{2}\{d(3,1)+d(5,1)\} = \frac{1}{2}(3+11) = 7 \\
d\{(3,5),2\} &= \frac{1}{2}\{d(3,2)+d(5,2)\} = \frac{1}{2}(7+10) = 8.5 \\
d\{(3,5),4\} &= \frac{1}{2}\{d(3,4)+d(5,4)\} = \frac{1}{2}(9+8) = 8.5
\end{aligned}
$$

y la correspondiente matriz de distancias es ahora

$\mathbf{D}_1$

	(3,5)	**1**	**2**	**4**
(3,5)	0	7	8.5	8.5
1	7	0	9	6
2	8.5	9	0	5
4	8.5	6	5	0

La distancia mínima en esta matriz corresponde a las observaciones $(2,4)$, que forman un nuevo grupo, que se une a *altura=5*. Nuevamente debemos calcular la distancia entre estos clusters mediante la liga promedio.

$$
\begin{aligned}
d\{(3,5),(2,4)\} &= \frac{1}{4}\{d(3,2)+d(5,4)+d(5,2)+d(5,4)\} = \frac{1}{4}(7+9+10+8) = 8.5 \\
d\{(3,5),1\} &= \frac{1}{2}\{d(3,1)+d(5,1)\} = \frac{1}{2}(3+11) = 7 \\
d\{(2,4),1\} &= \frac{1}{2}\{d(2,1)+d(4,1)\} = \frac{1}{2}(9+6) = 7.5
\end{aligned}
$$

que genera la matrz de distancias

$\mathbf{D}_2$

	(3,5)	**1**	**(2,4)**
(3,5)	0	7	8.5
1	7	0	7.5
(2,4)	8.5	7.5	0

cuya distancia mínima es *7* y corresponde a la unión de los grupos $(3,5)$ y *1*, que originan el grupo $(1,3,5)$. La distancia entre estos grupos es

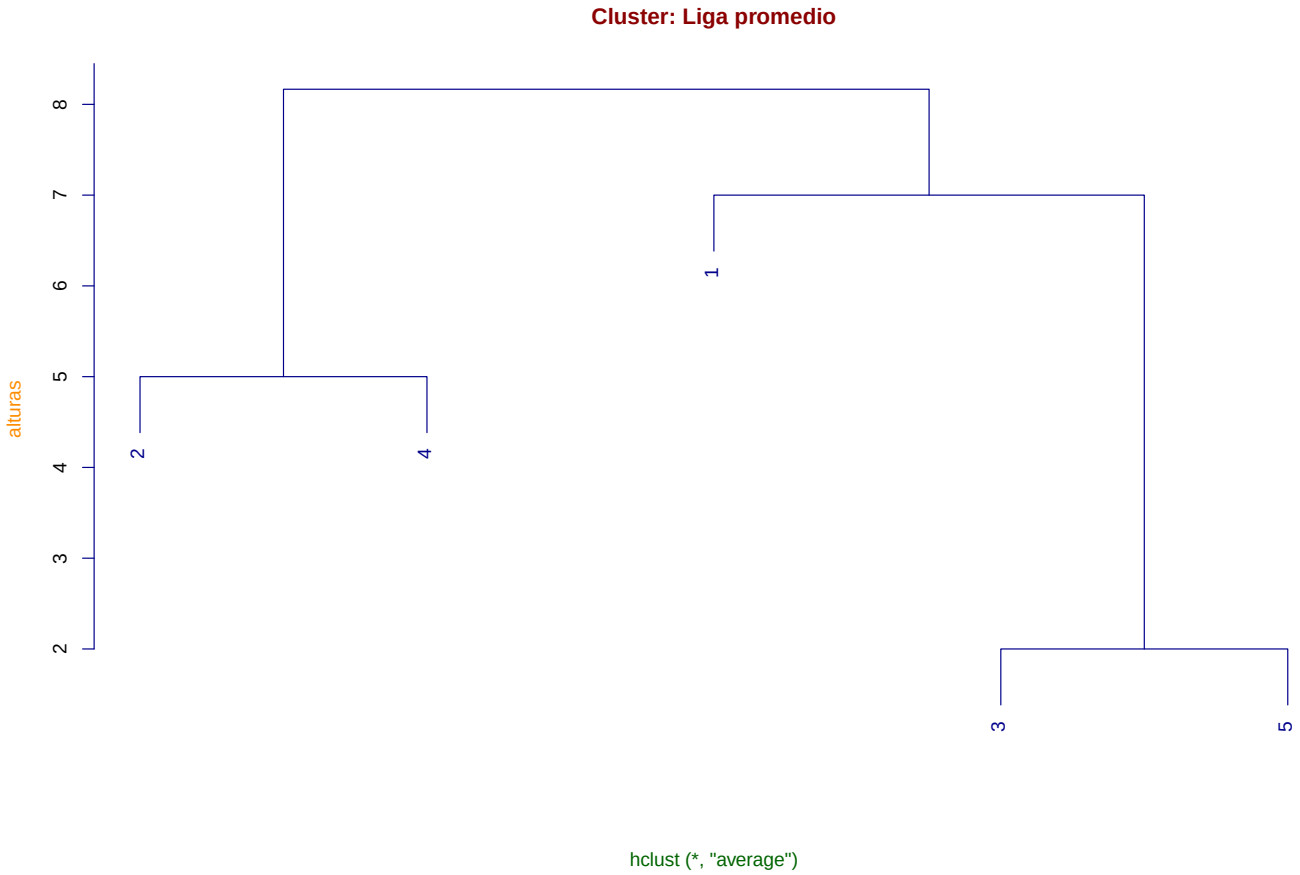
$$
\begin{aligned}
d\left\{(1,3,5),(2,4)\right\} &= \frac{1}{6}\left\{d(1,2)+d(3,2)+d(5,2)+d(1,4)+d(3,4)+d(5,4)\right\} \\
&= \frac{1}{6}(9+7+9+6+10+8) = 8.166
\end{aligned}
$$

