Manejo datos

Introducción

El análisis multivariado se analizan las relaciones entre dos o más conjuntos de mediciones que se hacen a cada objeto o individuo, en una o varias muestras.

Los datos los acomodaremos en la siguiente matriz:

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1p} \\ X_{21} & X_{21} & \cdots & X_{2j} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nj} & \cdots & X_{np} \end{bmatrix}$$

Cada renglón representa a un vector $\bar{X}_i \in \Re^p$ $i = 1, \ldots, n.$, tenemos n vectores renglón de dimensión p.

Los renglones corresponden a: personas, organizaciones, eventos, países, etc.

Las variables, son las características, o las propiedades que que nos interesan de los objetos.

Se tiene un universo de observaciones potenciales, pero de ese universo sólo se observará un subconjunto. A este subconjunto se le aplicará un modelo de medición (las condiciones e instrumento para medir) del que resultarán los datos. Se buscan asociaciones entre las variables a través de los distintos modelos multivariados.

Calidad en los datos

Para darnos una idea de la calidad de los datos podemos hacer:

1. Inspección visual.- Par ver si hay datos fuera de los rangos establecidos, conocer el máximo y mínimo de cada variable. Verificar que las codificaciones sean consistentes en toda la base.

2. Distribución de frecuencias.- De las variables de mayor interés. Algunas quisieramos que parecieran normales.

3. Gráficas de dispersión.- Identificar grupos u observaciones discrepantes.

4. Verificar métodos de recolección de los datos.- Detectar posibles fuentes de sesgo.

5. Manejo de las Observaciones Faltantes.- Tratar de rastrearlas ir a registros originales, razones de su omisión. Definir que se hará con estas observaciones, se puede usar algún valor de reemplazo o imputación o seleccionar cuáles si se desechan. Los valores faltantes generan sesgo este tema es de suma importancia

OJO Cuidado con el número de dígitos a guardar, puede perderse precisión o al revés desperdiciar espacio. Tener control sobre los estándares de medición.

Un grupo de datos de poca calidad no merece un análisis muy detallado.

Fases del análisis

- Manipulación inicial de los datos. Reunir los datos en forma conveniente.

- Análisis preliminar. Se intenta aclarar la forma de los datos y ver que dirección debe tomar el análisis

- Análisis definitivo que dará las bases para las conclusiones.

- Presentación de conclusiones

Escalas de Medición

Escala de medición	Operaciones	Cambios permitidos	Ejemplo	Valores
Nominal	Pertenencia a categoría	de nombre	Sexo Estado civil	Masc., Fem. Casado Divor
Ordinal	Grado de intensidad	que mantengan orden	Calificaciones sabor	NA, S, B, MB Bno reg malo
Intervalo	Igualdad de intervalos	de escala y origen	Temperatura Tiempo	Enteros, reales.
Razón	Igualdad de proporciones	de escala pero no de origen	Concentración sustancias	Enteros, reales.
Absoluta	Conteo de elementos	No escala no origen	Número de hijos	Enteros

Datos Faltantes

Los valores faltantes son observaciones que en un se tenía la intención de hacerlas, pero por distintas razones no se obtuvieron.

Continuar con el análisis estadístico y hacer como si no faltaran observaciones

afecta las propiedades de los estimadores (sesgos),

las longitudes de los intervalos de confianza,

la potencia de pruebas.

MECANISMOS DE PÉRDIDA

Denotando a Y como la información con intención de ser recoletada.

$$Y = \{Y_o, Y_m\}.$$

Donde Y_o es la información observada y Y_m la información perdida.

El indicador de información perdida R se define como

$$R = \begin{cases} 1 & \text{si Y se observa} \\ 0 & \text{si Y no se observa} \end{cases}$$

R está en relación a Y.

La validez del análisis dependerá del mecanismo de pérdida, que a su vez está definido por

$$P(R|y_o, y_m)$$

Observaciones Perdidas Completamente al Azar (MCAR)

Cuando la probabilidad de pérdida no depende de lo observado ni de lo no observado

$$P(r|y_o, y_m) = Pr(r)$$

En este caso puede hacerse la inferencia como si no hubiese faltantes, pues no hay sesgo, aunque si se pierde precisión al ser menor la muestra.

Ejemplo

Una muestra de laboratorio accidentalmente se cae y rompe, el baumanómetro se descompone, el doctor se salta una pregunta en la historia clínica, etc.

OJO Algunas pareden MCAR pero analizándolas bien podrían no serlo.

- En un estudio longitudinal alguien no prosigue pues tiene un accidente de "caída bajo el autobús". (Si el estudio fuera de avance en aprendizaje de un idioma parece que no continuar en el estudio no tiene nada que ver con la variable de aprendizaje, pero si se trata de un estudio psiquiátrico sobre depresión, podría ser que el sujeto no esté respondiendo

al tratamiento, entonces la no respuesta (accidente) si tiene que ver con la variable respuesta.)

- Un cuestionario no es respondido pues es robado en la oficina de correos, (podría No ser aleatorio, pues probablenete este relacionado con la variable zona donde la oficina está localizada.)

Observaciones Perdidas al Azar (MAR)

En este caso el mecanismo de pérdida no depende de la información perdida pero si de la observada.

$$P(r|y_o, y_m) = P(r|y_o).$$

ESto equivale a decir que que el comportamiento de dos unidades que comparten valores observados tienen el mismo comportamiento estadístico en las otras variables ya sea que se observen o no.

Por ejemplo:

	variables					
unidad	1	2	3	4	5	6
1	1	3	4.3	3.5	1	4.46
2	1	3	?	3.5	?	?

Las unidades 1 y 2 tienen los mismos valores observados, dados estos valores observados, bajo MAR, las variables 3,5 y 6 de

2 tienen la misma **distribución** que las variables 3,5 y 6 de la unidad 1.

Un caso especial de MAR, es la respuesta uniforme dentro de clases.

Ejemplos

Si en un estudio se buscan datos de ingreso y categoría de pago de impuestos. Es común que los que ganen más se nieguen a revelar su ingreso. Promediar los datos completos nos llevaría a subestimar el ingreso.

Ahora si se conoce para todos la categoría de pago de impuestos, dentro de cagda categoría la NO respuesta acerca del ingreso es

aleatoria, así el mecanismo de pérdida depende de la categoría de pago de impuestos y la pérdida de datos de ingreso NO depende de valor del ingreso mismo.

Para estimar convendría primero calcular el promedio por categoría de impuestos, dado que dentro de cada categoría la pérdida es aleatoria, el promedio es un estimador válido y luego para estimar el ingreso medio de toda la población, se combinan estos estimadores en un promedio ponderando de manera proporcional al tamaño de las categorías de impuestos.

