APRENDE CON ELI

ESTADÍSTICA MULTIVARIANTE

Estadística Multivariante

¿QUÉ ES LA ESTADÍSTICA MULTIVARIANTE?

En casi cualquier situación real, por ejemplo, las características físicas de nosotros como personas: la altura, el peso, la edad, el sexo, el lugar donde hemos nacido, el lugar donde vivimos actualmente, todas esas características, que son variables entre las personas, nos definen como individuos dentro de toda la población.

Pensemos en las propiedades de una imagen: su color, su brillo, su contraste. Son características que la definen y la hacen única.

Estos son dos ejemplos sencillos de lo que serían datos multivariantes, es decir, observaciones sobre las que se miden varias variables.

El análisis multivariante es entonces todo el conjunto de métodos, que veremos en este curso, y que nos permitirán explorar y describir las características de nuestros datos en función del conjunto de variables.

Por ejemplo, podremos resumir el conjunto de variables en unas pocas nuevas variables construidas como transformaciones de las variables originales, pero con la mínima pérdida de información para poder trabajar con ellos de una forma más eficiente. También encontrar grupos entre los datos, si existen. Clasificar nuevas observaciones en grupos que ya han sido definidos. Estas técnicas para describir, resumir y condensar información, clasificar datos, relacionar variables, se conocen a veces como técnicas de exploración de datos.

Los métodos para datos multivariantes se han popularizado en los últimos años en Ingeniería y Ciencias de la computación con el nombre de Minería de datos, nombre que indica la capacidad de estas técnicas para extraer

información a partir de los datos. Sin embargo, estas herramientas no permiten directamente obtener conclusiones generales respecto al proceso o sistema que genera los datos. Para ello necesitamos otro tipo de métodos para crear conocimiento respecto al problema mediante un modelo estadístico y la construcción de un modelo estadístico requiere del concepto de probabilidad. Las herramientas básicas para la construcción de modelos requieren estimar los parámetros del modelo a partir de los datos disponibles, contrastar hipótesis respecto a su estructura para lo cual es necesario conocer los fundamentos de la inferencia multivariante.

Así que los métodos van a estar agrupados en estos dos tipos de análisis: el análisis univariante, el más simple, que es cuando tenemos n observaciones (el tamaño muestral) pero de una sola variable. Por ejemplo, si nuestra variable es la altura de 100 estudiantes, la altura sería la variable aleatoria que estamos midiendo sobre esos 100 estudiantes, y 100 es el tamaño de nuestra muestra. Otra variable aleatoria podría ser el nivel de llenado de 50 botellas o por ejemplo la ciudad de nacimiento de los 1000 participantes de una encuesta.

Pero cuando hablamos de datos multivariantes vamos a tener n observaciones, pero más de una variable aleatoria a la vez, es decir, que sobre mis observaciones vamos a estar midiendo más de una sola variable, más de una característica. Y esas variables aleatorias van a ser características que se pueden medir simultáneamente sobre los n elementos.

Por ejemplo, las características que se miden sobre las personas cuando se les hace una encuesta, que se les preguntan varias cosas a la misma persona: altura, peso, edad, lugar de residencia, etc. Todas esas son características que van a ser medidas sobre cada uno de los elementos de mi muestra.

Entonces el aspecto de nuestros datos, en el caso multivariante, vamos a tener n filas en nuestra matriz de datos y vamos a tener p columnas, que van a ser las p variables.

NOTACIÓN

La matriz de datos se puede denotar de la siguiente forma, dependiendo de las filas:

$$\mathbf{x} = [\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_n}]$$

Donde $\mathbf{x_i} \in \mathbb{R}^p$, $\forall \mathrm{i} = 1, \dots, \mathrm{n}$

O de esta manera alternativa, dependiendo de las columnas:

$$\mathbf{x} = [\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_p}]$$

Donde $\mathbf{x_j} \in \mathbb{R}^n$, $\forall \mathrm{i} = 1, \dots, \mathrm{p}$

OBJETIVOS DEL ANÁLISIS MULTIVARIANTE

Los objetivos del análisis multivariante son comprender la estructura de los datos y resumirla de manera sencilla, entender las relaciones entre los datos o las variables, y tomar decisiones y hacer inferencias basadas en los datos.

TIPOS DE VARIABLES

Vamos a ver que existen dos tipos de variables que se pueden mezclar en nuestros datos cuando tenemos varias variables. Podemos encontrarnos **variables cuantitativas**, que son datos numéricos, por ejemplo, cuando nos preguntan la edad, la altura, los ingresos que tenemos. O **variables cualitativas,** que son atributos o categorías, como por ejemplo si nos preguntan el sexo o el color de ojos o la ciudad donde hemos nacido, todas esas variables toman valores que son categorías o atributos, no son números.

Ahora respecto a las variables cuantitativas, pueden ser continuas o discretas.

Las continuas van a ser valores reales dentro de un intervalo. Es decir, es muy raro que se repitan, por ejemplo, en el peso de las personas puede haber varias coincidencias, pero no muchas coincidencias y siempre van a estar definidas dentro de un intervalo de manera continua.

Sin embargo, las variables discretas son valores numerables o contables. Por ejemplo, el número de hermanos. Se puede tener cero hermanos, un hermano, dos hermanos, etc. Son valores que llegan hasta un límite y además son contables, numerables, es decir, nadie va a tener 1,5 hermanos. Esa es la diferencia entre continua y discreta.

Por tanto, vamos a asumir que estamos observando p variables aleatorias, cada una de ellas es una variable univariante, y se miden en un conjunto de elementos que son nuestra muestra. Luego, ese conjunto de variables aleatorias forma una variable aleatoria multivariante.

IRIS

Un ejemplo de datos reales multivariantes es el dataset de IRIS. Está formado por 150 flores y para cada flor (que cada una es un elemento muestral) se van a medir cuatro características: el largo y el ancho del sépalo y el largo y el ancho del pétalo.

VISUALIZACIÓN

¿Por qué es importante visualizar? Porque es una manera mucho más sencilla de obtener información. ¿Sobre qué vamos a obtener información? Sobre la distribución, sobre si hay simetría o asimetría, si hay multimodalidad, si hay presencia de atípicos o outliers, sobre si nuestros datos pueden agruparse, etc. Entonces la información gráfica va a ser un complemento de la información numérica que podemos obtener en un análisis exploratorio.

Ahora bien, el sistema de precepción humano, el que tenemos nosotros, tiene estas características importantes: somos capaces de entender e interpretar gráficos en dos o tres dimensiones, pero más de tres dimensiones no pueden visualizarse de manera tan sencilla.

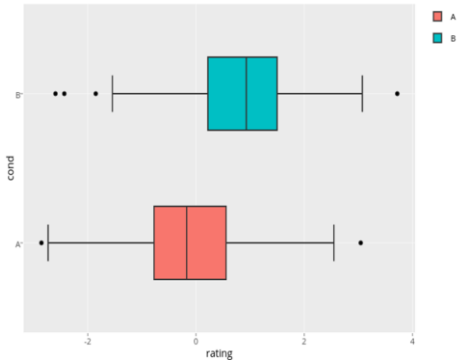
Por lo tanto, los gráficos que vamos a ver van a estar dibujados en dos o en tres dimensiones, pero aprenderemos también cómo representar datos de más de tres dimensiones en este tipo de gráficos.

¿QUÉ TIPO DE VARIABLES VAMOS A VISUALIZAR?

En realidad, vamos a graficar las variables de tipo cuantitativas, porque las cualitativas nos van a dar información adicional en los gráficos, no vamos a graficarlas como tal, sino que vamos a utilizar esa información, por ejemplo, para agrupar los datos.

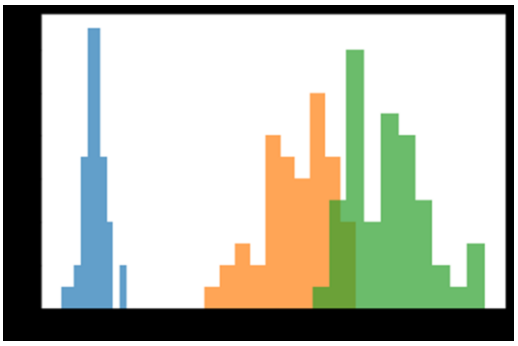
EJEMPLOS DE GRÁFICOS

Ejemplos de gráficos que podemos hacer son los Boxplots o diagramas de caja y bigotes.

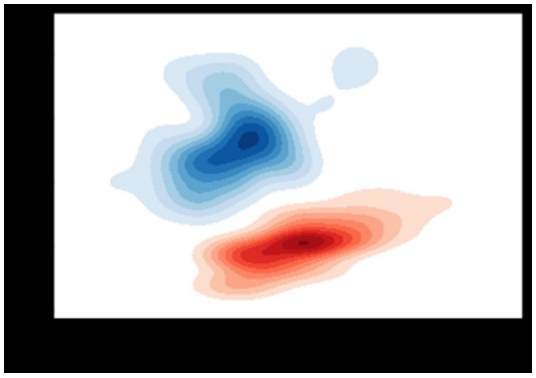


Además, vamos a ver el Bagplot que es una extensión del Boxplot.

También vamos a ver los Histogramas, que para variables múltiples que estén agrupadas en grupos podemos dibujar varios histogramas en dependencia del grupo y poner cada uno en diferente color para observar en un mismo gráfico los histogramas de varios grupos de datos.

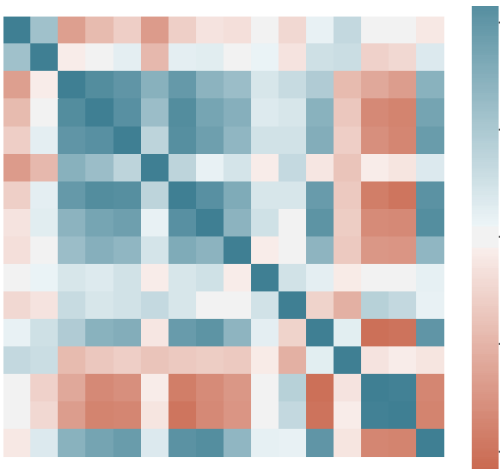


Vamos a ver los gráficos de densidades: la densidad Kernel.

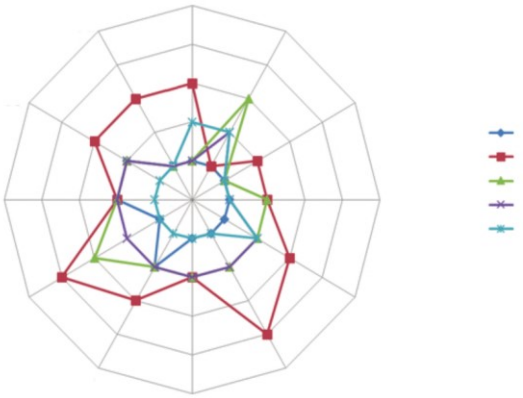


Los scatterplots, que son diagramas de dispersión, que me relacionan una variable con otra. Los scatterplot multivariantes o múltiples en los que vamos a ver estos gráficos de relaciones entre pares de variables, pero en un mismo gráfico.

Y también vamos a ver la matriz de correlaciones visual. La matriz de correlaciones es una matriz numérica, pero se puede estudiar de manera visual.



Y por último vamos a estudiar unos gráficos que se llaman de coordenadas paralelas y los gráficos radiales como éste que se ve en la imagen.



ANÁLISIS EXPLORATORIO – GRÁFICOS

BOXPLOT

Es un gráfico que nos da información sobre:

- Localización
- Desviación
- Asimetría
- Atípicos (outliers)

Para construir un boxplot es necesario calcular algunos estadísticos.

Para una variable aleatoria univariante

$$\mathbf{x} = [x_1, x_2, \dots, x_n],$$

los estadísticos de orden denotados como $x_{(1)}, \dots, x_{(n)}$ son las observaciones ordenadas en orden creciente. De aquí vamos a usar el mínimo y el máximo:

$$x_{(1)} = \min\{x_1, x_2, \dots, x_n\}$$

$$x_{(n)} = \max\{x_1, x_2, \dots, x_n\}$$

La mediana se puede definir en dependencia de los estadísticos de orden:

$$Mediana = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & , si\ n\ es\ impar \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & , si\ n\ es\ par \end{cases}$$

Los cuartiles dividen al conjunto de datos ordenados en 4 partes iguales.

El segundo cuartil Q_2 coincide con la mediana.

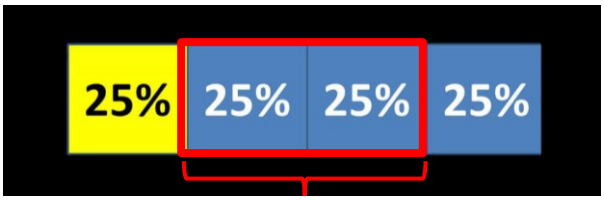
El primer y tercer cuartil se pueden calcular de esta forma:

$$Q_1 = x_{\left[\frac{n+1}{4}\right]}$$

$$Q_3 = x_{\left[\frac{3(n+1)}{4}\right]}$$

$[\cdot]$ denota la parte entera.

El rango intercuartílico se define como el bloque que recoge el 50% más central.



$$RI = Q_3 - Q_1$$

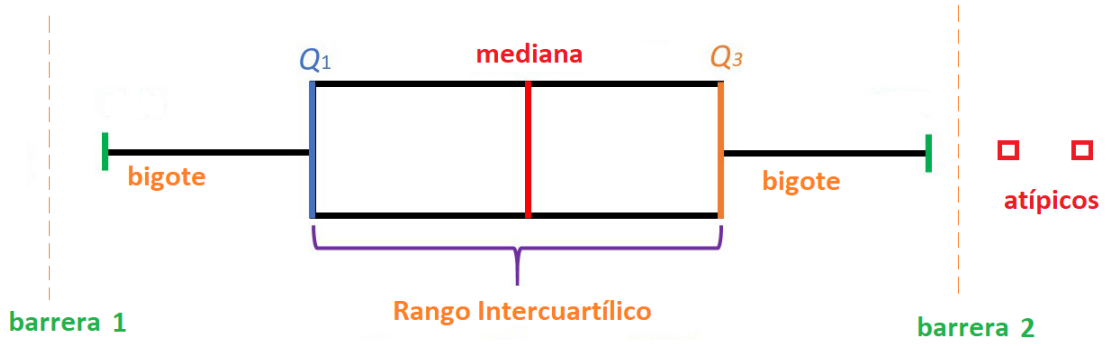
Las barreras se definen a partir de los cuartiles exteriores sumando o restando la misma cantidad $1.5\ RI$.

$$B_1 = Q_1 - 1.5\ RI$$

$$B_2 = Q_3 + 1.5\ RI$$

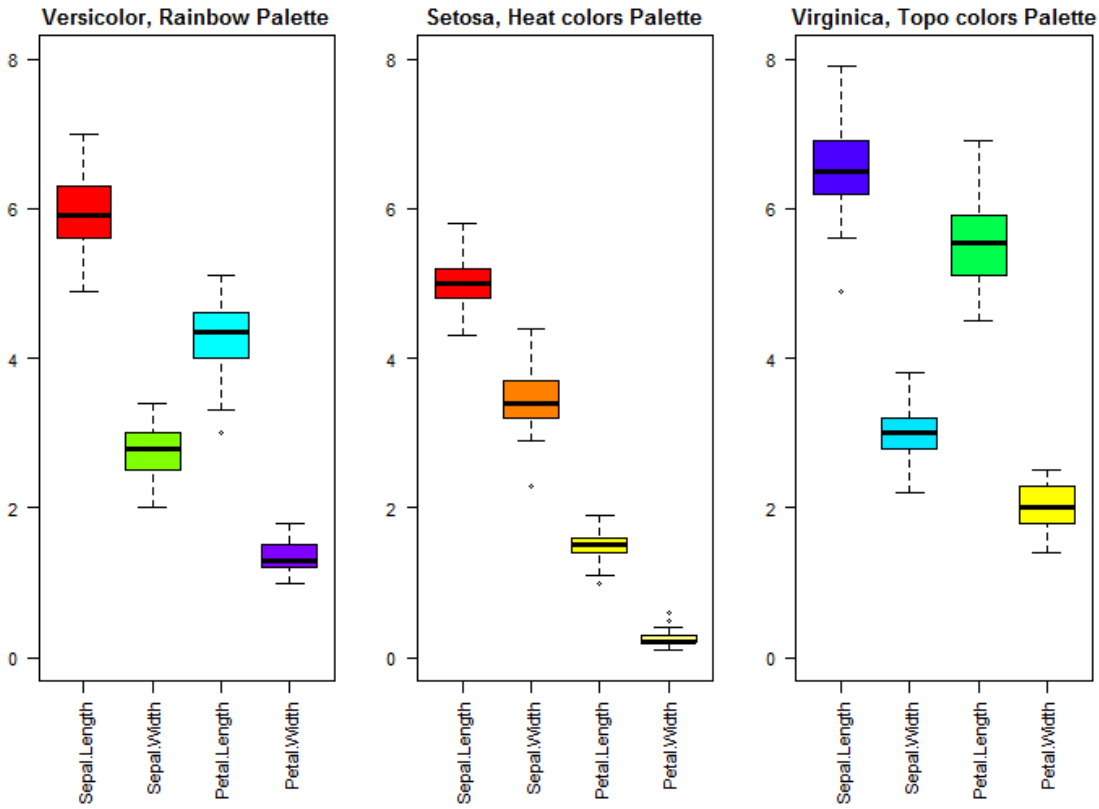
¿CÓMO CONSTRUIMOS EL BOXPLOT?

1. La caja estará formada por Q_1 y Q_3
2. Se dibuja una recta en el medio que es la mediana (coincide con Q_2)
3. Se detectan los atípicos con las barreras y se denotan con un símbolo especial, por ejemplo: □
4. Si no hay atípicos, los "bigotes" son el mínimo y el máximo.
5. Si hay atípicos, los "bigotes" son los puntos más remotos que no sean atípicos.



En el caso multivariante:

- Podemos hacer en un mismo gráfico, boxplots para todas las variables.
- Y dividirlos por grupos en dependencia de valores de variables categóricas.



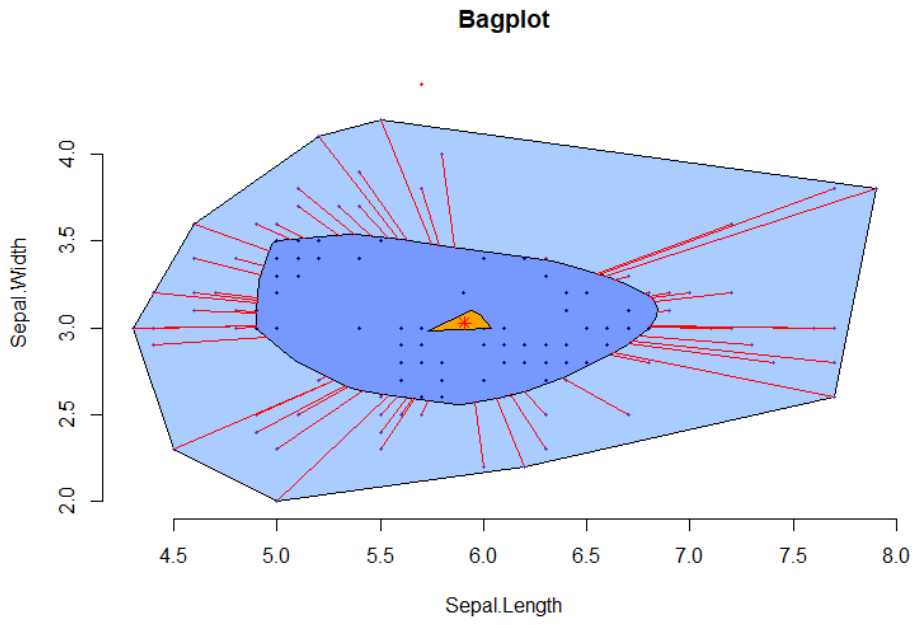
BAGPLOT

El bagplot es una extensión bivariante del boxplot:

- El boxplot nos informa sobre una variable aleatoria univariante.
- El bagplot nos informa sobre dos variables aleatorias, es decir, una variable aleatoria bivariante.

La estructura del bagplot es análoga a la del boxplot. Pero va a haber una "bolsa" (bag) en vez de una "caja".

La "bolsa" o "bag" es una región (un polígono) que contiene a lo sumo el 50% de datos más centrales. El polígono externo, llamado "cerca", no se dibuja como parte del diagrama, sino que se usa para detectar atípicos. Se forma al inflar la bolsa por un cierto factor (generalmente 3). Las observaciones fuera de la "cerca", se marcan como atípicos. Las observaciones que no sean marcadas como atípicos son las que formarán parte de la envolvente convexa que sería lo análogo a los "bigotes" en el boxplot. El centro se calcula con una medida de profundidad que se llama "Tukey Depth". Las profundidades son estimadores de distancia con el sentido opuesto, es decir: un punto es más profundo con respecto al centro, si tiene menos distancia hacia ese centro.

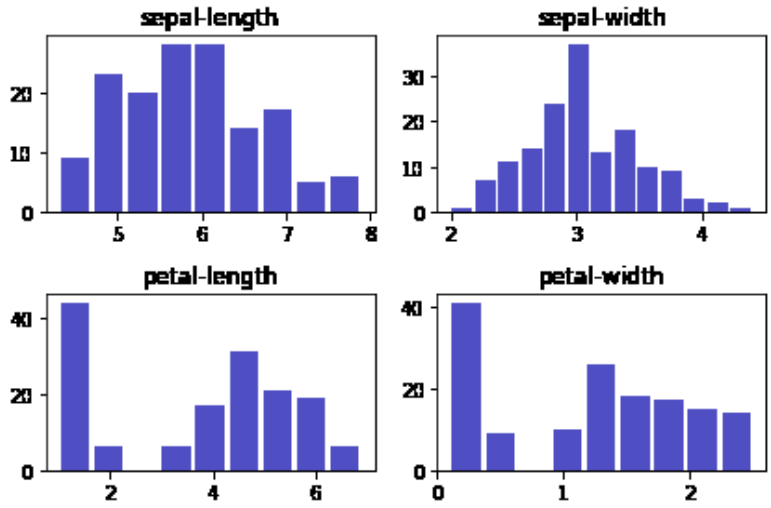


HISTOGRAMA

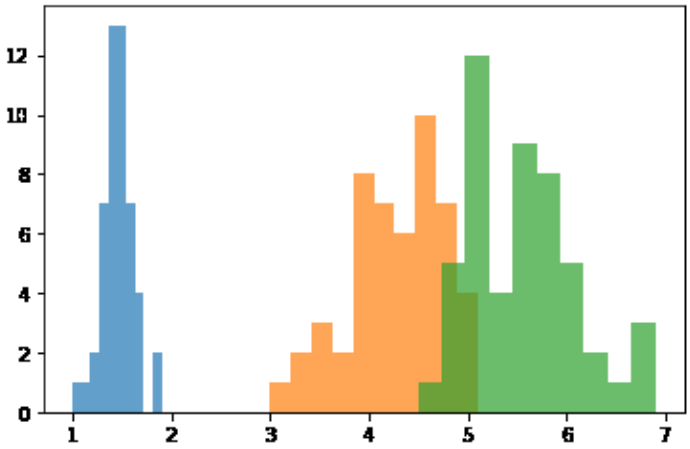
Es un gráfico que nos da información sobre:

- La distribución de una variable aleatoria
- Posible multimodalidad
- Asimetría
- Localización
- Dispersión
- Atípicos

El histograma es una representación gráfica de frecuencias. En el caso multivariante, podemos comparar los histogramas de todas las variables que tenemos.



Podemos comparar los histogramas de una variable dividida en los subgrupos que tenemos.

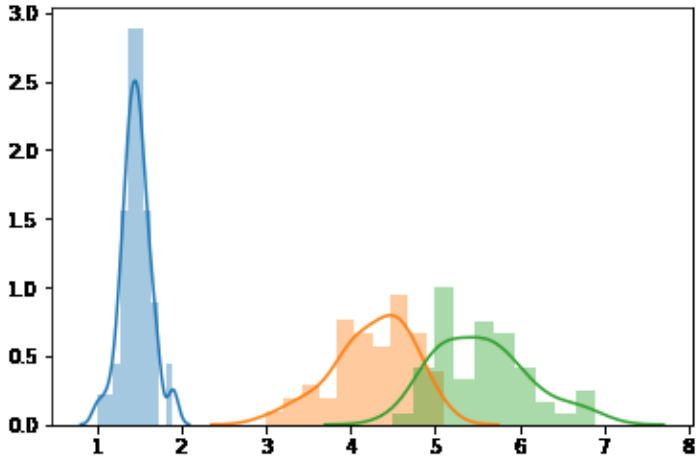


DENSIDAD KERNEL

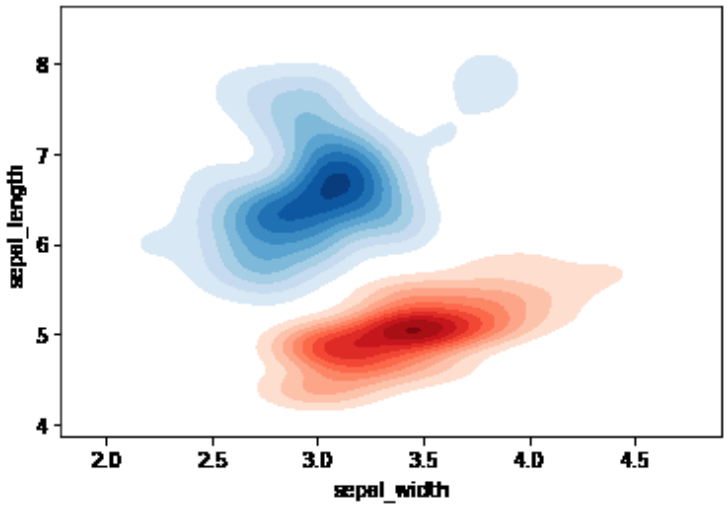
Es un gráfico que aporta:

- La misma información que un histograma.
- Además de estimar la función de densidad de la variable aleatoria y dibujarla sobre el histograma.

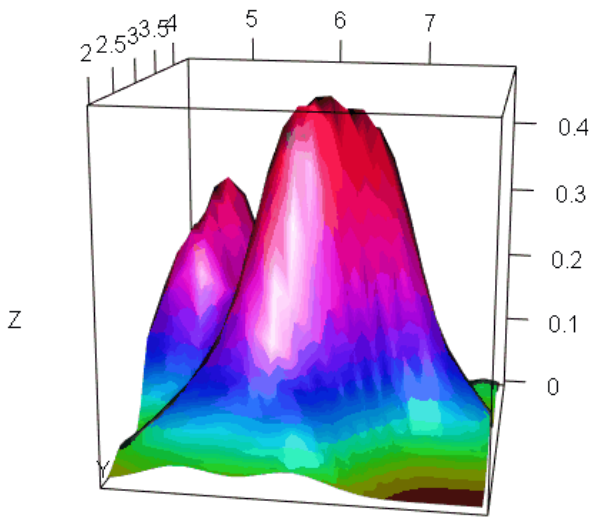
La densidad kernel está basada en una función que se llama kernel. Por ejemplo: Kernel Gaussiano, Uniforme, Triangular, Epanechnikov, etc. En el caso multivariante podemos hacer lo mismo que con los histogramas.



También podemos estimar la densidad kernel bivariada entre dos variables aleatorias y verlo en un gráfico bidimensional.



Podemos estimar la densidad kernel bivariada entre dos variables aleatorias y verlo en un gráfico bidimensional o tridimensional.



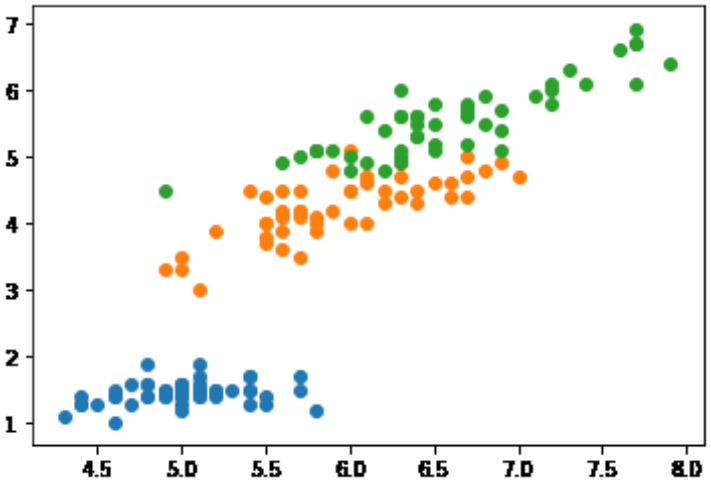
bivkde

SCATTERPLOT

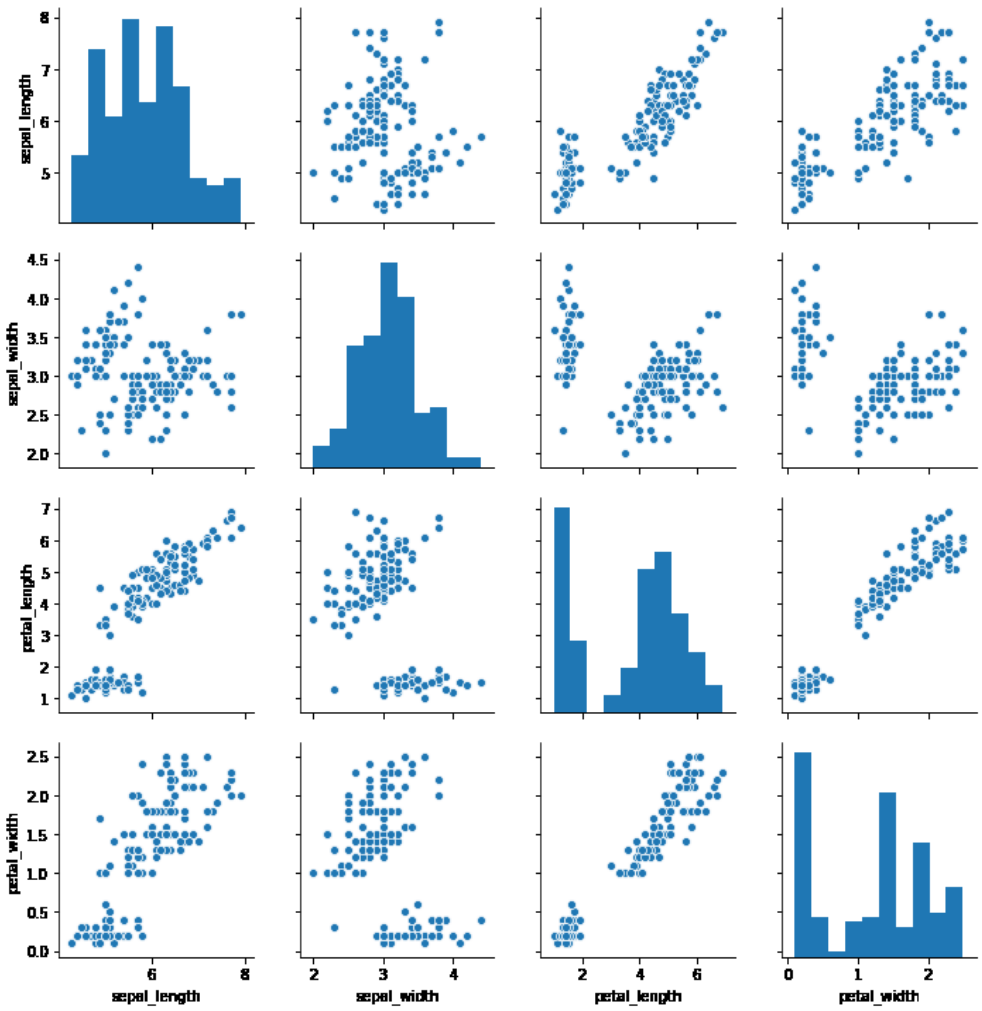
Es un gráfico que aporta:

- Visualizar la distribución bivariada (entre dos variables aleatorias)
- También se conoce como Diagrama de Dispersión.
- Nos ayuda a entender las relaciones entre pares de variables.
- Si se añade una 3ra variable puede hacerse un scatterplot 3-dimensional.
- El "scatterplot matrix" es un gráfico con múltiples 2D-scatterplots.

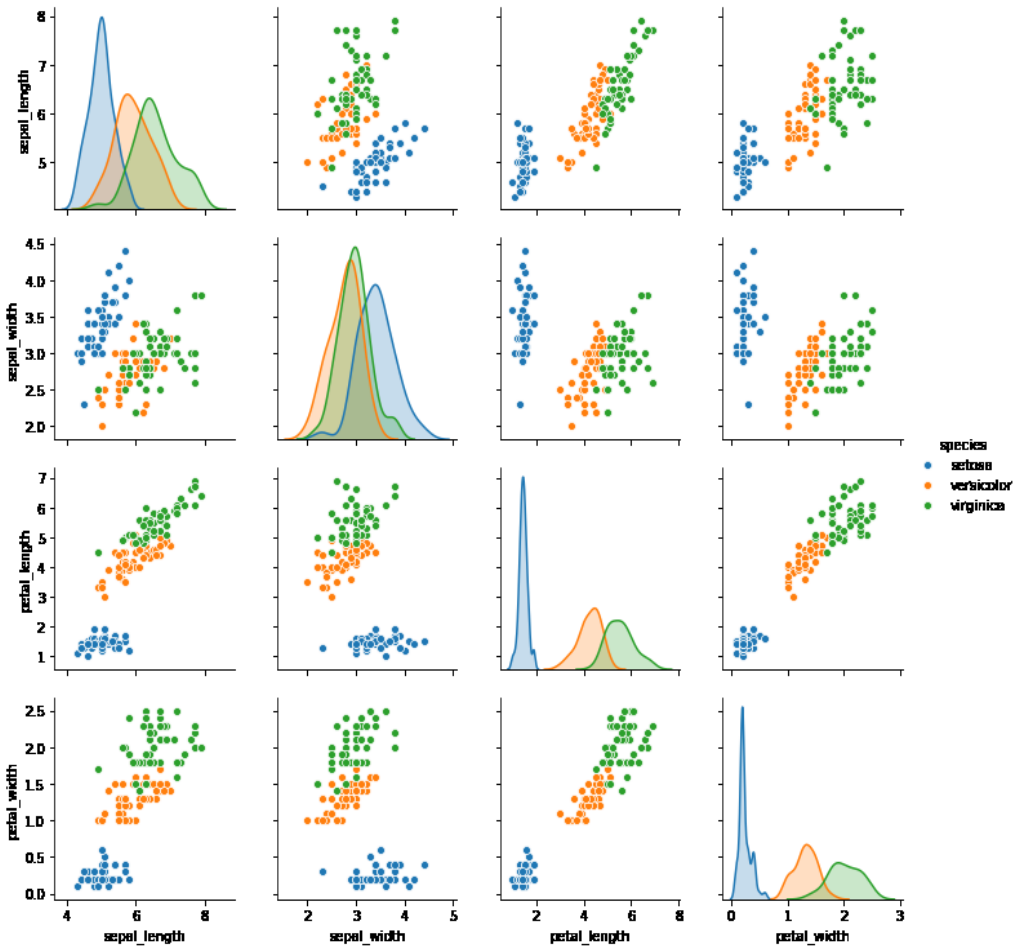
Si tenemos variables categóricas se pueden diferenciar los grupos.



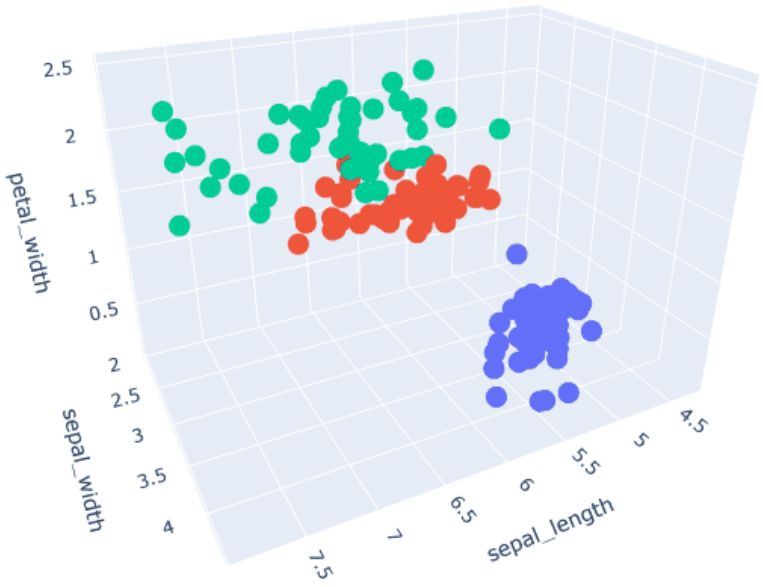
Si tenemos más de dos variables aleatorias, podemos ver los scatterplots de cada par de ellas en un mismo gráfico, el Scatterplot Matrix. En la diagonal del gráfico vienen los histogramas de cada variable.



Podemos diferenciar por grupos, y en vez de poner histogramas poner estimaciones de kernel densities.



Si ponemos en los tres ejes de coordenadas tres variables aleatorias de las que tengamos, podemos obtener un scatterplot 3-dimensional, incluso con movimiento.



MATRIZ DE CORRELACIONES GRÁFICA

La correlación es una medida numérica estadística que permite medir el grado de relación lineal entre un par de variables aleatorias. Cumple las siguientes propiedades:

- Toma valores en el rango de -1 a 1.
- Si vale -1 significa que la relación lineal es fuerte y negativa o inversa.
- Si vale 1 significa que la relación lineal es fuerte y positiva o directa.
- Si vale 0 significa que no hay relación lineal entre las dos variables.

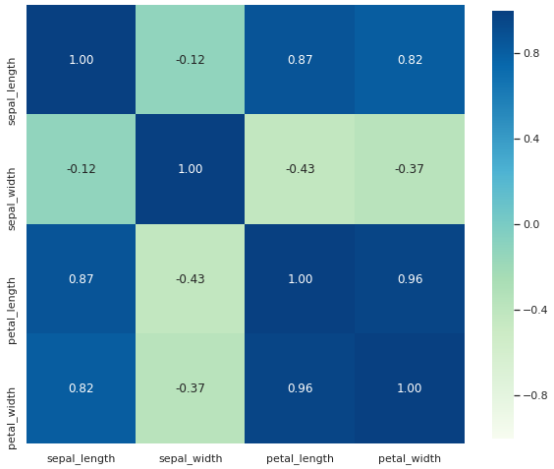
Supongamos que tenemos una matriz de datos en el espacio multivariante:

$\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_\mathrm{p}]$, donde $\mathbf{x_j} \in \mathbb{R}^{\boldsymbol{n}}$, para todas las $j = 1, \dots, p$.

La correlación entre dos variables $\mathbf{x}_i$ y $\mathbf{x}_j$ es igual a la covarianza estandarizada por las desviaciones de las variables:

$$corr(\mathbf{x}_i, \mathbf{x}_j) = \frac{cov(\mathbf{x}_i, \mathbf{x}_j)}{s_{\mathbf{x}_i} s_{\mathbf{x}_j}}$$

La matriz de correlaciones se define con cada elemento igual a la correlación entre las variables $\mathbf{x}_i$ y $\mathbf{x}_j$. En la diagonal vale 1 porque es la correlación entre cada variable consigo misma ($corr(\mathbf{x}_i, \mathbf{x}_i)$). Es simétrica porque $corr(\mathbf{x}_i, \mathbf{x}_j) = corr(\mathbf{x}_j, \mathbf{x}_i)$ para cualquier par de variables $\mathbf{x}_i$ y $\mathbf{x}_j$. Convirtiendo cada valor de correlación en un color dentro de una escala de colores en el rango de -1 a 1, la matriz de correlaciones se puede representar gráficamente:

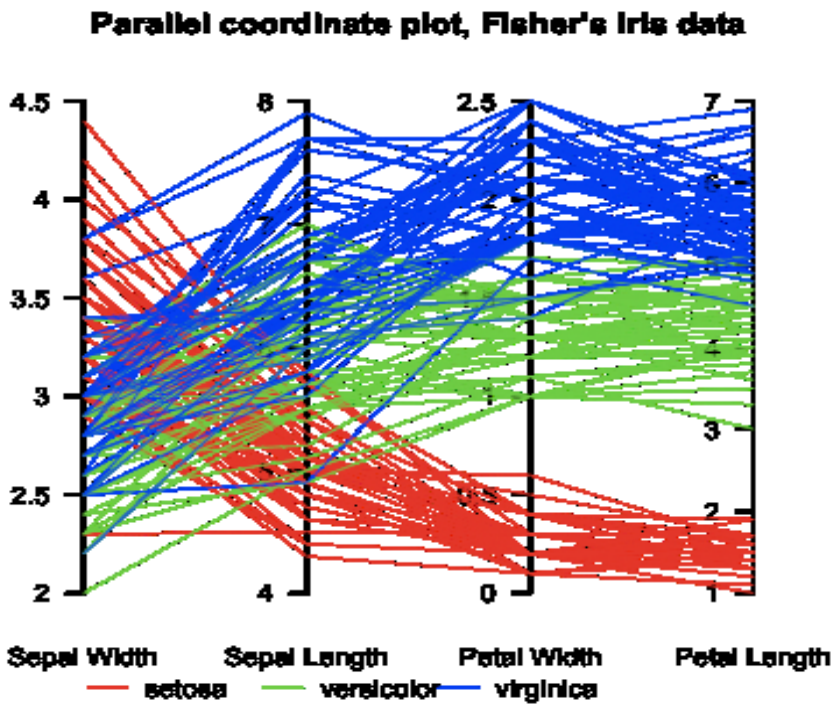


COORDENADAS PARALELAS

En vez de graficar las observaciones en un sistema de coordenadas ortogonales, podríamos poner cada coordenada en ejes paralelos y conectarlos con líneas:

- Las variables estarán en el eje horizontal y las observaciones en el eje vertical.
- Esta representación es muy útil en alta dimensión.
- Pero tiene la desventaja de que el orden de las variables influye en los posibles patrones que podrían mostrarse.

