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 pedadogía 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 simultaneamente 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 támbién 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 *pxp*, 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 $\mathbf{A}$ *pxp* 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 $\mathbf{A}$ *pxp* 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

$$\mathbf{A} = \lambda_1 \mathbf{e_1} \mathbf{e_1'} + \lambda_\mathbf{2} \mathbf{e_2} \mathbf{e_2'} + \cdots + \lambda_\mathbf{p} \mathbf{e_p} \mathbf{e_p'}$$

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

$\mathbf{\Lambda} = \mathbf{diag}\{\lambda_\mathbf{1}, \lambda_\mathbf{2}, ..., \lambda_\mathbf{p}\}$. Algunas veces se supone $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.

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

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

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 $\mathbb{E}(\mathbf{A\ X\ B}) = \mathbf{A}\mathbb{E}(\mathbf{X})\mathbf{B}$

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

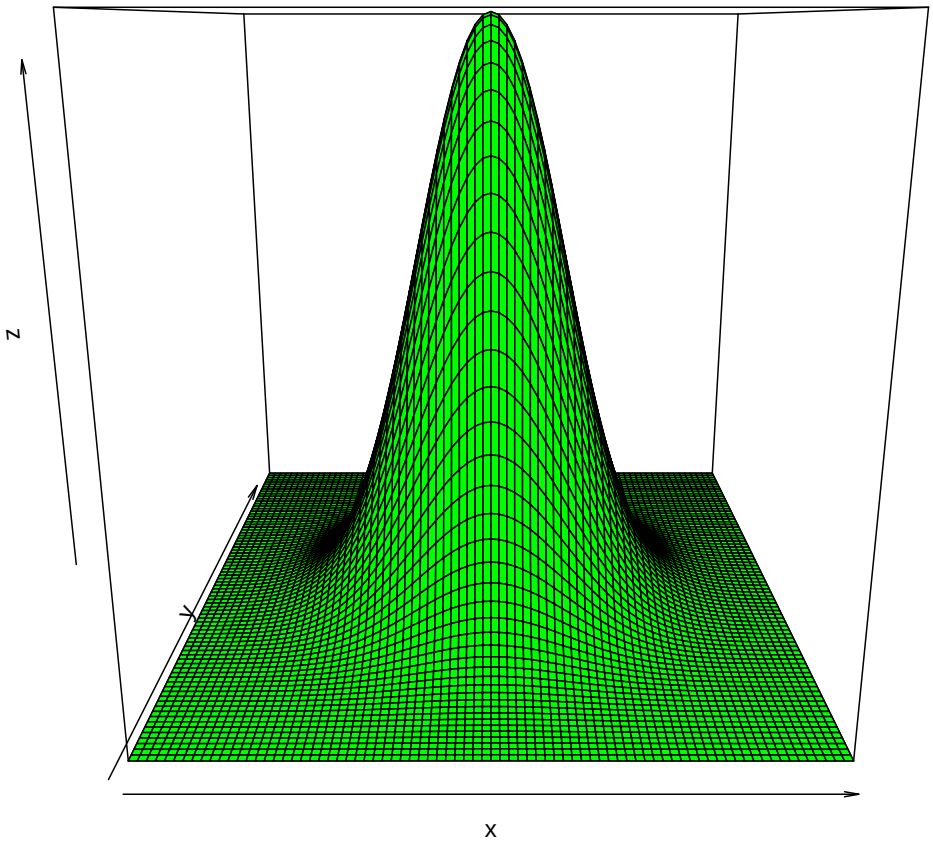
$$f(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} exp\left[-\frac{1}{2}(x_i - \mu)\Sigma^{-1}(x_i - \mu)^{'}\right]$$

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

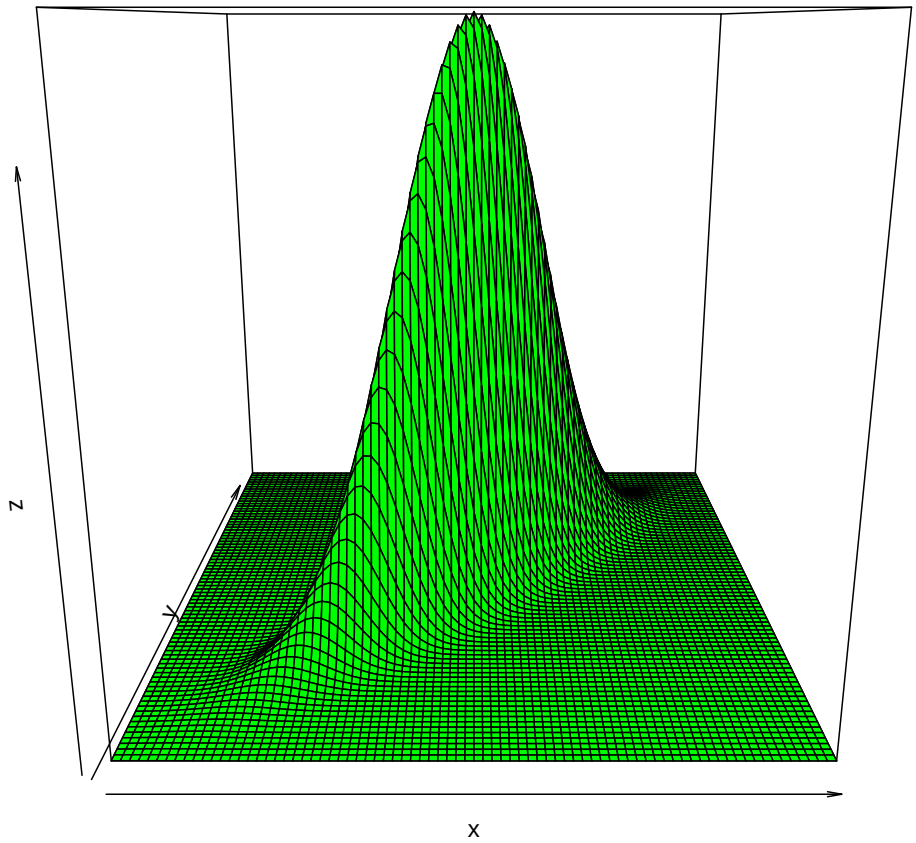
Resultados básicos

- $\mathbb{E}(\mathbf{X}) = \mu$
- $\mathbb{C}ov(\mathbf{X}) = \mathbf{\Sigma}$
- 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: $\Phi(\mathbf{t}) = exp\left[\mathbf{t}^{'}\mu + \frac{1}{2}\mathbf{t}^{'}\mathbf{\Sigma}\mathbf{t}\right]$