que genera la matriz

$$\mathbf{D}_3$$

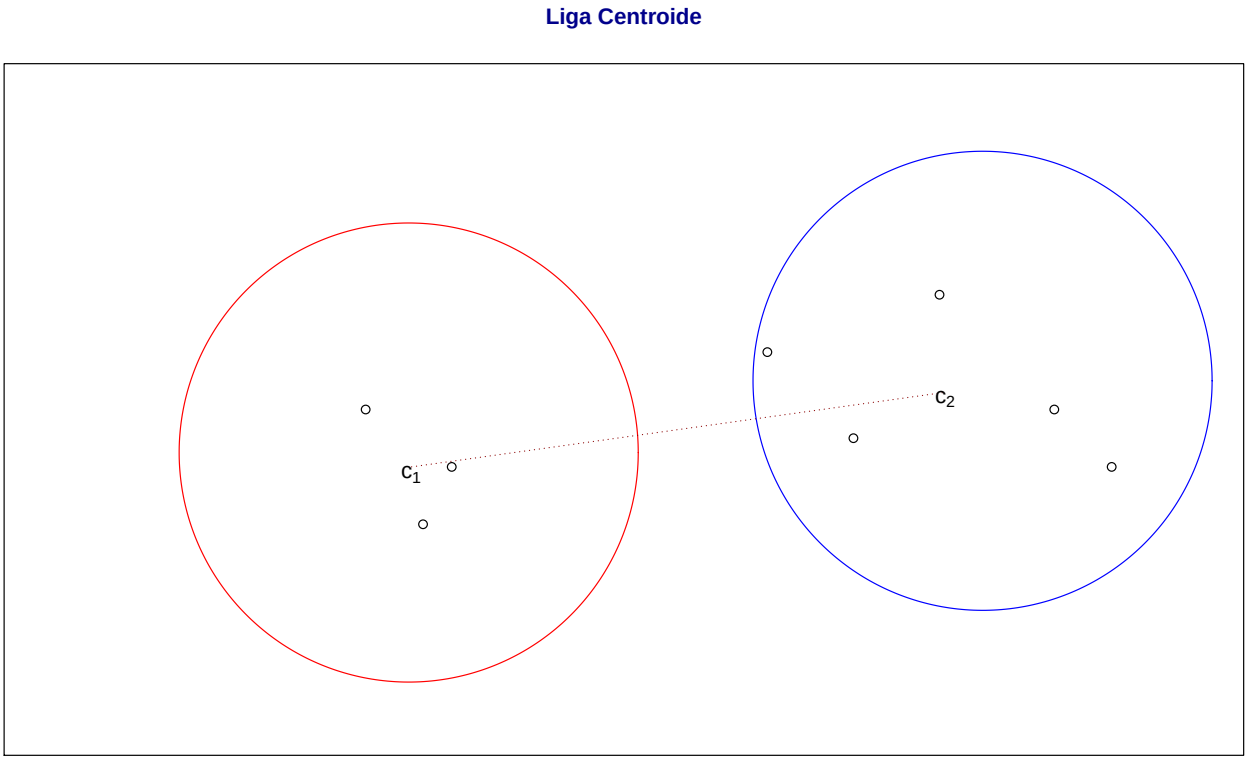
	(1,3,5)	**(2,4)**
(3,5)	0	8.166
(2,4)	8.166	0

