

Las técnicas del análisis multivariado tienen el objetivo de entender e interpretar datos que están medidos en espacios de gran dimensión, datos que son elementos de espacios multidimensionales.

Supóngase que tenemos un vector de variables \mathbf{x} en \mathbb{R}^p y sea x_1, x_2, \dots, x_n un conjunto de n observaciones de \mathbf{x} .

Tenemos

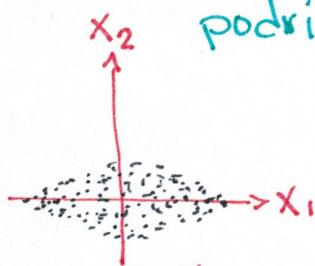
$$x_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \quad y,$$

asimismo

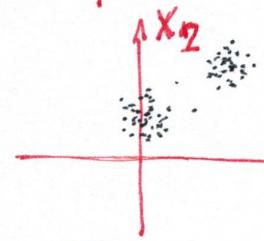
$$\mathbf{x} = (x_1, \dots, x_p)$$

Antes de pensar en llevar a cabo inferencias (estadísticas) a partir de x_1, \dots, x_n , deberíamos pensar en como describir, representar ó "ver" los datos. Si podemos visualizar ó representar gráficamente los datos, de forma que esto permita entender varias características importantes de los mismos, esto nos podría ayudar a seleccionar las herramientas matemáticas (estadísticas) adecuadas para hacer análisis e inferencias.

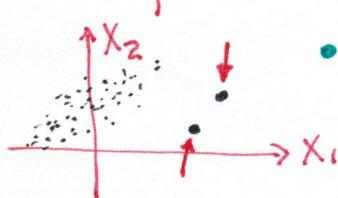
Algunas características de interés que los datos podrían presentar son:



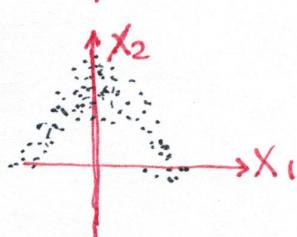
- Algunos de los componentes de \mathbf{X} tienen una dispersión mayor a otras componentes.



- Hay componentes en \mathbf{X} que indican la existencia de subgrupos ó aglomeraciones en los datos.



- Existente "datos aberrantes" (outliers) en alguna componente de \mathbf{X} .



- La distribución "normal" multivariada no es un "buen" modelo para \mathbf{X} .

- Hay combinaciones lineales (de "baja dimensión") de los componentes de \mathbf{X} que tienen un comportamiento alejado del modelo "normal".

Los dibujos a la izquierda, nos permiten pensar que para $P=2$, los "diagramas de dispersión" son de ayuda para nuestros objetivos. Existen técnicas computacionales y programas de cómputo modernos, que permiten visualizar datos tridimensionales ($P=3$) así como rotaciones de los mismos en

tiempo real. ¿Qué técnicas podríamos usar si $p > 3$?

Revisaremos algunas técnicas de representación gráfica para datos multivariados cuya finalidad es describir características de los datos, como las mencionadas en la lista anterior.

DIAGRAMAS Ó GRÁFICAS DE CAJAS (BOX PLOTS)

Los diagramas de cajas son gráficos para representar la distribución de las variables X_1, X_2, \dots, X_p . Con estos se puede vislumbrar características de la distribución de cada componente X_i , las características podrían ser: localización, sesgo, dispersión, longitud (pesadez) de las colas y valores "aberrantes" (observaciones que aparecen fuera del rango de posibles valores de X_i).

En particular, las cajas permiten comparar estas características para dos (X_i y $X_j; i \neq j$) ó más variables (X_i, X_j y $X_k \quad i \neq j; j \neq k; i \neq k$)

Las cajas usan un resumen de los datos que consta de cinco estadísticas (Five-number summaries):

4

- El cuartil superior F_U
- El cuartil inferior F_L
- = La mediana
- = El mínimo y el máximo de la muestra
(los valores extremos de la muestra).

Consideremos una muestra de observaciones x_1, \dots, x_n y sean $\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$ las correspondientes "estadísticas de orden" (los valores x_1, \dots, x_n ordenados de acuerdo a su magnitud). De esta forma $x_{(1)} = \min\{x_i : 1 \leq i \leq n\}$ y $x_{(n)} = \max\{x_i : 1 \leq i \leq n\}$

La mediana es el número que deja la mitad de los datos en $\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$ antes de él y la mitad de los datos en $\{x_{(1)}, \dots, x_{(n)}\}$ después de él. La mediana se puede definir como

$$M \equiv \begin{cases} x_{(\frac{n+1}{2})} & \text{si } n \text{ es impar,} \\ \frac{1}{2}\{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\} & \text{si } n \text{ es par.} \end{cases}$$

Para definir los cuartiles F_U y F_L :

1. Usemos M para dividir $\{x_{(1)}, \dots, x_{(n)}\}$ en

dos subconjuntos (los $x_{(i)}$'s antes y después de M).

(a) Si hay un número impar de datos en $\{x_{(1)}, \dots, x_{(n)}\}$, incluimos M en las dos listas:

$$C_1 = \{x_{(1)}, x_{(2)}, \dots, x_{(\frac{n+1}{2})} = M\},$$

$$C_2 = \{x_{(\frac{n+1}{2})} = M, x_{(\frac{n+1}{2} + 1)}, \dots, x_{(n)}\}.$$

(b) Si hay un número par de datos en $\{x_{(1)}, \dots, x_{(n)}\}$, la mediana M no es un elemento a considerar, es decir las listas son:

$$C_1 = \{x_{(1)}, x_{(2)}, \dots, x_{(\frac{n}{2})}\} \quad y$$

$$C_2 = \{x_{(\frac{n}{2} + 1)}, x_{(\frac{n}{2} + 2)}, \dots, x_{(n)}\}$$

2: El cuartil F_L es la mediana de los datos C_1 .
El cuartil F_U es la mediana de los datos C_2 .

El cálculo de F_L y F_U usando los pasos 1 y 2 se conoce en literatura como "Método de Tukey"

ejemplo: Datos de tamaños poblacionales de las quince ciudades más grandes en 2006.

City	Country	Pop. (10,000)	Order statistics
Tokyo	Japan	3,420	$x_{(15)}$
Mexico city	Mexico	2,280	$x_{(14)}$
Seoul	South Korea	2,230	$x_{(13)}$
New York	USA	2,190	$x_{(12)}$
Sao Paulo	Brazil	2,020	$x_{(11)}$
Bombay	India	1,985	$x_{(10)}$
Delhi	India	1,970	$x_{(9)}$
Shanghai	China	1,815	$x_{(8)}$
Los Angeles	USA	1,800	$x_{(7)}$
Osaka	Japan	1,680	$x_{(6)}$
Jakarta	Indonesia	1,655	$x_{(5)}$
Calcutta	India	1,565	$x_{(4)}$
Cairo	Egypt	1,560	$x_{(3)}$
Manila	Philippines	1,495	$x_{(2)}$
Karachi	Pakistan	1,430	$x_{(1)}$

Para estos datos, $n=15$, $M=x_{(8)}=1.815$, $F_L=1.610$

$F_U=2.105$, $x_{(1)}=1.430$, $x_{(15)}=3.420$.

La F-dispersión d_F se define como

$$d_F \equiv F_U - F_L .$$

Las barres externas se definen como

$$b_U \equiv F_U + 1.5 \cdot d_F , \quad b_L \equiv F_L - 1.5 \cdot d_F$$

b_U y b_L son los números (los fronteras) más allá de los cuales, un dato se considera ó se clasifica como un valor aberrante (outlier).

Para los datos de tamaños poblacionales de las ciudades tenemos

$$d_F = F_U - F_L = 2105 - 1610 = 495$$

$$b_L = F_L - 1.5 \cdot d_F = 1610 - 1.5 \cdot 495 = 867.5$$

$$b_U = F_U + 1.5 \cdot d_F = 2105 + 1.5 \cdot 495 = 2847.5$$

El diagrama de caja también reporta la media de los datos

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 1939.7 \leftarrow \text{para los datos de tamaños poblacionales}$$

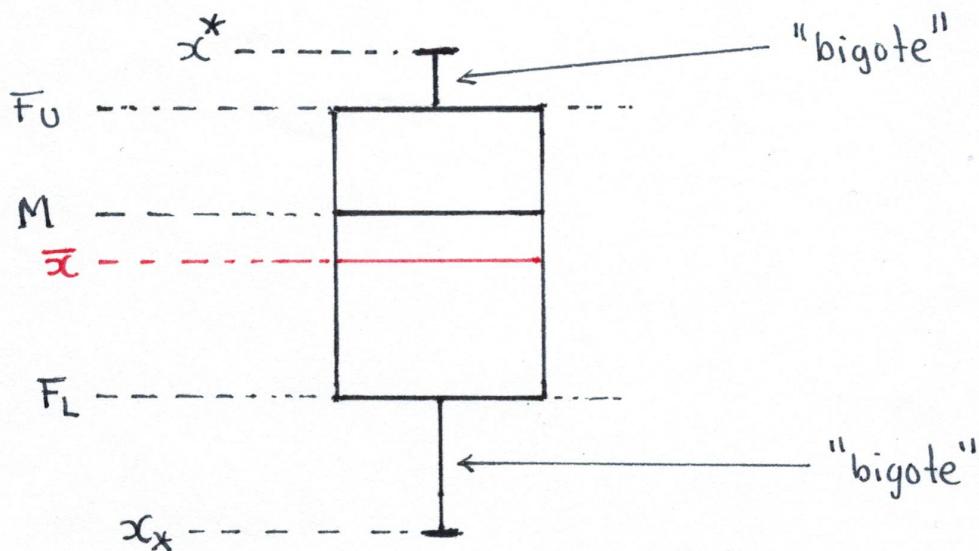
CONSTRUCCION DEL DIAGRAMA DE CAJA

- 1 Se dibuja una caja con bordes, arriba y abajo, en F_U y F_L respectivamente. El 50% de los datos quedan dentro de la caja.
- 2 Se dibuja una línea sólida en el valor de M y una línea punteada en el valor de \bar{x} .