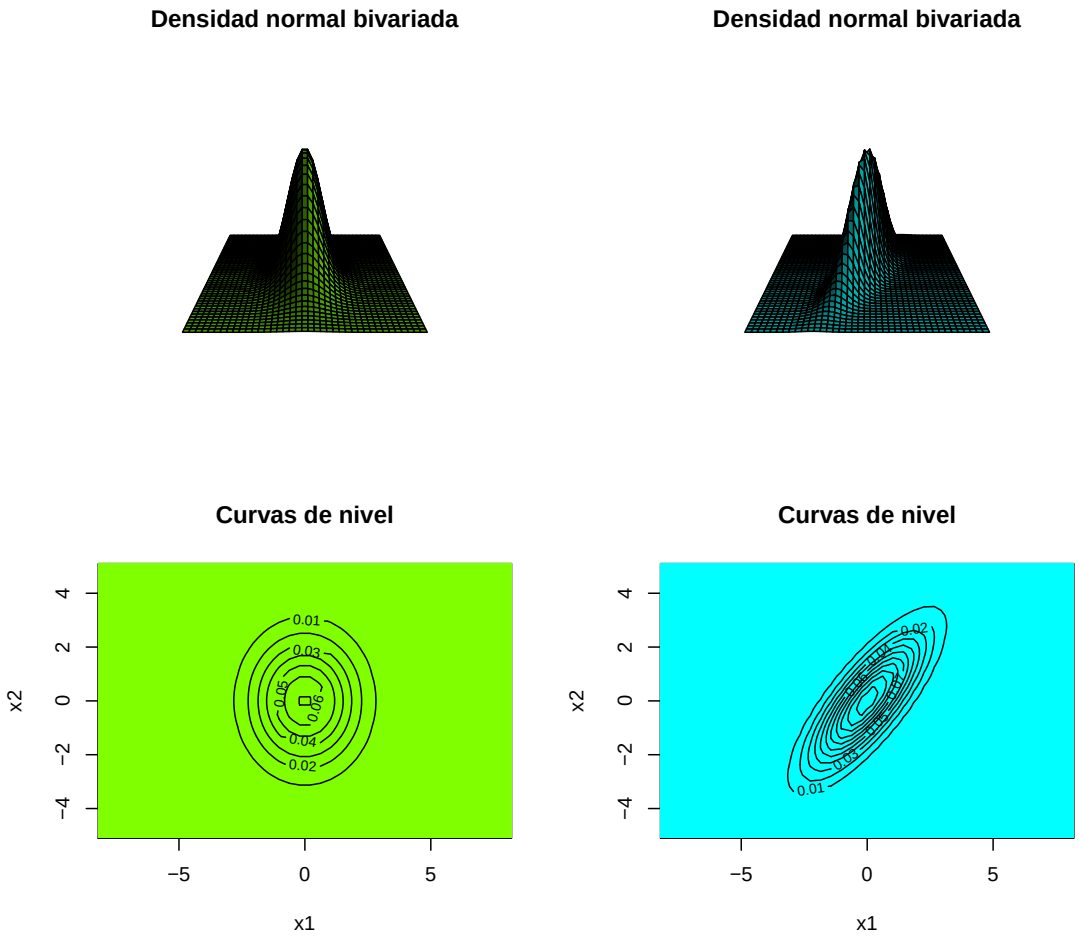
Normal bivariada estándar



Normal bivariada con correlación=0.9



Aspectos de una normal bivariada



Propiedades importantes de la normal multivariada

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

• Sea $\mathbf{Y}$=$\mathbf{C}$ $\mathbf{X}$ con $\mathbf{C}$ una matriz de de cxp con $Rango(\mathbf{C}) = k \leq p$. Entonces,

$\mathbf{Y} \sim N_k(\mathbf{C}\mu, \mathbf{C\Sigma C}^{'})$

• Todos los subconjunto 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

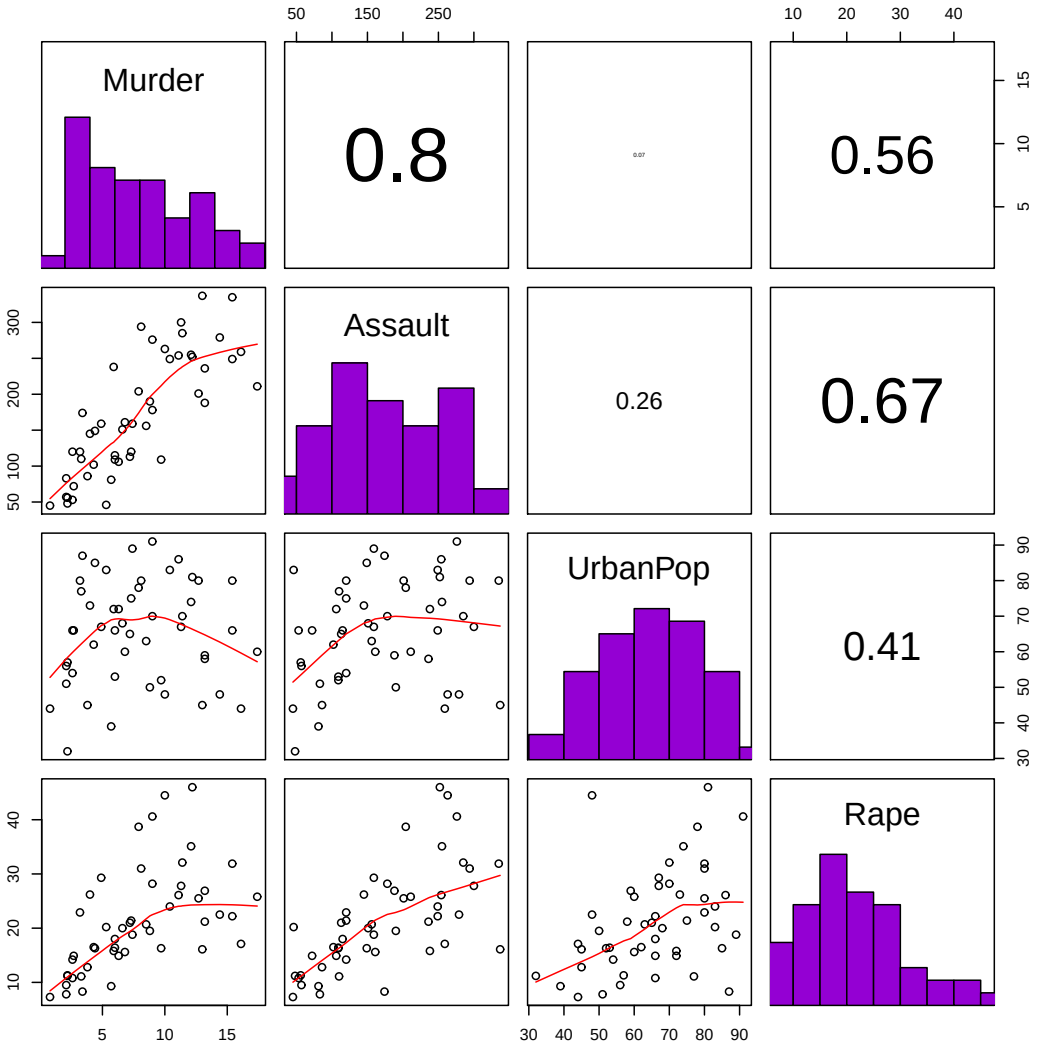
Una manera natural en estadística de mostrar la información contenida en un conjunto de datos, es a través de algunas representaciones gráficas de los mismos. Similar al análisis univariado estándar, se pueden hacer las representaciones gráficas que se considere necesarias, para cada variable. Pero, 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.

