UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES ACATLÁN

"Introducción al Análisis Multivariado Con *R*"

Tesina

QUE PARA OBTENER EL TÍTULO DE

Actuaria

PRESENTA

MARTIN NAVARRETE ROCIO GABRIELA

Asesor: Act. Mahil Herrera Maldonado

OCTUBRE 2012

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

AGRADECIMIENTOS

Para mí esta tesina es la culminación de una de las etapas más importantes en mi vida la universitaria, por eso agradezco en primer lugar a Dios por darme la oportunidad de haber estudiado, darme una familia, amigos y un trabajo.

Mención especial **a mi mamá, a mi papá y a mi hermana** que han estado ahí en todo momento, también de parte de mi familia materna a mi abuela María Luisa y a mis tíos que ayudaron a pagar la educación, en especial a mi tío Miguel por siempre preocuparse por mí.

Es imposible dejar de mencionar a mis amigos Angy, César, Diana, Erika, Eduardo, Nash, y Sandy como olvidar todas las veces que estudiamos, cuando jugamos UNO, el cómo me hacían reír en los momentos de estrés y muchos otros buenos recuerdos .A Juan José le agradezco por la paciencia y el apoyo que me ha dado durante el tiempo que he realizado la tesina.

Le doy las gracias a la UNAM por abrirme sus puertas en la FES Acatlán principalmente a la coordinación de Actuaría los cuales siempre se han preocupado por mejorar la carrera y por sus alumnos, especialmente a Mahil por dirigir esta tesina y a Víctor por creer en mí.

Quien no podía faltar es Angelita que durante la carrera siempre me apoyo.

A todos los que me dieron un consejo ya sea académico o personal y confiaron en mí en el transcurso de lo que empezó siendo un sueño y ahora es una realidad, MUCHAS GRACIAS!

Contenido

INTRODUCCIÓN

Actualmente, en las investigaciones es indispensable contar con una gran cantidad de información por lo que es necesario relacionarla ó reducirla, para ello son utilizados los métodos de Análisis Multivariado ya que ayudan al estudio de las características o atributos de los individuos u objetos.

Este método surgió en la Psicología a principios del siglo XX, pero con el tiempo ha sido utilizado en otros campos como: investigación de mercados, medicina, economía, estudio del entorno ambiental, estudio social, turismo, etc. Es por ser un análisis muy aplicable en varias áreas que se eligió para esta tesina.

Una herramienta eficaz para llevar a cabo este tipo de análisis es ***R***, éste es un lenguaje y entorno de programación para el análisis estadístico y gráfico. Es un ***software libre***, resultado de la implementación GNU del lenguaje S y S-Plus -versión comercial, el cual es desarrollado y mantenido por algunos de los más prestigiosos estadísticos, cuenta además, con la ventaja de ser gratuito y de fácil descarga e instalación sencilla por lo que es fácil de distribuir y hacer una formación en comunidad. Por las razones mencionadas anteriormente se eligió emplear ***R*** para esta tesina.

El presente trabajo tiene como objetivo introducir y guiar al lector a la aplicación del Análisis Multivariado en ***R*** y con ello complementar la tesina desarrollada previamente por la Actuaria Fabiola López González, quien proporciona una introducción a ***R*** y su aplicación en algunos temas estadísticos.

La tesina está dirigido principalmente a los alumnos de últimos semestres de la licenciatura de Actuaría u otras personas que se inician en el estudio del Análisis Multivariado y que tienen conocimientos básicos de ***R***.

El motivo para realizar este trabajo es que no existe una tesina en la carrera de Actuaría dedicada a ***R*** enfocada al Análisis Multivariado; además de hacer una compilación de las funciones de ***R*** del Análisis Multivariado lo cual es más fácil para las personas que lo llegaran a utilizar día a día.

Una de las aportaciones del trabajo es la descripción las funciones las cuales se encuentran en inglés, al igual que sus explicaciones cuando se llegan a investigar.

Otra razón importante por la que se realizó esta tesina es para que los alumnos que se especializan en el área de estadística que no pudieron tomar esta materia, tengan una noción de los métodos y su aplicación para complementar su conocimiento en dicha rama ya que es una materia muy práctica y les puede ayudar en el mundo laboral.

La tesina consta de cinco capítulos, estos temas fueron tomados del temario de la materia a excepción de componentes principales, es un tema que se incluyó al ser un método que presenta ***R*** y otros paquetes estadísticos.

El primer capítulo abarca el tema de "Componentes Principales" con sus pasos a seguir en *R* desde ver la matriz de correlación, su cálculo, el peso de cada uno, hasta la gráfica.

El apartado siguiente "Análisis Factorial" muestra a semejanza del anterior, los pasos para realizarlo como son: la revisión de la matriz de correlación para saber si tiene sentido aplicar la técnica, la extracción, determinación y rotación de factores con sus diferentes métodos.

El tercer segmento describe el "Análisis Discriminante" con sus diferentes reglas de decisión para conocer a que grupo pertenece el nuevo individuo y el comportamiento de éstas cuando la población tiene una distribución normal.

En la cuarta parte se desarrolla el tema de "Análisis de Conglomerados" inicia con la formulación del problema, continua con la selección de las medidas de similitud, para prosigue con clasificación de cúmulos y sus métodos para llegar a la elección de grupos y finaliza con la determinación de la validez y confianza.

Por último, en la sección cinco se revisa el "Escalamiento Multidimensional", su modelo y su par de métodos métrico y no métrico, el cual ayuda a visualizar las rutas más cortas.

Los capítulos constan de una parte teórica y otra práctica, en está última se exponen las instrucciones necesarias para desarrollar el método en ***R***, también se exponen una breve explicación de éstas y de los resultados.

Para hacer la práctica en ***R*** se emplearon dos bases de datos; la utilizada en los primeros cuatro capítulos fue construida a partir de los valores de las acciones de unas empresas participantes en la BMV del mes de agosto del 2011 y la segunda con distancias en km entre varios municipios del estado de México (ver anexo).

Es importante mencionar que para ejecutar los ejercicios es necesario saber importar datos de Excel a ***R*** *y* cargar librerías, para conocer más sobre el tema se sugiere consultar la tesina*: Lenguaje R: Un complemento libre para las asignaturas de estadísticas".*[1]

[1] López Gonzaléz, Fabiola, *Lenguaje R: Un complemento libre para las asignaturas de estadística*, México, UNAM,2008.

1. COMPONENTES PRINCIPALES

> El análisis de componentes principales es un procedimiento matemático que transforma un conjunto de variables posiblemente correlacionadas en un conjunto menor de variables no correlacionadas, llamadas *componentes principales* donde la componente principal explica la mayor variabilidad posible de los datos y cada componente subsecuente explica la mayor variabilidad posible restante no explicada por las componentes anteriores.[2]

Los componentes principales tienen como objetivo reducir la dimensionalidad de un conjunto de datos e interpretarlos, el cual ayuda a evitar redundancias en la información y destacar relaciones. Éstos pueden construir variables no observables a partir de variables observables.

A continuación se explicará el modelo.

1.1 MODELO

Se consideran una serie de variables x_i donde $i \epsilon (1, \dots . , p)$ sobre un grupo de objetos o individuos y se trata de calcular, a partir de ellas, un nuevo conjunto de variables $y_1, y_2, \dots, y_p$ incorrelacionadas entre sí, cuyas varianzas vayan decreciendo progresivamente.

Donde cada y_j $(j = 1,2, \dots p)$ es una combinación lineal de las $x_i{}'s$ originales, es decir:

$$y_j = a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jp}x_p = a_j{}'X$$

siendo $a'_j = (a_{1j}, a_{2j}, \dots, a_{pj})$un vector de constantes, y

[2] Luis Enrique Nieto Barajas**.** *Módulo6:Análisis multivariado*,pág 30,pdf http://allman.rhon.itam.mx/~lnieto/index_archivos/Modulo61.pdf. Revisado el 13 de mayo de 2011

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

Como lo que se quiere es la variabilidad entre los datos, por ende se busca maximizar la varianza; una forma simple podría ser aumentar los coeficientes a_{1j}, por ello, para mantener la ortogonalidad de la transformación se impone que el módulo del vector $a'_j = (a_{1j}, a_{2j}, \dots, a_{pj})$ sea 1. Es decir:

$$a'_j a_j = \sum_{k=1}^{p} a_{kj}^2 = 1$$

El primer componente se calcula eligiendo a_1 de modo que y_1 tenga la mayor varianza posible, sujeta a la restricción de que $a'_j a_j = 1$. El segundo componente principal se calcula obteniendo a_2 de modo que la variable y_2 obtenida, esté incorrelacionada con y_1, del misma forma se eligen $y_1, y_2, \dots, y_p$, incorrelaciondas entre sí, de manera que las variables aleatorias obtenidas tengan cada vez menor varianza. A continuación se explicará con mayor detalle el cálculo de los componentes.

a) Cálculo del primer componente principal

El primer componente principal se define "*como la combinación lineal de las variables originales que tienen varianza máxima*".[3] Los valores en este primer componente de los *n* individuos se representará por un vector y_1, dado por:

$$y_1 = a'_1 X$$

se quiere elegir a_1 de modo que se maximice la varianza de y_1 sujeta a la restricción

[3]Daniel Peña, "Componentes Principales" en *Análisis de datos multivariantes*, Madrid, McGraw-Hill/Interamericana de España, c2002,pág.137

$$a'_1 a_1 = 1$$

$$Var(y_1) = Var(a'_1 X) = a'_1 S a_1$$

donde S es la matriz de covarianzas.

El método habitual para maximizar una función de varias variables sujeta a restricción es el método de los multiplicadores de Lagrange. El problema consiste en maximizar la función $Var(y_1)$sujeta a la restricción $a'_1 a_1 = 1$ donde la incógnita es a_1.

Así se construye la función L:

$$L(a_1) = a'_1 S a_1 - \lambda(a'_1 a_1 - 1)$$

buscando el máximo, derivando e igualando a 0:

$$\frac{\partial L(a_1)}{\partial a_1} = 2Sa_1 - 2\lambda I a_1 = 0 \rightarrow (S - \lambda I)a_1 = 0$$

La ecuación anterior es un sistema lineal de ecuaciones, para que el sistema tenga una solución distinta de 0, la matriz $S - \lambda I$ tiene que ser singular, es decir, el determinante debe ser igual a cero:

$$|S - \lambda I| = 0$$

Lo que implica que λ es valor propio de S. La matriz de covarianzas S es de orden p y sí además es definida positiva, tendrá p autovalores o eigenvalores distintos, $\lambda_1, \lambda_2, \ldots, \lambda_p$ tales que$\lambda_1 > \lambda_2 > \cdots > \lambda_p$.

Desarrollando la expresión anterior:

$$(S - \lambda I)a_1 = 0$$

$$Sa_1 - \lambda I a_1 = 0$$

$$Sa_1 = \lambda I a_1$$

Sustituyendo en la varianza:

$$Var(y_1) = a'_1 S a_1 = a'_1 \lambda I a_1 = \lambda a'_1 a_1 = \lambda * 1 = 1\lambda$$

Luego se toma el mayor valor propio para maximizar la varianza y su correspondiente vector propio a_1.

b) Cálculo del segundo componente principal

El segundo componente principal se calcula de forma parecida, además se requiere que la $Cov(y_2, y_1) = 0$, es decir:

$$Cov(y_2, y_1) = Cov(a'_2 x, a'_1 x) = a'_2 E[(x - \mu)(x - \mu)']a'_1 = a'_2 S a_1 = 0$$

Como $Sa_1 = \lambda I a_1$ entonces $a'_2 S a_1 = a'_2 \lambda I a_1 = \lambda a'_2 a_1 = 0$ lo que equivale a que $a'_2 a_1 = 0$, esto significa que los vectores son ortogonales.

De este modo se tendrá que maximizar la varianza de y_2 con las siguientes restricciones:

- $a'_2 a_2 = 1$
- $a'_2 a_1 = 0$

Toma la función:

$$L(a_2) = a'_2 S a_2 - \lambda(a'_2 a_2 - 1) - \delta a'_2 a_1$$

Se deriva:

$$\frac{\partial L(a_2)}{\partial a_2} = 2Sa_2 - 2\lambda a_2 - \delta a_1 = 0$$

Multiplicando por a'_1 :

$$2a'_1 S a_2 - \delta = 0$$

Luego:

$$\delta = 2a'_1 S a_2 = 2a'_2 S a_1 = 0$$

De este modo:

$$\frac{\partial L(a_2)}{\partial a_2} = 2Sa_2 - 2\lambda a_2 - \delta a_1 = 2Sa_2 - 2\lambda a_2 = 0$$

$$(S - \lambda I)a_2 = 0$$

Usando los mismos razonamientos que antes, elegimos λ_2 como el segundo mayor valor propio de S con su vector propio asociado a_2.

Generalizando:

Los razonamientos anteriores se pueden extender, de modo que al *j* -ésimo componente le correspondería el *j* –ésimo autovalor.

En general, la matriz X (y por tanto la S) tiene rango *p*, existiendo entonces tantas componentes principales como variables que se obtendrán calculando los valores propios $\lambda_1, \lambda_2, \ldots, \lambda_p$, de la matriz de varianzas y covarianzas de las variables *S*, mediante:

$$|S - \lambda I| = 0$$

Y sus vectores asociados son:

$$(S - \lambda_i I) a_i = 0$$

Los términos λ_i son reales, al ser la matriz S simétrica, y positivos, ya que *S* es definida positiva. Por ser S simétrica sí λ_i y λ_h son dos raíces distintas sus vectores asociados son ortogonales.

Nombrando Y a la matriz cuyas columnas son los valores de los *p* componentes en los n individuos, estas nuevas variables están relacionadas con las originales mediante:

$$Y = XA$$

Donde $A'A = I$. Así al calcular los componentes principales equivale a aplicar una transformación ortogonal A a las variables X para obtener unas nuevas variables Y incorrelacionadas entre sí.

1.2 PROPIEDADES DE LOS COMPONENTES[4]

Los componentes principales son nuevas variables con las siguientes propiedades:

1. Conservan la variabilidad inicial: la suma de las varianzas de los componentes es igual a la suma de las varianzas de las variables originales, y la varianza generalizada de los componentes es igual a la original.

Como $Var(y_1) = y$ y la suma de los valores propios es la traza de la matriz:

$$tr(S) = Var(x_1) + \cdots + Var(x_p) = \lambda_1 + \lambda_2 + \cdots + \lambda_p$$

Por lo tanto:

$$\sum_{i=1}^{p} Var(x_i) = \sum \lambda_i = \sum_{i=1}^{p} Var(y_i)$$

Las nuevas variables tienen conjuntamente la misma variabilidad que las variables originales. Los componentes principales también conservan la varianza generalizada, (determinante de la matriz de covarianzas de las variables). Como el determinante es el producto de los valores propios, llamando S_y a la matriz de covarianzas de los componentes, que es la diagonal con términos λ_i.

$$|S_x| = \lambda_1 \ldots . \lambda_p = \prod_{I=1}^{p} Var(y_i) = |S_y|$$

2. La proporción de variabilidad explicada por un componente es el cociente entre su varianza y la suma de los valores propios de la matriz, por ejemplo para el componente *j* es $\frac{\lambda_j}{\sum \lambda_i}$.