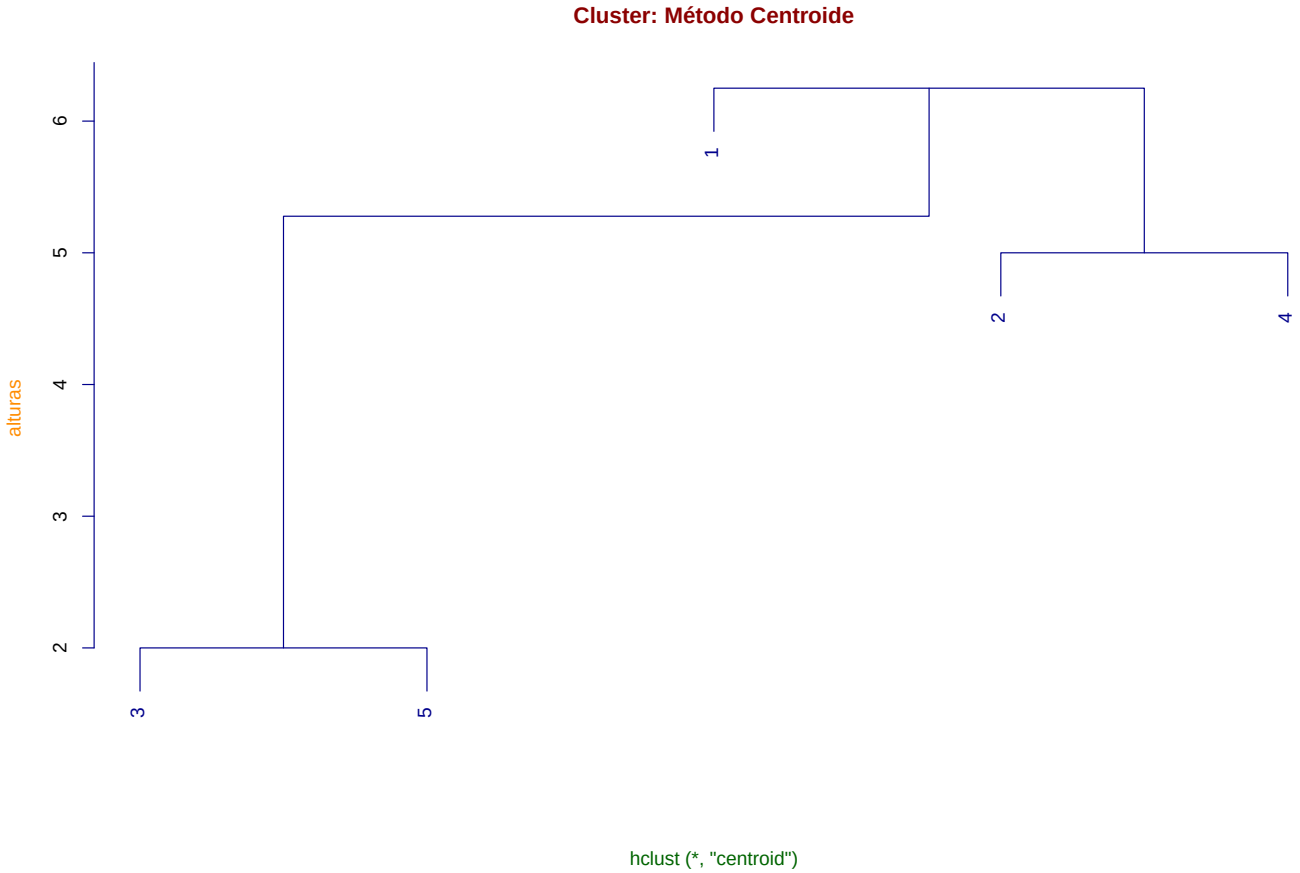
entonces, la distancia final a la que se unen todos los grupos es *8.166*. Observe que los grupos y las distancias a las que se unen, coinciden con la gráfica correspondiente a esta distancia promedio.