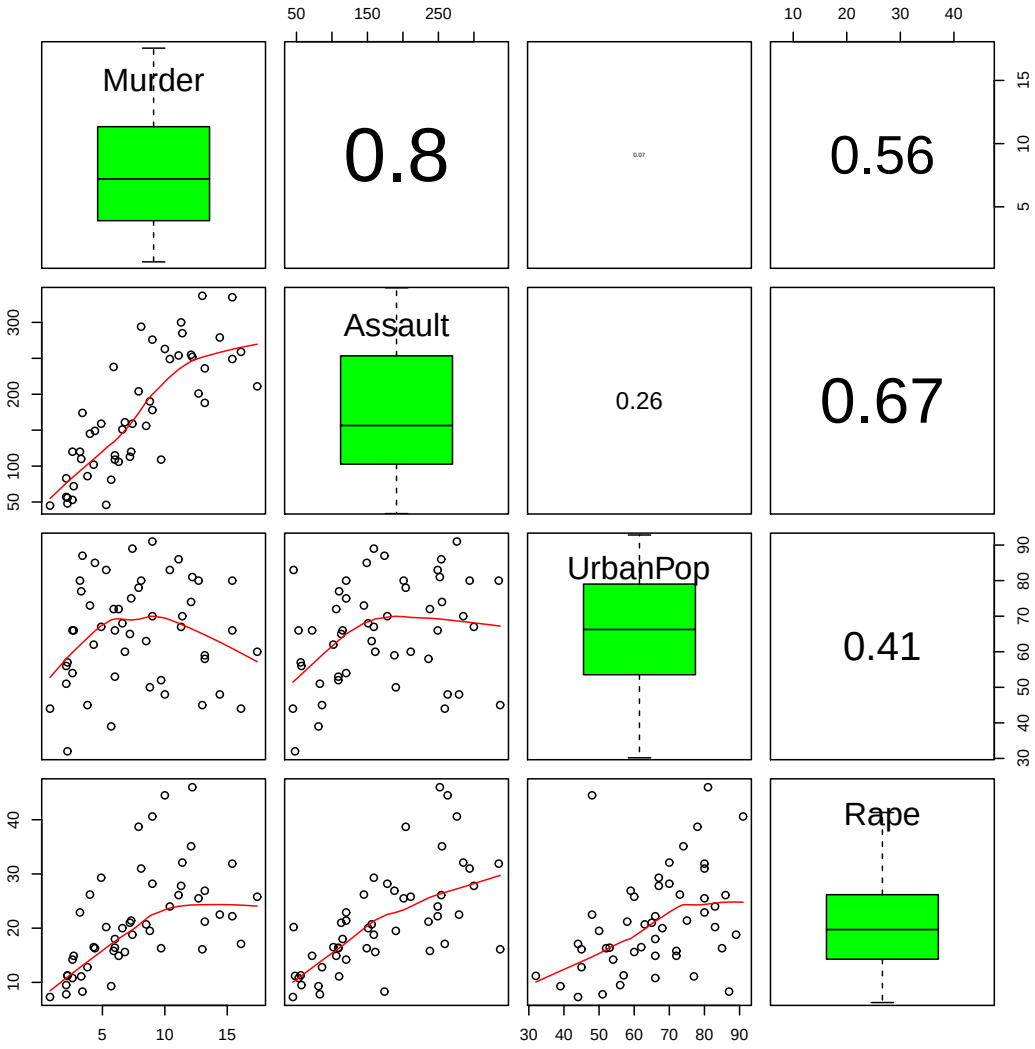
Existen diversas técnicas gráficas para desplegar datos multivariados. La finalidad esencial de éstas es tratar de identificar grupos similiares de sujetos, observaciones atípicas, dispersión de las variables, correlación entre ellas, etc.

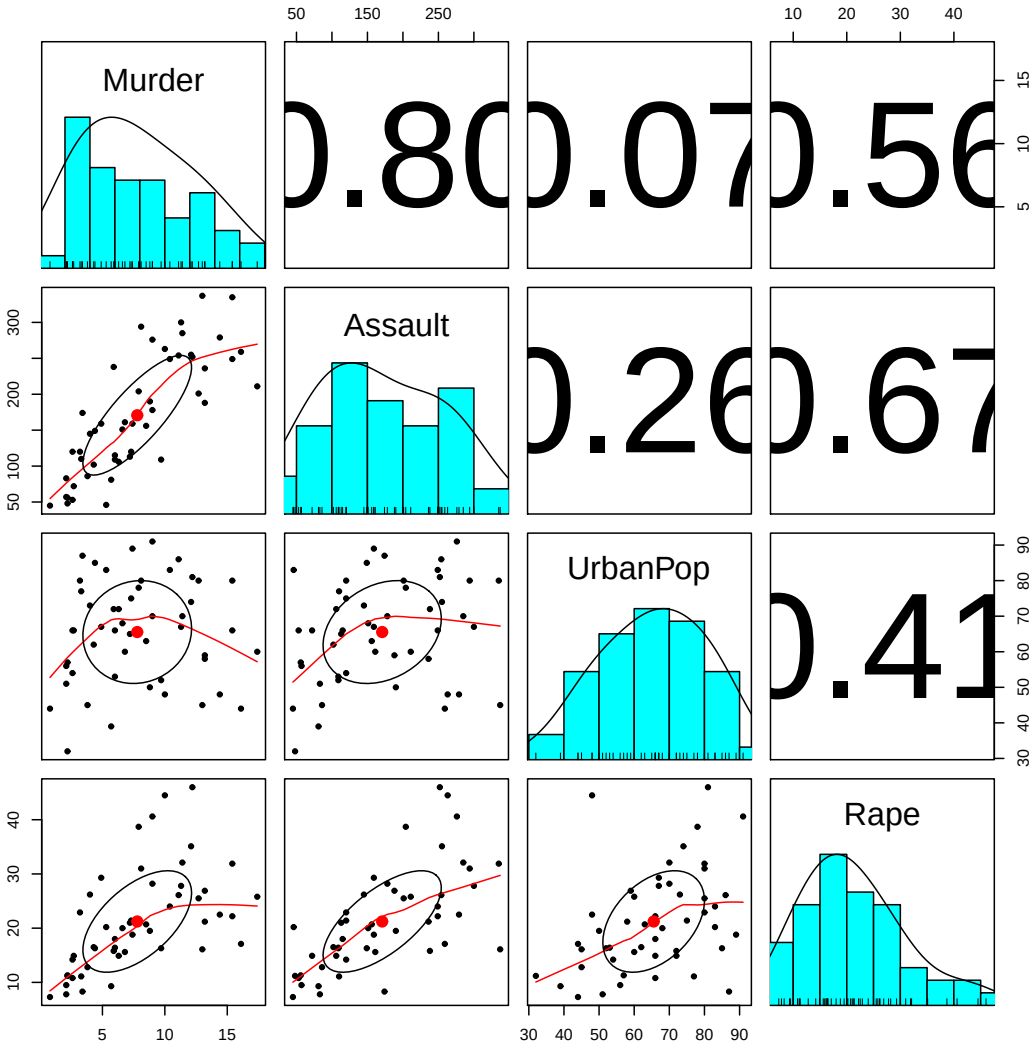
El uso de diagramas y gráficas ahorra tiempo, ya que las características esenciales de grandes volúmenes de datos estadísticos puede apreciarse de un solo vistazo.