3 Se dibujan líneas verticales ("bigotes") que van desde los bordes (arriba y abajo) de la caja y que llegan hasta los valores x^* y x_* dados por:

$$x^* = \max \{x \in \{x_{(1)}, \dots, x_{(n)}\} : x \leq b_U\},$$

$$x_* = \min \{x \in \{x_{(1)}, \dots, x_{(n)}\} : b_L \leq x\}.$$



4 Algunos paquetes de cómputo, muestran los datos aberrantes (outliers) con un carácter "*", si estos yacen fuera del intervalo (b_L, b_U) o con un carácter "●" si estos yacen fuera del intervalo $(F_L - 3 \cdot d_F, F_U + 3 \cdot d_F)$.

Para el ejemplo de los tamaños poblacionales de las ciudades tenemos que:

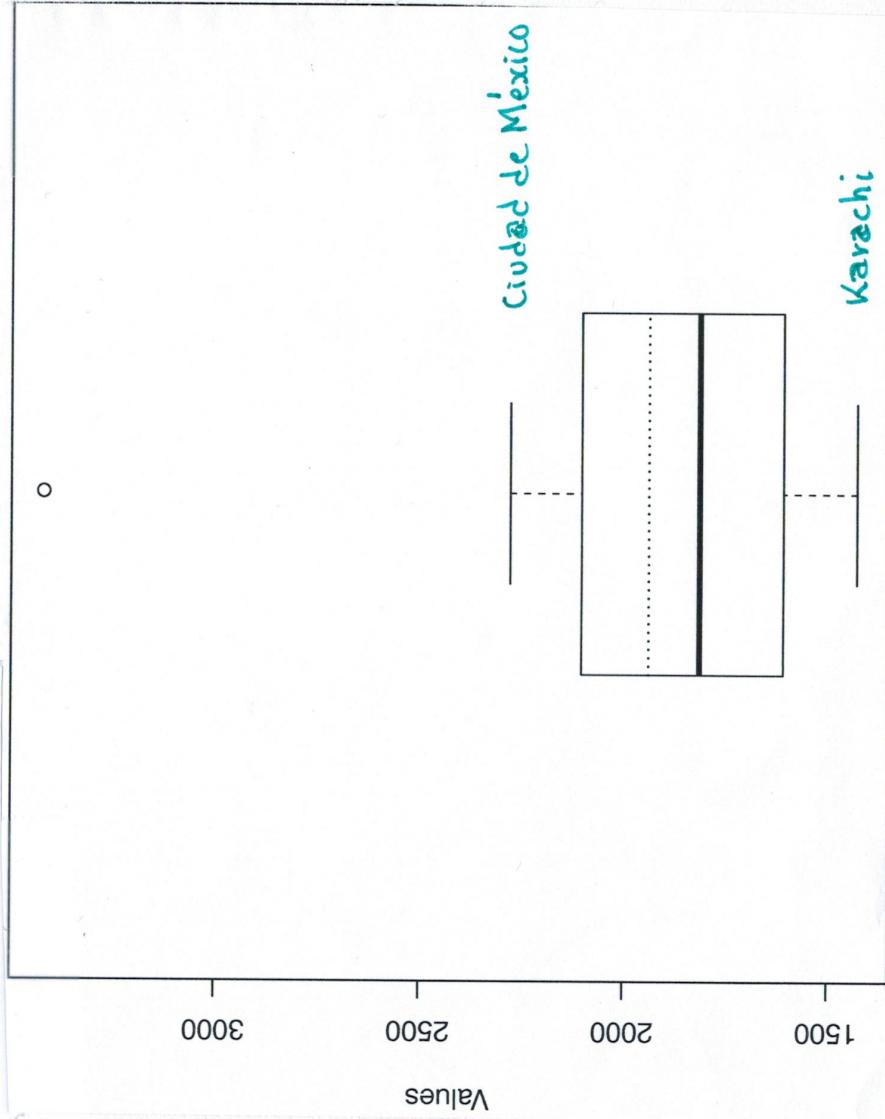
9

$$x^* = 2280 \quad (\text{Ciudad de México})$$

$$x_x = 1430 \quad (\text{Karachi})$$

Entonces los bigotes se dibujan desde el borde superior de la caja, hasta el valor x^* (ciudad de México), en la parte de arriba de la caja y, desde el borde inferior de la caja hasta el valor x_x (Karachi), en la parte de abajo de la caja.

Boxplot



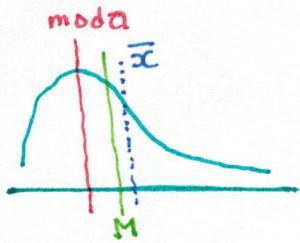
World Cities

El diagrama de caja muestra que los datos de tamaños poblacionales de las ciudades tienen un sesgo⁽¹⁾ hacia arriba (hay falta de simetría en la distribución de $X = \text{tamaño poblacional}$).

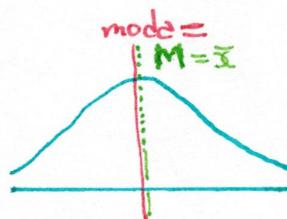
La mitad superior (los datos arriba de la mediana) de la muestra tiene una dispersión mayor que la mitad inferior. Hay un dato aberrante marcado con carácter "O" y este corresponde a la ciudad de Tokio. Debido a la sobredispersión en la mitad de los datos que están arriba de M la media, que no es una medida de tendencia central robusta, está desplazada hacia arriba y no coincide con la mediana.

- (1) Si $\bar{x} \neq M$ podemos decir que tenemos evidencia muestral que nos sugiere la hipótesis de que la distribución está sesgada (la distribución de X).

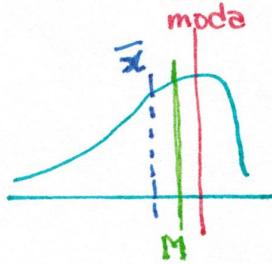
En el caso de distribuciones unimodales, en general pero **no** siempre, se tiene



(a) Sesgo positivo



(b) Simétrico



(c) Sesgo negativo

Ejercicio: Para los datos correspondientes al rendimiento, millas por galón de combustible, de los automóviles provenientes de Japón, Norte América y Europa, encontrar en cada caso, las estadísticas:

$$x^*, x^*, M, F_L, F_U, b_L, b_U, \bar{x}$$

Haga un análisis de estos datos usando los diagramas de cajas, desarrolle sus conclusiones.¹

¹ Las conclusiones son parte importante del ejercicio.

Ejemplo: Datos de billetes del banco de Suiza

$n=200$ observaciones x_1, \dots, x_{200} ,

$$x_i = (x_{i1}, x_{i2}, \dots, x_{i6}) ; i=1,2,\dots,200.$$

Las observaciones corresponden a $\mathbf{x} = (x_1, x_2, \dots, x_6)$
donde

x_1 = largo del billete

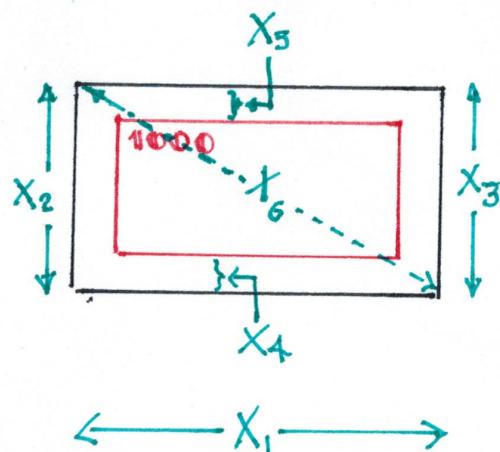
x_2 = Ancho del billete (izquierda)

x_3 = Ancho del billete (derecha)

x_4 = Distancia de la Figura en el billete
al borde inferior del billete.

x_5 = Distancia de la Figura en el billete
al borde superior del billete

x_6 = Longitud de la Diagonal del billete



Objetivo: Estudiar cómo estas mediciones podrían usarse para determinar si un billete es genuino ó si es falso.

Preguntarse como determinar si un billete es verdadero o falso usando estas seis mediciones esta relacionado con una de las características de interés que se mencionaron al inicio del tema del análisis exploratorio de datos:

Ver página →
2

¿Hay componentes de \mathbf{X} que indican la existencia de subgrupos ó aglomeraciones en los datos?

Dicho de otra forma, nos interesaría saber si una de las seis mediciones X_1, \dots, X_6 nos permite ver dos subgrupos dentro de los datos: los verdaderos y los falsos.