Una persona con depresión puede ser que tienda más a no contestar acerca de su ingreso, la gente con depresión a su vez en general tiene menos ingresos, entonces lo que ocurre es que si

hay un tasa alta de no respuesta entre las personas con depresión, la media real puede ser menor que la calculada con los datos existentes, es decir sin tomar en cuenta a los datos faltantes. Ahora si entre las personas con depresión la probabilidad de no contestar acerca de su ingreso no está relacionada con su nivel de ingreso, entonces los datos se consideran faltantes al azar, (MAR). Esto No significa que estos faltantes no produzcan sesgo y que se pueda uno olvidar del problema.

Observaciones NO Perdidas al Azar (MNAR)

Se dice que el mecanismo de pérdida NO es ignorable si los valores perdidos dependen de los valores no observados.

$$P(r|y_o, y_m) = Pr(r|y_m).$$

Y para hacer inferencia válida debe hacerse usando Y y R.

Ejemplos

Si se estudia una cierta enfermedad y las persona que padecen esa enfermedad son las que tienen una mayor probabilidad a no contestar a si la padecen, entonces los datos son faltantes no al azar, MNAR.

Claramente el estimador de la proporción que padece esa enfermedad será menor que la proporción que se obtendría con los datos completos.

Lo mismo ocurre en el caso de las personas con menor ingreso son las que tienden a no contestar su nivel de ingreso.

Esta falta de datos no al azar es un problema, la única manera de obtener un estimador insesgado es modelar la esa ausencia de datos y los valores mismos de las ausencias, esta tarea no es para nada simple.

Tratamiento de datos faltantes

Omisión total

Si los datos son MCAR las estimaciones obtenidas serán insesgadas si no son MCAR serán sesgadas, hay que tener en cuenta que esta pérdida de datos genera pérdida de potencia en las pruebas.

Omisión parcial

Por ejemplo en el cálculo de las correlaciones se usan las observaciones disponibles, pero entonces cada estimación está soportada por diferentes bases de datos. Puede ser el caso que se llegue a una matriz de correlaciones estimada NO definida positiva.

No hay que olvidar que hay que analizar a las observaciones NA y tratar de ver si se comportan (en ciertas variables) como la población total o si difieren.

Otra cosa importante es considerar que es lo que se tiene perdido. La situación de perder variables explicactivas es diferente a perder variables respuesta.

Sustitución

Hot deck, sustituir el caso por alguien semejante, (de donde sacamos a alguien semejante si ya acabó la encuesta, tener la providencia de guardar un montoncito extra para la sustitución??)

Imputación Simple

Sustituir los valores faltantes por la media (el estimador de mx́ima verosimilitud), pero eso tiene consecuencias sobre la estimación de la varianza. Pero siempre estaremos sustituyendo con el mismo valor.

O se puede sustituir usando una regresión, pero el problema sigue siendo que se sustituye por una media (esta vez condicionada)

SPPS permite sumar una variación aleatoria, se subsana en algo este tipo de problema.

O se puede usar el Algoritmo EM. En regresión si se conocieran los NA, estimar los parámetros del modelo sería fácil, y si se conocieran los parámetros del modelo de los datos sería sencillo hacer predicciones insesgadas de las observaciones faltantes. Este algoritmo es iterativo y va haciendo ambas cosas: con los datos existentes se estiman los paámetros del modelo de los datos, enseguida con estos parámetros se hacen estimaciones de los datos faltantes, y de nuevo se re-estiman los paámetros con las datos ya completados. Schafer (1997) hizo un programa NORM disponible en http://www.stat.psu.edu/ jls/misoftwa.html, SPSS tiene un procedimiento que hace imputación utilizando EM.

Imputación Múltiple

En imputación múltiple se generan valores para hacer la imputación basados en los datos existentes. Suponiendo que se estima y usando x, pero esta imputación se hace varias veces, es decir tendremos varios conjuntos de datos completados. Para hacer esto se usan métodos conocidos Markov Chain Monte Carlo. El programa NORM tambien su parte llamada data augmentation lo hace. SAS tiene dos procedimientos MI y MIANALYZE.

Schafer, J.L. & Olsden, M. K.. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research, 33, 545-571

En R esta el paquete MICE, material con referencia en: Van Buuren, S., Groothuis-Oudshoorn, K. (2011) MICE: Multivariate

Imputation by Chained Equations in R. Journal of Statistical Software.

http://www.stefvanbuuren.nl/publications/MICEinR-Draft.pdf

Observaciones Discrepantes

Estas observaciones también son conocidas como aberrantes, discordantes, contaminantes,sorprendentes, en inglés OUTLIER. Puede definírseles de varias formas, una de ellas es decir que es una observación que se encuentra a una distancia ANORMAL de las demás, y entonces hay que definir lo que es una distancia NORMAL, es decir la observación se encuentra fuera de la nube de datos. Estas observaciones pueden distorsionar la información, también pueden ser una señal de que el modelo de distribución

de los datos NO es el adecuado, o reflejar el haber encontrado una situación sorprendente o peculiar. Si la observación causa un impacto en el observador se le llama generalmente **discrepante**. Una observación **contaminante** será cualquiera que no corresponda a la distribución supuesta, y ésta puede no ser percibida por el observador.

Estas observaciones afectan fuertemente al estimador $\bar{X}$ de la media μ, y consecuentemente a los estimadores de $Var(X)$, de las de $Cov(X,Y)$ y de $Corr(X,Y)$. En análisis de regresión interesa identificar a las observaciones **influyentes**, que son aquellas que al omitirlas del análisis los valores de las $\widehat{\beta}$'s varían mucho.

Detectar estas observaciones puede ser una tarea bastante complicada, sobre todo cuando se tienen datos altamente multivariados.

En el caso univariado se les puede detectar muy fácilmentea través de gráficos boxplot o tambien al verificar si la media de los datos difiere mucho de la mediana.

Gráficas datos univariados

- gráfica de barras y de *pie* son solo para datos categóricos, debe haber espacios entre las barras.

- histograma debe tenerse cuidado con los anchos de barras y con los puntos que se consideran en el eje de las x.

- boxplot permite rápidamente ver observaciones discrepantes.

- stem, una versión de los histogramas pero permite ver los datos tal cual.

- series de tiempo

Gráficas datos multivariados

- estrellas. Convienen cuando no se tienen muchos atributos, pues con más de 10 o 12 aristas las confundimos en su forma.

- faces, debidas a Chernov, dado que el ojo humano esta muy entrenado para reconocer rostros humanos. A cada elemento

de la cara: pelo, ancho cara, largo nariz, tamaño de ojos se le asocia una característica.

- curvas de Andrews, a cada individuo se le asigna una curva de la siguiente manera: $t \in [-\pi, \pi]$ Si p es impar

$$f_i(t) = \frac{X_{i1}}{\sqrt{2}} + X_{i2}\sin(t) + X_{i3}\cos(t) + \ldots + X_{ip}\cos(\frac{(p-1)}{2}t)$$

Si p es par

$$f_i(t) = \frac{X_{i1}}{\sqrt{2}} + X_{i2}\sin(t) + X_{i3}\cos(t) + \ldots + X_{ip}\sin(\frac{p}{2}t)$$

Estas tres gráficas no son únicas, pues según ordenemos las variables darán origen a estrellas, curvas o caras distintas.