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

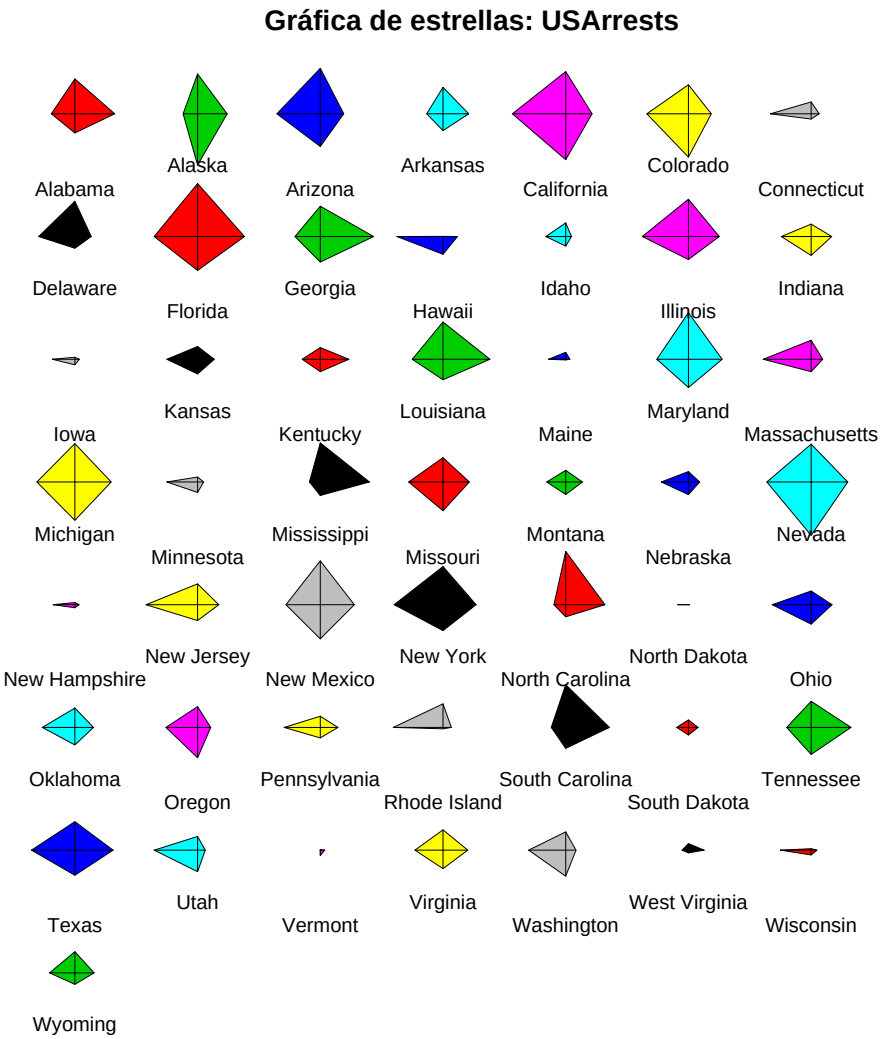

Murder
0.8
0.56
Assault
0.26
0.67
UrbanPop
0.41
Rape



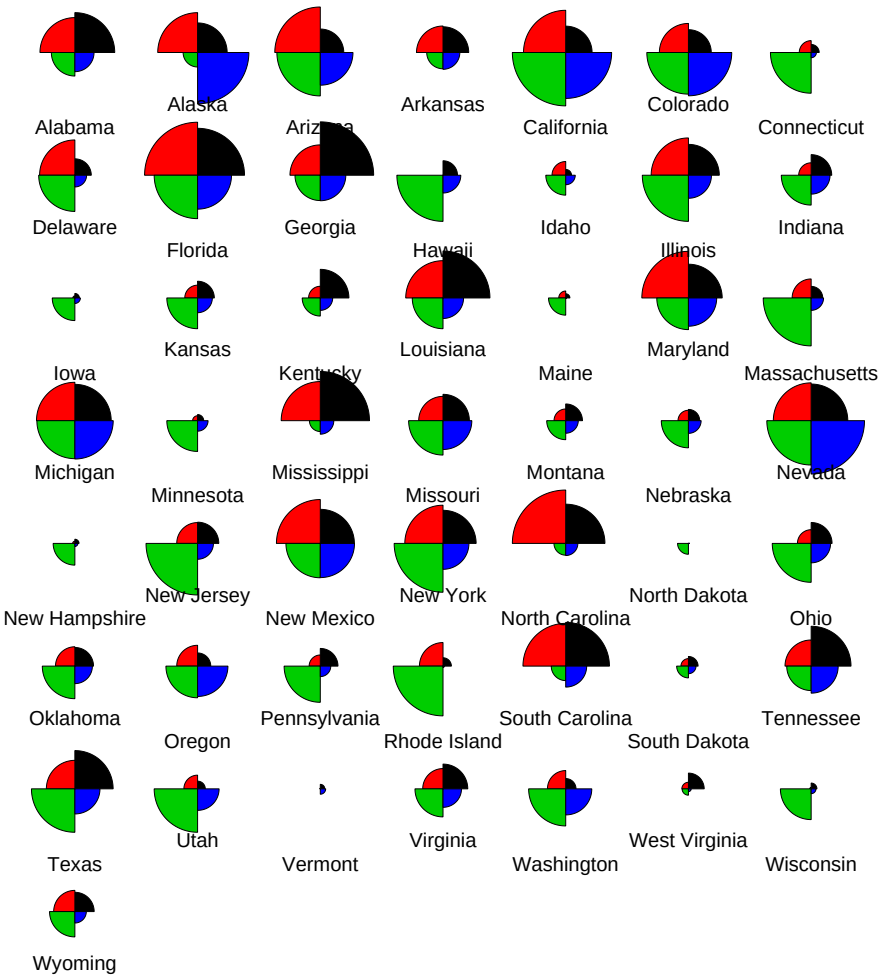
La gráfica anterior presenta características de la forma de la densidad de la variable (histograma y densidad tipo kernel) y de la correlación entre el grupo de variables. Pero no sería útil para descubrir qué estados son similares de acuerdo a este grupo de variables medidas. Para ello, recurriremos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



