Componentes Principales

Dra. Marta Quaglino
2023

Generalidades

- Componentes Principales así como otras técnicas multivariadas de reducción o síntesis de la información, fueron inicialmente desarrolladas a principios del siglo XX (Pearson, 1901) y retomado posteriormente en los años 30 (H.Hotelling). Sin embargo, se popularizan mucho después, con la aparición y acceso a las computadoras.
- La técnica se aplica en situaciones donde la matriz de información recoge mediciones <u>cuantitativas</u> (p) sobre cada individuo u objeto (n) (ya sean datos poblacionales o de una muestra)
- No reconoce poblaciones, ni diferentes roles entre las variables, y se aplica como análisis descriptivo complejo.

Objetivo

- El objetivo principal es: **Resumir** la información de la matriz de datos <u>conservando las diferencias entre individuos</u>
- Esta síntesis debe conducir a descubrir que aspectos (o factores complejos) diferencian a los individuos
- Las CP pretenden mostrar lo que se observaría en un grafico de los “individuos” en el espacio de las “variables”, si es que se pudiera visualizar un grafico en R^p
- Este enfoque, basado en los individuos u objetos, se sintetiza como análisis en R^p

Como se logra el objetivo?

- Se pretende reflejar la realidad p-dimensional en un “espejo" de menor dimensión (visible)
- Este espejo, es un sub-espacio de proyección que <u>conserve las diferencias y parecidos entre individuos</u> (inter-distancias)
- Tal sub-espacio deberá entonces conservar variabilidad entre individuos (porqué?)
- Al conservar variabilidad mostrara un resumen de la configuración original de puntos, con mínima deformación (por supuesto que también, con alguna perdida)

Búsqueda del sub-espacio

- La búsqueda se puede hacer en etapas: un eje por vez (se buscan varias direcciones ortogonales)
- La dirección del eje de proyección estará dada por un versor $\boldsymbol{\alpha}' = (\alpha_1, \alpha_2, \ldots \alpha_p)$, cuyas componentes son cosenos directores y deben cumplir con $\sum \alpha_i^2 = |\boldsymbol{\alpha}| = \boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$
- La proyección sobre la dirección del eje, de un punto del espacio original R^p se escribe como el producto:

$$\alpha' X = Y = CP$$

- Notar que *Y* es <u>una</u> nueva variable (por ahora) de dimensión 1.

Como encontrar el versor *a*?

- De modo que cumpla la condición impuesta sobre la dirección que define:
- MINIMIZAR LA DEFORMACION DE LA NUBE DE PUNTOS AL PROYECTARLA
- α / $\min \sum_{i=1}^{n} \sum_{j=1}^{n} \left[d(X_i, X_j) - d(Y_i - Y_j) \right]$
- Las interdistancias entre puntos en el espacio original $d(\boldsymbol{X}_i, \boldsymbol{X}_j)$ son constantes (y mas grandes que las distancias entre proyecciones) $\Rightarrow$
- α / $\max \sum_{i=1}^{n} \sum_{j=1}^{n} d^2(Y_i, Y_j) = \max Var(Y)$
- Demostrar, sabiendo que d^2(Yi,Yj)= mod(Yi-Yj)= =(Yi-Yj) (Yi-Yj)' (Ejercicio)

Extremos condicionados.

- Hay que maximizar una función de *p* variables, sujeta a una restricción, usamos multiplicadores de Lagrange: (usamos ***S*** pensando ***Σ*** no conocida)

$$\phi = Var(Y) - \lambda(\alpha'\alpha - 1) = \alpha' S\alpha - \lambda(\alpha'\alpha - 1)$$

- El sistema de derivadas parciales igualadas a cero es: $\frac{\partial\phi}{\partial\alpha} = 2S\alpha - \lambda 2\alpha = \underset{\sim}{0} \Rightarrow (S - \lambda I)\alpha = \underset{\sim}{0}$
- Por lo cual α es autovector normalizado de *S*, matriz de variancias y covariancias, asociado al autovalor λ. Cual de los *p* posibles?
- λ debe corresponder al mayor de los *p* autovalores de S: $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots\ldots \geq \lambda_p$

Porque el mayor autovalor de *S*?

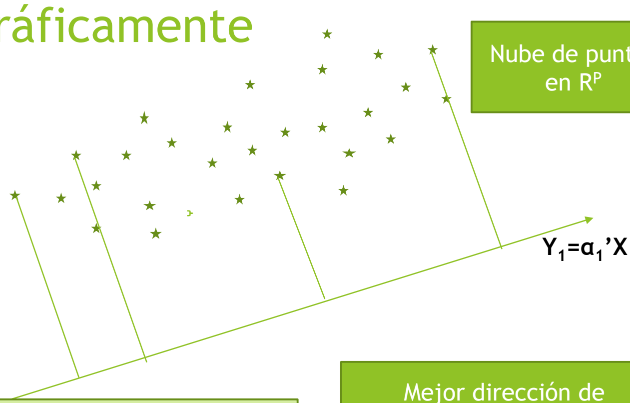
- Porque el autovalor coincide con la función a maximizar *Var(Y₁)*
- La ecuación que satisface α es: $(S - \lambda I)\alpha = \boldsymbol{0}$
- Si premultiplicamos ambos miembros por α', resulta: $\alpha'(S - \lambda I)\alpha = \alpha'\mathbf{0} = 0 \rightarrow (\alpha' S - \alpha'\lambda I)\alpha = \mathbf{0}$
- Por lo tanto, teniendo en cuenta la restricción: $\alpha' S\alpha - \lambda\, \alpha'\alpha = 0$ y $\alpha' S\alpha = Var(Y) = \lambda$
- Asi, para el primer eje de proyeccion $Y_1 = CP1$ usamos el autovalor mayor de S (λ_1) y su autovector asociado, normalizado (α_1)
- Asi nuestra primer $CP_1 = Y_1 = a_1'X$

Primera componente principal:

- El primer subespacio que encontramos, para resumir la información del conjunto original de puntos representados en el espacio *p*-dimensional, es un eje de proyección cuya dirección esta dada por el autovector normalizado (versor) asociado al mayor autovalor de la matriz de variancias y covariancias ***S*** (o ∑, matriz simétrica y definida positiva, por lo cual tiene *p* autovalores no nulos)
- Este 1er eje es el de la mayor variabilidad posible
- Esa variancia de los puntos proyectados coincide con el mayor autovalor de ***S*** → $Var(Y_1) = \lambda_1$

Gráficamente

Nube de puntos en R^p



$Var(Y_1) = \lambda_1$
Variabilidad de puntos proyectados

Mejor dirección de proyección….
Es la dirección de mayor variabilidad

“Espejo” de menor deformacion

Otras CP después de la primera….