Liga del centroide

Se usa generalmente sólo con variables continuas. Para esta liga, la distancia entre dos clusters es la distancia euclideana entre sus centros (centroides), que son los vectores de medias de las observaciones que pertenecen al grupo.





Método de Ward

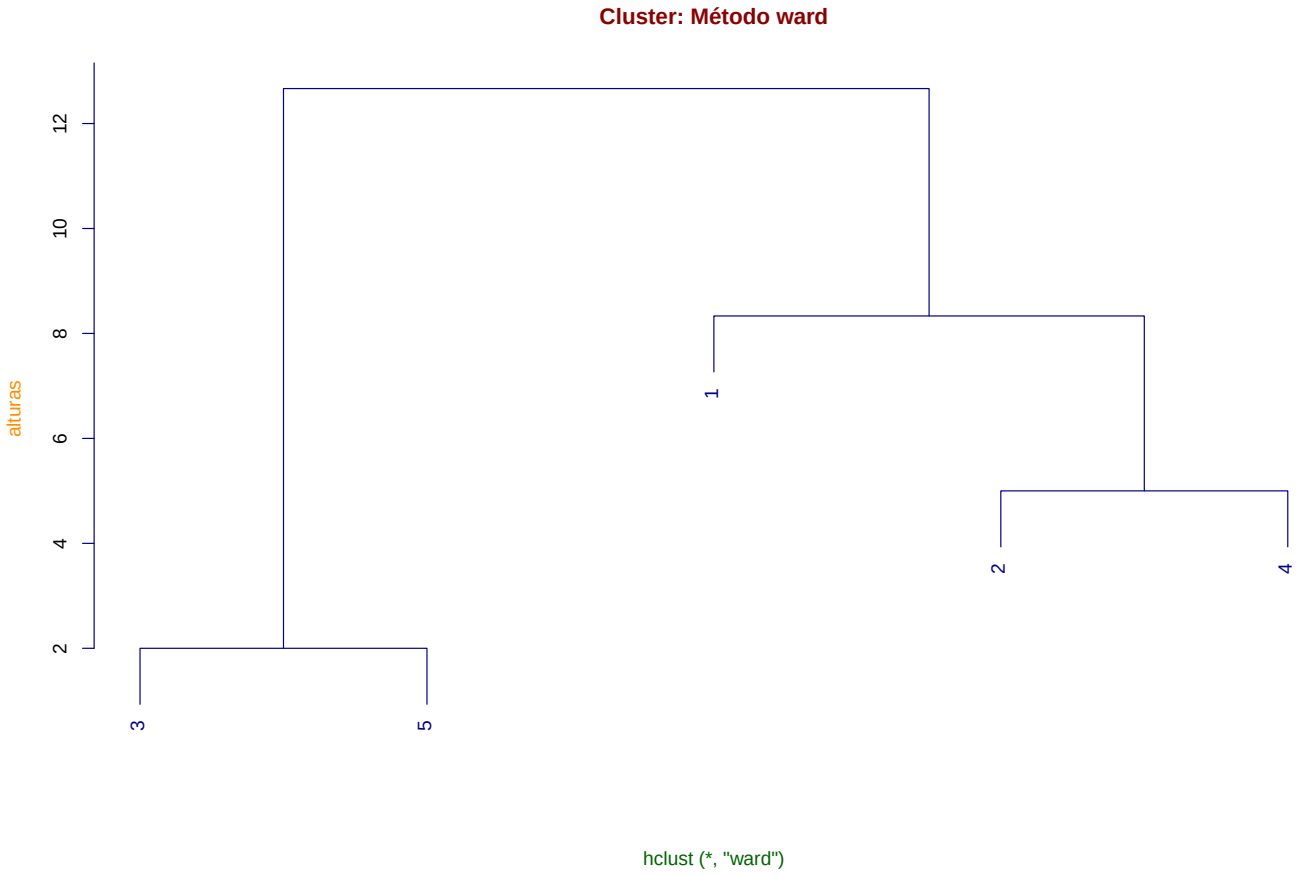
Este método es un proceso diferente a los anteriores, para construir un cluster jerárquico, y fue propuesto por Ward y Wishart. A diferencia con los métodos anteriores, aquí se parte de los elementos directamente, sin necesidad de construir la matriz de distancias, y se define una medida global de la heterogeneidad de observaciones aglutinadas en grupos. Dicha medida está dada por la suma de las distancias euclideanas al cuadrado entre cada elemento y la media de su grupo:

$$\mathbf{W} = \sum_{g} \sum_{i \in g} \left(\mathbf{X}_{ig} - \bar{\mathbf{X}}_g\right)' \left(\mathbf{X}_{ig} - \bar{\mathbf{X}}_g\right)$$

expresión ya conocida por nosotros. Donde $\bar{\mathbf{X}}_g$ es la media del grupo g.

El proceso de construcción inicia suponiendo que cada elemento forma un grupo, entonces, $g = n$ y por tanto $\mathbf{W}$ es cero. En el siguiente paso se unen los elementos que produzcan el incremento mínimo de $\mathbf{W}$. Obviamente, esto implica tomar los elementos más cercanos utilizando la distancia euclideana. Una vez concluido este paso, ahora tenemos $n-1$ grupos;

$n-2$ de ellos de un elemento y uno con dos elementos. Utilizamos el mismo criterio para decidir qué grupos debemos unir ahora para que **W** crezca lo menos posible, con lo que obtenemos $n-2$ grupos, y así sucesivamente hasta tener un único grupo (proceso aglomerativo). Los valores de **W** indican el crecimiento del criterio al formar grupos y pueden utilizarse para decidir cuántos grupos contienen nuestros datos, de forma natural.



¿Cuál es el mejor?

No existen reglas generales que lleven a la preferencia de una liga o método sobre otro; aunque los más utilizados son los tres últimos. Una recomendación sensata es tratar de analizar cuál es el criterio más razonable de acuerdo a la naturaleza de los datos por agrupar. De no funcionar esta estrategia, es conveniente probar con algunas ligas y comparar sus resultados.

El dendrograma

El *dendrograma*, también conocido como *árbol jerárquico*, es la representación gráfica del proceso de agrupamiento en forma de árbol. Los criterios para definir distancias entre grupos que hemos presentado, tienen la propiedad de que, si consideramos tres grupos, **A**, **B**, **C**, se cumple que

$$d(\mathbf{A},\mathbf{C}) \leq max\{d(\mathbf{A},\mathbf{B}) + d(\mathbf{B},\mathbf{C})\}$$

una medida de distancia que tiene esta propiedad se denomina *ultramétrica*. Esta propiedad de "desigualdad del triángulo" es más fuerte que la propiedad de desigualdad del triángulo estándar, ya que una ultramétrica es siempre una métrica. Ya que si $D^2(\mathbf{A},\mathbf{C})$ es menor o igual a que el máximo de $D^2(\mathbf{A},\mathbf{B})$ y $D^2(\mathbf{B},\mathbf{C})$, entonces, necesariamente será menor o igual que su suma: $D^2(\mathbf{A},\mathbf{B})+D^2(\mathbf{B},\mathbf{C})$. Entonces, el dendrograma es la representación gráfica de una ultramétrica, y su construcción es como sigue:

• En la parte inferior de la gráfica aparecen los n individuos u objetos a ser agrupados.

