U_1 / U_2

Estadística	Data Mining
• Procedimiento hipotético deductivo	• Procedimiento inductivo
• Supuestos iniciales	• Sin supuestos iniciales
• Sin herramientas informáticas	• Amplia difusión entre computadoras

Variables

- Categóricas ó Cualitativas: Son diferentes sin orden (sexo, color pelo, tipo auto, etc)
- Casicuantitativas u Ordinales: Son ordenables, pero no se puede establecer una distancia entre ellas. (estado enfermedad: I, II, III)
- Cuantitativas discretas: Valores numéricos sin completitud (cantidad de hijos, matrículas, plata, etc)
- Cuantitativas continuas: Valores numéricos con completitud (edad, peso, duración, etc)

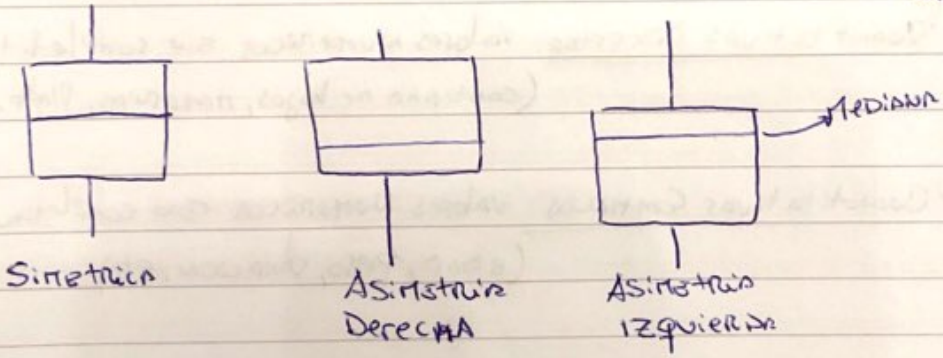
Frec. Absoluta: #observaciones de cada modalidad/intervalo

Frec. Relativa: $\frac{\text{Frec. Absoluta}}{\#\text{total observaciones}}$

M. Tendencia Central:
- Media / Promedio
- Media α-Podada
- Mediana
- Moda

M. Dispersión
- Rango Muestral (MAX-MIN)
- Varianza Muestral: Promedio de los cuadrados de las distancias de las obs. a Media
- Coeficiente de Variación
- MAD (Median Absolute Deviation)

M. Posición o estadísticos d/orden
- Cuantiles — Deciles (10)
  - Percentiles (100)
  - Cuartiles (4



INFO
- Mult.variada

	VAR-1	VAR 2	...	VAR-N
IND-1				
IND-2				
:				
IND-N				

N filas = #individuos

p-variables = #variables a analizar

- Grafico de Mosaico: Distr. Conjuntas Multivariadas

Singular = Det() = 0 = L. dependiente

Dispersograma: variables cuantitativas, vinculo de a pares de variables

G. Estrellas: Cuantitativas para detectar similitudes

C. Chernoff: ↗

Posición y Dispersión Multivariadas

Vector medias muestral: $\bar{x} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)$ donde $\bar{x}_i$ promedio de la i-esima variable.

Matriz de Varianzas y Covarianzas Muestral

$$\hat{\Sigma} = \frac{1}{m}(X - \bar{X})^t (X - \bar{X})$$

$\hat{\Sigma}$ tamaño $p \times p$, Simetrica, La diagonal son las varianzas muestrales de cada variable observada y fuera de la diagonal esta la covarianza de cada par de variables

$\hat{\Sigma}$ < Simetrica; Semi-definida positiva; Autovalores ≥ 0

q_{ij} < >0 Asoc. lineal (+); <0 Asoc. lineal (−); $=0$ No hay Asoc. lineal

Matriz De Correlación

↳ Mat. de Var y Cov Normalizada

$\frac{q_{ij} - \text{Media}}{\sigma}$ Normalizo si Media y σ son buenos estimadores

[La Cov Detecta Asociación lineal unicamente]

r_{ik} Sensibles y S_{ik} outliers

$$r_{ik} = \frac{S_{ik}}{\sqrt{S_{ii}}\sqrt{S_{kk}}}$$

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1N} \\ r_{21} & 1 & & \\ \vdots & & 1 & \\ r_{N1} & r_{N2} & \cdots & 1 \end{bmatrix}$$

Traza de Σ o R es (+)

Correlograma: Visualiza la fuerza y sentido de la correlación entre un conjunto de variables
- Azul: correlación (+)
- Rojo: correlación (−)

Alternativas Robustas

↳ Alternativa a las clásicas cuando hay presencia de outliers.

Efecto de Outliers →
- Enmascaramiento: un grupo esconde otro grupo.
- Inundación: una observación es outlier solo si hay presencia de otra.

/// Dist Mahalanobis: considera la correlación entre variables

/// Vector D/medianas: Sustituir el V de medias por el de medianas y calcular la mat. de covarianzas computando la menor Dist. Mahalanobis

/// MVE (Minimun Volume Elipsoid): Buscar el elipsoide de menor volumen que cubra m de las n observaciones.

/// MCD (Minimun Covariance Determinant): minimiza el Det de la matriz de covarianzas de m ~~observaciones~~ observaciones de las n que hay.