- bagplot parecida a un boxplot pero en dos dimensiones.

- gráfica de paralelas, se usan sobre todo cuando hay varias mediciones de una misma variable para un solo individuo.

- series de tiempo múltiples

Algunos conceptos de Estadística

La **media poblacional** se define como:

$$E[X_i] = \int x_i dF(x_i) = \mu_i$$

$$E(\mathbf{x}) = \begin{pmatrix} E(x_1) \\ E(x_2) \\ \vdots \\ E(x_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

La **varianza poblacional**, cuando existe,se define como:

$$Var[X_i] = \int (x_i - \mu_i)^2 f(x_i) dx_1 = \sigma_i^2 = \sigma_{ii}^2$$

La **covarianza poblacional** se define como:

$$cov(x_i, x_j) = \int \int (x_i - \mu_i)(x_j - \mu_j) f(x_i, x_j) dx_i dx_j = \sigma_{ij}$$

Estos valores se presentan dentro de la matriz de varianzas y covarianzas Σ de dimensiones $p \times p$

$$\Sigma = E[(X - E[X])(X - E[X])'] = var(x) \qquad (1)$$

$$= \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & \cdots & E[(X_1 - \mu_1)(X_p - \mu_p)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & \cdots & E[(X_2 - \mu_2)(X_p - \mu_p)] \\ \vdots & \vdots & \vdots \\ E[(X_p - \mu_p)(X_1 - \mu_1)] & \cdots & E[(X_p - \mu_p)(X_p - \mu_p)] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \sigma_{21} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{bmatrix}$$

Otra medida importante es la correlación entre dos variables y está dada por:

$\rho(x_i, x_j) == \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$

La **media muestral** de la j-ésima variable está dada por

$\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$

Denotaremos al conjunto de las medias en un vector de medias muestrales

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

La **varianza muestral** de la k-ésima variable se calcula como:

$S_{kk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2$

La **covarianza** entre la l-ésima variable y la k-ésima variable esta dada por

$$S_{lk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{il} - \bar{x}_l)(x_{ik} - \bar{x}_k)$$

La **matriz de covarianzas muestral** denotada por **S** de dimensiones $p \times p$, contiene a las varianzas y covarianzas.

$$\mathbf{S} = \frac{\mathbf{1}}{\mathbf{n-1}} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})' = \begin{bmatrix} {S_1}^2 & \dots & S_{1p} \\ & \ddots & \vdots \\ & & {S_p}^2 \end{bmatrix}$$

La **correlación** entre la l-ésima variable y la k-ésima variable está dada por:

$$cor(x_l, x_k) = s_{lk} / \sqrt{s_{ll} s_{kk}}.$$

Eigenvalores y eigenvectores Para poder hablar de eigenvalores y eigenvectores primero debemos de considerar una matriz cuadrada A de $(p \times p)$ y un vector $\mathbf{x}$ en $\Re^p$. Muchas aplicaciones se resuelven al encontrar vectores $\mathbf{x}$ tales que $\mathbf{x}$ y $A\mathbf{x}$ son paralelos. Para resolver este problema primero se presentarán algunas definiciones y conceptos.

Definición

Sea $\mathbf{A}$ una matriz de $p \times p$. El número real λ es llamado **eigenvalor** de $\mathbf{A}$, si existe un vector $\mathbf{x} \neq 0$ en $\Re^p$ tal que

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

Cada vector $\mathbf{x}$ diferente de cero que satisface la ecuación de arriba es llamado un eigenvector de A asociado al eigenvalor λ

A los eigenvalores también se les llama valores propios, valores característicos o valores latentes. Así mismo, a $det(\lambda I_n - A)$ se le llama polinomio característico de A.

Procedimientos para calcular valores y vectores propios

1. Encontrar $p(\lambda) = \det(A - \lambda \mathbf{I})$.

2. Calcular las raíces $\lambda_1, \lambda_2, \ldots, \lambda_p$ de $p(\lambda) = 0$.

3. Resolver el sistema homogéneo $(A - \lambda_i \mathbf{I})\mathbf{c}_i = \mathbf{0}$ que corresponde a cada valor característico de λ_i.

EJEMPLO para calcular los valores propios de una matriz.

Sea

$$A = \begin{bmatrix} 1 & 1 \\ -2 & 4 \end{bmatrix}$$

Se desean obtener todos los números reales λ y vectores $\mathbf{x} \neq 0$ que satisfacen $A\mathbf{x} = \lambda\mathbf{x}$, esto es,

$$\begin{bmatrix} 1 & 1 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Esta ecuación equivale a:

$$x_1 + x_2 = \lambda x_1$$

$$-2x_1 + 4x_2 = \lambda x_2$$

ó

$$(1 - \lambda)x_1 + x_2 = 0$$

$$-2x_1 + (4 - \lambda)x_2 = 0$$

Esta ecuación es un sistema homogéneo de dos ecuaciones con dos incógnitas. Este sistema tiene solución no trivial si y sólo si el determinante de la matriz de coeficientes es cero, esto es

$$\begin{vmatrix} 1-\lambda & 1 \\ -2 & 4-\lambda \end{vmatrix} = 0$$

Esto quiere decir que

$$(1-\lambda)(4-\lambda)+2=0$$

ó

$$\lambda^2 - 5\lambda + 6 = 0 = (\lambda-3)(\lambda-2).$$

por lo que

$$\lambda_1 = 2 \text{ y } \lambda_2 = 3$$

son los eigenvalores de A. De esta forma podemos calcular el eigenvector asociado a 2

$$\begin{bmatrix} 1 & 1 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Esta ecuación equivale a:

$$x_1 + x_2 = 2x_1$$

$$-2x_1 + 4x_2 = 2x_2$$

finalmente tenemos

$$x_1 - x_2 = 0$$

$2x_1 - 2x_2 = 0$

los vectores que satisfacen estas ecuaciones son de la forma $x_1 = x_2$ para $\lambda = 2$, por ejemplo, el vetor $\mathbf{x} = (1, 1)$ lo cumple.

De forma análoga los eigenvectores asociados $\lambda = 3$ son de la forma $x_1 = x_2/2$, por ejemplo el vector $\mathbf{x} = (1, 2)$ lo cumple.

Propiedades

Sea A una matriz de orden $p \times p$. Entonces

1. $\sum_{i=1}^{p} \lambda_i = \text{traza}(A)$.

2. $|A| = \det(A) = \prod_{i=1}^{p} \lambda_i = \lambda_1 * \lambda_2 * \cdots * \lambda_p$.