[4]Daniel Peña, "Componentes Principales" en *Análisis de datos multivariantes*, Madrid , McGraw-Hill/Interamericana de España, c2002,pág.145.

3. Las covarianzas entre cada componente principal y las variables de la columna de X vienen dadas por el producto de las coordenadas del vector propio y el valor propio asociado, esto es:

$$Cov(y_i, x_1 \ldots . x_p) = \lambda_i a_i = \lambda_i (a_{i1} \ldots . a_{ip})$$

Donde a_i es el i-ésimo vector propio que define al i-ésimo componente.

4. El coeficiente de correlación lineal entre la *i-ésima* componente y la *j-ésima* variable de la columna X, $\rho(y_i, x_j)$ está dado por:

$$\rho(y_i, x_j) = \frac{Cov(y_i, x_j)}{\sqrt{Var(y_i)Var(x_j)}} = \frac{\lambda_i a_{ij}}{\sqrt{\lambda_i s_j^2}} = a_{ij}\frac{\sqrt{\lambda_i}}{s_j}$$

5. Los r componentes principales $(r < p)$ proporcionan la predicción lineal óptima con r variables del conjunto de valores de variables X.

6. Si estandarizamos los componentes principales, dividiendo cada uno por su desviación típica, se obtiene la estandarización multivariante de los datos originales.

1.3 CAMBIOS DE ESCALAS

Sí las variables originales $(x_1, \ldots . , x_p)$ están incorrelacionadas, entonces carece de sentido calcular los componentes principales. Sí se hiciera, se obtendrían las mismas variables pero reordenadas de mayor a menor varianza. Para saber si $(x_1, \ldots . , x_p)$ están correlacionadas, se puede calcular la matriz de correlaciones.

El cálculo de los componentes principales de una serie de variables $(x_1, \ldots, x_p)$ depende normalmente de las unidades de medida empleadas. Si transformamos las unidades de medida, lo más probable es que cambien a su vez los componentes obtenidos.
Una solución frecuente es usar variables$(x_1, \ldots, x_p)$ tipificadas, con ello, se eliminan las diferentes unidades de medida y se consideran todas las variables implícitamente equivalentes en cuanto a la información recogida.

1.4 IDENTIFICACIÓN DE LOS COMPONENTES PRINCIPALES

Uno de los objetivos del cálculo de componentes principales es la identificación de los mismos, sin embargo este es un problema difícil. Habitualmente se conservan sólo aquellos componentes que recogen la mayor parte de la variabilidad, hecho que permite representar los datos según dos o tres dimensiones si se conservan dos o tres ejes factoriales, pudiéndose identificar entonces grupos naturales entre las observaciones.

1.5 INTERPRETACIÓN DE LOS COMPONENTES

La interpretación de los componentes debe de ser deducida tras observar la relación de los componentes con las variables iniciales (se tiene que estudiar tanto el signo como la magnitud de las correlaciones). Para que un componente sea fácilmente interpretable debe tener las siguientes características:

- Los coeficientes de cada componente deben ser próximos a 1.
- Una variable debe tener coeficientes elevados sólo en un componente.
- No deben existir componentes con coeficientes similares.

Una vez calculados los componentes principales, también se calcula la matriz de correlación entre las variables y los componentes, para ver que variables explican más que otras. Cuando existe una alta correlación entre las variables, en la matriz de correlación de

las variables y componentes, el primer componente tendrá todas las coordenadas del mismo signo (por lo general positivo) y se puede interpretar como un promedio ponderado de todas las variables, a dicha componente se le conoce como *componente de tamaño*. El resto de los componentes se interpretan como *componentes de forma* y por lo general tienen coordenadas de ambos signos, los cuales implican que contrastan unos grupos de variables frente a otros.

1.6 SELECCIÓN DEL NÚMERO DE COMPONENTES

Existen distintos métodos[5] para seleccionar el número de componentes:

1-Realizar un grafico de λ_i frente a a_i el procedimiento consiste en buscar un punto a partir del cual los valores propios se nivelan de forma horizontal, este punto es llamado **codo**, de manera que conforme se va nivelando, si los valores propios son muy cercanos a cero se pueden ignorar. En conclusión el criterio es quedarse con un número de componentes que excluyan a los que se asocian con el valor correspondiente mas pequeño, es decir, el número de componentes es igual al número de valores propios antes de que se nivele la grafica.

2-Seleccionar los componentes hasta que se cubra una proporción determinada de varianza, por lo general se considera entre un 80% a 90 %.

3- Descartar aquellos componentes asociados a valores propios inferiores a una cota, que suele fijarse como la varianza media, $\frac{\sum \lambda_i}{p}$. Cuando se usa la matriz de correlación en lugar de la de covarianza para llevar a cabo el análisis de componentes principales, el valor medio de los componentes es 1 y esta regla lleva a seleccionar los valores propios mayores a uno l.

[5]Daniel Peña, "Componentes Principales" en *Análisis de datos multivariantes*, Madrid, McGraw-Hill/Interamericana de España, c2002,pág.14

1.7 EJEMPLO EN *R*

Para hacer el ejemplo es necesario importar la base de datos, en este caso se utilizará la base de acciones de algunas de las empresas de la BMV, para ello se puede emplear las instrucción ***read.table ( )***[6]en la cual se escribe el nombre del archivo y si tienen título las columnas. (véase ejemplo 1.1)

Para designar el nombre a la base de datos antes de escribir la instrucción ***read.table ( )*** se pone el nombre que se le quiere dar (en este caso datos) y se forma una flecha por un signo de menor y un guión.

```
> datos<-read.table("acciones2.txt",header=TRUE)
> datos
   axtel cemex   geo soriana compartamos gruma gfamsa   kfol bachoco  cmr liverpool maseca
1   6.00  6.69 23.71   30.98       74.00 23.17  13.62 121.21   22.77 3.05     98.00  13.95
2   5.93  6.62 23.70   30.01       74.01 23.11  13.20 120.23   22.50 3.05     98.00  12.62
3   5.80  6.34 23.25   30.00       80.25 22.23  13.25 117.11   22.10 3.05     94.99  12.62
4   5.69  6.37 23.56   29.70       73.20 21.55  13.13 119.82   21.85 3.05     94.00  12.62
5   5.87  6.57 24.02   30.94       81.00 22.89  13.80 120.39   21.79 3.05     93.99  12.62
6   5.82  6.40 23.85   30.89       80.00 22.98  13.50 119.44   21.93 3.20     94.00  12.65
7   5.60  6.13 23.42   29.01       73.07 23.00  12.85 116.80   22.60 3.06     94.45  12.70
8   5.58  6.28 23.63   28.46       73.07 22.00  12.81 115.72   22.70 3.05     93.00  13.95
9   5.66  6.29 24.32   28.20       73.05 22.54  13.01 117.55   23.00 3.05     93.00  13.95
10  5.75  6.60 24.50   28.25       73.00 22.25  13.91 118.42   22.80 2.80     92.20  12.57
11  5.66  7.07 25.50   28.94       70.00 22.00  14.45 117.98   22.75 2.90     95.01  14.00
12  5.40  6.66 24.90   28.61       68.05 22.62  14.65 113.17   21.60 3.20     94.99  13.96
13  5.23  6.49 25.35   28.75       68.02 23.00  14.50 112.99   21.39 3.20     94.53  12.80
14  5.30  6.52 26.52   29.19       73.00 23.00  14.40 111.94   21.30 3.05     94.85  13.96
15  5.14  6.26 25.30   28.06       65.00 22.21  13.38 108.96   21.00 3.00     92.29  13.92
16  5.10  6.38 24.74   28.20       70.00 21.38  13.00 107.40   21.08 3.95     92.28  13.90
17  5.27  6.37 23.86   29.00       60.55 20.80  12.80 104.61   21.20 3.00     94.00  13.90
18  5.80  6.80 22.11   30.51       72.50 22.10  12.93 108.25   22.79 3.00     94.00  13.85
19  6.00  7.01 21.53   30.39       69.38 22.67  13.26 110.41   23.00 3.20     95.22  13.93
20  6.55  7.30 22.30   31.77       77.20 24.32  14.09 112.90   23.00 3.20     99.32  13.94
21  6.60  7.55 22.55   32.15       78.00 25.00  14.47 114.52   23.20 3.35     96.00  13.29
22  6.90  8.00 23.38   33.00       77.50 25.20  14.86 115.55   23.05 3.35     94.46  13.79
23  6.85  8.40 24.00   32.55       77.20 25.00  15.60 114.22   23.00 3.35     95.00  13.94
```

Ejemplo1.1

[6]Para mayor información sobre esta instrucción se puede consultar la tesina Fabiola López Gonzaléz,"Lenguaje R",*Lenguaje R: Un complemento libre para las asignaturas de estadística*, México,UNAM,2008.pág 25

Para el análisis de componentes principales se siguen los siguientes pasos:

1-Como se menciona anteriormente carecería de sentido realizar el análisis de componentes principales sino existiera una correlación entre las variables, es por eso que el primer paso es ver la correlación entre estas, para ello se utiliza la función ***cor ( )***

```
> cor(datos)
                axtel       cemex         geo      soriana compartamos       gruma     gfamsa        kfol     bachoco         cmr   liverpool
axtel       1.00000000  0.85488764 -0.56651149  0.908604791  0.64373268  0.83818289 0.5060068  0.30261533  0.7698939  0.13411027  0.49788613
cemex       0.85488764  1.00000000 -0.27613534  0.791434549  0.33393851  0.78662217 0.7621601 -0.01803411  0.5588849  0.29045629  0.36084386
geo        -0.56651149 -0.27613534  1.00000000 -0.545665653 -0.31699482 -0.24248490 0.2787616  0.01278454 -0.6053741 -0.07089945 -0.35119424
soriana     0.90860479  0.79143455 -0.54566565  1.000000000  0.64060365  0.79391904 0.4819617  0.19579284  0.5254244  0.22672240  0.54577745
compartamos 0.64373268  0.33393851 -0.31699482  0.640603648  1.00000000  0.58034259 0.2437971  0.63806026  0.4708987  0.09641522  0.28313796
gruma       0.83818289  0.78662217 -0.24248490  0.793919040  0.58034259  1.00000000 0.7022099  0.24436477  0.5500603  0.24340246  0.49136969
gfamsa      0.50600683  0.76216014  0.27876164  0.481961736  0.24379708  0.70220989 1.0000000  0.10496577  0.1776476  0.16570893  0.25146221
kfol        0.30261533 -0.01803411  0.01278454  0.195792841  0.63806026  0.24436477 0.1049658  1.00000000  0.3870336 -0.33613220  0.24970203
bachoco     0.76989392  0.55888489 -0.60537413  0.525424381  0.47089870  0.55006027 0.1776476  0.38703365  1.0000000 -0.14612568  0.38149349
cmr         0.13411027  0.29045629 -0.07089945  0.226722401  0.09641522  0.24340246 0.1657089 -0.33613220 -0.1461257  1.00000000 -0.03071525
liverpool   0.49788613  0.36084386 -0.35119424  0.545777452  0.28313796  0.49136969 0.2514622  0.24970203  0.3814935 -0.03071525  1.00000000
maseca      0.03153371  0.30926372  0.06242908 -0.001027291 -0.40634606  0.03814902 0.1606697 -0.55034913  0.0863552  0.21911303  0.04691426
                maseca
axtel       0.031533707
cemex       0.309263722
geo         0.062429076
soriana    -0.001027291
compartamos -0.406346057
gruma       0.038149021
gfamsa      0.160669676
kfol       -0.550349133
bachoco     0.086355192
cmr         0.219113030
liverpool   0.046914260
maseca      1.000000000
```

Ejemplo1.2

Los números están cercanos al -1 y 1 por lo que si existe una correlación y si se puede usar el método de componentes principales.

2-La instrucción ***var( )*** nos da la varianza entre las variables, si las varianzas tienen tamaño semejante ó las variables están medidas en las mismas unidades ó en unidades comparables se utiliza la matriz de covarianza.En caso de que las variables no cumplan con alguna de estas cualidades, se utilizaría la matriz de correlación con lo cual las variables serían estandarizadas.

```
> var(datos)
                axtel       cemex         geo       soriana compartamos       gruma     gfamsa         kfol     bachoco          cmr   liverpool
axtel       0.259107510  0.25197510 -0.33755474  0.6907322134   1.6348676  0.48619565 0.20062233  0.70670553  0.28787609  0.01533004  0.45771186
cemex       0.251975099  0.33528735 -0.18716561  0.6844130435   0.9647441  0.51904743 0.34374605 -0.04790830  0.23771996  0.03776858  0.37735494
geo        -0.337554743 -0.18716561  1.37022213 -0.9539306324  -1.8513336 -0.32345395 0.25416245  0.06865751 -0.52053992 -0.01863715 -0.74244625
soriana     0.690732213  0.68441304 -0.95393063  2.2304339921   4.7733263  1.35114881 0.56064783  1.34152589  0.57642075  0.07603794  1.47208340
compartamos 1.634867589  0.96474407 -1.85133360  4.7733262846  24.8928174  3.29954269 0.94743281 14.60515889  1.72583478  0.10802490  2.55127490
gruma       0.486195652  0.51904743 -0.32345395  1.3511488142   3.2995427  1.29856759 0.62327846  1.27755158  0.46044447  0.06228715  1.01125988
gfamsa      0.200622332  0.34374605  0.25416245  0.5606478261   0.9474328  0.62327846 0.60668854  0.37509229  0.10164289  0.02898478  0.35373478
kfol        0.706705534 -0.04790830  0.06865751  1.3415258893  14.6051589  1.27755158 0.37509229 21.04820791  1.30434051 -0.34630514  2.06895850
bachoco     0.287876087  0.23771996 -0.52053992  0.5764207510   1.7258348  0.46044447 0.10164289  1.30434051  0.53959684 -0.02410474  0.50610889
cmr         0.015330040  0.03776858 -0.01863715  0.0760379447   0.1080249  0.06228715 0.02898478 -0.34630514 -0.02410474  0.05042925 -0.01245711
liverpool   0.457711858  0.37735494 -0.74244625  1.4720833992   2.5512749  1.01125988 0.35373478  2.06895850  0.50610889 -0.01245711  3.26170198
maseca      0.009897826  0.11042372  0.04506166 -0.0009460474  -1.2501377  0.02680652 0.07716877 -1.55693557  0.03911542  0.03034130  0.05224585
                maseca
axtel       0.0098978261
cemex       0.1104237154
geo         0.0450616601
soriana    -0.0009460474
compartamos -1.2501377470
gruma       0.0268065217
gfamsa      0.0771687747
kfol       -1.5569355731
bachoco     0.0391154150
cmr         0.0303413043
liverpool   0.0522458498
maseca      0.3802328063
```

Ejemplo 1.3

En este caso la varianza entre las variables es de tamaño semejante, por ende se empleará la matriz de covarianza.

