Análisis de Componentes Principales (ACP)

Hamdi Raïssi

IES PUCV

hamdi.raissi@pucv.cl

Objetivos de esta parte

Objetivo 1 : "Ciencia de datos"

a- Resumir y representar información en un conjunto de datos con el mínimo de perdida posible.

b- Utilizar el ACP como herramienta para el *clustering* de datos.

Objetivo 2 : Métodologico : ACP como etapa previa a la regresión

a- Selección de regresores pertinentes.

b- Evitar problemas de colinealidad.

Objetivo 3 : Hacer clustering con un ACP

Origenes del ACP

- Pearson (1901) se interesó en approximar un conjunto de puntos en un espacio de dimensiones reducidas.
- No tuvo desarrollo grande durante mucho tiempo dado las capacidades de cálculo reducidas.
- Hay dos escuelas : Norte americana con supuestos Gaussianos y francesa con una perspectiva geometrica.

Bibliografía : Análisis de datos con R. F. Husson, S. Lê y J. Pagès.

El cuadro de los datos y anotaciones

Queremos estudiar un conjunto de datos con la estructura siguiente :

$\times$	Z_1	$\dots$	Z_k	$\dots$	Z_K
1	z_{11}	$\dots$	z_{1k}	$\dots$	z_{1K}
$\vdots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
i	z_{i1}	$\dots$	z_{ik}	$\dots$	z_{iK}
$\vdots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
I	z_{I1}	$\dots$	z_{Ik}	$\dots$	z_{IK}

- Consideramos K variables cuantitativas (=uso de correlaciones en ACP.... ! !).
- I individuos.

El cuadro de los datos y anotaciones

- Un individuo i esta descrito por los valores que toma al respeto de las K variables :

$$z_i = (z_{i1}, \ldots, z_{ik}, \ldots, z_{iK}),$$

  es un vector linea.
- Las variables Z_k son dadas por las columnas del conjunto de datos.

El cuadro de los datos y anotaciones

Los datos pueden ser vistos como :

- una nube de puntos en el espacio $\mathbb{R}^K$ al respeto de los individuos (vamos a considerar distancias entre individuos).
- Un conjunto de vectores en el espacio $\mathbb{R}^I$, al respeto de los variables.

El cuadro de los datos y anotaciones

Ejemplos.

- Bancos : variables=ahorros, creditos, sueldo, edad...etc. Individuos=clientes
- Salud : variables=tasa de grasa en el sangre, concentración de una proteina, ...etc. Individuos=pacientes.
- Internet : variables= Montos de compras en una página web, tiempo de navegación en la página, ...etc. Individuos=clientes en la web
- Agronomía : variables=grasa en la carne, acidez, cantidad de proteinas,...etc. Individuos=vacunos
- Y muchos más ejemplos.....

A veces las variables que describen los individuos pueden ser muchas (pensar en un cuestionario de una encuesta...)