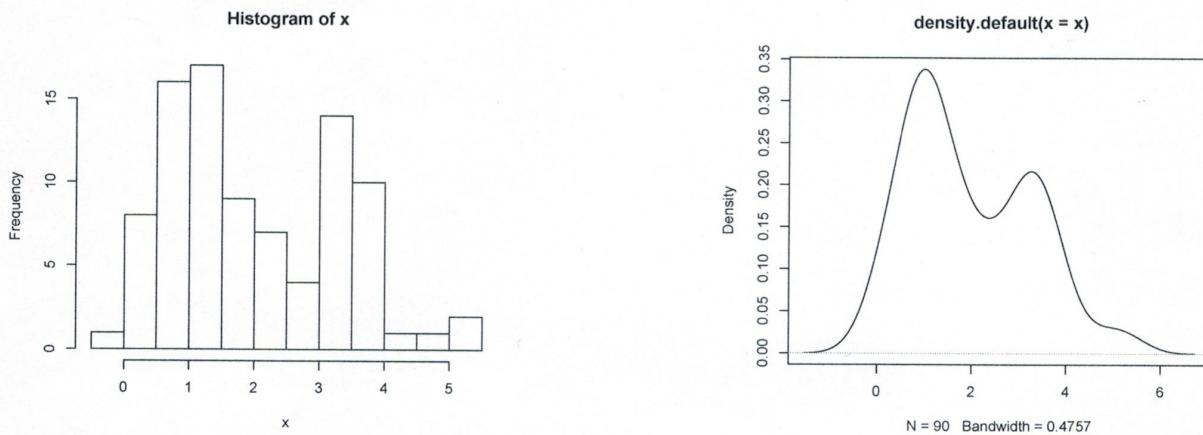
En este caso, partir directamente de los diagramas de cajas posiblemente no sea una idea muy útil, estos diagramas son representaciones unidimensionales de la distribución de una variable aleatoria (cualquier X_1, X_2, \dots, X_6). La forma

A

en que el diagrama de caja esta construido
no permite al usuario de esta representación
determinar si la distribución de X_i tiene
más de una moda. Por ejemplo, si simulamos
una muestra en donde hay bimodalidad

```
> x <- c(rnorm(50,1,0.5), rnorm(40,3,1))  
> hist(x)  
> plot(density(x))
```

Figura A datos simulados

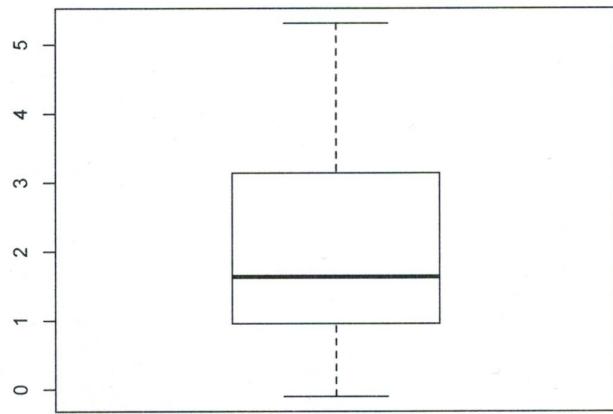


En la realidad, podemos tener una muestra como
esta, es decir una muestra en donde al parecer

hay dos subpoblaciones que juntas forman la población total. Pensemos por un momento que la muestra no fue simulada, que los datos fueron observados con la naturaleza que muestran las figuras, entonces al percetarnos de que hay dos modas, la idea de que hay dos tipos de individuos en la población es inmediata. A continuación tendríamos que investigar si las observaciones de uno de estos dos grupos provienen de individuos con una característica particular, por ejemplo si hablamos de los billetes, hay que averiguar si podemos "clasificar" los individuos de un grupo como billetes falsos y a los individuos del otro grupo como billetes verdaderos. Como ya se mencionó arriba los diagramas de caja no nos ayudan a determinar si hay dos (o más) grupos : La Figura B muestra el diagrama de caja correspondiente a los datos simulados en la Figura A

> boxplot(x, axes=TRUE, frame=TRUE)

Figura B datos simulados



Hasta el momento, hemos ilustrado que la posible existencia de subgrupos en un conjunto de datos, puede ser detectado (al menos en el caso unidimensional) usando histogramas y estimaciones de la densidad. Recomendamos al lector que revise estos temas en libros de Análisis Estadístico de Datos y Analysis de datos Multivariados, por ejemplo: "Data Analysis and Graphics Using R : An example-based approach", John Maindonald

and John Braun, Cambridge University Press.

De acuerdo a lo que hemos discutido, una opción para determinar si alguna componente ~~de~~ $\mathbf{X} = (X_1, X_2, \dots, X_n)$ nos ayuda a encontrar subgrupos en los datos, sería producir una gráfica con histogramas y/o densidades estimadas para todas las componentes en \mathbf{X} . No obstante, vamos a aprovechar para presentar otra gráfica de datos que suele usarse para estudiar comportamientos y posibles dependencias de datos con dimensión mayor a 1.

DIAGRAMAS DE DISPERSION (Scatterplots)

Los diagramas de dispersión son gráficos de varias componentes del vector $\mathbf{X} = (X_1, \dots, X_n)$.

Por ejemplo, para el caso de los billetes del banco Suizo $\mathbf{X} = (X_1, \dots, X_6)$ y

dados los datos x_{i1}, \dots, x_{i200} , donde
 $x_{i1} = (x_{i1}, \dots, x_{i6})$, podemos graficar
los puntos $(x_{i1}, x_{i2}) ; i=1,2,\dots,200$
en \mathbb{R}^2 , lo cual nos da idea de cómo varía
el vector (X_1, X_2)

```
> datos <- read.table("SwissBank 1.txt")
> x <- datos[,1]
> y <- datos[,2]
> plot(x,y)
```

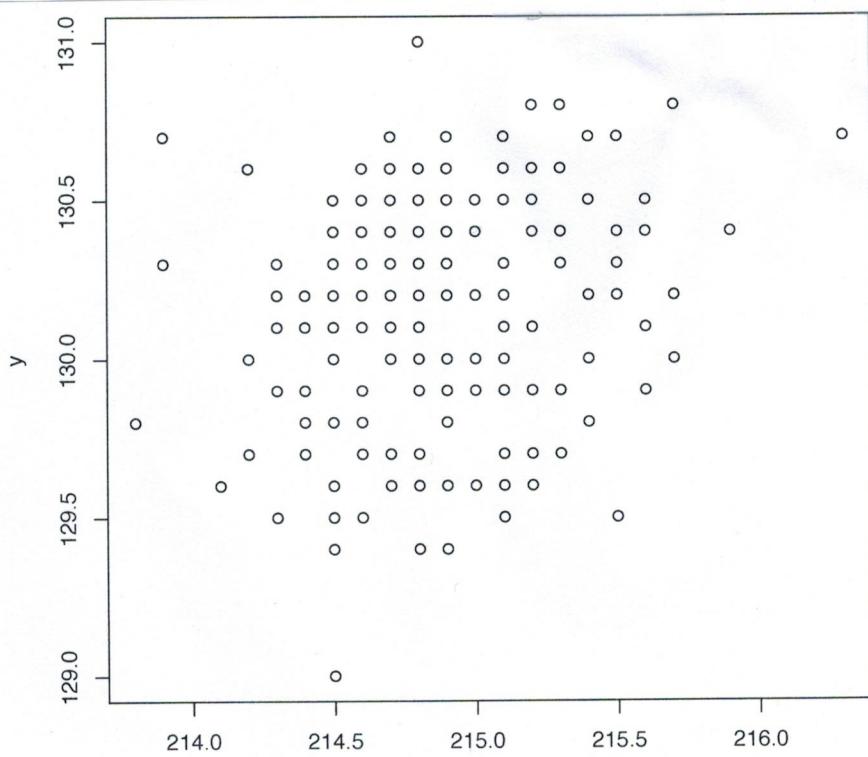


Figura C

Se pueden explorar todas las posibilidades, i.e. graficar (x_{ik}, x_{il}) ; $i=1,2,\dots,200$; $k \neq l$ $k,l \in \{1,2,\dots,6\}$. La buena noticia es que ya existen funciones en R para hacer esto