Podemos dibujar todas las observaciones con un mismo color, o diferenciar por subgrupos si se dispone de una variable categórica. En el ejemplo vemos que el comportamiento de las variables depende fuertemente de la especie de flor.



CURVAS DE ANDREWS

Es una versión suavizada de la representación en coordenadas paralelas. Cada observación se representa como una función de $-\pi$ a π. Las funciones se pueden interpretar como proyecciones de los datos en el espacio de las

Series de Fourier. Si hay alguna estructura entre los datos, puede ser visible con este gráfico.

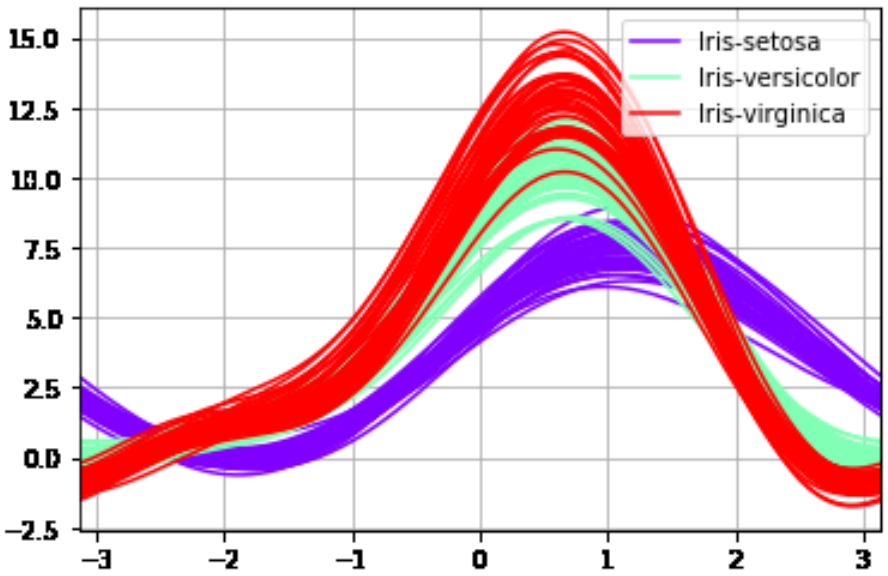
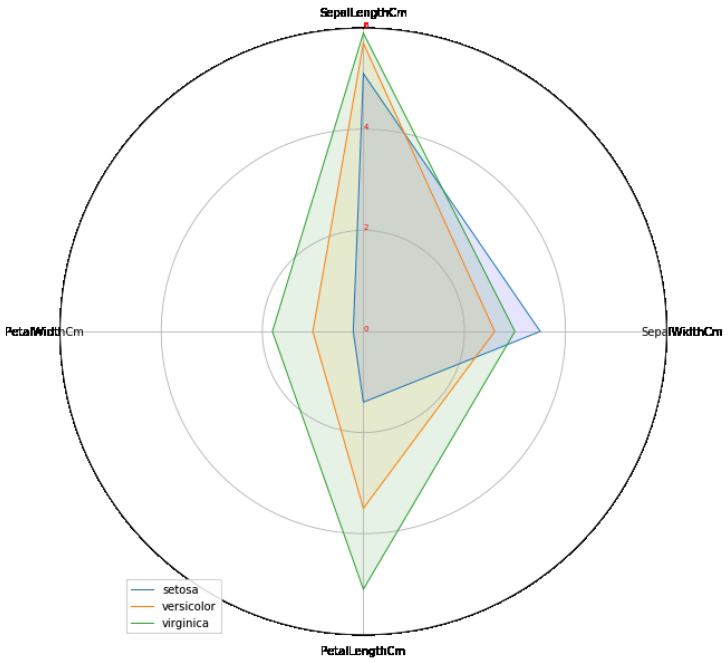


GRÁFICO DE RADAR

Es un gráfico bidimensional de tres o más variables cuantitativas representadas en ejes que comienzan desde el mismo punto. Es equivalente a una gráfica de coordenadas paralelas, con los ejes dispuestos radialmente. También se conoce como gráfico web, gráfico de araña, gráfico de estrellas, o gráfico polar Se pueden observar las formas de las diferentes observaciones y las diferencias entre las magnitudes para cada variable.



ANÁLISIS EXPLORATORIO – MEDIDAS DESCRIPTIVAS

LOCALIZACIÓN

La medida de localización clásica para describir datos multivariantes es el **vector de medias**, que tiene dimensión = cantidad de variables que tengamos. Recoge las medias de cada una de las p variables. Se calcula fácilmente mediante:

$$\bar{\mathbf{x}} = [\overline{\mathbf{x_1}}, \dots, \overline{\mathbf{x_p}}]$$

También existe el **vector de medianas**, resultado de hallar las medianas de cada variable univariante.

$$med(\mathbf{x}) = [\boldsymbol{med}(\mathbf{x_1}), \dots, \boldsymbol{med}(\mathbf{x_p})]$$

Sin embargo, pensemos en la definición de mediana univariante. El concepto univariante de la mediana se basa en nuestra capacidad para ordenar los datos. En el caso de datos multivariantes no hay un orden natural de los puntos de datos. Para desarrollar el concepto de la mediana en este caso, primero tenemos que acordar alguna convención para definir el "**orden**". En el espacio multivariante podemos definir el "**orden**" de varias maneras. Por ejemplo, podemos utilizar el concepto "**distancia**". Un ejemplo de distancia entre dos puntos del espacio multivariante es la **distancia Euclídea**.

Sean $\mathbf{x_i} = (x_{i1}, \dots, x_{ip})$ y $\mathbf{x_j} = (x_{j1}, \dots, x_{jp})$ dos observaciones p −dimensionales de nuestros datos, resultado de medir p variables en los individuos o elementos i y j. La **distancia Euclídea** entre dos observaciones $\mathbf{x_i}$ y $\mathbf{x_j}$ es:

$$d_E(\mathbf{x_i}, \mathbf{x_j}) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} = \sqrt{(\mathbf{x_i} - \mathbf{x_j})^t (\mathbf{x_i} - \mathbf{x_j})}$$

En el espacio **univariante** la mediana tiene la propiedad de minimizar la suma de distancias a los datos. La **mediana geométrica** es la idea análoga en el espacio de mayor dimensión, es decir, es el punto que minimiza la suma de las distancias euclídeas a los puntos de la muestra.

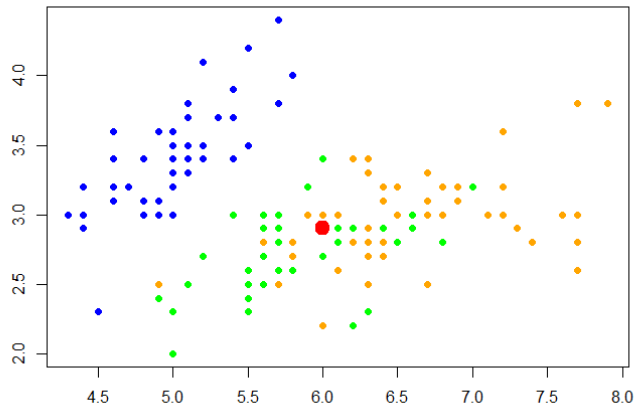
$$\mu_G = \underset{\mathbf{y}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} d_E(\mathbf{x_i}, \mathbf{y}) = \underset{\mathbf{y}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \|\mathbf{x_i} - \mathbf{y}\|_2$$

La mediana geométrica en el espacio de una sola dimensión (univariante) coincide con la mediana. También se le llama: Mediana espacial, o Mediana L_1

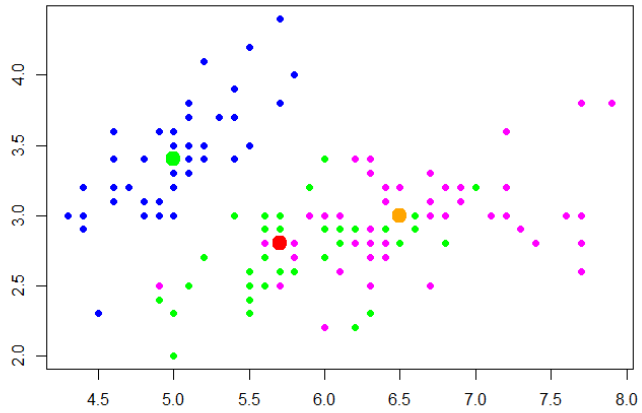
Otra forma de ordenar los datos en el espacio multivariante es con medidas de profundidad (Depth). La **medida de profundidad** de una observación, relativa a las otras observaciones, mide cuán profunda es esa observación dentro de la nube de datos. Esto proporciona un ordenamiento de las observaciones desde el centro hacia afuera, mientras mayor es la profundidad, más cerca del centro de los datos, y mientras menor es la profundidad, más alejada estará del centro. La observación más profunda (con profundidad máxima) puede definirse como la mediana multivariante.

La medida de profundidad de Tukey de una observación x_i respecto a las observaciones en $\mathbf{x}$, se define como la fracción mínima de observaciones en $\mathbf{x}$ contenidas en un semiespacio cerrado que contiene a x_i. El problema es que es muy difícil computacionalmente cuando la dimensión crece. Pero hay alternativas, como por ejemplo, generación aleatoria de los semiespacios.

Podemos calcular la profundidad de cada punto de nuestros datos hacia todos los demás. El centro multivariante será la observación con mayor profundidad.



Se puede dividir por subgrupos si disponemos de variables categóricas y calcular centros para cada grupo.



DISPERSIÓN Y DEPENDENCIA

En el espacio univariante, la varianza es la media de las distancias al cuadrado de los valores a la media.

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2$$

Vamos a usar la versión insesgada de la varianza, porque es mejor estimador.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2$$

La desviación típica es la raíz cuadrada de la varianza:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2}$$

En el caso multidimensional, cada variable tiene asociada una varianza y una desviación típica.

La covarianza es un valor que indica el grado de variación conjunta de dos variables aleatorias $\mathbf{x_1}$ y $\mathbf{x_2}$ respecto a sus medias, lo cual permite determinar si existe una dependencia entre ambas variables.

$$s_{12} = cov(\mathbf{x_1}, \mathbf{x_2}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \overline{\mathbf{x_1}})(x_{i2} - \overline{\mathbf{x_2}})$$

En el caso multidimensional, tendríamos una covarianza para cada par de variables.

Valores altos de una de las variables ⇒ valores altos de la otra, o viceversa, hay una dependencia lineal positiva entre esas dos variables ⇒ signo positivo de la covarianza.

Si mayores valores de una variable ⇒ menores valores de la otra, o viceversa, es decir, hay un comportamiento opuesto ⇒ signo negativo de la covarianza.

El signo de la covarianza, por lo tanto, expresa la tendencia en la relación lineal entre las variables.

Problema: El valor de la covarianza dependerá de las unidades de medida de las variables.

Solución: La versión normalizada de la covarianza, el coeficiente de correlación, indica la magnitud de la relación lineal.

El coeficiente de correlación muestral está definido como:

$$r = \frac{\sum_{i=1}^{n}(x_{i1} - \overline{\mathbf{x_1}})(x_{i2} - \overline{\mathbf{x_2}})}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \overline{\mathbf{x_1}})^2} * \sqrt{\sum_{i=1}^{n}(x_{i2} - \overline{\mathbf{x_2}})^2}}$$

Es decir, es una normalización de la covarianza, según las desviaciones típicas de las variables.

$$r_{12} = \frac{cov(\mathbf{x_1}, \mathbf{x_2})}{s_{\mathbf{x_1}} * s_{\mathbf{x_2}}}$$

El coeficiente de correlación toma valores entre -1 (relación $-$) y 1 (relación $+$), y si $r = 0$ significa que hay ausencia de relación lineal.

En el caso multidimensional tenemos que tener en cuenta todas las relaciones entre las variables. Las medidas que tenemos caracterizan la dispersión o variación a pares.

Variables	$\mathbf{x_1}$	$\mathbf{x_2}$	$\cdots$	$\mathbf{x_p}$
$\mathbf{x_1}$	s_1^2	s_{12}	$\cdots$	s_{1p}
$\mathbf{x_2}$	s_{21}	s_2^2	$\cdots$	s_{2p}
$\cdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\mathbf{x_p}$	s_{p1}	s_{p2}	$\cdots$	s_p^2

En el caso multidimensional, la matriz de varianzas y covarianzas muestrales va a caracterizar la dispersión conjunta y las dependencias de las variables que tenemos:

$$S_{\mathbf{x}} = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{bmatrix}$$

Es una matriz cuadrada de tamaño $p \times p$. En la diagonal están las varianzas (cada variable con ella misma) porque $cov(\mathbf{x}, \mathbf{x}) = var(\mathbf{x})$. Los elementos por arriba de la diagonal serán iguales que los elementos por debajo (matriz simétrica) porque $cov(\mathbf{x}, \mathbf{y}) = cov(\mathbf{y}, \mathbf{x})$. Usaremos los estimadores muestrales insesgados para que $\mathrm{E}(S_{\mathbf{x}}) = \Sigma$

S puede ser escrita en términos de la matriz de datos centrados $\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{1}_n \bar{\mathbf{x}}^t$:

$$S_{\mathbf{x}} = \frac{1}{n-1} \tilde{\mathbf{x}}^t \tilde{\mathbf{x}}$$

S es semi-definida positiva, es decir sus autovalores son no negativos $\lambda_j^S \geq 0,\ j = 1, \ldots, p.$

$$|S| = \prod_{j=1}^{p} \lambda_j^S \geq 0$$

Cuando $|S| = 0$ hay algunas variables que son combinación lineal de otras.

El rango de la matriz S es el número de variables linealmente independientes.

Si $|S| = 0$ es necesario eliminar las variables redundantes.

$$\mathrm{Tr(S)} = \mathrm{s}_1^2 + \cdots + \mathrm{s}_\mathrm{p}^2 = \lambda_1^S + \cdots + \lambda_p^S \geq 0$$

La matriz de datos centrados es $\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{1}_n \bar{\mathbf{x}}^t$

Ejemplo:

Si $\mathbf{x} = \begin{pmatrix} \mathbf{1} & \mathbf{-1} \\ \mathbf{1} & \mathbf{-2} \\ \mathbf{2} & \mathbf{0} \\ \mathbf{0} & \mathbf{-1} \end{pmatrix}$ con vector de medias $\bar{\mathbf{x}} = \begin{pmatrix} \mathbf{1} \\ \mathbf{-1} \end{pmatrix}$

$$\tilde{\mathbf{x}} = \begin{pmatrix} \mathbf{1} & \mathbf{-1} \\ \mathbf{1} & \mathbf{-2} \\ \mathbf{2} & \mathbf{0} \\ \mathbf{0} & \mathbf{-1} \end{pmatrix} - \begin{pmatrix} \mathbf{1} \\ \mathbf{1} \\ \mathbf{1} \\ \mathbf{1} \end{pmatrix} (\mathbf{1} \quad \mathbf{-1}) = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{-1} \\ \mathbf{1} & \mathbf{1} \\ \mathbf{-1} & \mathbf{0} \end{pmatrix}$$

Las covarianzas y la matriz de covarianza en general tienen un problema a la hora de interpretar la relación de dependencia entre las variables. La magnitud de los valores depende de las unidades de medida de las variables. Es por eso que, con la covarianza, sólo es posible interpretar si existe una relación de dependencia positiva o negativa, pero no podemos interpretar si es fuerte o débil. Si $s_{12} > 0$, hay dependencia directa (positiva), es decir, a grandes valores de $\mathbf{x}_1$ corresponden grandes valores de $\mathbf{x}_2$. Si $s_{12} < 0$, hay dependencia inversa (negativa), es decir, a grandes valores de $\mathbf{x}_1$ corresponden pequeños valores de $\mathbf{x}_2$. Si $s_{12} = 0$, se interpreta como la no existencia de una relación lineal entre las dos variables.

Una solución sería estandarizar las variables $\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_p$ para que no dependan de las unidades de medida:

$$z_{ij} = \frac{x_{ij} - \bar{\mathbf{x}}_j}{s_j}$$

La covarianza muestral entre las variables $\mathbf{z}_j$ y $\mathbf{z}_k$ sería entonces

$$\boldsymbol{r_{jk}} = \frac{1}{n-1} \sum_{i=1}^{n} \mathbf{z}_{ij}\, \mathbf{z}_{ik} = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_{ij} - \bar{\mathbf{x}}_j}{s_j} \frac{x_{ik} - \bar{\mathbf{x}}_k}{s_k} = \frac{s_{jk}}{s_j s_k}$$

A esto se le llama correlación muestral entre las variables $\mathbf{x}_j$ y $\mathbf{x}_k$. La correlación mide dependencia entre las dos variables sin que las unidades de medida influyan.

También se puede definir una matriz de correlaciones muestrales:

$$R_{\mathbf{x}} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

Es una matriz cuadrada de tamaño $p \times p$. En la diagonal están las correlaciones de cada variable con ella misma por lo tanto en la diagonal hay 1's. Los elementos por arriba de la diagonal serán iguales que los elementos por debajo (matriz simétrica) porque $r_{12} = r_{21}$.

- Si $r_{12} \approx 1$, hay dependencia positiva y fuerte.
- Si $r_{12} \approx -1$, hay dependencia negativa y fuerte.
- Si $r_{12} \approx 0$, se interpreta como la no existencia de una relación lineal entre las dos variables, o una dependencia débil.
- $r_{12} = 0$ si y sólo si $s_{12} = 0$, en cuyo caso diremos que las variables son incorreladas.

R puede ser escrita en términos de la matriz de covarianza:

$$R = D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$$

donde D es una matriz diagonal de tamaño $p \times p$ que contiene los elementos de la diagonal principal de S, es decir las varianzas $s_1^2, \ldots, s_p^2$.

R es semi-definida positiva, es decir, sus autovalores son no negativos $\lambda_j^R \geq 0,\ j = 1, \ldots, p$.

$$|R| = \prod_{j=1}^{p} \lambda_j^R \geq 0$$

Cuando $|R| = 0$ hay algunas variables que son combinación lineal de otras.

El rango de la matriz R es el número de variables linealmente independientes.

Si $|R| = 0$ es necesario eliminar las variables redundantes.

$$\mathrm{Tr(R)} = 1 + \cdots + 1 = \lambda_1^R + \cdots + \lambda_p^R = p$$

TRANSFORMACIONES LINEALES

Supongamos que tenemos nuestra matriz de datos $\mathbf{x}$ de tamaño $n \times p$ y sea $\mathbf{c} = (c_1, \ldots, c_p)^t$ un vector columna de tamaño $p \times 1$. El resultado de multiplicar la matriz de datos y el vector columna es $\mathbf{y} = \mathbf{xc}$, una nueva variable que es combinación lineal de las variables de $\mathbf{x}$.

$$\mathbf{xc} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_p \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \mathbf{y}$$

donde $y_i = c_1 x_{i1} + \cdots + c_p x_{ip}$ para $i = 1, \ldots, n$.

La media y la varianza de la nueva variable son:

$$\bar{\mathbf{y}} = \mathbf{c}^{\mathrm{t}} \bar{\mathbf{x}} \quad \text{y} \quad s_y^2 = \mathbf{c}^t S_{\mathbf{x}} \mathbf{c}$$

donde $\bar{\mathbf{x}}$ y $S_{\mathbf{x}}$ son la media muestral y la matriz de covarianza de $\mathbf{x}$.

Supongamos que C es una matriz de tamaño $p \times r$, entonces $\mathbf{y} = \mathbf{x}C$ es una transformación lineal de $\mathbf{x}$.

$$\mathbf{y} = \mathbf{x}C = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} c_{11} & \cdots & c_{1r} \\ \vdots & \ddots & \vdots \\ c_{p1} & \cdots & c_{pr} \end{pmatrix} = \begin{pmatrix} y_{11} & \cdots & y_{1r} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nr} \end{pmatrix}$$

donde $y_{ij} = c_{1j} x_{i1} + \cdots + c_{pj} x_{ip}$ para $i = 1, \ldots, n$ y j= $1, \ldots, p$

La media y la matriz de covarianza de la nueva variable son:

$$\bar{\mathbf{y}} = C^{\mathrm{t}}\bar{\mathbf{x}} \quad \text{y} \quad S_{\mathbf{y}} = C^{t} S_{\mathbf{x}} C$$

EJEMPLO DATASET DE IRIS

El vector de medias del data-set de Iris es:

$$\bar{\mathbf{x}} = (5.84, 3.05, 3.75, 1.19)^{t}$$

La matriz de covarianza muestral es:

$$S_{\mathbf{x}} = \begin{pmatrix} 0.68 & -0.04 & 1.27 & 0.51 \\ -0.04 & 0.18 & -0.32 & -0.12 \\ 1.27 & -0.32 & 3.11 & 1.29 \\ 0.51 & -0.12 & 1.29 & 0.58 \end{pmatrix}$$

Queremos crear dos nuevas variables:

- La suma de la longitud del Sépalo y del Pétalo para cada flor.
- La suma del ancho del Sépalo y del Pétalo para cada flor.

Lo que queremos es crear una nueva variable $\mathbf{y} = \mathbf{x}C$, donde:

$$C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

El vector de medias sería:

$$\bar{\mathbf{y}} = C^{t}\bar{\mathbf{x}} = \begin{pmatrix} 9.60 \\ 4.25 \end{pmatrix}$$

Y la matriz de covarianzas:

$$S_{\mathbf{y}} = C^{t} S_{\mathbf{x}} C = \begin{pmatrix} 6.35 & 1.43 \\ 1.43 & 0.52 \end{pmatrix}$$

ESTANDARIZACIÓN INDIVIDUAL O UNIVARIANTE

Recordemos la matriz de datos centrada: $\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{1}_n \bar{\mathbf{x}}^{t}$

La estandarización individual de $\mathbf{x}$ puede escribirse como:

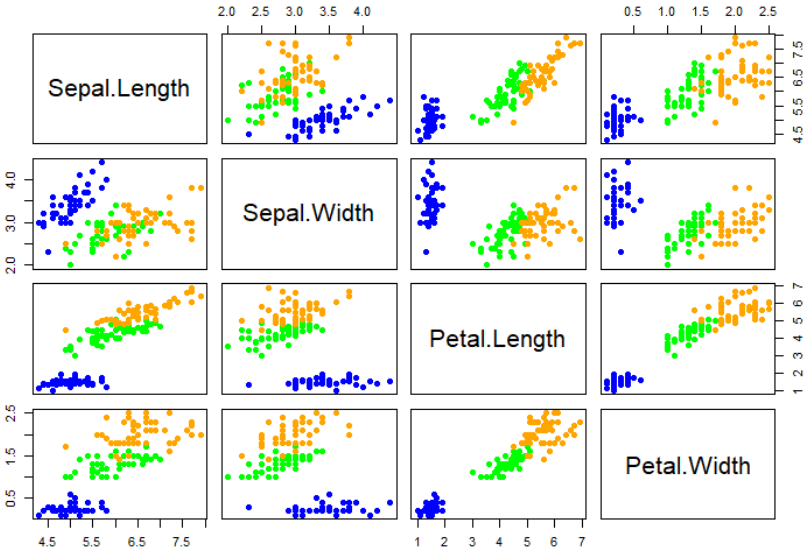
$$\mathbf{y} = \tilde{\mathbf{x}} D_{\mathbf{x}}^{-1/2}$$

Donde $D_{\mathbf{x}}$ es la matriz diagonal de tamaño $p \times p$ formada por elementos de la diagonal principal de $S_{\mathbf{x}}$, es decir, las varianzas $s_1^2, \ldots, s_p^2$.

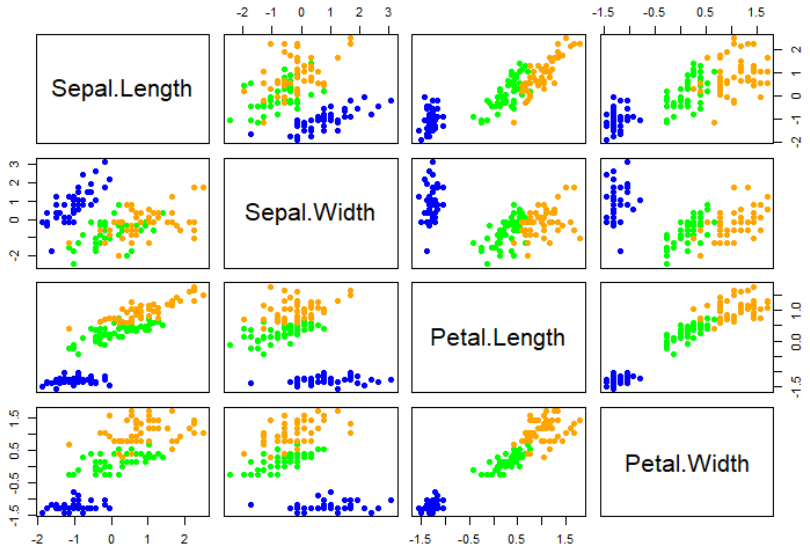
El vector de medias y la matriz de covarianza serían: $\bar{y} = \mathbf{0}_p$ y $S_{\mathbf{y}} = D_{\mathbf{x}}^{-1/2} S_{\mathbf{x}} D_{\mathbf{x}}^{-1/2} = R_{\mathbf{x}}$

Entonces, la estandarización univariante elimina la media y estandariza las varianzas.

IRIS SIN TRANSFORMACIÓN



IRIS CON TRANSFORMACIÓN UNIVARIANTE



ESTANDARIZACIÓN MULTIVARIANTE

La estandarización multivariante de X es:

$$\mathbf{y} = \tilde{\mathbf{x}} S_{\mathbf{x}}^{-1/2}$$

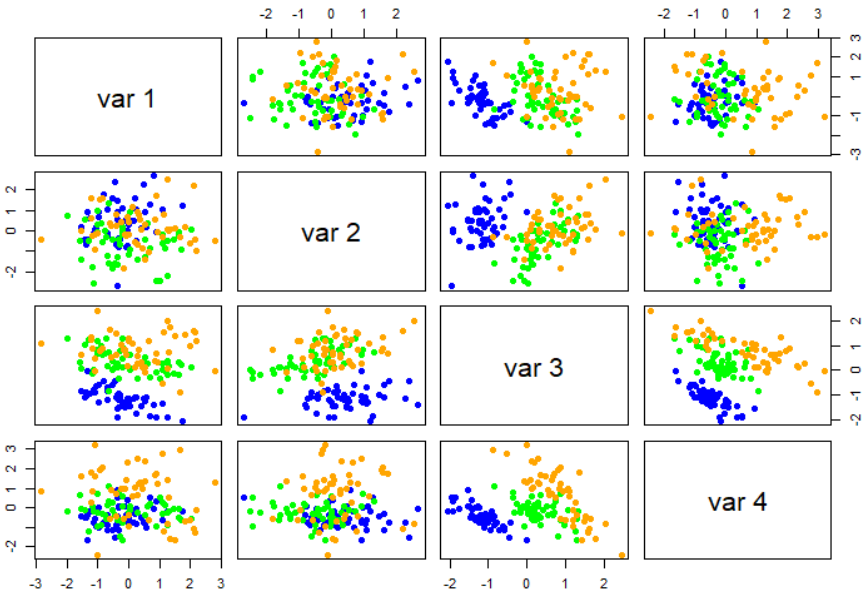
El vector de medias y la matriz de covarianza serían:

$$\bar{\mathbf{y}} = \mathbf{0}_p$$

$$S_{\mathbf{y}} = \left(S_{\mathbf{x}}^{-1/2}\right)^t S_{\mathbf{x}} \left(S_{\mathbf{x}}^{-1/2}\right) = I_p$$

Entonces, la estandarización multivariante elimina la media y también las correlaciones entre las variables, y además estandariza las varianzas.

IRIS CON TRANSFORMACIÓN MULTIVARIANTE



VARIABLE ALEATORIA MULTIVARIANTE

Cuando se mide un conjunto de variables aleatorias en cada elemento de la población, diremos que se ha definido una variable aleatoria multivariante o multidimensional. Las variables que se miden pueden ser cualitativas o cuantitativas. Algunos ejemplos de variables multivariantes son los siguientes:

- Para cada estudiante de una universidad, medimos: la edad, el sexo, la nota media del curso, la ciudad de residencia.
- Para cada una de las empresas de una zona industrial medimos: el número de trabajadores, la facturación, el sector industrial.
- Para cada país del mundo medimos: 10 indicadores de desarrollo.

En la matriz de datos cada fila representa un elemento de la población y cada columna los valores de cada variable en todos los elementos observados. La matriz tendrá n filas y p columnas, si hay n elementos en la población y se han medido p variables en cada elemento. Si llamamos a la matriz de datos $\mathbf{x}$, entonces el elemento x_{ij} representa el valor de la variable j en el individuo i, donde $i = 1, \ldots, n$ y $j = 1, \ldots, p$.

$$\mathbf{x} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}]$$

cada variable $\mathbf{x_j}$ es un vector columna de tamaño $n \times 1$ que representa la variable j, medida en los n elementos de la población.

Por ejemplo, medimos la edad: $\mathbf{x_1}$, género: $\mathbf{x_2}$, nota promedio del curso: $\mathbf{x_3}$, y ciudad de residencia: $\mathbf{x_4}$ de 17 estudiantes.

Es decir, tenemos $n = 17$ observaciones y $p = 4$ variables.

$$\mathbf{x} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} \\ x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} \\ x_{3,1} & x_{3,2} & x_{3,3} & x_{3,4} \\ \vdots & \vdots & \ddots & \vdots \\ x_{17,1} & x_{17,2} & x_{17,3} & x_{17,4} \end{bmatrix} = [\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4}]$$

DISTRIBUCIÓN CONJUNTA Y MARGINALES

Podemos decir que tenemos la distribución conjunta de una variable aleatoria multivariante cuando se especifica:

i. El espacio muestral de los posibles valores que toma la variable, que, en general, es un subconjunto de $\mathbb{R}^p$.

ii. Las probabilidades de cada resultado posible del espacio muestral.

Considere el vector aleatorio $[\mathbf{x}, \mathbf{y}]$, su función de distribución F_{xy} tiene las siguientes propiedades:

i. $0 \leq F_{\mathrm{xy}} \leq 1$

ii. $F_{\mathrm{xy}}(x, y)$ es monótona creciente con respecto a $\mathbf{x}$ e $\mathbf{y}$

iii. $\lim_{x,y\to-\infty} F_{\mathrm{xy}}(x, y) = 0$, $\lim_{x\to-\infty} F_{\mathrm{xy}}(x, y) = 0$, $\lim_{y\to-\infty} F_{\mathrm{xy}}(x, y) = 0$

iv. $\lim_{x,y\to+\infty} F_{\mathrm{xy}}(x, y) = 1$

v. $\lim_{x\to+\infty} F_{\mathrm{xy}}(x, y) = F_{\mathrm{y}}(y)$: es la distribución marginal de $\mathbf{y}$

vi. $\lim_{y\to+\infty} F_{\mathrm{xy}}(x, y) = F_{\mathrm{x}}(x)$: es la distribución marginal de $\mathbf{x}$

El vector aleatorio $[\boldsymbol{x}, \boldsymbol{y}]$ es continuo si existe una función de densidad conjunta f_{xy} de modo que la función de distribución conjunta pueda expresarse como:

$$F_{\mathrm{xy}}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{\mathrm{xy}}(u, v) du\, dv \ , \qquad (x, y) \in \mathbb{R}^2$$

Las funciones de densidad marginales son:

$$f_{\mathrm{x}}(x) = \int_{-\infty}^{+\infty} f_{\mathrm{xy}}(x, y) dy$$

$$f_{\mathrm{y}}(y) = \int_{-\infty}^{+\infty} f_{\mathrm{xy}}(x, y) dx$$

DISTRIBUCIÓN CONDICIONAL

Sean $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ e $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_q]$ dos variables aleatorias multivariantes con funciones de densidad f_{x} y f_{y}, respectivamente, y función de densidad conjunta f_{xy}.

La función de densidad condicional de $\mathbf{y}$ dado $\mathbf{x}$ es:

$$f_{\mathbf{y}|\mathbf{x}}(y|x) = \frac{f_{\mathbf{xy}}(x,y)}{f_{\mathbf{x}}(x)}$$

La función de densidad condicional de $\mathbf{y}$ dado $\mathbf{x}$ es:

$$f_{\mathbf{x}|\mathbf{y}}(x|y) = \frac{f_{\mathbf{xy}}(x,y)}{f_{\mathbf{y}}(y)}$$

TEOREMA DE BAYES

De la definición anterior podemos definir una expresión para la función de densidad conjunta dependiendo de la condicional:

$$f_{\mathbf{xy}}(x,y) = f_{\mathbf{y}|\mathbf{x}}(y|x)f_{\mathbf{x}}(x) = f_{\mathbf{x}|\mathbf{y}}(x|y)f_{\mathbf{y}}(y)$$

De la última igualdad, sale que:

$$f_{\mathbf{y}|\mathbf{x}}(y|x) = \frac{f_{\mathbf{x}|\mathbf{y}}(x|y)f_{\mathbf{y}}(y)}{f_{\mathbf{x}}(x)} = \frac{f_{\mathbf{x}|\mathbf{y}}(x|y)f_{\mathbf{y}}(y)}{\int f_{\mathbf{xy}}(x,y)dy} = \frac{f_{\mathbf{x}|\mathbf{y}}(x|y)f_{\mathbf{y}}(y)}{\int f_{\mathbf{x}|\mathbf{y}}(x|y)f_{\mathbf{y}}(y)dy}$$

La expresión anterior define el Teorema de Bayes, uno de los resultados más importantes de Estadística y la base de la Inferencia Bayesiana.

INDEPENDENCIA

Las variables multivariantes $\mathbf{x}$ *e* $\mathbf{y}$, son independientes si, y sólo si:

$$f_{\mathbf{xy}}(x,y) = f_{\mathbf{x}}(x)f_{\mathbf{y}}(y), \qquad \forall x,y$$

Esto es equivalente a usar las funciones de distribución:

$$F_{\mathbf{xy}}(x,y) = F_{\mathbf{x}}(x)F_{\mathbf{y}}(y), \qquad \forall x,y$$

Si $\mathbf{x}$ *e* $\mathbf{y}$ son independientes, entonces las condicionales:

$$f_{\mathbf{y}|\mathbf{x}}(y|x) = f_{\mathbf{y}}(y)$$

$$f_{\mathbf{x}|\mathbf{y}}(x|y) = f_{\mathbf{x}}(x)$$

Es decir, si son independientes, condicionar una con respecto a la otra no produce ningún cambio.

VALOR ESPERADO

Sea $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ una variable aleatoria multivariante $p-$dimensional, donde cada variable está medida en n elementos: $\mathbf{x}_j \in \mathbb{R}^n$, para toda $j = 1, \dots, p$.

El valor esperado, esperanza matemática o vector de medias de $\mathbf{x}$ es un vector cuyos componentes son los valores esperados de cada uno los componentes del vector aleatorio, es decir, de cada variable univariante que lo compone:

$$\mu_{\mathbf{x}} = E[\mathbf{x}] = \begin{pmatrix} E[\mathbf{x}_1] \\ \vdots \\ E[\mathbf{x}_p] \end{pmatrix}$$

donde

$$E[\mathbf{x}_j] = \int_{-\infty}^{+\infty} \mathbf{x}_j \, f_{\mathbf{x}_j}(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j$$

Y $f_{\mathbf{x}_j}(\mathbf{x}_j)$ es la densidad marginal de $\mathbf{x}_j$, para cada $j = 1, \dots, p$.

COVARIANZA Y CORRELACIÓN

La matriz de covarianza de una variable aleatoria multivariante $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ con valor esperado $\mu_{\mathbf{x}}$, es una matriz simétrica y semi-definida positiva:

$$\Sigma_{\mathbf{x}} = E[(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})^t]$$

Los elementos dentro de la diagonal de $\Sigma_{\mathbf{x}}$ son las varianzas de cada variable:

$$\sigma^2_{\mathbf{x}_j} = \int_{-\infty}^{+\infty} \left(\mathbf{x}_j - \mu_{\mathbf{x}_j}\right)^2 f_{\mathbf{x}_j}(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j$$

para cada $j = 1, \dots, p$.

Los elementos fuera de la diagonal de $\Sigma_{\mathbf{x}}$ son las covarianzas entre pares de variables:

$$\sigma_{\mathbf{x}_j\mathbf{x}_k} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left(\mathbf{x}_j - \mu_{\mathbf{x}_j}\right)\left(\mathbf{x}_k - \mu_{\mathbf{x}_k}\right) f_{\mathbf{x}_j\mathbf{x}_k}(\mathbf{x}_j, \mathbf{x}_k)\mathrm{d}\mathbf{x}_j\mathrm{d}\mathbf{x}_k$$

para cada $j, k = 1, \ldots, p$ y $j \neq k$.

Los elementos fuera de la diagonal de $\Sigma_{\mathbf{x}}$ son las covarianzas entre pares de variables:

$$\sigma_{\mathbf{x}_j\mathbf{x}_k} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left(\mathbf{x}_j - \mu_{\mathbf{x}_j}\right)\left(\mathbf{x}_k - \mu_{\mathbf{x}_k}\right) f_{\mathbf{x}_j\mathbf{x}_k}(\mathbf{x}_j, \mathbf{x}_k)\mathrm{d}\mathbf{x}_j\mathrm{d}\mathbf{x}_k$$

para cada $j, k = 1, \ldots, p$ y $j \neq k$.

Otra forma de ver la definición de covarianzas es:

$$\sigma_{\mathbf{x}_j\mathbf{x}_k} = E\left[\left(\mathbf{x}_j - E\left[\mathbf{x}_j\right]\right)\ \left(\mathbf{x}_k - E[\mathbf{x}_k]\right)\ \right] = E\left(\mathbf{x}_j\mathbf{x}_k\right) - E\left(\mathbf{x}_j\right)E(\mathbf{x}_k)$$

Donde el valor esperado $E\left(\mathbf{x}_j\mathbf{x}_k\right)$ es:

$$E\left(\mathbf{x}_j\mathbf{x}_k\right) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathbf{x}_j\mathbf{x}_k\, f_{\mathbf{x}_j\mathbf{x}_k}(\mathbf{x}_j, \mathbf{x}_k)\mathrm{d}\mathbf{x}_j\mathrm{d}\mathbf{x}_k$$

La matriz de correlación de una variable aleatoria multivariante $\mathbf{x} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_p\right]$ con matriz de covarianza $\Sigma_{\mathbf{x}}$, es:

$$\varrho_{\mathbf{x}} = \Delta_{\mathbf{x}}^{-1/2}\Sigma_{\mathbf{x}}\Delta_{\mathbf{x}}^{-1/2}$$

Donde $\Delta_{\mathbf{x}}$ es una matriz diagonal con las varianzas de los componentes de $\mathbf{x}$.

Los elementos fuera de la diagonal de la matriz $\varrho_{\mathbf{x}}$ son las correlaciones entre pares de variables:

$$\rho_{\mathbf{x}_j\mathbf{x}_k} = \frac{\sigma_{\mathbf{x}_j\mathbf{x}_k}}{\sigma_{\mathbf{x}_j}\sigma_{\mathbf{x}_k}}$$

para cada $j, k = 1, \ldots, p$ y $j \neq k$.