3. Si A es positiva definida , entonces $\lambda_i > 0$ $(i = 1, \ldots, p)$

4. Si A es una matriz de número reales simétrica, entonces sus eigenvalores y eigenvectores son reales.

5. si A es positiva semidefinida de rango r, entonces exactamnete r de los λ_i son positivos y $(p - r)$ son cero.

6. Si $\lambda_i \neq \lambda_j$ entonces los eigenvectores asociados son ortogonales,

$\mathbf{x_i}.\mathbf{x_j} = \mathbf{0}$.

Es decir si todos los λ_i son distintos, entonces L la matriz que tiene como columnas a los eigenvectores $\mathbf{x_i}$ es ortogonal $LL' = I$.

Diagonalización de matrices

Se dice que una matriz A es **diagonalizable** si puede escribirse como:

$$A = PDP^{-1}$$

donde P es una matriz invertible cuyos vectores columna son los eigenvectores de A y D es una matriz diagonal formada por los eigenvalores de A.

Una matriz A es diagonalizable si todas las raíces de su polinomio característico son reales y diferentes.

Si además la matriz P es ortogonal se dice entonces que la matriz A es **diagonalizable ortogonalmente**, pudiendo escribirse como

$$A = PDP'$$

Nota:

Sí todas las raíces del polinomio característico de A son reales y no todas diferentes, entonces A puede o no ser diagonalizable. El polinomio característico de A puede escribirse como el producto de n factores, cada uno de la forma $\lambda - \lambda_j$, donde λ_j es una raíz del polinomio característico. Ahora los eigenvalores de A son las

raíces reales del polinomio característico de A. De aquí que el polinomio característico se pueda escribir como

$$(\lambda - \lambda_1)^{k_1}(\lambda - \lambda_2)^{k_2} \ldots (\lambda - \lambda_r)^{k_r}$$

donde $\lambda_1, \lambda_2, \ldots, \lambda_r$ son los distintos eigenvalores de A,

$k_1 + k_2 + \ldots + k_r = n$

k_j es número entero

A k_j es la multiplicidad de λ_i. Se puede demostrar que si las raíces del polinomio característico de A son todas reales, entonces A

puede ser diagonalizada si y solo si para cada eigenvalor λ_i de multiplicidad k_j podemos encontrar k_j eigenvectores linealmente independientes.

Cualquier matriz cuadrada simétrica con coeficientes reales es ortogonalmente diagonalizable. A este resultado se le conoce como Teorema Espectral.

Ejemplo. Sea

$$A = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{bmatrix}.$$

Los eigenvalores y eigenvectores de A son : 5, -1 y -1

$$1/\sqrt{3}\begin{bmatrix}1\\1\\1\end{bmatrix},\ 1/\sqrt{6}\begin{bmatrix}-2\\1\\1\end{bmatrix}\ y\ 1/\sqrt{2}\begin{bmatrix}0\\-1\\1\end{bmatrix}.$$

respectivamente.

Entonces por la descomposición espectral se tiene que

$$A=\begin{bmatrix}1/\sqrt{3} & -2/\sqrt{6} & 0\\ 1/\sqrt{3} & 1/\sqrt{6} & -1/\sqrt{2}\\ 1/\sqrt{3} & 1/\sqrt{6} & 1/\sqrt{2}\end{bmatrix}\begin{bmatrix}5 & 0 & 0\\ 0 & -1 & 0\\ 0 & 0 & -1\end{bmatrix}\begin{bmatrix}1/\sqrt{3} & -2/\sqrt{6} & 0\\ 1/\sqrt{3} & 1/\sqrt{6} & -1/\sqrt{2}\\ 1/\sqrt{3} & 1/\sqrt{6} & 1/\sqrt{2}\end{bmatrix}$$

.

Matrices no-negativas definidas

Definición

Cuando $\mathbf{x}'\mathbf{A}\mathbf{x} > \mathbf{0}$ para todo $\mathbf{x}$ diferente a $\mathbf{x} = \mathbf{0}$ se dice que se $\mathbf{x}'\mathbf{A}\mathbf{x}$ es una forma cuadrática positiva definida, y A es una matriz positiva definida Definición

Cuando $\mathbf{x}'\mathbf{A}\mathbf{x} \geq \mathbf{0}$ para todo $\mathbf{x}$ y $\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{0}$ para algunos $\mathbf{x} \neq \mathbf{0}$ entonces $\mathbf{x}'\mathbf{A}\mathbf{x}$ es una forma cuadrática semi-positiva definida, y A es una matriz semipositiva definida

Definición

Los dos tipos de matrices tomados juntos, positivas definidas y positivas semi-definidas, son llamados no-negativas definidas

Si la matriz es simétrica tiene las siguientes propiedades:

1. Todos sus eigenvalores son reales.

2. Es diagonalizables.

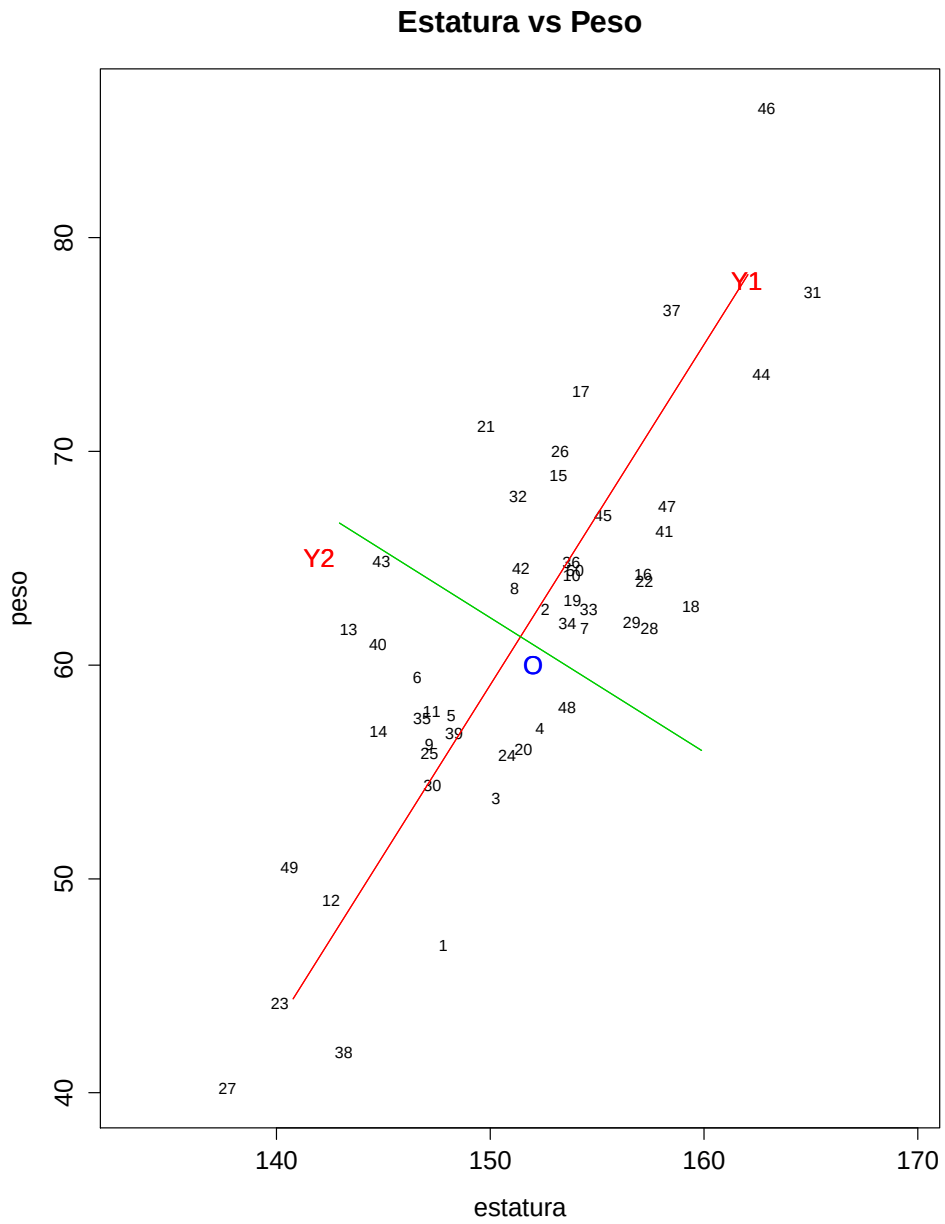
3. El rango iguala al número de eigenvalores diferentes de cero.