El cuadro de los datos y anotaciones

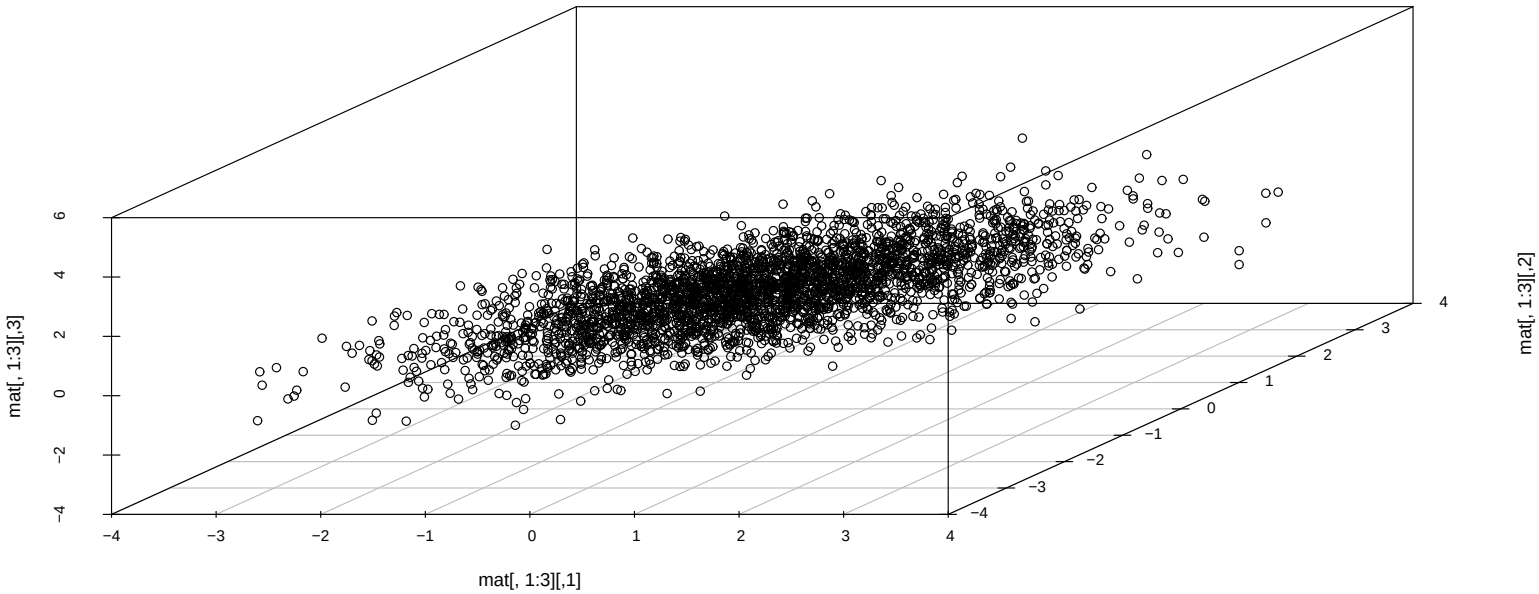


Figure – A muchas veces, es dificil ver correlciones entre variables o grupos/tendencias en los individuos en 3 dimensiones. En 4 dimensiones es imposible.

ACP : representar nubes de puntos en 2 dimensiones en general.

El cuadro de los datos y anotaciones

- Necesitamos resumenes para poder constestar a preguntas claves sobre los datos.
- Entregar una visualización fiel de los datos.

Tarea del ACP :

- Representar la nube de puntos original en un espacio de dimensión reducido con el mínimo de distorciones (en el sentido de distancia entre puntos).
- Representar las variables originales de acuerdo a la reducción de la dimensión.

El cuadro de los datos y anotaciones

- Hay dos maneras de resumir los datos : Por las columnas y por las filas.
- Por los individuos : Podemos identificar grupos
- Por las variables : Podemos ver las variables las más pertinentes, estudiar relaciones entre variables.

Es importante desarrollar esos dos aspectos de manera relacionada (no de manera independiente) :

- Permite de identificar caracteristicas sobresalientes de los individuos en según las diferentes variables.

Transformaciones previas : Anotaciones

- El promedio de la variable k :

$$\bar{z}_k = \frac{1}{I}\sum_{i=1}^{I} z_{ik}.$$

- La desviación estandar :

$$\mathfrak{s}_k = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(z_{ik} - \bar{z}_k)^2}$$

Transformaciones previas al aplicar un ACP

- Queremos identificar informaciones redundantes entre variables.
- Se considera correlaciones.
- Centrar y normalizar los datos para poder manejar correlaciones entre variables :

$$corr(z_k, z_l) = \frac{\frac{1}{I}\sum_{j=1}^{I}(z_{jk} - \bar{z}_k)(z_{jl} - \bar{z}_l)}{\mathfrak{s}_k \mathfrak{s}_l}$$

- Siempre se hace en ACP.

Transformaciones previas al aplicar un ACP

Ventajas de normalizar los datos (dividir por la estandar desviación) :

- Permite de evitar efectos de unidades (gramos, kilogramos...)
- Permite de dar la misma importancia a todas las variables.
- Cuando no se hace la normalización hablamos de "ACP no estandardizada".

En la gran mayoria de los casos se hace normalización $\Rightarrow$ Nos vamos a considerar solo esta situación.