Para $j = k$, estamos con los elementos de la diagonal de la matriz de correlaciones $\varrho_{\mathbf{x}}$:

$$\rho_{\mathbf{x}_j\mathbf{x}_j} = \frac{\sigma_{\mathbf{x}_j\mathbf{x}_j}}{\sigma_{\mathbf{x}_j}\sigma_{\mathbf{x}_j}} = \frac{\sigma_{\mathbf{x}_j}^2}{\sigma_{\mathbf{x}_j}^2} = 1$$

para todo $j = 1, \dots, p$.

Sean $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_\mathrm{p}]$ e $\mathbf{y} = [\mathbf{y}_1, \dots, \mathbf{y}_\mathrm{q}]$ dos variables aleatorias multivariantes con vectores de media $\mu_\mathbf{x}$ y $\mu_\mathbf{y}$, y matrices de covarianza Σ_x y Σ_y, respectivamente.

La matriz de covarianza entre $\mathbf{x}$ e $\mathbf{y}$ es una matriz de tamaño $\mathrm{p} \times \mathrm{q}$ dada por:

$$cov[\mathbf{x}, \mathbf{y}] = E\left[(\mathbf{x} - \mu_\mathbf{x})(\mathbf{y} - \mu_\mathbf{y})^t\right]$$

Similarmente, la matriz de correlaciones entre $\mathbf{x}$ e $\mathbf{y}$ es una matriz de tamaño $\mathrm{p} \times \mathrm{q}$ dada por:

$$cor[\mathbf{x}, \mathbf{y}] = \Delta_x^{-1/2} cov[\mathbf{x}, \mathbf{y}] \Delta_\mathrm{y}^{-1/2}$$

Donde Δ_x y Δ_y son matrices diagonales con elementos iguales a los elementos de la diagonal de Σ_x y Σ_y, respectivamente.

ESPERANZA CONDICIONAL

Sean $\mathbf{x}$ e $\mathbf{y}$ dos variables aleatorias con funciones de densidad $f_\mathbf{x}$ y $f_\mathbf{y}$, respectivamente, y sea $f_{\mathbf{y}|\mathbf{x}}$ la función de densidad condicional de $\mathbf{y}$ dado $\mathbf{x}$.

El valor esperado condicional o esperanza condicional de $\mathbf{y}$ dado $\mathbf{x}$ es:

$$E_{\mathbf{y}|\mathbf{x}}[y|x] = \int y f_{\mathbf{y}|\mathbf{x}}(y|x) dy = \int y \frac{f_{\mathbf{xy}}(x, y)}{f_\mathbf{x}(x)} dy$$

El valor esperado condicional o esperanza condicional de $\mathbf{x}$ dado $\mathbf{y}$ es:

$$E_{\mathbf{x}|\mathbf{y}}[x|y] = \int x f_{\mathbf{x}|\mathbf{y}}(y|x) dx = \int x \frac{f_{\mathbf{xy}}(x, y)}{f_\mathbf{y}(x)} dx$$

El valor esperado condicional de $\mathbf{y}$ dado $\mathbf{x}$ depende de x.

El valor esperado condicional de $\mathbf{x}$ dado $\mathbf{y}$ depende de y.

Podemos hallar $E_\mathbf{y}[y]$ calculando primero la esperanza condicional y luego hallandole a eso la esperanza con respecto a la distribución de $\mathbf{x}$:

$$E_{\mathbf{y}}[y] = E_{\mathbf{x}}\left[E_{\mathbf{y}|\mathbf{x}}[\, y|x\,]\right]$$

La matriz de covarianza y la matriz de correlación condicionales de $\mathbf{y}$ dado $\mathbf{x}$ son las matrices de covarianza y correlación de la variable aleatoria condicional $\mathbf{y}|\mathbf{x}$.

En particular, la matriz de covarianza condicional contiene las varianzas condicionales: $var_{\mathbf{y}_j|\mathbf{x}}[y_j|x]$ y las covarianzas condicionales: $cov_{\mathbf{y}_j,\mathbf{y}_k|\mathbf{x}}[y_j y_k|x]$

PROPIEDADES DE LA ESPERANZA CONDICIONAL

Sean a y b constantes, g una función de valor real, y supongamos que $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ son variables conjuntamente distribuidas:

1) $E[a|\mathbf{y}] = a$
2) $E[a\mathbf{x} + b\mathbf{z}|\mathbf{y}] = aE[\mathbf{x}|\mathbf{y}] + bE[\mathbf{z}|\mathbf{y}]$
3) $E[\mathbf{x}|\mathbf{y}] \geq \mathbf{0}, \quad si \quad \mathbf{x} \geq 0$
4) $E[\mathbf{x}|\mathbf{y}] = E[\mathbf{x}]$ si $\mathbf{x}$ e $\mathbf{y}$ son independientes.
5) $E[\mathbf{x}g(\mathbf{y})|\mathbf{y}] = g(\mathbf{y})E[\mathbf{x}|\mathbf{y}]$
6) $E\big[E[\mathbf{x}|\mathbf{y}]\big] = E[\mathbf{x}]$

LEY DE LA VARIANZA TOTAL

$$var_{\mathbf{y}_j}[y_j] = E_{\mathbf{x}}\left[var_{\mathbf{y}_j|\mathbf{x}}[y_j|x]\,\right] + var_{\mathbf{x}}\left[E_{\mathbf{y}_j|\mathbf{x}}[y_j|x]\right]$$

NORMAL MULTIVARIANTE

NORMAL UNIVARIANTE

La función de densidad de una variable aleatoria $\mathbf{x}$ Normal univariante con media $\mu_x = E[\mathbf{x}]$ y varianza $\sigma_x^2 = var[\mathbf{x}]$ es:

$$f_x(x) = (2\,\pi\sigma_x^2)^{-1/2}\exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\}$$

Notación:

$$\mathbf{x} \sim N(\mu_{\mathrm{x}}, \sigma_{\mathrm{x}}^2)$$

NORMAL BIVARIANTE

La Normal multivariante es una generalización a dos o más dimensiones de la Normal o Gaussiana univariante. En el caso bidimensional o bivariante tenemos un vector aleatorio formado por dos variables aleatorias $[\mathbf{x}, \mathbf{y}]$. Si el vector aleatorio fuera Normal multivariante, la función de densidad conjunta sería:

$$f_{\mathrm{xy}}(x,y) = \frac{1}{2\,\pi\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_{\mathrm{x}})^2}{\sigma_{\mathrm{x}}^2} - 2\rho\,\frac{(x-\mu_{\mathrm{x}})(y-\mu_{\mathrm{y}})}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} + \frac{(y-\mu_{\mathrm{y}})^2}{\sigma_{\mathrm{y}}^2}\right]\right\}$$

donde:

$$\mu_{\mathrm{x}} = E[\boldsymbol{x}], \quad \mu_{\mathrm{y}} = E[\mathbf{y}], \quad \sigma_{\mathbf{x}}^2 = var[\mathbf{x}], \quad \sigma_{\mathbf{y}}^2 = var[\mathbf{y}], \quad \rho = \frac{cov(\mathbf{x},\mathbf{y})}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}}$$

Si $\rho = 0$, entonces:

$$f_{\mathrm{xy}}(x,y) = \frac{1}{2\,\pi\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}}\exp\left\{-\frac{1}{2}\left[\frac{(x-\mu_{\mathrm{x}})^2}{\sigma_{\mathrm{x}}^2}+\frac{\left(y-\mu_{\mathrm{y}}\right)^2}{\sigma_{\mathrm{y}}^2}\right]\right\}$$

$$= \frac{1}{\sqrt{2\,\pi}\sigma_{\mathrm{x}}}\exp\left\{-\frac{1}{2}\frac{(x-\mu_{\mathrm{x}})^2}{\sigma_{\mathrm{x}}^2}\right\} * \frac{1}{\sqrt{2\,\pi}\sigma_{\mathrm{y}}}\exp\left\{-\frac{1}{2}\frac{\left(y-\mu_{\mathrm{y}}\right)^2}{\sigma_{\mathrm{y}}^2}\right\}$$
$$= f_{\mathrm{x}}(x) * f_{\mathrm{y}}(y)$$

Donde $f_{\mathrm{x}}(x)$, $f_{\mathrm{y}}(y)$ son dos funciones de densidad de Normales: $N(\mu_{\mathrm{x}}, \sigma_{\mathrm{x}}^2)$ and $N(\mu_{\mathrm{y}}, \sigma_{\mathrm{y}}^2)$, respectivamente.

Si el vector aleatorio $[\mathbf{x}, \mathbf{y}]$ es Normal bivariante y el coeficiente de correlación es cero, entonces las variables marginales $\mathbf{x}$ e $\mathbf{y}$ son Normales también, y son independientes.

En este caso el concepto de incorrelación equivale al de independencia.

Si el vector aleatorio $\mathbf{z} = [\mathbf{x}, \mathbf{y}]$ es Normal bivariante con vector de medias $\mu_z = \left(\mu_{\mathrm{x}}, \mu_{\mathrm{y}}\right)$, y matriz de covarianza $\Sigma_{\mathrm{z}} = \begin{bmatrix} \sigma_{\mathrm{x}}^2 & \rho\sigma_{\mathrm{x}}\sigma_{\mathrm{y}} \\ \rho\sigma_{\mathrm{x}}\sigma_{\mathrm{y}} & \sigma_{\mathrm{y}}^2 \end{bmatrix}$. Entonces su función de densidad puede expresarse en forma matricial como:

$$f_{\mathbf{z}} = \frac{1}{2\pi\,\det(\Sigma_z)^{1/2}}\exp\left\{-\frac{1}{2}(z-\mu_z)^t\Sigma_z^{-1}(z-\mu_z)\right\}$$

Si el vector aleatorio $\mathbf{z}$ tiene p variables aleatorias, es Normal multivariante con vector de medias μ_z, y matriz de covarianza Σ_z. Entonces su función de densidad puede expresarse en forma matricial como:

$$f_{\mathbf{z}} = \frac{1}{(2\,\pi)^{p/2}\,\det(\Sigma_z)^{1/2}}\exp\left\{-\frac{1}{2}(z-\mu_z)^t\Sigma_z^{-1}(z-\mu_z)\right\}$$

NORMAL ESTÁNDAR

Si el vector aleatorio con distribución Normal multivariante $\mathbf{z} = [\mathbf{z}_1, \ldots, \mathbf{z}_p]$ tiene vector de medias $\mu_{\mathrm{z}} = \mathbf{0}_p$ y matriz de covarianza $\Sigma_{\mathrm{z}} = I_p$

$\mathbf{0}_p$: es el vector de longitud p formado por ceros.

I_p: es la matriz identidad de tamaño $p \times p$.

Entonces $\mathbf{z}$ es Normal estándar:

$$\mathbf{z} \sim N_p(\mathbf{0}_p, I_p)$$

ESTANDARIZACIÓN

¿Cómo se relaciona la normal estándar con una distribución $N_p(\boldsymbol{\mu}_p, \Sigma_p)$?

A través de una transformación lineal:

Si $\mathbf{y} \sim N_p(\boldsymbol{\mu}_\mathrm{y}, \Sigma_\mathrm{y})$, y hacemos una transformación

$$\mathbf{z} = \Sigma_\mathrm{y}^{-\mathbf{1/2}}(\mathbf{y} - \boldsymbol{\mu}_y)$$

Entonces la transformación $\mathbf{z}$ es Normal estándar $\mathbf{z} \sim N_p(\mathbf{0}_p, I_p)$

¿Cómo podemos crear $N_p(\boldsymbol{\mu}_p, \Sigma_p)$ a partir de $N_p(\mathbf{0}_p, I_p)$?

Usando la transformación inversa:

Si $\mathbf{z}$ es Normal estándar $\mathbf{z} \sim N_p(\mathbf{0}_p, I_p)$:

$$\mathbf{y} = \Sigma_\mathrm{y}^{\mathbf{1/2}}\mathbf{z} + \boldsymbol{\mu}_y$$

Entonces $\mathbf{y} \sim N_p(\boldsymbol{\mu}_\mathrm{y}, \Sigma_\mathrm{y})$

TRANSFORMACIÓN LINEAL

Si $\mathbf{x} \sim N_p(\boldsymbol{\mu}_\mathrm{x}, \Sigma_\mathrm{x})$

A es una matriz de tamaño $q \times p$

$\mathbf{b}$ es un vector de tamaño $q \times 1$

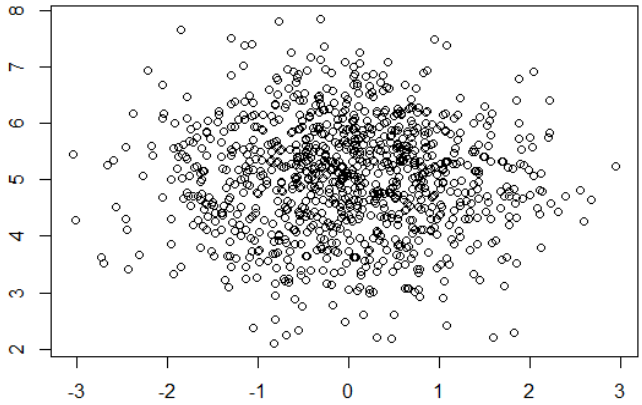
Entonces: la transformación lineal $\mathbf{y} = A\mathbf{x} + \mathbf{b}$ tiene distribución:

$$\mathbf{y} \sim N_q(A\boldsymbol{\mu}_\mathrm{x} + \mathbf{b},\ \ A\Sigma_\mathrm{x}A^t)$$

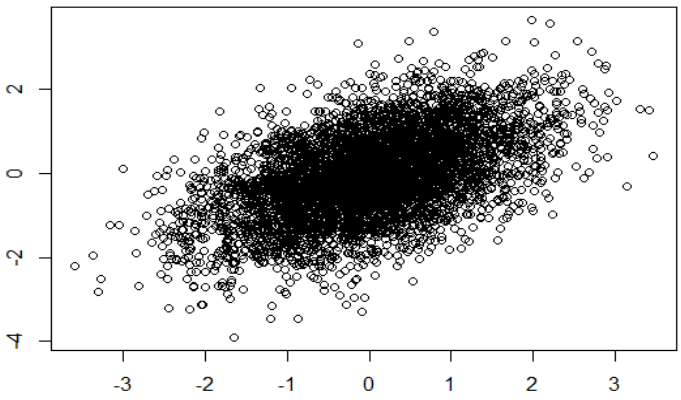
La transformación lineal de una Normal es Normal.

EJEMPLOS

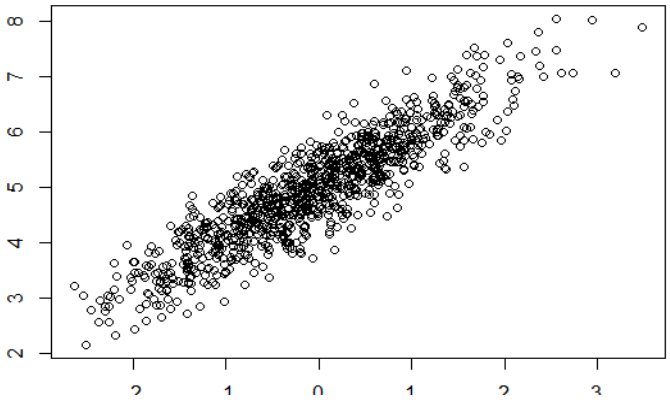
$$[\mathbf{x}, \mathbf{y}] \sim N_2\left(\mu = (0,5), \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$



$$[\mathbf{x}, \mathbf{y}] \sim N_2\left(\mu = (0,0), \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$$



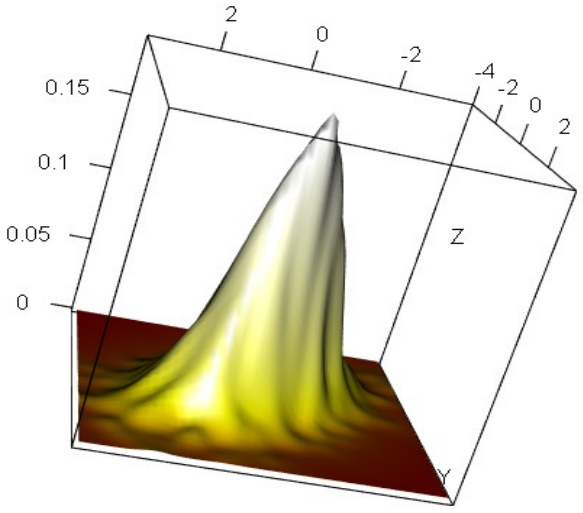
$$[\mathbf{x}, \mathbf{y}] \sim N_2\left(\mu = (0,5), \Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}\right)$$



CURVAS DE NIVEL

Imaginemos que tenemos una variable aleatoria bivariante con distribución Normal, cuya función de densidad se representa de la siguiente forma:

$$[\boldsymbol{x}, \boldsymbol{y}] \sim N\left(\mu = (0,0), \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$$

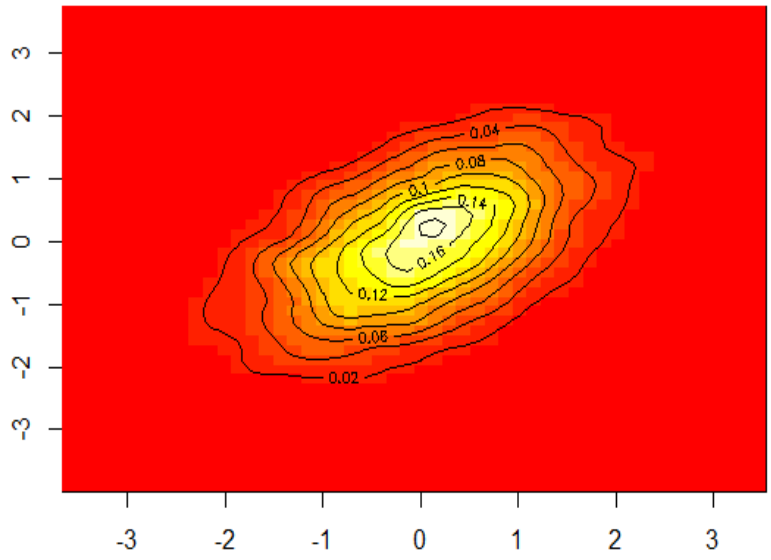


Si cortamos con un plano a la función de densidad quedaría una curva que delimita el borde exterior de la función a una determinada altura.

Las curvas de nivel o contornos son el resultado de cortar con planos paralelos a la función de densidad.

Mientras más altura tenga ese plano, más pequeño será el contorno porque estaremos más cerca del centro de agrupación de los datos.

En dimensión mayor a 2 tendremos hiperplanos.



Si el vector aleatorio $\mathbf{x} = [\mathbf{x_1}, \dots, \mathbf{x_p}]$ es Normal multivariante con vector de medias $\boldsymbol{\mu}_\mathbf{x}$, y matriz de covarianza $\Sigma_\mathbf{x}$. Entonces igualando su función de densidad a una contante queda:

$$f_\mathbf{x} = \frac{1}{(2\,\pi)^{p/2}\,\det(\Sigma_\mathbf{x})^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})^t \Sigma_\mathbf{x}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})\right\} = k$$

$$\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})^t \Sigma_\mathbf{x}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})\right\} = k\,(2\,\pi)^{p/2}\,\det(\Sigma_\mathbf{x})^{1/2}$$

$$(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})^t \Sigma_\mathbf{x}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x}) = -2\log\left\{k\,(2\,\pi)^{p/2}\,\det(\Sigma_\mathbf{x})^{1/2}\right\} = c$$

$$(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})^t\,\Sigma_\mathbf{x}^{-1}\,(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x}) = c$$

En el caso Gaussiano multivariante, la ecuación de los contornos viene dada por:

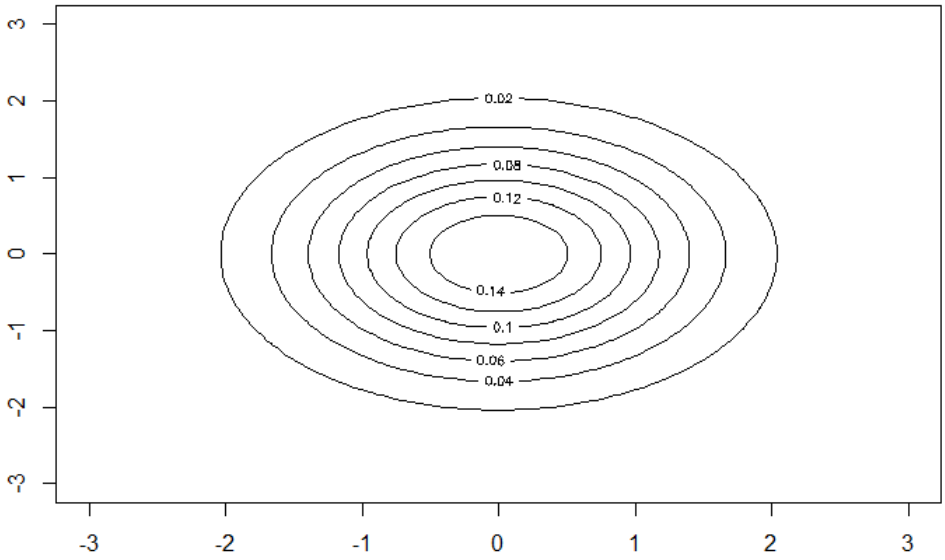
$$(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})^t\,\Sigma_\mathbf{x}^{-1}\,(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x}) = c$$

donde c es una constante.

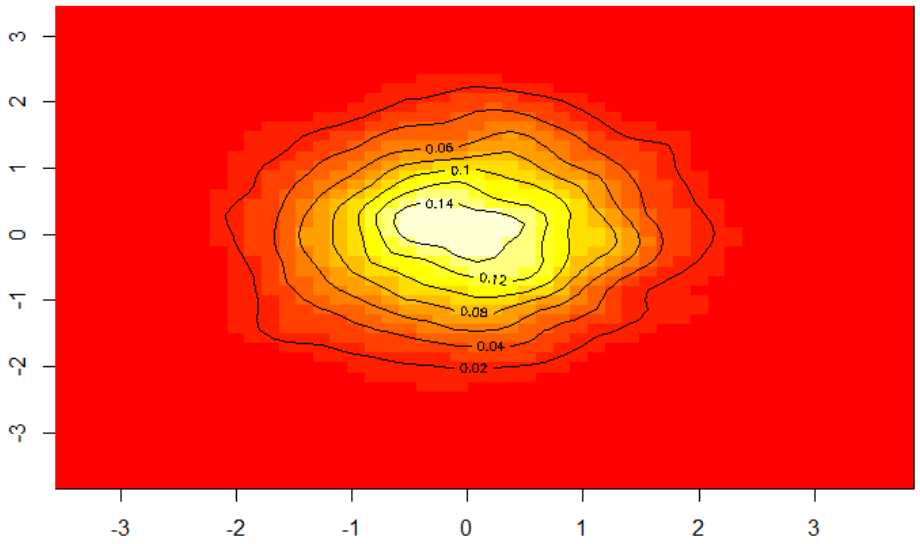
- Para diferentes valores de c se tienen diferentes contornos.
- En el caso bivariante, las curvas de nivel son elipses.

- En el caso de dimensión $p > 2$, las curvas de nivel son elipsoides.

Teóricamente quedarían elipses, si el contorno se estima con la densidad teórica:



Si el contorno se estima con la densidad estimada quedaría:

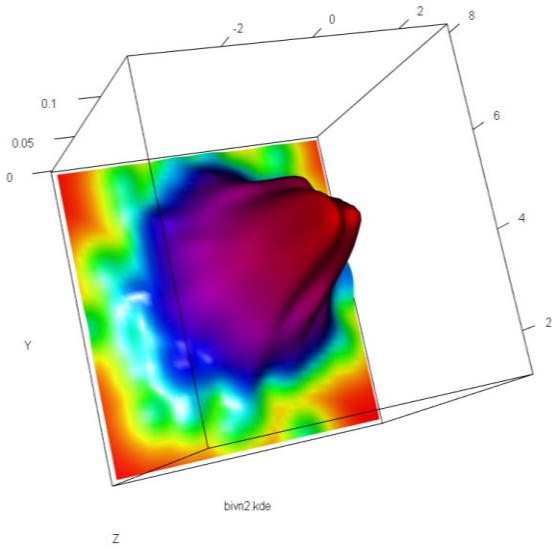


DENSIDAD KERNEL MULTIVARIANTE

Imaginemos que tenemos una variable aleatoria bivariante con distribución Normal, cuya función de densidad se representa de la siguiente forma:

$$[\mathbf{x}, \mathbf{y}] \sim N\left(\mu = (0,5), \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

Esta sería la llamada "densidad kernel" de esa variable bivariante:



ATÍPICOS

¿Cómo detectar atípicos en el espacio multivariante?

Tenemos el centro multidimensional de los datos. Necesitamos una generalización multivariada de la idea univariada de medir a cuántas desviaciones estándar está una observación alejada del centro. Esto se resuelve con el concepto de distancia, que es cero si los datos están en el centro y aumenta a medida que se aleja de ese centro. Veamos la distancia euclídea con más detalle.

DISTANCIA EUCLÍDEA

Consideremos $\mathbf{x} = [x_1, \dots, x_p]$ e $\mathbf{y} = [y_1, \dots, y_p]$ dos observaciones p −dimensionales de nuestros datos.

$$d_E(\mathbf{x}, \mathbf{y}) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + (y_3 - x_3)^2 + \cdots + (y_p - x_p)^2}$$

La distancia euclídea entre dos observaciones p −dimensionales $\mathbf{x}$ e $\mathbf{y}$, se puede calcular considerándolas como vectores.

$$d_E(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^t(\mathbf{x} - \mathbf{y})}$$

El teorema de Pitágoras asume que las componentes x_1 y x_2 tienen las mismas unidades de medida. No podemos calcular una distancia entre dos características con diferentes unidades de medida. En las aplicaciones de las Estadísticas Multivariadas, las personas, las marcas, los negocios, etc. se caracterizan por vectores de medición. Una persona puede caracterizarse por: su altura ($\mathbf{x_1}$), edad ($\mathbf{x_2}$), género ($\mathbf{x_3}$), nivel socioeconómico ($\mathbf{x_4}$), peso ($\mathbf{x_5}$), marca de jabón más utilizada ($\mathbf{x_6}$), etc. De esa manera, la persona k está representada en la muestra por un vector $\mathbf{x}_k$. Estas variables se miden en diferentes escalas (metros, años, escalas nominales, kilogramos, etc.

DISTANCIA DE MAHALANOBIS

¿Qué desventaja tiene la Distancia Euclídea?

En el espacio multivariante, las variables pueden estar correlacionadas, y es importante tener esto en cuenta. La distancia euclídea asume que no lo están. Una distancia que sí tiene esto en cuenta es la Distancia de Mahalanobis, muy usada en el caso multivariante.

Supongamos que tenemos n datos de una variable aleatoria multivariante

$$\mathbf{x} = \left[\mathbf{x}_1, \dots, \mathbf{x}_p\right]$$

Con media $\boldsymbol{\mu} = \left[\mu_1, \dots, \mu_p\right]$ y matriz de covarianza Σ.

La distancia de Mahalanobis (MD) de una observación multivariante p −dimensional

$$\mathbf{x}_i = \left[\mathbf{x}_{i1}, \dots, \mathbf{x}_{ip}\right]$$

Se define como:

$$MD(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})}$$

Si la matriz de covarianza es la matriz de identidad $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, la distancia Mahalanobis se reduce a la distancia euclídea.

En la práctica no conocemos el centro o la estructura de la distribución poblacional de la que obtuvimos los datos, por lo tanto, se usan los estimadores muestrales para estimar las distancias, la media de la muestra: $\bar{\mathbf{x}}$ y la matriz de covarianza muestral: S

$$MD(\mathbf{x}) = \sqrt{(\mathbf{x} - \bar{\mathbf{x}})^t \mathrm{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}})}$$

Con las distancias tenemos una medida de “lejanía" o "cercanía" para cada observación multivariante.

Si MD si es demasiado alta, eso significa que esta observación está muy lejos del centro de la distribución, por lo tanto, posiblemente sea un atípico.

¿CÓMO DETECTAR ATÍPICOS CON LA DISTANCIA DE MAHALANOBIS?

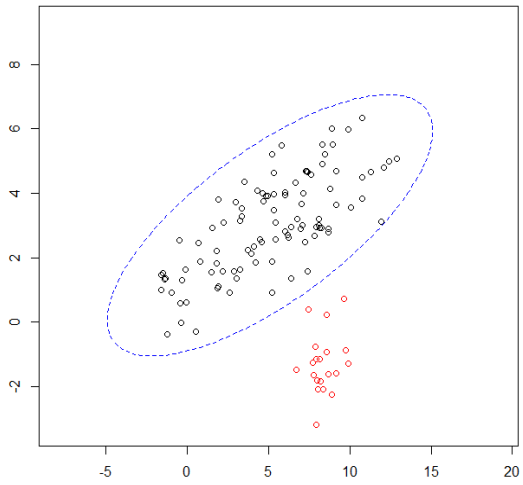
Primero necesitamos saber a partir de qué momento un valor de la distancia Mahalanobis podría ser considerado como un valor muy alto.
Se sabe que para datos normalmente distribuidos, la distribución de la distancia Mahalanobis al cuadrado, con los estimadores clásicos media muestral y matriz de covarianza muestral, es aproximadamente una chi-cuadrado con $\boldsymbol{p}$ grados de libertad, donde *p* es el número de variables que tenemos.

La notación es: χ^2_p

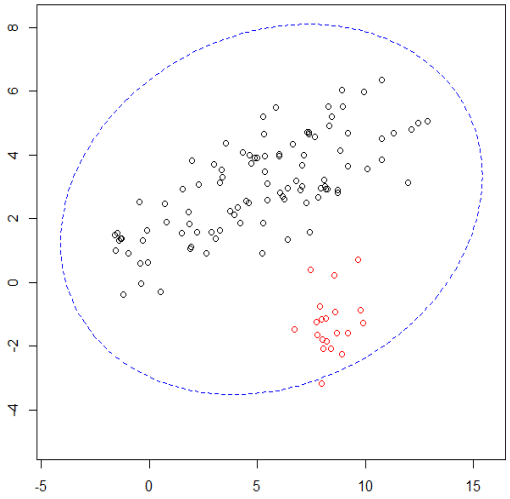
Por lo general, el valor de corte es el cuantil 97.5% de la distribución chi-cuadrado con *p* d.f. (degrees of freedom=grados de libertad):

$$\chi^2_{p;0.975}$$

Este valor de corte y los otros cuantiles de la chi-cuadrado se pueden observar en el caso bidimensional como elipses centradas en el centro bivariado de nuestros datos. Y elipsoides en el caso $p > 2$. Esas observaciones cuya MD al cuadrado es más alta que $\chi^2_{p;0.975}$ se etiquetan como valores atípicos.



Sin embargo, aquí hay un grave problema. La media muestral y la matriz de covarianza muestral son estimadores que se dejan influenciar por la presencia de los valores atípicos (observaciones rojas), por tanto, en la práctica cuando se calcula la MD, los valores se calculan incorrectamente porque dependen de los valores de los estimadores que se afectan con los atípicos, así que el método para detectarlos no logra identificarlos correctamente:

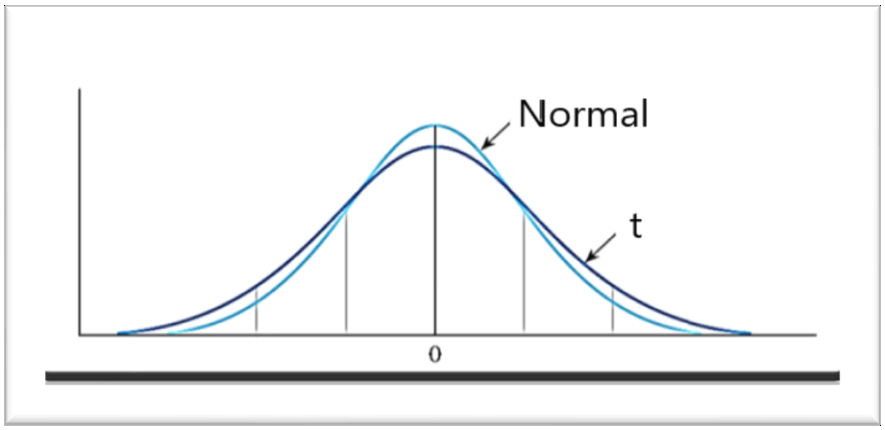


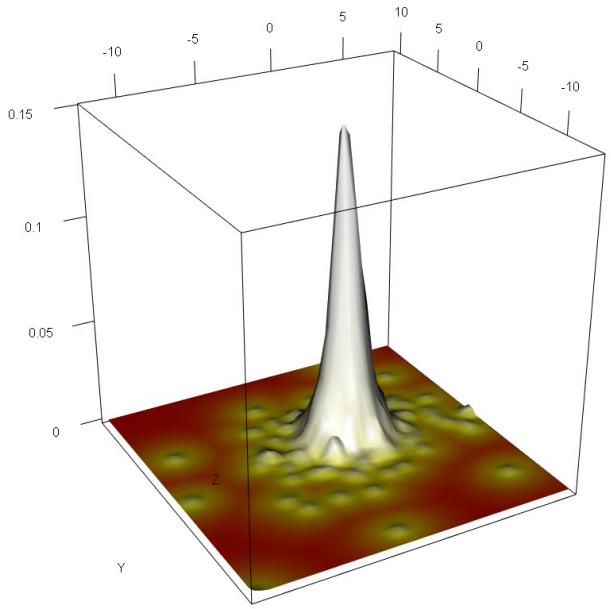
La solución es considerar estimadores robustos de localización y dispersión en la Distancia de Mahalanobis. Esto forma parte del **análisis robusto de datos**. Si quieres saber más sobre qué estimadores se pueden usar, cómo se puede definir una distancia de Mahalanobis robusta, cómo detectar atípicos eficazmente en el espacio univariante y multivariante, puedes ver mi Curso avanzado de atípicos y outliers en R y Matlab donde vemos toda la teoría, todos los métodos tanto en el espacio univariante, multivariante, como en regresión.

OTRAS DISTRIBUCIONES

T-STUDENT

La distribución t-student multivariante es una extensión de la t-student univariante para el caso $p > 2$. La t-student univariante es $\mathbf{x} \sim t_n$. Se obtiene de la Normal y no tiene relación con la realidad. Depende de un parámetro n llamado "grados de libertad". Su gráfico es simétrico y en forma de campana, parecido al de la Normal. Pero tiene mayor dispersión que la Normal. Su varianza tiende a 1 a medida que n aumenta. Se acerca a la Normal, a medida que n aumenta. Valor esperado y varianza: $E[\mathbf{x}] = 0$ *y* $var(\mathbf{x}) = \frac{n}{n-2}$.





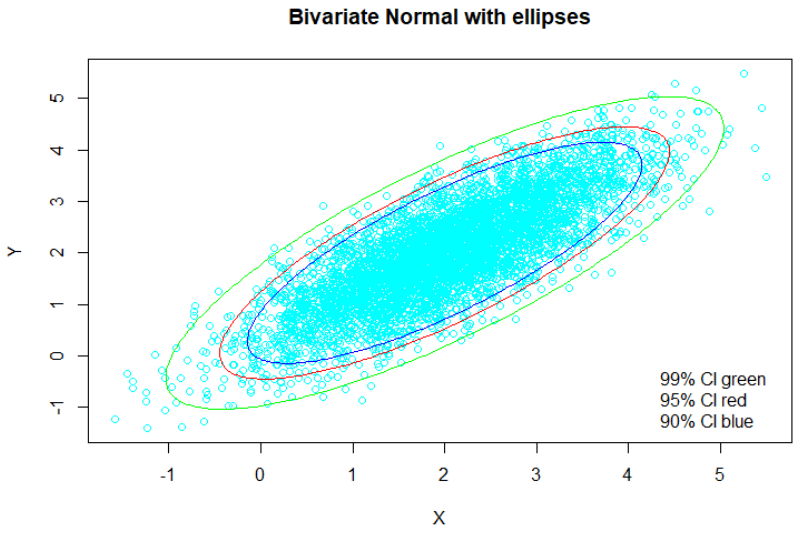
La t-student multivariante también pertenece a la familia de distribuciones elípticas. El valor de n también significa los "grados de libertad". La distribución t-student es de la familia de distribuciones de cola pesada, o heavy-tailed, porque tiene mayor densidad de probabilidad en la cola comparada con una Normal con igual vector de medias y matriz de covarianza. Hay más distribuciones de cola pesada, por ejemplo: mixturas de distribuciones.

DISTRIBUCIONES ELÍPTICAS Y ESFÉRICAS

Una variable aleatoria $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ sigue una distribución elíptica si su función de densidad solo depende de la variable a través de

$$(\mathbf{x} - \mathbf{m})^{\mathbf{t}} \mathrm{V}^{-1} (\mathbf{x} - \mathbf{m})$$

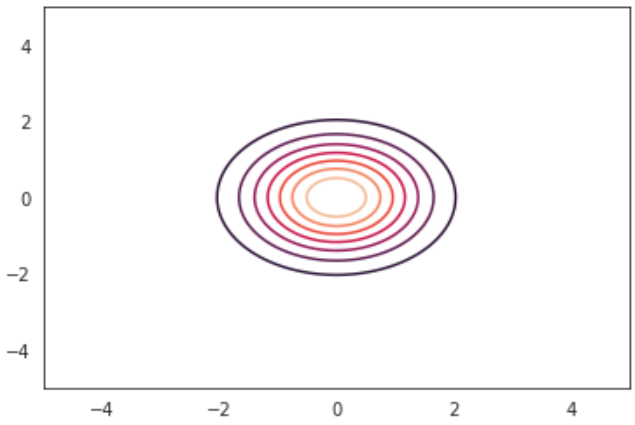
donde $\mathbf{m}$ es un vector de tamaño $p \times 1$ y V es una matriz (no necesariamente la media y la matriz de covarianza de $\mathbf{x}$). Las curvas de nivel de las distribuciones elípticas son elipsoides centrados en $\mathbf{m}$. En el caso bidimensional, serían elipses centradas en $\mathbf{m}$. La distribución Gaussiana o Normal multivariante es un caso de distribución elíptica: $N_p(\boldsymbol{\mu}_{\mathrm{x}}, \Sigma_{\mathrm{x}})$.



Una variable aleatoria $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ sigue una distribución esférica si su función de densidad solo depende de la variable a través de

$$\mathbf{x}^t\mathbf{x}$$

Las curvas de nivel de las distribuciones esféricas son esferas centrados en el origen. La distribución es invariante por rotaciones, es decir, si definimos $\mathbf{y} = C\mathbf{x}$, donde C es una matriz ortogonal, la densidad de la variable $\mathbf{y}$ es la misma que la de $\mathbf{x}$. En el caso bidimensional, las curvas de nivel serían circunferencias centradas en $(0,0)$. La distribución Normal estándar multivariante es un caso de distribución esférica: $N_p(\mathbf{0}_p, I_p)$.



PROPIEDADES DE LAS DISTRIBUCIONES ELÍPTICAS

- Tienen propiedades en común con la Normal.

- Las marginales y las condicionales son también elípticas.
- Las medias condicionales son una función lineal de las variables que la determinan.
- Sin embargo, la Normal es la única distribución en la familia que tiene la propiedad de que si la matriz de covarianza es diagonal (covarianzas cero) todas las variables que componen al vector aleatorio son independientes.

MIXTURA DE DISTRIBUCIONES

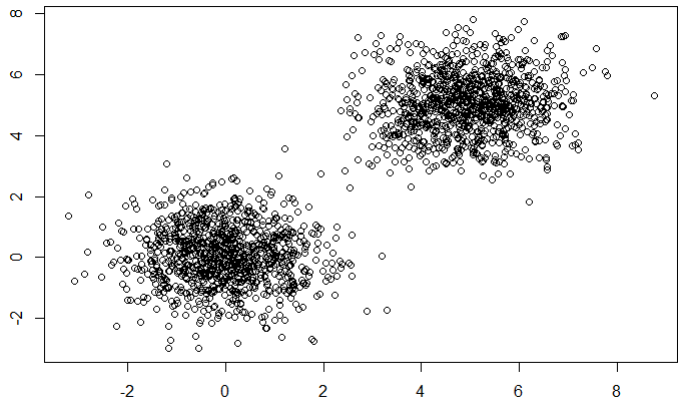
Modelado de mixtura de distribuciones es una técnica que consiste en modelar una distribución multivariante mediante una mixtura o suma ponderada de diferentes distribuciones. Una variable aleatoria multivariante $\mathbf{x} = [\mathbf{x_1}, \dots, \mathbf{x_p}]$ tiene una mixtura de distribuciones como distribución cuando:

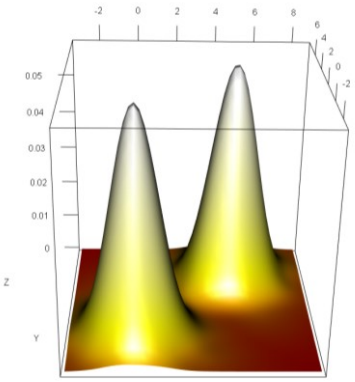
$$\mathbf{f}_x(x) = \sum_{g=1}^{G} \pi_g f_{\mathrm{x},g}(x)$$

- $\pi_1, \dots, \pi_G$ son pesos tales que $\sum_{g=1}^{G} \pi_g = 1$.
- $f_{\mathrm{x},1}, \dots, f_{\mathrm{x},G}$ son funciones de densidad multivariantes.