Los eigenvalores de una matriz simétrica son todos no-negativos si y solo si la matriz es no-negativa definida.

Componentes principales

Si se tienen mediciones x_1 como alturas y x_2 como pesos de un grupo de personas.

El análisis de componenetes principales tiene como objetivo reducir la dimensión y conservar en lo posible su estructura, es decir la ''forma''de la nube de datos.


Estatura vs Peso
peso
estatura
40
50
60
70
80
140
150
160
170
Y1
Y2
O

En este ejemplo se buscará proyectar los datos sobre un eje que reproduzca de la mejor manera la "forma" de la nube de datos.

El primer paso es centrar los datos en el centroide $(\bar{X}_1, \bar{X}_2)$ y después se hace una rotación, de manera que la "las proyecciones" sean lo más parecidas posibles a los vectores originales. En la figura de arriba puede verse que los individuos que quedan a la izquierda en el eje $OY1$, son los más pequeños en **talla**, y a la derecha los más grandes. Tomado en cuenta el otro eje $OY2$, los sujetos que quedan por encima de este son los aquellos que tienen un peso mayor dada la estatura, y por debajo los que tienen poco peso dada su estatura, es decir este eje habla de habla de la **forma** de los sujetos. Para estos sujetos ocurre que varían mucho en talla y hay poca variación en la forma.

Para este procedimiento se requiere girar los ejes, esto se consigue aplicando una transformación lineal a los datos, esto es

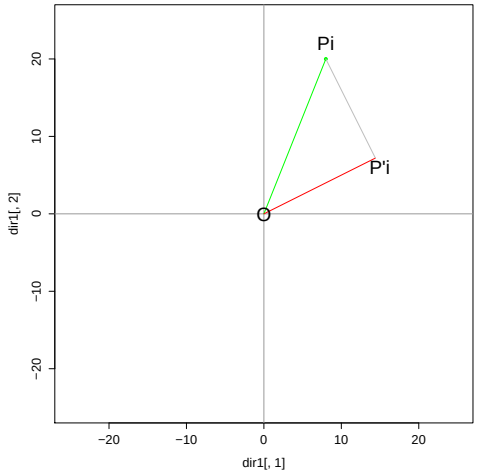
$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} X_1\cos\alpha + X_2\sin\alpha \\ -X_1\sin\alpha + X_2\cos\alpha \end{bmatrix}$$

Las proyecciones de los puntos sobre el eje $OY1$ son una buena aproximación a los datos, ya que en la otra dirección hay poca variación.

Entonces se puede usar únicamente

$Y_1 = X_1\cos\alpha + X_2\sin\alpha$, y así la nueva variable Y_1 resume a las otras dos.

Proyección de un punto sobre un nuevo eje



Por Teorema de Pitágoras se tiene que:

$$(OP_i)^2 = (OP_i')^2 + (P_iP_i')^2.$$

La cantidad $(P_iP_i')^2$ puede ser vista como error.

Si se hace la sumatoria sobre todos los sujetos desde $i = 1, \ldots, n$ y se divide entre $n - 1$ se tiene

$$C = \frac{\Sigma(OP_i)^2}{n-1} = \frac{\Sigma(OP_i')^2}{n-1} + \frac{\Sigma(P_iP_i')^2}{n-1}$$

El objetivo es entonces minimizar la cantidad $\frac{\Sigma(P_iP_i')^2}{n-1}$

Ahora $\Sigma(OP_i)^2$ es una cantidad fija, no depende de los ejes de coordenadas, y por tanto minimizar $\frac{\Sigma(P_iP_i')^2}{n-1}$ es equivalente a maximizar $\frac{\Sigma(OP_i')^2}{n-1}$, esta última cantidad coincide con la **varianza de las proyecciones sobre eje** $OY1$,es decir que el ángulo de rotación que se busca es aquel que MAXIMIZE la varianza de las proyecciones.

En general se tienen datos en un espacio de p dimensiones entonces se busca la transformación lineal $a'x = y_1$ de manera que tenga máxima varianza, se conoce como primer componente principal. Si el vector x tiene matriz de varianzas y covarianzas Σ entonces $var(y_1) = var(a'x) = a'\Sigma a$

Se buscan también otras combinaciones lineales Y_i de las variables originales

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p$$

que sean ortogonales entre si (y_i ortogonal a y_j) y que de manera sucesiva vayan maximizando la varianza.

Digamos

$y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p.$

Elegir a_1 de manera que la $var(y_1)$ sea de máxima varianza.

Si a se toma de norma muy grande, entonces la varianza de y_1 puede ser tan grande como se quisiera, de manera que se deben imponer condiciones a a_1 para acotar el tamaño de varianza, la condición que se impone es

$\|a_1\| = a_1' \cdot a_1 = 1$

De esta manera obtenemos

$$var(y_1) = var(a_1'\bar{x}) = a_1'\Sigma a_1$$

Esta es la función objetivo, es decir, debemos encontrar que vector a_1 es el que maximiza $a_1' \Sigma a_1$ y además $a_1' \cdot a_1 = 1$.