Variables Ordered and Colored by Correlation

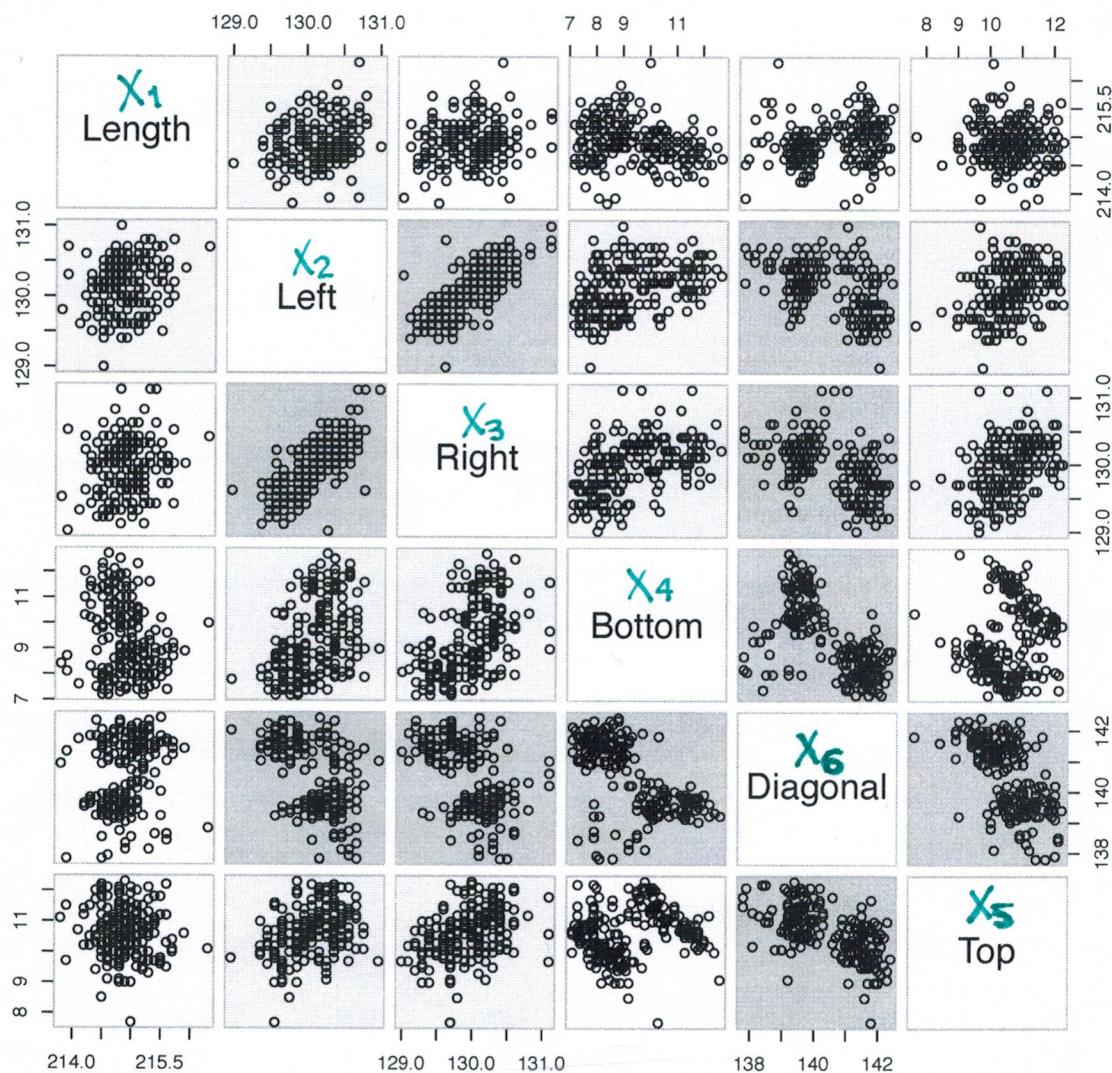


FIGURA D

```

notes1.R <- function(){
# produces a scatter plot of swiss banknote data
# first load package uskewFactors v2.0, this contains
# Swiss banknote data. Then invoke data(banknote)
# This function requires package gclus

pdf("scatter banknotes1.pdf")
library(uskewFactors)
data(banknote)
library(gclus)
dta <- banknote[c(1,2,3,4,5,6)]
dta.r <- abs(cor(dta))
dta.o <- order.single(dta.r)
dta.col <- dmat.color(dta.r)
cpairs(dta, dta.o, panel.colors=dta.col, gap=.5, main="Variables Ordered and Colored by
Correlation" )
dev.off()

}

```

La Figura D contiene los diagramas de dispersión de todos los vectores (x_i, x_j) ; $i \neq j$; $i, j \in \{1, 2, \dots, 6\}$ para las componentes x_i y x_j de \mathbf{x} . Observemos que el renglón 5 y la columna 5 de esta matriz muestran diagramas de dispersión para los cuales hay una clara separación de todos (los 200 individuos) los billetes en dos subgrupos. Este renglón y columna corresponden a la interacción de las componentes x_1, x_2, x_3, x_4 y x_5 con

La componente X_6 = longitud de la diagonal del billete. Lo anterior nos hace pensar que es esta componente (esta característica de los billetes) la que nos puede ayudar a determinar 2 subpoblaciones⁽¹⁾: Los billetes genuinos y los billetes falsos.

El problema de clasificación en la estadística, consiste en determinar qué individuos en la muestra pertenecen a un subgrupo y cuáles a otro subgrupo. Además, es de interés estudiar si el método empleado servirá para clasificar nuevos individuos, provenientes de la misma población, que se observen a futuro.

En principio, nosotros vamos a revisar metodología para clasificar en este curso.

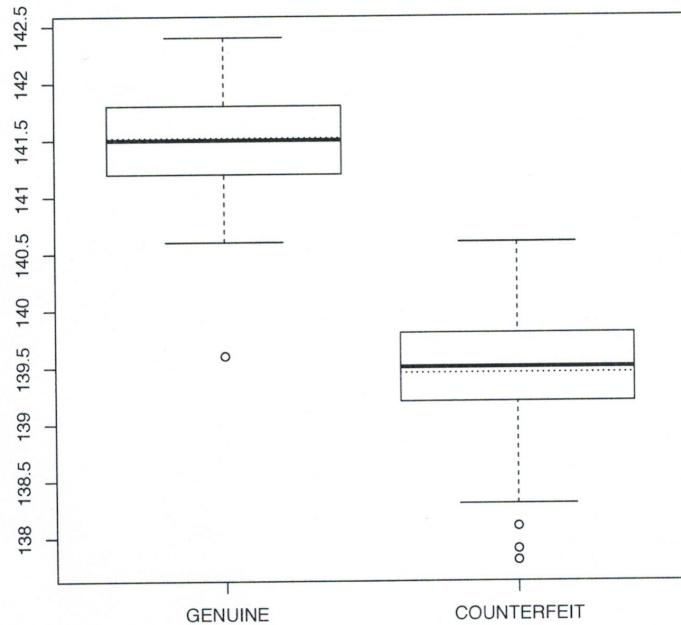
(1) El nombre correcto aquí sería subgrupos porque al principio hablamos de una muestra

Al momento, nosotros intuimos que es la componente $X_6 = \text{longitud de la diagonal}$, la que nos permitirá clasificar⁽¹⁾, pero no tenemos (todavía) conocimiento de métodos de clasificación. En espera de aprender sobre métodos de clasificación posteriormente en el curso, supóngase que ya tenemos información de la muestra, sobre qué valores de $x_{1,6}, \dots, x_{200,6}$ corresponden a billetes falsos y, qué valores corresponden a billetes genuinos. Asumiendo que $x_{1,6}, \dots, x_{100,6}$ corresponden a billetes verdaderos y $x_{101,6}, \dots, x_{200,6}$ corresponden a billetes falsos, la función de R "MVA boxbank6.R" produce un gráfico de cejas para comparar estos dos subgrupos.

- (1) Conocer esto, es producto de nuestras técnicas para graficar, representar y explorar los datos. Este conocimiento se usará como insumo de los métodos de clasificación.

FIGURA E

Swiss Bank Notes



A la izquierda se representa la distribución de la longitud de la diagonal X_6 para aquellos billetes genuinos en la muestra. A la derecha la distribución de X_6 para aquellos billetes falsos.

Como se puede observar la longitud de la diagonal para los billetes genuinos tiende a ser mayor que la longitud de la diagonal para los billetes falsos. Si uno repite este gráfico pero usando el largo de los billetes, es decir, la componente X_1 , veremos que la distinción (las diferencias

entre los dos grupos) no son tan claras (usar la función de R "MVAboxbank1.R"). De la Figura E, notamos que casi todas las observaciones de la longitud de la diagonal correspondientes a los billetes genuinos están por encima de las longitudes correspondientes a los billetes falsos. ¿Es la longitud de la diagonal en los billetes del banco de Suiza una forma de distinguir entre billetes falsos y verdaderos?

Las estadísticas M , F_U , F_L , d_f , b_U , b_L , $x_{\bar{x}}$ y x^* se pueden obtener, para estos datos, con la función de R "BoxStats.R".

Ejercicio: - Producir un diagrama de cajas para los dos grupos: billetes genuinos y billetes falsos usando la componente X_1 de \mathbf{X} .

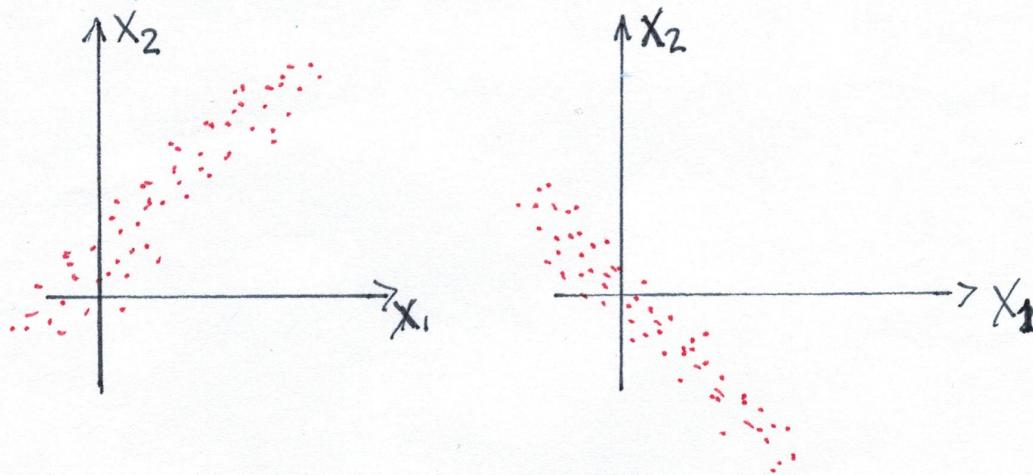
- Calcular las estadísticas M , F_U , F_L , b_U , b_L , $x_{\bar{x}}$ y x^* para los dos grupos usando la componente X_6 .

- Comentar y comparar los dos análisis:
Las cajas para X_1 y las cajas para X_6

- Haga un resumen respecto a cómo funcionan los histogramas y las estimaciones de la densidad para unos datos. Explique con cuidado qué es lo que los paquetes dibujan y describa un ejemplo con datos.

De regreso a la discusión sobre los diagramas de dispersión, estos tienen utilidad más allá de ayudarnos a identificar subpoblaciones.