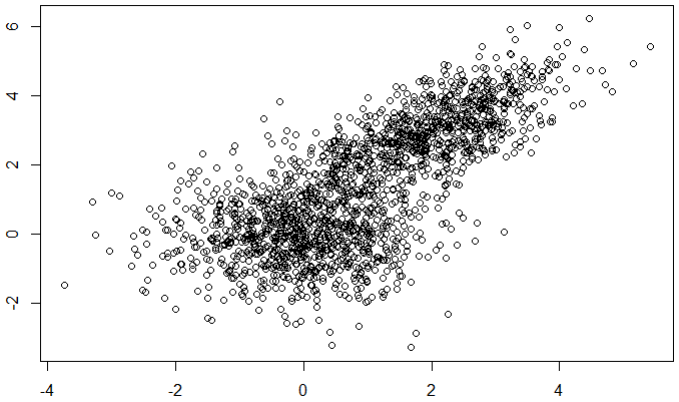
Puede interpretarse en términos de una población heterogénea. Asumiendo que la variable aleatoria multivariante $\mathbf{x}$ está definida en una población heterogénea que se puede subdividir en G grupos homogéneos. Entonces $\pi_1, \dots, \pi_G$ puede verse como la proporción de elementos de los grupos $1, \dots, G$, mientras que $f_{\mathrm{x},1}, \dots, f_{\mathrm{x},G}$ son las funciones de densidad asociadas a cada subgrupo de la población.

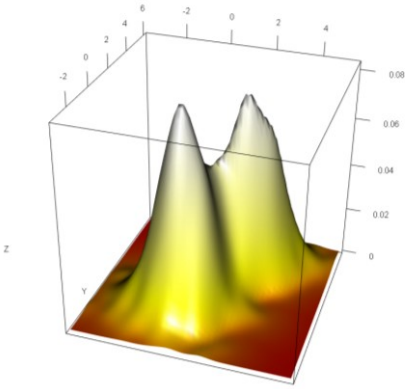
$$\mathbf{x} = [\mathbf{x_1}, \mathbf{x_2}] \sim 1000N(0_2, I_2) + 1000N(\mu = (5{,}5),\ I_2)$$

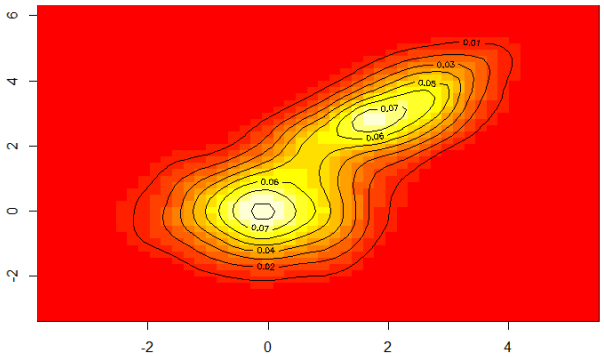




$$\mathbf{x} = [\mathbf{x_1}, \mathbf{x_2}] \sim 1000N(0_2, I_2) + 800N\left(\mu = (2,3),\ \Sigma = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}\right)$$







CÓPULAS

Si tenemos un vector aleatorio $[\mathbf{x}, \mathbf{y}]$ cuya función de distribución conjunta es F_{xy}, podemos marginalizar y obtener las funciones de distribución marginales de $\mathbf{x}$ e $\mathbf{y}$.

$$F_{xy} \longrightarrow F_x, F_y$$

En general, con las funciones de distribución marginales F_x, F_y no podemos obtener la función de distribución conjunta F_{xy} (excepto cuando se tiene independencia).

El concepto de cópula sirve para conectar funciones de densidad marginales con funciones de densidad conjunta. Por simplicidad vamos a centrarnos en el caso bidimensional $p = 2$. Una cópula bidimensional es una función $C: [0,1]^2 \rightarrow [0,1]$ con las siguientes propiedades:

1. Para cada $\mathrm{u}, \mathrm{v} \in [0,1]: C(u, 0) = C(0, v) = 0$
2. Para cada $\mathrm{u}, \mathrm{v} \in [0,1]: C(u, 1) = u,\ C(1, v) = v$
3. Para cada $(u_1, u_2), (v_1,\ v_2) \in [0,1] \times [0,1]$ con $u_1 \leq u_1$ *y* $v_1 \leq v_2$*:*

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$$

TEOREMA DE SKLAR

Consideremos el vector aleatorio bidimensional $[\mathbf{x}, \mathbf{y}]$. Sea F_{xy} la función de distribución conjunta con marginales F_{x} y F_{y}, existe una función cópula C_{xy} tal que:

$$F_{xy}(x, y) = C_{xy}\left(F_x(x), F_y(y)\right)$$

para todo $x, y \in \mathbb{R}^2$.

$F_{xy} \longleftarrow F_x, F_y$

La función cópula enlaza la distribución multivariante con las distribuciones marginales univariantes.

Si C_{xy} es una función cópula y F_{x} y F_{y} son funciones de distribución, entonces F_{xy} definida como

$$F_{\mathrm{xy}}(x, y) = C_{\mathrm{xy}}\left(F_{\mathrm{x}}(x), F_{\mathrm{y}}(y)\right)$$

es una función de distribución conjunta con marginales F_{x} y F_{y}.

Si $\boldsymbol{F}_{\mathbf{x}}$ y $\boldsymbol{F}_{\mathbf{y}}$ son continuas, entonces $\boldsymbol{C}_{\mathbf{xy}}$ es única.

INDEPENDENCIA

Si $\mathbf{x}$ e $\mathbf{y}$ son dos variables aleatorias con funciones de distribución F_{x} y F_{y}, y la función de distribución multivariante (conjunta) es F_{xy}. Entonces, $\mathbf{x}$ e $\mathbf{y}$ son independientes si y sólo si:

$$C_{\mathrm{xy}}(F_{\mathrm{x}}, F_{\mathrm{y}}) = F_{\mathrm{x}}F_{\mathrm{y}}$$

Esta función cópula se llama cópula de independencia.

CÓPULA GAUSSIANA

Consideremos F_z la función de distribución de una Normal univariante estándar.

Sea F_x la función de distribución conjunta de $\mathbf{x}$, un vector aleatorio p −dimensional, con vector de medias $\boldsymbol{\mu}_{\mathrm{x}}$ y matriz de correlaciones R_{x}.

Entonces la función:

$$C^{Gauss}_{\mathrm{x},R_{\mathrm{x}}}(u) = F_x\left(F_z^{-1}(u_1), \dots, F_z^{-1}(u_p)\right)$$

es la cópula Gaussiana p −dimensional con matriz de correlaciones R_{x}, donde $(u_1, \dots, u_p) \in [0,1]^p$.

Si $R_{\mathrm{x}} \neq I_p$ entonces la cópula permite generar dependencia simétrica.

INFERENCIA

En la práctica, los datos que vamos a tener conforman una muestra. La Inferencia Estadística se basa en obtener información sobre la población, a partir de esos datos muestrales. Usualmente no sabremos cuál es la distribución de nuestros datos. En el caso multivariante vamos a tener más de una variable, con lo cual, se dificulta más el proceso. El promedio, o la media, de una muestra no tiene por qué coincidir con la media poblacional.

Eso es porque es una estimación puntual. No sabemos cuán lejos o cerca está de la media poblacional.

¿Y si queremos tener más seguridad sobre el valor real del parámetro poblacional?

Podríamos usar estimadores muestrales con buenas propiedades que nos aseguren que esa **estimación puntual** está bastante cerca de ese valor poblacional.

Podríamos dar un rango de valores en el que, con una determinada "confianza" el valor real poblacional va a estar incluido en ese intervalo. Eso sería obtener un "**intervalo de confianza**" para el parámetro que estamos estimando.

Otra idea es **contrastar una hipótesis** en base a la información muestral que tenemos.

ESTADÍSTICOS MUESTRALES

Supongamos que $\mathbf{x}$ es una variable aleatoria multivariante:

$$\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_p]$$

De la cual tenemos una muestra:

$$\mathbf{x_{i\cdot}} = [\mathbf{x}_{i1}, \dots, \mathbf{x}_{ip}]$$

para $i = 1, \dots, n$ resumida en una matriz de datos.

Si no sabemos la distribución de $\mathbf{x}$, podemos usar alguna función de la muestra para obtener información sobre las propiedades de la distribución en la población.

Esas funciones de la muestra son los estadísticos muestrales:

- Los estadísticos van a ser también variables aleatorias con su propia distribución.
- Necesitamos su distribución para saber la relación entre el estadístico y su contraparte poblacional.

Supongamos que tenemos una muestra aleatoria de la variable aleatoria poblacional $\mathbf{x}$:

$$\mathbf{x_1}., \dots, \mathbf{x_n}.$$

Y supongamos que $E[\mathbf{x}] = \boldsymbol{\mu}_{\mathrm{x}}$ y $cov[\mathbf{x}] = \Sigma_{\mathrm{x}}$

Entonces el vector de media muestral $\bar{\mathbf{x}}$ y la matriz de covarianza muestral S_x verifican estas propiedades:

1. $E[\bar{\mathbf{x}}] = \boldsymbol{\mu}_{\mathrm{x}}$
2. $cov[\bar{\mathbf{x}}] = \frac{\mathbf{1}}{\boldsymbol{n}}\Sigma_{\mathrm{x}}$
3. $E[S_{\mathrm{x}}] = \Sigma_{\mathrm{x}}$

TEOREMA CENTRAL DEL LÍMITE

CASO UNIVARIANTE

El Teorema central del límite en el caso univariante es el siguiente:

Si la variable aleatoria $\mathbf{x} = [x_1, \dots, x_n]$ sigue una distribución Normal con media μ_{x} y varianza σ_{x}^2, es decir:

$$\mathbf{x} \sim Normal(\mu_{\mathrm{x}}, \sigma_{\mathrm{x}}^2)$$

Por la propiedad aditiva de la Normal ("la suma de Normales es Normal"), se tiene que la media muestral sigue una distribución Normal.

$$\bar{\mathbf{x}} \sim Normal(\mu_{\mathrm{x}}, \frac{1}{n}\sigma_{\mathrm{x}}^2)$$

El resultado de la acción conjunta de otras variables aleatorias:

$$\mathbf{y} = \mathbf{x_1} + \mathbf{x_2} + \cdots + \mathbf{x_n}$$

Cumpliéndose:

- Las $\mathbf{x}_i$ son independientes entre sí, $\forall i = 1, \dots, n$.
- Las $\mathbf{x}_i$ tienen media μ_i y varianza σ_i^2 finitas (no necesariamente iguales).

- n es suficientemente grande.

Entonces:

$$\mathbf{y} \sim \mathrm{N}\left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mu_{\mathrm{i}} \ , \ \sum_{\mathrm{i}=1}^{\mathrm{n}} \sigma_{\mathrm{i}}^{2}\right)$$

Cuando el resultado de un experimento es debido a un conjunto muy grande de causas independientes, que actúan sumando sus efectos, es esperable que el resultado siga una distribución Normal.

CASO PARTICULAR: LA MEDIA

Sea la variable aleatoria univariante $\mathbf{x} = [x_1, \dots, x_n]$, con media μ y varianza $\sigma^2 \neq 0$, cada elemento puede considerarse a su vez como una variable aleatoria, y la media de ellos también será una variable aleatoria:

$$\bar{\mathbf{x}} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Como se cumple que:

- Las $\boldsymbol{x_i}$ son independientes entre sí, $\forall \mathrm{i} = 1, \dots, \mathrm{n}$.
- Las $\boldsymbol{x_i}$ tienen media μ y varianza σ^2, las mismas que $\mathbf{x}$.
- Si n es suficientemente grande:

Entonces:

$$\bar{\mathbf{x}} \sim N\left(\mu \, , \frac{\sigma^2}{n}\right)$$

CASO MULTIVARIANTE

Dada una colección de vectores aleatorios $\mathbf{x_1}, \mathbf{x_2}, \ \dots, \ \mathbf{x_n}$ independientes e idénticamente distribuidos, con vector medio $\boldsymbol{\mu}$ y matriz de covarianza Σ, entonces el vector medio muestral $\bar{\mathbf{x}}$ se distribuye aproximadamente como una Normal multivariante para muestras suficientemente grandes.

$$\bar{\mathbf{x}} \sim \mathrm{N}\left(\boldsymbol{\mu},\ \frac{1}{\mathrm{n}}\Sigma\right)$$

En multivariante tenemos n vectores aleatorios $\mathbf{x_1}, \mathbf{x_2}, \ \ldots, \ \mathbf{x_n}$ cada uno de los cuales es p −dimensional, es decir, $\mathbf{x}_i \in \mathbb{R}^p$, $\forall i = 1, \ldots, n$:

$$\mathrm{x}_i = [\mathrm{x}_{i(1)}, \ldots, \mathrm{x}_{i(p)}]$$

Entonces el sumatorio de las x_i es:

$$\sum_{i=1}^{n} \mathrm{x}_i = [\mathrm{x}_{1(1)}, \ldots, \mathrm{x}_{1(p)}] + \cdots + [\mathrm{x}_{n(1)}, \ldots, \mathrm{x}_{n(p)}] = \left[\sum_{i=1}^{n} \mathrm{x}_{i(1)}, \ldots, \sum_{i=1}^{n} \mathrm{x}_{i(p)}\right]$$

Y la media sería:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_i = \left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_{i(1)}, \ldots, \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_{i(p)}\right] = [\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_p]$$

MÉTODO DE MÁXIMA VEROSIMILITUD

Si suponemos que conocemos la distribución de la variable aleatoria multivariante $\mathbf{x}$, entonces, el objetivo principal de la inferencia estadística es estimar los parámetros que sean desconocidos de esta distribución.

Entonces, sea el vector de parámetros $\theta = [\theta_1, \ldots, \theta_r]$ de una distribución determinada con función de densidad $f(\cdot\,|\theta)$.

El objetivo será estimar el vector θ a partir de la muestra i.i.d. de la que se disponga.

El método más importante para llevar a cabo esta tarea es el de máxima verosimilitud (en inglés: Maximum Likelihood Estimation MLE).

Supongamos que tenemos una variable aleatoria multivariante $\mathbf{x}$ y de ella obtenemos una muestra con elementos i.i.d. (independientes e idénticamente distribuidos):

$$\mathbf{x_1}., \ \mathbf{x_2}., \ldots, \mathbf{x_n}.$$

Entonces, por la propiedad de independencia y porque los elementos muestrales se distribuyen con la misma distribución, la función de densidad conjunta está dada por:

$$\boldsymbol{f}(\mathbf{x_1}., \mathbf{x_2}., \dots, \mathbf{x_n}.|\, \theta) = \prod_{i=1}^{n} f(\mathbf{x_i}.|\theta)$$

donde cada $\mathbf{x_i}. = (\mathbf{x_{i1}}, \dots, \mathbf{x_{ip}}),\ \forall i = 1, \dots, n.$

Note que en la función de densidad conjunta los datos $\mathbf{x}$ son conocidos, mientras que θ es desconocido

$$\boldsymbol{f}(\mathbf{x_1}., \mathbf{x_2}., \dots, \mathbf{x_n}.|\, \theta) = \prod_{i=1}^{n} f(\mathbf{x_i}.|\theta)$$

En el estudio de MLE se considera a θ una variable y a $\mathbf{x}$ fijo.

La función de verosimilitud es entonces:

$$l(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(\mathbf{x_i}.|\theta)$$

El estimador máximo verosímil de θ, denotado como $\hat{\theta}$, es el valor de θ que maximiza la función de verosimilitud $l(\theta|\mathbf{x})$:

$$\hat{\theta} = \arg\max_{\boldsymbol{\theta}} l(\theta|\mathbf{x})$$

En otras palabras, $\hat{\theta}$ es el valor de θ que maximiza la probabilidad de obtener la muestra bajo estudio.

Es equivalente, y comúnmente más sencillo, maximizar la función log-verosimilitud (log-likelihood) o función soporte:

$$L(\theta|\mathbf{x}) = \log l(\theta|\mathbf{x})$$

Entonces:

$$\hat{\theta} = \arg\max_{\boldsymbol{\theta}} l(\theta|\mathbf{x}) = \arg\max_{\boldsymbol{\theta}} L(\theta|\mathbf{x})$$

Usualmente, el proceso de maximización es demasiado complejo y no puede hacerse analíticamente.

En estos casos, se utilizan técnicas de optimización no lineales.

Sin embargo, veremos algunos ejemplos donde sí podemos hallar analíticamente el estimador MLE de una variable aleatoria multivariante.

DISTRIBUCIÓN ASINTÓTICA

El siguiente resultado nos dice la distribución asintótica del MLE que resulta ser Gaussiana:

Teorema:

Supongamos que tenemos una muestra i.i.d. $\mathbf{x_1}., \mathbf{x_2}., \dots, \mathbf{x_n}.$

Si $\hat{\theta}$ es el MLE de $\theta \in \mathbb{R}^r$, bajo algunas condiciones de regularidad, cuando $n \to \infty$:

$$\sqrt{n}(\hat{\theta} - \theta) \to_d N(\mathbf{0}_r, F^{-1})$$

Donde F denota la matriz de información de Fisher dada por:

$$F = -\frac{1}{n} E\left[\frac{\partial^2}{\partial\theta\, \partial\theta^t} L(\theta|\mathbf{x})\right]$$

Consecuencias del Teorema:

1. El MLE es asintóticamente insesgado
2. Eficiente (varianza mínima)
3. Se distribuye como una Normal
4. $\hat{\theta}$ es un estimador consistente de θ, muy buenas propiedades.

ESTIMADORES MLE DE UNA NORMAL

Sea $\mathbf{x_1}., \mathbf{x_2}., \dots, \mathbf{x_n}.$ una muestra aleatoria de la variable multivariante $\mathbf{x} \sim N(\boldsymbol{\mu}, \Sigma)$

Entonces, la función de densidad conjunta, que coincide con la verosimilitud, es:

$$l(\boldsymbol{\mu}, \Sigma|\mathbf{x}) = \boldsymbol{f}(\mathbf{x_1}., \mathbf{x_2}., \dots, \mathbf{x_n}.|\boldsymbol{\mu}, \Sigma)$$
$$= \prod_{i=1}^{n} \left\{ (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left(-\frac{(\mathbf{x}_i. - \boldsymbol{\mu})^t \Sigma_{\mathrm{x}}^{-1} (\mathbf{x}_i. - \boldsymbol{\mu})}{2} \right) \right\}$$

La función soporte, el logaritmo de la verosimilitud, es:

$$L(\boldsymbol{\mu}, \Sigma|\, \mathbf{x}) = \log l(\boldsymbol{\mu}, \Sigma|\mathbf{x})$$
$$= -\frac{np}{2} \log 2\pi - \frac{n}{2} \log|\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i. - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}_i. - \boldsymbol{\mu})$$

Tenemos que hallar un MLE para los parámetros desconocidos $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$.

Empezamos con $\boldsymbol{\mu}$, en la expresión de la función soporte sólo la última expresión de $\boldsymbol{\mu}$

$$L(\boldsymbol{\mu}, \Sigma|\, \mathbf{x}) = -\frac{np}{2} \log 2\pi - \frac{n}{2} \log|\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i. - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}_i. - \boldsymbol{\mu})$$

Aplicando el siguiente resultado:

$$\frac{\partial (w^t A w)}{\partial w} = 2Aw$$

si el vector w no depende A y A es una matriz simétrica.

Entonces, considerando $w = (\mathbf{x}_i. - \boldsymbol{\mu})$ y $A = \Sigma^{-1}$ y aplicando la propiedad:

$$\frac{\partial L}{\partial \boldsymbol{\mu}} = -\frac{1}{2} \sum_{i=1}^{n} 2\Sigma^{-1} (\mathbf{x}_i. - \boldsymbol{\mu})$$

Simplificando y cambiando el signo:

$$\frac{\partial L}{\partial \boldsymbol{\mu}} = -\frac{1}{2} \sum_{i=1}^{n} 2\Sigma^{-1} (\mathbf{x}_i. - \boldsymbol{\mu}) = \sum_{i=1}^{n} \Sigma^{-1} (\boldsymbol{\mu} - \mathbf{x}_i.)$$

Como el objetivo es maximizar, igualamos la derivada a cero, y como Σ^{-1} es definida positiva nunca va a ser igual a cero, así que queda:

$$0 = \sum_{i=1}^{n} (\boldsymbol{\mu} - \mathbf{x}_{i.}) = \sum_{i=1}^{n} (\boldsymbol{\mu}) - \sum_{i=1}^{n} (\mathbf{x}_{i.}) = n\boldsymbol{\mu} - \sum_{i=1}^{n} (\mathbf{x}_{i.})$$

Despejando $\boldsymbol{\mu}$:

$$n\boldsymbol{\mu} - \sum_{i=1}^{n} (\mathbf{x}_{i.}) = 0$$

$$\boldsymbol{\mu} = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} (\mathbf{x}_{i.}) = \bar{\mathbf{x}}$$

Entonces, el MLE del parámetro $\boldsymbol{\mu}$ es el vector media muestral:

$$\hat{\mu} = \bar{\mathbf{x}}$$

Ahora queremos hallar el MLE para $\boldsymbol{\Sigma}$.

Vamos a usar las siguientes propiedades:

1. La traza es invariante bajo permutaciones cíclicas de productos:

$$\mathrm{tr(ACB) = tr(CAB) = tr(BCA)}$$

2. Como $w^t Aw$ es un escalar (tamaño 1x1) podemos tomar su traza y obtener el mismo valor:

$$\mathrm{w^t Aw = tr(w^t Aw) = tr(ww^t A)}$$

3. La derivada del logaritmo del determinante es:

$$\frac{\partial \log|\mathrm{A}|}{\partial \mathrm{A}} = \mathrm{A^{-t}}$$

4. Y la última propiedad:

$$\frac{\partial \,\mathrm{tr(AX^{-1}B)}}{\partial \mathrm{X}} = -(\mathrm{X^{-1}BAX^{-1}})^{\mathrm{t}}$$

Recordemos la función soporte que tenemos que derivar con respecto a $\boldsymbol{\Sigma}$:

$$L(\boldsymbol{\mu}, \Sigma|\, \mathbf{x}) = -\frac{np}{2}\log 2\pi - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_{i\cdot} - \boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}_{i\cdot} - \boldsymbol{\mu})$$

El primer término es constante con respecto a Σ.

Vamos a llamarle C a lo que es constante, sacar factor común $-1/2$, y vamos a llamar al sumatorio $S = \sum_{i=1}^{n}(\mathbf{x}_{i\cdot} - \boldsymbol{\mu})(\mathbf{x}_{i\cdot} - \boldsymbol{\mu})^t$:

$$\begin{aligned} L(\boldsymbol{\mu}, \Sigma|\, \mathbf{x}) &= C - \frac{1}{2}\left(n\log|\Sigma| + \sum_{i=1}^{n}(\mathbf{x}_{i\cdot} - \boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}_{i\cdot} - \boldsymbol{\mu}) \right) \\ &= C - \frac{1}{2}\left(n\log|\Sigma| + \sum_{i=1}^{n} tr\left[(\mathbf{x}_{i\cdot} - \boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}_{i\cdot} - \boldsymbol{\mu})\right] \right) \\ &= C - \frac{1}{2}\left(n\log|\Sigma| + \sum_{i=1}^{n} tr\left[(\mathbf{x}_{i\cdot} - \boldsymbol{\mu})(\mathbf{x}_{i\cdot} - \boldsymbol{\mu})^t \Sigma^{-1}\right] \right) \\ &= C - \frac{1}{2}(n\log|\Sigma| + tr\,[S\Sigma^{-1}]) \end{aligned}$$

Es decir, la log-verosimilitud es:

$$L(\boldsymbol{\mu}, \Sigma|\, \mathbf{x}) = C - \frac{1}{2}(n\log|\Sigma| + tr\,[S\Sigma^{-1}])$$

El primer término es constante.

Para derivar el segundo término usaremos la propiedad 3, usando $\mathrm{A} = \Sigma$ y sabiendo que Σ es simétrica ($\Sigma^t = \Sigma$):

$$\frac{\partial \log|\Sigma|}{\partial \Sigma} = \Sigma^{-\mathrm{t}} = \Sigma^{-1}$$

Para derivar el tercer término usaremos la propiedad 4, con $A = S$ y $B = I$, y así obtenemos:

$$\frac{\partial\, \mathrm{tr}(\mathrm{S}\Sigma^{-1})}{\partial \Sigma} = -(\Sigma^{-1}\mathrm{S}\Sigma^{-1})^{\mathrm{t}} = -(\Sigma^{-1}\mathrm{S}\Sigma^{-1})$$

La última igualdad es porque S y Σ son simétricas.

Entonces, derivando e igualando a cero, queda:

$$\frac{\partial L}{\partial \Sigma} = -\frac{1}{2}\left(n\Sigma^{-1} - (\Sigma^{-1}\mathrm{S}\Sigma^{-1})\right) = 0$$

$$n\Sigma^{-1} - (\Sigma^{-1}\mathrm{S}\Sigma^{-1}) = 0$$

$$(n - \Sigma^{-1}\mathrm{S})\Sigma^{-1} = 0$$

Como Σ^{-1} es definida positiva:

$$(n - \Sigma^{-1}\mathrm{S}) = 0$$

Despejando Σ, y sabiendo que $\Sigma\Sigma^{-1} = I$:

$$(\Sigma n - \Sigma\Sigma^{-1}\mathrm{S}) = 0$$

$$\Sigma n = S$$

$$\Sigma = \frac{1}{n}S = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_{i\cdot} - \boldsymbol{\mu})(\mathbf{x}_{i\cdot} - \boldsymbol{\mu})^t$$

Entonces, sabiendo que el MLE para $\boldsymbol{\mu}$ es:

$$\widehat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$$

Obtenemos que el MLE para Σ es:

$$\hat{\Sigma} = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}(\mathbf{x}_{i\cdot} - \bar{\mathbf{x}})(\mathbf{x}_{i\cdot} - \bar{\mathbf{x}})^t$$

TEST DE HIPÓTESIS MULTIVARIANTES

Con frecuencia se desea comprobar si una muestra dada puede provenir de una distribución con ciertos parámetros conocidos.

> *Ejemplo:* En control de calidad se contrasta si el proceso está en estado de control, lo que supone contrastar si las muestras provienen de una población normal con ciertos valores de los parámetros.

También puede interesar comprobar si varias muestras multivariantes provienen o no de la misma población.

Ejemplo: Queremos comprobar si ciertos mercados son igualmente rentables o si varios medicamentos producen efectos similares.

Otro test muy útil es realizar un contraste para ver si la hipótesis de Normalidad no es rechazada por los datos observados, porque estos test de inferencia se basan en la hipótesis de Normalidad de los datos.

Es importante saber:

1. Todo lo que estamos asumiendo sobre nuestros datos.
2. La información muestral que nos proporcionan los datos.

MÉTODO DE RAZÓN DE VEROSIMILITUDES

Para realizar contrastes de hipótesis sobre parámetros vectoriales se suele aplicar el **método de razón de verosimilitudes**. Esta teoría proporciona pruebas estadísticas con buenas propiedades para tamaños muestrales grandes. Lo cual significa que usualmente se obtienen resultados buenos y fiables.

Vamos a asumir:

1. Que los datos provienen de una población con distribución Normal Multivariante.
2. Que esa distribución tiene un parámetro $\boldsymbol{\theta} \in \mathbb{R}^p$, es decir, es un parámetro vectorial p −dimensional.
3. Que $\boldsymbol{\theta}$ toma valores en un subconjunto Ω, donde Ω es un subconjunto de $\mathbb{R}^p$.

Queremos contrastar:

La hipótesis nula (H_0) sobre que $\boldsymbol{\theta}$ está contenido en una región Ω_0 del espacio paramétrico, frente a una hipótesis alternativa (H_1) de que no se encuentra allí:

$$H_0\colon \boldsymbol{\theta} \in \Omega_0$$

$$H_1\colon \boldsymbol{\theta} \in \Omega - \Omega_0$$

Ejemplo:

Consideremos que el vector aleatorio $p-$ dimensional $\mathbf{x}$ sigue una distribución Normal multivariante:

$$\mathbf{x} \sim N(\boldsymbol{\mu}_{\mathbf{x}}, \Sigma_{\mathbf{x}})$$

En este caso, supongamos que queremos contrastar algo sobre el vector de medias, es decir, $\boldsymbol{\theta} = \boldsymbol{\mu}_{\mathbf{x}}$

Queremos contrastar la hipótesis nula de que $\boldsymbol{\mu}_{\mathbf{x}}$ es igual a un vector con valores fijos $\boldsymbol{\mu}_0$, entonces el test sería:

$$H_0: \ \boldsymbol{\mu}_{\mathbf{x}} = \boldsymbol{\mu}_0$$

$$H_1: \ \boldsymbol{\mu}_{\mathbf{x}} \neq \boldsymbol{\mu}_0$$

Entonces, en este caso $\Omega_0 = \{\boldsymbol{\mu}_0\}$ y $\Omega \in \mathbb{R}^p$

Para comparar estas dos hipótesis, hay que comparar las probabilidades de obtener los datos bajo ambas hipótesis. Pero, calcular estas probabilidades requiere conocer el valor del vector de parámetros, que es desconocido. El método de razón de verosimilitudes resuelve este problema tomando el valor que hace más probable obtener la muestra observada y que es compatible con la hipótesis.

En concreto, la máxima probabilidad de obtener la muestra observada bajo H_0 se obtiene como sigue:

- Si Ω_0 determina un valor único para el vector de parámetros: $\Omega_0 = \{\boldsymbol{\theta}_0\}$, entonces se calcula la probabilidad de los datos suponiendo $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.
- Si Ω_0 permite muchos valores, se elige el valor del parámetro que haga máxima la probabilidad de obtener la muestra.
    - ✓ Como la probabilidad de la muestra observada es proporcional a la distribución conjunta de las observaciones, al sustituir en esta función los datos, lo que resulta es la función de verosimilitud.

- ✓ Calculando el máximo de esta función en Ω_0, se obtiene el máximo valor de la verosimilitud compatible con H_0, que representaremos por $f(H_0)$.

La máxima probabilidad de obtener la muestra observada bajo H_1 se calcula:

- ✓ Obteniendo el máximo absoluto de la función sobre todo el espacio paramétrico Ω.
- ✓ Estrictamente debería calcularse en el conjunto $\Omega - \Omega_0$ pero es más simple hacerlo sobre todo el espacio ya que en general se obtiene el mismo resultado.
- ✓ Particularizando la función de verosimilitud en su máximo, que corresponde al estimador MV de los parámetros, se obtiene una cantidad que representaremos como $f(H_1)$.

A continuación, compararemos $f(H_0)$ y $f(H_1)$.

Para eliminar las constantes y hacer la comparación invariante ante cambios de escala de las variables, construimos su cociente, que llamaremos razón de verosimilitudes (RV):

$$RV = \frac{f(H_0)}{f(H_1)}$$

Por construcción $RV \leq 1$

Rechazaremos H_0 cuando RV sea suficientemente pequeño.

La región de rechazo de H_0 se define por:

$$RV \leq \alpha$$

donde α se determinará imponiendo que el nivel de significación del test sea α.

El valor de α determina también el nivel de confianza con el que obtendremos los resultados.

Relación entre nivel de significación (α) y nivel de confianza (NC):

- ✓ El nivel de significación α es determinado a priori, y es el valor con el que se compara la razón de verosimilitudes para definir si es un

valor suficientemente pequeño como para poder rechazar la hipótesis nula H_0 con cierta confianza.

- ✓ Ese nivel de confianza es determinado a su vez por el nivel de significación α, ya que:

$$NC + \alpha = 1$$

- ✓ Valores usuales del nivel de significación con su correspondiente nivel de confianza:
  - $\alpha = 0.01 \quad NC = 0.99$
  - $\alpha = 0.05 \quad NC = 0.95$
  - $\alpha = 0.1 \quad NC = 90$

Para calcular el valor de α:

- ✓ Es necesario conocer la distribución de RV cuando H_0 es cierta, lo que suele ser difícil en la práctica.
- ✓ Sin embargo, cuando el tamaño muestral es grande, el doble de la diferencia de soportes entre la hipótesis alternativa y la nula, cuando H_0 es cierta, se distribuye asintóticamente como una χ^2_m con un número de grados de libertad m igual a la diferencia de dimensión entre los espacios Ω y Ω_0.
- ✓ Es decir:

$$\lambda = -2 \log RV = 2\big(L(H_1) - L(H_0)\big) \sim \chi^2_m$$

donde $L(H_i) = \log\ f(H_i)$, $i = 0{,}1$.

Es frecuente que la dimensión de Ω sea p y la dimensión de Ω_0 sea $p - r$, siendo r el número de restricciones lineales sobre el vector de parámetros.

Entonces, el número de grados de libertad de la diferencia de soportes, $\lambda \sim \chi^2_m$, es:

$$m = \dim(\Omega) - \dim(\Omega_0) = p - (p - r) = r$$

Es decir, es igual al número de restricciones lineales sobre el vector de parámetros, impuestas por H_0.

CONTRASTE PARA LA MEDIA

Consideremos la muestra $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \sim N_p(\boldsymbol{\mu}, \Sigma)$

Se desea realizar un contraste de hipótesis para la media:

$$H_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0$$

$$H_1\colon \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

En este caso no se asume nada sobre Σ.

Para construir un contraste de razón de verosimilitudes, tenemos que calcular el máximo de la función de verosimilitud bajo H_0 y bajo H_1.

La función soporte es:

$$L(\boldsymbol{\mu}, \Sigma\,|\mathbf{x}) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x} - \boldsymbol{\mu})^t\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

Es decir, tenemos que obtener los estimadores MLE de $\boldsymbol{\mu}$ y Σ bajo H_0 y bajo H_1.

$$\mathrm{L}(\mu, \Sigma\,|\mathrm{x}) = -\frac{\mathrm{n}}{2}\log|\Sigma| - \frac{1}{2}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x} - \mu)^{\mathrm{t}}\Sigma^{-1}(\mathrm{x} - \mu)$$

Ya sabemos que en el caso de la distribución Normal estos estimadores, bajo H_1, son el vector de media muestral $\bar{\mathbf{x}}$ y la matriz de covarianza muestral S.

Según la definición equivalente:

$$\mathrm{L}(\mu, \Sigma\,|\mathrm{x}) = -\frac{\mathrm{n}}{2}\log|\Sigma| - \frac{n}{2}tr\,\Sigma^{-1}\mathrm{S} - \frac{n}{2}(\bar{\mathbf{x}} - \mu)^t\Sigma^{-1}(\bar{\mathbf{x}} - \mu)$$

Y sustituyendo μ por $\bar{\mathbf{x}}$ y Σ por S, el soporte para H_1 es:

$$\mathrm{L}(H_1) = -\frac{n}{2}\log|S| - \frac{np}{2}$$

Bajo H_0 el estimador de $\boldsymbol{\mu}$ es directamente $\boldsymbol{\mu}_0$, con lo cual queda:

$$\mathrm{L}(H_0) = -\frac{n}{2}\log|S_0| - \frac{np}{2}$$

donde $S_0 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \mu_0)(\mathbf{x}_i - \mu_0)^t$

Como bajo H_1 el soporte era:

$$\mathrm{L}(H_1) = -\frac{n}{2}\log|S| - \frac{np}{2}$$

La diferencia de soportes será:

$$\lambda = 2\big(L(H_1) - L(H_0)\big) = 2\left(-\frac{n}{2}\log|S| - \frac{np}{2} + \frac{n}{2}\log|S_0| + \frac{np}{2}\right)$$
$$= \frac{2n}{2}(\log|S_0| - \log|S|) = n\log\frac{|S_0|}{|S|}$$

Rechazaremos H_0 cuando RV sea suficientemente pequeño.

En términos de λ, rechazamos H_0 cuando λ es suficientemente grande, es decir, cuando el soporte de los datos para H_1 es suficientemente mayor que para H_0.

$$\lambda = -2\log(RV) = 2\big(L(H_1) - L(H_0)\big) = n\log\frac{|S_0|}{|S|}$$

Esta condición equivale a que la varianza generalizada bajo H_0: $|S_0|$ sea significativamente mayor que bajo H_1: $|S|$.

La distribución de λ es una χ^2, con grados de libertad igual a la diferencia de las dimensiones del espacio en que se mueven los parámetros bajo ambas hipótesis.

La dimensión del espacio paramétrico bajo H_0 es (como $\boldsymbol{\mu}_0$ está fija):

$$p(p\ +\ 1)/2$$

el número de términos distintos en $\boldsymbol{\Sigma}$.

La dimensión del espacio paramétrico bajo H_1 es:

$$p + p(p+1)/2$$

La diferencia es p, que serán los grados de libertad de la χ^2

La distribución χ^2 es una distribución asintótica porque hemos asumido que n es grande.

Pero en este caso, podemos obtener la distribución exacta del ratio de verosimilitudes, no siendo necesaria la distribución asintótica.

Se sabe que el ratio

$$\frac{|S_0|}{|S|} = 1 + \frac{T^2}{n-1}$$

donde el estadístico $T^2 = (n-1)(\bar{\mathbf{x}} - \boldsymbol{\mu})^t S^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})$ sigue una distribución llamada T^2 de Hotelling con p y $n-1$ grados de libertad.

Existe además una relación entre el estadístico T^2 y la distribución F, la cual podemos usar para calcular los percentiles de T^2.

La diferencia de soportes λ es una función monótona de T^2:

$$\lambda = n \log \frac{|S_0|}{|S|} = n \log \left(1 + \frac{T^2}{n-1}\right)$$

Entonces, podemos utilizar directamente este estadístico en lugar de la razón de verosimilitudes.

Rechazaremos H_0 cuando T^2 sea suficientemente grande.

CONTRASTE PARA LA MATRIZ DE COVARIANZA

El contraste de razón de verosimilitudes también se aplica para contrastes de matrices de covarianzas, de forma similar a cuando lo hicimos para contraste de medias.

Vamos a ver tres contrastes en este caso:

1. En el primero se contrasta que la matriz tome un valor específico.
2. En el segundo, que la matriz es diagonal y las variables están incorreladas (contraste de independencia).
3. En el tercero, que las variables además tienen la misma varianza (contraste de esfericidad).

CONTRASTE DE UN VALOR ESPECÍFICO

Supongamos que tenemos una muestra de una Normal multivariante:

$$\mathbf{x} \sim N(\boldsymbol{\mu},\ \Sigma)$$

Se desea realizar el siguiente contraste de hipótesis:

$$H_0\colon\ \Sigma = \Sigma_0$$

$$H_1\colon\ \Sigma \neq \Sigma_0$$

Aquí no estaríamos asumiendo nada sobre $\boldsymbol{\mu}$

Para construir un contraste de razón de verosimilitudes, calcularemos el máximo de la función de soporte bajo H_0 y bajo H_1.

Utilizando la expresión del soporte:

$$\mathrm{L}(\mu, \Sigma\ |\mathrm{x}) = -\frac{\mathrm{n}}{2}\log|\Sigma| - \frac{n}{2}tr\ \Sigma^{-1}\mathrm{S} - \frac{n}{2}(\bar{\mathbf{x}} - \mu)^t\Sigma^{-1}(\bar{\mathbf{x}} - \mu)$$

Bajo H_1, los estimadores son $\bar{\mathbf{x}}$ y S y el soporte queda:

$$\mathrm{L}(H_1) = -\frac{n}{2}\log|S| - \frac{np}{2}$$

Bajo H_0, el valor Σ queda especificado por Σ_0 y $\boldsymbol{\mu}$ se estimará mediante $\bar{\mathbf{x}}$, con lo que:

$$\mathrm{L}(H_0) = -\frac{\mathrm{n}}{2}\log|\Sigma_0| - \frac{n}{2}tr\ \Sigma_0^{-1}\mathrm{S}$$

Bajo H_1, los estimadores son $\bar{\mathbf{x}}$ y S y el soporte queda:

$$\mathrm{L}(H_1) = -\frac{n}{2}\log|S| - \frac{np}{2}$$

La diferencia de soportes será:

$$\lambda = 2\big(\mathrm{L}(H_1) - \mathrm{L}(H_0)\big) = n\log\frac{|\Sigma_0|}{|S|} + n\ tr\ \Sigma_0^{-1}\mathrm{S} - np$$

La distribución del estadístico λ es una χ^2, con grados de libertad igual a la diferencia de las dimensiones del espacio en que se mueven los parámetros bajo ambas hipótesis que es $p(p\ +\ 1)/2$, el número de términos distintos

en Σ, porque bajo H_0 se estiman p y bajo H_1 $p + p(p\ +\ 1)/2$, con lo cual la diferencia es $p(p\ +\ 1)/2$.