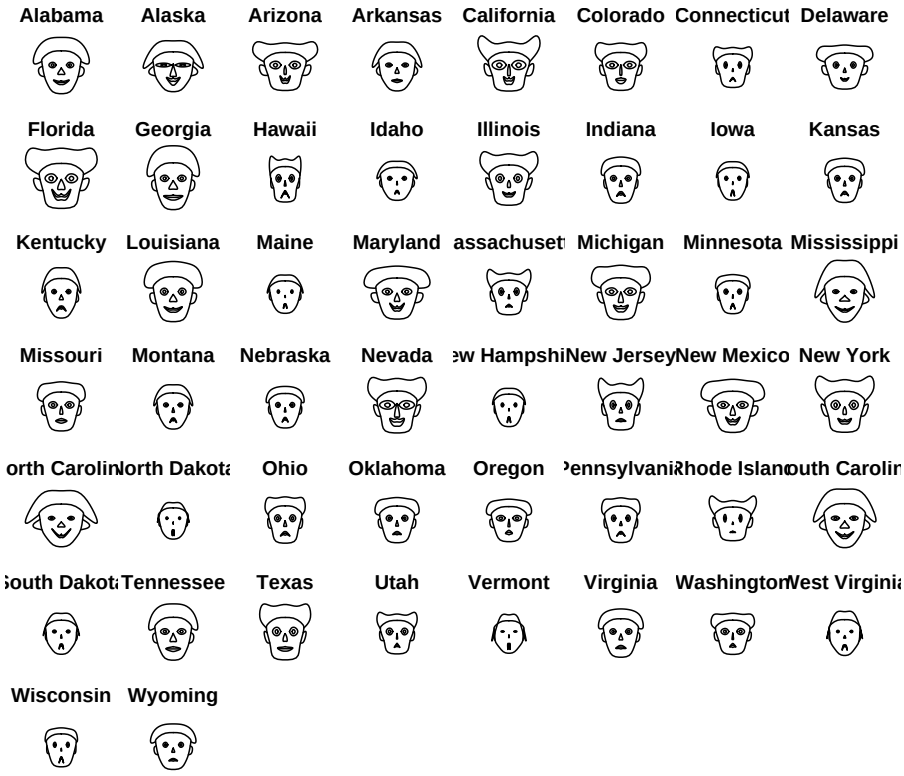
Gráfica de estrellas: USArrests

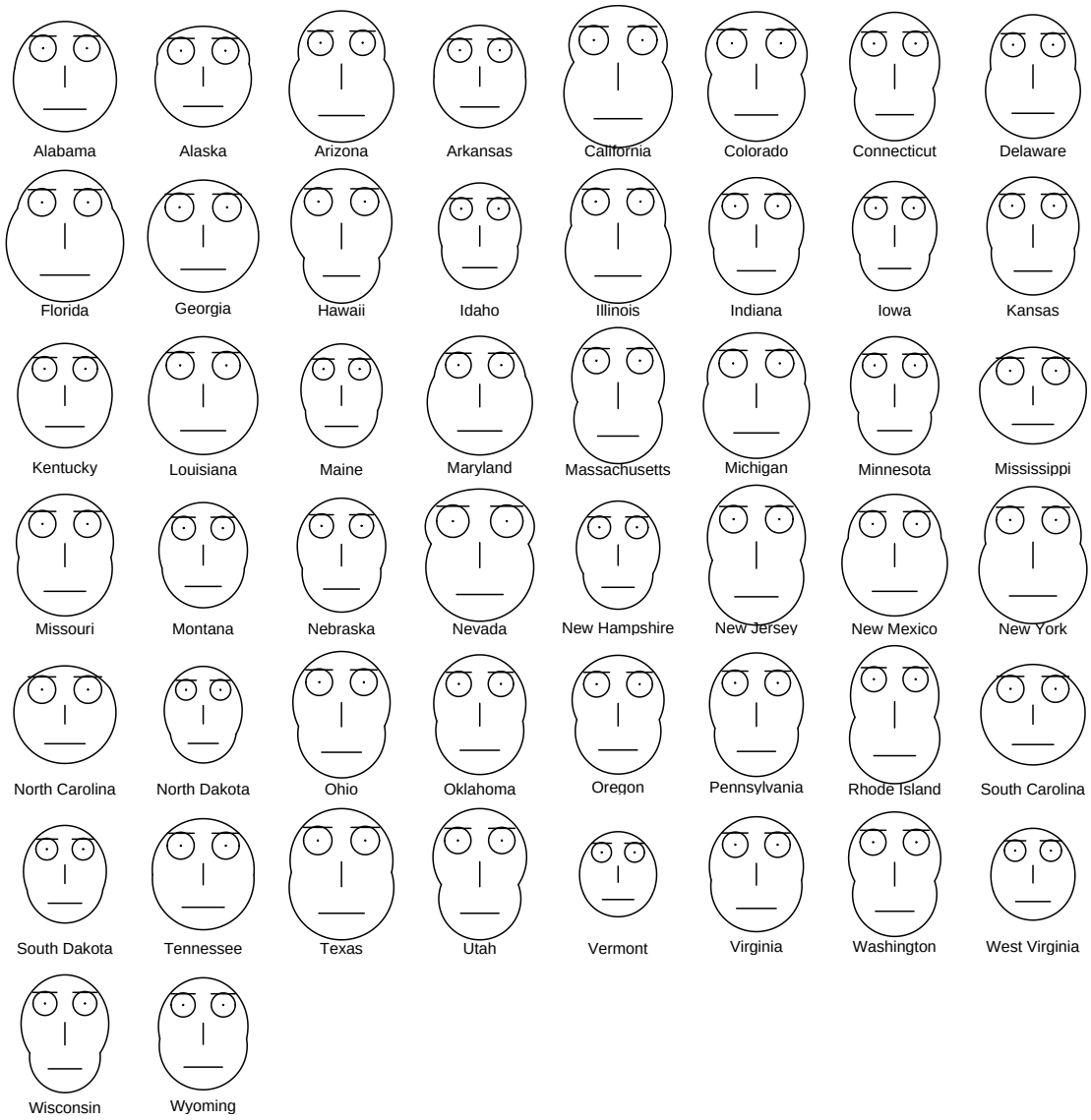


Caritas de Chernoff

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

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.