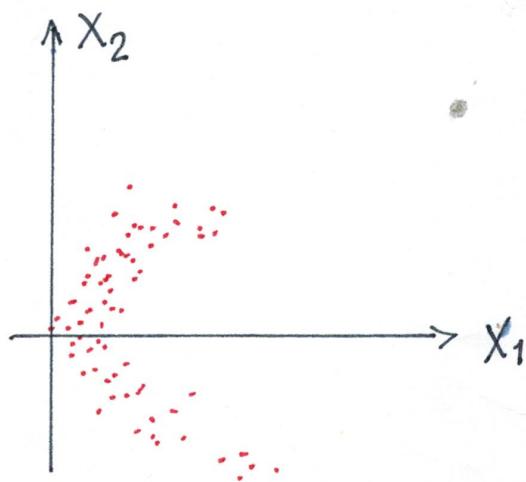
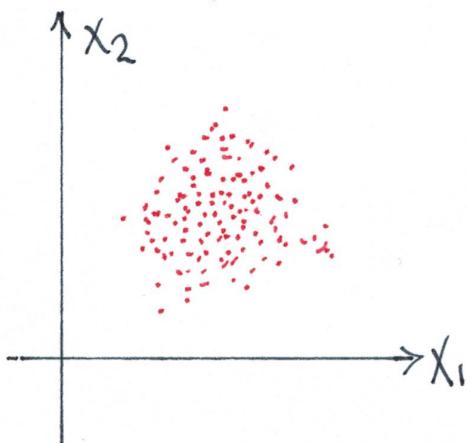
FIGURA
F



En la figura F, los diagramas de dispersión nos muestran que la hipótesis de una posible relación funcional $X_2 = f(X_1)$ tendría sentido.

Asimismo, un diagrama de dispersión podría echar por tierra esta hipótesis

Figura
G



De regreso a los datos de los billetes, en ocasiones un diagrama de dispersión en dimensión mayor que 2 puede ayudar a visualizar los datos.

Las figuras D y E nos permiten intuir que la componente $X_6 = \text{Longitud de la diagonal del billete}$ contiene información para distinguir dos subgrupos en la muestra (los billetes falsos y los verdaderos).

La figura H presenta un diagrama de dispersión en tres dimensiones, en donde se grafican los vectores (X_4, X_5, X_6)

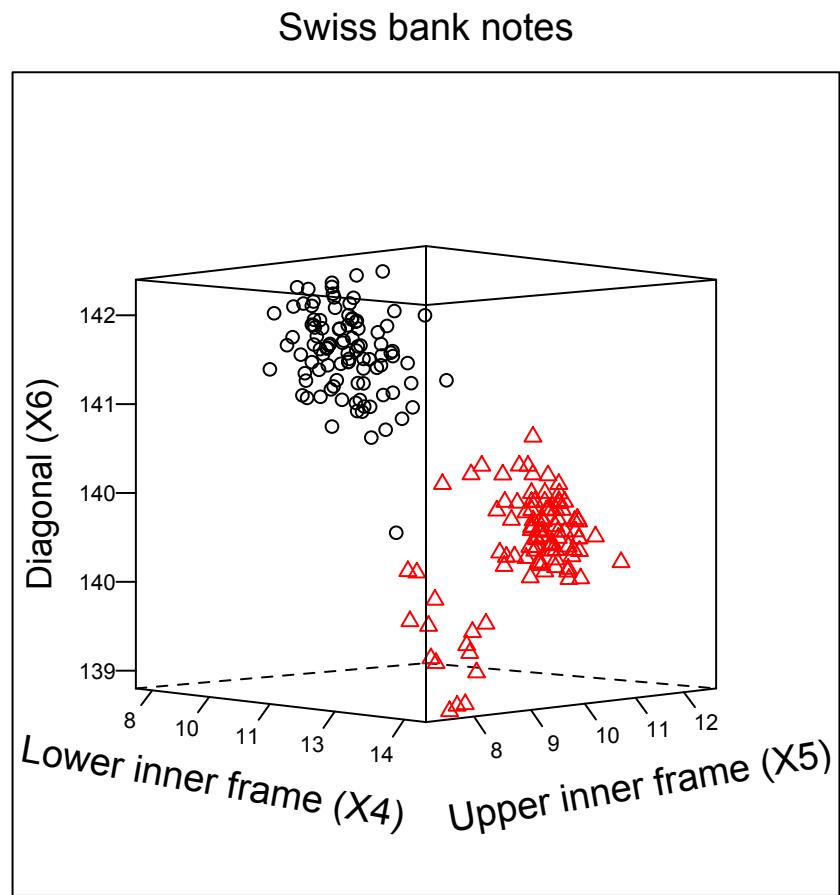


Figura H-1

Swiss bank notes

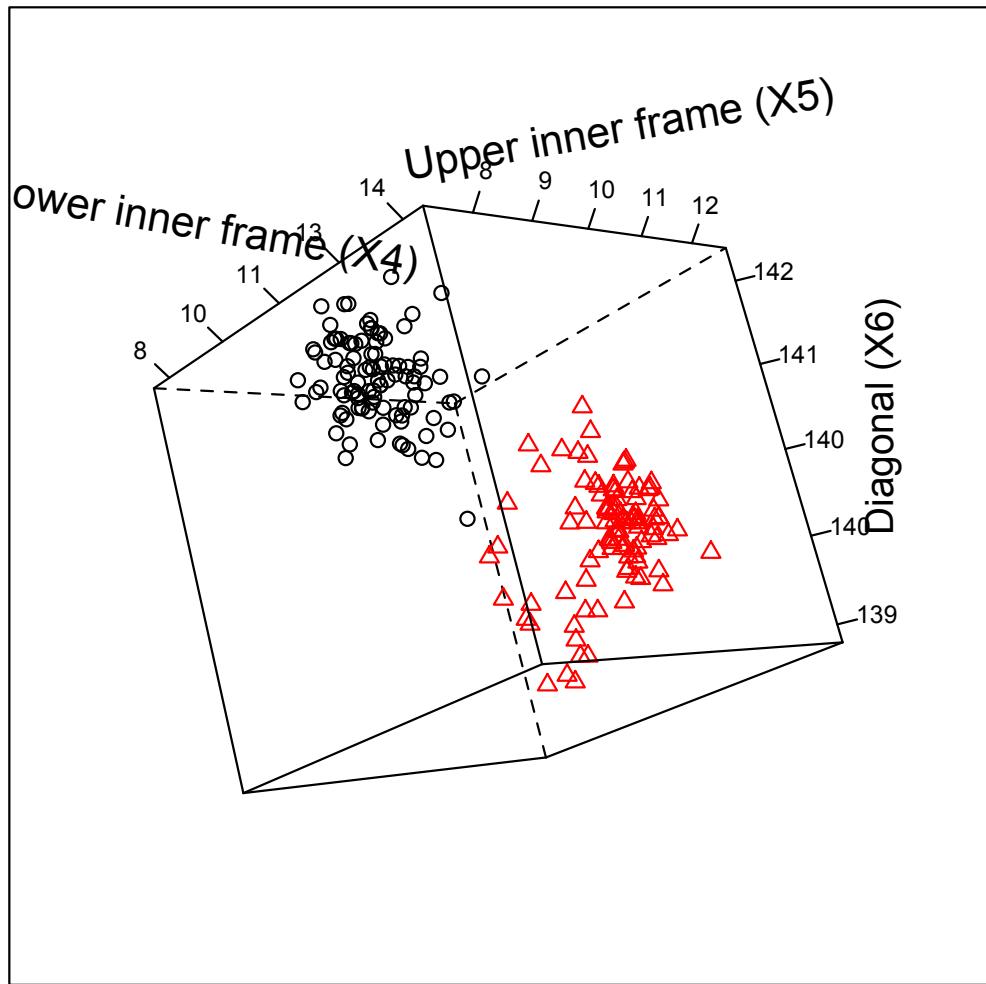


Figura H-2

ALGUNOS GRAFICOS PARA REPRESENTAR DATOS

x_i, \dots, x_n , donde $x_i = (x_{i1}, \dots, x_{ip})$ y $p > 3$

Idee: Transformar cada $x_i \in \mathbb{R}^p$ en un objeto que esté en dos dimensiones (ó tres dimensiones).

1 LAS CARAS DE CHERNOFF-FLURY

Estas representaciones gráficas, condensan la información de x_i en una dibujo de una cara.

El tamaño y color de los rasgos en una cara, como lo son: pupilas, ojos, nariz, etc., se asignan dudos los valores de $x_{i1}, x_{i2}, \dots, x_{ip}$. En algunos paquetes como R, se pueden usar los siguientes rasgos en el dibujo de la cara

- 1 Tamaño del ojo derecho.
- 2 Tamaño de la pupila en el ojo derecho.
- 3 Posición de la pupila en el ojo derecho.
- 4 Inclinación en el ojo derecho.
- 5 Posición horizontal del ojo derecho.
- 6 Posición vertical del ojo derecho.

- 7 curvatura de la ceja derecha .
- 8 densidad de la ceja derecha .
- 9 posición horizontal de la ceja derecha .
- 10 Posición vertical de la ceja derecha .
- 11 Línea superior del cabello (derecha) .
- 12 Línea inferior del cabello (derecha) .
- 13 Línea derecha de la cara
- 14 Tono del cabello (derecha) .
- 15 Inclinación del cabello (derecha)
- 16 Línea de la nariz (derecha)
- 17 Tamaño de la boca (derecha)
- 18 Curvatura de la boca (derecha)

Rasgos del 19 al 36, son los mismos que 1-18 pero para el lado izquierdo.

Cada variable que se asignará a algún rasgo de la cara se transforma primero al intervalo (escala) (0,1), de forma que el mínimo

de los posibles valores de esa variable corresponde a 0 y el máximo de los posibles valores corresponde a 1.⁽¹⁾ Si por ejemplo el rasgo de la cara al cual fue asignada la variable es el tono del cabello, el individuo de la muestra cuyo valor en la variable sea el de mayor magnitud, podría aparecer en los gráficos como la cara con el tono de cabello más oscuro.