Si en particular se quiera contrastar con la matriz identidad $\Sigma_0 = I$, quedaría:

$$\lambda = -n \log|S| + n\ tr\ \mathrm{S} - np$$

CONTRASTE DE INDEPENDENCIA

Otro contraste de interés es el de independencia, donde suponemos que la matriz Σ_0 es diagonal.

$$H_0\colon \Sigma = \Sigma_0\ diagonal$$

$$H_1\colon \Sigma \neq \Sigma_0\ diagonal$$

En este caso, la estimación máximo verosímil de Σ_0 es $\hat{\Sigma}_0 = diag(S)$ que es una matriz diagonal con términos s_{ii} iguales a los de la matriz S.

El estadístico λ se reduce a:

$$\lambda = n \log \frac{\prod s_{ii}}{|S|} + n\ tr\hat{\Sigma}_0^{-1} S - np$$

Podemos reducir la expresión:

$$tr\hat{\Sigma}_0^{-1} S = tr\hat{\Sigma}_0^{-1/2} S \hat{\Sigma}_0^{-1/2} = trR = p$$

donde $\mathrm{R} = \hat{\Sigma}_0^{-1/2} \mathrm{S}\ \hat{\Sigma}_0^{-1/2}$

Así, el estadístico se reduce a:

$$\lambda = -n \log |R|$$

El estadístico suele escribirse en función a los valores propios λ_i de R, lo que da una forma equivalente:

$$\lambda = -n \sum_{i=1}^{p} \log \lambda_i$$

Su distribución asintótica será una χ^2, con grados de libertad igual a:

$$p + \frac{p(p\ +\ 1)}{2} -\ p\ - p = p(p\ -\ 1)/2.$$

CONTRASTE DE ESFERICIDAD

Un caso particular importante del contraste anterior es suponer que todas las variables tienen la misma varianza y están incorreladas. En este caso no ganamos nada por analizarlas conjuntamente, ya que no hay información común. Este contraste equivale a suponer que la matriz Σ_0 es escalar, es decir $\Sigma_0 = \sigma^2 I$. Se denomina de esfericidad, ya que la distribución de las variables tiene curvas de nivel que son esferas: hay una total simetría en todas las direcciones en el espacio.

El contraste es:

$$H_0\colon \Sigma = \sigma^2 I$$

$$H_1\colon \Sigma \neq \sigma^2 I$$

Sustituyendo $\Sigma_0 = \sigma^2 I$ en la función soporte bajo H_0, y sabiendo que:

1. El determinante de una matriz diagonal es el producto de los elementos de su diagonal: $|\sigma^2 I| = \sigma^{2p}$
2. $trI^{-1} = trI = p$

$$\mathrm{L}(H_0) = -\frac{\mathrm{n}}{2}\log|\Sigma_0| - \frac{n}{2} tr\ \Sigma_0^{-1}\mathrm{S} = -\frac{np}{2}\log\sigma^2 - \frac{n}{2\sigma^2} trS$$

El estimador máximo verosímil para σ^2 sería:

$$\hat{\sigma}^2 = trS/p$$

Es decir, el promedio de las varianzas.

Bajo H_1 la función soporte es la misma que en el caso anterior,

$$\mathrm{L}(H_1) = -\frac{n}{2}\log|S| - \frac{np}{2}$$

Y la diferencia de soportes queda:

$$\lambda = 2\big(\mathrm{L}(H_1) - \mathrm{L}(H_0)\big) = 2\left(-\frac{n}{2}\log|S| - \frac{np}{2} + \frac{np}{2}\log\sigma^2 + \frac{n}{2\sigma^2}trS\right)$$
$$= n\log\sigma^{2p} - n\log|S| + \frac{n}{\sigma^2}trS - np$$
$$= n\log\frac{\sigma^{2p}}{|S|} + \frac{n}{\sigma^2}trS - np$$

Sustituyendo $\hat{\sigma}^2 = trS/p$ el contraste se reduce a:

$$\lambda = n\log\frac{\sigma^{2p}}{|S|} + \frac{n}{\sigma^2}trS - np = np\log\frac{\hat{\sigma}^2}{|S|} + \frac{np}{trS}trS - np$$
$$= np(\log\hat{\sigma}^2 - \log|S|)$$

Que tiene una distribución asintótica χ^2 con grados de libertad:

$$p + \frac{p(p\ +\ 1)}{2} -\ p\ - 1 = p(p\ +\ 1)/2\ -\ 1\ =\ (p\ +\ 2)(p\ -\ 1)/2$$

ANÁLISIS DE VARIANZA MULTIVARIANTE

CONTRASTE DE IGUALDAD DE VARIAS MEDIAS

Supongamos que tenemos una muestra de tamaño n de una variable aleatoria p −dimensional $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_\mathrm{p}]$

Supongamos que podemos agrupar los datos en G grupos con $n_1, \dots, n_G$ observaciones cada uno.

Se tiene que cumplir $n_1 + \cdots + n_G = n$

Vamos a asumir también Normalidad en cada grupo, y además que la matriz de covarianza es la misma para todos los grupos: $\boldsymbol{\Sigma}$.

Nos interesa contrastar si las medias de los grupos son iguales o no:

$$H_0\colon\ \mu_1, \dots, \mu_G = \mu$$

$$H_1\colon\ no\ todos\ los\ \mu_g\ son\ iguales$$

Este problema también se conoce como análisis de varianza multivariante.

La función de verosimilitud bajo H_0 de una muestra Normal homogénea ya la hemos calculado y sabemos que alcanza su máximo en $\widehat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ y $\hat{\Sigma} = S$.

Sustituyendo estas estimaciones en la función soporte bajo H_0:

$$L(H_0) = -\frac{\mathrm{n}}{2}\log|S| - \frac{np}{2}$$

Bajo H_1 la muestra es heterogénea, y las n observaciones se subdividen en grupos.

La función de verosimilitud bajo H_1 será:

$$f(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_G, \Sigma|\mathbf{x}) = \Sigma^{-\mathrm{n}/2}(2\pi)^{-np/2}\exp\left\{-\frac{1}{2}\sum_{g=1}^{G}\sum_{h=1}^{n_g}(\mathbf{x}_{hg} - \boldsymbol{\mu}_g)^t\Sigma^{-1}(\mathbf{x}_{hg} - \boldsymbol{\mu}_g)\right\}$$

donde $\mathbf{x}_{hg}$ es el vector de variables del grupo g, y $\boldsymbol{\mu}_g$ su media.

En este caso la estimación de la media de cada grupo g será la media muestral en dicho grupo:

$$\widehat{\boldsymbol{\mu}}_g = \bar{\mathbf{x}}_g$$

La estimación de la matriz de covarianzas común se estima por:

$$\hat{\Sigma}_g = S_w = \frac{1}{n}W$$

En este caso W es la matriz de suma de cuadrados dentro de los grupos:

$$W = \sum_{g=1}^{G}\sum_{h=1}^{n_g}(\mathbf{x}_{hg} - \bar{\mathbf{x}}_g)(\mathbf{x}_{hg} - \bar{\mathbf{x}}_g)^t$$

Sustituyendo esto en la función soporte tendremos que:

$$\mathrm{L}(H_1) = -\frac{n}{2}\log|S_w| - \frac{np}{2}$$

La diferencia de soportes será:

$$\lambda = n\log\frac{|S|}{|S_w|}$$

Rechazaremos H_0 cuando esta diferencia λ sea suficientemente grande.

Es decir, cuando la variabilidad suponiendo H_0 cierta, medida por $|S|$, sea mucho mayor que la variabilidad cuando permitimos que las medias de los grupos sean distintas, medida por $|S_w|$.

Su distribución es, asintóticamente, una χ^2_m donde los grados de libertad m, se obtienen por la diferencia entre ambos espacios paramétricos.

H_0 determina una región Ω_0 donde hay que estimar los p componentes del vector de medias común y la matriz de covarianzas, en total $p + p(p+1)/2$ parámetros.

Bajo la hipótesis H_1 hay que estimar G vectores de medias más la matriz de covarianzas lo que supone $Gp + p(p+1)/2$ parámetros.

La diferencia es:

$$m = \dim(\Omega) - \dim(\Omega_0) = p(G-1)$$

Entonces, λ sigue una distribución asintótica:

$$\lambda = n\log\frac{|S|}{|S_w|} \sim \chi^2_{p(G-1)}$$

Para muestras pequeñas puede mejorarse, si expresamos el estadístico λ como:

$$\lambda = h\log\frac{|S|}{|S_w|}$$

donde $h = (n-1) - (p+G)/2$

Este contraste es la generalización multivariante del análisis de varianza y puede deducirse alternativamente como sigue.

Llamemos variabilidad total de los datos a:

$$T = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^t$$

T mide las desviaciones respecto a una media común.

Vamos a descomponer la matriz T como suma de dos matrices:

$$T = W + B$$

La primera, W, es la matriz de las desviaciones respecto a las medias de cada grupo:

$$W = \sum_{g=1}^{G} \sum_{h=1}^{n_g} (\mathbf{x}_{hg} - \bar{\mathbf{x}}_g)(\mathbf{x}_{hg} - \bar{\mathbf{x}}_g)^t$$

La segunda, B, medirá la variabilidad explicada por las diferencias entre las medias, es como una sumas de cuadrados entre grupos:

$$B = \sum_{g=1}^{G} n_g (\bar{\mathbf{x}}_g - \bar{\mathbf{x}})(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})^t$$

Esta descomposición generaliza al caso vectorial la descomposición clásica de análisis de la varianza.

Para obtenerla lo que se ha hecho es sumar y restar las medias de grupo en la expresión de T:

$$T = \sum_{g=1}^{G} \sum_{h=1}^{n_g} (\mathbf{x}_{hg} - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \bar{\mathbf{x}}_g)(\mathbf{x}_{hg} - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \bar{\mathbf{x}}_g)^t$$

Desarrollando se comprueba que el doble producto se anula y resulta:

$$T = W + B$$

La descomposición anterior puede expresarse como:

$$T \text{ (variabilidad total)} = W \text{ (variabilidad residual)} + B \text{ (variabilidad explicada)}$$

Para hacer un contraste de medias iguales, podemos comparar el tamaño de las matrices. Por ejemplo, si B es grande, como T es fija, entonces W es pequeña.

La medida de tamaño adecuada es el determinante.

Podemos basar el contraste en el cociente $|T|/|W|$.

Para tamaños moderados el contraste es similar al de la razón de verosimilitudes:

$$\lambda_0 = h\ \log\frac{|T|}{|W|} = h\log\frac{|W+B|}{|W|} = h\ \log\ |I + W^{-1}B| = h\sum \log(1+\lambda_i)$$

donde λ_i son los vectores propios de la matriz $W^{-1}B$.

ANÁLISIS DE COMPONENTES PRINCIPALES

Un problema muy importante en Estadística Multivariante es la maldición de la dimensionalidad: si el ratio n/p no es suficientemente grande, algunos problemas no se pueden abordar tan fácilmente.

Que el ratio n/p sea grande significa que n tiene que ser mucho mayor que p: $n >> p$.

Por ejemplo, supongamos que tenemos n datos de una Normal p −dimensional:

$$\mathbf{x} \sim N(\mu, \Sigma)$$

En este caso, el número de parámetros a estimar es $p\ +\ p(p+1)/2$

Si $p = 5$ o $p = 10$ hay 20 o 65 parámetros respectivamente.

Mientras más alto sea p, mucho mayor tiene que ser el número de observaciones para poder obtener estimaciones fiables.

Hay varias técnicas de reducción de la dimensión que tratan de contestar la misma pregunta:

¿Es posible describir con precisión los valores de las p variables mediante un número menor de variables $r < p$?

La respuesta es sí, y con ello se habrá reducido la dimensión del problema a costa de una pequeña pérdida de información.

Vamos a ver dos técnicas:

1. Análisis de Componentes Principales (PCA)
2. Análisis Factorial (FA)

Dadas n observaciones de p variables, se analiza si es posible representar adecuadamente esta información con un número menor de variables construidas como combinaciones lineales de las originales.

Por ejemplo, con variables con alta dependencia es frecuente que un pequeño número de nuevas variables (menos del 20% de las originales) expliquen la mayor parte (más del 80%) de la variabilidad original.

Su utilidad es doble:

1. Permite representar óptimamente en un espacio de dimensión pequeña, observaciones de un espacio general p −dimensional. En este sentido componentes principales es el primer paso para identificar posibles variables latentes no observadas, que están generando la variabilidad de los datos.
2. Permite transformar las variables originales, en general correladas, en nuevas variables incorreladas, facilitando la interpretación de los datos.

El problema que se desea resolver es cómo encontrar un espacio de dimensión más reducida que represente adecuadamente los datos.

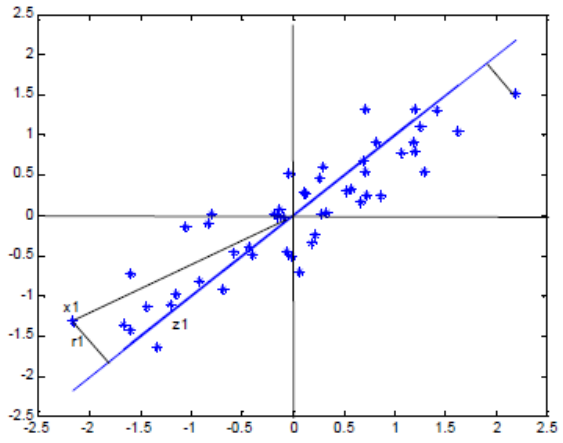
El problema puede abordarse desde tres perspectivas equivalentes:

1. Enfoque descriptivo
2. Enfoque estadístico
3. Enfoque geométrico

ENFOQUE DESCRIPTIVO

Se desea encontrar un subespacio de dimensión menor que p, tal que, al proyectar los puntos sobre él, estos conserven su estructura con la menor distorsión posible. Veamos cómo convertir esta noción intuitiva en un criterio matemático operativo. Consideremos primero el caso de dos dimensiones $(p = 2)$ y un subespacio de dimensión uno, una recta. Se desea que las proyecciones de los puntos bidimensionales sobre esta recta mantengan, lo mejor posible, sus posiciones relativas.

La figura muestra los puntos en un diagrama de dispersión, y una recta que. Esta recta, intuitivamente, proporciona un buen resumen de los datos, ya que las proyecciones de los puntos sobre ella indican aproximadamente la situación de los puntos en el plano. La representación es buena porque la recta pasa cerca de todos los puntos y estos se deforman poco al proyectarlos.



Al proyectar cada punto sobre la recta se forma un triángulo rectángulo. La hipotenusa es la distancia del punto al origen $(\mathbf{x}_i^t \mathbf{x}_i)^{1/2}$. Los catetos son la proyección del punto sobre la recta (z_i) y la distancia entre el punto y su proyección (r_i). Por el Teorema de Pitágoras, podemos escribir:

$$\mathbf{x}_i^t \mathbf{x}_i = z_i^2 + r_i^2$$

Sumando esta expresión para todos los puntos, se obtiene:

$$\sum_{i=1}^{n} \mathbf{x}_i^t \mathbf{x}_i = \sum_{i=1}^{n} z_i^2 + \sum_{i=1}^{n} r_i^2$$

Minimizar $\sum_{i=1}^{n} r_i^2$, la suma de las distancias a la recta de todos los puntos es equivalente a maximizar $\sum_{i=1}^{n} z_i^2$, la suma al cuadrado de los valores de las proyecciones.

Como las proyecciones z_i son variables de media cero, maximizar la suma de sus cuadrados equivale a maximizar su varianza. Este resultado es intuitivo: la recta parece adecuada porque conserva lo más posible la variabilidad original de los puntos. Si consideramos una dirección de proyección perpendicular a la de la recta en esta figura, los puntos tendrían muy poca variabilidad y perderíamos la información sobre sus distancias en el espacio.

ENFOQUE ESTADÍSTICO

Representar puntos $p-$ dimensionales con la mínima pérdida de información en un espacio de dimensión uno es equivalente a sustituir las p variables originales por una nueva variable, $\mathbf{z}_1$, que resuma óptimamente la información. Esto supone que la nueva variable debe tener globalmente máxima correlación con las originales o, en otros términos, debe permitir prever las variables originales con la máxima precisión. Esto no será posible si la nueva variable toma un valor semejante en todos los elementos, es decir, usaremos la variable de máxima variabilidad.

En la figura anteriormente vista, la recta no es la línea de regresión de ninguna de las variables con respecto a la otra, que se obtienen minimizando las distancias verticales u horizontales, sino que al minimizar las distancias ortogonales o de proyección se encuentra entre ambas rectas de regresión. Este enfoque puede extenderse para obtener el mejor subespacio resumen de los datos de dimensión 2. Para ello calcularemos el plano que mejor aproxima a los puntos. Estadísticamente esto equivale a encontrar una segunda variable $\boldsymbol{z}_2$, incorrelada con la anterior, y que tenga varianza

máxima. En general, la componente $\mathbf{z}_r$ $(r < p)$ tendrá varianza máxima entre todas las combinaciones lineales de las p variables originales, con la condición de estar incorrelada con las $\mathbf{z}_1, \ldots, \mathbf{z}_{r-1}$ previamente obtenidas.

ENFOQUE GEOMÉTRICO

Si consideramos la nube de puntos de la figura vemos que los puntos se sitúan siguiendo una elipse y podemos describir su orientación dando la dirección del eje mayor de la elipse y la posición de los punto por su proyección sobre esta dirección. Puede demostrarse que este eje es la recta que minimiza las distancias ortogonales y volvemos al problema que ya hemos resuelto. En mayores dimensiones tendremos elipsoides y la mejor aproximación a los datos es la proporcionada por el eje mayor del elipsoide. Considerar los ejes del elipsoide como nuevas variables originales supone pasar de variables correladas a variables ortogonales.

CÁLCULO DE LOS COMPONENTES

El primer componente principal será la combinación lineal de las variables originales que tenga varianza máxima.

Los valores de este primer componente en los n individuos se representarán por un vector $\mathbf{z}_1$, dado por:

$$\mathbf{z}_1 = \mathbf{x}\mathbf{a}_1$$

Estamos suponiendo sin pérdida de generalidad que ya $\mathbf{x}$ tiene los datos centrados.

Como las variables originales tienen media cero también $\mathbf{z}_1$ tendrá media nula.

Su varianza será:

$$var(\mathbf{z}_1) = \frac{1}{n}\mathbf{z}_1^t\mathbf{z}_1 = \frac{1}{n}\mathbf{a}_1^\mathbf{t}\mathbf{x}^\mathbf{t}\mathbf{x}\mathbf{a}_1 = \mathbf{a}_1^\mathbf{t}S\mathbf{a}_1$$

donde S es la matriz de covarianza de las observaciones.

Es obvio que podemos maximizar la varianza sin límite aumentando el módulo del vector $\mathbf{a}_1$.

Para que la maximización tenga solución debemos imponer una restricción al módulo del vector $\mathbf{a}_1$, y, sin pérdida de generalidad, impondremos que $\mathbf{a}_1^t\mathbf{a}_1 = 1$.

Introduciremos esta restricción mediante el multiplicador de Lagrange:

$$M = \mathbf{a}_1^{\mathbf{t}} S\mathbf{a}_1 - \boldsymbol{\lambda}(\mathbf{a}_1^t\mathbf{a}_1 - 1)$$

donde S es la matriz de covarianza de las observaciones.

Maximizaremos esta expresión de la forma habitual derivando respecto a los componentes de $\mathbf{a}_1$ e igualando a cero.

Entonces:

$$\frac{\partial M}{\partial \mathbf{a}_1} = 2S\mathbf{a}_1 - 2\lambda\mathbf{a}_1 = 0$$

Cuya solución es:

$$S\mathbf{a}_1 = \lambda\mathbf{a}_1$$

Esto implica que $\mathbf{a}_1$ es un vector propio de la matriz S, y λ su correspondiente valor propio.

Para determinar qué valor propio de S es la solución de la ecuación, tendremos en cuenta que, multiplicando por la izquierda por $\mathbf{a}_1^t$ y usando la restricción $\mathbf{a}_1^t\mathbf{a}_1 = 1$, queda:

$$\mathbf{a}_1^t S\mathbf{a}_1 = \lambda\mathbf{a}_1^t\mathbf{a}_1 = \lambda$$

Como vimos anteriormente, $var(\mathbf{z}_1) = \mathbf{a}_1^t S\mathbf{a}_1$

Entonces λ es la varianza de $\mathbf{z}_1$.

Como esta es la cantidad que queremos maximizar, λ será el mayor valor propio de la matriz S.

Su vector propio asociado, $\mathbf{a}_1$, define los coeficientes de cada variable en el primer componente principal.

Vamos a obtener el mejor plano de proyección de las variables originales.

Lo calcularemos estableciendo como función objetivo que la suma de las varianzas de $\mathbf{z}_1 = \mathbf{x}\mathbf{a}_1$ y $\mathbf{z}_2 = \mathbf{x}\mathbf{a}_2$ sea máxima donde $\mathbf{a}_1$ y $\mathbf{a}_2$ son los vectores que definen el plano.

La función objetivo será:

$$\phi = \mathbf{a}_1^t S \mathbf{a}_1 + \mathbf{a}_1^t S \mathbf{a}_1 - \lambda_1(\mathbf{a}_1^t \mathbf{a}_1 - 1) - \lambda_2(\mathbf{a}_2^t \mathbf{a}_2 - 1)$$

que incorpora las restricciones de que las direcciones deben de tener módulo unitario $\mathbf{a}_\mathrm{i}^t \mathbf{a}_\mathrm{i} = 1$, para $i = 1{,}2$.

Derivando e igualando a cero:

$$\frac{\partial \phi}{\partial \mathbf{a}_1} = 2S\mathbf{a}_1 - 2\lambda_1 \mathbf{a}_1 = 0$$

$$\frac{\partial \phi}{\partial \mathbf{a}_2} = 2S\mathbf{a}_2 - 2\lambda_2 \mathbf{a}_2 = 0$$

La solución de este sistema es:

$$S\mathbf{a}_1 = \boldsymbol{\lambda_1}\mathbf{a}_1$$

$$S\mathbf{a}_2 = \boldsymbol{\lambda_2}\mathbf{a}_2$$

que indica que $\mathbf{a}_1$ y $\mathbf{a}_2$ deben ser vectores propios de S.

Tomando los vectores propios de norma uno y sustituyendo, se obtiene que, en el máximo, la función objetivo es:

$$\phi = \lambda_1 + \lambda_2$$

Entonces:

- ✓ λ_1 y λ_2 deben ser los dos auto-valores mayores de la matriz S
- ✓ $\mathbf{a}_1$ y $\mathbf{a}_2$ deben ser sus correspondientes auto-vectores.

Observemos que la covarianza entre $\mathbf{z}_1$ y $\mathbf{z}_2$, dada por $\mathbf{a}_1^t S \mathbf{a}_2$, es cero ya que

$$\mathbf{a}_1^t \mathbf{a}_2 = 0$$

Entonces, las variables $\mathbf{z}_1$ y $\mathbf{z}_2$ estarán incorreladas.

Puede demostrarse que si en lugar de maximizar la suma de varianzas, que es la traza de la matriz de covarianzas de la proyección, se maximiza la varianza generalizada (el determinante de la matriz de covarianzas) se obtiene el mismo resultado.

Puede demostrarse análogamente que el espacio de dimensión r que mejor representa a los puntos viene definido por los vectores propios asociados a los r mayores autovalores de S. Estas direcciones se denominan direcciones principales de los datos y a las nuevas variables definidas por esas direcciones se les llama componentes principales. En general, la matriz de datos $\mathbf{x}$ (y por tanto la S) tienen rango p, existiendo entonces tantas componentes principales como variables que se obtendrán calculando los valores propios $\lambda_1, \ldots, \lambda_p$ de la matriz de covarianzas S, mediante:

$$|S - \lambda I| = 0$$

Y sus vectores asociados son:

$$(S - \lambda_i I)\, \mathbf{a}_\mathrm{i} = 0$$

Los términos λ_i son reales, al ser la matriz S simétrica.

Serán además positivos, ya que S es definida positiva.

Los vectores propios asociados a dos valores propios diferentes serán ortogonales.

Si S fuese semi-definida positiva de rango menor que p, habría algunos autovalores positivos y el resto serían ceros.

Llamando $\mathbf{z}$ a la matriz cuyas columnas son los valores de los p componentes en los n individuos, estas nuevas variables están relacionadas con las originales mediante:

$$\mathbf{z} = \mathbf{x}A$$

donde la matriz A cumple: $A^t A = I$.

Calcular los componentes principales equivale a aplicar una transformación ortogonal A a las variables $\mathbf{x}$ (ejes originales) para obtener unas nuevas

variables $\mathbf{z}$ incorreladas entre sí. Esta operación puede interpretarse como elegir unos nuevos ejes coordenados, que coincidan con los ejes naturales de los datos. La transformación ortogonal A es de tamaño $p \times r$ si se están hallando los r primeros componentes principales. Sus columnas serán los r primeros auto-vectores de la matriz de covarianzas de los datos S. La matriz de covarianza de $\mathbf{z}$ sería S_{z}, y es una matriz diagonal con elementos $\lambda_1, \dots, \lambda_r$, es decir, los primeros auto-valores de S.

$$\mathrm{S}_{\mathrm{z}} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_r \end{pmatrix}$$

Entonces, la utilidad de PCA se puede ver de las siguientes formas:

- ✓ Permite una representación óptima, en un espacio de dimensiones reducidas, de las observaciones originales.
- ✓ Permite que las variables correlacionadas originales se transformen en nuevas variables no correlacionadas, facilitando la interpretación de los datos.

PROPIEDADES DE LOS COMPONENTES

1. Conservan la variabilidad inicial:

   La suma de las varianzas de los componentes es igual a la suma de las varianzas de las variables originales, y la varianza generalizada de los componentes es igual a la original.

   En efecto, la varianza del componente $\mathbf{z}_h$ es:

   $$var(\mathbf{z}_h) = \lambda_h$$

   La suma de los auto-valores es la traza de la matriz de covarianza:

   $$traza(S) = var(\mathbf{x}_1) + \cdots + var(\mathbf{x}_p) = \lambda_1 + \cdots + \lambda_p$$

   Entonces:

   $$\sum_{i=1}^{p} var(\mathbf{x}_i) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} var(\mathbf{z}_i)$$

Esto significa que las nuevas variables $\mathbf{z}_i$ tienen conjuntamente la misma variabilidad que las variables originales, pero su distribución es muy distinta en los dos conjuntos.

Para comprobar que también conservan la varianza generalizada, valor del determinante de S, como el determinante es el producto de los λ_i:

$$|S_{\mathbf{x}}| = \lambda_1 \dots \lambda_p = \prod_{i=1}^{p} var(\mathbf{z}_i) = |S_{\mathbf{z}}|$$

2. La proporción de variabilidad explicada por un componente es el cociente entre su varianza, el valor propio asociado al vector propio que lo define, y la suma de los valores propios de la matriz.

   La varianza del componente h es λ_h, el valor propio que define el componente.

   La suma de todas las varianzas de las variables originales es $\sum_{i=1}^{p} \lambda_i$, que es igual a la suma de las varianzas de los componentes.

   Entonces la proporción de variabilidad total explicada por el componente h es:

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

3. Las covarianzas entre cada componente principal y las variables originales $\mathbf{x}$ vienen dadas por el producto entre las coordenadas del vector propio que define al componente y el valor propio:

$$cov(\mathbf{z}_i; \mathbf{x}_1, \dots, \mathbf{x}_p) = \lambda_i \mathbf{a}_i = (\lambda_i a_{i1}, \dots, \lambda_i a_{ip})$$

   donde $\mathbf{a_i}$ es el vector de coeficientes de la componente $\mathbf{z_i}$

   Para justificar este resultado, vamos a calcular la matriz $p \times p$ de covarianzas entre los componentes y las variables originales. Esta matriz es:

$$cov(\mathbf{z}, \mathbf{x}) = \frac{1}{n} \mathbf{z}^t \mathbf{x}$$

La primera fila de la matriz $cov(\mathbf{z}, \mathbf{x})$ proporciona las covarianzas entre la primera componente principal y las p variables originales. Como $\mathbf{z} = \mathbf{x}A$, sustituyendo:

$$cov(\mathbf{z}, \mathbf{x}) = \frac{\mathbf{1}}{\boldsymbol{n}} A^{\boldsymbol{t}} \mathbf{x}^{\mathbf{t}} \mathbf{x} = A^t S = DA^t$$

donde A contiene en sus columnas los vectores propios de S, y D es la matriz diagonal de los valores propios.

En consecuencia, la covarianza entre, por ejemplo, el primer componente principal y las p variables vendrá dada por la primera fila de $A^t S$, es decir, $\mathbf{a}_1^t S$, o también $\lambda_1 \mathbf{a}_1^{\mathbf{t}}$, donde $\mathbf{a}_1$ es el vector de coeficientes de la primera componente principal.

4. Las correlación entre un componente principal y una variable $\mathbf{x}$ es proporcional al coeficiente de esa variable en la definición del componente, y el coeficiente de proporcionalidad es el cociente entre la desviación típica del componente y la desviación típica de la variable.

   Para comprobarlo:

$$corr(\mathbf{z}_i, \mathbf{x_j}) = \frac{cov(\mathbf{z}_i, \mathbf{x_j})}{\sqrt{var(\mathbf{z}_i)var(\mathbf{x_j})}} = \frac{\lambda_{\boldsymbol{i}} a_{ij}}{\sqrt{\lambda_i s_j^2}} = a_{ij} \frac{\sqrt{\lambda_i}}{s_j}$$

5. Las r componentes principales ($r < p$) proporcionan la predicción lineal óptima con r variables del conjunto de variables $\mathbf{x}$.

   Esta afirmación puede expresarse de dos formas. La primera demostrando que la mejor predicción lineal con r variables de las variables originales se obtiene utilizando las r primeras componentes principales.

   La segunda demostrando que la mejor aproximación de la matriz de datos que puede construirse con una matriz de rango r se obtiene

construyendo esta matriz con los valores de los r primeros componentes principales.

6. Si estandarizamos los componentes principales, dividiendo cada uno por su desviación típica, se obtiene la estandarización multivariante de los datos originales.

   Estandarizando los componentes $\mathbf{z}$ por sus desviaciones típicas, se obtienen las nuevas variables:

$$\mathbf{y_c} = \mathbf{z}D^{-1/2} = \mathbf{x}AD^{-1/2}$$

   donde $D^{1/2}$ es la matriz que contiene las inversas de las desviaciones típicas de las componentes.

   La estandarización multivariante de una matriz de variables $\mathbf{x}$ de media cero viene dada por:

$$\mathbf{y_s} = \mathbf{x}AD^{-1/2}A^{\boldsymbol{t}}$$

   Por tanto, la estandarización multivariante puede interpretarse como obtener los componentes principales y estandarizarlos para que tengan todos la misma varianza.

PCA NORMADO

Los componentes principales se obtienen maximizando la varianza de la proyección. En términos de las variables originales esto supone maximizar:

$$M = \sum_{i=1}^{p} a_i^2 s_i^2 + 2\sum_{i=1}^{p}\sum_{j=i+1}^{p} a_i a_j s_{ij}$$

Con la restricción $\mathbf{a}^t\mathbf{a} = 1$

Si alguna de las variables originales, por ejemplo $\mathbf{x_1}$, tiene una varianza s_1^2 mayor que las demás, la manera de aumentar M es hacer tan grande como podamos la coordenada a_1 asociada a esta variable.

En el límite si una variable tiene una varianza mucho mayor que las demás el primer componente principal coincidirá muy aproximadamente con esta variable.

Cuando las variables tienen unidades distintas esta propiedad no es conveniente: si disminuimos la escala de medida de una variable cualquiera, de manera que aumenten en magnitud sus valores numéricos (pasamos por ejemplo de medir en km a medir en metros), el peso de esa variable en el análisis aumentará, ya que:

1. Su varianza será mayor y aumentará su coeficiente en el componente, ya que contribuye más a aumentar M.
2. Sus covarianzas con todas las variables aumentarán, con el consiguiente efecto de incrementar a_i.
3. En resumen, cuando las escalas de medida de las variables son muy distintas, la maximización de M dependerá decisivamente de estas escalas de medida y las variables con valores más grandes tendrán más peso en el análisis.

Si queremos evitar este problema, conviene estandarizar las variables antes de calcular los componentes, de manera que las magnitudes de los valores numéricos de las variables sean similares.

La estandarización resuelve otro posible problema.

Si las variabilidades de las $\mathbf{x}$ son muy distintas, las variables con mayor varianza van a influir más en la determinación de la primera componente.

Este problema se evita al estandarizar las variables, ya que entonces las varianzas son la unidad, y las covarianza son iguales a los coeficientes de correlación.

La ecuación a maximizar se transforma en:

$$M' = 1 + 2\sum_{i=1}^{p}\sum_{j=i+1}^{p} a_i\, a_j\, r_{ij}$$

donde r_{ij} es el coeficiente de correlación lineal entre las variables $\mathbf{x_i}$ y $\mathbf{x_j}$.

En consecuencia, la solución depende de las correlaciones y no de las varianzas.

Los componentes principales normados se obtienen calculando los vectores y valores propios de la matriz R de correlaciones.

Llamando λ_i^R a los valores propios de esa matriz, que suponemos no singular, se verifica que:

$$\sum_{i=1}^{p} \lambda_i^R = traza(R) = p$$

Las propiedades de los componentes extraídos de R son:

1. La proporción de variación o variabilidad explicada por λ_p^R será:

$$\frac{\lambda_p^R}{p}$$

2. Las correlaciones entre cada componente $\mathbf{z}_j$ y las variables $\mathbf{x}$ originales vienen dadas directamente por $\mathbf{a}_j^t\sqrt{\lambda_j}$ siendo $\mathbf{z}_j = \mathbf{x}\mathbf{a}_j$

Conclusiones:

1. Cuando las variables $\mathbf{x}$ originales están en distintas unidades conviene aplicar el análisis de la matriz de correlaciones o análisis normado.
2. Cuando las variables tienen las mismas unidades, ambas alternativas son posibles.
3. Si las diferencias entre las varianzas de las variables son informativas y queremos tenerlas en cuenta en el análisis, no debemos estandarizar las variables.
4. Por ejemplo, supongamos dos índices con la misma base, pero uno fluctúa mucho y el otro es casi constante. Este hecho es informativo, y para tenerlo en cuenta en el análisis, no se deben estandarizar las variables, de manera que el índice de mayor variabilidad tenga más peso.

5. Por el contrario, si las diferencias de variabilidad no son relevantes podemos eliminarlas con el análisis normado. En caso de duda, conviene realizar ambos análisis, y seleccionar aquel que conduzca a conclusiones más informativas.

INTERPRETACIÓN

En la práctica, los datos pueden ser representados por los "principal component scores", que son los valores de las nuevas variables.

Si hemos usado la matriz de covarianza muestral S de los datos originales **x,** la matriz que contiene a los "principal component scores" es:

$$\mathbf{z} = \tilde{\mathbf{x}}\, V_r^{S_x}$$

donde $V_r^{S_x}$ es la matriz que contiene a los auto-vectores $a_1^S, \dots, a_r^S$ de S_x asociados a los r mayores auto-valores $\lambda_1^S, \dots, \lambda_r^S$.

Si hemos usado la matriz de correlaciones muestrales R de los datos originales **x,** la matriz que contiene a los "principal component scores" es:

$$\mathbf{z} = \mathbf{y}\, V_r^{S_x} = \tilde{\mathbf{x}}\, D_{S_x}^{-1/2}\, V_r^{R_x}$$

donde $V_r^{R_x}$ es la matriz que contiene a los auto-vectores $a_1^R, \dots, a_r^R$ de la matriz de correlaciones de los datos originales R_x, asociados a los r mayores mayores auto-valores $\lambda_1^R, \dots, \lambda_r^R$. Notemos que $\mathbf{y} = \tilde{\mathbf{x}}\, D_{S_x}^{-1/2}$ son los datos originales normados, donde D_{S_x} es la matriz diagonal que contiene las varianzas muestrales de **x**.

INTERPRETACIÓN

Cuando existe una alta correlación positiva entre todas las variables, el primer componente principal tiene todas sus coordenadas del mismo signo y puede interpretarse como un promedio ponderado de todas las variables.

El primer componente se interpreta entonces como un factor global de "tamaño".

Los restantes componentes se interpretan como factores “de forma” y típicamente tienen coordenadas positivas y negativas, que implica que contraponen unos grupos de variables frente a otros.

Estos factores “de forma” pueden frecuentemente escribirse como medias ponderadas de dos grupos de variables con distinto signo y contraponen las variables de un signo a las del otro.

SELECCIÓN DE LOS COMPONENTES

Se han sugerido distintas reglas para seleccionar el número de componentes a mantener:

1. Realizar un gráfico (scree-plot) de λ_i frente a i . Comenzar seleccionando componentes hasta que los restantes tengan aproximadamente el mismo valor de λ_i. La idea es buscar un “codo” en el gráfico, es decir, un punto a partir del cual los valores propios λ_i son aproximadamente iguales. El criterio es quedarse con un número de componentes que excluya los asociados a valores pequeños y aproximadamente del mismo tamaño.

2. Seleccionar componentes hasta cubrir una proporción determinada de varianza, como el 80% o el 90%. Esta regla es arbitraria y debe aplicarse con cierto cuidado. Por ejemplo, es posible que un único componente de “tamaño” recoja el 90% de la variabilidad y sin embargo pueden existir otros componentes que sean muy adecuados para explicar la “forma” las variables.

3. Desechar aquellos componentes asociados a valores propios inferiores a una cota, que suele fijarse como la varianza media $\sum \lambda_i / p$, En particular, cuando se trabaja con la matriz de correlación R, el valor medio de los componentes es 1, y esta regla lleva a seleccionar los valores propios mayores que la unidad. De nuevo esta regla es arbitraria: una variable que sea independiente del resto puede tener un valor propio mayor que la unidad. Sin embargo, si está incorrelada con el resto puede ser una variable poco relevante

para el análisis, y no aportar mucho a la comprensión del fenómeno global.

INTERPRETACIÓN GRÁFICA

La interpretación de los componentes principales se favorece representando las proyecciones de las observaciones sobre un espacio de dimensión dos, definido por parejas de los componentes principales más importantes. La representación habitual es tomar dos ejes ortogonales que representen los dos componentes considerados, y situar cada punto sobre ese plano por sus coordenadas con relación a estos ejes, que son los valores de los dos componentes para esa observación. Por ejemplo, en el plano de los dos primeros componentes, las coordenadas del punto $\mathbf{x}_i$ son $\mathbf{z}_{1i} = \mathbf{a}_1^t \mathbf{x}_i$ y $\mathbf{z}_{2i} = \mathbf{a}_2^t \mathbf{x}_i$

ANÁLISIS FACTORIAL

El análisis factorial tiene por objetivo explicar un conjunto de variables observadas por un pequeño número de variables latentes, o no observadas, que llamaremos factores. Por ejemplo, supongamos que hemos tomado veinte medidas físicas del cuerpo de una persona: estatura, longitud del tronco y de las extremidades, anchura de hombros, peso, etc. Es intuitivo que todas estas medidas no son independientes entre sí, y que conocidas algunas de ellas podemos prever con poco error las restantes porque las dimensiones del cuerpo humano dependen de ciertos factores, y si estos fuesen conocidos podríamos prever con poco error los valores de las variables observadas.

Otro ejemplo, supongamos que estamos interesados en estudiar el desarrollo humano en los países del mundo. Disponemos de muchas variables económicas, sociales y demográficas, en general dependientes entre sí, que están relacionadas con el desarrollo. Podemos preguntarnos si el desarrollo de un país depende de un pequeño número de factores tales que, conocidos sus valores, pudiéramos prever el conjunto de las variables económicas que teníamos de cada país.