- Se buscan mas ejes de proyección, con la condición de que sean ortogonales entre si , *(a_i' a_j)=0, $i \neq j$*
- La ortogonalidad geométrica, equivale a no correlación estadística (lineal), *Corr(Y_i, Y_j)= 0, $i \neq j$* (Ej.)
- Además cada nuevo eje deberá tener la máxima variabilidad posible, después de la ya evidenciada por el primero, primero y segundo, etc.
- Habría que plantear otra función "variancia" a maximizar, ahora sujeta a dos restricciones

$$\phi = \alpha' S\alpha - \lambda(\alpha'\alpha - 1) - \nu(\alpha'\alpha_1)$$

$$\frac{\partial \phi}{\partial \alpha} = 2S\alpha - \lambda 2\alpha - \nu\alpha_1 = \vec{0}$$

Trabajando la ecuación….

$$\frac{\partial \phi}{\partial \alpha} = 2S\alpha - \lambda 2\alpha - \nu\alpha_1 = \underset{\sim}{0}$$

- Pre-multiplicamos por α_1 ambos miembros:

$$\alpha_1'(\cancel{2}S\alpha - \lambda\cancel{2}\alpha - (1/2)\nu\alpha_1) = \alpha_1'\underset{\sim}{0}$$

$$\alpha_1' S\alpha - \lambda\alpha_1'\alpha - (\nu/2)\alpha_1'\alpha_1 = 0$$

$$Cov(Y_1, Y) - \lambda\cos(\alpha_1, \alpha) - (\nu/2)\|\alpha_1\| = 0 \Rightarrow \nu = 0$$

- Si $\nu = 0$, la ecuación a resolver, queda igual que en el primer paso y su solución será similar….utilizando el autovalor que sigue a λ_1, en orden descendiente (λ_2) y asi sucesivamente

Hasta cuantos ejes de proyección?

- Un tercer eje requeriria mas condiciones:
  - Maximizar variabilidad (despues de las halladas)
  - Ser ortogonal con las dos direcciones anteriores
  - Tener norma 1 para formar la proyeccion
- La función Φ con 3 multiplicadores de Lagrange, tiene 2 nulos (condiciones agregadas al 1er paso) y vuelve a quedar la misma ecuación.
- Se termina cuando no hay mas autovalores, y estos son p, dimensión de $\boldsymbol{S}$
- Todos son positivos (propiedades), pero puede haber iguales….que significa sobre la nube de puntos original?

En síntesis, los pasos a seguir, son:

$\underset{\sim}{X} \rightarrow \underset{\sim}{\Sigma} \rightarrow \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$

$\downarrow$

$\alpha_1 \quad \alpha_2 \quad \ldots \quad \alpha_p$

$\downarrow$

$CP_1 \quad CP_2 \quad \ldots \quad CP_p$

Y=C'X con C ortonormal define una rotacion rigida del espacio original / $\Sigma_Y = C'\Sigma C = \Lambda = \mathrm{diag}(\lambda_i)$

Expresión matricial de todas las CP

- Cada Componente Principal, se expresa:
- $Y_i = \boldsymbol{\alpha}_i{}'\boldsymbol{X}_j = \alpha_{i1}\ X_1 + \alpha_{i2}\ X_2 + \ldots + \alpha_{ip}\ X_p$
- Pueden representarse todas juntas en otro vector de dimensión *p* utilizando la matriz ***C*** de autovectores normalizados, o matriz que diagonaliza a ***S*** o **Σ**
- $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix}$ $\quad C'X = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots \boldsymbol{\alpha}_p)'\ X$
- Siendo ***C*** tal que $\boldsymbol{C}'\ \boldsymbol{\Sigma}\ \boldsymbol{C} = \mathrm{diag}(\lambda_i) = \boldsymbol{\Lambda}$
- Esta matriz Λ es tambien la matriz de variancias y covariancias del nuevo vector de variables ***Y***

Matriz de variancias y cov. de *Y*

- Sin perdida de generalidad, se puede suponer que el vector de variables originales ***X*** esta centrado (E(***X***)=***0***)
- En ese caso, el vector ***Y*** también esta centrado:
- $E(\boldsymbol{Y}) = E(\boldsymbol{C}'\boldsymbol{X}) = \boldsymbol{C}'\ E(\boldsymbol{X}) = \boldsymbol{C}'\ \boldsymbol{0} = \boldsymbol{0}$
- Entonces $\mathrm{Var}(\boldsymbol{Y}) = \boldsymbol{\Sigma}_Y = \boldsymbol{E(YY')} = \boldsymbol{E(C'X\ (C'X)')} =$

  $\boldsymbol{E(C'XX'C)} = \boldsymbol{C'\ E(XX')C} = \boldsymbol{C'\ \Sigma\ C} = \boldsymbol{\Lambda}$
- Por lo tanto los autovalores que estan en la diagonal principal de $\boldsymbol{\Lambda}$, son las variancias de cada CP (hecho que sabíamos por la deducción de las CP), y además, los elementos no diagonales o covariancias, son nulas, lo cual equivale a incorreción entre ellas (condición que impusimos en su deducción)

Necesidad de transformaciones

- Componentes principales usa la información de la matriz de datos, a través de las variancias y covariancias (***S*** o ***Σ***)
- Las variancias y covariancias se ven afectadas por diferencias en las escalas de las variables
- Si una variancia fuera dominante eso se traslada a los autovalores y en consecuencia a las CP
- Por lo tanto, cuando la información original de las variables, esta dada en mediciones no-conmensurables (distintas escalas, distintas unidades o variabilidades) conviene hacer un paso intermedio de estandarización de las variables $\Rightarrow$ obtener autovectores de la matriz de correlaciones

R o ***Σ*** ??

- Los resultados e interpretaciones de las CP calculadas sobre matriz de correlaciones o matriz de variancias, son diferentes
- No tienen relación funcional entre ellos, por lo cual no se puede pasar de una solución a la otra
- La elección por alguna de estas alternativas, debe basarse en consideraciones teóricas del problema en cuestión
- Lo estándar es aplicar CP sobre ***R*** para "homogeneizar" la influencia de todas las variables (variancias unitarias para todas), pero en ciertos campos los investigadores argumentan en contra de esta estrategia

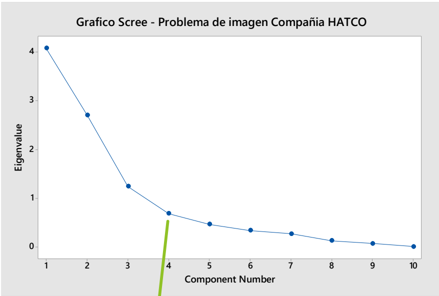
Como se analizan los resultados?

- Sabemos como se calculan las CP…pero como se interpretan? Podemos dar algunos pasos:
- 1) Se selecciona el numero de CP a interpretar,
- 2) Se interpretan las CP seleccionadas, analizando los coeficientes de la c.l.: ***a*** (o las correlaciones variable-componente: ***a****, $\rho=\alpha_{ij}\sqrt{\lambda_i}/\ \sigma_j$)
- 3) Se grafican variables e individuos en planos de pares de CP (plano de factores de variabilidad)
- 4) Puede enriquecerse el grafico con características cualitativas, pertenencia a grupos, proyectarse variables o individuos suplementarios, etc.
- Ampliemos cada aspecto…..