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

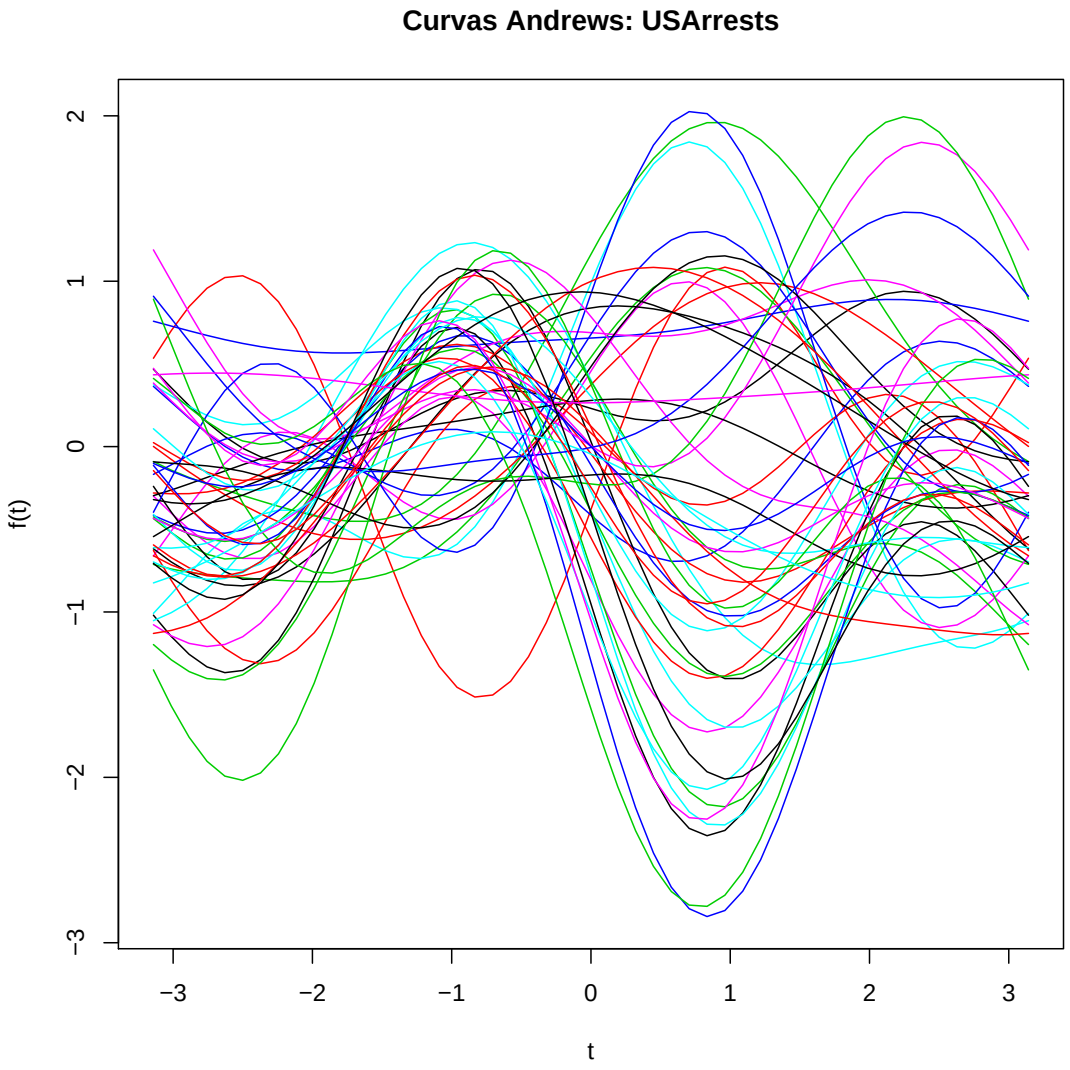
ii) Preserva distancias

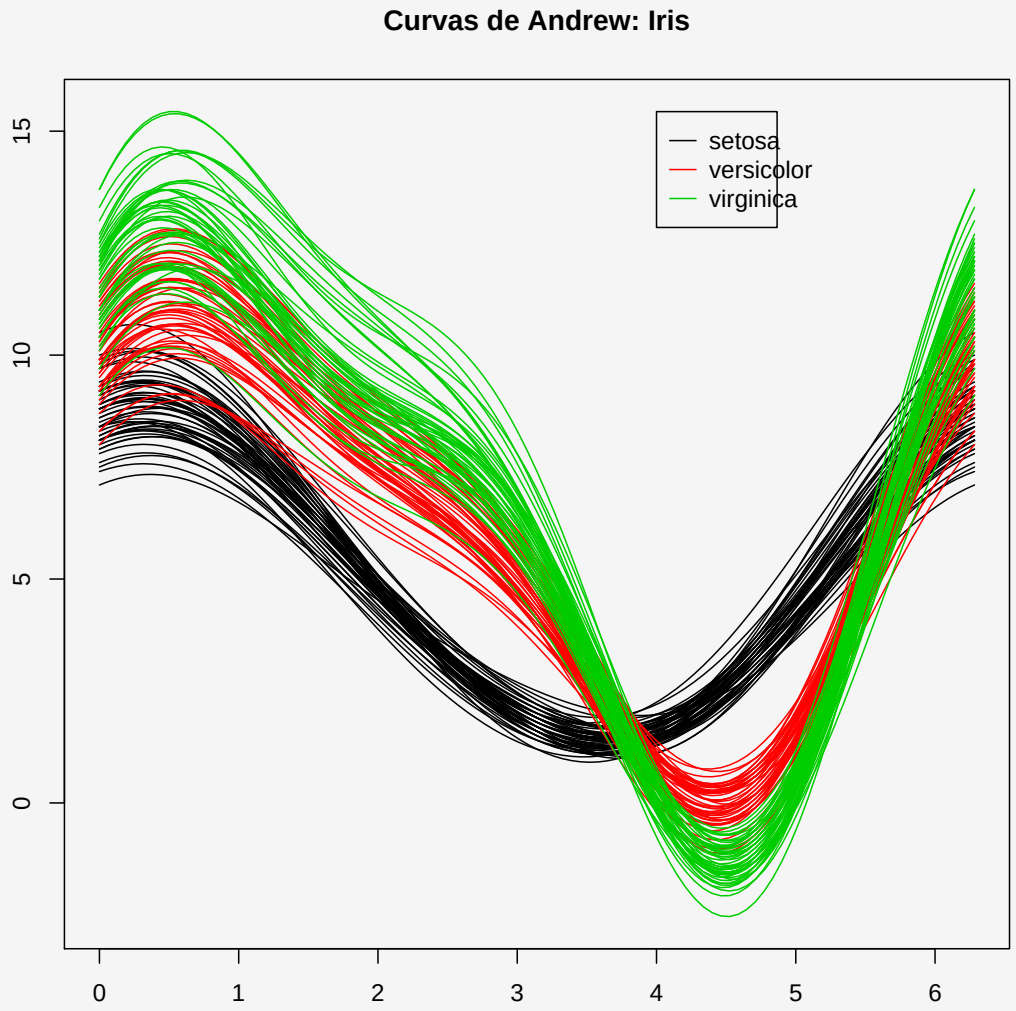
$$\left\| f_{X_i}(t) - f_{X_j}(t) \right\|^2 = \int_{-\pi}^{\pi} \left(f_{X_i}(t) - f_{X_j}(t) \right)^2 dt = \pi \sum_{k=1}^{p} (X_{ik} - X_{jk})^2$$

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 allas.