3-Se hace el análisis de componentes

a) Se ven los componentes principales con ***princomp ( )***, la cual saca los eigeinvalores de la matriz de correlación.

```
> summary(princomp(datos))
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5      Comp.6      Comp.7      Comp.8       Comp.9      Comp.10      Comp.11
Standard deviation     6.1368718 3.0395254 1.91270669 1.20289918 1.01420803 0.598566374 0.469854392 0.345544098 0.2231821841 0.1783564517 0.1029386005
Proportion of Variance 0.6996759 0.1716384 0.06796723 0.02688198 0.01910983 0.006656217 0.004101374 0.002218246 0.0009253838 0.0005909905 0.0001968608
Cumulative Proportion  0.6996759 0.8713143 0.93928151 0.96616349 0.98527332 0.991929540 0.996030914 0.998249160 0.9991745442 0.9997655346 0.9999623955
                          Comp.12
Standard deviation     4.499026e-02
Proportion of Variance 3.760451e-05
Cumulative Proportion  1.000000e+00
```

Ejemplo 1.4

Nota: En caso de que se utilizará la matriz de correlación la instrucción sería ***(princomp(datos[,2:12]),cor=T)***

b) Para obtener la proporción explicada y acumulada utilizamos el comando ***Summary( )*** ésta produce resúmenes de los resultados de las diversas funciones, en este caso para la función de ***princomp ( ).***

```
> princomp(datos)
Call:
princomp(x = datos)

Standard deviations:
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
6.13687179 3.03952536 1.91270669 1.20289918 1.01420803 0.59856637 0.46985439 0.34554410 0.22318218 0.17835645
   Comp.11    Comp.12
0.10293860 0.04499026

 12  variables and  23 observations.
```

Ejemplo 1.5

Con esta información podemos seleccionar los componentes, como se había mencionado en el apartado teórico, por lo general se considera entre un 80% a 90 % de la proporción de la varianza.

c) Se observa el peso de los componentes con ***loadings( )***

```
> loadings(princomp(datos))

Loadings:
           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12
axtel                     0.106        -0.206                              0.189   0.657   0.671
cemex               0.103  0.123 -0.219 -0.225         0.150  0.311         0.438   0.374  -0.649
geo                -0.178 -0.167 -0.648  0.462         0.180 -0.114 -0.452  0.211
soriana     -0.130  0.299  0.309 -0.136 -0.354  0.607  0.362 -0.186 -0.226         -0.257
compartamos -0.744  0.516 -0.341         0.220
gruma               0.177  0.228 -0.415 -0.205 -0.181 -0.613 -0.511                       -0.109
gfamsa                     0.114 -0.533                        0.607  0.381 -0.340  -0.181   0.187
kfol        -0.631 -0.737  0.144        -0.155
bachoco                    0.135        -0.372 -0.589         0.224 -0.466  0.136  -0.427   0.117
cmr                                                        -0.120  0.509  0.729  -0.366   0.241
liverpool           0.104  0.790  0.174  0.557
maseca                           -0.109        -0.480  0.658 -0.384  0.324 -0.229

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000   1.000   1.000   1.000
Proportion Var  0.083  0.083  0.083  0.083  0.083  0.083  0.083  0.083  0.083   0.083   0.083   0.083
Cumulative Var  0.083  0.167  0.250  0.333  0.417  0.500  0.583  0.667  0.750   0.833   0.917   1.000
```

Ejemplo 1.6

d) Se grafican los eigenvalores, para ello se emplea la instrucción ***screeplot( )*** la cual muestra los componentes principales en el eje de las abcisas y los eigenvalores en el de las ordenadas.

```
> screeplot(princomp(datos))
```

Ejemplo 1.7

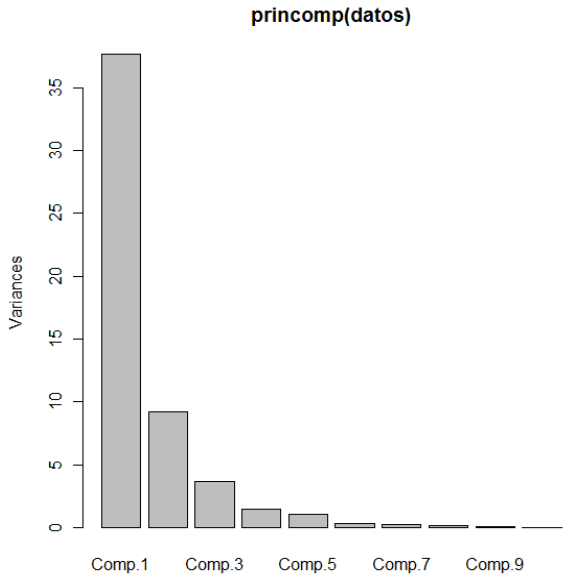

Figura 1.1 Gráfica de los eigenvalores

e) Se sacan las puntuaciones con **$*scores***, en este caso se tienen para los dos primeros componentes.

```
> princomp(datos) $scores[,1:2]
          Comp.1     Comp.2
 [1,] -5.2709006 -3.4619162
 [2,] -4.5601290 -3.1585326
 [3,] -6.8345135  1.9136815
 [4,] -3.0590521 -4.1114655
 [5,] -9.5224322  0.0900117
 [6,] -8.1799682  0.2801414
 [7,] -1.1780309 -1.8433486
 [8,] -0.1227393 -1.4916840
 [9,] -1.2795920 -2.9234166
[10,] -1.7768226 -3.7441666
[11,]  0.4822374 -4.5351064
[12,]  5.0148219 -1.9715364
[13,]  5.1337008 -1.9880441
[14,]  2.1104255  1.3695378
[15,] 10.4309698 -1.2052389
[16,]  7.7395112  2.5401032
[17,] 16.2664605  0.1658693
[18,]  4.5421136  4.8007591
[19,]  5.2841100  1.9739535
[20,] -2.8775311  5.2922127
[21,] -4.3313788  4.3855049
[22,] -4.5531773  3.4669474
[23,] -3.4580830  4.1557333
```

Ejemplo 1.8

f) Se gráfican los valoresde las puntuaciones.

Antes de graficar se construye un vector con los días de las acciones, para la construcción de este se emplea la función ***c ( )***, ésta concatena los valores dados como componentes de un vector como se muestra a continuación.

```
> fechas<-c(31,30,29,26,25,24,23,22,19,18,17,16,15,12,11,10,9,8,5,4,3,2,1)
```

Ejemplo 1.9

Para la gráfica se emplea la instrucción ***plot ( ),*** entre los argumentos que tiene esta función están***:***

- *x*: los valores en el eje x
- y: los valores del eje y
- type: el otro argumento va el tipo de gráfico ;por ejemplo :"p" para puntos, "l" para líneas, "h" para líneas verticales, entre otras.
- main: el nombre del título de la gráfica
- sub: el nombre del subtitulo de la gráfica.
- xlab: el nombre de los ejes de las x
- ylab: el nombre de los eje de las x

```
> plot(princomp(datos)$scores[,1:2],type="n")
```

Ejemplo 1.10

En este caso se grafica las puntuaciones y no se ponen líneas o puntos con el argumento "n".

Para completar el gráfico y así ponerle los días correspondientes al valor de las acciones se usa la función ***text( )*** la cual agrega texto dado por labels en las coordenadas

```
> text(princomp(datos)$scores[,1:2],label=fechas,col=fechas)
```

Ejemplo 1.11

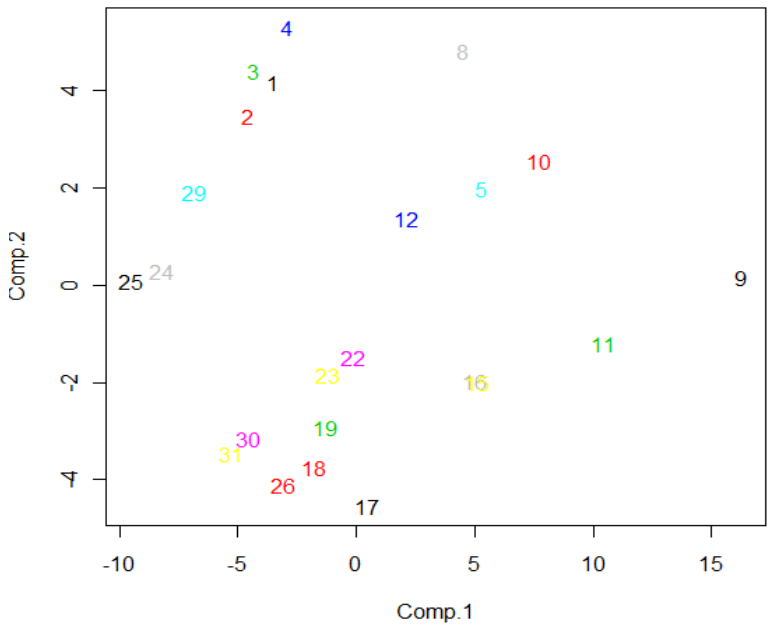

Figura 1.2 Gráfica de puntuaciones

En la figura 1.2 se observa que las fechas del 9, 10, 8, 11,4 y 16 se encuentran lejos del resto, por lo que se les pueden considerar como un *outliers*.

g) Para complementar el análisis se puede emplear la función ***biplot ( )*** y así verla proporción de la varianza explicada por las primeras dos componentes y los pesos (cargas ó *loadings*) relativos sobre el primer y segundo componente.

```
> biplot(princomp(datos))
```

Ejemplo 1.12

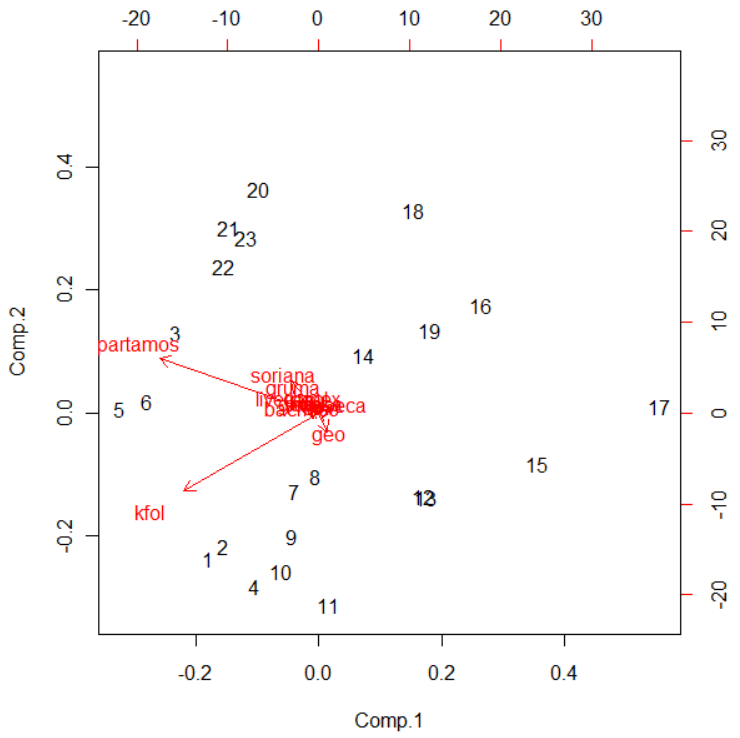

Figura 1.3 Gráfica de *biplot*

2. ANÁLISIS FACTORIAL

> El análisis factorial es una técnica que tiene por cometido reducir un conjunto de p variables aleatorias (interrelacionadas), en un grupo de m factores latentes independientes, de tal manera que el número de factores sean menores a las p variables, estos factores representan a las variables originales con una pérdida mínima de información.[7]

Aunque pueden realizarse análisis factoriales con variables discretas y/o ordinales lo habitual es que las variables sean cuantitativas continuas. Es aconsejable que se tenga una idea más o menos clara de cuáles son los factores comunes que quiere medir y que se elija las variables de acuerdo con ello.

Sí las unidades de medida de las variables no son comparables se deben estandarizar los datos antes de realizar el análisis.

El análisis factorial puede ser exploratorio o confirmatorio.

- **Análisis factorial exploratorio (AFE):** El investigador no tiene *a priori* una hipótesis acerca del número de factores comunes estos se seleccionan durante el análisis.

  Su propósito es determinar que variables directamente observables contribuyen a medir cada una de las variables latentes incluidas en el modelo.

- **Análisis factorial confirmatorio (AFC):** La hipótesis es que existen un número determinados factores preestablecidos los cuales tienen un significado y que cada uno de ellos está asociado con un determinado subconjunto de las variables. El análisis factorial confirmatorio arroja un nivel de confianza para poder aceptar o rechazar dicha hipótesis.

[7]Magdalena Ferrán Aranaz**,"**SPSS para Windows, Programación y Análisis Estadístico*".* España ,McGraw Hill, 1996. pág. 421.

2.1 MODELO

En el modelo del análisis factorial, sí los factores son inferidos a partir de las variables observadas, cada una será expresada como una combinación lineal de factores no observables directamente, por lo que un conjunto de variables aleatorias $X_1, \ldots., X_n$ se explicarán por un conjunto de factores comunes $F_1, \ldots., F_m$ y n factores únicos $U_1, \ldots., U_n$, como se puede ver en el modelo factorial lineal:

$$X_1 = a_{11}F_1 + \cdots + a_{1m}F_m + d_1U_1$$

$$X_2 = a_{21}F_1 + \cdots + a_{2m}F_m + d_2U_2$$

$$X_n = a_{n1}F_1 + \cdots + a_{nm}F_m + d_nU_n$$

Donde:

$a_{11}, \ldots, a_{1m}$: Son los pesos factoriales de los factores comunes

$F_1, \ldots., F_m$: m Factores comunes

d_n: Es el peso factorial del factor único

U_i: Es el factor único

Los principios fundamentales del modelo factorial lineal son dos:

1. La varianza total de la variable *i* (S_i^2) puede ser explicada por la suma de tres tipos de varianzas independientes entre sí:

- Varianza total explicada por los factores de 1 a *k* $(S_1^2, \ldots, S_k^2)$
- Varianza de los factores únicos (S_u^2), es la varianza no explicada por los factores comunes y corresponde a la varianza específica propia de la variable.
- Varianza del error debida a errores aleatorios

La notación matemática de la varianza total de una variable es:

$$S_i^2 = S_1^2 + S_2^2 + \cdots + S_k^2 + S_u^2$$

Dividiendo entre S_i^2:

$$1 = a_{i1}^2 + a_{i2}^2 + \cdots + a_{ik}^2 + S_{iu}^2$$

Donde:

- $a_{i1}^2 + a_{i2}^2 + \cdots + a_{ik}^2$ es la proporción de varianza total de la variables explicadas por los factores 1 a k

- S_{iu}^2 es la proporción de la varianza total de la variable explicada por la unicidad.

Lo que se anota como:

$$1 = h_i^2 + E_i^2 + e_i^2$$

En donde:

- $h_i^2 = a_{i1}^2 + a_{i2}^2 + \cdots + a_{ik}^2$ es la comunalidad o parte de la varianza total de la variable que se relaciona con el resto de las variables.

- $S_{iu}^2 = E_i^2 + e_i^2$ es la varianza total explicada por la especificidad de la variable total de la variable explicada por la unicidad (E_i^2) y la varianza del error (e_i^2).