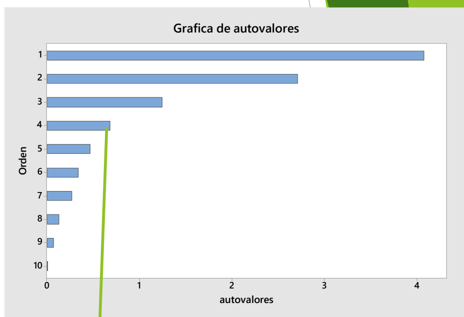
Número de CP a retener (Paso 1)

- Hay varios criterios y es conveniente utilizarlos en conjunto
- Proporción de variabilidad explicada = $\sum_{i=1}^{m} \lambda_i/\mathrm{tr}(\Sigma)$. Indica la calidad del resumen. Basada en la propiedad: *"la variabilidad total de los datos, dada por la tr(Σ), es reproducida por todas las CP (tr(Λ))"*. (Ejercicio)
- Magnitudes relativas de autovalores consecutivos: Test scree o gráfico de lineas con los autovalores ordenados o bien su "histograma". Se retienen hasta el λi a partir del cual la línea es horizontal
- Retener las CP que representen mayor variabilidad, que cada variable por separado $\mathrm{Var}(Y_j)= \lambda_j \geq \mathrm{Var}(X_i)$ (si las CP se obtienen de ***R***, retener CP_i si $\lambda i>1$)

Ejemplos de test scree e histograma de autovalores



Punto de quiebre a partir del cual la línea casi es horizontal



Autovalor a partir del cual las diferencias consecutivas son parecidas y pequeñas

Otros aspectos a tener en cuenta para decidir cuantas CP retener

- Cada CP retenida, debe tener una representación conceptual interpretable. Sino tiene sentido su interpretación, significa que solo representa "ruido" y se descarta
- Hay un par de test de hipótesis, uno basado en la teoría paramétrica y dependiente de la normalidad conjunta de las variables y el otro no paramet. Los dos prueban si la nube restante, luego de retener "*i*" componentes, es esférica Ho) $\lambda_{i+1}=\ldots=\lambda p$
- Paramétrico: Test de esfericidad
- No paramétrico: Test del bastón roto, compara con un fraccionamiento aleatorio de la $tr(\Lambda)$

Interpretación de las CP (Punto 2)

- Para interpretar cada "factor" de variabilidad que representa una CP que se decidió retener, deben interpretarse los coeficientes de la combinación lineal (componentes del autovector normalizado):
- $Y_i = \alpha_{i1} X_1 + \alpha_{i2} X_2 + \ldots + \alpha_{ip} X_p$
- Si las variables están estandarizadas, los coeficientes de cada variable, indicarán su aporte en la CP, según su signo y magnitud (coeficientes crudos)
- Si no lo están, es importante "limpiar su efecto y leer los coeficientes "estandarizados", porque la mayor o menor variabilidad afecta a su importancia relativa en la lectura de la combinación lineal
- En estos casos se usa la corr(X_i, $\boldsymbol{Y_j}$)

Correlación variable-componente

- La $CP_i = Y_i = a_1{}'X$ y la variable $x_j = (0\ldots1\ldots0)\boldsymbol{X}$
- La $Cov(x_j, Y_i) = Cov((0\ldots1\ldots0)\boldsymbol{X},\ a_i{}'X) = (0..1..0)\ \Sigma\ a_i$
- Y como a_i es autovector de Σ, $\Sigma\ a_i = \lambda_i\ a_i$
- Y la $Cov(x_j, Y_i) = (0..1..0)\ \Sigma\ a_i = \lambda_i\ (0..1..0)\ a_i = \lambda_i\ a_{ij}$
- De aquí la $Corr(x_j, Y_i) = \dfrac{Cov(x_j, Y_i)}{\sigma_j \sqrt{\lambda_i}} = \dfrac{\lambda_i\ a_{ij}}{\sigma_j \sqrt{\lambda_i}} = \dfrac{\sqrt{\lambda_i}\ a_{ij}}{\sigma_j}$
- Es decir, la correlación entre la variable y la componente es directamente proporcional al coeficiente y al autovalor (a su raíz cuadrada) e inversamente proporcional al desvio estándar de la variable
- Las correlaciones o "cargas" suelen representarse en planos de las CP (dos primeras, primera y tercera, etc.)

Gráfico de los individuos en planos de CP (scores)- (Punto 3)

- Las CP son combinaciones lineales de las variables medidas. La matriz de información contiene por líneas, las coordenadas de cada individuo (u objeto), por tanto es directo valorar a cada individuo en cada CP_j: $Y_{ij} = a`X_j$
- Se seleccionan para graficar solamente a las CP que se decidan retener y como las CP se han generado no correlacionadas, se pueden analizar por separado. Se opta por hacerlo en planos, de dimension 2
- Se eligen CP1 con CP2, CP1 con CP3, etc.
- El gráfico resultante es el RESUMEN buscado de la información multivariada y allí podrán identificarse los parecidos y diferencias entre individuos, recurriendo a las interpretaciones de cada componente

Información adicional de los planos de proyección

- El gráfico sobre las CP tomadas de a dos (planos de proyección), puede utilizarse para observar la relación de ciertas variables adicionales (muchas veces cualitativas), con los individuos. Puede ser la pertenencia a una población o grupo particular.
- Se etiquetan con un color, numero, letra o cualquier carácter en el plano y se observa su distribución y/o agrupamiento en la nube de puntos
- También pueden ubicarse "individuos promedio" calculados en subgrupos o sub-poblaciones, lo cual puede evidenciar diferencias entre ellos….
- Veremos otro tema mas especifico para diferenciar poblaciones (Análisis Discriminante)

Objetivos posibles en análisis de CP

- Descubrir subgrupos de individuos
- Detectar outliers, por inspección visual o aplicando test que utilizan los autovalores menores
- Construir índices o variables sintéticas, de resumen, como promedios ponderados
- Analizar el movimiento de un conjunto de variables a través del tiempo (versión filtrada de una serie multivariada)
- Detectar dimensiones redundantes (variables que muestran lo mismo)
- Plantear estrategias de control de procesos en función de objetivos de calidad múltiples
- Hacer análisis de multi-colinealidad en Regresión Múltiple, con la posibilidad de corregirla utilizando CP sobre variables independientes, asociadas auto-valores cercanos a cero

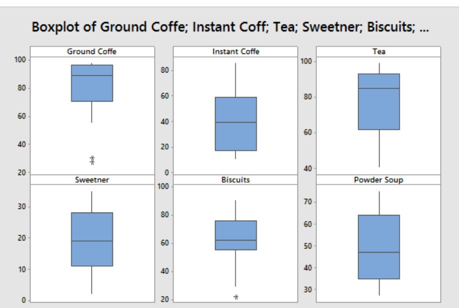
PROBLEMA 1: ¿Cómo varía el consumo de comidas entre países? (Data set:Food.xls)

- El propósito del estudio fue analizar los parecidos y diferencias entre países europeos, respecto del consumo de diferentes productos alimenticios
- El conjunto de datos a analizar, corresponden al consumo relativo de 20 tipos de comidas diferentes, en 16 países de Europa.
- Cada medición varía de 0 a 100, ya que representa un porcentaje de consumo.
- La variables son todas cuantitativas. Podríamos comenzar aplicando las estrategias gráficas y numéricas de resumen, en la búsqueda de esas características que hacen que los países se parezcan o no
- Los análisis multivariados permiten esta descripción considerando TODAS las variables a la vez.