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 unos factores, no directamente 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 Karl Pearson y Charles Spearman en comprender las dimensiones de la inteligencia humana en los años 30, y muchos de sus avances se han producido en el área de la psicometría. El análisis factorial está relacionado con los componentes principales, pero existen ciertas diferencias. En primer lugar, los componentes principales se construyen para explicar las varianzas, mientras que los factores se construyen para explicar las covarianzas o correlaciones entre las variables. En segundo lugar, componentes principales es un herramienta descriptiva, mientras que el análisis factorial presupone un modelo estadístico formal de generación de la muestra dada.

MODELO FACTORIAL

Supondremos que observamos un vector de variables $\mathbf{x}$, de dimensiones $p \times 1$, en n elementos de una población. El modelo de análisis factorial establece que este vector de datos observados se genera mediante la relación:

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda\mathbf{f} + \mathbf{u}$$

$\mathbf{f}$: es un vector $m \times 1$ de variables latentes o factores no observados. Supondremos que siguen una distribución $N_m(\mathbf{0}, I)$, es decir, los factores son variables de media cero e independientes entre sí y con distribución Normal.

Λ: es una matriz de tamaño $p \times m$ de constantes desconocidas $(m < p)$. Contiene los coeficientes que describen cómo los factores $\mathbf{f}$ afectan a las variables observadas $\mathbf{x}$ y se denomina matriz de carga (loading matrix).

Supondremos que observamos un vector de variables $\mathbf{x}$, de dimensiones $p \times 1$, en n elementos de una población.

El modelo de análisis factorial establece que este vector de datos observados se genera mediante la relación:

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda\mathbf{f} + \mathbf{u}$$

$\mathbf{u}$: es un vector de tamaño $p \times 1$ de perturbaciones no observadas. Recoge el efecto de todas las variables distintas de los factores que influyen sobre $\mathbf{x}$. Supondremos que $\mathbf{u} \sim N_p(\mathbf{0}, \psi)$ donde ψ es una matriz diagonal, y también que las perturbaciones están incorreladas con los factores $\mathbf{f}$, es decir, $cov[\mathbf{f}, \mathbf{u}] = \mathbf{0}$.

Con estas tres hipótesis deducimos que:

- $\boldsymbol{\mu}$ es la media de las variables de las variables $\mathbf{x}$, ya que tanto los factores como las perturbaciones tienen media cero.
- $\mathbf{x}$ tiene distribución Normal, al ser suma de variables Normales, y llamando S a su matriz de covarianzas

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}, S)$$

- La ecuación del modelo implica que dada una muestra aleatoria simple de n elementos generada por el modelo factorial, cada dato x_{ij} puede escribirse como:

$$x_{ij} = \mu_j + \lambda_{j1} f_{1i} + \cdots + \lambda_{jm} f_{mi} + u_{ij}$$

que descompone x_{ij}, el valor observado en el individuo i de la variable j, como suma de $m + 2$ términos.

El primer término es la media de la variable j, μ_j

Del segundo término al término $m + 1$ recogen el efecto de los m factores.

Y el último término, $m + 2$, es una perturbación específica de cada observación, u_{ij}.

Los efectos de los factores sobre x_{ij} son el producto de los coeficientes $\lambda_{j1}, \ldots, \lambda_{jm}$ que dependen de la relación entre cada factor y la variable j (y

que son los mismos para todos los elementos de la muestra), por los valores de los m factores en el elemento muestral i, $f_{1i}, \ldots, f_{mi}$.

Poniendo juntas las ecuaciones para todas las n observaciones, la matriz de datos $\mathbf{x}$ de tamaño $n \times p$ puede escribirse como:

$$\mathbf{x} = \mathbf{1}\boldsymbol{\mu}^t + \mathbf{F}\Lambda^t + \mathbf{U}$$

donde $\mathbf{1}$ es un vector $n \times 1$ de unos, $\mathbf{F}$ es una matriz $n \times m$ que contiene los m factores para los n elementos, Λ^t es la transpuesta de la matriz de carga que resulta ser $m \times p$ cuyos coeficientes constantes relacionan las variables y los factores, y $\mathbf{U}$ es una matriz $n \times p$ de perturbaciones.

PROPIEDADES

La matriz de carga Λ contiene las covarianzas entre los factores y las variables observadas.

En efecto, la matriz de covarianzas Λ entre las variables y los factores se obtiene multiplicando por $\mathbf{f}^t$ por la derecha y tomando esperanzas en el modelo factorial:

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda\mathbf{f} + \mathbf{u}$$

$$\mathbf{x} - \boldsymbol{\mu} = \Lambda\mathbf{f} + \mathbf{u}$$

$$(\mathbf{x} - \boldsymbol{\mu})\mathbf{f}^t = \Lambda\mathbf{f}\mathbf{f}^t + \mathbf{u}\mathbf{f}^t$$

$$\boldsymbol{E}[(\mathbf{x} - \boldsymbol{\mu})\mathbf{f}^t] = \Lambda\mathrm{E}[\mathbf{f}\mathbf{f}^t] + \boldsymbol{E}[\mathbf{u}\mathbf{f}^t] = \Lambda$$

Ya que, por hipótesis, los factores están incorrelados, es decir, $\mathrm{E}[\mathbf{f}\mathbf{f}^t] = I$, y tienen media cero y están incorrelados con las perturbaciones, es decir, $cov[\mathbf{u}, \mathbf{f}] = \boldsymbol{E}[\mathbf{u}\mathbf{f}^t] = 0$.

Entonces Λ se define como:

$$\Lambda = \boldsymbol{E}[(\mathbf{x} - \boldsymbol{\mu})\mathbf{f}^t]$$

Esta ecuación indica que los términos λ_{ij} de la matriz de carga Λ, representan la covarianza entre la variable $\mathbf{x}_i$ y el factor $\mathbf{f}_j$ y al tener los

factores varianza unidad, son los coeficientes de regresión cuando explicamos las variables observadas por los factores.

En el caso particular en que las variables $\mathbf{x}$ estén estandarizadas, los términos λ_{ij} son también las correlaciones entre las variables y los factores.

La matriz de covarianzas entre las observaciones verifica, según

$$S = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\boldsymbol{t}}] = \Lambda E[\mathbf{f}\mathbf{f}^{\mathrm{t}}]\Lambda^{\mathrm{t}} + E[\mathbf{u}\mathbf{u}^{t}]$$

Ya que $E[\mathbf{f}\mathbf{u}^{t}] = 0$ al estar incorrelados los factores y el ruido.

Entonces, se obtiene la propiedad fundamental:

$$S = \Lambda\Lambda^{t} + \psi$$

Que establece que la matriz de covarianzas de los datos observados admite una descomposición como suma de dos matrices:

1. La primera $\Lambda\Lambda^{t}$ es una matriz simétrica de rango $m < p$. Esta matriz contiene la parte común al conjunto de las variables y depende de las covarianzas entre las variables y los factores.

2. La segunda ψ, es una matriz diagonal que contiene la parte específica de cada variable, que es independiente del resto.

Esta descomposición $S = \Lambda\Lambda^{t} + \psi$ implica que las varianzas de las variables observadas pueden descomponerse como:

$$\sigma_i^2 = \sum_{j=1}^{m} \lambda_{ij}^2 + \psi_i^2$$

con $i = 1, \ldots, p$, donde el primer término es la suma de los efectos de los factores y el segundo término es el efecto de la perturbación.

Llamando:

$$h_i^2 = \sum_{j=1}^{m} \lambda_{ij}^2$$

a la suma de los efectos de los factores que llamaremos “comunalidad”, tenemos que:

$$\sigma_i^2 = h_i^2 + \psi_i^2$$

para $i = 1, \dots, p$

En esta descomposición:

$$\sigma_i^2 = h_i^2 + \psi_i^2$$

- h_i^2 se llama la i −ésima *"comunalidad".*
- ψ_i^2 es la varianza del i −ésimo elemento de la perturbación $\mathbf{u}$ y se llaman "$unicidades$".

Esta igualdad puede interpretarse como una descomposición de la varianza en:

$$\sigma_i^2 = h_i^2 + \psi_i^2$$

$$Varianza\ observada = Variabilidad\ com\acute{u}n\ (Comunalidad) + Variabilidad\ espec\acute{i}fica\ (Unicidad)$$

Esto es análogo a la descomposición clásica de la variabilidad de los datos en una parte explicada y otra no explicada que se realiza en el análisis de la varianza.

En el modelo factorial la parte explicada es debida a los factores y la no explicada es debido al ruido o componente aleatorio.

UNICIDAD

En el modelo factorial, ni la matriz de carga, Λ, ni los factores, $\mathbf{f}$, son observables.

Esto plantea un problema de indeterminación: dos representaciones $(\Lambda, \mathbf{f})$ y $(\Lambda^*, \mathbf{f}^*)$ serán equivalentes si

$$\Lambda\mathbf{f} = \Lambda^*\mathbf{f}^*$$

Esta situación conduce a dos tipos de indeterminación.

1. Un conjunto de datos puede explicarse con la misma precisión con factores incorrelados o correlados.

2. Los factores no quedan determinados de manera única.

Vamos a analizar estas dos indeterminaciones. Para mostrar la primera, si H es cualquier matriz no singular, la representación usual puede también escribirse como

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda H H^{-1}\mathbf{f} + \mathbf{u}$$

Y llamando $\Lambda^* = \Lambda H$ a la nueva matriz de carga y $\mathbf{f}^* = H^{-1}\mathbf{f}$ a los nuevos factores:

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda^*\mathbf{f}^* + \mathbf{u}$$

Donde los nuevos factores tienen ahora una distribución:

$$\mathbf{f}^* \sim N(\mathbf{0}, H^{-1}(H^{-1})^t)$$

Y por lo tanto están correlados.

Análogamente, partiendo de factores correlados $\mathbf{f} \sim N(\mathbf{0},\, S_f)$ siempre podemos encontrar una expresión equivalente de las variables mediante un modelo con factores incorrelados.

En efecto, sea A una matriz tal que $S_f = AA^t$, que siempre existe si S_f es definida positiva, entonces $\mathrm{A}^{-1}S_f(\boldsymbol{A^{-1}})^{\boldsymbol{t}} = \boldsymbol{I}$.

Entonces:

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda A A^{-1}\mathbf{f} + \mathbf{u}$$

Y tomando $\Lambda^* = \Lambda A$ como la nueva matriz de coeficientes de los factores y $\mathbf{f}^* = A^{-1}\mathbf{f}$ como los nuevos factores, el modelo es equivalente a otro con factores incorrelados.

Esta indeterminación se ha resuelto en las hipótesis del modelo tomando siempre los factores como incorrelados.

En segundo lugar, si H es ortogonal, los modelos:

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda\mathbf{f} + \mathbf{u}$$

y

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda H H^{\mathrm{t}} \mathbf{f} + \mathbf{u}$$

son indistinguibles porque $H = H^t = H^{-1}$ y $HH^{\mathrm{t}} = I$

Ambos contienen factores incorrelados, con matriz de covarianzas la identidad. En este sentido, decimos que el modelo factorial está indeterminado ante rotaciones. Esta indeterminación se puede resolver imponiendo restricciones sobre los componentes de la matriz de carga.

CRITERIO VARIMAX

Como vimos anteriormente, la matriz de carga no está identificada ante multiplicaciones por matrices ortogonales, que equivalen a rotaciones. En análisis factorial está definido el espacio de las columnas de la matriz de carga, pero cualquier base de este espacio puede ser una solución. Para elegir entre las posibles soluciones, se tienen en cuenta la interpretación de los factores. Intuitivamente, será más fácil interpretar un factor cuando se asocia a un bloque de variables observadas. Esto ocurrirá si las columnas de la matriz de carga, que representan el efecto de cada factor sobre las variables observadas, contienen valores altos para ciertas variables y pequeños para otras. Esta idea puede plantearse de distintas formas que dan lugar a distintos criterios para definir la rotación. Los coeficientes de la matriz ortogonal que define la rotación se obtendrán minimizando una función objetivo que expresa la simplicidad deseada en la representación conseguida al rotar. El criterio más utilizado es el Varimax, que exponemos a continuación.

La interpretación de los factores se facilita si los que afectan a algunas variables no lo hacen a otras y al revés. Este objetivo conduce al criterio de maximizar la varianza de los coeficientes que definen los efectos de cada factor sobre las variables observadas. Para precisar este criterio, llamemos:

- δ_{ij} a los coeficientes de la matriz de carga asociados al factor j en las $i = 1, \dots, p$ ecuaciones después de la rotación.

- $\boldsymbol{\delta}_j$ al vector que es la columna j de la matriz de carga después de la rotación.

Se desea, que la varianza de los coeficientes al cuadrado de este vector sea máxima. Se toman los coeficientes al cuadrado para prescindir de los signos, ya que interesa su valor absoluto.

Llamando $\bar{\delta}_{.j} = \sum \delta_{ij}^2/p$ a la media de los cuadrados de los componentes del vector $\boldsymbol{\delta}_j$, la variabilidad para factor j es:

$$\frac{1}{p}\sum_{i=1}^{p}\left(\delta_{ij}^2 - \bar{\delta}_{.j}\right)^2 = \frac{1}{p}\sum_{i=1}^{p}\delta_{ij}^4 - \left(\frac{1}{p}\right)^2\left(\sum_{i=1}^{p}\delta_{ij}^2\right)^2$$

Y el criterio es maximizar la suma de las varianzas para todos los factores, dada por:

$$VC = \left(\frac{1}{p}\right)\sum_{j=1}^{m}\sum_{i=1}^{p}\delta_{ij}^4 - \left(\frac{1}{p}\right)^2\sum_{j=1}^{m}\left(\sum_{i=1}^{p}\delta_{ij}^2\right)^2$$

Sea Λ la matriz de carga estimada inicialmente. El problema es encontrar una matriz ortogonal M tal que la matriz δ dada por:

$$\delta = \Lambda M$$

Y cuyos coeficientes δ_{ij} vienen dados por

$$\delta_{\mathrm{ij}} = \boldsymbol{\lambda}_i^t \mathbf{m}_j$$

Siendo $\boldsymbol{\lambda}_i^t$ la fila i de la matriz Λ, y siendo $\mathbf{m}_j$ la columna j de la matriz M que buscamos.

Los términos de la matriz M se obtendrán derivando la ecuación de VC respecto a cada uno de sus términos m_{ij} teniendo en cuenta las restricciones de ortogonalidad $\mathbf{m}_i^{\mathrm{t}}\mathbf{m}_i = 1$ y $\mathbf{m}_i^t\mathbf{m}_j = 0$ para $i \neq j$. El resultado obtenido es la rotación varimax.

En resumen, empezando con una matriz de carga Λ, podemos considerar matrices de carga rotadas $\Lambda^* = \Lambda H$ donde H es una matriz ortogonal cuadrada.

El criterio Varimax selecciona la matriz ortogonal tal que:

$$H^* = \arg\max VC(\Lambda H)$$

Lo cual lleva al resultado $\Lambda^* = \Lambda H^*$

La rotación óptima varimax lleva a visualizaciones de la matriz de carga que permiten una interpretación más sencilla que las matrices de carga sin rotar.

PCFA

Si tenemos nuestros datos $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ y sabemos que las variables univariantes tienen diferentes unidades de medidas, es preferible estandarizar las variables.

La estandarización univariante de las variables resulta en una nueva variable aleatoria multivariante

$$\mathbf{y} = \boldsymbol{\Delta}_{\mathbf{x}}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{x}})$$

donde Δ_x es una matriz diagonal, y sus elementos son las varianzas de las variables $\mathbf{x}_i$, $i = 1, \dots, p$.

La variable aleatoria multivariante estandarizada $\mathbf{y}$ tendrá media $\mathbf{0}_p$ y matriz de covarianza $S_y = R_x$ igual a la matriz de correlaciones de $\mathbf{x}$.

Entonces, si $\mathbf{x}$ sigue el modelo factorial:

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda\mathbf{f} + \mathbf{u}$$

Su estandarización $\mathbf{y}$ seguirá el modelo factorial:

$$\mathbf{y} = \boldsymbol{\Delta}_{\mathbf{x}}^{-1/2}\Lambda\mathbf{f} + \boldsymbol{\Delta}_{\mathbf{x}}^{-1/2}\mathbf{u}$$

Y la matriz de covarianzas de $\mathbf{y}$ podrá descomponerse como:

$$S_y = R_x = \boldsymbol{\Delta}_{\mathbf{x}}^{-1/2}\Lambda\Lambda^{\mathrm{t}}\,\boldsymbol{\Delta}_{\mathbf{x}}^{-1/2} + \boldsymbol{\Delta}_{\mathbf{x}}^{-1/2}\psi_{\boldsymbol{u}}\boldsymbol{\Delta}_{\mathbf{x}}^{-1/2}$$

donde $\psi_{\boldsymbol{u}}$ es la matriz diagonal de covarianzas de los errores $\mathbf{u}$.

Como consecuencia, el modelo factorial de $\mathbf{y}$ es similar al modelo factorial de $\mathbf{x}$ con:

1. Una matriz de carga $M = \mathbf{\Delta_x^{-1/2}}\Lambda$
2. Un conjunto de factores (que son los mismos) $\mathbf{f}$
3. Un conjunto de errores $\epsilon = \mathbf{\Delta_x^{-1/2}u}$ con matriz de covarianzas diagonal $\mathrm{S}_\epsilon = \mathbf{\Delta_x^{-1/2}}\psi_{\boldsymbol{u}}\mathbf{\Delta_x^{-1/2}}$

En otras palabras, tenemos que el modelo factorial resulta:

$$\mathrm{y} = \mathrm{Mf} + \epsilon$$

Con descomposición de la matriz de covarianzas:

$$\mathrm{S_y} = \mathrm{R_x} = \mathrm{MM^t} + \mathrm{S}_\epsilon$$

Trabajar con esta descomposición de S_y es útil porque las entradas de la diagonal van a ser iguales a 1, y la suma de las comunalidades será igual a 1. Sus interpretaciones son más sencillas.

Sobre los métodos no-distribucionales, los más conocidos son:

1. Principal Component Factor Analysis (PCFA)
2. Principal Factor Analysis (PFA)
3. Ambos están basados en:
4. La descomposición de la matriz de covarianza dada por $S_x = \Lambda\Lambda^t + \psi$ si trabajamos con los datos $\mathbf{x}$ originales.
5. La descomposición de la matriz de correlaciones dada por $R_x = MM^t + \mathrm{S}_\epsilon$ si trabajamos con los datos $\mathbf{y}$ estandarizados.

Por simplicidad vamos a presentar los métodos sólo para el primer caso, con los datos $\mathbf{x}$ y la descomposición de su matriz de covarianza, porque resultados similares son fáciles de obtener para $\mathbf{y}$.

Primero, hagamos una descomposición espectral de S_x:

$$S_x = V_p D_p V_P^t = \Lambda\Lambda^t + \psi$$

Donde V_p es la matriz que contiene los auto-vectores de S_x y D_p es una matriz diagonal de tamaño $p \times p$ que contiene los auto-valores de S_x.

Por simplicidad vamos a presentar los métodos sólo para el primer caso, con los datos $\mathbf{x}$ y la descomposición de su matriz de covarianza, porque resultados similares son fáciles de obtener para $\mathbf{y}$.

Si asumimos $\psi = 0_{p\times p}$ entonces

$$V_p D_p V_P^t = \Lambda\Lambda^t$$

Como Λ tiene dimensión $p \times m$, $\Lambda\Lambda^t$ es una matriz de rango $m < p$.

Como consecuencia, D_p contiene $p - m$ autovalores iguales a cero, y en ese caso:

$$\Lambda = V_m D_m^{1/2}$$

donde $\mathrm{V_m}$ es la matriz que contiene los auto-vectores de $\mathrm{S_x}$ asociados a los m auto-valores de $\mathrm{S_x}$ diferentes de cero. Y D_m es la matriz diagonal de tamaño $m \times m$ que contiene esos auto-valores.

La idea para seleccionar m es como en PCA, usando la varianza explicada por los componentes.

EJEMPLO

Consideremos las siguientes variables univariantes medidas en 50 estados de USA:

- $\mathbf{x_1}$: estimación de población (en miles) a fecha 1ro Julio, 1975
- $\mathbf{x_2}$: ingresos per cápita (en dólares) en el año 1974
- $\mathbf{x_3}$: analfabetismo (porcentaje) en el año 1970
- $\mathbf{x_4}$: esperanza de vida entre los años 1969-1971
- $\mathbf{x_5}$: tasa de homicidios y homicidios no negligentes (por 100000) 1976
- $\mathbf{x_6}$: porcentaje de graduados de secundaria 1970
- $\mathbf{x_7}$: número medio de días con temperatura mínima bajo cero (1931-1960) en la capital o gran ciudad

- $\mathbf{x_8}$: área de terreno en millas cuadradas

Tomamos logaritmos para la 1ra, 3ra y 8va. Adicionalmente, consideremos la matriz de correlaciones en vez de la de covarianzas. Los 3 primeros auto-vectores explican el 76.69% de la variabilidad total.

Usando los 3 primeros auto-valores, los de mayor valor, de la matriz de correlaciones, y sus auto-vectores asociados, podemos estimar la matriz de carga M.

Para hallarla:

$$M = V_3^{R_x}\left(D_3^{R_x}\right)^{-1/2}$$

donde:

- $V_3^{R_x}$ es la matriz con los primeros tres auto-vectores de R_x.

$D_3^{R_x}$ es la matriz diagonal con los tres auto-valores más grandes de R_x.

La matriz de carga estimada es igual a:

$$\widehat{M} = \begin{pmatrix} -0.44 & -0.47 & 0.53 \\ 0.52 & -0.54 & 0.26 \\ -0.87 & 0.04 & 0.13 \\ 0.76 & -0.05 & 0.41 \\ -0.84 & -0.31 & -0.23 \\ 0.80 & -0.41 & -0.07 \\ 0.69 & 0.25 & -0.40 \\ -0.06 & -0.67 & -0.61 \end{pmatrix}$$

Si usamos la rotación Varimax, la estimación final de la matriz de carga sería:

$$\widehat{M}^* = \begin{pmatrix} 0.00 & 0.00 & 0.84 \\ 0.78 & -0.10 & 0.12 \\ -0.66 & 0.00 & 0.58 \\ 0.76 & 0.38 & -0.16 \\ -0.57 & -0.52 & 0.51 \\ 0.82 & -0.20 & -0.32 \\ 0.28 & 0.00 & -0.79 \\ 0.00 & -0.90 & 0.00 \end{pmatrix}$$

El primer factor distingue entre estados fríos con riqueza, educación, longevidad y no violentos, de estados cálidos con pobreza, poco educados, con poca longevidad y violentos.

El segundo factor distingue a los grandes estados con poblaciones ricas y educadas pero violentas y de poca longevidad, de los pequeños estados con personas muy longevas.

El tercer factor distingue los estados poblados y violentos de los estados menos poblados, fríos, no violentos, pero de poca longevidad.

Una vez que hemos estimado la matriz de carga, es posible estimar la matriz de covarianza de los errores, como la matriz diagonal:

$$\hat{\psi} = R_x - \widehat{M}^*\left(\widehat{M}^*\right)^t$$

Lo cual nos da las siguientes "unicidades": 0.28, 0.35, 0.21, 0.24, 0.12, 0.17, 0.29, 0.16.

Entonces, las variables mejor explicadas por el modelo son:

- $\mathbf{x_5}$**:** tasa de homicidios
- $\mathbf{x_6}$**:** porcentaje de graduados de secundaria
- $\boldsymbol{log}(\mathbf{x_8})$**:** logaritmo del área de terreno

PFA

En Principal Factor Analysis (PFA) también empezaremos con la igualdad:

$$S_x = \Lambda\Lambda^t + \psi$$

Entonces, $\Lambda\Lambda^t = S_x - \psi$ tiene que ser una matriz de rango $m < p$ porque Λ tiene dimensión $p \times m$.

Como consecuencia tiene $p - m$ auto-valores iguales a cero.

La descomposición espectral de $S_x - \psi$ es dada por:

$$S_x - \psi = U_m \Omega_m U_m^t$$

Donde

- U_m es la matriz que contiene los auto-vectores de $S_x - \psi$ asociados a los m auto-valores de $S_x - \psi$ diferentes de cero.

Ω_m es la matriz diagonal $m \times m$ que contiene estos auto-valores.

Entonces podemos poner:

$$\Lambda = U_m \Omega_m^{1/2}$$

El problema es que hemos empezado por $S_x - \psi$ y de esta expresión ψ es desconocida.

ψ tiene que ser estimada. Por ejemplo podríamos usar la estimación de ψ obtenida por PCFA:

$$\hat{\psi} = R_x - \hat{M}^* \left(\hat{M}^*\right)^t$$

Así, los auto-valores y auto-vectores de $S_x - \psi$ serán reemplazados por aquellos de $S_x - \hat{\psi}$

EJEMPLO

Continuando con el ejemplo anterior, los primeros tres auto-vectores de $R_x - \hat{\psi}$ explican el 88.8% de la variabilidad total. Entonces, usando los tres auto-valores más grandes de la matriz de correlaciones y sus auto-vectores asociados, podemos estimar la matriz de carga M:

Para hallarla:

$$M = U_3^{R_x - \hat{\psi}} \left(\Omega_3^{R_x - \hat{\psi}}\right)^{-1/2}$$

donde:

- $U_3^{R_x - \hat{\psi}}$ es la matriz con los primeros tres auto-vectores de $R_x - \hat{\psi}$.

$\Omega_3^{R_x-\widehat{\psi}}$ es la matriz diagonal con los tres auto-valores más grandes de $R_x - \hat{\psi}$.

La matriz de carga estimada es igual a:

$$\widehat{M} = \begin{pmatrix} -0.42 & -0.27 & 0.56 \\ 0.48 & -0.36 & 0.32 \\ -0.84 & 0.08 & 0.10 \\ 0.73 & 0.02 & 0.36 \\ -0.84 & -0.34 & -0.12 \\ 0.78 & -0.40 & 0.04 \\ 0.66 & 0.12 & -0.40 \\ -0.06 & -0.77 & -0.36 \end{pmatrix}$$

Si usamos la rotación Varimax, la estimación final de la matriz de carga sería:

$$\widehat{M}^* = \begin{pmatrix} 0.00 & 0.00 & 0.75 \\ 0.67 & 0.00 & 0.00 \\ -0.65 & 0.00 & 0.56 \\ 0.74 & 0.30 & -0.17 \\ -0.59 & -0.47 & 0.51 \\ 0.80 & -0.21 & -0.30 \\ 0.28 & 0.00 & -0.73 \\ 0.00 & -0.85 & 0.00 \end{pmatrix}$$

Notemos que ambas matrices están muy cerca de las estimaciones hechas con PCFA. Consecuentemente, la interpretación es similar a la obtenida por el método anterior.

Una vez que hemos estimado la matriz de carga, podemos obtener una nueva estimación de la matriz de covarianza de los errores $\hat{\psi}$ con la matriz diagonal: $R_x - \widehat{M}^*\left(\widehat{M}^*\right)^t$.

En este caso, las unicidades son: 0.42, 0.52, 0.26, 0.31, 0.15, 0.21, 0.38, 0.26.

Entonces las variables originales mejor explicadas por los factores son:

- $\log(\mathbf{x}_3)$**:** logaritmo del analfabetismo
- $\mathbf{x}_6$**:** porcentaje de graduados de secundaria
- $log(\mathbf{x}_8)$**:** logaritmo del área de terreno

ESTIMACIÓN DE LOS FACTORES

Vamos a ver el método Bartlett, que supone que el vector de valores de los factores para cada observación es un parámetro a estimar.

El vector de tamaño $p \times 1$ de valores de las variables en el individuo i, $\mathbf{x}_i$, tiene una distribución Normal con media $\Lambda \mathbf{f}_\mathrm{i}$, donde $\mathbf{f}_i$ es el vector de tamaño $m \times 1$ de factores para el elemento i en la muestra, y matriz de covarianza ψ, es decir:

$$\mathbf{x}_i \sim N_p(\Lambda \mathbf{f}_\mathrm{i}, \psi)$$

Los parámetros $\mathbf{f}_\mathrm{i}$ pueden estimarse por máxima verosimilitud.

Si conocemos Λ, el modelo factorial:

$$\mathbf{x}_i = \mu_x + \Lambda \mathbf{f}_i + \mathbf{u}_i$$

es un modelo de regresión con variable dependiente $\mathbf{x}_i$, variables explicativas las columnas de Λ, y parámetros desconocidos $\mathbf{f}_i$.

Como la perturbación, $\mathbf{u}_i$, no se distribuye como $N(0, I)$ sino $N(0, \psi)$, tendremos que utilizar mínimos cuadrados generalizados:

$$\hat{\mathbf{f}}_\mathrm{i} = \left(\hat{\Lambda}^t \hat{\psi}^{-1} \hat{\Lambda}\right)^{-1} \hat{\Lambda}^t \hat{\psi}^{-1} (\mathbf{x}_i - \hat{\mu}_x)$$

Si los datos han sido escalados:

$$\hat{\mathbf{f}}_\mathrm{i} = \left(\widehat{M}^t \widehat{S_\epsilon}^{-1} \widehat{M}\right)^{-1} \widehat{M}^t \widehat{S_\epsilon}^{-1} (\mathbf{y}_i)$$

El análisis factorial de componentes principales (PCFA) y el análisis factorial principal (PFA) son procedimientos no paramétricos y, por lo tanto, pueden aplicarse a los datos sin requerir conocimiento de la distribución subyacente de los datos.

Si sabemos que los datos son gaussianos o no muy diferentes de los gaussianos, explotar este conocimiento adicional puede conducir a mejores estimadores para la matriz de carga y los factores.

Ahora, consideramos el caso en el que $\mathbf{x}$ y los errores $\mathbf{u}$ están Normalmente distribuidos, y usaremos el método de máxima verosimilitud para estimar los parámetros del modelo.

Dada la matriz de datos generada por un modelo factorial Gaussiano, la log-verosimilitud de los parámetros del modelo es:

$$l(\boldsymbol{\mu}_x, \Sigma_x|\mathbf{x}) = -\frac{np}{2}\log 2\pi - \frac{n}{p}\log|\Sigma_x| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}_x)^t\Sigma_x^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_x)$$

donde $\Sigma_x = \Lambda\Lambda^{\mathrm{t}} + \psi$

Como hemos visto antes, el MLE (maximum likelihood estimator) de $\boldsymbol{\mu}_x$ es: $\widehat{\boldsymbol{\mu}}_x = \bar{\mathbf{x}}$

Reemplazando esto en la ecuación:

$$l(\Sigma_x|\mathbf{x}, \widehat{\boldsymbol{\mu}}_x = \bar{\mathbf{x}}) = -\frac{np}{2}\log 2\pi - \frac{n}{p}\log|\Sigma_x| - \frac{n-1}{2}traza(\Sigma_x^{-1}S_x)$$

Ahora reemplazamos $\Sigma_x = \Lambda\Lambda^{\mathrm{t}} + \psi$:

$$l(\Lambda, \psi|\mathbf{x}) = -\frac{np}{2}\log 2\pi - \frac{n}{p}\log|\Lambda\Lambda^{\mathrm{t}} + \psi| - \frac{n-1}{2}traza((\Lambda\Lambda^{\mathrm{t}} + \psi)^{-1}S_x)$$

No hay forma explícita de obtener el MLE de Λ y de ψ, a no ser que se impongan algunas restricciones para la forma de esas matrices. Por lo general, lo que se utilizan son métodos de optimización para estimar los MLE's numéricamente. En cualquier caso, un estimador MLE es invariante ante transformaciones lineales de las variables. Así que las soluciones obtenidas para los datos originales y los datos escalados son equivalentes.

NÚMERO DE FACTORES

Para determinar el número de factores, asumiendo Normalidad, podemos usar el test de hipótesis del ratio de verosimilitudes para las hipótesis:

H_0: el número de factores es m

H_1: el número de factores no es m

El estadístico es igual a:

$$\lambda = n \log\left(\frac{\left|\widehat{\Lambda}\widehat{\Lambda}^{\mathrm{t}} + \hat{\psi}\right|}{\left|\hat{\Sigma}_x\right|}\right) - np + (n-1)\, traza\left(\left(\widehat{\Lambda}\widehat{\Lambda}^{\mathrm{t}} + \hat{\psi}\right)^{-1} S_x\right)$$

donde $\hat{\Sigma}_x$ es el MLE de Σ_x y S_x es la matriz de covarianza muestral.

El estadístico λ bajo la hipótesis nula H_0 tiene una distribución χ^2 con grados de libertad:

$$\frac{1}{2}\left((p-m)^2 - (p+m)\right)$$

La idea es aplicar el test secuencialmente, es decir, empezar con $m = 1$, y si el test es rechazado, considerar el caso siguiente $m = 2$, y así.

El número máximo de factores que podemos considerar tiene que verificar que:

$$(p-m)^2 - (p+m) > 0$$

Entonces existe un máximo número de factores que podemos usar bajo MLE.

El método de regresión de factores (regression factor scores) es usado usualmente con MLE para estimar los "scores" de los factores. Asume que los factores son variables aleatorias y busca un predictor lineal que minimice el error cuadrático medio (MSE) de predicción. El par $(\mathbf{f}_i, \mathbf{x}_i)$ tiene distribución Normal multivariante. Entonces es posible mostrar que el predictor lineal que minimiza el error cuadrático medio de predicción es:

$$E[\mathbf{f}_i|\mathbf{x}_i] = (I_m + \Lambda^t \psi^{-1} \Lambda)^{-1} \Lambda^t \psi^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_x)$$

El estimador final se obtiene remplazando $\boldsymbol{\mu}_x$, Λ y ψ por sus respectivas estimaciones MLE.

EJEMPLO

Consideremos el mismo ejemplo de los 50 estados de USA. Estos datos raramente serán Normales, pero apliquemos MLE para estimar el modelo factorial. Los valores del ratio de verosimilitudes, el estadístico λ son: 89.63, 46.53, 20.57 y 7.39. Los p-valores asociados son todos menores que 0.01 excepto el último que es igual a 0.0249. Si consideramos ese nivel de

significación podemos escoger tres factores, si ponemos un nivel más alto 0.05 por ejemplo podríamos coger incluso 4 factores. Sin embargo, estos datos no son Normales así que estos resultados deben tomarse con cuidado.

ESCALAMIENTO MULTIDIMENSIONAL

Las técnicas de escalado multidimensional (MDS) son una generalización de la idea de componentes principales.

En lugar de disponer de una matriz de observaciones por variables de tamaño $n \times p$, como en componentes principales, se dispone de una matriz, D, cuadrada $n \times n$ de distancias, valores de proximidad, similaridades, o disimilaridades entre los n elementos.

Por ejemplo, esa matriz D puede representar:

1. Las similitudes o distancias entre n productos fabricados por una empresa.
2. Las distancias percibidas entre n candidatos políticos.
3. Las diferencias entre n preguntas de un cuestionario o las distancias o similitudes entre n sectores industriales.

Estas distancias en D pueden haberse obtenido a partir de ciertas variables. O pueden ser el resultado de una estimación directa, por ejemplo, opiniones de ciertos jueces sobre las similaridades entre los elementos considerados. La dimensionalidad del problema es desconocida a priori.

El objetivo que se pretende es representar esta matriz mediante un conjunto de variables ortogonales $\mathbf{y}_1, \ldots, \mathbf{y}_p$, donde $p < n$. De manera que las distancias euclídeas entre las coordenadas de los elementos respecto a estas variables sean iguales (o lo más próximas posibles) a las distancias o disimilaridades de la matriz original. Es decir, a partir de la matriz D se pretende obtener una matriz X, de dimensiones $n \times p$, que pueda interpretarse como la matriz de p variables en los n individuos, y donde la distancia euclídea entre los elementos reproduzca, aproximadamente, la matriz de distancias D inicial. Cuando $p > 2$, las variables pueden

ordenarse en importancia y suelen hacerse representaciones gráficas en dos y tres dimensiones para entender la estructura existente.

En general no es siempre posible encontrar p variables que reproduzcan exactamente las distancias iniciales. Sin embargo, es frecuente encontrar variables que reproduzcan aproximadamente las distancia iniciales. Por otro lado, si la matriz de distancias se ha generado calculando las distancias euclídeas entre las observaciones definidas por ciertas variables, sí podremos recuperar las componentes principales de estas variables.

El escalado multidimensional comparte con componentes principales el objetivo de describir e interpretar los datos. Si existen muchos elementos, la matriz de similaridades será muy grande y la representación por unas pocas variables de los elementos nos permitirá entender su estructura: qué elementos tienen propiedades similares, si aparecen grupos entre los elementos, si hay elementos atípicos, etc. Además, si podemos interpretar las variables aumentará nuestro conocimiento del problema, al entender cómo se han generado los datos.

El escalado multidimensional representa un enfoque complementario a componentes principales en el sentido siguiente. Componentes principales considera la matriz $p \times p$ de correlaciones (o covarianzas) entre variables, e investiga su estructura. El escalado multidimensional considera la matriz $n \times n$ de correlaciones (o covarianzas) entre individuos, e investiga su estructura. Ambos enfoques están claramente relacionados, y existen técnicas gráficas, como el biplot que estudiaremos en esta sección, que aprovechan esta dualidad para representar conjuntamente las variables y los individuos en un mismo gráfico.

Los métodos existentes se dividen en:

- Métricos, cuando la matriz inicial es propiamente de distancias.
- No métricos, cuando la matriz es de similaridades.

Los métodos métricos, también llamados coordenadas principales, utilizan las diferencias entre similitudes.

Los no métricos parten de que, si A es más similar a B que a C, entonces A está más cerca de B que de C, pero las diferencias entre las similitudes AB y AC no tienen interpretación.

ESCALADO MÉTRICO O COORDENADAS PRINCIPALES

Dada una matriz de datos, $\mathbf{x}$, de individuos por variables, de tamaño $n \times p$, podemos obtener los datos centrados (media cero) mediante:

$$\tilde{\mathbf{x}} = \left(\boldsymbol{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^{\boldsymbol{t}} \right) \mathbf{x}$$

A partir de $\tilde{\mathbf{x}}$ podemos construir dos tipos de matrices cuadradas y semi-definida positivas:

1. La matriz de covarianzas, definida por $\tilde{\mathbf{x}}^t \tilde{\mathbf{x}}/n$
2. La matriz de productos cruzados, $Q = \tilde{\mathbf{x}} \tilde{\mathbf{x}}^t$
3. La última, Q, puede interpretarse como una matriz de similitud (covarianzas) entre los n elementos.

En efecto, los términos de la matriz Q, q_{ij}, contienen el producto escalar por pares de elementos:

$$\mathrm{q_{ij}} = \sum_{\mathrm{s=1}}^{\mathrm{p}} \tilde{\mathrm{x}}_{\mathrm{is}} \tilde{\mathrm{x}}_{\mathrm{js}} = \tilde{\mathrm{x}}_{\mathrm{i}}^{\mathrm{t}} \tilde{\mathrm{x}}_{\mathrm{j}}$$

donde $\tilde{\mathrm{x}}_{\mathrm{i}}^{\mathrm{t}}$ es la fila i de la matriz $\tilde{\mathrm{x}}$.

El producto escalar tiene la siguiente expresión alternativa:

$$\mathrm{q_{ij}} = |\tilde{\mathrm{x}}_{\mathrm{i}}| \left| \tilde{\mathrm{x}}_{\mathrm{j}} \right| \cos \theta_{\mathrm{ij}}$$

1. Si los dos elementos tienen coordenadas similares, $cos\ \theta_{ij} \approx 1$ y q_{ij} será grande.
2. Si los dos elementos son muy distintos, $cos\ \theta_{ij} \approx 0$ y q_{ij} será pequeño.

En este sentido podemos interpretar la matriz $Q = \tilde{\mathbf{x}}\tilde{\mathbf{x}}^t$ como la matriz de similitud entre elementos.

Las distancias entre las observaciones se deducen inmediatamente de la matriz de similitud.