Transformaciones previas al aplicar un ACP

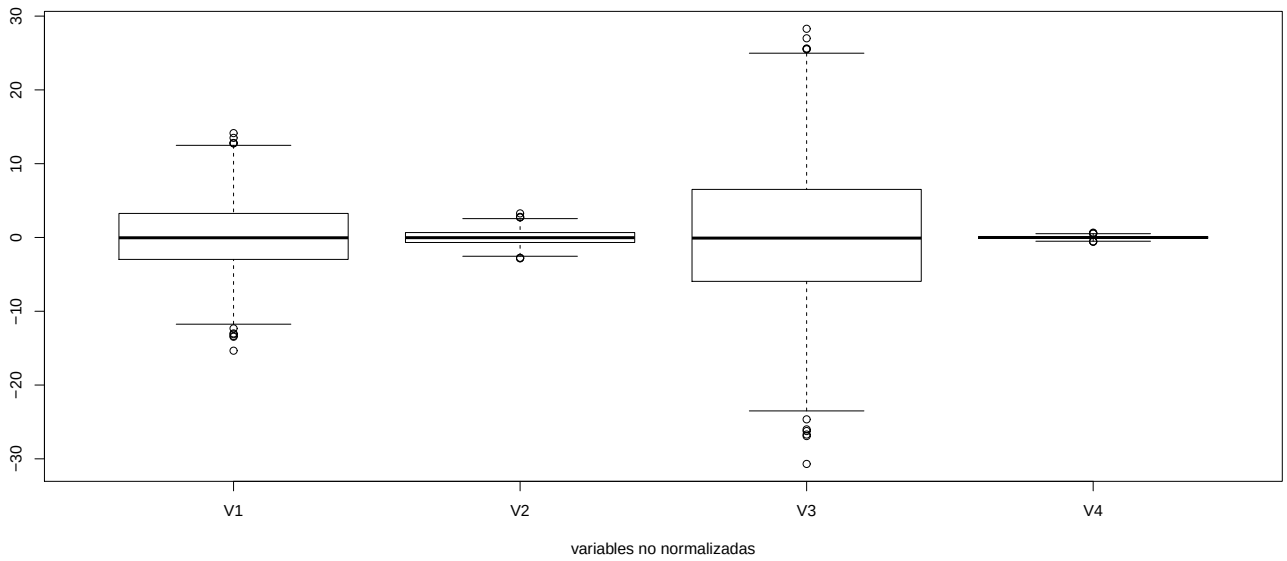


Figure – En este ejemplo, si los datos no son estandardizados, entonces el ACP se resume a eliminar V2 y V4 !

- La normalización evita que las variables con varianzas altas "aplastan" a las variables con varianzas baja.
- Sin embargo V2 y V4 pueden tener una información pertinente ...

El cuadro de los datos y anotaciones

En adelante consideramos sólo las variables normalizadas :

$\times$	X_1	$\ldots$	X_k	$\ldots$	X_K
1	x_{11}	$\ldots$	x_{1k}	$\ldots$	x_{1K}
$\vdots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
i	x_{i1}	$\ldots$	x_{ik}	$\ldots$	x_{iK}
$\vdots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
I	x_{I1}	$\ldots$	x_{Ik}	$\ldots$	x_{IK}

Proyección de los puntos

- La distancia entre dos individuos i y j en la nube normalizada :

$$d_{\mathbf{X}}(i,j) = \sqrt{\sum_{k=1}^{K}(x_{ik} - x_{jk})^2}$$

- Queremos hacer la proyección de la nube de puntos de los individuos del espacio R^K un espacio de dimensión R^s, con $s \ll K$ ($s = 1, 2$ a muchas veces).
- La matématica nos dice que la proyección ortogonal minimiza las distorciones de las distancias entre los individuos.

Proyección ortogonal.

- Dado que las distancias son reducidas aplicando la proyección ortogonal (teorema de Pitágoras), entonces debemos elegir el espacio que maximiza la varianza de la representación de la nube de puntos.