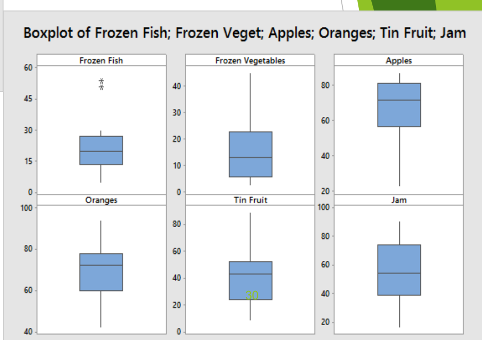
Algunas miradas por variable (análisis univariado)

- Si se pretende hacer un resumen de la información, tenemos opciones gráficas y numéricas.
- Las variables son cuantitativas, pero hay solo 16 (países) mediciones p/variable, con Box-Plot, se podrán ver las tendencias de consumo de cada tipo de alimento, mediante la ubicación de algunos percentiles
- Y si quisiéramos ubicar a los países? No se puede
- Y son muchos gráficos individuales!!
- También podríamos obtener una tabla de medidas descriptivas
- ……..todo será insuficiente para captar globalmente

Mirada univariada: Algunos Box-Plot



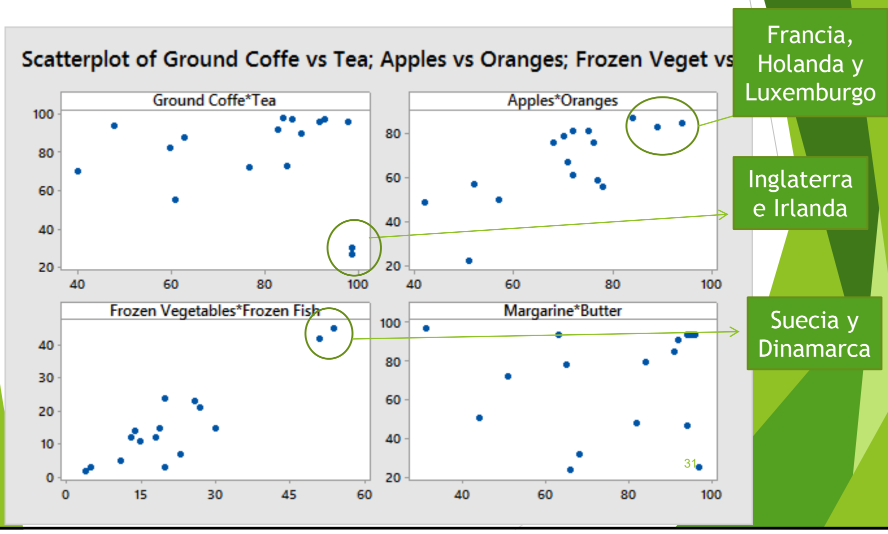
- ¿Qué alimentos se consumen mas frecuentemente en Europa?
- ¿Cuáles con menos frecuencia?
- Qué alimentos tienen distribuciones más concentradas (hay mas parecidos entre países en su consumo?
- O mas variables?



- Para qué alimentos hay países con consumo atípico al de Europa?
- ¿Cuáles son las medianas de consumo relativo?

Análisis bivariados

- Se podrían observar las asociaciones entre consumos de algunos alimentos e identificar los países (caracterizarlos) acercando el cursor……pera hay varios pares!!



Análisis multivariado: Componentes Principales

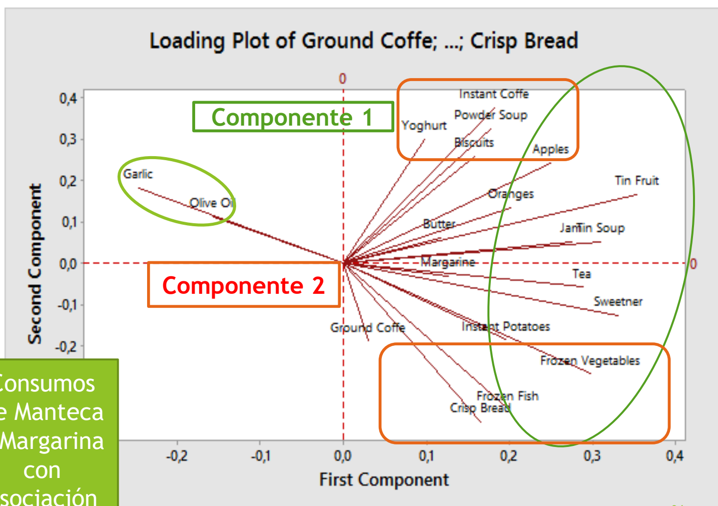
- El análisis de ***Componentes Principales***, es una de las técnicas multivariadas de resumen, más utilizadas para explorar la información conjunta escondida en una matriz de datos que represente los valores de variables cuantitativas medidas sobre un conjunto de unidades
- Permite la **caracterización** de los objetos, estudiando la estructura de asociaciones entre las variables.
- Crea un resumen condensando de la información, que puede ser graficado para captar con simplicidad que países tienen consumos parecidos y cuáles son muy diferentes, permitiendo obtener una descripción de esos países según sus consumos típicos
- Estas particularidades se descubren a través de dos gráficos, el de scores y el de cargas, donde el concepto intuitivo que nos permite analizarlos, es el concepto de distancias.
- La técnica trata tanto a los objetos como a las variables en grupos homogéneos, es decir no reconoce, si los hubiera, la pertenencia a distintos grupos o un rol diferente entre las variables

Gráficos de scores y de cargas

- El ***gráfico de cargas*** representa a las variables en un plano y muestra un resumen de las relaciones entre las variables. Es equivalente a mirar las correlaciones en una tabla (en análisis sobre R, coincide con mirar los pesos o coeficientes)
- Si dos variables están muy asociadas, forman un ángulo muy pequeño
- Si el ángulo entre dos variables es cercano a 90º, no hay asociación entre las variables
- Si el ángulo es de 180º, están inversamente correlacionadas
- El ***gráfico de scores*** representa a las observaciones (países) en un plano y muestra los parecidos y las diferencias entre ellos
- Lo que se observa en el gráfico de cargas, permite interpretar el gráfico de scores
- Los dos gráficos pueden mirarse superpuestos, la posición relativa de las observaciones y las variables, permiten deducir las características de las unidades en relación a las variables

Gráfico de cargas para los alimentos

¿Qué miramos?



Consumos de Manteca y Margarina con asociación + con CP2 y - con CP1

Puntos lejos del origen y dirección en la que se alejan, para interpretar las CP

Independencia Correlacion + o -

Otra visión del grafico de cargas:

Coeficientes (autovector αi)			
Variable	PC1	PC2	PC3
Café_granos	0,022	-0,193	-0,453
Café_instant	0,201	0,367	0,057
Te	0,281	-0,062	0,241
Dulces	0,251	-0,116	-0,174
Galletas	0,175	0,255	0,057
Sopa_polvo	0,192	0,317	0,016
Sopa_lata	0,321	0,035	0,178
Pure_instant	0,204	-0,210	-0,161
Pescado_cong	0,195	-0,375	-0,159
Verduras_congel	0,295	-0,289	-0,096
Manzanas	0,265	0,226	-0,243
Naranjas	0,220	0,113	-0,440
Fruta_lata	0,370	0,148	-0,032
Mermelada	0,284	0,039	0,298
Ajo	-0,242	0,187	-0,330
Manteca	0,115	0,067	0,109
Margarina	0,123	-0,031	-0,066
Aceite_Oliva	-0,148	0,109	-0,219
Yoghurt	0,106	0,284	-0,303
Pan_crujiente	0,157	-0,395	0,002

Correlaciones variable-CP ($\sqrt{\lambda i}\alpha i$)			
Variable	PC1	PC2	PC3
Café_granos	0,053	-0,388	-0,758
Café_instant	0,486	0,736	0,095
Te	0,680	-0,123	0,403
Dulces	0,606	-0,233	-0,290
Galletas	0,424	0,512	0,094
Sopa_polvo	0,464	0,635	0,025
Sopa_lata	0,775	0,070	0,297
Pure_instant	0,492	-0,421	-0,268
Pescado_cong	0,472	-0,752	-0,266
Verduras_congel	0,713	-0,579	-0,160
Manzanas	0,640	0,453	-0,407
Naranjas	0,531	0,226	-0,737
Fruta_lata	0,894	0,296	-0,054
Mermelada	0,686	0,078	0,498
Ajo	-0,581	0,374	-0,551
Manteca	0,278	0,133	0,182
Margarina	0,297	-0,061	-0,109
Aceite_Oliva	-0,355	0,219	-0,365
Yoghurt	0,256	0,570	-0,506
Pan_crujiente	0,378	-0,792	0,003

Las diferencias entre correlaciones son mas marcadas que entre los coeficientes del autovector normalizado Y Se puede agregar un criterio de selección: retener correlaciones superiores de…0,50?