La distancia euclídea al cuadrado entre dos elementos es:

$$d_{ij}^2 = \sum_{s=1}^{p} \left(\tilde{x}_{is} - \tilde{x}_{js}\right)^2 = \sum_{s=1}^{p} \tilde{x}_{is}{}^2 + \sum_{s=1}^{p} \tilde{x}_{js}{}^2 - 2\sum_{s=1}^{p} \tilde{x}_{is}\tilde{x}_{js}$$

La distancia d_{ij}^2 puede calcularse en función de la matriz Q:

$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij}$$

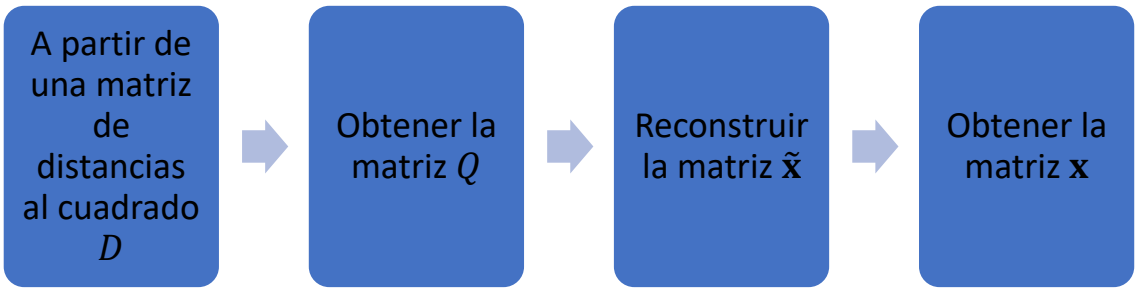
Por tanto, dada la matriz $\tilde{\mathbf{x}}$ de datos centrados podemos construir la matriz de similitud $Q = \tilde{\mathbf{x}}\tilde{\mathbf{x}}^t$ y a partir de ella, la matriz D de distancias al cuadrado entre dos elementos.

En forma matricial:

$$D = diag(Q)1^t + 1\, diag(Q)^t - 2Q$$

donde $diag(Q)$ es el vector que contiene los términos de la diagonal de la matriz Q y $\mathbf{1}$ es el vector de unos.

El problema que vamos a abordar es el inverso:



En primer lugar, observemos que no hay pérdida de generalidad en suponer que las variables tienen media cero.

Esto es consecuencia de que las distancias entre dos puntos no varían si expresamos las variables en desviaciones hacia la media:

$$d_{ij}^2 = \sum_{s=1}^{\mathrm{p}} \left(\tilde{\mathrm{x}}_{\mathrm{is}} - \tilde{\mathrm{x}}_{\mathrm{js}}\right)^2 = \sum_{s=1}^{\mathrm{p}} \left((x_{\mathrm{is}} - \bar{x}_s) - (x_{\mathrm{is}} - \bar{x}_s)\right)^2 = \sum_{s=1}^{\mathrm{p}} \left(x_{\mathrm{is}} - x_{\mathrm{js}}\right)^2$$

Vamos a buscar una matriz $\tilde{\mathbf{x}}$ con variables de media cero.

Esto significa que $\tilde{\mathbf{x}}^{\mathrm{t}}\mathbf{1} = \mathbf{0}$

Porque desglosando el vector de medias de la matriz $\tilde{\mathbf{x}}$:

$$\begin{pmatrix} \overline{\tilde{\mathrm{x}}_{:,1}} \\ \vdots \\ \overline{\tilde{\mathrm{x}}_{:,\mathrm{p}}} \end{pmatrix} = \frac{1}{n}\begin{pmatrix} \tilde{\boldsymbol{x}}_{11} + \tilde{\boldsymbol{x}}_{21} + \cdots + \tilde{\boldsymbol{x}}_{n1} \\ \vdots \\ \tilde{\boldsymbol{x}}_{1p} + \tilde{\boldsymbol{x}}_{2p} + \cdots + \tilde{\boldsymbol{x}}_{np} \end{pmatrix} = \frac{1}{n}\begin{pmatrix} \tilde{\boldsymbol{x}}_{11} & \cdots & \tilde{\boldsymbol{x}}_{1p} \\ \vdots & \ddots & \vdots \\ \tilde{\boldsymbol{x}}_{n1} & \cdots & \tilde{\boldsymbol{x}}_{np} \end{pmatrix}^t \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \frac{1}{\mathrm{n}}\tilde{\mathrm{x}}^{\mathrm{t}}1$$

Si el vector de medias es el vector de ceros $\mathbf{0}$ de tamaño $p \times 1$, entonces

$$\frac{1}{\mathrm{n}}\tilde{\mathrm{x}}^{\mathrm{t}}1 = \mathbf{0}$$

Y consecuentemente: $\tilde{\mathbf{x}}^{\boldsymbol{t}}\mathbf{1} = \mathbf{0}$ porque $n \neq 0$.

La condición: $\tilde{\mathbf{x}}^{\boldsymbol{t}}\mathbf{1} = \mathbf{0}$ implica que también $Q\mathbf{1} = \mathbf{0}$ porque:

$$Q1 = \tilde{\mathrm{x}}\tilde{\mathrm{x}}^{\mathrm{t}}1 = \tilde{\mathrm{x}}0 = \mathbf{0}$$

La condición $Q\mathbf{1} = \mathbf{0}$ quiere decir que la suma de todos los elementos de una fila de la matriz de similitudes Q (y de una columna porque Q es simétrica), debe ser cero.

$$Q1 = \begin{pmatrix} q_{11} + q_{12} + \cdots + q_{1n} \\ \vdots \\ q_{n1} + q_{n2} + \cdots + q_{nn} \end{pmatrix} = \begin{pmatrix} q_{11} + q_{21} + \cdots + q_{n1} \\ \vdots \\ q_{1n} + q_{2n} + \cdots + q_{nn} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

Sabemos, por lo visto anteriormente que la distancia d_{ij}^2 puede calcularse en función de la matriz Q:

$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij}$$

Para imponer la restricción $Q\mathbf{1} = \mathbf{0}$ sumemos la ecuación anterior por filas:

$$\sum_{i=1}^{n} d_{ij}^2 = \sum_{i=1}^{n} q_{ii} + \sum_{i=1}^{n} q_{jj} - 2\sum_{i=1}^{n} q_{ij} = traza(Q) + nq_{jj}$$

donde hemos utilizado que:

- $Q\mathbf{1} = \mathbf{0}$ implica $\sum_{i=1}^{n} q_{ij} = 0$
- $\sum_{i=1}^{n} q_{ii} = traza(Q)$
- $\sum_{i=1}^{n} q_{jj}$ lo de dentro no depende de i por lo tanto se suma esa cantidad n veces.

Sumando la ecuación por columnas:

$$\sum_{j=1}^{n} d_{ij}^2 = traza(Q) + nq_{ii}$$

Y sumando lo anterior por filas de nuevo:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 = \sum_{i=1}^{n}(traza(Q) + nq_{ii}) = nt + nt = 2nt$$

Ya que:

$t = traza(Q)$ es un número y $\sum_{i=1}^{n} q_{ii} = traza(Q) = t$

Por lo anterior sabemos que:

$$q_{jj} = \frac{\sum_{i=1}^{n} d_{ij}^2 - t}{n}$$

$$q_{ii} = \frac{\sum_{j=1}^{n} d_{ij}^2 - t}{n}$$

Sustituyendo q_{ii} y q_{jj} en:

$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij} = \frac{\sum_{j=1}^{n} d_{ij}^2 - t}{n} + \frac{\sum_{i=1}^{n} d_{ij}^2 - t}{n} - 2q_{ij}$$
$$= \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 - \frac{t}{n} + \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 - \frac{t}{n} - 2q_{ij}$$

Denotemos las medias por columnas, por filas, y la media de todos los elementos de D:

$$d_{.j}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2$$

$$d_{i.}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2$$

$$d_{..}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2$$

Y vamos a usar que $\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 = 2nt$ implica que $t = \frac{n}{2} d_{..}^2$

Y que esto a su vez implica que

$$-\frac{2}{n}t = d_{..}^2$$

Sustituyendo en la ecuación de los elementos de D:

$$d_{ij}^2 = d_{i.}^2 + d_{.j}^2 - \frac{2}{n}t - 2q_{ij} = d_{i.}^2 + d_{.j}^2 - d_{..}^2 - 2q_{ij}$$

Ahora podemos hallar los elementos de la matriz Q en base a los elementos de la matriz D:

$$q_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2\right)$$

De esta expresión:

$$q_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2\right)$$

Puede comprobarse que:

$$Q = -\frac{1}{2}\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^t\right) D \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^t\right) = -\frac{1}{2} PDP$$

donde P es la matriz de proyección:

$$P = I - \frac{1}{n}\mathbf{1}\mathbf{1}^t$$

Pasemos ahora al problema de obtener la matriz $\mathbf{x}$ dada la matriz Q.

Suponiendo que la matriz de similitud es definida positiva de rango p, puede representarse por la descomposición en valores y vectores propios:

$$Q = \Lambda V \Lambda^{\mathrm{t}}$$

La matriz V es de tamaño $n \times p$ y contiene los vectores propios correspondientes a valores propios no nulos de Q.

La matriz Λ es diagonal de tamaño $p \times p$ y contiene los valores propios.

La matriz V^t es la transpuesta de V y es de tamaño $p \times n$.

Escribiendo:

$$Q = \left(V\Lambda^{1/2}\right)\left(\Lambda^{1/2}V^t\right)$$

Y tomando:

$$Y = V\Lambda^{1/2}$$

Queda:

$$Q = YY^t$$

¿Es Y la matriz que buscábamos $\tilde{\mathbf{x}}$ o $\mathbf{x}$?

- Hemos obtenido una matriz Y de tamaño $n \times p$ con p variables incorreladas que reproducen la métrica inicial.
- Pero sabemos que $Q = \tilde{\mathbf{x}}\tilde{\mathbf{x}}^t = \tilde{\mathbf{x}}AA^t\tilde{\mathbf{x}}^t = \tilde{\mathbf{x}}A(\tilde{\mathbf{x}}\boldsymbol{A})^{\boldsymbol{t}}$ para cualquier matriz A ortogonal, porque esta matriz es invariante ante rotaciones de las variables.
- Por lo tanto, de D solo es posible obtener una rotación de los datos $\tilde{\mathbf{x}}$ dada por la matriz Y la cual llamaremos matriz de coordenadas principales.

CONSTRUCCIÓN DE LAS COORDENADAS PRINCIPALES

Para poder construir $Q = \left(V\Lambda^{1/2}\right)\left(\Lambda^{1/2}V^t\right)$ es necesario calcular la raíz cuadrada de la matriz Λ que contiene los valores propios de Q, entonces estos deberían ser no-negativos.

Diremos que una matriz de distancias D es compatible con una métrica euclídea si la matriz de similitud Q que se obtiene a partir de ella es semi-definida positiva.

Esta condición es necesaria y suficiente, es decir:

1. Si D se ha construido a partir de una métrica euclídea Q es no negativa
2. Si Q es no negativa es posible encontrar una métrica euclídea que reproduzca D.

En general, la matriz de distancias no tiene por qué ser compatible con una métrica euclídea, pero es frecuente que la matriz de similitud obtenida a partir de ella tenga p valores propios positivos y más grandes que el resto. Si los restantes $n - p$ valores propios no nulos son mucho menores que los demás, podemos obtener una representación aproximada de los puntos utilizando los p vectores propios asociados a p valores propios positivos de la matriz de similitud. En este caso, las representaciones gráficas conservarán sólo aproximadamente la distancia entre los puntos.

Supongamos que tenemos una matriz de distancias al cuadrado D. El procedimiento para obtener las coordenadas principales es:

1. Construir $Q = -\frac{1}{2}PDP$.
2. Obtener los valores propios de Q. Tomar los r mayores valores propios, donde r se escoge de manera que los restantes $n - r$ valores propios sean próximos a cero.
3. Obtener las coordenadas de los puntos en las variables mediante $\mathbf{v}_i\sqrt{\lambda_i}$ donde λ_i es un valor propio de Q y $\mathbf{v}_i$ su vector propio asociado. Esto implica aproximar Q por:

$$Q \approx \left(V_r \Lambda_r^{1/2}\right)\left(\Lambda_r^{1/2} V_r^t\right)$$

1. Definir la matriz de las r coordenadas principales:

$$Y_r = V_r \Lambda_r^{1/2}$$

Se puede definir una medida de precisión para este ajuste a partir de los r autovalores positivos:

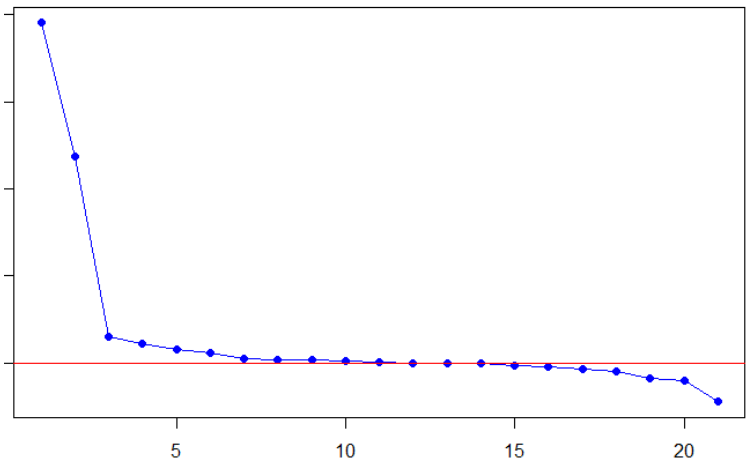
$$m_r = \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{n} |\lambda_i|}$$

La cantidad $100 \times m_r$ nos da un porcentaje de lo bien que representa la matriz estimada de coordenadas principales a los datos.

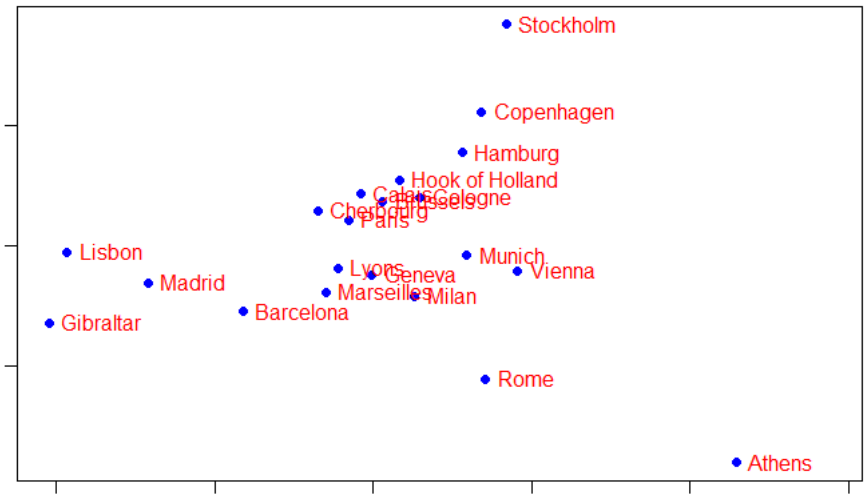
EJEMPLO

Tenemos una matriz D que contiene las distancias (en km) entre 21 ciudades de Europa. El problema que queremos resolver es crear un mapa con el que se han generado estas distancias. Si obtenemos Q a partir de D como vimos anteriormente y calculamos los autovalores de Q, vemos que podríamos coger los dos primeros autovalores $r = 2$, para con ellos construir la matriz de coordenadas principales. Después de los dos primeros autovalores los que quedan son cercanos a cero, además notemos que hay autovalores negativos (por debajo de la línea roja)

Después de los dos primeros autovalores los que quedan son cercanos a cero, además notemos que hay autovalores negativos (por debajo de la línea roja)



Al hacer la representación bidimensional de los datos en las coordenadas estimadas vemos lo similar que es a la representación de las ciudades en un mapa real.



La medida de precisión en este ejemplo es igual a:

$$m_2 = \frac{\sum_{i=1}^{2} \lambda_i}{\sum_{i=1}^{21} |\lambda_i|} = 0.7537$$

Lo cual nos dice que la representación en estas coordenadas paralelas de los datos de los que provienen las distancias que teníamos, explica un 75% de su variabilidad. Es una representación adecuada.

Cuando los datos originales forman una matriz $\tilde{\mathbf{x}}$ de individuos por variables y construimos la matriz D de distancias utilizando las distancias euclídeas entre los puntos a partir de dichas variables originales, las coordenadas principales obtenidas de la matriz D son equivalentes a los componentes principales de las variables.

En efecto, con variables de media cero los componentes principales son los auto-vectores de $1/\mathrm{n}\,\tilde{\mathbf{x}}^t\tilde{\mathbf{x}}$, mientras que como hemos visto antes, las coordenadas principales son los auto-vectores estandarizados por $\sqrt{\lambda_i}$ correspondientes a los auto-valores de $Q = \tilde{\mathbf{x}}\tilde{\mathbf{x}}^t$.

Esto es porque $\tilde{\mathbf{x}}\tilde{\mathbf{x}}^t$ y $\tilde{\mathbf{x}}^t\tilde{\mathbf{x}}$ tienen el mismo rango y los mismos auto-valores no nulos.

El análisis en coordenadas principales o escalado multidimensional, está muy relacionado con componentes principales. En ambos casos tratamos de reducir la dimensionalidad de los datos.

En componentes principales partimos de la matriz $\tilde{\mathbf{x}}^t\tilde{\mathbf{x}}$, obtenemos sus valores propios, y luego proyectamos las variables sobre estas direcciones para obtener los valores de los componentes, que son idénticas a las coordenadas principales, que se obtienen directamente como vectores propios de la matriz $\tilde{\mathbf{x}}\tilde{\mathbf{x}}^t$.

Se puede comprobar que si $Q = \tilde{\mathbf{x}}\tilde{\mathbf{x}}^t$ el componente principal de $\mathbf{x}$, llamémosle Z_r, es proporcional a la coordenada principal $Y_r = aZ_r$ para algún valor a. Si la matriz de similaridades proviene de una métrica euclídea ambos métodos conducirán al mismo resultado. Sin embargo, el concepto de coordenadas principales puede aplicarse a una gama más amplia de problemas que PCA, ya que las coordenadas principales pueden obtenerse siempre, aunque las distancias de partida no hayan sido exactamente generadas a partir de variables. Esto lo veremos en el caso de escalado no métrico.

ESCALADO NO MÉTRICO

En los problemas de escalado no métrico se parte de una matriz de diferencias o disimilitudes entre objetos que se ha obtenido generalmente por consultas a jueces, o a partir de procedimientos de ordenación de los elementos.

Por ejemplo, el escalado no métrico se ha aplicado para:

- Estudiar las semejanzas entre las actitudes, preferencias o percepciones de personas sobre asuntos políticos o sociales.
- Evaluar las preferencias respecto a productos y servicios en marketing y en calidad.

Los valores de una tabla de similaridades o distancias se obtienen habitualmente por alguno de los procedimientos siguientes:

1. Estimación directa. Un juez, o un conjunto de jueces, estiman directamente las distancias entre los elementos. Una escala muy utilizada es la escala $0 - 100$, de manera que la distancia o disimilaridad entre un elemento y sí mismo sea cero y la distancia

entre dos elementos distintos refleje la percepción de sus diferencias. Con n elementos esto requiere $n(n\ -\ 1)/2$ evaluaciones.

2. Estimación de rangos. Se selecciona un elemento y se pide al juez, o grupo de jueces, que ordene los $n - 1$ restantes por mayor o menor proximidad al seleccionado. A continuación, se selecciona el siguiente y se ordenan los $n\ -\ 2$ restantes, y así sucesivamente. Existen algoritmos de cálculo que transforman estas ordenaciones en una matriz de distancias.

3. Rangos por pares. Se presentan al juez los $n(n\ -\ 1)/2$ pares posibles y se le pide que los ordene de mayor a menor distancia. Por ejemplo, con cuatro objetos supongamos que se obtienen los resultados en orden de distancia: (3,4), (2,3), (2,4), (1,4), (1,2) y (1,3). Entonces, los más próximos son los objetos 3 y 4, y a esta pareja se le asigna el rango 1. A la pareja siguiente, (2,3), se le asigna rango 2 y así sucesivamente hasta la pareja de los elementos más alejados, el 1 y el 3, que reciben rango $n(n\ -\ 1)/2$, que en este caso es 6. A continuación se calcula un rango medio para cada objeto, promediando los rangos de los pares donde aparece. Por ejemplo, el objeto 1 aparece en pares que tienen rango 4, 5 y 6, con lo que el rango del objeto 1 es:

$$rango(1) = \frac{4 + 5 + 6}{3} = 5$$

Igualmente obtenemos que:

$$rango(2) = \frac{2 + 3 + 5}{3} = 3$$

Y los demás $rango(3) = 3$ y $rango(4) = 2.7$

Las diferencias entre los rangos se toman ahora como medidas de distancia entre los objetos.

Se supone que la matriz de similaridades está relacionada con una matriz de distancias, pero de una manera compleja.

Es decir, se acepta que los jueces utilicen en las valoraciones ciertas variables o dimensiones, pero los datos incluyen elementos de error y variabilidad personal.

Por tanto, las variables que explican las similitudes entre los elementos comparados determinarán una distancias euclídeas entre ellos, d_{ij}, que están relacionadas con las similitudes dadas, δ_{ij}, mediante una función desconocida:

$$\delta_{ij} = f(d_{ij})$$

Las similitudes están relacionadas con las distancias euclídeas mediante:

$$\delta_{ij} = f(d_{ij})$$

La única condición que se impone es que f sea una función monótona, es decir:

$$\delta_{ij} > \delta_{ih} \Leftrightarrow d_{ij} > d_{ih}$$

El objetivo que se pretende es encontrar unas coordenadas (usualmente se selecciona $r = 2$ para ver gráficamente) que sean capaces de reproducir estas distancias a partir únicamente de la condición de monotonía.

Para ello hay que definir:

1. Un criterio de bondad del ajuste que sea invariante ante transformaciones monótonas de los datos.
2. Un algoritmo para obtener las coordenadas, optimizando el criterio establecido.

CONSTRUCCIÓN DE LAS COORDENADAS PRINCIPALES

Para ello el método usual es el algoritmo de Shepard-Kruskal:

1. Usar el MDS métrico en la matriz de disimilaridades D para obtener un conjunto inicial de coordenadas principales.
2. Calcular las distancias euclídeas entre las coordenadas obtenidas.

3. Hacer una regresión entre los valores de las distancias euclídeas y las disimilaridades, teniendo en cuenta la restricción de monotonicidad.

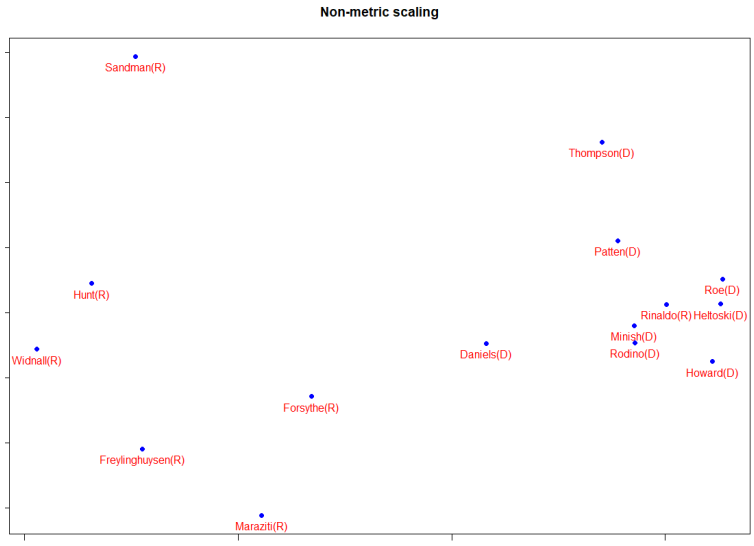
4. Comparar las distancias euclídeas originales con los valores predichos $\hat{d}_{ij}$ usando la regresión anterior, con el método STRESS:

$$STRESS = \frac{\sum_{i<j}\left(\delta_{ij}^2 - \hat{d}_{ij}\right)^2}{\sum_{i<j}\delta_{ij}^2}$$

5. Remplazar las distancias euclídeas originales con las predichas y repetir el proceso a partir del punto (3) hasta que el valor de STRESS sea muy pequeño.

EJEMPLO

- Vamos a considerar un conjunto de datos que muestra la cantidad de veces que 15 congresistas de Nueva Jersey votaron de manera diferente en la Cámara de Representantes sobre 19 proyectos de ley ambientales.
- Las abstenciones no se registran.
- La pregunta es si las afiliaciones de los partidos se pueden detectar en los datos.
- Aplicamos un escalamiento no métrico al comportamiento de votación que se muestra en el conjunto de datos.
- Un gráfico bidimensional del resultado nos ayudaría a comprender si el comportamiento de votación es similar al objetivo de los partidos.
- La respuesta es que sí, aunque hay más variación entre los republicanos.
- El comportamiento de votación de uno de los republicanos (Rinaldo) parece estar más cerca de sus colegas demócratas que del comportamiento de votación de otros republicanos.



CLUSTERIZACIÓN

El análisis de clúster o conglomerados tiene como objetivo agrupar elementos en grupos homogéneos en función de las similitudes o similaridades entre ellos. Normalmente se agrupan las observaciones, aunque puede también aplicarse para agrupar variables. Estos métodos se conocen también con el nombre de métodos de clasificación automática o no supervisada, o de reconocimiento de patrones sin supervisión.

MÉTODOS

- Partición de los datos:
  - Algoritmo k-medias
- Métodos jerárquicos:
  - Métodos aglomerativos
  - Métodos divisivos
- Clúster basado en modelos

¿QUÉ ESTUDIA?

El análisis de conglomerados estudia tres tipos de problemas:

Partición de los datos: Disponemos de datos que sospechamos son heterogéneos y se desea dividirlos en un número de grupos prefijado, de manera que:

1. cada elemento pertenezca a uno y solo uno de los grupos.
2. todo elemento quede clasificado.
3. cada grupo sea internamente homogéneo.

Por ejemplo: se dispone de una base de datos de compras de clientes y se desea hacer una tipología de estos clientes en función de sus pautas de consumo.

Construcción de jerarquías: Deseamos estructurar los elementos de un conjunto de forma jerárquica por su similitud.

Por ejemplo: tenemos una encuesta de atributos de distintas profesiones y queremos ordenarlas por similitud. Una clasificación jerárquica implica que los datos se ordenan en niveles, de manera que los niveles superiores contienen a los inferiores. Este tipo de clasificación es muy frecuente en biología, al clasificar animales, plantas etc. Estrictamente, estos métodos no definen grupos, sino la estructura de asociación en cadena que pueda existir entre los elementos. Sin embargo, como veremos, la jerarquía construida permite obtener también una partición de los datos en grupos.

Clúster basado en modelos: En este caso, se parte de la hipótesis de que los datos han sido generados de una mixtura de distribuciones. Entonces lo que hay que estimar son los parámetros desconocidos de esta mixtura usando el método de máxima verosimilitud. Las observaciones son asignadas al grupo que tiene mayor probabilidad de haberlas generado.

¿QUÉ NECESITAMOS?

Los métodos de partición utilizan la matriz de datos. Los algoritmos jerárquicos utilizan la matriz de distancias o similitudes entre elementos. Los

algoritmos basados en un modelo utilizan algoritmos de optimización para estimar los parámetros desconocidos.

K-MEDIAS

Supongamos que tenemos una muestra de n elementos con p variables. El objetivo es dividir esta muestra en un número de grupos prefijado, K.

El algoritmo de K −medias requiere las cuatro etapas siguientes:

1. Seleccionar K puntos como centros de los grupos iniciales. Esto puede hacerse:

    a) asignando aleatoriamente los objetos a los grupos y tomando los centros de los grupos formados

    b) tomando como centros los K puntos más alejados entre sí

    c) construyendo los grupos con información a priori, o bien seleccionando los centros a priori.

2. Calcular las distancias euclídeas de cada elemento al centro de los K grupos, y asignar cada elemento al grupo más próximo. La asignación se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalculan las coordenadas de la nueva media de grupo.

3. Definir un criterio de optimalidad y comprobar si reasignando uno a uno cada elemento de un grupo a otro mejora el criterio.

4. Si no es posible mejorar el criterio de optimalidad, terminar el proceso.

IMPLEMENTACIÓN DEL ALGORITMO

El criterio de homogeneidad que se utiliza en el algoritmo de K −medias es la *suma de cuadrados dentro de los grupos (SCDG)* para todas las variables,

que es equivalente a la suma ponderada de las varianzas de las variables en los grupos:

$$SCDG = \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{i=1}^{n_k} \left(x_{ijk} - \bar{\mathbf{x}}_{jk}\right)^2$$

donde x_{ijk} es el valor de la variable j en el elemento i del grupo k, y $\bar{\mathbf{x}}_{jk}$ es la media de esta variable en el grupo.

El criterio se escribe:

$$\min SCDG = \min \sum_{k=1}^{K} \sum_{j=1}^{p} n_k s_{jk}^2$$

donde n_k es el número de elementos del grupo k y s_{jk}^2 es la varianza de la variable j en dicho grupo.

La varianza de cada variable en cada grupo es claramente una medida de la heterogeneidad del grupo y al minimizar las varianzas de todas las variables en los grupos obtendremos grupos más homogéneos.

Un posible criterio alternativo de homogeneidad sería minimizar las distancias al cuadrado entre los centros de los grupos y los puntos que pertenecen a ese grupo. Si medimos las distancias con la norma euclídea, este criterio se escribe:

$$\min \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{\mathbf{x}}_k)^t (x_{ik} - \bar{\mathbf{x}}_k) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} d^2(i, g)$$

donde $d^2(i, g)$ es el cuadrado de la distancia euclídea entre el elemento i del grupo g y su media de grupo.

Es fácil comprobar que ambos criterios son idénticos. Sabemos que un escalar es igual a su traza, entonces podemos escribir el último criterio como:

$$\min \sum_{k=1}^{K} \sum_{i=1}^{n_k} traza[d^2(i, g)] = traza\left[\sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{\mathbf{x}}_k)^t (x_{ik} - \bar{\mathbf{x}}_k)^t\right]$$

Y llamando W a la matriz de SCDG:

$$W = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{\mathbf{x}}_k)(x_{ik} - \bar{\mathbf{x}}_k)^t$$

Tenemos que

$$\min traza(W) = \min SCDG$$

Como la traza es la suma de los elementos de la diagonal principal ambos criterios coinciden.

Este criterio se denomina criterio de la traza, y fue propuesto por Ward (1963). El criterio de la traza tiene dos propiedades importantes. La primera es que no es invariante ante cambios de medida en las variables. Cuando las variables vayan en unidades distintas conviene estandarizarlas, para evitar que el resultado del algoritmo dependa de cambios irrelevantes en la escala de medida. Cuando vayan en las mismas unidades suele ser mejor no estandarizar, ya que es posible que una varianza mucho mayor que el resto sea precisamente debida a que existen dos grupos de observaciones en esa variable, y si estandarizamos podemos ocultar la presencia de los grupos.

La segunda propiedad del criterio de la traza es que minimizar la distancia euclídea produce grupos aproximadamente esféricos. Por otro lado, este criterio está pensado para variables cuantitativas. Aunque puede aplicarse si existe un pequeño número de variables binarias, si una parte importante de las variables son atributos, es mejor utilizar los métodos jerárquicos.

La maximización de este criterio requeriría calcularlo para todas las posibles particiones en el número de grupos especificado. Esto es computacionalmente muy costoso, salvo para valores de n muy pequeños. Por eso, sólo encontraremos mínimos locales de SCDG, con lo cual, se recomienda aplicar el algoritmo usando diferentes configuraciones iniciales.

El algoritmo de k–medias busca la partición óptima con la restricción de que en cada iteración sólo se permite mover un elemento de un grupo a otro. El algoritmo funciona como sigue:

1. Partir de una asignación inicial.

2. Comprobar si moviendo algún elemento se reduce W.

3. Si es posible reducir W, mover el elemento, recalcular las medias de los dos grupos afectados por el cambio y volver al punto 2. Si no es posible reducir W terminar.

En consecuencia, el resultado del algoritmo puede depender de la asignación inicial y del orden de los elementos. Por eso siempre conviene repetir el algoritmo con distintos valores iniciales y permutando los elemento de la muestra.

NÚMERO DE GRUPOS

En la aplicación habitual del algoritmo de K −medias hay que fijar el número de grupos, K. Obviamente, este número no puede estimarse con un criterio de homogeneidad ya que la forma de conseguir grupos muy homogéneos y minimizar la $SCDG$ es hacer tantos grupos como observaciones, con lo que siempre $SCDG = 0$. Se han propuesto distintos métodos para seleccionar el número de grupos.

Un procedimiento aproximado que se utiliza bastante es realizar un test F aproximado de reducción de variabilidad.

Consiste en comparar la $SCDG$ de K grupos con la de $K + 1$, y calcular la reducción proporcional de variabilidad que se obtiene aumentando un grupo adicional.

El test es:

$$F = \frac{SCDG(K) - SCDG(K+1)}{SCDG(K+1)/(n-K-1)}$$

Se está comparando la disminución de variabilidad al aumentar un grupo con la varianza promedio. El valor de F suele compararse con una distribución $F_{p,\ p(n-K-1)}$. Esta regla no está muy justificada porque los datos no tienen por qué verificar las hipótesis necesarias para aplicar la distribución F. Una regla empírica que da resultados razonables, sugerida por Hartigan (1975), e implantada en algunos programas informáticos, es introducir un grupo más si este cociente es mayor que 10.

K-MEDOIDES

La partición alrededor de los “medoides” (se suele llamar también Partition around medoids - PAM) es otro algoritmo de partición. Esencialmente, PAM es una modificación del algoritmo $K-$medias. Este algoritmo busca K "objetos representativos" en lugar de los centroides entre las observaciones en el conjunto de datos. Entonces, se espera que el método sea más robusto ante anomalías o atípicos. Una desventaja del algoritmo es que, aunque funciona bien en pequeños conjuntos de datos, no son lo suficientemente eficientes como para utilizarlos para agrupar grandes conjuntos de datos.

El algoritmo es el siguiente:

1. Sea $\mathbf{x}_i$ para $i = 1, \dots, n$ el conjunto de observaciones de la matriz de datos.

2. Calcular $D = \{d_{ij},\ tal\ que\ i, j = 1, \dots, n\}$, una matriz que contiene las distancias entre las n observaciones.

3. Elegir K observaciones como los “medoides” de los K grupos iniciales.

4. Asignar cada observación a su “medoide” más cercano usando la matriz de distancias D.

5. Para cada clúster, buscar la observación $\mathbf{x}_j$ (si existe) que proporciona la mayor reducción de

$$SCDG = \sum_{k=1}^{K} \sum_{c(i)=k} d_{ij}^2$$

   donde $c(i)$ es el clúster que contiene a $\mathbf{x}_i$

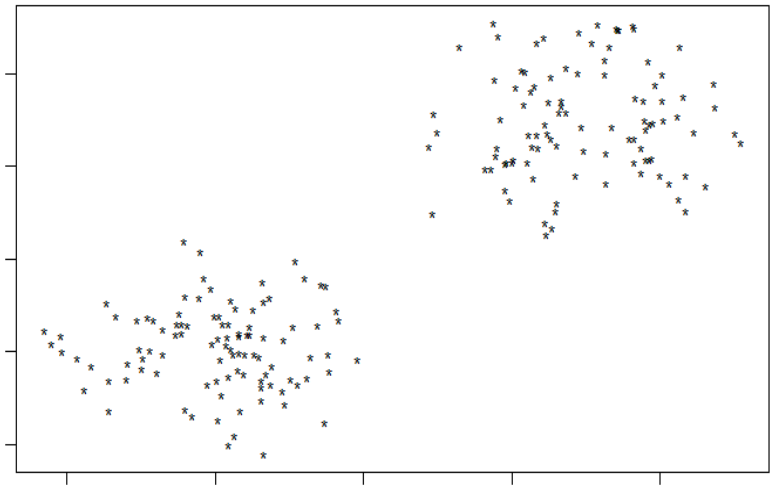
   Seleccionar la observación $\mathrm{x_j}$ que reduce $SCDG$ como el nuevo “medoide” para ese clúster (note que $SCDG$ sólo considera las distancias entre las observaciones del clúster en cuestión y su medoide).

6. Repetir pasos (4) y (5) hasta que no se reduzca $SCDG$.

MÉTODOS JERÁRQUICOS

Los métodos jerárquicos parten de una matriz de distancias o similaridades entre los elementos de la muestra y construyen una jerarquía basada en las distancias. Si todas las variables son continuas, la distancia más utilizada es la euclídea entre las variables estandarizadas. En general, no es recomendable utilizar las distancias de Mahalanobis, ya que la única matriz de covarianzas disponible es la de toda la muestra, y puede mostrar unas correlaciones muy distintas de las que existen entre las variables dentro de los grupos.

Por ejemplo, generemos dos grupos de variables normales independientes de medias (0,0) y (5,5) y varianza la matriz identidad. La posición de los grupos genera en el conjunto de puntos una correlación positiva fuerte, que desaparece si consideramos cada uno de los grupos por separado.



Correlación grupo 1 ---> 0.0009943199

Correlación grupo 2 ---> 0.05546129

Correlación grupo 1 + 2 ---> 0.8614679

Hay dos tipos de métodos jerárquicos para Clusterización:

1. **Aglomerativos:** empezamos con n (el número de observaciones) clústeres y los vamos agrupando en menos cantidad de clústeres que

serán de mayor tamaño porque los iniciales solo contienen una observación.

2. Divisivos: empezamos con un solo clúster que contiene a todas las observaciones, y vamos dividiendo en clústeres más pequeños.

DENDOGRAMA

Al final del análisis jerárquico, nos encontraremos con un gráfico llamado dendograma. Los resultados dependen de las distancias consideradas. En particular, si los datos tienen diferentes unidades de medida, y la distancia usada no tiene en cuenta esto, es mejor estandarizar las variables.

ESTANDARIZAR

Para decidir si estandarizar las variables o no antes del análisis conviene tener en cuenta el objetivo del estudio. Si no estandarizamos, la distancia euclídea dependerá sobre todo de las variables con valores más grandes, y el resultado del análisis puede cambiar completamente al modificar su escala de medida. Si estandarizamos, estamos dando a priori un peso semejante a las variables, con independencia de su variabilidad original, lo que puede no ser siempre adecuado.

VARIABLES BINARIAS

Cuando en la muestra existen variables continuas y atributos el problema se complica. Supongamos que la variable $\mathbf{x}_1$ es binaria. La distancia euclídea entre dos elementos de la muestra en función de esta variable es $(\mathbf{x}_{i1} - \mathbf{x}_{h1})^2$ que tomará valor cero si $\mathbf{x}_{i1} = \mathbf{x}_{h1}$, es decir, cuando el atributo está o no en ambos elementos (o sea, cuando valen cero o valen uno al mismo tiempo). Valdrá 1 si una vale cero y la otra 1, da igual el orden, es decir, el atributo solo está en uno de los elementos. Sin embargo, la distancia entre dos elementos correspondiente a una variable continua estandarizada, será $(\mathbf{x}_{i1} - \mathbf{x}_{h1})^2/s_1^2$ que puede ser mucho mayor que $\mathbf{1}$. Con lo que las variables continuas van en general a pesar mucho más que las binarias. Esto puede ser aceptable en muchos casos, pero cuando, por la

naturaleza del problema, esta situación no sea deseable, la solución es trabajar con similaridades.

SIMILARIDADES

El coeficiente de similaridad según la variable $j = 1, \dots, p$ entre dos elementos muestrales i y h se define como una función s_{jih} no negativa y simétrica.

1. $s_{jii} = 1$
2. $0 \leq s_{jih} \leq 1$
3. $s_{jih} = s_{jhi}$

Si obtenemos las similaridades para cada variable entre dos elementos podemos combinarlas en un coeficiente de similaridad global entre los dos elementos.

El coeficiente de similaridad propuesto por Gower es:

$$s_{ih} = \frac{\sum_{j=1}^{p} w_{jih}\, s_{jih}}{\sum_{j=1}^{p} w_{jih}}$$

donde $\mathrm{w_{jih}}$ es una variable ficticia que es igual a uno si la comparación de estos dos elementos muestrales i y h mediante la variable j tiene sentido, y será cero si no queremos incluir esa variable en la comparación entre los elementos.

Por ejemplo, si la variable $\mathrm{x_1}$ es: 1 si una persona pide un crédito, 0 si no, y la variable $\mathrm{x_2}$ es: 1 si lo devuelve y 0 si no. Si una persona no ha pedido crédito $(\mathrm{x_1 = 0})$ no tiene sentido preocuparse por $\mathrm{x_2}$. EN este caso, al comparar los individuos $(\mathrm{i, h})$ si uno cualquiera de los dos tiene un valor cero en $\mathrm{x_1}$, asignaremos a la variable $\mathrm{w_{2ih}}$ el valor cero.

Las similaridades entre elementos en función de las variables cualitativas pueden construirse individualmente o por bloques.

La similaridad entre dos elementos por una variable binaria será uno, si ambos tienen el atributo, y cero en caso contrario.

Alternativamente, podemos agrupar las variables binarias en grupos homogéneos y tratarlas conjuntamente.