• Las uniones entre elementos están compuestas por tres lineas rectas. Dos verticales originadas en los elementos que se unen, y que son perpendiculares al eje de los elementos y una paralela a este eje que se sitúa al nivel (distancia o altura) a la que se unen.

• El proceso se repite hasta que todos los elementos estén unidos por lineas rectas.

Si se corta el dendrograma a un nivel o altura específico, se obtiene una visualización del número de grupos existentes a ese nivel, y los elementos que los conforman. Por supuesto, si se hacen cortes a alturas diferentes en el dendrograma, es posible que se tengan agrupaciones diferentes, por lo que esta decisión es fundamental para este análisis.

El dendrograma es útil cuando los grupos tienen una clara definición, pero puede ser poco útil si no es así. Ya que es necesario un despliegue gráfico, si el número de elementos por clasificar es muy grande, será muy difícil, casi imposible, identificar grupos visualmente bajo este procedimiento.

Métodos no jerárquicos

Métodos de partición

Otro grupo importante de procesos para construir clustes, son los métodos *de partición o partitivos*. Igual que en los métodos aglomerativos, existen diferentes algoritmos, no obstante, el algoritmo conocido como *K-medias* es el más importante de ellos. Es conveniente decir que este algoritmo requiere que el número de grupos, K, se fije o decida de antemano. El algoritmo de K-medias sigue una lógica enteramente distinta a la de los métodos jerárquicos. Este algoritmo no se basa en medidas de distancia, pero utiliza la variación *intra-cluster* como una medida para formar grupos homogéneos. Específicamente, el procedimiento tiene como objetivo la segmentación de los datos de tal manera que la variación dentro del grupo se reduzca al mínimo. Por consiguiente, no es necesario utilizar una medida de distancia en el primer paso del análisis.

Este procedimiento para formar clusters, inicia con la asignación aleatoria de los sujetos al número de clusters definido inicialmente. Este paso inicial puede requerir

- Seleccionar K puntos arbitrariamente como centros de los grupos iniciales y, posteriormente, asignar los sujetos más cercanos a dichos puntos

- Tomar como centros los K puntos más alejados entre sí.

- Determinar los grupos con información a priori, o bien seleccionar los centros a priori.

Una vez realizado este paso inicial, los sujetos son reasignados sucesivamente a otros cluster para minimizar la variación dentro de clusters que, básicamente, es el cuadrado de la distancia de cada observación al centro de su cluster asociado. Si la asignación de un sujeto a otro cluster decrece la variación intra clusters, esta observación es reasignada a dicho cluster, y se recalculan las medias de los grupos. Este proceso genera clusters internamente homogéneos y externamente heterogéneos

Con los métodos jerárquicos, una observación permanece en el cluster una vez que ha sido asignada a él, pero con K-medias, la pertenencia a los clusters puede cambiar en el curso del proceso de agrupamiento. En consecuencia, K-medias no trabaja como un procedimiento jerárquico.

Es necesario realizar un análisis previo para decidir cuántos grupos subyacen al conjunto de datos que queremos agrupar.

En este caso, el criterio de homogeneidad sería minimizar las distancias al cuadrado entre los centros de los grupos y los puntos que pertenecen a ese grupo. Si medimos las distancias con la norma euclideana, este criterio se escribe:

$$\min \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(X_{ik} - \bar{X}_k\right)' \left(X_{ik} - \bar{X}_k\right) = \min \sum_{k=1}^{K} \sum_{i=1}^{n_k} d^2\left(i,k\right)$$

donde $d^2\left(i,k\right)$ es el cuadrado de la distancia euclideana entre la observación i del grupo k y su media de grupo.