Como ejemplo, vamos a considerar los billetes $x_{q1}, x_{q2}, \dots, x_{q10}$ en el caso del banco suizo.

Los primeros diez individuos corresponden a billetes verdaderos y los segundos diez corresponden a billetes falsos (de acuerdo a la información que recibimos para hacer la figura E).

(1) Supóngase que se selecciona la variable j $j \in \{1, 2, \dots, p\}$ entonces $x_0 = \min\{x_{1j}, x_{2j}, \dots, x_{nj}\}$
 $x_1 = \max\{x_{1j}, x_{2j}, \dots, x_{nj}\}$, la transformación lleva (x_0, x_1) a $(0, 1)$.

Para estos datos, vamos a asignar las componentes X_1, X_2, \dots, X_6 a los siguientes rasgos en las caras

$X_1 \rightarrow 1, 19$ tamaños de los ojos.

$X_2 \rightarrow 2, 20$ tamaños de las pupilas.

$X_3 \rightarrow 4, 22$ inclinaciones de los ojos.

$X_4 \rightarrow 11, 29$ líneas superiores del cabello.

$X_5 \rightarrow 12, 30$ líneas inferiores del cabello.

$X_6 \rightarrow 13, 14, 31, 32$ líneas de la cara y tono del cabello

usando estas especificaciones y la función de R "MVAfacebank10.R" obtenemos la figura I

Las caras correspondientes a los dos últimos renglones en la figura I provienen de billetes falsos. Es claro que los dos primeros renglones tienen caras con expresión más "feliz" que las caras de los últimos dos renglones.

```

# -----
# Book:      MVA3
# -----
# See also:  MVAfacebank50
# -----
# Quantnet:  MVAfacebank10
# -----
# Description: MVAfacebank10 computes Flury faces for
#               the Swiss bank notes data ("bank2.dat").
#               Interesting plot is obtained for obs 91-110.
# -----
# Author:    2006-09-15 Julia Wandke
# -----


rm(list=ls(all=TRUE))
graphics.off()

#install.packages("aptpack")
library(aptpack)

# Load data
# The data file should be located in the same folder as this Qlet
# Set the R working directory to this directory using setwd()
# setwd("C:/...")      # set working directory if windows
# setwd("~/Users/...") # set working directory if mac
x = read.table("SwissBank 1.txt")
xx = x[91:110,]

ncolors=15

faces(xx, nrow = 4, face.type=1, scale=TRUE, col.nose = rainbow(ncolors), col.eyes = rainbow(ncolors,
  start = 0.6, end = 0.85), col.hair = terrain.colors(ncolors),
  col.face = heat.colors(ncolors), col.lips = rainbow(ncolors,
  start = 0, end = 1), col.ears = rainbow(ncolors, start = 0,
  end = 0.8), plot.faces = TRUE)

```

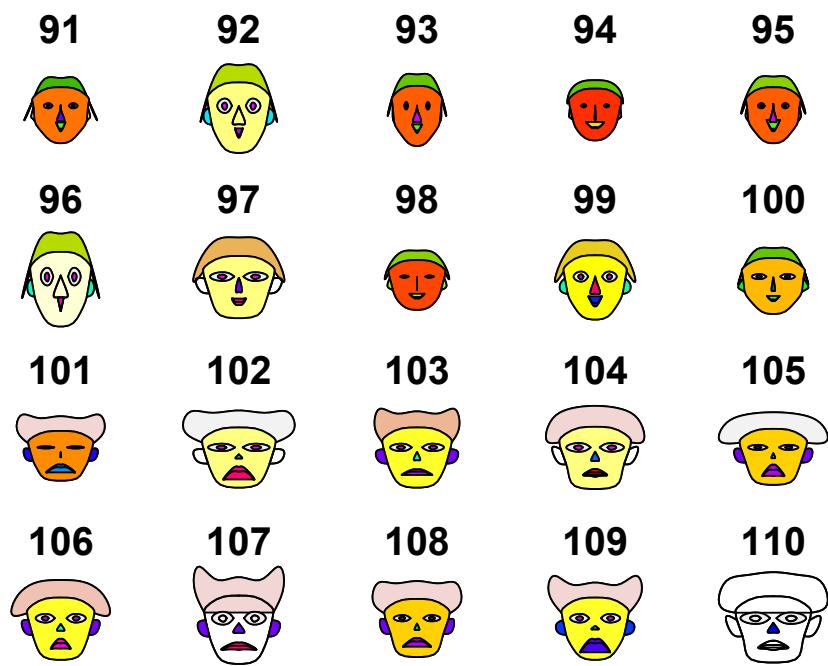


Figura I

Asimismo, el tono del cabello y las líneas de la cara también son diferentes entre estos dos grupos.

Notas: • La función "MVAfacebook10.R" manda un mensaje donde dice que X_6 está asociada con sonrisa. Esto se aprecia en las caras.

- Para clasificación, las caras similares podrían formar subgrupos en los datos.
- Outliers se podrían identificar en caras con rasgos extremos.

2 Las curvas de Andrews

De nuevo es la idea de representar vectores en \mathbb{R}^p ($p > 3$) como objetos en 2 dimensiones.

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \leftrightarrow f_i(t)$$

donde la función $f_i(t)$ está dada por

$$f_i(t) = \begin{cases} \frac{x_{i1}}{\sqrt{2}} + x_{i2} \sin(t) + x_{i3} \cos(t) + \cdots + x_{ip-1} \sin\left(\frac{p-1}{2}t\right) + x_ip \cos\left(\frac{p-1}{2}t\right) & \text{si } p \text{ es impar,} \\ \frac{x_{i1}}{\sqrt{2}} + x_{i2} \sin(t) + x_{i3} \cos(t) + \cdots + x_ip \sin\left(\frac{p}{2}t\right) & \text{si } p \text{ es par.} \end{cases}$$

La idea de usar esta función impone de considerar una serie de Fourier:

$$(A) \dots g(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \{a_n \cos(nt) + b_n \sin(nt)\}; -\pi \leq t \leq \pi$$

La función $f_i(t)$ se puede escribir como en (A)

$$\text{si } a_0 = x_{i1}\sqrt{2}; a_1 = x_{i3}; b_1 = x_{i2}; \dots$$

$$a_{\frac{p-1}{2}} = x_{ip}; b_{\frac{p-1}{2}} = x_{ip-1}; a_{\frac{p-1}{2}+1} = 0 = b_{\frac{p-1}{2}+1};$$

$$a_{\frac{p-1}{2}+2} = 0 = b_{\frac{p-1}{2}+2}; \dots \text{ cuando } p \text{ es impar}$$

Y cuando p es par

$$a_0 = x_{i1}\sqrt{2}; a_1 = x_{i3}; b_1 = x_{i2}; \dots$$

$$a_{\frac{p}{2}-1} = x_{ip-1}; b_{\frac{p}{2}-1} = x_{ip-2}; a_{\frac{p}{2}} = 0; b_{\frac{p}{2}} = x_{ip}$$

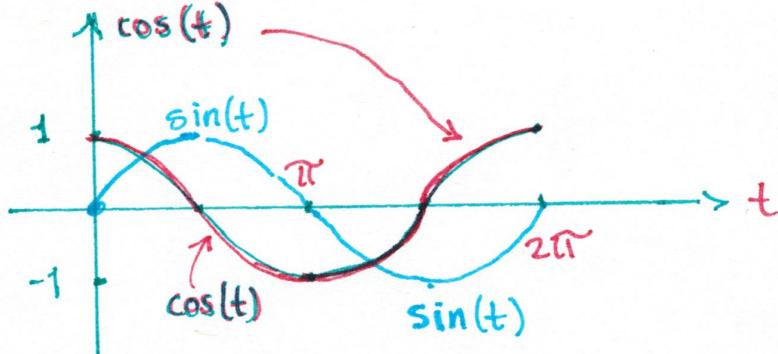
$$a_{\frac{N}{2}+1} = 0 = b_{\frac{N}{2}+1}; a_{\frac{N}{2}+2} = 0 = b_{\frac{N}{2}+2}; \dots$$

Sobre el intervalo $[0, 2\pi]$:

En la serie de Fourier (A), para valores "pequeños" de n , los coeficientes a_n y b_n están asociados a funciones que "no oscilan mucho"

P. ej. $n=1$ a_1 re-escalada el valor de $\cos(t)$

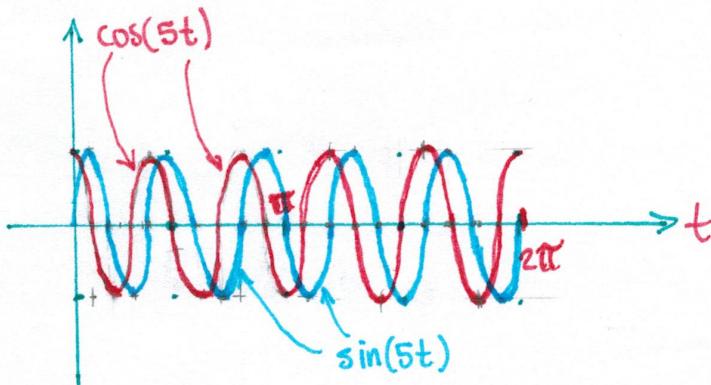
"Frecuencias Bajas"



$n=1$ b_1 re-escalada el valor de $\sin(t)$.