Proyección ortogonal : ilustración con la reducción de 2 a 1 dimensión

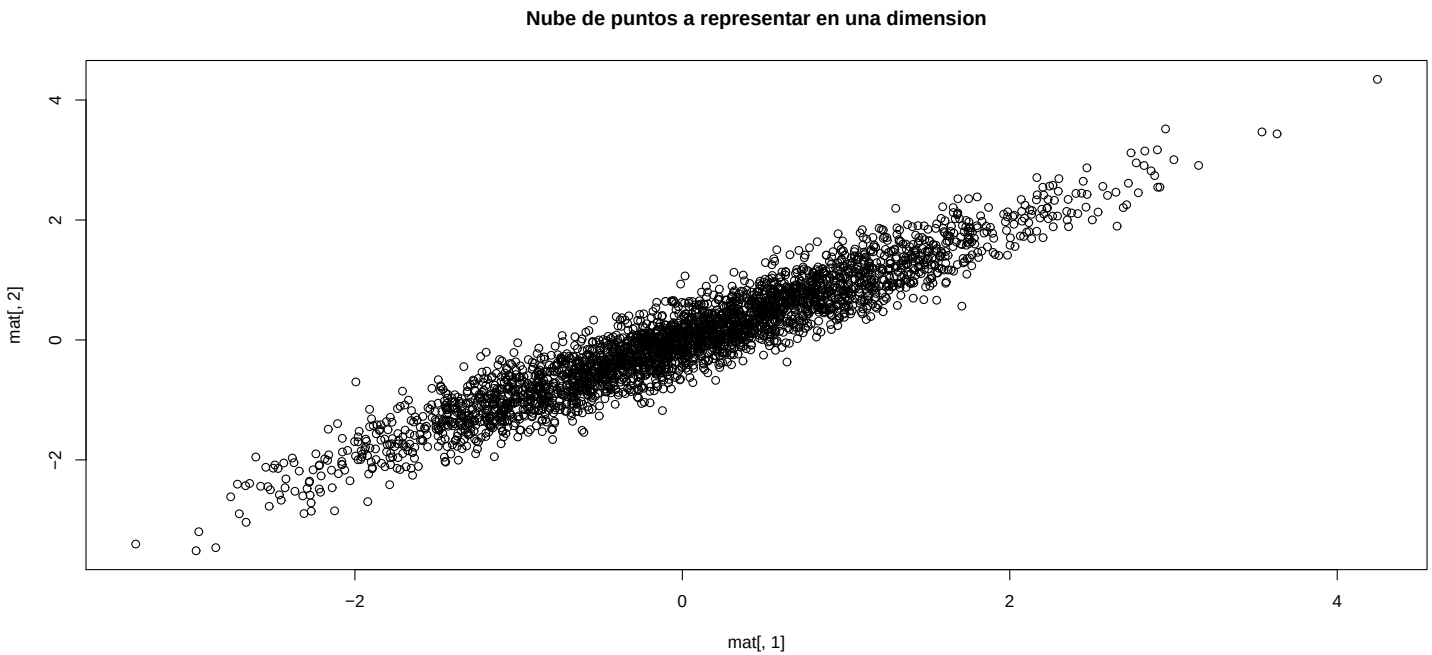
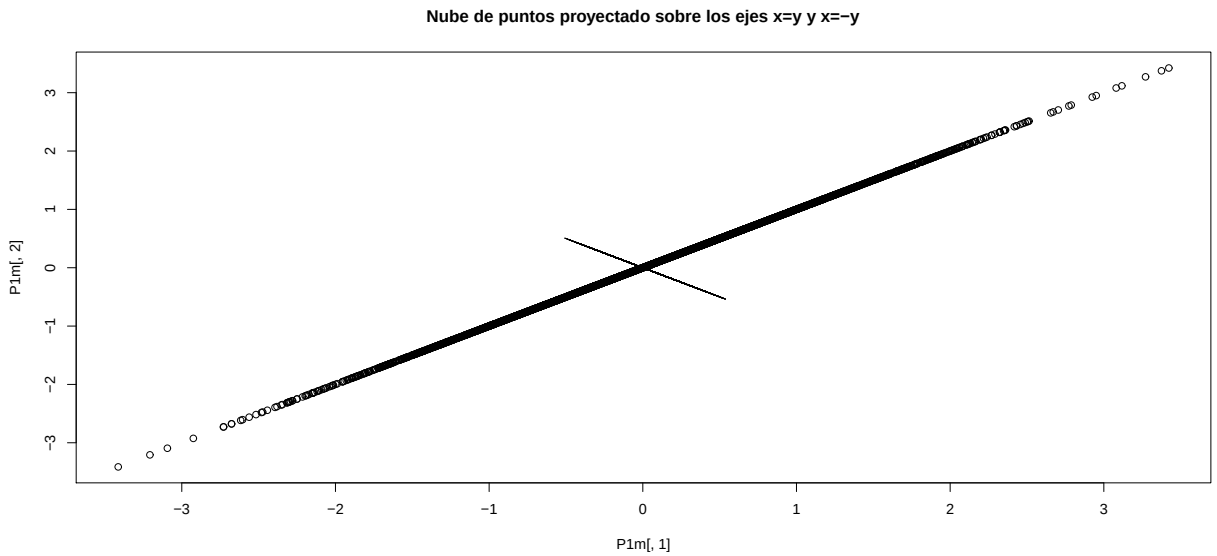