2. La proporción de la varianza total de una variable explicada por cada uno de los factores comunes puede ser expresada como un coeficiente de determinación(r^2).La raíz cuadrada de esta proporción nos da un coeficiente de correlación (saturación) entre el factor y la variable.

La ecuación anterior quedaría:

$$1 = a_{i1}^2 + a_{i2}^2 + \cdots + a_{ik}^2 + E_i^2 + e_i^2$$

En donde:

- $a_{i1}^2 + a_{i2}^2 + \cdots + a_{ik}^2$ son los coeficientes de correlación de la variable i con los factores $1,2,3, \ldots, k$.

De lo anterior se deduce que la correlación entre dos variables es igual a la suma de los productos de sus saturaciones para cada uno de los factores, queda de la siguiente forma:

$$r_{xy} = r_{x1}r_{y1} + r_{x2}r_{y2} + \cdots + r_{xk}r_{yk}$$

De manera generalizada se anota:

$$r_{xy} = \sum a_{xy}a_{yk}$$

Así las correlaciones estimadas entre los factores y las variables pueden ser utilizadas como estimación de las correlaciones entre las variables.

2.2 SECUENCIA METODOLÓGICA EN EL ANÁLISIS FACTORIAL

En este cuadro se expone la secuencia del análisis factorial[8]:

Paso	Objetivo	Indicadores/Métodos
1.-Matriz de correlación	Examinar la correlación y la asociación lineal entre las variables	El determinante, Test de esfericidad de Barlett, Indice de KMO, Correlación anti-imagen, Medida de adecuación de la muestra y coeficientes de Correlación Múltiple.
2-Extracción de factores	Resumir el conjunto de variables en un subconjunto de factores, de tal manera que aun siendo en número menor, ofrezcan la misma información	Método centroide, Componentes principales, Factor Principal, Factorización de ejes principales, Máxima verosimilitud, Mínimos cuadrados no ponderados, Mínimos cuadrados no generalizados, etc
3.-Determinación del número de factores	Elegir el número de factores que se necesitan	Regla de Kaiser, MétodoMap ,Razón de verosimilitud, Aprior, etc
4.-Rotación de Factores	Pretende seleccionar la solución factorial más sencilla e interpretable	Ortogonales y Oblicuas
5.-Puntuaciones Factoriales	Permite determinar en que medida los factores seleccionados se dan en los individuos o en otras unidades.	

A continuación se desarrollará la metodología

[8]María José Rodríguez Jaime.“Análisis factorial” , Modelo *sociodemográficos:Atlas social de la cuidad de Alicate*. España , Biblioteca Miguel de Cervantes, 2001,pág 194.

2.2.1 MATRIZ DE CORRELACIÓN

Para que el análisis factorial tenga sentido las variables deben de estar asociadas linealmente. Se espera que las variables que tienen correlación muy alta entre sí la tengan con el mismo factor o factores.

Para medir el grado de asociación entre las variables existe un número importante de coeficientes estadísticos. De todos ellos, el coeficiente de correlación múltiple es el más conocido. Este indicador mide el grado de intercorrelación, de tal manera que cuando éstos sean bajos, las variables podrán ser eliminadas; y cuando sean altos, la matriz puede ser considerada para el análisis factorial. No obstante, no en todas las ocasiones una correlación baja implica que inexistencia de factores compartidos. Por ello existen test como el determinante de la matriz de correlaciones, esfericidad de Barlettó y Kaiser-Meyer-Olkin.

2.2.2 EXTRACCIÓN DE FACTORES

Para proceder a la extracción de los coeficientes factoriales se parte de la identidad fundamental del Análisis Factorial:

$$R = AA' + S_u^2$$

Donde:

- R: es la matriz de correlaciones entre las variables.
- A:es la matriz de cargas factoriales
- S_u^2: es la matriz de varianzas y covarianzas de los factores únicos.

Se considera una transformación de R: $R - S_u^2 = AA' = R^*$ (matriz de correlación reducida) cuyos elementos diagonales son las comunalidades y el resto, los coeficientes de correlación lineal entre las variables originales. La idea consiste en determinar A partiendo de alguna estimación para R^* y a partir de ella calcular los coeficientes de la matriz A.

Existen distintos métodos de estimación de los coeficientes de la matriz de cargas factoriales A; los más comunes son los siguientes:

a. Componentes Principales

El método consiste en estimar las puntuaciones factoriales mediante las puntuaciones tipificadas de las k primeras componentes principales y la matriz de cargas factoriales mediante las correlaciones de las variables originales con dichas componentes. La ventaja de este método es que siempre proporciona una solución, sin embargo, al no estar basado en el modelo de Análisis Factorial puede llevar a estimadores muy sesgados de la matriz de cargas factoriales.

b. Ejes Factoriales

Este método se basa en que sólo una parte de la variabilidad total de cada variable depende de factores comunes y, por tanto, la comunalidad inicial no será 1, estima dichas comunalidades mediante los coeficientes de determinación múltiple de cada variable con el resto. Se sustituyen estos valores en la diagonal principal de la matriz R^* y se procede a efectuar un Análisis de Componentes Principales. Una vez obtenido el resultado, se estiman de nuevo las comunalidades, se vuelven a sustituir en la diagonal principal de la matriz R^* y el proceso se retroalimenta hasta alcanzar un criterio de parada.

El inconveniente del método de Ejes Factoriales es que el cálculo de las comunalidades requiere mucho tiempo y recursos informáticos y, además, no siempre se pueden estimar o, incluso, pueden ser no válidas (comunalidades menores que 0 o mayores que 1).

c. Método de la máxima verosimilitud

Es un método de extracciones que tiene como hipótesis que la muestra procede de una distribución normal multivariada. Las correlaciones se ponderan por el inverso de la unicidad de las variables y se emplea un algoritmo iterativo. Este método genera un estadístico de bondad de ajuste chi-cuadrado que permite contrastar la bondad del modelo para explicar la matriz de correlaciones.

d. Mínimos cuadrados no ponderados

Este método extracción que minimiza la suma de los cuadrados de las diferencias entre las matrices de correlación observada y reproducida, ignorando los elementos de la diagonal.

e. Mínimos cuadrados generalizados

Al igual que el método anterior minimiza la suma de los cuadrados de las diferentes matrices ente las matrices de correlaciones observada y reproducida, pero en éste las correlaciones se ponderan por el inverso de su unicidad, de manera que las variables cuya unicidad es alta, reciben un peso menor que aquellas cuyo valor es bajo. Este método genera un estadístico de bondad de ajuste chi-cuadrada que permite contrastar la hipótesis nula que la matriz residual es la matriz nula.

f. Comparación entre los distintos métodos de extracción

1. Cuando las comunalidades son altas (mayores que 0.6) todos los procedimientos tienden a dar la misma solución.
2. Cuando las comunalidades son bajas para algunas de las variables el método de componentes principales tiende a dar soluciones muy diferentes del resto de los métodos, con cargas factoriales mayores.
3. Si el número de variables es alto (mayor que 30), las estimaciones de la comunalidad tienen menos influencia en la solución obtenida y todos los métodos tienden a dar el mismo resultado.
4. Si el número de variables es bajo todo depende del método utilizado para estimar las comunalidades y sí éstas son altas más que del método utilizado para estimarlas.

2.2.3 DETERMINACIÓN DEL NÚMERO DE FACTORES

La matriz factorial puede presentar un número de factores superior al necesario para explicar la estructura de los datos originales. Generalmente hay un conjunto reducido de factores, los primeros explican la mayor parte de la variabilidad total. Los otros factores suelen contribuir relativamente poco. Existen diversos criterios para determinar el número de factores a conservar.

a. Regla de Kaiser

Indica sólo conservar aquellos factores cuyos valores propios sean mayores a la unidad. Este criterio lo suelen utilizar los programas estadísticos por defecto. Sin embargo, este criterio es generalmente inadecuado tendiendo a sobre estimar el número de factores.

b. Método MAP

El método MAP (Minimum Average Partial) implica calcular el promedio de las correlaciones parciales al cuadrado después de que cada uno de los componentes ha sido parcializado de las variables originales. Cuando el promedio de las correlaciones parciales al cuadrado alcanza un mínimo no se extraen más componentes. Este mínimo se alcanza cuando la matriz residual se acerca más a una matriz identidad. Un requisito para utilizar esta regla es que cada uno de los componentes retenidos deben tener al menos dos variables con pesos altos en ellos.

c. Razón de Verosimilitud

El método de razón de verosimilitud, se trata de un criterio de bondad de ajuste pensado para la utilización del método de extracción de máxima verosimilitud, que se distribuye según Ji-cuadrado. La lógica de este procedimiento es comprobar si el número de factores extraído es suficiente para explicar los coeficientes de correlación observados.

d. Determinación *a priori*

La determinación *a priori* es más fiable si los datos y las variables están bien elegidos y el investigador conoce a fondo el problema, lo ideal es plantear el análisis factorial con una idea previa de cuántos factores hay y cuáles son.

2.2.4 ROTACIÓN DE LOS FACTORES

La rotación de factores pretende transformar la matriz inicial en una que sea más fácil de interpretar; esta rotación puede ser ortogonal u oblicua.

Cuando varios factores tienen una carga grande respecto a varias variables, resulta muy difícil determinar la forma como difieren los factores. La rotación no afecta la bondad de la solución factorial, y aunque la matriz factorial cambia, las comunalidades y los porcentajes de la varianza total explicada no cambian, pero si cambian los porcentajes atribuibles a cada factor. La rotación redistribuye la varianza explicada por los factores individuales. Así que diferentes métodos de rotación pueden conducir a la identificación de factores diferentes.

a. Rotación Ortogonal[9]

En ésta los ejes de coordenadas se rotan manteniendo un ángulo de noventa grados entre ellos y eso supone que los factores identificados no se relacionan entre sí. Existen varios métodos:

- *Método Varimax*: Es el más utilizado y trata de minimizar el número de variables que tienen alta carga en un factor.
- *Método Quartimax:* El objetivo de este método es que cada variable tenga correlaciones elevadas con un pequeño número de factores. Para ello busca maximizarla varianza de las cargas factoriales al cuadrado de cada variable en los factores.
- *Método Equamax*: Es una combinación de los dos anteriores, que simplifica los factores y las variables. No tiene mucha aceptación por lo tanto es utilizado pocas veces.

[9]http://tgrajales.net/estfactorial.pdf

b. Rotación oblicua

En ésta los ejes que se rotan conservan entre sí un ángulo diferente a noventa grados y supone cierto grado de relación entre los factores que lleguen a conformarse. Existen diferentes tipos de rotación:

- Rotación Oblimín: Trata de encontrar una estructura simple sin que importe el hecho de que las rotaciones sean ortogonales, esto es, las saturaciones no representan ya las correlaciones entre los factores y las variables. Se considera un parámetro que controla el grado de correlación entre los factores, con valores preferentemente entre −0,5 y 0,5.

  Se puede considerar que una rotación es sólo un medio para conseguir unos ejes que permitan describir los puntos de la muestra de la manera más simple posible.[10]

- Método Promax*:* Consiste en alterar los resultados de una rotación ortogonal hasta crear una solución con cargas factoriales lo más próximas a la estructura ideal. Dicha estructura supone que se obtiene elevando las cargas factoriales conseguidas en una rotación ortogonal, a una potencia que suele estar entre 2 y 4.

2.2.5 CÁLCULO DE LAS PUNTUACIONES FACTORIALES

Una vez determinados los factores rotados el siguiente paso es calcular la matriz de puntuaciones factoriales *F*, este cálculo es importante cuando se somete, con posterioridad, en otro tipo de análisis como de *cluster* ó conglomerados.

[10]Juan Manuel Marin Diazarque, *Análisis Factorial*, pág 6, pdf,http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema4am.pdf. Revisado el 3 de julio 2012

2.3 EJEMPLO EN *R*

En este ejemplo también se empleará la matriz de la BMV, la cual ya esta importada en (véase ejemplo 1.1).

Para el análisis factorial se siguen los siguientes pasos:

1-Al igual que en el método de componentes principales se observa si existe correlación entre las variables para ello se utiliza la función ***cor( )*** (véase ejemplo 1.2).

```
> cor(datos)
                axtel       cemex         geo      soriana compartamos       gruma     gfamsa        kfol    bachoco         cmr   liverpool
axtel       1.00000000  0.85488764 -0.56651149  0.908604791  0.64373268  0.83818289 0.5060068  0.30261533  0.7698939  0.13411027  0.49788613
cemex       0.85488764  1.00000000 -0.27613534  0.791434549  0.33393851  0.78662217 0.7621601 -0.01803411  0.5588849  0.29045629  0.36084386
geo        -0.56651149 -0.27613534  1.00000000 -0.545665653 -0.31699482 -0.24248490 0.2787616  0.01278454 -0.6053741 -0.07089945 -0.35119424
soriana     0.90860479  0.79143455 -0.54566565  1.000000000  0.64060365  0.79391904 0.4819617  0.19579284  0.5254244  0.22672240  0.54577745
compartamos 0.64373268  0.33393851 -0.31699482  0.640603648  1.00000000  0.58034259 0.2437971  0.63806026  0.4708987  0.09641522  0.28313796
gruma       0.83818289  0.78662217 -0.24248490  0.793919040  0.58034259  1.00000000 0.7022099  0.24436477  0.5500603  0.24340246  0.49136969
gfamsa      0.50600683  0.76216014  0.27876164  0.481961736  0.24379708  0.70220989 1.0000000  0.10496577  0.1776476  0.16570893  0.25146221
kfol        0.30261533 -0.01803411  0.01278454  0.195792841  0.63806026  0.24436477 0.1049658  1.00000000  0.3870336 -0.33613220  0.24970203
bachoco     0.76989392  0.55888489 -0.60537413  0.525424381  0.47089870  0.55006027 0.1776476  0.38703365  1.0000000 -0.14612568  0.38149349
cmr         0.13411027  0.29045629 -0.07089945  0.226722401  0.09641522  0.24340246 0.1657089 -0.33613220 -0.1461257  1.00000000 -0.03071525
liverpool   0.49788613  0.36084386 -0.35119424  0.545777452  0.28313796  0.49136969 0.2514622  0.24970203  0.3814935 -0.03071525  1.00000000
maseca      0.03153371  0.30926372  0.06242908 -0.001027291 -0.40634606  0.03814902 0.1606697 -0.55034913  0.0863552  0.21911303  0.04691426
                maseca
axtel       0.031533707
cemex       0.309263722
geo         0.062429076
soriana    -0.001027291
compartamos -0.406346057
gruma       0.038149021
gfamsa      0.160669676
kfol       -0.550349133
bachoco     0.086355192
cmr         0.219113030
liverpool   0.046914260
maseca      1.000000000
```

Ejemplo 2.1

Como la mayoría de los valores están cercanos a 1 o -1 por lo tanto están correlacionadas las acciones y si tiene sentido hacer un análisis factorial.

2-Se extraen los de factores, para ello se emplea la instrucción ***factanal( )*** la cual tiene en el primer argumento la matriz a utilizar, en el segundo el número de factores que se quieren sacar y el tercero podemos elegir el tipo de rotación a usar .El método por default que emplea ***R*** para la extracción de factores es el de máxima verosimilitud explicado con anterioridad en el capítulo.