Pero si $n=5$ a_5 re-escalada el valor de $\cos(5t)$

"Frecuencias altas"

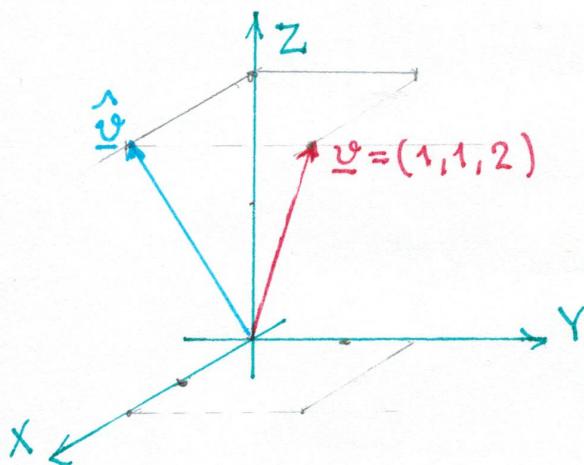


b_5 re-escalada el valor de $\sin(5t)$

Si n no es muy "pequeña" a_n y b_n están asociados a funciones que "oscilan más" que cuando n es "pequeña".

En análisis de Fourier se dice que los primeros términos en (A) están asociados a "frecuencias bajas" y mientras más grande sea n , el correspondiente término se asocia a una "frecuencia más alta". (1)

Ahora, recordemos el concepto de proyección ortogonal en espacios vectoriales de dimensión finita



$$\mathbb{H} = \mathbb{R}^3$$

$$\mathcal{V} = \{(x, y, z) \in \mathbb{R}^3 : \exists \alpha, \beta \in \mathbb{R} \text{ con } (x, y, z) = \alpha(1, 0, 0) + \beta(0, 0, 1)\}$$

$$\mathcal{V} = \{(x, y, z) \in \mathbb{R}^3 : (x, y, z) = (\alpha, 0, \beta); \alpha, \beta \in \mathbb{R}\}$$

- (1) En Física se define la frecuencia como el número de veces que se repite un ciclo en una unidad de tiempo. Las funciones seno y coseno repiten un ciclo en el intervalo $[0, 2\pi]$

Para calcular la proyección \hat{v} del vector v en el subespacio V_1 , necesitamos encontrar α' y β' en \mathbb{R} tales que

$$\hat{v} = \alpha'(1,0,0) + \beta'(0,0,1). \quad \dots (B)$$

Debido a que $\{(1,0,0), (0,0,1)\}$ es una base ortonormal para V_1 , podemos calcular

$$\langle v, (1,0,0) \rangle = 1 \cdot 1 + 1 \cdot 0 + 2 \cdot 0 = 1 = \alpha',$$

$$\langle v, (0,0,1) \rangle = 1 \cdot 0 + 1 \cdot 0 + 2 \cdot 1 = 2 = \beta'.$$

Entonces, por (B) $\hat{v} = (1,0,2)$.

Los coeficientes α' y β' en (B) representan re-escalamientos de $(1,0,0)$ y de $(0,0,1)$ necesarios para obtener \hat{v} .

En general, si H es un espacio vectorial de dimensión finita y $\{u_1, \dots, u_n\}$ es una base ortonormal de H , entonces para cada $v \in H$ existen $\alpha_1, \dots, \alpha_n$ en \mathbb{R} t.q.

$$v = \sum_{i=1}^n \alpha_i u_i \quad \dots \quad (B')$$

Los coeficientes a_1, \dots, a_n en (B') representan re-escalamientos de $\underline{u}_1, \dots, \underline{u}_n$ necesarios para obtener \underline{v} .

En análisis de Fourier, la representación (A) de una función $g(t)$ se puede entender en forma similar a lo dicho arriba para combinaciones lineales de vectores ortonormales que generen a un espacio vectorial.

Para cada n :

- a_n y b_n son los re-escalamientos de $\cos(nt)$ y $\sin(nt)$ necesarios para obtener $\underline{g}(t)$
- Si la magnitud de a_n es "grande", entonces la función $\cos(nt)$ es importante para describir a $g(t)$ a través de (A) , (análogamente si b_n tiene magnitud grande, $\sin(nt)$ es importante para describir $g(t)$ a través de (A)).
- Como consecuencia del punto anterior

Si la magnitud de a_n es "grande"⁽¹⁾ y n es pequeña, la función $g(t)$ en (A) tiene un comportamiento basado en frecuencias bajas (También si b_n es "grande").

Si la magnitud de a_n es "grande" y n es grande, la función $g(t)$ en (A) tiene un comportamiento basado en frecuencias altas (También si la magnitud de b_n es grande).

Como resultado de estas observaciones, si usamos $f_i(t)$ en la página 35 para representar a $x_{G_i} = (x_{i1}, \dots, x_{ip})$, las primeras componentes de este vector estarán asociadas con frecuencias bajas y las últimas componentes con frecuencias altas en la gráfica de $f_i(t)$.

Notemos que si permutamos las componentes de x_{G_i} para representar a $x_{G'_i} = (x_{i\pi_1}, x_{i\pi_2}, \dots, x_{i\pi_p})$

(1) La magnitud de a_n ó b_n

a través de $f_i(t)$, esta función tendrá un aspecto diferente a la $f_i(t)$ obtenida de x_{i1} , en otras palabras, la representación de x_{i1} como una curva de Andrews depende del orden en que aparecen las componentes de x_{i1} .

Si $x_{i1} = (x_{i11}, x_{i12}, x_{i13})$ y $x'_{i1} = (x_{i13}, x_{i11}, x_{i12})$

las funciones

$$f_i(t) = \frac{x_{i11}}{\sqrt{2}} + x_{i12} \sin(t) + x_{i13} \cos(t)$$

y

$$f'_i(t) = \frac{x_{i13}}{\sqrt{2}} + x_{i11} \sin(t) + x_{i12} \cos(t)$$

pueden ser muy diferentes.

Ejemplo: Consideremos los datos de mediciones de los billetes del banco Suizo.

En particular, nos concentraremos en las observaciones $x_{96}, x_{97}, \dots, x_{105}$.

La figura J muestra una gráfica de las curvas de Andrews correspondientes. Sabemos que $x_{96}, x_{97}, \dots, x_{100}$ corresponden

```

rm(list = ls(all = TRUE))
graphics.off()

#install.packages("tourr")
library(tourr)

# setwd("C:/...")      # set working directory if windows
# setwd("~/Users/...") # set working directory if mac

data = read.table("SwissBank 1.txt")

x = data[96:105,]
y = NULL
i = 1

while(i <= 6){
z = (x[, i] - min(x[, i])) / (max(x[, i]) - min(x[, i])) # zero-one scaling
y = cbind(y, z)
i = i + 1
}

Type = c(rep(1, 5), rep(2, 5))
f = as.integer(Type)
#grid = seq(0, 2 * pi, length = 1000)
grid = seq(0,1,length=1000)

#plot(grid, 2*pi*andrews(y[1,])(2*pi*grid), type = "l", lwd = 1.5, main = "Andrews curves (Bank data)",
# axes = FALSE, frame = TRUE, ylim = c(-0.3, 0.5), ylab = "", xlab = "")

plot(grid, 2*pi*andrews(y[1,])(2*pi*grid), type = "l", lwd = 1.5, main = "Andrews curves (Bank data)",
axes = FALSE, frame = TRUE, ylim = c(-2, 3), ylab = "", xlab = "")

for (i in 2:5){
lines(grid, 2*pi*andrews(y[i,])(2*pi*grid), col = "black", lwd = 1.5)
}
for (i in 6:10){
lines(grid, 2*pi*andrews(y[i,])(2*pi*grid), col = "red3", lwd = 1.5, lty = "dotted")
}
#axis(side = 2, at = seq(-.5, .5, .25), labels = seq(-.5, .5, .25))
#axis(side = 1, at = seq(0, 7, 1), labels = seq(0, 7, 1))

axis(side = 2, at = seq(-2, 3, 1), labels = seq(-2, 3, 1))
axis(side = 1, at = seq(0, 1, 0.2), labels = seq(0, 1, 0.2))

```

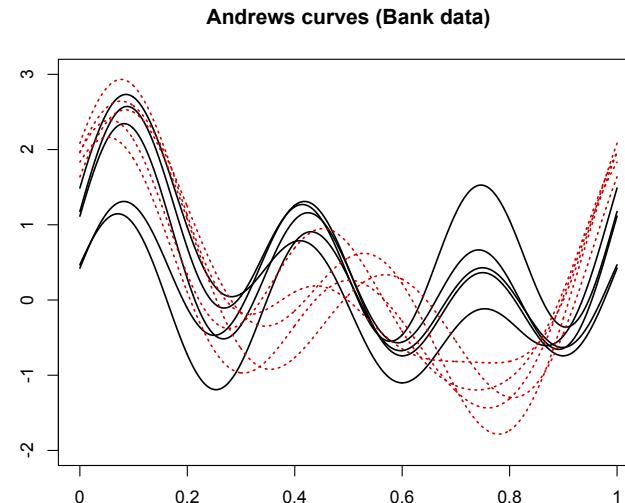


Figura J

a billetes verdaderos y $x_6^{101}, \dots, x_6^{105}$ corresponden a billetes falsos. En la figura J podemos ver que las curvas con linea punteada en color rojo guardan un comportamiento que las distingue de las demás, no obstante no es muy evidente la separación en dos grupos. Para esta figura el color negro corresponde a billetes verdaderos y el rojo corresponde a billetes falsos.