Para lograr maximizar estas condiciones se usan multiplicadores de Lagrange:

$f(x_1, x_2, \ldots, x_p)$ sujeto a $g(x_1, x_2, \ldots, x_p) = c$,

donde f es una función diferenciable. Existe una λ tal que $\frac{\partial f}{\partial x_i} - \lambda \frac{\partial g}{\partial x_i} = 0$. en este caso es

$$L(a_1) = a_1' \Sigma a_1 - \lambda(a_1' a_1 - 1)$$

$$\frac{\partial L}{\partial a_1} = 2\Sigma a_1 - 2\lambda a_1$$

al derivar e igualar a cero, se tiene $(\Sigma - \lambda I)a_1 = 0$ entonces resulta que a_1 es eigenvector de Σ y λ su eigenvalor, que equivale a decir que

$|\Sigma - \lambda I| = 0.$

Sean $\lambda_1, \lambda_2, \ldots, \lambda_p$ los eigenvalores que satisfacen

$\lambda_1 > \lambda_2 > \ldots > \lambda_p \geq 0$ y $\Sigma = A\Lambda A'$ donde $\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix}$

NOTA. Esto se da pues la matriz es simétrica y semi positiva definida.

¿Cual de ellos determina a la primer componente?

$var(a_1' X) = a_1' \Sigma a_1$

$= a_1' \lambda I a_1$

$= \lambda$

Como interesa que sea el que maximize la varianza, λ es el mayor de los λ_i, digamos λ_1.

Entonces a_1 es el eigenvector asociado a λ_1.

Para buscar la segunda componente $y_2 = a_2' X$, se impone también la condición $a_2' a_2 = 1$, y además y_2 debe ser no correlacionada a y_1, entonces su covarianza debe ser cero. Haciendo el desarrollo se tiene

$$cov(y_2, y_1) = cov(a_2' X, a_1' X)$$

$$= E[a_2'(x - \mu)(x - \mu)' a_1]$$

$$= a_2' \Sigma a_1$$

Se sabe que

$\Sigma a_1 = \lambda_1 a_1$ y sustituyendo en la expresión anterior se tiene

$$\lambda_1 a_2' a_1 = a_2' \lambda_1 a_1 = 0$$

$$\Leftrightarrow \;\; a_2' a_1 = 0$$

es decir que a_1 es perpendicular a a_2.

$$L(a_2) = a_2' \Sigma a_2 - \lambda(a_2' a_2 - 1) - \delta a_2' a_1$$

$$\frac{\partial L}{\partial a_2} = 2(\Sigma - \lambda I) a_2 - \delta a_1 = 0$$

Premultiplicando esto por a_1' y operando

$$2 a_1' \Sigma a_2 - \delta = 0$$

como $a_1' a_2 = 0$ y como también $a_1' \Sigma a_2 = 0$ (no correlacionado), se tiene que δ debe ser cero, entonces la ecuación que interesa es $(\Sigma - \lambda I)a_2 = 0$ y de acuerdo a esto λ_2 corresponde al segundo eigenvalor y a_2 al segundo eigenvector.

Cuando hay eigenvalores iguales se eligen eigenvectores ortogonales.

Sea $A = \left[\begin{array}{cccc} \bar{a}_1 & \bar{a}_2 & \ldots & \bar{a}_p \end{array} \right]$

Sea $Y_{p \times 1}$, el vector de las componentes principales.

$Y = A'X$

La matriz de covarianzas de Y es Λ y esta dada por

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ & & \ddots & & \\ 0 & 0 & 0 & \dots & \lambda_p \end{bmatrix}$$

$var(Y) = A'\Sigma A = A'A\Lambda A'A = \Lambda$

$traza(\Lambda) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} var(y_i)$

$traza(\Lambda) = traza(A'\Sigma A) = traza(\Sigma AA') = traza(\Sigma) = \sum_{i=1}^{p} var(x_i)$

$\sum_{i=1}^{p} var(y_i) = \sum_{i=1}^{p} var(x_i)$

Esto es útil para determinar el número de componentes a utilizar.

Si se considera como **varianza generalizada** a $\sum_{i=1}^{p} {\sigma^2}_i = traza(\Sigma)$, entonces

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

De esta forma tenemos que

$\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}$ nos dice el porcentaje de la varianza generalizada que es explicado por la componente j-ésima y

$\frac{\sum_{i=1}^{j} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$ nos da el porcentaje de la varianza generalizada dado por las primeras j componentes principales.

La covarianza entre x_i y y_j es el vector $\Sigma a_j = \lambda_j a_j$ entonces

$cov(x_i, y_j) = \lambda_j a_{ij}$ y

$corr(x_i, y_j) = \frac{\lambda_j a_{ij}}{\sqrt{\lambda_j}\sqrt{\sigma^2}_i}$

$= \frac{\sqrt{\lambda_j} a_{ij}}{\sigma_i}$

NOTA.- Como se desconoce Σ, todo en la práctica se hace con su estimador S.

Cuando se trabaja con distintas unidades conviene hacer el análisis con la matriz R de correlaciones. Los eigenvectores de R y S no coinciden, y no hay una forma de obtener unos a partir de los otros.

Distancias y disimilitudes

Para entrar al tema de análisis de conglomerados es muy importante introducir los conceptos de distancia y disimilitud.

Todos conocemos la llamada distancia euclideana Si $x = (x_1, x_2, \ldots, x_p)$ y $y = (y_1, y_2, \ldots, y_p)$ entonces la distancia entre x y y es:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_p - y_p)^2}$$

Producto punto

Sean $\bar{x} = [x_1, x_2, \ldots, x_p]$ y $\bar{y} = [y_1, y_2, \ldots, y_p]$ dos vectores p-dimensionales. Entonces el producto punto entre los vectores

se define como $\bar{x} \cdot \bar{y} = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n = \sum_{i=1}^{m} x_i y_j$. Al producto punto también se le llama producto interior o escalar.

Ejemplo el producto punto

$$\begin{aligned} x \cdot y &= (4,1,2,3) \cdot (3,1,7,2) \\ &= (4)(3) + (1)(1) + (2)(7) + (3)(2) \\ &= 33. \end{aligned}$$

Se tiene que: $d(x,y)^2 = (x-y) \cdot (x-y)$.

La distancia puede generalizarse introduciendo una matriz$A > 0$ es decir una matriz positiva definida con dimensiones $p \times p$ esto es: $(x-y)' A (x-y)$, cuando $A = S^{-1}$ donde **S** es la matriz de varianzas y covarianzas de las variables que conforman a los vectores.

Hay otras distancias como la *city block* que corresponde a:

$$d(x, y) = |(x_1 - y_1)| + \ldots + |(x_p - y_p)|$$

Y también existe la generalización:

$$d(x, y) = ((x_1 - y_1)^\alpha + \ldots + (x_p - y_p)^\alpha)^{1/\alpha}$$

con α un entero,conocida como distancia Minkowski.