Para un factor:

```
> datosfact1<-factanal(datos, factors=1, rotation="varimax")
> datosfact1

Call:
factanal(x = datos, factors = 1, rotation = "varimax")

Uniquenesses:
     axtel      cemex        geo    soriana compartamos      gruma     gfamsa       kfol    bachoco        cmr   liverpool     maseca
     0.005      0.267      0.685      0.171      0.586      0.292      0.738      0.910      0.416      0.981      0.748      0.999

Loadings:
            Factor1
axtel        0.998
cemex        0.856
geo         -0.562
soriana      0.911
compartamos  0.644
gruma        0.842
gfamsa       0.512
kfol         0.300
bachoco      0.764
cmr          0.139
liverpool    0.502
maseca

               Factor1
SS loadings      5.203
Proportion Var   0.434

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 137.23 on 54 degrees of freedom.
The p-value is 3.5e-09
```

Ejemplo 2.2

Lo que arroja ***R*** es:

- *Loadings:* Los pesos de los factores para cada variable.
- *Uniquenesses*: Es la unicidad de cada uno de las variables.
- *SS loadings:* Es la suma de los cuadrados de los pesos.
- *Proportin Var:* La proporción de la varianza.
- *El test de significancia y el p-value*: La hipótesis nula que utiliza ***R*** se refiere si es adecuado el número de factores, es decir, si el *p-value* es muy pequeño se rechaza la hipótesis, por lo que ese número de factores no son los adecuados como es el caso para este primer factor.

Así se revisa la solución para dos factores

```
> datosfact2<-factanal(datos, factors=2, rotation="varimax")
> datosfact2

Call:
factanal(x = datos, factors = 2, rotation = "varimax")

Uniquenesses:
      axtel       cemex         geo     soriana compartamos       gruma      gfamsa        kfol     bachoco         cmr   liverpool      maseca
      0.005       0.123       0.241       0.170       0.579       0.192       0.005       0.909       0.351       0.967       0.748       0.970

Loadings:
            Factor1 Factor2
axtel        0.943   0.326
cemex        0.683   0.640
geo         -0.753   0.437
soriana      0.853   0.320
compartamos  0.638   0.119
gruma        0.688   0.578
gfamsa       0.199   0.978
kfol         0.298
bachoco      0.806
cmr                  0.154
liverpool    0.476   0.159
maseca               0.172

               Factor1 Factor2
SS loadings      4.544   2.195
Proportion Var   0.379   0.183
Cumulative Var   0.379   0.562

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 93.28 on 43 degrees of freedom.
The p-value is 1.41e-05
```

Ejemplo 2.3

Se prueba para tres factores:

```
> datosfact3<-factanal(datos, factors=3, rotation="varimax")
> datosfact3

Call:
factanal(x = datos, factors = 3, rotation = "varimax")

Uniquenesses:
      axtel       cemex         geo     soriana compartamos       gruma      gfamsa        kfol     bachoco         cmr   liverpool      maseca
      0.005       0.010       0.190       0.170       0.270       0.185       0.026       0.278       0.348       0.845       0.733       0.534

Loadings:
            Factor1 Factor2 Factor3
axtel        0.915   0.396
cemex        0.653   0.654  -0.369
geo         -0.797   0.411
soriana      0.828   0.379
compartamos  0.598   0.234   0.564
gruma        0.637   0.634
gfamsa       0.133   0.972  -0.107
kfol         0.250   0.166   0.795
bachoco      0.801
cmr                  0.134  -0.358
liverpool    0.459   0.193   0.140
maseca                      -0.676

               Factor1 Factor2 Factor3
SS loadings      4.289   2.398   1.720
Proportion Var   0.357   0.200   0.143
Cumulative Var   0.357   0.557   0.701

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 52.83 on 33 degrees of freedom.
The p-value is 0.0157
```

Ejemplo 2.4

Por último para cuatro factores:

```
> datosfact4<-factanal(datos, factors=4, rotation="varimax")
> datosfact4

Call:
factanal(x = datos, factors = 4, rotation = "varimax")

Uniquenesses:
      axtel       cemex         geo     soriana compartamos       gruma      gfamsa        kfol     bachoco         cmr   liverpool      maseca
      0.014       0.008       0.078       0.053       0.231       0.212       0.095       0.147       0.005       0.713       0.715       0.489

Loadings:
            Factor1 Factor2 Factor3 Factor4
axtel        0.831   0.521   0.139
cemex        0.941   0.211  -0.234
geo                 -0.956
soriana      0.767   0.490   0.236  -0.249
compartamos  0.470   0.293   0.669   0.126
gruma        0.849   0.193   0.176
gfamsa       0.869  -0.375
kfol         0.215           0.731   0.521
bachoco      0.506   0.606           0.609
cmr          0.209          -0.168  -0.462
liverpool    0.376   0.315   0.210
maseca       0.158          -0.697

               Factor1 Factor2 Factor3 Factor4
SS loadings      4.378   2.203   1.706   0.951
Proportion Var   0.365   0.184   0.142   0.079
Cumulative Var   0.365   0.548   0.691   0.770

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 31.13 on 24 degrees of freedom.
The p-value is 0.15
```

Ejemplo 2.5

La solución con tres factores (ejemplo 2.4) resulta ser la más apropiada por el obtener el mayor valor en el *p-value*, lo que implica que no se rechaza la hipótesis nula.

3-En caso de que se quieran saber las puntuaciones factoriales, se cambia la parte de la instrucción ***rotation*** por ***scores*** el cual puede ser por dos métodos: regresión ("*regression*") o mínimos cuadrados de Bartlet ("*Bartlett*").También al final de la instrucción de

factanal( ) le agregamos ***$scores.***

```
> scores<-factanal(datos, factors=4, scores="regression")$scores
> scores
           Factor1     Factor2     Factor3     Factor4
 [1,]  0.135818000  0.33025932  0.84306331  0.68254224
 [2,] -0.074218018  0.26707093  0.89208447  0.35720088
 [3,] -0.541860517  0.46815534  1.32577918 -0.33006843
 [4,] -0.552611972  0.12324225  1.06646501 -0.51811190
 [5,]  0.111736888 -0.26084819  1.87579205 -0.80631250
 [6,] -0.219006188  0.06352719  1.83785648 -0.56810356
 [7,] -0.982051607  0.55343758  0.55013656  0.98839083
 [8,] -0.902110213  0.44677375 -0.29090524  1.22585260
 [9,] -0.652121131  0.08088542 -0.15547101  2.04800468
[10,] -0.012252569 -0.55616224 -0.07145239  1.73239285
[11,]  0.687371231 -1.29273321 -1.11953222  1.68809105
[12,]  0.002946352 -1.30189776 -0.57135955 -0.24032186
[13,] -0.171843156 -1.55121141 -0.16043527 -0.32706117
[14,]  0.075380671 -1.92752584  0.11526987 -0.35219997
[15,] -0.868093568 -1.07146368 -0.57946430 -1.06462913
[16,] -0.978201023 -0.57773915 -1.03077982 -1.34195864
[17,] -1.222099898  0.27399665 -1.30225300 -1.71150558
[18,] -0.591191009  1.60117906 -1.21415690 -0.06637915
[19,] -0.272163035  1.74249678 -1.38409013  0.02032165
[20,]  0.797158455  1.32609642  0.12410313 -0.35558429
[21,]  1.284561863  1.00354307  0.01994313 -0.02440280
[22,]  2.160919240  0.52064615  0.06389889 -0.57136630
[23,]  2.783931203 -0.26172843 -0.83449226 -0.46479149
```

Ejemplo 2.6

Para la representación de las puntuaciones factoriales se emplea la instrucción ***pairs( )*** la cual su resultado es una matriz de dispersión.

```
> pairs(scores)
```

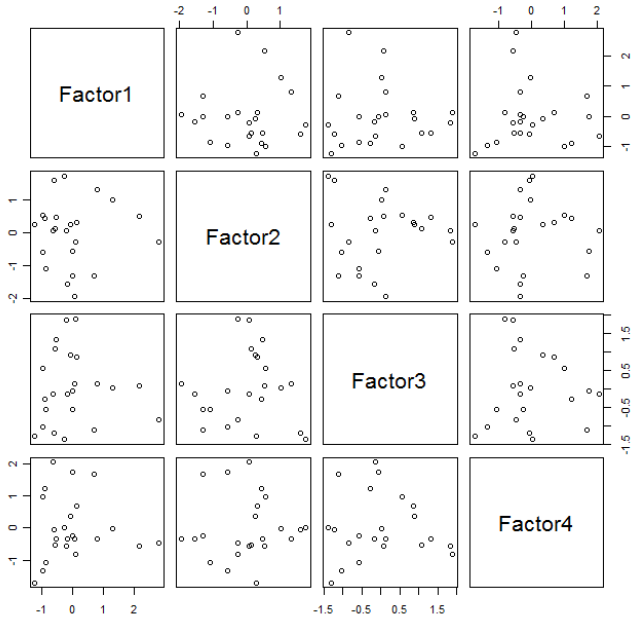

Figura2.1

4-Para graficar las puntaciones factoriales se utilizan el siguiente conjunto de instrucciones:

- ***par( )***[11]se emplea para establecer los parámetros de los gráficos de forma permanente; es decir; para establecer el color del fondo del gráfico, el color de la letra, dividir la ventana gráfica, entre otros. Para este ejemplo se utiliza el parámetro ***pty*** que especifica el tipo de zona de dibujo que se utilizará; "s" genera una región cuadrada trazado y la "m" genera la región máxima.
- ***plot( )*** (véase ejemplo 1.10)
- ***text ( )*** **(**véase ejemplo 1.11)

[11]Para mayor información sobre esta instrucción consultar ver tesina Fabiola López Gonzaléz,"Lenguaje R",*Lenguaje R: Un complemento libre para las asignaturas de estadística*,México ,UNAM,2008.pág 51

```
> par(pty="s")
> plot(scores[,1],scores[,2],
+ ylim=range(scores[,2]),
+ xlab="Factor 1",ylab="Factor 2",type="n",lwd=2)
> text(scores[,1],scores[,2],
+ labels=abbreviate(row.names(datos),minlength=8),cex=0.6,lwd=2)
```

Ejemplo 2.5

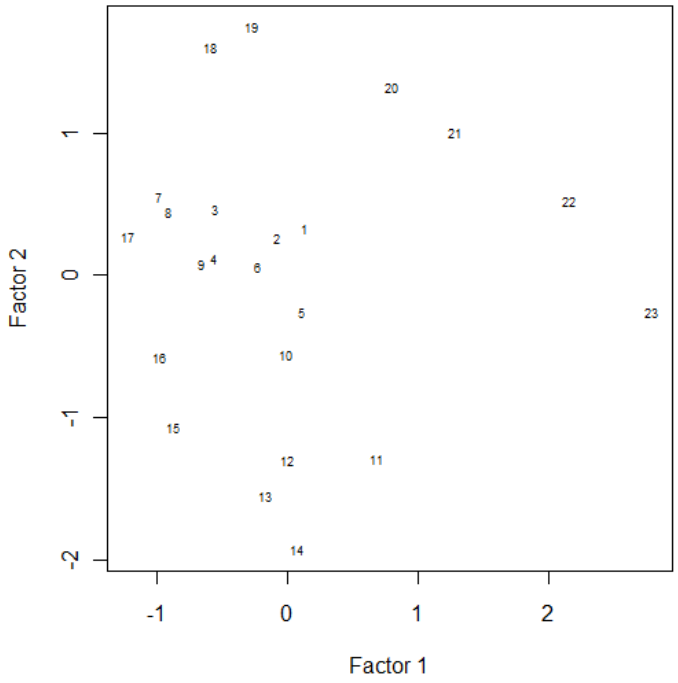

Figura 2.2

En este caso es para los factores 1 y 2, bien se puede hacer para las combinaciones de los factores 2 vs 3,1 vs 3, entre otras. Con esta gráfica podremos conocer que días son los más raros o extremos puede ayudar a detectar casos atípicos.

3. ANÁLISIS DISCRIMINANTE

El análisis discriminante es una técnica de clasificación y asignación de individuos a un grupo conocidas sus características. En este se disponen de una serie de grupos definidos
a priori con una serie de observaciones para cada individuo referidas a un conjunto de variables relevantes. En base a la información se llega a calcular una función discriminante que se puede utilizar para predicciones futuras[12].

Existen dos ramas dentro del análisis discriminante: el descriptivo y el predictivo.

- **Análisis Discriminante Predictivo**: Se utiliza cuando se tiene individuos de origen desconocido, éste nos proporciona una regla de clasificación lo más precisa posible para determinar a qué grupo pertenece una nueva observación.

- **Análisis Discriminante Descriptivo**: Desea saber cuáles son las variables que más diferencian a los grupos, cuales son importantes y cuáles no dan efectos de clasificar los sujetos.

3.1 MODELO

En general se tienen poblaciones $\pi_j, j = 1,2,\dots,k$, conocidas y en cada una de ellas existen observaciones de una muestra de cierto vector $X = (x_1,\dots.,x_p)'$. Se trata de encontrar funciones discriminantes o reglas de decisión $f = f(x_1,\dots.,x_p)$ cuyos valores en los distintos grupos (o poblaciones) estén lo más separados posible. En algunos casos, tales funciones definen regiones $R_1,..,R_n$ del espacio euclideano R^n.

Dado un nuevo individuo w cuya población de procedencia se desconoce, y sobre el cual se pueden medir las variables, $x_1,\dots.,x_p$ el problema de clasificación trata de asignar al

[12]Rafael Bisquerra Alzina, "Análisis discriminante" en *Introducción conceptual al análisis multivariable : Un enfoque informatica con los paquetes spss-x, bmdp, lisrel y spad,vol 1*, Barcelona , Promociones y publicaciones universitarias,1989,pág 243

individuo alguna de las poblaciones $\pi_j, j = 1,2, \dots, k$.Para ello se utilizan las funciones discriminantes.

3.2 CLASIFICACIÓN PARA DOS POBLACIONES

3.2.1 REGLA DE LA MÁXIMA VEROSIMILITUD

Sea la densidad de la población $\pi_1 f_1(x)$. La regla de la máxima verosimilitud dada para colocar x en π_j maximizando la verosimilitud es $L_1(x) = f_1(x) = \max_1 f_1(x)$. La regla de máxima verosimilitud está definida de la siguiente forma:

$$x \in \pi_1 \quad \text{si } L_1(x) > L_2(x)$$

$$x \in \pi_2 \text{si } L_1(x) < L_2(x)$$

3.2.2 REGLA DE BAYES

Supongamos que se conoce la probabilidad $q_1 = P(\pi_1)$ de que un individuo x pertenezca a π_1.Conocida $x = (x_1, \dots., x_n)$ de verosimilitud $L_1(x)$, la probabilidad de que x sea de π_1 se obtiene el Teorema de Bayes:

$$P(\pi_1|x) = \frac{q_1 L_1(x)}{q_1 L_1(x) + q_2 L_2(x)}$$

El criterio queda:

$$x \in \pi_1 \text{ si } q_1 L_1(x) > q_2 L_2(x)$$

$$x \in \pi_1 \text{ si } q_1 L_1(x) < q_2 L_2(x)$$

La regla de Máxima verosimilitud coincide con la de Bayes cuando $q_1 = q_2 = \frac{1}{2}$

Las probabilidades de clasificación errónea para cuando j =2 son p_{21} y p_{21} donde:

- $p_{21} = P(X \epsilon R_2 | \pi_1) = \int_{R_2} f_1(x)dx =$ Probabilidad de poner un individuo en el grupo dos dado que pertenezca a la población uno.
- $p_{12} = P(X \epsilon R_1 | \pi_2) = \int_{R_1} f_2(x)dx =$ Probabilidad de poner un individuo en el grupo uno dado que pertenezca a la población dos.