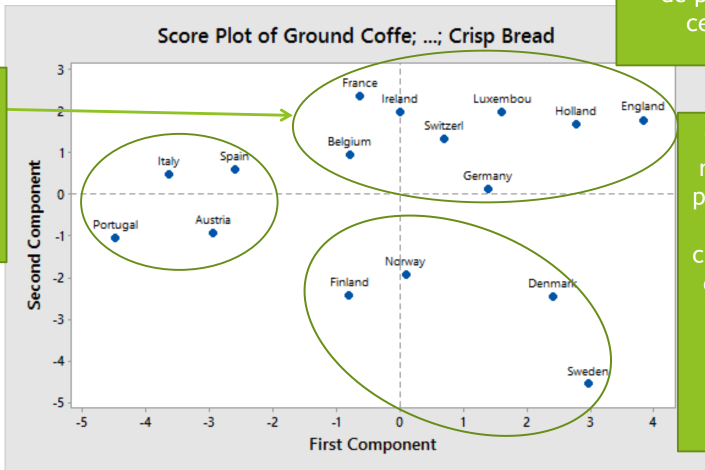
Interpretación de las Componentes

- ***El gráfico de cargas (o la Tabla) representa un resumen de las relaciones entre las variables*** y esta interpretación será un medio para interpretar el gráfico de scores, porque muestra el significado de cada Componente
- **Componente 1:** Esta CP1, muestra correlación baja con sólo una variable: el consumo de café en grano. Tiene correlación negativa con consumo de ajo y aceite de oliva y correlación positiva con el resto de las variables que representan el consumo, especialmente sopas, frutas, te, dulces, vegetales…
- Esta es una Componente global de consumo, sin incluir el café en grano (variable que está muy poco asociada)
- Un país que tuviera en el gráfico de scores un valor alto de la CP1, puede ser descripto como un país con altos consumos en todas las variables, excepto en café en grano, y que tiene poco consumo de ajo y aceite de oliva.
- **Componente 2:** Esta negativamente asociada a café en granos, mermeladas, vegetales y pescado congelados y pan crujiente. Y positivamente con café instantáneo, sopa en polvo, galletitas y mermeladas.
- Un país con alta CP2 tiene bajo consumo en café=granos, mermeladas, congelados…¿?… y alto en café instantáneo, galletitas……¿?

Gráfico de scores

Se pueden reconocer grupos de países por cercanía

Un grupo o dos?



Su ubicación relativa en el plano nos dará sus características de consumo.

Puntos cercanos al origen: promedios

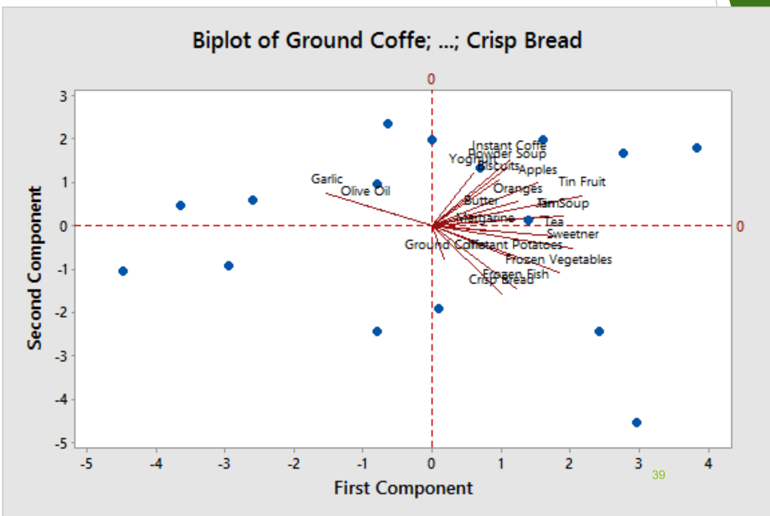
Caracterizamos el consume de los paises segun su posicion sobre las CP

Interpretación del gráfico de scores

- ***Presenta un resumen de las relaciones entre las observaciones*** (países Europeos), mostradas en un mapa, donde países cercanos tiene similares hábitos de consumo de comidas
- Una dirección en el gráfico de scores se corresponde con la misma dirección en el gráfico de cargas
- Los países nórdicos (Noruega, Suecia , Finlandia y Dinamarca) comparten hábitos de consumo. ¿Cuáles? Según el de cargas, tienen altos consumos de productos como: pan crujiente, vegetales y peces congelados, puré en polvo y café en granos. Estos países tiene también poco consumo de ajo y aceite de oliva, relativo a otros.
- Otro agrupamiento que resalta es el de los países del sur, Portugal, España e Italia, al que se le agrega Austria. En estos países los habitantes prefieren el ajo y aceite de oliva, en cambio comen pequeñas cantidades de frutas y sopas en lata y edulcorante.
- El grupo conformado por Luxemburgo, Inglaterra, Suiza, Holanda , Francia, Irlanda y Bélgica, se caracterizan especialmente por comer yogurt, manzanas, galletas, café instantáneo, sopa en polvo (CP2 alta)
- Alemania queda en una posición intermedia con valor nulo en la CP2 y bajo en CP1, lo cual indica consumos promedios

Biplot

- Otra opción para ver los dos gráficos simultáneamente, es superponerlos. Si son muchas variables y/o muchas observaciones, podría no verse con claridad



¿Dos Componentes son suficientes?

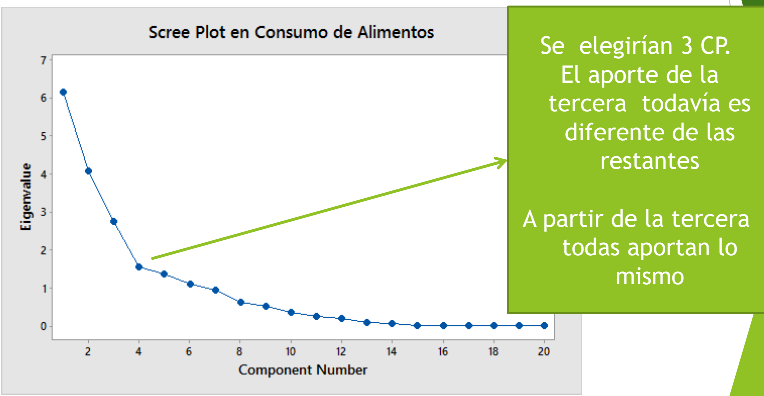
- Un criterio para elegir con cuántas componentes describir la información de la matriz de datos, es comprobar cuanta variabilidad de los datos explican. Es usual fijar un porcentaje de variabilidad que se pretenda explicar.
- El conjunto total de las Componentes reproduce toda la variabilidad
- ¿Cuanto reproducen dos CP? Hay que revisar la tabla con las variancias explicadas y los aportes porcentuales de 1, 2, 3, 4, etc. componentes

Dos Componentes explican algo más del 60% de la variabilidad total

Orden	1	2	3	4	5	6...					
Autovalor	5,853	4,029	2,800	1,529	1,362	1,185	0,935	0,675	0,568	0,411	0,240
Propor.	0,293	0,201	0,140	0,076	0,068	0,059	0,047	0,034	0,028	0,021	0,012
Prop.cum	0,293	0,494	0,634	0,711	0,779	0,838	0,885	0,918	0,947	0,967	0,979

Otra visión: el Gráfico Scree

- Se grafican las variancias de cada componente contra su orden de importancia, marcando una línea quebrada.