Como sabemos que la componente X_6 contiene información para separar dos grupos en los datos, procedemos a graficar las curvas de Andrews de los vectores

$$x_i^1 = (X_6, X_5, X_4, X_3, X_2, X_1)$$

La figura K muestra $f_{96}^1(t), \dots, f_{105}^1(t)$

Apreciación subjetiva: La segunda gráfica resulta más difícil de interpretar. En la primer gráfica X_6 está asociada a frecuencias altas, según parece, esto se refleja en el hecho de

```

# -----
# Description: MVAandcur2 computes Andrew's Curves for the observations
#             96-105 of the Swiss bank notes data (bank2.dat).
#             Here we changed the order of the variables.
# -----
# Output: Plot of Andrew's Curves for the observations
#         96-105 of the Swiss bank notes data.
# -----

rm(list=ls(all=TRUE))
graphics.off()

#Install.packages("tourr")
library(tourr)
install.packages("matlab")
library(matlab)

# Load data
# The data file should be located in the same folder as this Qlet
# Set the R working directory to this directory using setwd()
# setwd("C:/...")           # set working directory if windows
# setwd("/Users/...")        # set working directory if mac
data = read.table("SwissBank 1.txt")

x = data[96:105,]
y = NULL
i = 1

while(i <= 6){
  z = (x[,i]-min(x[,i]))/(max(x[,i])-min(x[,i]))
  y = cbind(y,z)
  i = i+1}

y = flipr(y) # change the order

Type = c(rep(1.5), rep(2.5))
f = as.integer(Type)
#grid <- seq(0, 2*pi, length = 1000)
grid = seq(0,1,length=1000)

plot(grid, 2*pi*andrews(y[1, ])(2*pi*grid), type = "l", lwd = 1.5, main = "Andrews curves (Bank data)",
      axes = FALSE, frame = TRUE, ylim = c(-2, 3), ylab = "", xlab = "")

for (i in 2:5){
  lines(grid, 2*pi*andrews(y[i,])(2*pi*grid), col="black",lwd=1.5)
}
for (i in 6:10){
  lines(grid, 2*pi*andrews(y[i,])(2*pi*grid), col="red3",lwd=1.5,lty="dotted")
}

axis(side = 2, at = seq(-2, 3, 1), labels = seq(-2, 3, 1))
axis(side = 1, at = seq(0, 1, 0.2), labels = seq(0, 1, 0.2))

```

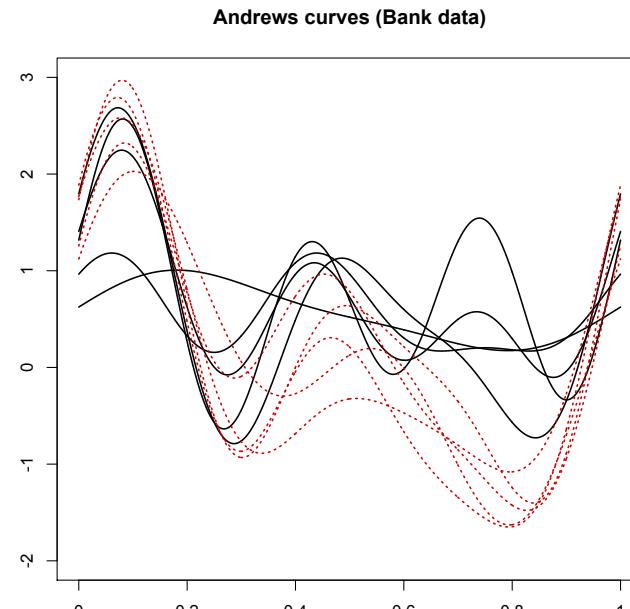


Figura K

de que las curvas negras oscilan en forma similar (seis cambios de concavidad) pero en forma diferente a la oscilación que presenten las curvas rojas. (4 cambios de concavidad). En la segunda figura X₆ esta asociada al término constante, hay dos curvas negras que ya no tienen la misma oscilación que las otras (curvas negras). Si en este gráfico no se hubieran asignado colores, sería muy difícil proponer grupos.

Para este tipo de gráficos (las curvas de Andrews) se espera que un outlier en alguna componente de $\mathbf{x} = (x_1, \dots, x_p)$ se refleje en que una curva tendrá un comportamiento muy diferente al resto de las curvas. Si hay subconjuntos de curvas que son similares, estos agrupamientos se pueden usar para hacer grupos en los datos x₁, ..., x_n. Se sugiere usar la metodología conocida como "Componentes Principales", para encontrar un orden adecuado de las componentes de \mathbf{x} y luego

usar las curvas de Andrews.

Gráficas de coordenadas Paralelas (PCP)

- Para este tipo de gráficas primero se re-escalan todas las variables x_{i1}, \dots, x_{ip} ; $i=1, \dots, n$ para que estén en el intervalo $[0,1]$. Por ejemplo, para la variable j ; $j \in \{1, 2, \dots, p\}$

$$x_{i0j} = \min\{x_{ij} : i=1, 2, \dots, n\}$$

$$x_{i1j} = \max\{x_{ij} : i=1, 2, \dots, n\}$$

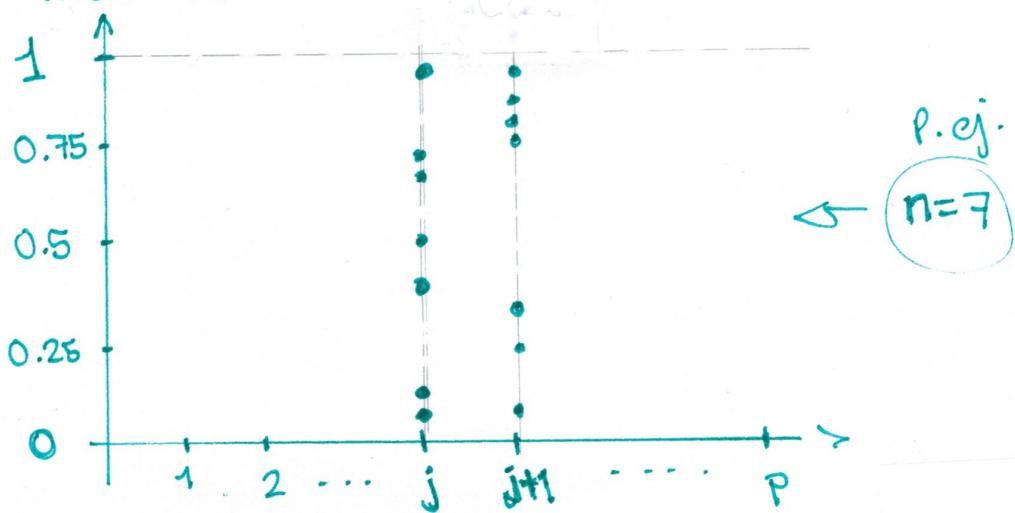
$$g_j(x) \equiv \frac{x - x_{i0j}}{x_{i1j} - x_{i0j}} \quad \text{es tal que:}$$

- 1) Para la variable j , $\tilde{x}_{ij} \equiv g_j(x_{ij}) \in [0,1]$.
- 2) $g_j(x)$ monótona no decrec.
- 3) Lineal. $\tilde{x}_{ij} \equiv g_j(x_{ij})$

- Una vez re-escaladas las variables, sobre el eje horizontal se dibujen los índices de las variables (se coloca una escala con los índices) $j=1, 2, \dots, P$

- Para cada valor de j se grafican los n puntos (pares ordenados) (j, \tilde{x}_{ij}) $i=1, 2, \dots, n$.

En el plano, estos puntos yacen sobre una linea horizontal



- Los puntos $(1, \tilde{x}_{i1}), (2, \tilde{x}_{i2}), \dots, (p, \tilde{x}_{ip})$, correspondientes al individuo i , se conectan con segmentos de linea recta

La figura L muestra un gráfico PCP para los datos de los billetes del banco Suizo, en específico los individuos $x_{96}, x_{97}, \dots, x_{105}$

```

# -----
# Book:      MVA
# -----
# Quantlet:  MVAparrcoo1
# -----
# Description: MVAparrcoo1 computes a parallel coordinate plot for the
#               observations 96-105 of the Swiss bank notes data
#               (bank2.dat).
# -----
# -----
# Output:    Parallel coordinate plot for the observations 96-105 of
#               the Swiss bank notes data (bank2.dat).
# -----


rm(list=ls(all=TRUE))
graphics.off()

install.packages("MASS")
library(MASS)

# Load data
# The data file should be located in the same folder as this Qlet
# Set the R working directory to this directory using setwd()
# setwd("C:/...")      # set working directory if windows
# setwd("~/Users/...") # set working directory if mac
data = read.table("SwissBank 1.txt")
x = data[96:105,]
ir = rbind(x[,1], x[,2], x[,3], x[,4], x[,4], x[,6])
parcoord(log(ir[, c(1, 2, 3, 4, 5, 6)]), lwd =2,
        col = c(1,1,1,1,2,2,2,2), lty=c(rep(1,5),rep(4,5)),
        main="Parallel coordinates plot (Bank data)", frame=TRUE, ablines=FALSE)
axis(side=2, at=seq(0,1,0.2), labels=seq(0,1,0.2))

```

Parallel coordinates plot (Bank data)

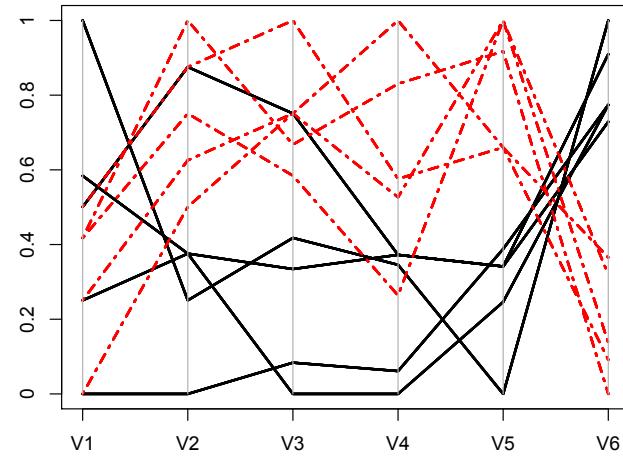


Figura L