Otro criterio de homogeneidad que se utiliza en el algoritmo de K-medias, es la suma de cuadrados dentro de los grupos ($\mathbf{W}$) para todas las variables, que es equivalente a la suma ponderada de las varianzas de las variables en los grupos:

$$\mathbf{W} = \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{i=1}^{n_k} \left(X_{ijk} - \bar{X}_{jk}\right)^2$$

con X_{ijk} el valor de la variable j de la observación i en el grupo k, y $\bar{X}_{jk}$ la media de esta variable en el grupo. El criterio se escribe como

$$\mathbf{W} = \sum_{k=1}^{K} \sum_{j=1}^{p} n_k s_{jk}^2$$

donde n_k es el número de elementos del grupo k y s_{jk}^2 es la varianza de la variable j en dicho grupo. La varianza de cada variable en cada grupo es claramente una medida de la heterogeneidad del grupo y al minimizar las varianzas de todas las variables en los grupos obtendremos grupos más homogéneos.

Podemos comprobar que ambos criterios son idénticos. Ya que un escalar es igual a su traza, podemos escribir el primer criterio como

$$\min \sum_{k=1}^{K} \sum_{i=1}^{n_k} traza \left[d^2 (i,k) \right] = \min traza \left[\sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(X_{ik} - \bar{X}_k \right) \left(X_{ik} - \bar{X}_k \right)' \right]$$

si denotamos como $\mathbf{W}_1$ a la matriz de suma de cuadrados dentro de los grupos,

$$\mathbf{W}_1 = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(X_{ik} - \bar{X}_k \right) \left(X_{ik} - \bar{X}_k \right)'$$

tenemos que

$$min\ traza\left(\mathbf{W}_1\right) = min\ traza\left(\mathbf{W}\right)$$

Como la traza es la suma de los elementos de la diagonal principal, ambos criterios coinciden. Este criterio se denomina *criterio de la traza*, propuesto por Ward (1963).

El proceso de maximización de este criterio implicaría calcularlo para todas las posibles particiones, que resulta obviamente imposible, salvo para valores de n muy pequeños. El algoritmo de k medias busca la partición óptima con la restricción de que en cada iteración sólo se permite mover un elemento de un grupo a otro. El algoritmo funciona como sigue

1. Arrancar con una partición inicial

2. Verificar si cambiando algún elemento se reduce $\mathbf{W}$

3. Si es posible reducir $\mathbf{W}$ reasignar el elemento; recalcular las medias de los dos grupos afectados por el cambio y volver a *(2)*. Si no es posible reducir $\mathbf{W}$, terminar.

Es claro que el resultado del algortimo puede depender de la asignación de elementos en la partición inicial y del orden de los elementos. Es conveniente siempre repetir el algoritmo con distintos valores iniciales y permutando los elemento de la muestra. El efecto del orden de las observaciones suele ser pequeño, pero conviene asegurarse en cada caso de que no está afectando el procedimiento.

Escalamiento Multidimensional

La última de las técnicas estadísticas de análisis multivariado que mostraremos, es la conocida como *Escalamiento Multidimensional (EM)* (*Multidimensional Scaling (MDS)*). Las técnicas de escalamiento multidimensional son una generalización de la idea de componentes principales cuando, en lugar de disponer de una matriz de observaciones por variables, como en componentes principales, se dispone de una matriz de *distancias* o *similaridades* o *proximidades* o *disimilaridades*. Los objetos de esta matriz pueden ser similaridades, δ_{ij}, en las que un valor más grande de δ_{ij} corresponde a una mayor semejanza entre los objetos i y j; o disimilaridades, donde un valor grande indicaría muy poca semejanza (o una mayor desemejanza) entre ellos.

Por ejemplo, esta matriz puede representar las similitudes o distancias entre n productos fabricados por una empresa; las distancias percibidas entre n candidatos políticos; las diferencias entre n preguntas de un cuestionario o examen o las distancias o similitudes entre n sectores industriales. Estas distancias pueden haberse obtenido a partir de ciertas variables, o pueden ser el resultado de una estimación directa, por ejemplo, preguntando a un grupo de jueces por sus opiniones sobre las similaridades entre los elementos sometidos a su juicio.