Se podría agregar el análisis de la Componente 3

PROBLEMA 2: Un caso en Medicina Deportiva (Data set: nadadores.xls)

- Hace unos años, se diseño un Plan de Trabajo con el fin de mejorar la flexibilidad y elongación artro-muscular de nadadores de competición en un Club de Rosario (Centro de Investigaciones Medico-Deportivas)
- El plan esta basado en la hipótesis de que la flexibilidad y elongación artro-muscular son aptitudes físicas capaces de ser perfeccionadas mediante la practica de ejercicios de estiramiento

Objetivos planteados

- Evaluar si el Plan de Trabajo propuesto por el CIMED causo una mejoría significativa en la flexibilidad de los 20 nadadores de competición del Club
- Generar un indicador global de flexibilidad univariado a través de todas las mediciones de flexibilidad utilizadas (si es razonable)

Variables medidas antes y después del programa de entrenamiento

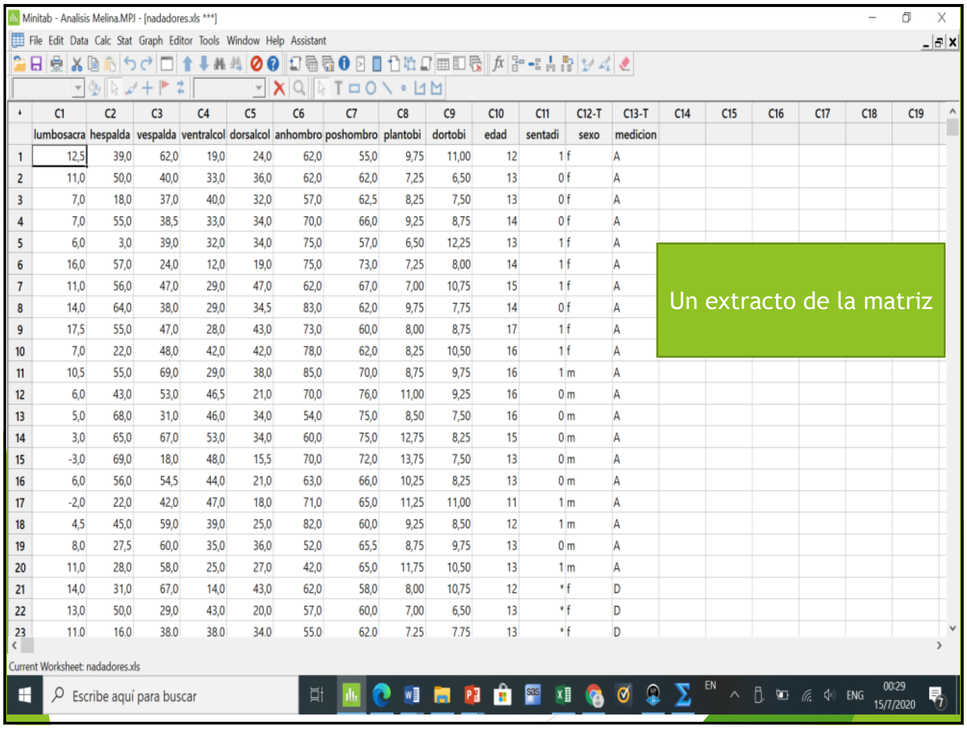
- Flexibilidad de la articulación lumbo-sacra (con elongación máxima de músculos isquio-tibiales)(+)
- Flexibilidad horizontal de la espalda (-)
- Flexibilidad vertical de la espalda (+)
- Flexibilidad ventral de la columna dorso-lumbar (-)
- Flexibilidad dorsal de la columna dorso-lumbar (+)
- Flexibilidad anterior de la articulación del hombro (+)
- Flexibilidad posterior de la articulación del hombro(+)
- Flexibilidad del bloque articular del tobillo y pie en flexión plantar (-)
- Flexibilidad del bloque articular del tobillo y pie en flexión dorsal (+)
- Edad en años (en primera medición)
- Prueba de sentadilla (pasa o no pasa)
- Sexo

Cuestiones practicas a resolver

- Las unidades de medida son conmensurables? Usamos $\sum$ o R?
- Hacemos dos análisis por separado (mediciones antes y después) o consideramos a los individuos "duplicados"….. o a las variables?
- No tenemos que perder de vista que el objetivo es evaluar la eficacia del Plan de Trabajo
- Podremos considerar conjuntamente con las mediciones cuantitativas el sexo y la prueba de sentadilla?

La matriz de datos

- La matriz de datos contiene las 9 mediciones de flexibilidad (cuantitativas) mas el éxito-fracaso de la prueba de sentadilla, la edad y el sexo, para cada uno de los 20 nadadores. Todas las mediciones de flexibilidad se tomaron al principio del programa y al final.
- Al estar repetidas las medidas….como se arma la matriz de datos?
- Alternativas (matrices yuxtapuestas)
  - Grabar 20 filas (n) y 22 columnas (p=10+10+2)
  - Grabar 40 filas, repitiendo individuos según medición antes/despues y 12 columnas
- Se opto por repetir individuos….la segunda opcion



Minitab - Analisis Melina.MPJ - [nadadores.xls ***]

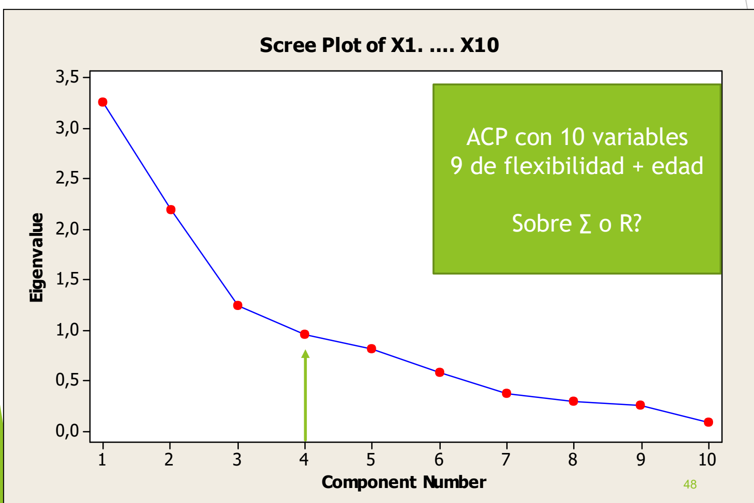
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12-T	C13-T
	lumbosacra	hespalda	vespalda	ventralcol	dorsalcol	anhombro	poshombro	plantobi	dortobi	edad	sentadi	sexo	medicion
1	12,5	39,0	62,0	19,0	24,0	62,0	55,0	9,75	11,00	12	1	f	A
2	11,0	50,0	40,0	33,0	36,0	62,0	62,0	7,25	6,50	13	0	f	A
3	7,0	18,0	37,0	40,0	32,0	57,0	62,5	8,25	7,50	13	0	f	A
4	7,0	55,0	38,5	33,0	34,0	70,0	66,0	9,25	8,75	14	0	f	A
5	6,0	3,0	39,0	32,0	34,0	75,0	57,0	6,50	12,25	13	1	f	A
6	16,0	57,0	24,0	12,0	19,0	75,0	73,0	7,25	8,00	14	1	f	A
7	11,0	56,0	47,0	29,0	47,0	62,0	67,0	7,00	10,75	15	1	f	A
8	14,0	64,0	38,0	29,0	34,5	83,0	62,0	9,75	7,75	14	0	f	A
9	17,5	55,0	47,0	28,0	43,0	73,0	60,0	8,00	8,75	17	1	f	A
10	7,0	22,0	48,0	42,0	42,0	78,0	62,0	8,25	10,50	16	1	f	A
11	10,5	55,0	69,0	29,0	38,0	85,0	70,0	8,75	9,75	16	1	m	A
12	6,0	43,0	53,0	46,5	21,0	70,0	76,0	11,00	9,25	16	0	m	A
13	5,0	68,0	31,0	46,0	34,0	54,0	75,0	8,50	7,50	16	0	m	A
14	3,0	65,0	67,0	53,0	34,0	60,0	75,0	12,75	8,25	15	0	m	A
15	-3,0	69,0	18,0	48,0	15,5	70,0	72,0	13,75	7,50	13	0	m	A
16	6,0	56,0	54,5	44,0	21,0	63,0	66,0	10,25	8,25	13	0	m	A
17	-2,0	22,0	42,0	47,0	18,0	71,0	65,0	11,25	11,00	11	1	m	A
18	4,5	45,0	59,0	39,0	25,0	82,0	60,0	9,25	8,50	12	1	m	A
19	8,0	27,5	60,0	35,0	36,0	52,0	65,5	8,75	9,75	13	0	m	A
20	11,0	28,0	58,0	25,0	27,0	42,0	65,0	11,75	10,50	13	1	m	A
21	14,0	31,0	67,0	14,0	43,0	62,0	58,0	8,00	10,75	12	*	f	D
22	13,0	50,0	29,0	43,0	20,0	57,0	60,0	7,00	6,50	13	*	f	D
23	11,0	16,0	38,0	38,0	34,0	55,0	62,0	7,25	7,75	13	*	f	D