Figure – Nube de puntos a dos dimensiones que queremos representar en una.

Proyección ortogonal : ilustración con la reducción de 2 a 1 dimensión



- La proyección sobre el eje $x = y$ nos permite obtener el maximo de variabilidad.
- La proyección sobre el eje $x = -y$ conlleva a la peor perdida de variabilidad.
- Importante bien elegir los ejes para minimizar la perdida de información !

Eligir los ejes de proyección de los datos.

- Sea X de dimensiones $I \times K$ que contiene los datos (columnas=variables normalizadas, filas=individuos).
- La varianzas y correlaciones empiricas se escriben $\frac{1}{I}X'X$.
- Suponemos $\frac{1}{I}X'X$ definida positiva.
- Entonces la matriz de varianza-covarianza empirica es diagonalisable.

$$\frac{1}{I}X'X = PDP',$$

- P contiene los vectores propios.
- $D = diag(\lambda_1, \lambda_2, \ldots, \lambda_K)$ es diagonal y contiene los valores propios $\lambda_1 > \lambda_2 > \cdots > \lambda_K$.

Eligir los ejes de proyección de los datos.

Propriedad

- El sub-espacio de dimensión s que asegura una varianza maxima de los puntos proyectados de la nube de dimensión K corresponde al espacio generado por las s primeras columnas de P (vectores propios de la matriz de varianza-covarianza).
- La inercia del eje j, $0 \leq j \leq K$ (o sea la varianza explicada por el eje j), es dada por λ_j.

Eligir los ejes de proyección de los datos.

Idea intuitiva de la propriedad anterior :

- Construimos el espacio de dimensión s que maximiza la varianza de la nube proyectada eje por eje.
- Una vez que el primer eje es identificado, es lógico que este eje será encontrado en el resultado final : o sea a dentro del espacio que maximiza la varianza.
- El segundo eje que maximiza la varianza debe ser ortogonal al primer eje (sino es ortogonal "explicamos de nuevo", y entonces de manera inutil, parte de la variabilidad).
- Nota : P es una matriz ortogonal, o sea sus columnas son ortogonales. (entonces cumple con el punto anterior)

Eligir los ejes de proyección de los datos.

Idea intuitiva de la propriedad anterior :

- La solución encontrada (vectores propios, valores propios) es el resultado de la maximización de

$$\frac{z'X'Xz}{z'z}, \quad \text{o de manera equivalente} Var(Xa), \text{ con } \|a\| = 1$$

  al respeto de la variable $z \in \mathbb{R}^K$ (ver ejercicio de la guía).
- La variabilidad explicada por la proyección al espacio de dimensión s es dada por :

$$\frac{\lambda_1 + \cdots + \lambda_s}{\lambda_1 + \cdots + \lambda_K} \times 100 = \frac{\lambda_1 + \cdots + \lambda_s}{K} \times 100.$$

  (Es dada en $\%$ en general)

Eligir los ejes de proyección de los datos.

Definición

Los ejes que maximizan la varianza de la nube proyectada se llaman **componentes principales**.

Eligir los ejes de proyección de los datos.

Como elegir el número de componentes s pertinentes ? Dos reglas posibles entre otras :

A- La dimensión s se elige observando la caída entre los λ_j succesivos.

B- Regla de Kaiser :

- En el caso estandardizado, las varianzas de las variables de partida son iguales a 1.
- Si construimos un componente principal con varianza > 1, su poder explicativo es superior a las variables originales.

$\Rightarrow$ Quedamos nos con los componentes principales con una varianza superior a 1.