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 éstas 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 componente 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:

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

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, $\|\mathbf{a}_1\| = \mathbf{a}_1^{'}\mathbf{a}_1 = 1$, de tal manera que:

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

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

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

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

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

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-ésimo* 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 genearal $\mathbb{C}ov(Z_i, Z_j) = 0$ 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 q de ellos

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

• 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

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

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 i en la componente j, 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?

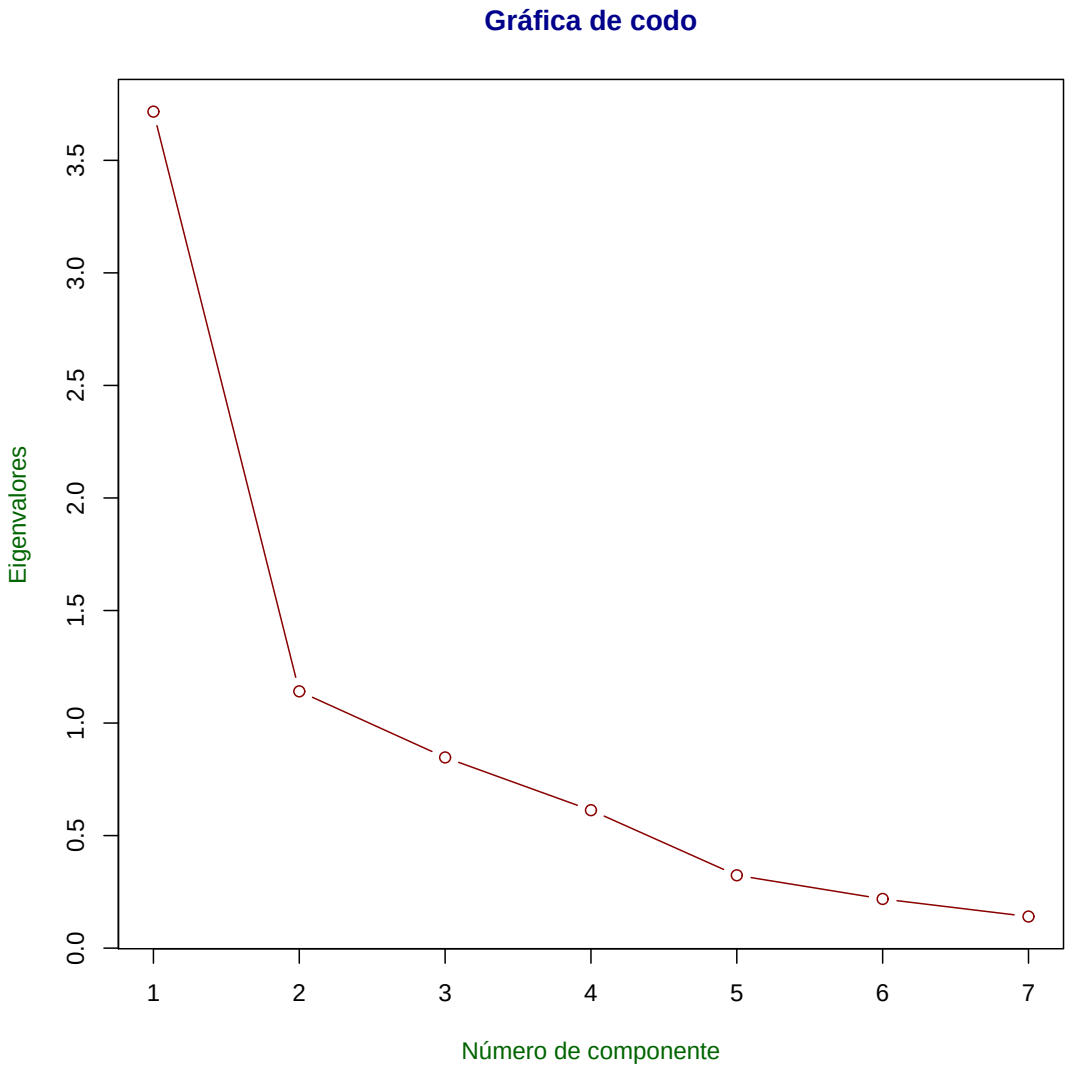
• Porcentaje de varianza explicada requerido (Matriz de varianza-covarianza)

• 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.



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.

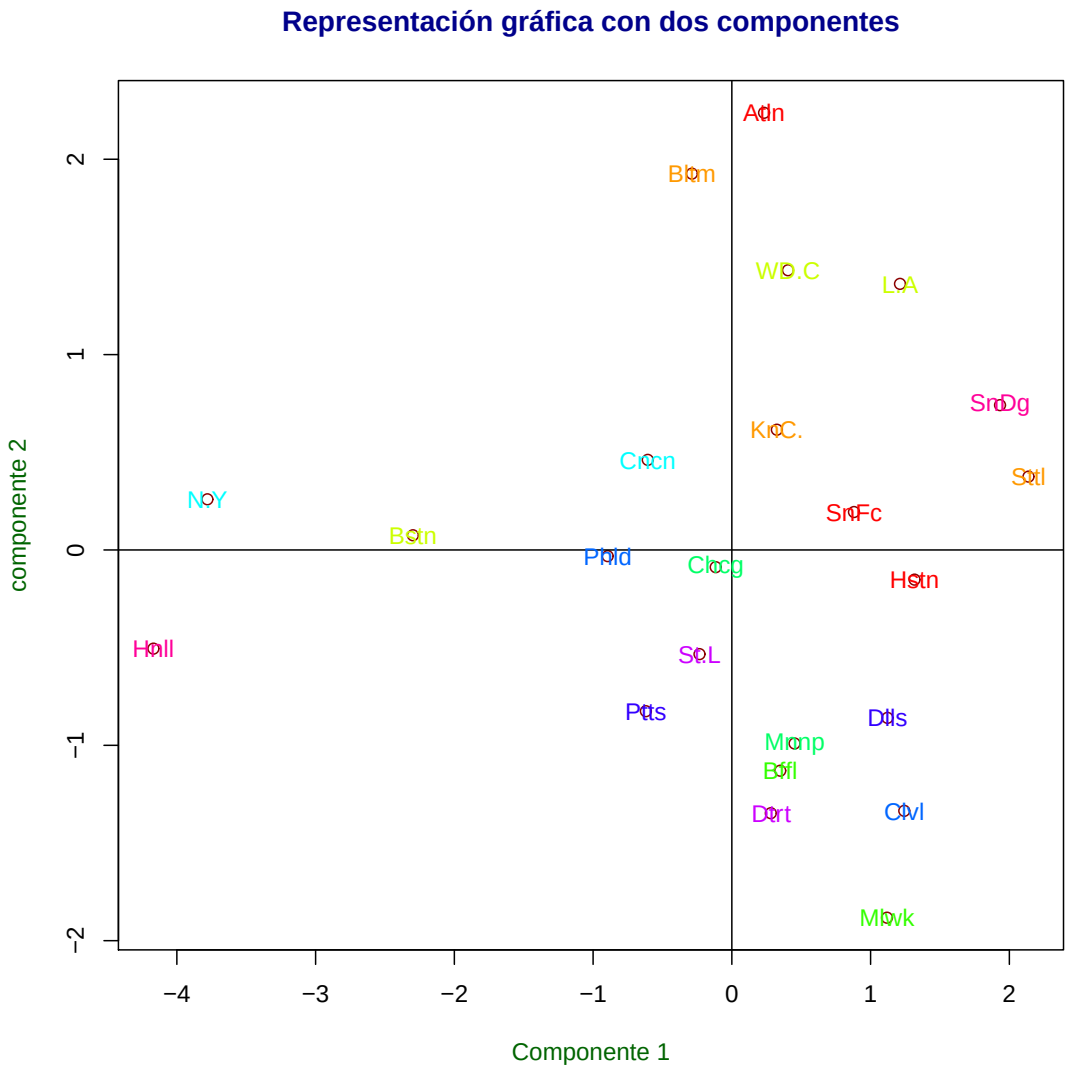
Denotemos por $\mathbf{Z}_i$: Vector de c.p. del individuo $\mathbf{X}_i$ y por $\mathbf{Z}_j$: Vector de c.p. del individuo $\mathbf{X}_j$. Entonces, se trata de mostrar que la distancia entre estas componentes es igual a la distancia entre los vectores originales de los sujetos.