Current Worksheet: nadadores.xls

Un extracto de la matriz

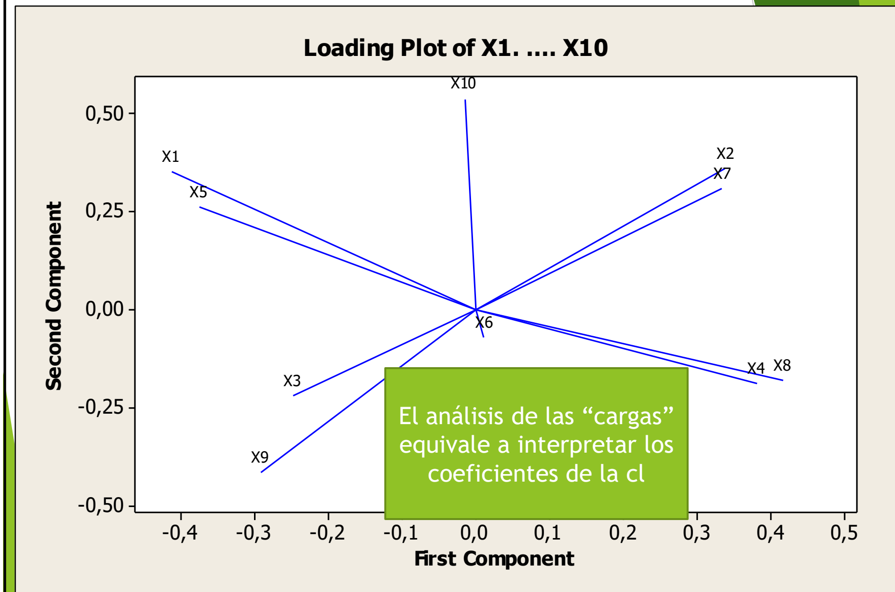
Eigenanalysis of the Correlation Matrix

Eigenvalue	3,2482	2,192	1,237	0,956	0,806	0,570	0,363	0,2905
Proportion	0,325	0,219	0,124	0,096	0,081	0,057	0,036	0,029
Cumulative	0,325	0,544	0,668	0,763	0,844	0,901	0,938	0,967

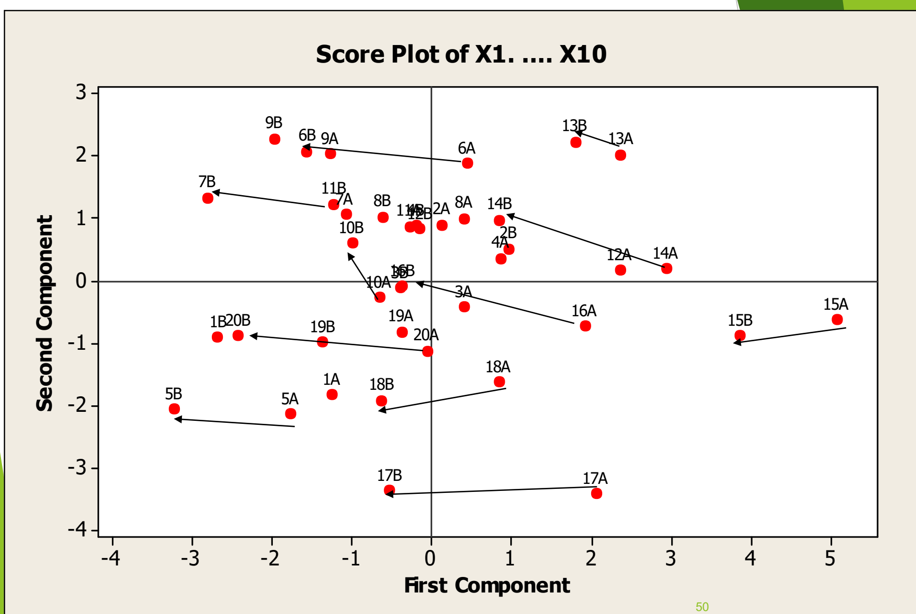


ACP con 10 variables
9 de flexibilidad + edad

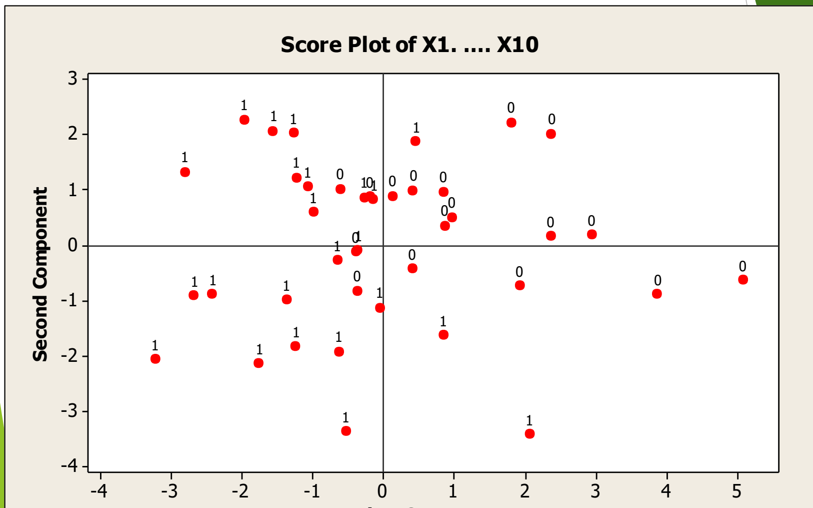
Sobre ∑ o R?