- Aplicación : Hacer una ACP de una nube de puntos en R (ver los datos "autos.txt", nos vamos a estudiar más en detalle este conjunto de datos en la guía).
- Sin embargo se queda pendiente una pregunta : ¿Como interpretar la proyección de los puntos a la luz de las variables originales ?

Representación de las variables a la luz de la proyección de la nube de puntos.

$\times$	V_1	$\dots$	V_l	$\dots$	V_s
1	v_{11}	$\dots$	v_{1l}	$\dots$	v_{1s}
$\vdots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
i	v_{i1}	$\dots$	v_{il}	$\dots$	v_{is}
$\vdots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
I	v_{I1}	$\dots$	v_{Il}	$\dots$	v_{Is}

- El individuo i tiene componente v_{il} en el eje l, $l = 1, \dots, s$.
- Sea $(v_{1l}, \dots, v_{Il})'$ los componentes de los I individuos en el eje l.
- Así podemos formar un vector $V_l = (v_{1l}, \dots, v_{Il})' \in \mathbb{R}^I$.

Representación de las variables a la luz de la proyección de la nube de puntos.

- Podemos calcular la correlación entre X_k, $k = 1, \ldots, K$ y los V_l, $l = 1, \ldots, s$:

$$corr(X_k, V_l) = \cos(\theta_{kl}),$$

  donde $\cos(\theta_{kl})$ es el angulo entre los vectores X_k y V_l.
- Obtenemos una representación de las variables X_k en un espacio de dimensión s.
- Se puede mostrar que este espacio maximiza la inercia (la variabilidad) de representación de las K variables en un espacio de dimensión s.
- En este sentido las representaciones de las variables y de los puntos son relacionados.

Representación de las variables a la luz de la proyección de la nube de puntos.

Interpretación :

- Caso 1 : Suponemos una cierta variable X_k y un cierto vector V_l tales que

$$corr(X_k, V_l) \approx 1.$$

- Si $v_{il} \gg 0$ entonces $x_{ik} \gg 0$, $v_{il} \ll 0$ entonces $x_{ik} \ll 0$.
- "El eje l es una buena approximación de la variable X_k".

Representación de las variables a la luz de la proyección de la nube de puntos.

Interpretación :

- Caso 2 : Suponemos una cierta variable X_k y un cierto vector V_l tales que
$$corr(X_k, V_l) \approx -1.$$
- Si $v_{il} \gg 0$ entonces $x_{ik} \ll 0$, $v_{il} \ll 0$ entonces $x_{ik} \gg 0$.
- "El eje l es una buena approximación de la variable $-X_k$".

Representación de las variables a la luz de la proyección de la nube de puntos.

Interpretación :

- Caso 3 : Suponemos una cierta variable X_k y un cierto vector V_l tales que

$$corr(X_k, V_l) \approx 0.$$

- El componente l en el espacio proyectado no se interpreta en terminos de la variable X_k.
- Si $corr(X_k, V_l) \approx 0$, para todos los V_l, $l = 1, \ldots, s$, con

$$\frac{\lambda_1 + \cdots + \lambda_s}{\lambda_1 + \cdots + \lambda_K} \times 100 \approx 100,$$

entonces la variable X_k no tiene un gran poder explicativo de la variabilidad de la nube de puntos original.

- Aplicación : Interpretar la ACP de los datos "auto.txt" al respecto de las variables originales.

Agregar otras variables.

Variables cuantitativas :

- ACP como etapa previa de una regresión lineal : Hacemos la proyección de la variable dependiente sobre las variables X para identificar los regresores pertinentes.

Variables cualitativas :

- No se pueden incluir las variables cualitativas en una ACP de partida ! ! (correlaciones....)
- Sin embargo, visualizamos variables cualitativas en la nube proyectada para mejor comprensión.

Agregar otras variables.

Ejemplo :

- Tenemos la variable "peso" de personas en nuestro estudio.
- Sale que la variable peso tiene correlación cerca 1 con el eje 1.
- Los puntos con la modalidad "hombres" deberían salir a la derecha y las mujeres a la izquierda (variable cualitativa "sexo").

Agregar nuevos individuos.

Detección de outliers.

- Un individuo es sospechoso (error en el valor, fraude....).
- Cuando $K > 3$ es muy muy dificil ver lo en la nube de puntos....