Si suponemos que todos los atributos tienen el mismo peso, podemos construir una medida de similaridad entre dos elementos A y B respecto a todos estos atributos contando el número de atributos que están presentes:

a) En ambos

b) En A y no en B

c) En B y no en A

d) En ninguno de los dos elementos

Estas cuatro cantidades forman una tabla de asociación entre elementos, y servirán para construir medidas de similitud o similaridad entre los dos elementos comparados.

En esta tabla se verifica que $n_\alpha = a + b + c + d$, donde n_α es el número de atributos.

Por ejemplo, la siguiente tabla presenta una posible matriz de datos con siete atributos binarios y con ella vamos a construir una tabla de asociación que presenta la distribución conjunta de los valores 0 y 1 para los elementos A y B.

Elementos	variables (atributos)						
	x_1	x_2	x_3	x_4	x_5	x_6	x_7
A	0	1	1	0	0	0	1
B	1	0	1	1	1	1	0
C	1	0	0	1	1	1	1
.	.	.	.	.	.	.	

El elemento A tiene 3 valores 1 en el conjunto de variables binarias y de estos tres casos, en una ocasión también el elemento B tiene el valor 1, y en otras dos tiene el valor 0. El elemento A toma 4 veces el valor 0, ninguna coincidiendo con B porque en este caso B toma el valor uno.

La tabla de asociación sería la siguiente:

		B		
		1	0	
A	1	1 (a)	2 (b)	3
	0	4 (c)	0 (d)	4
Suma		5	2	7

La suma de los totales de filas y columnas debe ser siempre el número de atributos binarios considerados.

Para calcular un coeficiente de similitud entre dos individuos a partir de su tabla de asociación se utilizan los dos criterios principales siguientes:

1. Proporción de coincidencias. Se calcula como el número total de coincidencias sobre el número de atributos totales:

$$s_{ij} = \frac{a+d}{n_\alpha}$$

   Por ejemplo, la similitud de A y B es $1/7$ y la de B y C es $5/7$.

2. Proporción de apariciones. Cuando la ausencia de un atributo no es relevante, podemos excluir las ausencias y calcular sólo la proporción de veces donde el atributo aparece en ambos elementos. El coeficiente se define por:

$$s_{ij} = \frac{a}{a+b+c}$$

   Por ejemplo, con este criterio la similitud entre A y B es también $1/7$, y la de B y C es $4/6$.

Aunque las dos propuestas anteriores son las más utilizadas puede haber situaciones donde sean recomendables otras medidas. Por ejemplo, podemos querer dar peso doble a las coincidencias, con lo que resulta:

$$s_{ij} = \frac{2(a+d)}{2(a+d)+b+c}$$

O también se podría tener en cuenta sólo las coincidencias y tomar:

$$s_{ij} = \frac{a}{b+c}$$

Finalmente, los coeficientes de similitud o similaridad entre dos elementos i y h para una variable continua j se construye mediante:

$$s_{jih} = 1 - \frac{|\mathbf{x}_{ij} - \mathbf{x}_{hj}|}{rango(\mathbf{x}_j)}$$

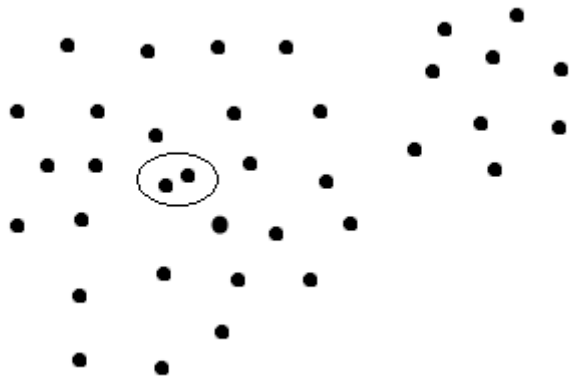
Donde $rango(\mathbf{x}_j) = \max(\mathbf{x}_j) - \min(\mathbf{x}_j)$.

De esta manera el coeficiente resultante estará siempre entre cero y uno. Una vez obtenida la similaridad global entre los elementos, podemos transformar los coeficientes en distancias. Lo más simple es definir la distancia mediante $d_{ih} = 1 - s_{ih}$, pero esta relación puede no verificar la propiedad triangular. Sin embargo, si definimos la distancia por: $d_{ih} = \sqrt{2(1 - s_{ih})}$ entonces sí se verifica.

MÉTODO AGLOMERATIVO

Los algoritmos aglomerativos (agnes: agglomerative nesting) que se utilizan tienen siempre la misma estructura y sólo se diferencian en la forma de calcular las distancias entre grupos. Su estructura es:

1. Comenzar con tantas clases como elementos, n. Las distancias entre clases son las distancias entre los elementos originales.



2. Seleccionar los dos elementos más próximos en la matriz de distancias y formar con ellos una clase.

3. Sustituir los dos elementos utilizados en el punto (2) para definir la clase, por un nuevo elemento que represente la clase construida. Las distancias entre este nuevo elemento y los anteriores se calculan con uno de los criterios para definir distancias entre grupos que comentaremos luego.

4. Volver a (2), y repetir (2) y (3) hasta que tengamos todos los elementos agrupados en una clase única.

CRITERIOS PARA DEFINIR DISTANCIAS ENTRE GRUPOS

Supongamos que tenemos un grupo A con n_a elementos, y un grupo B con n_b elementos, y que ambos se fusionan para crear un grupo (AB) con $n_a + n_b$ elementos.

La distancia del nuevo grupo, (AB), a otro grupo C con n_c elementos, se calcula habitualmente por alguna de las cinco reglas siguientes:

1. (Single linkage) Encadenamiento simple o vecino más próximo: La distancia entre los dos nuevos grupos es la menor de las distancias entre grupos antes de la fusión. Es decir:

$$d(C, AB) = \min(d_{CA}, d_{CB})$$

   Este criterio es invariante ante transformaciones monótonas. Tiende a producir grupos alargados.

2. (Complete linkage) Encadenamiento completo o vecino más alejado: La distancia entre los dos nuevos grupos es la mayor de las distancias entre grupos antes de la fusión. Es decir:

$$d(C, AB) = \max(d_{CA}, d_{CB})$$

   Este criterio es invariante ante transformaciones monótonas. Tiende a producir grupos esféricos.

3. (Average linkage) Media de grupos: La distancia entre los dos nuevos grupos es la media ponderada entre las distancias entre grupos antes de la fusión:

$$d(C, AB) = \frac{n_a}{n_a + n_b} d_{CA} + \frac{n_b}{n_a + n_b} d_{CB}$$

Como se ponderan los valores de las distancias, este criterio no es invariante ante transformaciones monótonas de las distancias.

4. (Centroid linkage) Método del centroide: Se aplica generalmente sólo con variables continuas. La distancia entre dos grupos se hace igual a la distancia euclídea entre sus centros, donde se toman como centros los vectores de medias de las observaciones que pertenecen al grupo. Cuando se unen dos grupos se pueden calcular las nuevas distancias entre ellos sin utilizar los elementos originales. Puede demostrarse que el cuadrado de la distancia euclídea de un grupo C a la unión de los grupos A, con n_a elementos y B con n_b es:

$$d^2(C, AB) = \frac{n_a}{n_a + n_b} d_{CA}^2 + \frac{n_b}{n_a + n_b} d_{CB}^2 - \frac{n_a n_b}{(n_a + n_b)^2} d_{AB}^2$$

5. (Ward linkage) Método de Ward: $d(C, AB)$ es la distancia euclídea al cuadrado entre el vector media muestral de los elementos en ambos clústeres.

DENDOGRAMA

El dendrograma, o árbol jerárquico, es una representación gráfica del resultado del proceso de agrupamiento en forma de árbol.

Los criterios que hemos presentado para definir distancias tienen la propiedad de que, si consideramos tres grupos, A, B, C, se verifica que:

$$d(A, C) \leq \max\{d(A, B), d(B, C)\}$$

Una medida de distancia que tiene esta propiedad se denomina ultramétrica.

Esta propiedad es más fuerte que la propiedad triangular, ya que una ultramétrica es siempre una distancia. En efecto, si $d^2(A, C)$ es menor o igual que el máximo de $d^2(A, B)$ y $d^2(B, C)$, forzosamente será menor o igual que la suma $d^2(A, B) + d^2(B, C)$.

El dendrograma es la representación de una ultramétrica, y se construye como sigue:

1. En la parte inferior del gráfico se disponen los n elementos iniciales.
2. Las uniones entre elementos se representan por tres líneas rectas. Dos dirigidas a los elementos que se unen y que son perpendiculares al eje de los elementos y una paralela a este eje que se sitúa al nivel en que se unen.
3. El proceso se repite hasta que todos los elementos están conectados por líneas rectas.

El dendrograma es útil cuando los puntos tienen claramente una estructura jerárquica, pero puede ser engañoso cuando se aplica ciegamente, ya que dos puntos pueden parecer próximos cuando no lo están, y pueden aparecer alejados cuando están próximos.

Particularmente, el dendograma muestra las distancias en las que los clústeres se combinan para formar nuevos clústeres.

Clústeres similares se combinan en distancias pequeñas, mientras que clústeres diferentes se combinan en distancias grandes.

Consecuentemente, la diferencia en distancias define cuán cerca están unos clústeres de otros.

Para obtener una partición de los datos en un número específico de grupos o clústeres, podemos cortar el dendograma en una distancia apropiada.

Si cortamos el dendrograma a un nivel de distancia dado, obtenemos una clasificación del número de grupos existentes a ese nivel y los elementos que los forman.

El número de líneas verticales K, cortado por una línea horizontal en el dendograma en un nivel de distancia, identifica una solución K −clústeres.

Los elementos localizado en el final de cada rama por debajo de la línea horizontal constituyen los miembros de cada clúster.

Para saber si una solución de clústeres es apropiada o no, podemos usar la "silhouette". Sea:

- $a(\mathbf{x}_i)$ la distancia media de $\mathbf{x}_i$ hacia los otros puntos en su clúster.
- $b(\mathbf{x}_i)$ la menor distancia media de $\mathbf{x}_i$ hacia clústeres de los que $\mathbf{x}_i$ no es miembro.

La silhouette de $\mathbf{x}_i$ es:

$$sil(\mathbf{x}_i) = \frac{a(\mathbf{x}_i) - b(\mathbf{x}_i)}{\max\{a(\mathbf{x}_i), b(\mathbf{x}_i)\}}$$

La silhouette varía entre -1 y 1, un valor positivo significa que el elemento está bien identificado con su clúster, y un valor negativo, lo contrario.

La media de las silhouettes nos da una medida global de cuán buena es la configuración (mientras más positiva mejor).

MÉTODO DIVISIVO

En análisis de clúster divisivo (diana: divisive analysis), la idea es que en cada paso, las observaciones se dividan en un grupo A y otro grupo B.

1. El grupo A se inicializa extrayendo la observación que tiene la mayor distancia media hacia las otras observaciones. En un primer paso esa observación es la que conforma el nuevo clúster A, y el resto conforman el clúster B.
2. Dada la separación entre A y B, para cada observación en B se calculan las siguientes cantidades:
    a) La distancia media entre esa observación concreta de B y las demás observaciones en B.
    b) La distancia media de esa observación concreta de B y todas las observaciones de A.
3. Luego, se calculan las diferencias entre los resultados de (a) y (b) del punto (2) para cada observación de B.

4. Van a haber dos posibilidades:

    i. Si todas las diferencias son negativas, detendremos el algoritmo.

    ii. Si alguna diferencia es positiva, tomaremos la observación de B con la diferencia positiva más grande y la moveremos al clúster A y repetiremos todos los cálculos.

Este algoritmo proporciona una división binaria, es decir, en clústeres A y B, pero luego se puede aplicar dentro de los propios clústeres A y B para dividirlos.

CLÚSTER BASADO EN MODELOS

Este análisis es también conocido como clúster basado en mixturas. Para ello, se asume que los datos han sido generados por una mixtura de K distribuciones desconocidas. Para estimar los parámetros desconocidos de la mixtura se utiliza el método de máxima verosimilitud. Más específicamente, un algoritmo que se llama Expectation-Maximization (EM). Una vez que se estiman los parámetros de la mixtura, cada observación es asignada a la parte (clúster) con mayor probabilidad de generar esa observación.

Como estamos asumiendo que los datos son generados por una mixtura, su función de densidad conjunta es igual a:

$$f(\mathbf{x}|\theta) = \sum_{k=1}^{K} \pi_k \, f_{\mathbf{x},k} \, (\mathbf{x}|\theta_k)$$

donde θ es un vector que contiene a todos los parámetros del modelo que son desconocidos, es decir, incluyendo a π_k que son los pesos, y a los parámetros de las distribuciones θ_k.

Entonces, para una matriz de datos $\mathbf{x}$, con observaciones $\mathbf{x}_i = [\mathbf{x}_{1i}, \dots, \mathbf{x}_{ip}]$, la función de verosimilitud es:

$$l(\theta|\mathbf{x}) = \prod_{i=1}^{n} f_x(x_i|\theta_k) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k \, f_{\mathbf{x},k} \, (\mathbf{x}_i|\theta_k) \right)$$

El logaritmo de la verosimilitud es:

$$L(\theta|\mathbf{x}) = \log l(\theta|\mathbf{x}) \; = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k \, f_{\mathbf{x},k} \, (\mathbf{x}_i|\theta_k) \right)$$

Hallar explícitamente el MLE de los parámetros de la mixtura no es posible, incluso cuando los datos provienen de una distribución Normal multivariante.

Por eso, se utiliza un algoritmo de optimización para hallar la solución, es decir, los parámetros que minimizan la log-verosimilitud anterior, y este método se llama Expectation-Maximization (EM).

Una vez que han sido estimados $\hat{\pi}_1, \ldots, \hat{\pi}_K$ y $\hat{\theta}_1, \ldots, \hat{\theta}_K$, la probabilidad a posteriori estimada de que cada observación x_i pertenezca a una población específica $g - \acute{e}sima$, se obtiene con el Teorema de Bayes:

$$\Pr(g|\mathbf{x}_i) = \frac{\hat{\pi}_g f_{x,g}\left(x_i \middle| \hat{\theta}_g\right)}{\sum_{k=1}^{K} \hat{\pi}_k f_{x,k} \left(x_i \middle| \hat{\theta}_k\right)}$$

Las observaciones son asignadas a la densidad (clúster) g con mayor probabilidad $\Pr(g|\mathbf{x}_i)$.

En el análisis basado en modelos, es posible seleccionar el número de clústeres K que mejor funciona.

La idea es compara las soluciones para diferentes valores de $K = 1,2, \ldots$ y escoger el mejor resultado.

Para ello, se usa un criterio de selección de modelos, como por ejemplo el Akaike Information Criterion (AIC) o el Bayesian Information Criterion (BIC).

Por ejemplo, el BIC selecciona el número de clústeres que minimiza:

$$BIC(k) = -2 \times L_k\left(\hat{\theta}\middle|\mathbf{x}\right) + \log(n) \times q$$

Donde $L_k(\hat{\theta}|\mathbf{x})$ es la log-verosimilitud maximizada asumiendo k grupos, y q es el número de parámetros del modelo.

M-clust es un método popular para hacer clúster basado en modelos, que asume densidades Normales y selecciona el mejor modelo basado en el criterio BIC.

Para reducir el número de parámetros a estimar, M-clust trabaja con la descomposición espectral de las matrices covarianza de las densidades Normales Σ_k para $k = 1, \dots, K$ dadas por:

$$\Sigma_k = \lambda_{1,k} V_k \widetilde{\Lambda}_k V_k^t$$

donde $\lambda_{1,k}$ es el auto-valor más grande, V_k es la matriz que contiene los auto-vectores de Σ_k y $\widetilde{\Lambda}_k$ es la matriz diagonal de auto-valores divididos por $\lambda_{1,k}$.

Esta descomposición permite diferentes configuraciones:

1. Esférica con igual volumen.
2. Esférica con volumen desigual.
3. Diagonal con igual volumen y forma.
4. Diagonal, variando volumen, con igual forma.
5. Diagonal, variando volumen y forma.
6. Elipsoidal, con igual volumen, forma y orientación.
7. Elipsoidal, con igual volumen y forma.
8. Elipsoidal, con igual forma.
9. Elipsoidal, variando volumen, forma y orientación.

Esférica, diagonal y elipsoidal son relativas a las matrices de covarianza.

Igual volumen significa que $\lambda_{1,1} = \cdots = \lambda_{1,K}$

Igual forma significa que $\widetilde{\Lambda}_1 = \cdots = \widetilde{\Lambda}_K$

Igual orientación significa que $V_1 = \cdots = V_K$

CLASIFICACIÓN

¿QUÉ ES?

Clasificar datos es asignar a las observaciones un grupo determinado asumiendo que los datos son heterogéneos y se pueden dividir en grupos.

El problema es el siguiente:

- Tenemos un conjunto de elementos que provienen de dos o más poblaciones.
- Observamos una variable $p-$dimensional $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ en estos elementos.
- Queremos clasificar un nuevo elemento, con valores conocidos para esta variable pero desconocemos de cuál población proviene.

SUPERVISADA VS NO SUPERVISADA

Este problema también se conoce como clasificación supervisada, porque nosotros vamos a conocer de antemano la clasificación de los elementos muestrales en los datos. Eso va a servir de información para clasificar nuevos elementos. Notemos que clasificación no supervisada sería el análisis clúster, porque ahí no sabemos cuál es la agrupación de los datos de la muestra.

- Finanzas: credit scoring, el problema de decidir si darle un crédito o no a una persona, basándose en sus ingresos, salud, edad, etc.
- Control de calidad: clasificar componentes en buen o mal estado: lámparas, televisiones, etc.
- Ingeniería: diseño de maquinarias capaces de clasificar automáticamente billetes, monedas, códigos postales, etc.
- Otros ejemplos: reconoces declaraciones de impuestos fraudulentas, negocios fraudulentos, procesos de manufacturación deficientes, etc.

DEFINICIÓN DEL PROBLEMA DE CLASIFICACIÓN

Tenemos una variable aleatoria $p-$dimensional $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ definida sobre elementos muestrales que pertenecen a G poblaciones, P_g, para $g = 1, \dots, G$. La variable respuesta $\mathbf{y}$ contiene las clasificaciones de los elementos muestrales, es decir, toma valores en $g = 1, \dots, G$. Tenemos una matriz de datos $\mathbf{x}$ de tamaño $n \times p$ con observaciones $\mathbf{x_i}$ para $i = 1, \dots, n$ con probabilidades de pertenencias a grupos conocidas. Desconocemos las probabilidades π_g para $g = 1, \dots, G$, de que un elemento seleccionado al azar provenga de la población g. Se cumple que $\pi_1 + \cdots + \pi_G = 1$. Queremos clasificar un nuevo elemento con valores conocidos en las variables $\mathbf{x_0} = [\mathbf{x}_{01}, \dots, \mathbf{x}_{0p}]$ dentro de alguna población G.

Hay muchas técnicas de clasificación para abordar este problema.

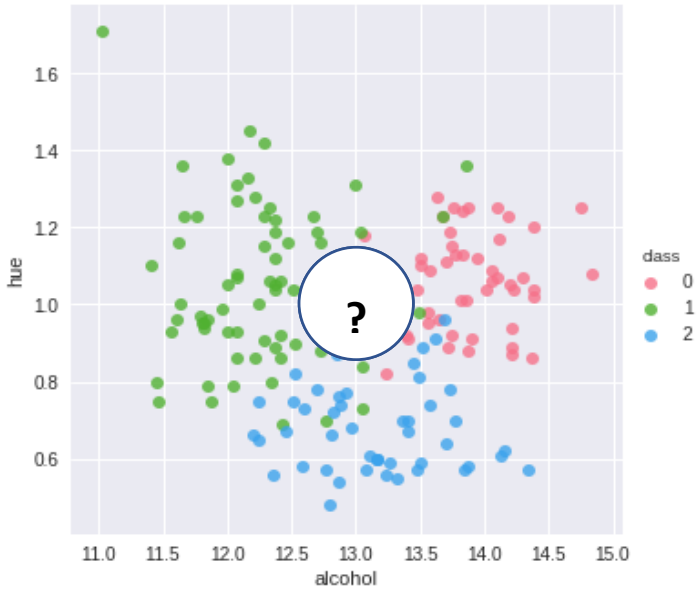
Vamos a estudiar las tres técnicas más conocidas:

1. k vecinos más cercanos (kNN: k-Nearest Neighbors)
2. Clasificador bayesiano
3. Regresión logística

KNN

k-Nearest Neighbors o k vecinos más cercanos: es un método de clasificación no paramétrico, es decir, no requiere asumir ninguna distribución para la variable aleatoria $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_p]$. Este método no requiere estimar las probabilidades desconocidas π_g para $g = 1, \dots, G$, de que un elemento seleccionado al azar provenga de la población g.

La idea es buscar, para la nueva observación que queremos clasificar, sus k vecinos más cercanos, es decir, las k observaciones más cercanas respecto a una medida de distancia.



El algoritmo es el siguiente:

1. Definimos una medida de distancia adecuada para las observaciones.
2. Calculamos la distancia entre la nueva observación x_0 que queremos clasificar, y las observaciones que tenemos en nuestra matriz de datos $\mathbf{x}$.
3. Seleccionamos las k observaciones más cercanas a $\mathbf{x_0}$, y miramos a qué grupo pertenecen.
4. Clasificamos $\mathbf{x_0}$ en la población a la que pertenece una mayor proporción de sus $\boldsymbol{k}$ vecinos.

Cuando las variables tienen distintas unidades de medida es recomendable escalar o normalizar para que las distancias no se vean influenciadas por las magnitudes.

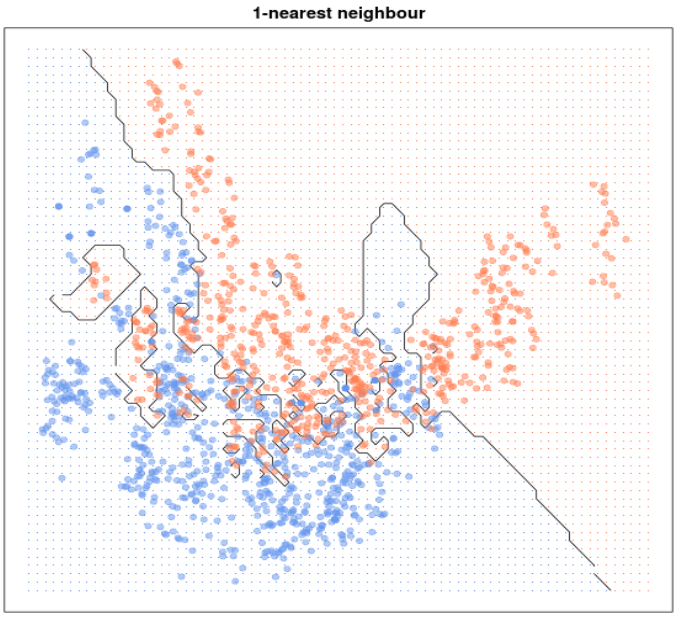
Escalar resulta en un rango entre 0 y 1:

$$\mathbf{x}' = \frac{\mathbf{x} - \min(\mathbf{x})}{\max(\mathbf{x}) - \min(\mathbf{x})}$$

Normalizar devuelve una variable centrada en media cero con desviación típica 1:

$$\mathbf{x}' = \frac{\mathbf{x} - \bar{\mathbf{x}}}{\sigma}$$

Una decisión que hay que tomar es: ¿qué valor de k se debe seleccionar?



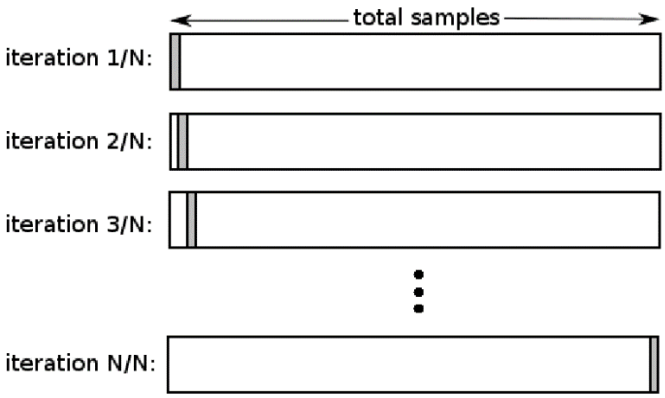
VALIDACIÓN CRUZADA

La validación cruzada o cross-validation es un método muy útil para evaluar el rendimiento de los métodos estadísticos, en este caso, el de clasificación.

Consiste en, dada una muestra y un modelo que depende de ciertos de parámetros:

1. Dividimos la muestra en dos sub-muestras.
2. Usamos la primera sub-muestra para estimar los parámetros del modelo.
3. Usamos la segunda sub-muestra para validar el rendimiento del método o modelo con los parámetros que han sido estimados con la primera sub-muestra.

Leave-one-out cross validation es un caso particular de validación cruzada, donde la segunda sub-muestra consiste en sólo una observación de la muestra.



El parámetro que queremos encontrar en este caso es k, así que repetiríamos el procedimiento para $k = 1, \dots, k_{max}$

El k óptimo es el que dé menor error de "observaciones mal clasificadas" (missclassification rate - MR):

$$MR = \frac{total\ de\ observaciones\ mal\ clasificadas}{total\ de\ observaciones}$$

Para poder llegar a conclusiones en este caso, tendríamos que hacer n veces el leave-one-out cross validation, una vez por cada una de las n observaciones de la muestra.

Otro método de validación cruzada es el k-fold cross validation. Pero no confundir este k con el del método kNN. Es muy común el 10-fold cross validation.

CLASES BALANCEADAS

Notemos que kNN no tiene en cuenta los números de observaciones en cada grupo ni las probabilidades de pertenencia de cada grupo.

Si las clases no están balanceadas, por ejemplo, una o más clases tiene muchos más elementos que otra clase, kNN al igual que otros métodos de clasificación, puede sesgarse y clasificar las nuevas observaciones hacia las poblaciones más grandes.

En estos casos, lo que se suele hacer es balancear las clases con algún método, por ejemplo:

1. Añadir copias muestrales de observaciones seleccionadas aleatoriamente de las poblaciones con menor representación (over-sampling).
2. Eliminar observaciones de las poblaciones con más observaciones (under-sampling).

CLASIFICACIÓN BAYESIANA

Se basa en el Teorema de Bayes, por lo que es un modelo probabilístico.

La idea es: clasificar o asignar la nueva observación a la población con mayor probabilidad de haberla generado. Supongamos que conocemos las funciones de densidades para cada una de las G poblaciones $f_1, \dots, f_G$. Usando el Teorema de Bayes la probabilidad de que la nueva observación $\mathbf{x}_0$ haya sido generada por la población P_k es:

$$P(y = k|\mathbf{x} = \mathbf{x}_0) = \frac{\pi_k f_k(\mathbf{x}_0)}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_0)}$$

Donde π_g para $g = 1, \dots, G$ son las probabilidades de que una observación provenga de la población g-ésima.

En otras palabras, clasificamos a $\mathbf{x}_0$ en la población k-ésima P_k si $P(y = k|\mathbf{x} = \mathbf{x}_0)$ es la probabilidad máxima dentro de todas las posibles considerando el total de G poblaciones.

Este criterio es equivalente a decir que $\pi_k f_k(\mathbf{x_0})$ es máximo. En particular, si se cumple $\pi_1 = \cdots = \pi_G$, entonces las condiciones para clasificarla en P_k es que $f_k(\mathbf{x_0})$ sea máxima. Esto quiere decir que clasificaremos a $\mathbf{x_0}$ en la población que tenga mayor valor de densidad en esa observación $\mathbf{x_0}$. Veamos qué pasa si asumimos que las funciones de densidad de las poblaciones $f_1, \ldots, f_G$ son Normales. Primero supongamos que tienen diferentes vectores de media $\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_G}$ pero la misma matriz de covarianza Σ. Con estas suposiciones, en el grupo de la población k la variable aleatoria multivariante se distribuye como:

$$\mathbf{x} \sim N(\boldsymbol{\mu}_k, \Sigma)$$

Según la regla de clasificación bayesiana, clasificaríamos el nuevo elemento $\mathbf{x_0}$ en la población P_k que maximice $\pi_k f_k(\mathbf{x_0})$ que es dada por:

$$\pi_k f_k(\mathbf{x_0}) = \pi_k (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left(-\frac{(\mathbf{x_0} - \mu_k)^t \Sigma^{-1} (\mathbf{x_0} - \mu_k)}{2}\right)$$

Tomando logaritmos y eliminando constantes esto es equivalente a clasificar el nuevo elemento en la población que maximice la expresión:

$$2 \log \pi_k - (\mathbf{x_0} - \mu_k)^t \Sigma^{-1} (\mathbf{x_0} - \mu_k)$$

La expresión anterior es igual a

$$2 \log \pi_k - \mathbf{x_0} \Sigma^{-1} \mathbf{x_0} + 2\mu_k^t \Sigma^{-1} \mathbf{x_0} - \mu_k^t \Sigma^{-1} \mu_k$$

Consecuentemente, la regla bayesiana se reduce a clasificar el nuevo elemento en la población que maximice:

$$p_k(\mathbf{x_0}) = 2\mu_k^t \Sigma^{-1} \mathbf{x_0} - \mu_k^t \Sigma^{-1} \mu_k + 2 \log \pi_k$$

porque $\mathbf{x_0} \Sigma^{-1} \mathbf{x_0}$ no depende de la población P_k.

Como la expresión $p_k(\mathbf{x_0})$ depende linealmente de $\mathbf{x_0}$ el método se denomina **análisis discriminante lineal.**

Cuando, además, asumimos que las probabilidades $\pi_1 = \cdots = \pi_G$, el clasificador o análisis discriminante lineal, tiene una interesante interpretación.

Si las probabilidades son iguales, la expresión a maximizar

$$2\log\pi_k - (\mathbf{x}_0 - \mu_k)^t\Sigma^{-1}(\mathbf{x}_0 - \mu_k)$$

resulta equivalente a minimizar

$$(\mathbf{x}_0 - \mu_k)^t\Sigma^{-1}(\mathbf{x}_0 - \mu_k)$$

Que quiere decir que clasificaremos $\mathbf{x}_0$ a la población cuyo vector de medias es el más cercano en términos de la Distancia de Mahalanobis.

De hecho, la regla discriminante lineal se puede poner en función de la Distancia de Mahalanobis:

$$\begin{aligned} P(y = k|\mathbf{x} = \mathbf{x}_0) &= \frac{\pi_k f_k(\mathbf{x}_0)}{\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_0)} \\ &= \frac{\pi_k \exp\left(-\frac{(\mathbf{x}_0 - \mu_k)^t\Sigma^{-1}(\mathbf{x}_0 - \mu_k)}{2}\right)}{\sum_{g=1}^{G}\pi_g \exp\left(-\frac{(\mathbf{x}_0 - \mu_g)^t\Sigma^{-1}(\mathbf{x}_0 - \mu_g)}{2}\right)} \\ &= \frac{\pi_k \exp\left(-\frac{1}{2}D_M^2(\mathbf{x}_0, \mu_k)\right)}{\sum_{g=1}^{G}\pi_g \exp\left(-\frac{1}{2}D_M^2(\mathbf{x}_0, \mu_g)\right)} \end{aligned}$$

Donde $D_M^2(\mathbf{x}_0, \mu_g) = (\mathbf{x}_0 - \mu_g)^t\Sigma^{-1}(\mathbf{x}_0 - \mu_g)$ es la distancia de Mahalanobis al cuadrado entre la nueva observación $\mathbf{x}_0$ y la media del grupo en cuestión μ_g

En resumen, en el análisis discriminante lineal estamos asumiendo que conocemos las probabilidades $\pi_1, \ldots, \pi_G$ y los parámetros de las distribuciones Normales, es decir, las medias $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G$ y la matriz de covarianza común $\boldsymbol{\Sigma}$.

En la práctica esto no siempre es así, por lo que al final tendremos que estimar estas cantidades. Para hacer estas estimaciones, podemos considerar que la matriz de datos puede subdividirse en G matrices correspondientes a las G poblaciones. Entonces si tenemos una matriz de datos de tamaño $n \times p$ la subdividiremos por filas separando los grupos cuyos tamaños muestrales son n_g tales que $n = \sum_{g=1}^{G} n_g$.

Las probabilidades $\pi_1, \ldots, \pi_G$ pueden estimarse con las proporciones de los datos que pertenecen a cada grupo:

$$\hat{\pi}_g = \frac{n_g}{n}$$

Los vectores de medias en cada grupo pueden estimarse como la media muestral dentro del grupo:

$$\bar{\mathrm{x}}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} x_i^g$$

Donde x_i^g son las observaciones dentro del grupo g. La matriz de covarianza común $\boldsymbol{\Sigma}$ puede estimarse como:

$$\boldsymbol{S_w} = \sum_{g=1}^{G} \left(\frac{n_g - 1}{n - G} \right) S_g$$

Donde S_g es la matriz de covarianza muestral de los elementos del grupo g.

$$S_g = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} \left(x_i^g - \bar{\mathrm{x}}_g \right) \left(x_i^g - \bar{\mathrm{x}}_g \right)^t$$

Donde S_g es la matriz de covarianza muestral de los elementos del grupo g. Usando estos estimadores se puede calcular ya la regla de análisis discriminante lineal. Para evaluarla habría que hacer validación cruzada.

ANÁLISIS DISCRIMINANTE CUADRÁTICO

Cuando las hipótesis que hemos asumido antes cambian, y asumimos que las densidades son Normales pero que las medias y las matrices de covarianza son distintas entre los grupos, estamos en presencia del análisis discriminante cuadrático.

Entonces en la población P_k, la variable se distribuye: $\mathbf{x} \sim N(\boldsymbol{\mu}_k, \Sigma_{\mathrm{k}})$

Y el criterio se convierte en maximizar:

$$\pi_k f_k(\mathbf{x}_0) = \pi_k (2\pi)^{-p/2} |\Sigma_{\mathrm{k}}|^{-1/2} \exp\left(-\frac{(\mathbf{x}_0 - \mu_k)^t \Sigma^{-1} (\mathbf{x}_0 - \mu_k)}{2}\right)$$

Lo cual se puede reducir a:

$$p_k(\mathbf{x}_0) = -\mathbf{x}_0 \Sigma_k^{-1} \mathbf{x}_0 + 2\mu_k^t \Sigma_{\mathrm{k}}^{-1} \mathbf{x}_0 - \mu_k^t \Sigma_{\mathrm{k}}^{-1} \mu_k + 2\log \pi_k - \log|\Sigma_k|$$

Como $p_k(\mathbf{x}_0)$ depende cuadráticamente de $\mathbf{x}_0$, el método se denomina análisis discriminante cuadrático.

Lo cual lleva a la siguiente expresión:

$$P(y = k|\mathbf{x} = \mathbf{x}_0) = \frac{\pi_k |\Sigma_k|^{-1/2} \exp\left(-\frac{1}{2} D_M^2(\mathbf{x}_0, \mu_k)\right)}{\sum_{g=1}^{G} \pi_g |\Sigma_g|^{-1/2} \exp\left(-\frac{1}{2} D_M^2(\mathbf{x}_0, \mu_g)\right)}$$

donde $D_M^2(\mathbf{x}_0, \mu_g) = (\mathbf{x}_0 - \mu_g)^t \Sigma_g^{-1} (\mathbf{x}_0 - \mu_g)$ es la distancia de Mahalanobis al cuadrado entre $\mathbf{x}_0$ y μ_g

REGRESIÓN LOGÍSTICA

Uno de los problemas de la clasificación bayesiana es que se asume que la variable aleatoria tiene una distribución específica en la población P_g. Esto restringe el tipo de variables que pueden usarse en las reglas discriminantes lineal y cuadrática. Sin embargo, un marco probabilístico aporta un soporte muy fuerte a las decisiones, la pregunta es si se podrían calcular las probabilidades $P(y = k|\mathbf{x} = \mathbf{x}_0)$ sin tener conocimiento a priori de las densidades.

Para evitar usar el Teorema de Bayes, una posibilidad es asumir que esa probabilidad es igual a cierta función positiva de $\mathbf{x_0}$:

$$P(y = k|\mathbf{x} = \mathbf{x}_0) = p_{0k}$$

tal que $p_{01} + \cdots + p_{0G} = 1$

La pregunta es qué funciones $p_{01}, \ldots, p_{0G}$ son las más apropiadas para obtener buenas clasificaciones. Consideremos el problema de discriminación o clasificación entre dos poblaciones. Y supongamos que además de las variables explicativas tenemos una variable respuesta que toma valor cero para una de esas poblaciones y valor uno para la otra. En caso de que sean clases no numéricas siempre se pueden transformar a estos valores.

El primer enfoque para tratar de obtener un modelo que explique la respuesta es considerar un modelo de regresión:

$$\mathbf{y} = \beta_0 + \boldsymbol{\beta}_1^t \mathbf{x} + \mathbf{u}$$

Pero esto presenta problemas. Primero tomemos esperanzas en la expresión anterior para $\mathbf{x} = \mathbf{x_0}$, quedaría, por un lado:

$$E[\mathbf{y}|\mathbf{x}_0] = \beta_0 + \boldsymbol{\beta}_1^t \mathbf{x_0}$$

Ahora, sabemos que la variable $\mathbf{y}$ es binomial y toma valores cero y uno, supongamos que toma valor uno con probabilidad p_0 y valor cero con probabilidad $1 - p_0$, donde $p_0 = P(y = 1|\mathbf{x}_0)$.

La esperanza de $\mathbf{y}$ será:

$$E[\mathbf{y}|\mathbf{x}_0] = p_0 \times 1 + (1 - p_0) \times 0 = p_0$$

Con lo cual podemos concluir que

$$p_0 = \beta_0 + \boldsymbol{\beta}_1^t \mathbf{x_0}$$

Esto tiene el problema de que, si estimamos el modelo lineal, la predicción $\hat{y} = \hat{p}_0$ será la probabilidad de que un individuo con características definidas por $\mathbf{x} = \mathbf{x_0}$ pertenezca a la segunda población ($y = 1$).

Sin embargo, no hay ninguna garantía de que esa estimación esté entre cero y uno, y debe estarlo porque es una probabilidad. Esto es un problema porque incluso pueden aparecer valores negativos.

Si queremos que el modelo construido para discriminar nos proporcione directamente la probabilidad de pertenecer a cada población, debemos transformar la variable respuesta para garantizar que la respuesta prevista esté entre cero y uno:

$$p_0 = F(\beta_0 + \boldsymbol{\beta}_1^t \mathbf{x_0})$$

De esta manera p_0 estará entre cero y uno si escogemos una F para que tenga esa propiedad. Como, por ejemplo, la función logística:

$$p_0 = \frac{1}{1 + e^{-\beta_0 - \beta_1^t \mathbf{x}_0}}$$

Esta función tiene varias ventajas, una de ellas su continuidad.

MODELO LOGIT

Además, como:

$$1 - p_0 = \frac{e^{-\beta_0 - \beta_1^t \mathbf{x}_0}}{1 + e^{-\beta_0 - \beta_1^t \mathbf{x}_0}} = \frac{1}{1 + e^{\beta_0 + \beta_1^t \mathbf{x}_0}}$$

resulta que:

$$g = \log \frac{p_0}{1 - p_0} = \beta_0 + \boldsymbol{\beta}_1^t \mathbf{x}_0$$

A g se le llama *variable logit* o *modelo logit*, y representa en una escala logarítmica la diferencia entre las probabilidades de pertenecer a ambas poblaciones.

MODELO MULTILOGIT

El modelo logit puede generalizarse para más de dos poblaciones, es decir, para variables cualitativas con más de dos niveles posibles. Supongamos G

poblaciones, llamando p_{0k} a la probabilidad de que el elemento $\mathbf{x_0}$ pertenezca a la clase k, podemos escribir:

$$p_{0k} = P(y = k|\mathbf{x} = \mathbf{x}_0) = \frac{e^{\beta_{0k} + \beta_{1k}^t \mathbf{x}_0}}{1 + \sum_{g=1}^{G-1} e^{\beta_{0g} + \beta_{1g}^t \mathbf{x}_0}}$$

para $k = 1, \ldots, G-1$. Y para $k = G$:

$$p_{0G} = P(y = G|\mathbf{x} = \mathbf{x}_0) = \frac{1}{1 + \sum_{g=1}^{G-1} e^{\beta_{0g} + \beta_{1g}^t \mathbf{x}_0}}$$

Notemos primero que $\boldsymbol{\beta}_{1k}$ es un vector p-dimensional.

Y que:

$$\sum_{k=1}^{G} p_{0k} = \sum_{k=1}^{G} P(y = k|\mathbf{x} = \mathbf{x}_0) = 1$$

El modelo logístico puede aplicarse cuando las variables explicativas no son Normales, incluyendo variables discretas y variables categóricas que pueden incluirse en el modelo vía variables dummy (0/1).