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

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.



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

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

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 jésimo renglón de $\mathbf{H}$ (por tanto, la j-ésima columna de $\mathbf{H}^{'}$). Si *q=2*, los **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

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

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

$\mathbf{U} = (n-1)^{-1/2}\mathbf{Z}\mathbf{L}^{-1/2} = (n-1)\mathbf{X}\mathbf{A}\mathbf{L}^{-1/2}$ $\Big($cuya k-ésima columna es $(n-1)^{-1/2}\ell_k^{-1/2}\mathbf{X}\mathbf{a}_k$, 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

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

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

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

5.- $Corr\,(X_i, X_j) = \dfrac{\mathbf{h}_i'\mathbf{h}_j}{\|\mathbf{h}_i\|\,\|\mathbf{h}_j\|}$, 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. Partiendo 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 $\mathbf{g}$ y $\mathbf{h}$, 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.

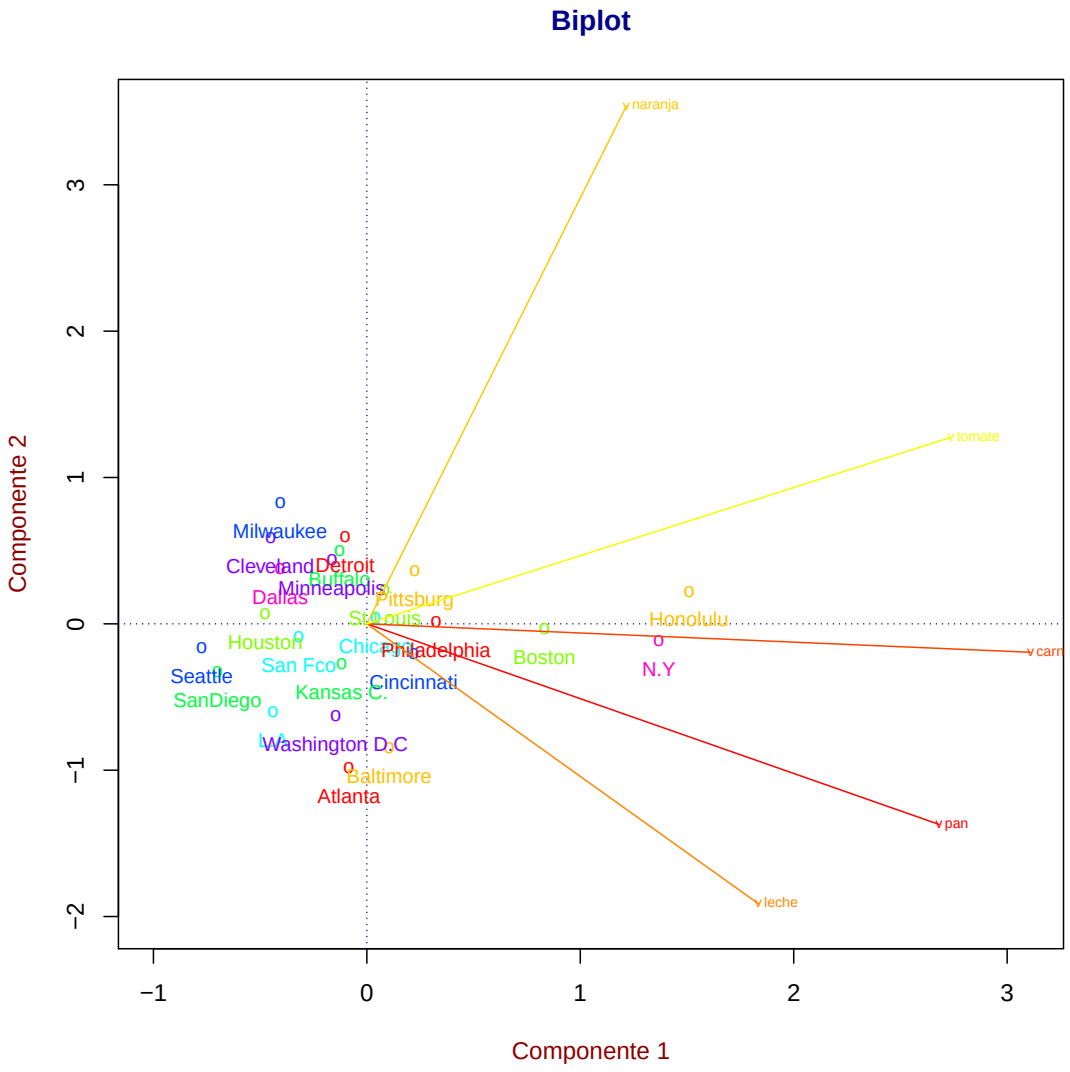
• Individuos semejantes representarán puntos cercanos en la gráfica

• 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 $\mathbf{X}$ 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
naranja
tomate
carn
pan
leche
Milwaukee
Cleveland
Detroit
Buffalo
Minneapolis
Dallas
Pittsburg
Honolulu
Houston
Chic
Philadelphia
Boston
N.Y
San Fco
Seattle
Cincinnati
Kansas C.
SanDiego
Washington D.C
Baltimore
Atlanta
Componente 2
Componente 1
3
2
1
0
−1
−2
−1
0
1
2
3

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 Sperman (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

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

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 diagonales.

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 modelo*, 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 $\Sigma = \Sigma(\underline{\theta})$. 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:

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

con G alguna función específica. Los valores en $\Sigma(\underline{\theta})$ 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

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

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 *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 policorica 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 para la identificación de cargas factoriales significativas, basadas en el tamaño de la muestra	
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.