$\Rightarrow$ Lo agregamos a la salida del ACP : Si sale de la nube proyectada, es que es un "outlier"

En la mayoría de los casos en la practica :

1- Hay un punto que sale raro, y que tiene una contribución (en $\%$) a la construcción de los ejes demasiado grande.

2- Lo sacamos, hacemos la ACP, y lo agregamos de nuevo (sin que sirve a la construcción de los ejes !).

El punto 3 tiene la misma idea que los residuos estudentizados "externos" en regresión.

Problema de colinealidad : Recordatorio

- Hacemos la regresión :

$$Y = X\beta + \epsilon,$$

- El estimador MCO :

$$\hat{\beta} = (X'X)^{-1}X'Y,$$

- Si tenemos casi-colinealidad entre variables, podemos invertir $X'X$ pero vamos a tener mala precisión ! ! (ver el código "col.problem.txt")
- Demasiado variables explicativas : "over parameterization"

Etapa previa a la regresión : más vale prevenir que curar

- Sirve a detectar variables que pueden ser fuentes de colinealidad antes de hacer la regresión

$\Rightarrow$ Si hay dos variables que tienen correlación fuerte (**angulo pequeño y cerca el circulo unitario**), entonces pueden ser fuente de colinealidad.

Después la regresión : Curar...

Hicimos diagnostico de colinealidad. Soluciones posibles :

- Sacamos una variable :

$\Rightarrow$ Perdida de información descontralada y no minimizada...,

$\Rightarrow$ Difícil hacerlo cuando tenemos muchas variables

- Aplicar una "ridge regresión" $(X'X + cI_K)^{-1}$ en vez de $(X'X)^{-1}$

$\Rightarrow$ Introduce sesgo en la estimación. (igual solución competitiva con el ACP y PLS...

Después la regresión : Curar...

Regresión por componentes principales :

$\Rightarrow$ Hacemos la regresión :

$$Y = V\widetilde{\beta} + u,$$

donde $V = (V_1 : V_2 : \ldots : V_s)$, con s no necesariamente $s \ll K$ (no estamos tratando de resumir a una o dos dimensiones en este caso pero de arreglar el problema de colinealidad).

- Sin embargo se puede perder el sentido de las variables a veces.

Regresión PLS : Partial Linear Regression

A veces podemos dudar que la *regresión por componentes principales* sea la buena solución :

- En el caso anterior construimos los regresores de manera independientes a la variable explicada.
- Sin embargo eso puede parecer contradictorio a la idea de predicción de la variable explicada.
- El espacio obtenido puede ser no explicativo de manera optimal por la variable explicada ! !

Regresión PLS

A veces tenemos más de una variable explicada

$\Rightarrow$ Queremos construir componentes principales de las variables explicativas que explican las variables explicadas.

- Sin embargo no de manera independiente, pero maximizando la covarianza de variables explicativas y variables explicadas.

Ejemplo en Química (aplicación más común de la PLS) :

- Reacción química : $A + B + C + D + E \rightarrow W + X + Y + Z$
- Hacemos CP de las variables explicativas A, B, C, D, E al respeto de W, X, Y, Z

Regresión PLS

Notas :

- Dado que hay varias variables explicadas, Y es una matriz
- $Y = X\beta + \epsilon$, es tal que β es una matriz y ϵ también.

Regresión PLS

Algoritmo : Nos vamos a construir una sequencia de regresores X_i, $i = 0, 1, 2, ...$

1- Empezar con $X_0 = X$.

2- Por la etapa $\mathbf{a}$, buscar un eje de X_{a-1}, (que se escribe $X_{a-1}w_a$) que maximiza la correlación con un eje de Y (que se escribe Yu_a) :

$$(w_a, u_a) = \arg\max_{u,w}\{< X_{a-1}w, Yu >: \|u\| = 1, \|w\| = 1\}.$$

3- Definir el nuevo componente ortogonal de norma 1 :

$$t_a = X_{a-1}w_a / \|X_{a-1}w_a\|$$

4- Definir el X_a por el siguiente paso que sea ortogonal a t_a :

$$X_a = X_{a-1} - t_a(t_a' X_{a-1}).$$

5- Parrar cuando un cierto criterio es alcanzado.