Las probabilidades de clasificación errónea crean un costo $C(i|j)$ cuando a π_i se le asigna un individuo perteneciente a π_j siendo el criterio de clasificación:

$$R_j = \{x: C(i|j) q_j L_j(x) > C(j|i) q_i L_i(x) \text{ para } i = 1, \ldots, k i \neq j\}$$

La esperanza del costo de clasificación está dada por:

$$ECM = C(2|1) p_{21} \pi_i + C(1|2) p_{12} \pi_j$$

3.2.3 CRITERIO GEOMETRICO

La distancia de Mahalanobis entre el individuo de coordenadas $x = (x_1, \ldots, x_n)'$ y la poblaciónπ_i es:

$$\delta^2(x, \mu_1) = (x - \mu_1) S^{-1} (x - \mu_2)'$$

La regla de decisión es:

$$x \epsilon \pi_1 \text{ si } \delta^2(x, \mu_1) > \delta^2(x, \mu_2) \text{ para } i = 1, \ldots, k;\ i \neq j$$

$$x \epsilon \pi_2 \text{ si } \delta^2(x,\mu_1) < \delta^2(x,\mu_2) \text{para } i = 1,\ldots,k;\ i \neq j$$

Se define la función discriminante como:

$$F(x) = (\mu_1 - \mu_2)'\mathrm{S}^{-1}\mathrm{x} = \mathrm{d}_1\mathrm{x}_1 + \cdots + \mathrm{d}_\mathrm{n}\mathrm{x}_\mathrm{n}$$

3.2.4 DISCRIMINADOR LINEAL DE FISHER

En las aplicaciones μ_1 ,μ_2 y S son desconocidas, siendo estimadas por los vectores de medias muestrales ,$\overline{X_1}$,$\overline{X_2}$y la matriz de covarianzas muestrales *S*. Indicando, $\overline{X} = (\overline{X_1} + \bar{X}_2)$, el discriminante lineal es:

$$\hat{F}(x) = (\overline{X_1} - \overline{X_2})'S^{-1}x$$

La regla de decisiones:

$$x \epsilon \pi_1 \text{ si } \hat{F}(x) > \hat{F}(\overline{x})$$

$$x \epsilon \pi_2 \text{ si } \hat{F}(x) < \hat{F}(\overline{x})$$

3.2.5 CLASIFICACIÓN DE POBLACIONES NORMALES

Sea $X_1 \epsilon \pi_1$ y $X_1 \sim N(\mu_1, S_1)$, además $X_2 \epsilon \pi_2$ y $X_2 \sim N(\mu_2, S_2)$

$$x \epsilon \pi_1 \text{ si } L_1(x) > L_2(x)$$

$$x \epsilon \pi_2 \text{si } L_1(x) < L_2(x)$$

Donde $L_i(x) = (2\pi)^{-\mathrm{n}/2}|S_i|\mathrm{e}^{[1/2(\mathrm{x}-\mu_i)'S^{-1}(\mathrm{x}-\mu_i)]}\ i = 1,2$

3.3CLASIFICACIÓN DE *K* POBLACIONES

3.3.1 REGLA DE LA MÁXIMA VEROSIMILITUD

La regla de máxima verosimilitud consiste en asignar x ala población con la mayor verosimilitud, por lo que el criterio de decisión quedaría:

$$x\epsilon\,\pi_j \text{ si } L_j(x) = \max_i\{|L_i(x)|\} \text{ para } i = 1,\ldots,k,$$

3.3.2 REGLA DE BAYES

Generalizando para la regla de Bayes queda de la siguiente manera:

$$x\epsilon\,\pi_j \text{ si } q_j L_j(x) = \max_i\{q_i|L_i(x)|\} \text{ para } i = 1,\ldots,k,$$

3.3.3 CRITERIO GEOMÉTRICO

Reside en asignar x a la población más próxima por lo tanto la regla de decisión es :

$$x\epsilon\pi_j \text{ si } \delta^2(x,\pi_j) = \min_i\{\delta^2(x,\pi_i)\}$$

3.3.4 DISCRIMINADOR LINEAL DE FISHER

La regla discriminante lineal se basa en la maximización de la relación de la entre la varianza dentro de una proyección $\propto' x$.

Supongamos que se tienen muestras $X_j, j = 1,\ldots,k$, donde k son las poblaciones. Sea $Y = X \propto$ y $Y_j = X_j \propto$ que denotan combinaciones lineales. La suma de cuadrados intragrupo está dada por :

$$\sum_{j=1}^{J} Y_j^{'\prime} H_j Y_j = \sum_{j=1}^{J} \alpha X_j^{'\prime} H_j X_j \alpha = \alpha' W \alpha$$

Donde H_j es una matriz semidefinida positiva de$(n_i \times n_j)$. La suma de cuadrados entre grupos es:

$$\sum_{j=1}^{J} n_j (\bar{y}_j - \overline{y})^2 = \sum_{j=1}^{J} n_j \{\propto'(\bar{x}_j - \overline{x})\}^2 = \propto' \beta \propto$$

Donde $\bar{\boldsymbol{y}}_{\boldsymbol{j}}$ y $\bar{\mathbf{x}}_{\boldsymbol{j}}$ son la media de $\boldsymbol{Y_j}$ y $\boldsymbol{X_j}$ y $\bar{\boldsymbol{y}}$ y $\bar{\boldsymbol{x}}$ son la media de $\boldsymbol{Y}$ y $\mathbf{X}$ respectivamente.

Finalmente, siendo el criterio de clasificación queda:

$$R_j = \{x: |\propto' (x - \bar{x}_j)| \leq |\propto' (x - \bar{x}_i)| \text{para } i = 1, \ldots, j\}$$

3.4 EJEMPLO EN *R*

Para este ejemplo se utilizará la base de acciones de algunas empresas de la BMV que se empleo en los capítulos pasados con el nombre que se ha estado asignado **datos** (véase ejemplo 1.1)

Antes de hacer el análisis discriminante se le anexa a la base el vector con las semanas del mes, ya que estos serán los grupos. Luego se unen con la base con la función

data frame( ).

```
> semana<-c(5,5,5,4,4,4,4,4,3,3,3,3,3,2,2,2,2,2,1,1,1,1,1)
> datos=data.frame(datos,semana)
```

Ejemplo 3.1

1-El primer paso para hacer el análisis discriminante es llamar a la librería ***MASS***[13] la cual contiene un conjunto de funciones de la estadística aplicada moderna.

```
> library(MASS)
```

Ejemplo 3.2

2- Se escribe la función para el análisis discriminante lineal ***lda( )***

```
> lda(semana~.,datos[,1:12])
Call:
lda(semana ~ ., data = datos[, 1:12])

Prior probabilities of groups:
        1         2         3         4         5
0.2173913 0.2173913 0.2173913 0.2173913 0.1304348

Group means:
  axtel cemex      geo soriana compartamos    gruma   gfamsa     kfol  bachoco   cmr liverpool   maseca
1 6.580 7.652 22.75200  31.972    75.85600 24.43800 14.45600 113.5200 23.05000 3.290  96.00000 13.77800
2 5.322 6.466 24.50600  28.992    68.21000 21.89800 13.30200 108.2320 21.47400 3.200  93.48400 13.90600
3 5.540 6.622 24.91400  28.550    70.42400 22.48200 14.10400 116.0220 22.30800 3.030  93.94600 13.45600
4 5.712 6.350 23.69600  29.800    76.06800 22.48400 13.21800 118.4340 22.17400 3.082  93.88800 12.90800
5 5.910 6.550 23.55333  30.330    76.08667 22.83667 13.35667 119.5167 22.45667 3.050  96.99667 13.06333

Coefficients of linear discriminants:
                LD1         LD2         LD3         LD4
axtel      -7.2906360  0.08012637 -4.43174176  0.56260686
cemex       2.5534767 -4.79114127  7.40849736  1.79632618
geo         2.0020413 -0.92938051 -0.62214182  0.98097318
soriana     1.7002034 -0.25046082  0.06327215 -0.46808089
compartamos 0.1832792 -0.15596277  0.23232181  0.04635991
gruma      -0.3016218 -0.96599461  1.44547691 -0.52004631
gfamsa     -3.4507925  3.38571004 -5.18146047 -1.05915785
kfol       -0.2210198  0.59854805  0.34856533 -0.11870561
bachoco     0.4021939  1.17331151 -3.70359494  0.43159902
cmr        -3.1279720  2.03481747 -2.88387154 -0.03547511
liverpool  -0.1558792 -0.07332115  0.28344408  0.82185380
maseca     -0.1533407 -0.40098093  0.59393113 -0.56413270

Proportion of trace:
   LD1    LD2    LD3    LD4
0.4289 0.3845 0.1675 0.0191
```

Ejemplo 3.3

Ésta devuelve:

- *Priori probabilities of group**s***: Las probabilidades *a priori* de los grupos, que se calculan utilizando la proporción de elementos de cada grupo (en este caso el primer grupo contiene 5 elementos de los 23 ,por lo tanto la probabilidad queda como 5/23 =02.1739)
- *Group means*: La media de cada grupo

.

[13]Autor Brian Ripley

- *Coeficients of linear discriminant:* Los coeficientes de los discriminantes lineales se usan por la función para decidir a que grupo pertenece cada fecha.

- *Proportion of trace*: da una idea de la importancia de cada eje discriminante. Al ser LD1 y LD2 mucho mayores que los otros, esto quiere decir que las semanas se pueden clasificar bien utilizando dos ejes discriminantes.

Si sólo se quiere verlas probabilidades *a priori* se escribe:

```
> datos.lda$prior
        1         2         3         4         5
0.2173913 0.2173913 0.2173913 0.2173913 0.1304348
```

Ejemplo 3.4

3-Si llega nueva información para predecir a que grupo pertenecen, lo primero que se debe de hacer es una matriz con los nuevos registros esto se puede lograr con la función

rbind ( ) que une los vectores horizontalmente, como se muestra a continuación:

```
>nuevosdatos<rbind(c(6.12,6.63,23.97,30.11,74.01,24,13.79,122,22.7,3.3,96.24,13.75),c(6.05,6.46,23.9,29.79,73,23.5,13.5,121.42,23.7,3.3,93.57,13.95))
```

Ejemplo 3.5

En este caso se anexan los dos últimos días de la quinta semana, es decir, 1° y 2 de septiembre. Para ponerle nombre a cada una de las columnas de esta nueva la función ***colnames ( ).***

```
> colnames(nuevosdatos)<-colnames(datos[,-13])
```

Ejemplo 3.6

Lo que hace el ejemplo anterior es con ***colnames(datos[,-13])*** extrae los nombres de las columnas de nuestra base de acciones,para asignarselos a ***nuevosdatos***,osea la nueva base.

Se hace un marco de datos:

```
> nuevosdatos<-data.frame(nuevosdatos)
```

Ejemplo 3.7

Luego se predice con ***predict( ) $class. Predict*** es una función genérica para las predicciones de los resultados de las funciones del modelo de ajuste diferentes, al ponerle *$class* al final hace que prediga el número de grupo.

```
> predict(datos.lda,nuevosdatos)$class
[1] 5 3
Levels: 1 2 3 4 5
```

Ejemplo 3.8

En este caso la predicción no es del todo correcta ya que para el segundo registro que se introdujo arrojo que pertenecía a la tercera semana cuando en realidad pertenece a la quinta.

Una forma alternativa de proporcionar el conjunto de datos de la función ***lda( )*** es:

```
> lda(datos[,1:12],semana)
Call:
lda(datos[, 1:12], semana)

Prior probabilities of groups:
        1         2         3         4         5
0.2173913 0.2173913 0.2173913 0.2173913 0.1304348

Group means:
  axtel cemex      geo soriana compartamos    gruma   gfamsa     kfol
1 6.580 7.652 22.75200  31.972    75.85600 24.43800 14.45600 113.5200
2 5.322 6.466 24.50600  28.992    68.21000 21.89800 13.30200 108.2320
3 5.540 6.622 24.91400  28.550    70.42400 22.48200 14.10400 116.0220
4 5.712 6.350 23.69600  29.800    76.06800 22.48400 13.21800 118.4340
5 5.910 6.550 23.55333  30.330    76.08667 22.83667 13.35667 119.5167
   bachoco   cmr liverpool   maseca
1 23.05000 3.290  96.00000 13.77800
2 21.47400 3.200  93.48400 13.90600
3 22.30800 3.030  93.94600 13.45600
4 22.17400 3.082  93.88800 12.90800
5 22.45667 3.050  96.99667 13.06333

Coefficients of linear discriminants:
                  LD1         LD2         LD3         LD4
axtel       -7.2906360  0.08012637 -4.43174176  0.56260686
cemex        2.5534767 -4.79114127  7.40849736  1.79632618
geo          2.0020413 -0.92938051 -0.62214182  0.98097318
soriana      1.7002034 -0.25046082  0.06327215 -0.46808089
compartamos  0.1832792 -0.15596277  0.23232181  0.04635991
gruma       -0.3016218 -0.96599461  1.44547691 -0.52004631
gfamsa      -3.4507925  3.38571004 -5.18146047 -1.05915785
kfol        -0.2210198  0.59854805  0.34856533 -0.11870561
bachoco      0.4021939  1.17331151 -3.70359494  0.43159902
cmr         -3.1279720  2.03481747 -2.88387154 -0.03547511
liverpool   -0.1558792 -0.07332115  0.28344408  0.82185380
maseca      -0.1533407 -0.40098093  0.59393113 -0.56413270

Proportion of trace:
   LD1    LD2    LD3    LD4
0.4289 0.3845 0.1675 0.0191
```

Ejemplo 3.9

4 ANÁLISIS DE CONGLOMERADOS (AC)

> El análisis de conglomerados (*cluster* en inglés) o Taxonomía Numérica o Análisis de Clasificación, es una técnica estadística multivariante cuya idea básica es agrupar un conjunto de observaciones en un número dado de grupos o conglomerados.[14]

Éste es una técnica de análisis exploratorio puesto que no utiliza ningún tipo de modelo estadístico para llevar a cabo el proceso de clasificación, este agrupamiento se basa en la idea de distancia o similitud.