U3// : ACP

$A \in \mathbb{R}^{m \times n}$ con autovalores $\lambda_1, \lambda_2, \dots, \lambda_m$

$$\text{Traza}(A) = \sum_{i=1}^{m} \lambda_i$$

$$\text{Det}(A) = \prod_{i=1}^{m} \lambda_i$$

Reducción de Dimensión → ACP

- Transforma un conjunto de variables correlacionadas en un cjto de variables **No** correlacionadas, de menor dimensión que se obtiene a partir de combinaciones lineales de las variables originales

- Técnica descriptiva, exploratoria, no tiene supuestos ⇒ "Siempre puedo aplicarse"

- Elige comb. lineales de las variables que **maximizan** la varianza (minimizan la pérdida de info inicial)

- Todas las variables juegan el mismo papel, no existen indep y dep.
- Descarta información redundante
- Alternativa para visualizar info multidimensional
- Solo tiene sentido si las variables originales están fuertemente correlacionadas.

- Es recomendable el ACP con la M. de correlaciones (R) ya que Σ se ve influenciada por la variabilidad del conjunto.

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \dots + a_{1N}X_N$$

$$Y_2 = a_{21}X_1 + a_{22}X_2 + \dots + a_{2N}X_N$$

donde $\underline{a_i} = (a_{i1}, a_{i2}, \dots, a_{ip}) \in \mathbb{R}^N$ es el vector de **cargas/loadings**

Buscan comb. lineal de variables que maximicen variabilidad.

La variabilidad de Y_1 es máxima cuando el vector de cargas es el autovector asociado al mayor autovalor de Σ (mat de var y cov)

Y_2: Hago lo mismo con las variables originales NO correlacionadas con Y_1.

Proporción de Variabilidad de Y_1: $\frac{\lambda_1}{tr(\Sigma)} = \frac{\lambda_1}{\lambda_1+\lambda_2+\dots+\lambda_N}$

Comp Principales

Criterios de Parada

1: Porcentaje de variabilidad explicada.

2. Criterio de Kaiser: Retener las m primeras componentes / sus autovalores resultan ≥ 1

$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N \geq 1$

Algunos autores recomiendan ≥ 0.7

Se puede extender a Σ como $\geq \frac{tr(\Sigma)}{p}$

3: Criterio del bastón roto: (Gráfico S/Sedimentación) Se para cuando la variabilidad explicada se estabiliza a partir de un cierto valor.

4: Prueba de esfericidad: En un punto deja de haber direcciones de máxima variabilidad ⇒ la distribución tiene forma / empieza a tener forma de esfera.

Estimación, se usa $\hat{\Sigma} = S$

Autovalores corresponden a la varianza de la componente $\Rightarrow$ (Desvío estandar)^2

Comúnmente usar R por sobre Σ

CARGAS

- De y_i es ⊕ $\Rightarrow$ La variable $\wedge$ la componente tienen correlación ⊕
- Si es ⊕ $\Rightarrow$ un individuo que tenga puntuación alta en esta variable tendrá valores más alto de esta variable que otro individuo
- y_i es ⊖ $\Rightarrow$ correlación ⊖
- Dos individuos, con puntuaciones similares en las restantes variables, el que tenga más alta en esta se ubicará en un valor menor de la componente.

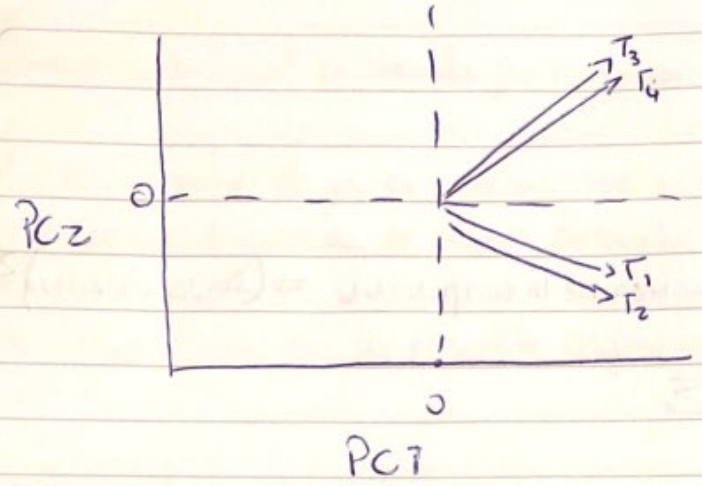
Primera Componente tiene todas cargas positivas (o negativas) $\Rightarrow$ Componente de Tamaño: Individuo con alto en esta tiende en todas las trazas (rapidez)

Segunda componente con valores intercalados (⊕ $\wedge$ ⊖) $\Rightarrow$ Componente de Forma contrasta las primeras 2 trazas con las ultimas dos.

- Si tengo pocos individuos los grafico todos, con muchos agrupo

biPlot: Las proyecciones sobre las componentes nos dan los loadings

- $\measuredangle$ chico $\Rightarrow$ mucha correlación
- variables ($T_1, T_2, T_3 \ldots$) muy o poco correlacionadas.



- Las 4 proyectadas sobre PC_1 son ⊕ "Tamaño"
- T_3 ∧ T_4 proyectadas sobre PC_2 son ⊕ ∧ T_1 ∧ T_2 son ⊖ "Forma"
- Ptos cerca del origen son "promedio"

Tamaño ∧ Forma son no correlacionadas

Sino tengo la muestra y sólo tengo Σ y R no puedo obtener los scores (Puntos en la PC)

ACP-Robustas

- Outliers pueden distorsionar la Mat. de Cov Muestral
- Métodos con menos supuestos pero computacionalmente costosos.

MCD (Mínimum Covariance Det)
Estimador Stahel-Donoho
MVE (Mínimum Volume Elipsoide)

} Opciones para reemplazar el vector de medias y la matriz de Cov.