Para que una función entre dos puntos sea considerada distancia debe cumplir con las siguientes propiedades:

$$d(x, y) > 0 \; \forall x \neq y$$

$$d(x, y) = 0 \Leftrightarrow x \equiv y$$

$$d(x, y) \leq d(x, z) + d(z, y) \; \forall x, y, z$$

Se puede tener otro enfoque, que tal que la disimilitud que quiere calcularse es entre (**columnas** de X), en este caso el coeficiente de correlación es una buena medida. Antes de construir las medidas hay que decidir si un coeficiente de correlación alto pero negativo significa un acercamiento grande entre las variables, o un total alejamiento.

¿Qué hacer si se tienen exclusivamente variables binarias???,a continuación se presentan varis medidas de disimilitud. Si se tienen dos sujetos, i y j y si se denota por

$a =$ todas las coincidencias $++$

$b =$ todas las coincidencias $-$

$c =$ todas las coincidencias +-

$d =$ todas las coincidencias -+

p es el total de características $a + b + c + d = p$.

disimilitud entre i y j es

i) $d_{ij} = 1 - \frac{coincidencias}{p} = 1 - \frac{a+d}{p} = \frac{b+c}{p}$

esta **disimilitud** (en ocasiones no cumple con las propiedades de las distancias) corresponde a la proporción de variables que no coinciden.

Hay quienes sostienen que no deben tomarse en cuenta las coincidencias en ausencias, entonces: ii) Coeficiente de Jaccard

$d_{ij} = \frac{b+c}{a+b+c}$

d no se toma en cuenta pues la ausencia de cierta característica no ayuda a decir si son o no parecidos.

iii)Coeficiente de Czekanowski

$d_{ij} = \frac{b+c}{2a+b+c}$ si se quiere compensar el echar fuera a b

Para datos cualitativos

$c_{ij} =$ coeficiente de coincidencias de Sneath

$$c_{ij} = \frac{\text{número de atributos en los que las unidades coinciden}}{p}$$

$$d_{ij} = 1 - c_{ij}$$

Se puede utilizar el índice de Gower para crear distancias cuando se consideran datos con varios tipos de escalas de medición.

Para cada variable x_k la similitud entre individuo i y el j se escribe como

$\delta_{ijk} = 1$ si puedo comparar a i contra j en la variable k o

$\delta_{ijk} = 0$ en otro caso.

y a $s_{ijk} = 1$ si son iguales o $s_{ijk} = 0$ si son diferentes.

La **similitud** entre i y j esta dada por

$$c_{ij} = \frac{\sum_{k=1}^{p} s_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}$$

La disimilitud como $d_{ij} = 1 - c_{ij}$

Si las x_k son cuantitativas

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{rango(x_k)}.$$

Si x_k es cualitativa

$s_{ijk} = 1$ si los individuos concuerdan en la variable k y $s_{ijk} = 0$ si no.

Si se requiere construir disimilitud entre variables y se trabaja con datos cualitativos

Se construye una tabla de contingencias tamaño 2×2 donde $a + c + b + d = n$

$$d_{kl} = \chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(a+c)(c+d)(b+d)}$$

pero depende del tamaño de la muestra $\chi^2 \leq n$.

Otra medida que puede usarse es

$$d_{kl} = 1 - \sqrt{\frac{\chi^2}{a+b+c+d}}.$$

Análisis De Conglomerados

El objetivo de este análisis es formar grupos de observaciones, de manera que todas las unidades en un grupo sean similares entre ellas pero que sean diferentes a aquellas de otros grupos. La parte interesante es definir que es similar. Si por ejemplo el rango de las disimilitudes corre de 0 a C, podríamos definir que dos unidades son consideradas similares si su disimilitud es menor a $\frac{1}{2}C$.

Hay métodos **jerárquicos y no- jerárquicos**. Dentro de los jerárquicos están los aglomerativos y los divisivos.

Los métodos jerárquicos aglomerativos dan origen a un gráfico llamado **dendograma** . Estos método son iterativos y en cada paso debe recalcularse la matriz de distancias, en los casos en

que se tienen muchas observaciones esto puede llevar mucho tiempo de cómputo.

A continuación se presentan varios de estos métodos:

- **Liga sencilla o del vecino más cercano**: La distancia entre dos conglomerados, es la mínima de todas las posibles distancias entre objetos que pertenecen a distintos conglomerados. Se unen los dos conglomerados más cercanos. Este método tiende a tener un buen desempeño cuando hay grupos de forma elongada, conocidos como tipo cadena.

- **Liga completa o del vecino más lejano**: La distancia entre dos conglomerados, es la máxima de todas las posibles

distancias entre objetos que pertenecen a distintos conglomerados. Se unen los dos conglomerados más cercanos.Se desempeña bien cuando los conglomerados son de forma circular.

- **Liga promedio**: La distancia entre dos conglomerados, es el promedio de todas las posibles distancias entre objetos que pertenecen a distintos conglomerados.Se unen los dos conglomerados más cercanos. Funciona bien para conglomerados tanto tipo cadena como circulares.

- **Liga centroide**: La distancia entre dos conglomerados está definida como la diferencia entre las medias (centroide) de cada conglomerado.Se unen los dos conglomerados más cercanos.

- **Liga mediana**: La distancia entre dos conglomerados está definida como la diferencia ponderada entre las medias (centroide) de cada conglomerado, la ponderación está dada por el tamaño (número de unidades del conglomerado) de los conglomerados.Se unen los dos conglomerados más cercanos. Se espera un mejor rendimiento que el del centroide cuando los grupos varían mucho en tamaño.

- **Método de Ward**: Se unen los dos conglomerdos que arrojen la menor varianza intragrupo (witih group) (con respecto al centroide de cada conglomerado).

Sólo las dos primeras ligas son invariantes ante transformaciones monótonas de las disimilitudes d_{ij} En algunos caso se puede

trabajar con sólo la matriz de disimilitudes, en otros se requieren los datos originales.

Se construye el dendograma y se hace un corte a la altura máxima de la disimilitud que se fije para considerar a dos sujetos como similares.

En los métodos divisivos R tiene uno llamado DIANA.

Entre los no jerárquicos están:

- **k medias (k means)**: se requiere de antemano dar el número de grupos existente. Es un método iterativo, por lo que requiere de una solución inicial, y de allí se va optimizando una

función objetivo. En cada paso se reasignan los elementos del conglomerado de manera tal que la suma de distancias al cuadrado de los puntos al centro de su conglomerado sea mínima.

- **basado en un modelo de mezcla de normales**: éste considera además diferentes estructuras de covarianza, es un método bayesiano, y usa estimación tipo EM.

- **de dos pasos**: se recomienda cuando se tienen muchos datos.

  1. **Formación de Preconglomerados**. La idea es reducir el tamaño de la matriz de distancias entre todos los posibles elementos. Los preconglomerados son conglomerados

de los datos originales que se usan después en un método jerárquico. Se lee un caso, el algoritmo decide si se une a los conglomerados ya existentes, o forma un nuevo conglomerado. Una vez acabado el proceso de preaconglomeramiento, cada conglomerado es tratado como un elemento, la nueva matriz de distancias se reduce al número de preconglomerados.