4.1 SELECCIÓN DE UNA MEDIDA DE SIMILITUD

Se seleccionan las variables en las que se basa la agrupación y se establece una medida de proximidad o disimilariad entre ellos que cuantifique el grado de similaridad entre cada par de objetos. Las medidas de proximidad, similitud o semejanza miden el grado de semejanza entre dos objetos de forma que, cuanto mayor (menor) es su valor, mayor (menor) es el grado de similaridad existente entre ellos y con más (menos) probabilidad los métodos de clasificación tenderán a ponerlos en el mismo grupo.

Las medidas de disimilitud, desemejanza o distancia miden la distancia entre dos objetos de forma que, cuanto mayor (menor) sea su valor, más (menos) diferentes son los objetos y menor (mayor) la probabilidad de que los métodos de clasificación los pongan en el mismo grupo.

Por lo que una medida de proximidad, es un índice que cuantifica el grado de asociación, similaridad (s_{ij}) o disimilaridad (δ_{ij}), de un par de objetos o estímulos. Esta medida surge de dos maneras: directas o derivadas. Sí los individuos (ya sean encuestados, empresas, países, etc.) fueron cuestionados para examinar pares de objetos y en consecuencia cuantificar explícitamente la (dis)similaridad de cada par, entonces se dice que los datos son proximidades directas de juicios. En cambio las derivadas son estimadas o calculadas a

[14]Luis Antonio Chamba Eras,“ Análisis de conglomerados”,*Exploración y Análisis de datos.* País Vasco ,Universidad del país Vasco, 2010, pág. 25.

partir de la matriz de datos original, generalmente la matriz de proximidades se construye a partir de la matriz de covarianza o de correlación.

4.1.1. COEFICIENTES DE DISIMILARIDAD

Según el tipo de datos que existen diferentes formas de medir las proximidades. A continuación se describen los principales índices de distancia y similitud que se usan.

Sea Ω un conjunto formado por k objetos, una distancia sobre Ω es una función

$d_{ij}: \Omega \times \Omega \to R^{+} \cup \{0\}$ tal que cumple con las siguientes propiedades:

1. $d_{ij} \geq 0$
2. $d_{ii} = 0$
3. $d_{ij} = d_{ji}$ (simetría)
4. $d_{ij} \leq d_{it} + d_{jt}$ (desigualdad triangular)
5. $d_{ij} = 0 \leftrightarrow i = j$
6. $d_{ij} \leq \max\{d_{it}, d_{jt}\}$ (desigualdad ultramétrica)

Una distancia recibe diferentes denominaciones según las propiedades que verifica; si cumple las propiedades 1-3 se dice que esdisimilaridad,1-5 es métrica, y 1-6 ultramétrica.

a) Datos del tipo binario

Para las medidas de disimilitud son necesarios los coeficientes de similaridad, los más sencillos y comúnmente utilizados son: a, b, c, d, n y p los cuales, dados dos individuos i, j se pueden calcular de la siguiente forma:

- $a_{ij} = \sum_{r=1}^{p} x_{ir}\, x_{jr}$: es el número de atributos que posee el individuo i y j
- $b_{ij} = \sum_{r=1}^{p} x_{ir}\,(1 - x_{jr})$: es el número de atributos que posee el individuo i y no j

- $c_{ij} = \sum_{r=1}^{p}(1 - x_{ir})\, x_{jr}$: es el número de atributos que posee j pero no i
- $d_{ij} = \sum_{r=1}^{p}(1 - x_{ir})(1 - x_{jr})$: es cuando ninguno posee el atributo
- $n_i = a_{ij} + b_{ij} = \sum_{r=1}^{p} x_{ir}$: es el número de atributos que posee i
- $n_j = a_{ij} + c_{ij} = \sum_{r=1}^{p} x_{jr}$: es el número de atributos que posee j

Claramente $p = a_{ij} + b_{ij} + c_{ij} + d_{ij}$

Las principales medidas de disimilitud para datos de tipo binario son las siguientes: Jaccard, Czekanowski, Dice & Sorense, Socal & Minchesner, Russel & Rao, Kulenzinski, y Euclideana Ponderada

b) Datos cuantitativos

Se supone sobre las variables cuantitativas o bien numéricas, ya sean de tipo discreto o continuo pero no binario.

Las principales distancias en este contexto son: Euclideana, Euclideana Ponderada, Karl Pearson, Mahalanobis, Métrica City Block (o Manhattan) y Métrica de Minkowski.

c) Datos cualitativos

Al tener características cualitativas de los objetos entonces tenemos datos del tipo cualitativo, los que se pueden reescribir con ceros y unos, por lo que se consideran variables de tipo binarias y así se pueden usar los índices de disimilitud mencionados anteriormente para este tipo de datos.

d) Datos mixtos

Si en la base de datos existen diferentes tipos de variables: binarias, categóricas, ordinales, cuantitativas no existe una solución universal al problema de cómo combinarlas para construir una medida de distancia. Algunas soluciones sugeridas son:

- Expresar todas las variables en una escala común, habitualmente binaria, transformando el problema en uno de los ya contemplados anteriormente. Esto tiene costos en términos de pérdida de información.

- Realizar análisis por separado utilizando variables del mismo tipo y utilizar el resto de las variables como instrumentos para interpretar los resultados obtenidos.

4.1.2. COEFICIENTES DE SIMILARIDAD

Una medida de similaridad sobre Ω es una función tal que s_{ij}: $\Omega \times \Omega \rightarrow R^{+} \cup \{0\}$ la cual cumple con las siguientes propiedades:

1. $s_{ij} \geq 0$
2. $s_{ij} = s_{ji}$
3. s_{ij} crece cuando la similaridad entre ie $j$crece.

La mayoría de los coeficientes de similaridad varían de cero a uno siendo cero cuando todo carácter de un individuo no se presenta en el otro y uno cuando todo carácter del primer individuo se presenta en el otro.

En situaciones, como en la taxonomía es muy común usar medidas de similaridad entre los puntos.

a) Datos del tipo binario

Los coeficientes de similaridad más sencillos y comúnmente más utilizados son los de variables dicotómicas que son iguales a los de disimilaridad.

Las principales medidas de similitud, son las siguientes: Jaccard, Czekanowski, Dice - Sorense ,Socal - Minchesner y Rogers

b) Datos cuantitativos

Éstas podrían abordarse convirtiéndolas en variables binarias y usar los coeficientes descritos, lo que supone obviamente una pérdida de información, lo más lógico es

considerar medidas de similaridad que puedan aplicarse directamente. Una de tales medidas es el coeficiente de correlación de Pearson.

c) Datos cualitativos

Al igual que en las disimilitudes se puede rescribir la tabla mediante columnas de ceros y uno, que se pueden considerar como variables binarias, por lo que se pueden usar los índices de similitud de los datos binarios.

d) Datos tipo mixto

Un coeficiente de similaridad sugerido por Gower (1971) es:

$$S_{ij} = \frac{\sum_{k=1}^{p} w_{ijk} S_{ijk}}{\sum_{k=1}^{p} w_{ijk}}$$

Donde:

w_{ijk}: Es el peso de cero que se asignan cuando la variable k es desconocida para uno o ambos individuos.

S_{ij}: En los datos categóricos los toman los valores de uno cuando los dos individuos tienen el mismo valor y cero en otro caso

S_{ijk}: En variables cuantitativas $1 - \frac{|x_{ik} - x_{jk}|}{R_k}$

4.1.3. Transformaciones de similaridades a disimilaridades

Es posible convertir las similaridades en disimilaridades y viceversa, mediante transformaciones sencillas. Las transformaciones más utilizadas son las que convierten las

similaridades en disimilaridades ya que éstas suelen ser las más solicitadas por la mayoría de los programas. Las transformaciones más utilizadas son:

- $\delta_{ij} = 1 - s_{ij}$ Suele utilizarse cuando las similaridades están medidas en una escala de 0 a 1 como las matrices cuyos elementos son en realidad proporciones.

- $\delta_{ij} = c - s_{ij}$ Se utilizan cuando las similaridades están medidas en una escala de 0 a c.

4.2 CLASIFICACIÓN DE CÚMULOS

Habiendo obtenido los datos en una matriz de disimilaridad o similaridad, el siguiente procedimiento es la clasificación de los conglomerados.
El conglomerado jerárquico se caracteriza por el desarrollo de una jerarquía o estructura en forma de árbol. Los no jerárquicos, consisten en generar un número fijo de grupos.

4.2.1. MÉTODOS JERÁRQUICOS

Estos métodos se suelen representar por medio de una gráfica bidimensional llamada dendograma. Se dividen en método de uniones o aglomerativo y en método divisivo:

4.2.1.1. Métodos de uniones o aglomerativo

Se parte tomando los n individuos como un cúmulo, los cúmulos más similares se agrupan entre sí y así sucesivamente hasta que la disimilaridad entre distintos cúmulos va decreciendo quedando así un solo cúmulo. Los métodos de aglomeración más comunes son:

a) **Método del vecino más cercano (liga sencilla):** En el método del vecino más cercano la distancia entre dos cúmulos es el mínimo de las distancias entre un objeto de un cúmulo y un objeto del otro.

$$d_{rp+q} = \min(d_{rp}, d_{rq})$$

b) Método del vecino más lejano (liga completa): La distancia entre dos cúmulos es el máximo de las distancias entre un objeto de un cúmulo y un objeto del otro.

$$d_{rp+q} = \max(d_{rp}, d_{rq})$$

c) Método de la distancia promedio (liga promedio): Se define como el promedio de todas las distancias entre dos elementos.

$$d_{rp+q} = \frac{1}{2} d_{rp} + \frac{1}{2} d_{rq}$$

d) Método del centroide: La distancia entre dos conglomerados es la distancia entre sus centroides.

e) Método de Ward ó de la Suma de Cuadrados: Los nuevos conglomerados se crean de tal manera de que se minimice la suma de cuadrados total de las distancias dentro de cada *cluster*.

4.2.1.2. Métodos de divisiones

Estos métodos trabajan en sentido opuesto a los anteriores. Se inicia tomando a todos los individuos en un sólo grupo, este se va particionando para tener grupos más pequeños hasta que hay el mismo número de cúmulos que individuos.

4.2.2. MÉTODOS NO JERÁRQUICOS

Este tipo de métodos consiste producir un número fijo de cúmulos, por ejemplo k; éste puede estar preestablecido o puede ser obtenido como parte del proceso. Este tipo de métodos puede iniciar con una selección inicial de grupos semilla que van a formar el centroide de los cúmulos. El método más común es el método de k –medias.

4.2.2.1. Método de k -medias

Este es un método iterativo que consiste en los siguientes procedimientos:

1. Particionar el conjunto de individuos en k grupos iniciales arbitrarios y calcular el centroide (media) de cada cúmulo.
2. Calcular la distancia (euclidiana) de cada individuo a cada uno de los k centroides. Reasigna a cada individuo al cúmulo cuya distancia al centroide sea la menor.
3. Repetir el paso 2 hasta que ningún individuo sea reasignado a un cúmulo nuevo.

4.3 ELECCIÓN DEL NÚMERO DE GRUPOS

Un aspecto importante en el análisis de conglomerados es decidir el número de éstos. Aunque no se han desarrollado formalmente pruebas estadísticas, algunas tienen una aceptación relativamente amplia. La prueba de Lee se basa en la siguiente ecuación:

$$C_p = \max\left\{\frac{|T|}{|E|}\right\}$$

La maximización se hace sobre todas las posibles particiones de los datos en dos grupos. El uso de esta prueba es limitada porque es aplicable únicamente en caso univariado Peck, Fisher y Van encuentran un intervalo de confianza para el número de conglomerados, a través de un procedimiento "bootstrap". El procedimiento consiste en definir una función criterio que dependa de dos tipos de costos, un costo asociado con el número de

conglomerados y un costo asociado con la descripción de un individuo por su respectivo conglomerado (homogeneidad del conglomerado); se busca entonces un intervalo de confianza para k, el número de conglomerados, que minimice la función criterio ó bien las consideraciones teóricas, conceptuales o prácticas pueden sugerir un número determinado de grupos.

4.4 DETERMINACIÓN DE LA CONFIANZA Y VALIDEZ

Dados los criterios generales que comprende el análisis de conglomerados, no debe aceptarse ninguna solución de agrupación sin una evaluación de su confianza y validez. Los procedimientos formales para evaluar la confianza y validez de las soluciones de agrupación son complejos y no por completo defendibles.

No obstante, los siguientes procedimientos ofrecen revisiones adecuadas de la calidad de los resultados de la agrupación:

- Métodos gráficos, como lo son: las caras de Chernoff, curvas DeAndrews, etc.
- Realizar el análisis de conglomerados con los mismos datos y utilizar distintas medidas de distancia. Comprarlos resultados con todas las medidas a fin de determinar la estabilidad de las soluciones.
- Utilizar diversos métodos de conglomerado y comparar los resultados.
- Eliminar las variables en forma aleatoria. Realizar la agrupación con base en el conjunto reducido de variables. Comparar los resultados basados en el conjunto completo con los que se obtuvieron al realizar el conglomerado.
- En el conglomerado, no jerárquico, la solución puede depender del orden de los casos en el conjunto de datos. Llevar a cabo corridas múltiples y utilizar distintos órdenes de los casos hasta que la solución se estabilice.

4.5 EJEMPLO EN *R*

4.5.1. MÉTODOS JERÁRQUICOS

Al igual que en los ejemplos anteriores se utilizará la matriz de los valores de las acciones de las empresasla cual ya esta importada en (véase ejemplo 1.1).