El análisis de las "cargas" equivale a interpretar los coeficientes de la cl

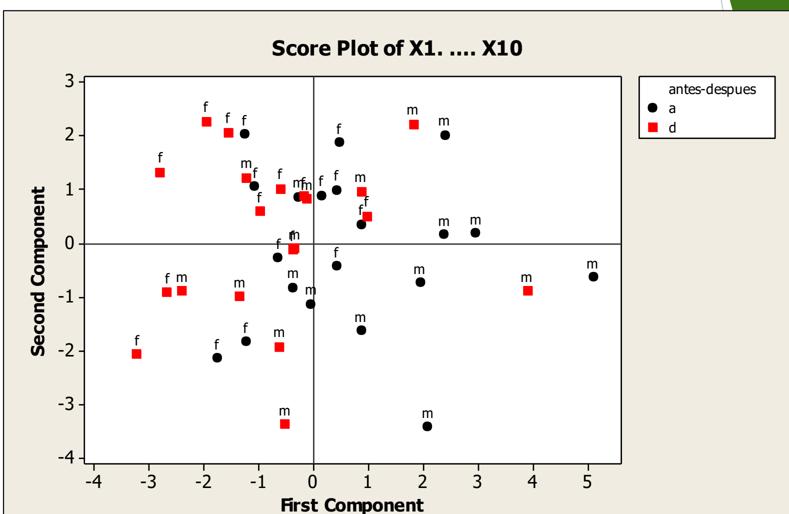


Prueba de sentadilla



51

Sexo



52

Resultados generales

- Se identifican dos indicadores univariados, no correlacionados, que indican aspectos particulares de la flexibilidad de los nadadores.
- La CP1 es la que mas diferencia en flexibilidad de izquierda a derecha, teniendo en cuenta los aspectos de….
- El indicador CP2 esta asociado con la edad (en el rango de los nadadores de competición del club) …
- Las mujeres son mas flexibles que los hombres
- El programa de entrenamiento produce una diferencia positiva en la flexibilidad

Individuos suplementarios

- Una vez analizadas las CP puede surgir interés en analizar un nuevo individuo (un nadador que se incorpora al staff)
- Puede usarse el resultado de las CP para proyectarlo en el gráfico y verlo comparativamente insertado en el total estudiado
- Simplemente se calcula la cl: $Y_i = \alpha` X_{nuev}$
- Tener en cuenta que los valores de X deben corresponder a la métrica usada para calcular las CP (datos crudos o estandarizados)

Variables suplementarias

- También es posible incorporar el aporte de OTRA variable al análisis, a posteriori de haber calculado las CP
- Se analiza a través de su correlación con las CP halladas….pero no hay un α_{ij} para esa variable. Como se logra?
- Estimando la correlación con los valores sobre cada individuo de la CP_j y la variable X_{nueva} :

$$Corr(X_{nueva}, Y_j) = Cov(X_{nueva}, Y_j)/\sigma_{Xnueva}\sigma_{Yj} =$$

$$(\Sigma\, x_{s,nueva}\, Y_{s,j}/n)\ sqrt(\lambda_j)$$

Recomendaciones

- No hay relación entre los análisis sobre R o Σ. La decisión depende de la investigación en particular
- Las correlaciones (o covariancias) miden asociaciones lineales, y las CP sólo explicarán estas asociaciones. Si se sabe de antemano que existen de otro tipo, conviene aplicar transformaciones para linearizar
- ***R*** y ***Σ*** son MUY afectadas por outliers, si ellos existen y se desconoce este hecho, los resultados estarán sesgados en su dirección
- Si hay muchas variables que miden el mismo concepto, las primeras CP identificarán ese aspecto. NO agregar variables que midan lo mismo

Extensiones del análisis de CP

- La técnica es la más antigua de las técnicas de análisis de datos multivariados
- Es muy flexible y robusta a distintas situaciones
- Una extensión muy útil fue aportada por la escuela francesa, aplicándola sobre datos categóricos: Análisis de Correspondencias Binario y Múltiple
- Transformando datos a matrices indicadoras, el ACP conduce a la representación de datos de frecuencias, resumen usual para variables categóricas

Datos binarios

- El ACP sobre información binaria logra excelentes resultados de clasificación
- A diferencia de los métodos de clasificación pura, permite evaluar la importancia relativa de cada variable en la clasificación
- *A multivariate approach to the proteomics of tomato fruit ripening*. Pratta, G. Quaglino, M. et all. (2010). Genes, Genomes and Genomics, 4, 48-51

Datos de Conjuntos Múltiples

- Las tablas multivariadas pueden incluir:
- Varios conjuntos de individuos, un solo conjunto de variables y varias condiciones: tablas yuxtapuestas por condiciones "a lo largo" (alumnos x variables académicas en distintas escuelas- rendimiento agronómico x variedades en distintos campos, flexibilidad x nadador en distintos tiempos)
- Varios conjuntos de variables, un solo conjunto de individuos y varias condiciones: tablas yuxtapuestas por variables "a lo ancho" (variables moleculares y agronómicas en poblaciones de maíz de varios campos, variables de rendimiento y fliares., sobre alumnos de varias facultades)

Efecto de yuxtaponer tablas

- Se pierde el análisis de algunas inter-relaciones
- Las coordenadas que se generan, representan una variabilidad general, que involucra tanto las diferencias entre grupos (o condiciones), como la variabilidad dentro de ellas
- A veces la estrategia de yuxtaponer (individuos) se reemplaza por considerar matrices promedio
- Hay técnicas especificas que consideran estas tres dimensiones. Veremos algunas de ellas como ultimo tema del curso

Datos de tres modos

- Se presentan cuando las variables, individuos y condiciones están totalmente cruzadas, formando "un cubo" de información multivariada:
- Un solo conjunto de individuos sobre los que se mide un mismo conjunto de variables en una variedad de situaciones
- Rendimiento x especies x ambientes
- Permanencia en sangre de varias dosis de un fármaco (mediciones repetidas) en mujeres y hombres de distintos grupos etarios
- Atributos de Comerciales evaluados por muestras de potenciales clientes en distintas ciudades

Análisis Bi-Plots

- Los biplots de Gabriel representan una descomposición de la matriz de datos similar a la de las CP, pero usando una métrica distinta (no usa los versores en la proyección)
- Logra representar individuos y variables en un mismo plano, de modo de ver similitudes entre individuos, entre variables y una visión de los individuos parcializada por variables de interés

Descomposición de matrices usadas en Bi-Plots

- Se denomina descomposición singular y surge de la relación entre vectores de matrices producto XX´ y X´X, ambas cuadradas simétricas
- Si u_f es autovector de XX´ asociado a λ_f ➔

  $XX'u_f = \lambda_f u_f$ (por definición)
- Pre-multiplicando por X´: $X'X(X'u_f) = \lambda_f (X'u_f)$
- De donde $X'u_f$ es autovector de X´X asociado al mismo autovalor
- Para hacerlo de norma 1, se lo divide por su norma= sqrt[$(X'u_f)'$ $(X'u_f)$] = sqrt (λ_f)

Reproducción de X a través de valores y vectores propios de XX´

- u_f es autovector de XX´ asociado a λ_f
- $W_f = X'u_f$ / sqrt(λ_f) es autovector de X´X
- Ambos son de norma 1
- $W_f u_f' = X'u_f u_f'$ / sqrt(λ_f) o bien:
- sqrt(λ_f) $W_f u_f' = X'u_f u_f'$
- Sumando a través de f (# autovalores no nulos): Σ sqrt(λ_f) $W_f u_f' = X'$
- Y cada sumando está ponderado por elementos de muy distinta magnitud y X podría aproximarse con sólo los dos primeros….