2. **Agrupación de preconglomerados**. Se usa un método jerárquico estándar usando como elementos a los preconglomerados.

En general cada método puede llevar a una agrupación diferente. Se recomienda usar varios métodos para identificar los individuos

que brincan de un grupo a otro según el método usado, tambien para definir los tamaños de los grupos.

Análisis de discriminante

A diferencia del análisis de conglomerados aquí se busca construir una regla de asignación de individuos a distintos grupos **ya dados**, entonces la muestra que se tiene puede considerarse como de entrenamiento, y cuando lleguen nuevos individuos ya se tiene la manera de asignarlos a uno de los grupos.

Veremos el análisis de **discriminante lineal** y el **discriminante cuadrático**. Para su derivación se supondrá que los datos son normales.

El primer caso, el de discriminante lineal considera que los grupos tienen distintas medias pero **comparten la misma matriz de varianzas y covarianzas**. Se verá que esto llevará a construir hiperplanos que corten al espacio de observaciones en los distintos grupos. En el segundo caso se considera que cada grupo tiene su propia matriz de varianza y covarianza, aquí los cortes en vez de hacerse de manera plana se hacen con curvas. Si se tiene un número g de grupos entonces se requieren $g-1$ hiperplanos (curvas) para partir al espacio.

Análisis de discriminante lineal

cuando hay dos grupos

Suponemos que la población π_i es normal multivartiada de dimensión p con $i = 1, 2$

$$f_i(x) = (2\pi)^{-p/2}|\Sigma_i|^{-1/2}\,\text{exp}[-1/2(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)]$$

$$\frac{f_1(x)}{f_2(x)} = |\Sigma_1|^{-1/2}|\Sigma_2|^{1/2}\,\text{exp}[-1/2\{x'(\Sigma_1^{-1}-\Sigma_2^{-1})x-2x'(\Sigma_1^{-1}\mu_1-\Sigma_2^{-1}\mu_2)+ \quad (2)$$

En el caso de que $\Sigma_1 = \Sigma_2 = \Sigma$ ocurre que

$$|\Sigma_1|^{-1/2}|\Sigma_2|^{1/2} = 1,$$

$$x'(\Sigma_1^{-1} - \Sigma_2^{-1})x = 0,$$

$-2x'(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2) = -2x'(\Sigma^{-1})(\mu_1 - \mu_2)$ y

$$\text{y}\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2 = (\mu_1+\mu_2)'(\Sigma_1^{-1})(\mu_1-\mu_2)$$

Entonces

$$\frac{f_1(x)}{f_2(x)} = \exp[-1/2\{(-2x + (\mu_1 + \mu_2))'(\Sigma^{-1})(\mu_1 - \mu_2)$$

Cuando$\frac{f_1(x)}{f_2(x)} \geq k$ donde $k \geq 1$ concuerda que debemos asignar a x a la población π_1.

Si $k = 1$ esta regla también puede escribirse como:

$$x'(\Sigma^{-1})(\mu_1 - \mu_2) \geq 1/2(\mu_1 + \mu_2)'(\Sigma^{-1})(\mu_1 - \mu_2)$$

Esta expresión es una función lineal en términos de x y de allí su nombre.

NOTA A esta misma expresión llegó Fisher **sin suponer la normalidad**, sólo hallando la dirección en la que la proyección de

los centros de ambos grupos se hace lo más alejada posible. Entonces cuando los datos se alejan mucho de la normalidad hace que la función discriminante no funcione muy bien.

¿Y si hay más grupos?

Se tienen varias posibles reglas de clasificación por parejas, algunas de ellas resultan **redundantes**, en realiidad con g grupos sólo $g-1$ son necesarias.

Análisis de discriminante cuadrático

Cuando $\Sigma_1 \neq \Sigma_2$ el término $x'(\Sigma_1^{-1} - \Sigma_2^{-1})x$ que está dentro la exponencial en la ecuación (1)no se cancela y este es un término

cuadrático y da lugar a que la regla de asignación involucre a una **curva**.

Una vez creada la regla de asignación, conviene construir medidas para ver que tan buena es, interesa entonces conocer cuántos de la muestra quedan bien o mal clasificados, también se usan medidas de desempeño tipo *jack knife*, y tipo validación cruzada.

Además de los discriminantes lineal y cuadrático, se puede utilizar la regresión logística para construir una regla de asignación cuando se trata de dos grupos sólamente o también se puede usar algún método bayesiano.

Análisis de discriminante Logístico

En un anális de regresión logística se tiene un modelo de la forma:

$$P(i \in \pi_1|\mathbf{x}_i) = \frac{\exp(\mathbf{x_i'}\beta)}{1+\exp(\mathbf{x_i'}\beta)} = \frac{\exp(\beta_o + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}{1+\exp(\beta_o + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}$$

y

$$P(i \in \pi_2|\mathbf{x}_i) = \frac{1}{1+\exp(\mathbf{x_i'}\beta)}$$

Una vez que se estiman las βs, se calculan las probabilidades y se asigna a la poblacion π_1 si $P(i \in \pi_1|\mathbf{x}_i) > 0{,}5$ y ala población π_2 en caso contrario.

Análisis de discriminante basado en funciones de probabilidad

Cuando es posible tener funciones de probabilidad f_1 y f_2, entonces la regla discriminante para x es: asignar a la población π_1 si $f_1(x) > f_2(x)$, ahora si se conocen las probabilidades *apriori* q_1 la probabilidad de que x provenga de la población π_1 y q_2, la de provenir de π_2 con $q_1 + q_2 = 1$ entonces asignar x a la población π_1 si

$$\frac{f_1(x)}{f_2(x)} > \frac{q_2}{q_1}$$

Es decir la observación se asigna a la población que tenga la verosimilitud más alta.

La probabilidad de clasificación errónea sería:

$$q_1 P(2|1) + q_2 P(1|2)$$

donde $P(2|1)$ y $P(1|2)$ son las probabilidades de clasificación errónea de cada población.

Usando teorema de Bayes la probabilidad *aposteriori* de que un individuo con valores observados x_o provenga de la población π_i es

$$q(\pi_i|x_o) = \frac{q_i f_i(x_o)}{q_1 f_1(x_o) + q_2 f_2(x_o)},$$

entonces se debe asignar a x_o a la población que tenga la probabilidad *aposteriori* más alta.

NOTA

Si de las $f_i(x)$ no se sabe gran cosa, éstas deben ser **estimadas**, por ejemplo con estimadores **no paramétricos** tipo kernel, tipo *spline*, suavizamientos,etc.

No es raro encontrar en ciertos anális de datos que se usa primero un análisis de conglomerados y después uno de discriminante para ver que tan bien quedaron formados los grupos.

Comandos en R

Hist, boxplot, stem, script.., barplot, pieplot, parallel, stars, faces, Andrews curves, bagplot. Dist, Mclust, hclust, Diana, clara, kmeans,agnes

dicrim