Regresión PLS

Observaciones :

- Paso 4 : Dado que $t_a' t_a = 1$ (t_a es de norma 1), es facil de ver que $t_a' X_a = 0$.
- Dado que $t_{a+1} = X_a w_{a+1} / \|X_a w_{a+1}\|$, y que X_a es ortogonal a t_a, es facil de ver que t_{a+1} es ortogonal a t_a que es ortogonal a t_{a-1}....etc
- Las predicciones son obtenidas así : $T = \{t_1, \ldots, t_a\}$.

$$\hat{\beta}_a = (T'T)^{-1} T'Y$$

$$\widehat{Y}_a = T\hat{\beta}_a$$

- Paso 5 : El criterio de `STOP` puede ser de validación cruzada : para cada individuo i : 1- Lo sacamos. 2- Calculamos su predicción $\widehat{Y}_a(i)$. 3- Calculamos el error $e_i = Y - \widehat{Y}_a(i)$. 4- Evaluamos $\sum e_i^2$.

Existe una versión donde se reduce también la dimensión de Y junto a X siguiendo pasos similares.

Porqué con ACP ?

Dado que el clustering se hace sobre datos resumidos por el ACP :

- El clustering se puede hacer con una visualización de los datos.
- Permite intervenir si queremos cambiar cosas al resultado final.
- El ACP se hace sobre información pertinente seleccionada por ACP.

Cómo funciona ?

La nube de puntos tiene una varianza total.

- Maximizar la varianza inter-grupos (varianza entre los centros ponderados de los grupos).
- Queremos hacer clusters o grupos con minimo de varianza intra-grupos.

La idea es hacer grupos los más homogeneos posibles.

Cómo funciona ?

Formula de Huygens :

$$\text{Varianza total} = \text{Varianza inter clusters} + \text{Varianza intra clusters}$$

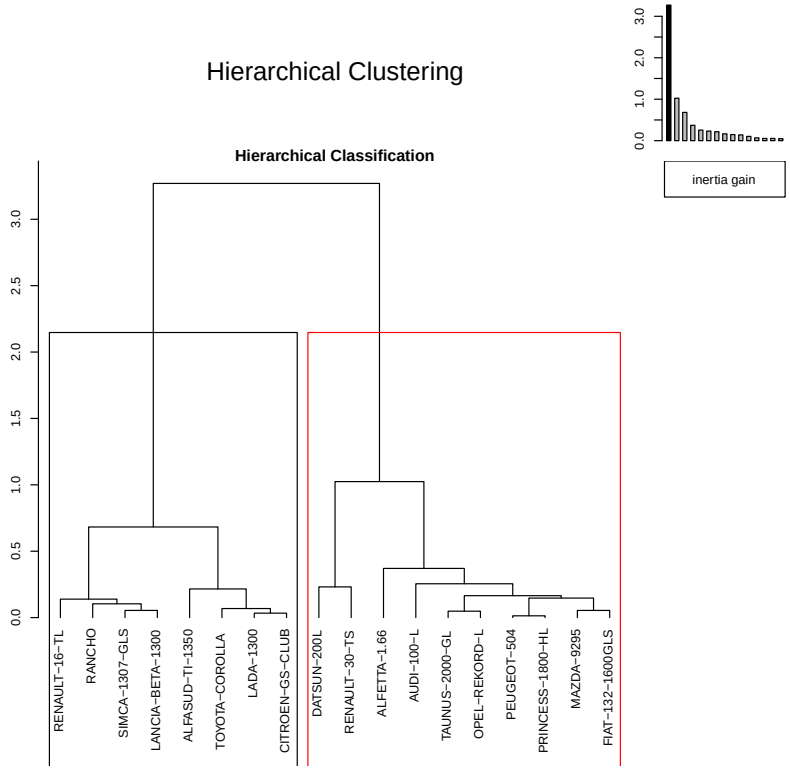
- Varianza total es fija, dada por los datos
- Cada vez que agregamos un cluster nuevo :

$\Rightarrow$ Varianza inter clusters $\uparrow$

$\Rightarrow$ Varianza intra clusters $\downarrow$

Encontrar el número minimo de clusters tal que los cambios $\uparrow$ y $\downarrow$ se estabilizan.

Ejemplo



El eje de las ordenadas y las barras arriba representan las ganancias por la varianza inter-clusters. Hacer dos clusters parece optimal