Antes de poner la instrucción se construye un vector empleando la función ***c( )*** (véase ejemplo 1.9) con el número de días.

```
> dias<-c(31,30,29,26,25,24,23,22,19,18,17,16,15,12,11,10,9,8,5,4,3,2,1)
```

Ejemplo 4.1

1-Para el método jerárquico se usa la instrucción ***hclus ( )*** la cual utiliza el método Lance-Williams que calcula y actualiza en cada paso la disimilaridad entre *clusters*, dentro de ésta se escribe la función ***dist ( )*** la cual calcula la matriz de distancia y ***"ave"*** para que sea por el método liga promedio**.**

```
> hc<-hclust(dist(datos),"ave")
```

Ejemplo 4.2

2-Se grafica con ***plot ( )*** (véase ejemplo 1.10)

```
> plot(hc,label=dias)
```

Ejemplo 4.3

Como resultado se tiene:

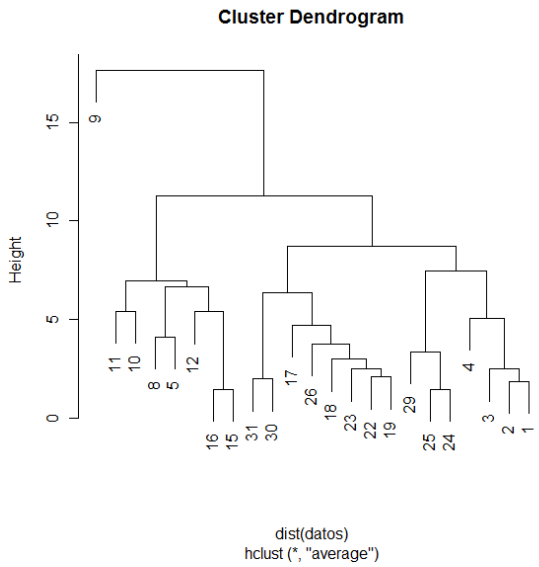

Figura 4.1

Para cambiar de método al del vecino más cercano, se cambia ***"ave"*** por ***"sin"*** como se puede ver a continuación

```
> hc<-hclust(dist(datos),"sin")
> plot(hc,label=días)
```

Ejemplo 4.4

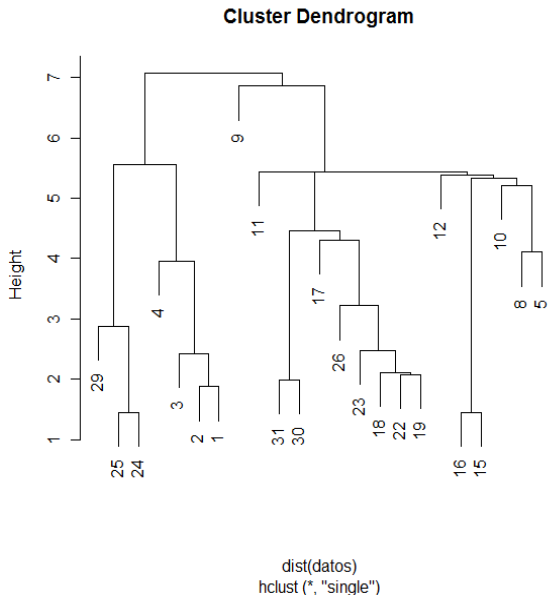

Figura 4.2

Para el caso del vecino más cercano o liga completa no es necesaria la instrucción ***R*** utiliza este método por *default*.

```
> hc<-hclust(dist(datos))
> plot(hc,label=dias)
```

Ejemplo 4.5

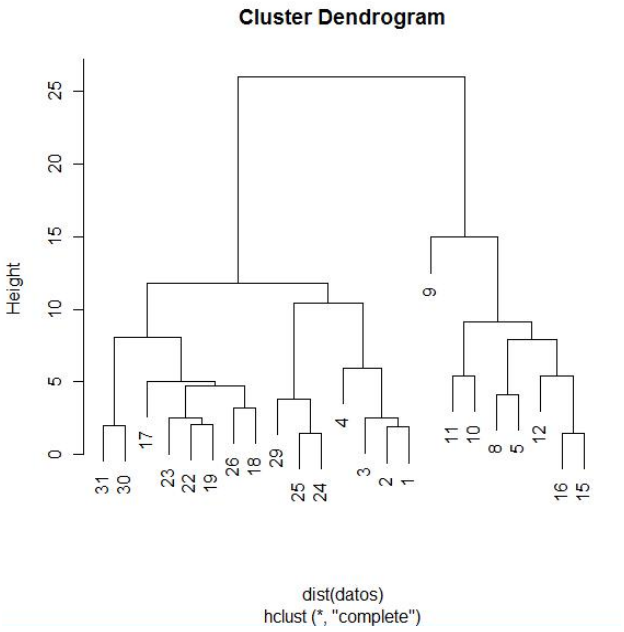

Figura 4.3

La instrucción ***identify( )*** ayuda a ver como se forman los grupos ya que esta lee la posición del puntero del gráfico cuando el botón del ratón es presionado.

A continuación busca en las coordenadas dadas en *x* e *y* para el punto más próximo al puntero.

```
> x<-identify(hc)
```

Ejemplo 4.6

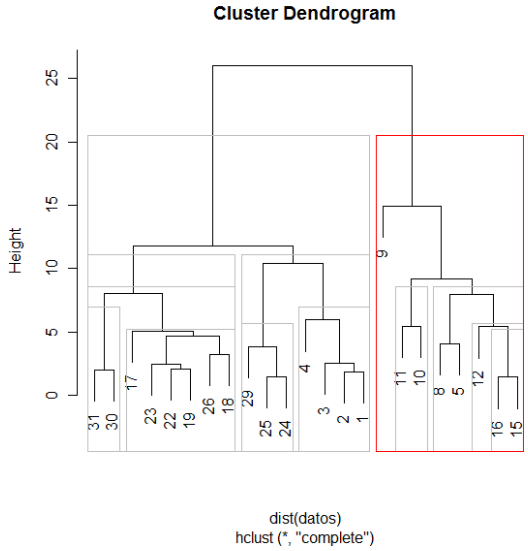

Figura 4.4

Si se quiere agrupar en un cierto número de grupos se siguen los siguen pasos:

Se empieza utilizando la instrucción ***cutree ( )*** la cual significa *cuts a tree*, es decir, corta un árbol, del resultado de hclust, esta instrucción corta, por el número deseado de los grupos o la altura de corte, en este caso lo dividimos en grupos de cinco y queda así la instrucción:

```
> memb<-cutree(hc,k=5)
```

Ejemplo 4.7

Si se hubiera querido cortar la altura la instrucción quedaría:

```
> memb<-cutree(hc,h=5)
```

Ejemplo 4.8

Después de haberlo dividió en grupos se obtiene los valores promedio por cada uno, auxiliándonos de la instrucción ***colmeans( )***

```
> cent<-NULL
> for(k in 1:5){
+ cent<-rbind(cent,colMeans(datos[2:12][memb== k, , drop=FALSE]))}
```

Ejemplo 4.9

Luego se aplica nuevamente el método jerárquico a los "nuevos" datos y se gráfica.

```
> hc1<-hclust(dist(cent),,members=table(memb))
> plot(hc1)
```

Ejemplo 4.10

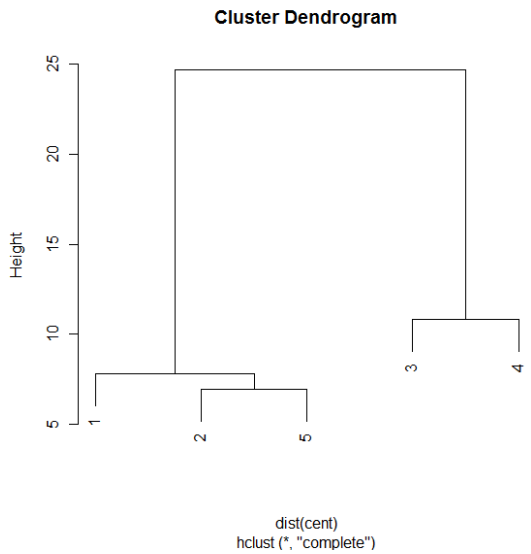

Figura 4.5

Se puede comparar el dendograma de los cinco grupos con el original con la función ***par ( )*** poniendo dentro ***mfrow*** la cual divide la ventana gráfica según el número que se le escribe en la instrucción ***c ( )***,en este caso una fila y dos columnas.

```
> opar <- par(mfrow = c(1, 2))
> plot(hc, main = "Datos originales")
> plot(hc1, main = "Sólo 5 clusters")
> par(opar)
```

Ejemplo 4.11

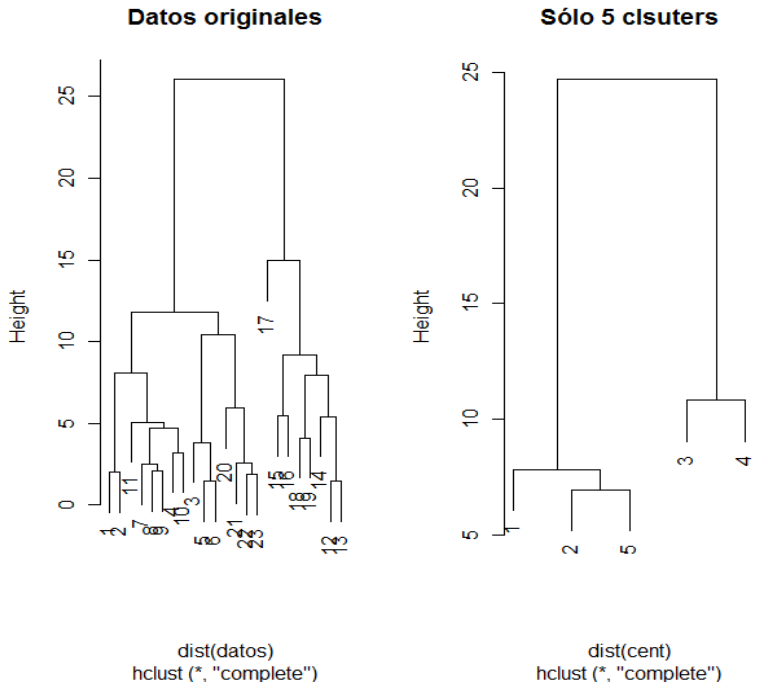

Figura 4.6

4.5.2. MÉTODOS NO JERÁRQUICOS

4.5.2.1. k-medias

Utilizando la instrucción ***kmeans( )*** , se aplica el método de las k -medias en este caso para 3 grupos:

```
> datos.kmeans<-kmeans(datos[2:12],3)
```

Ejemplo 4.12

Se compara el método jerárquico con el k -medias, primero se obtienen 3 grupos por el método jerárquico utilizando la instrucción ***cutree ( )*** (véase ejemplo 4.6) y después se unen las columnas con la instrucción ***cbind( ).***

```
> memb<-cutree(hc,k=3)
> cbind(memb,datos.kmeans$cluster)
      memb
 [1,]    1 1
 [2,]    1 1
 [3,]    1 2
 [4,]    1 1
 [5,]    1 2
 [6,]    1 2
 [7,]    1 1
 [8,]    1 1
 [9,]    1 1
[10,]    1 1
[11,]    1 1
[12,]    2 3
[13,]    2 3
[14,]    2 3
[15,]    2 3
[16,]    2 3
[17,]    3 3
[18,]    2 3
[19,]    2 3
[20,]    1 2
[21,]    1 2
[22,]    1 2
[23,]    1 2
```

Ejemplo 4.13

Al comparar los métodos se ve que el método jerárquico (columna izquierda) y el k –medias (columna derecha), el día 31 (correspondiente al primer renglón) lo puso en el primer *cluster* el método jerárquico, al igual que en el método k –medias, así sucesivamente se puede ver la comparación para cada uno de los 23 días.

Otra forma de realizar el método jerárquico es empleando la función ***daisy( )*** que ayuda a calcular la disimilaridad la diferencia entre ésta con ***dist( )*** es que con la primera se puede usar para las variables de tipo mixtas pero antes se debe de cargar la librería ***cluster.***

```
> library(cluster)
> hc<-hclust(daisy(datos),"ave")
> plot(hc)
```

Ejemplo 4.14

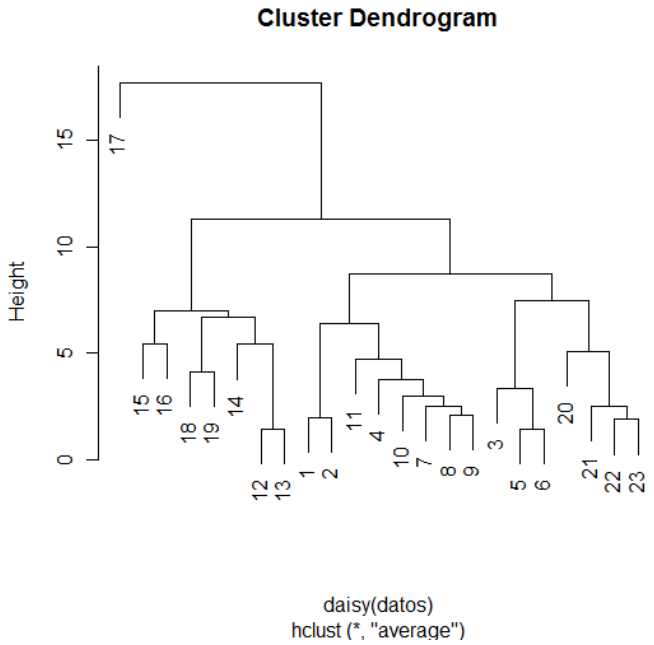

Figura 4.15

Otro gráfico aparte del dendograma es el llamado bandera (*banner*), ésta grafica la misma información que en el dendograma.

```
> library(cluster)
> agn1 <- agnes(daisy(datos[2:12]), metric = "average", stand = TRUE)
> plot(agn1)
```

Ejemplo 4.16

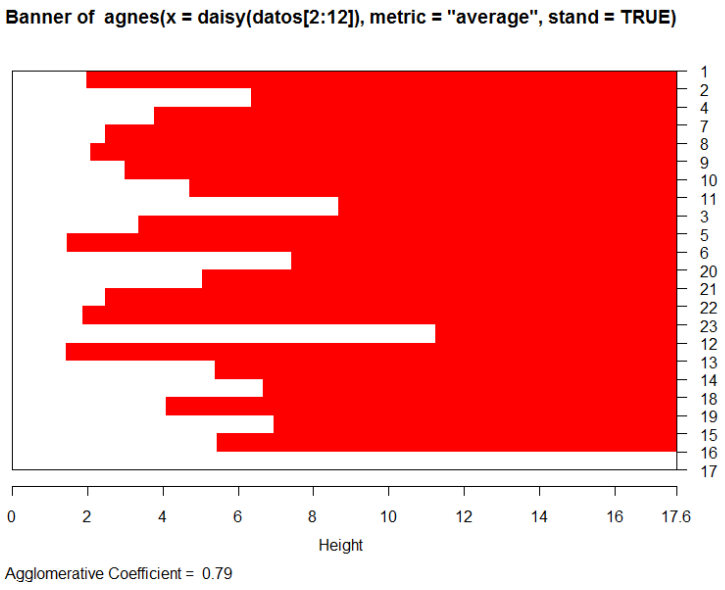

Figura 4.8

5. ESCALAMIENTO MULTIDIMENSIONAL (MDS)

El escalamiento multidimensional (MDS) es una herramienta matemática que utiliza proximidades entre las observaciones para producir su representación espacial para ello se auxilia de una matriz $(n\ x\ n)$ de disimilaridades o distancias, $D_{n\times n}$.[15]

El objetivo es encontrar un conjunto de puntos en una menor dimensión (por lo general dos dimensiones). De este modo el escalamiento multidimensional se centra en el problema de construir una configuración de n puntos en un espacio k -dimensional, generalmente el espacio euclidiano, usando la información acerca de las distancias entre los n objetos.

La medida en la cual se basa el escalamiento multidimensional para poder establecer la cercanía o lejanía entre las observaciones, objetos o estímulos es la *medida de proximidad*.

2.1. MODELO

El MDS empieza con una matriz de proximidades, $\Delta \in M_{n\times n}$, donde n es el número de estímulos. Cada elemento δ_{ij} de Δ representa la proximidad entre el estímulo i y el estímulo j.

$$\Delta = \begin{pmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n1} & \delta_{n2} & \cdots & \delta_{nn} \end{pmatrix}$$

[15] Hardle Wolfgang, *Applied Multivariate Statisticval*, New York ,Springer , pág 205.

A partir de esta matriz de proximidades el MDS proporciona como salida una matriz, $X \in M_{n \times n}$ donde n, al igual que antes, es el número de estímulos, y m es elnúmero de dimensiones. Cada valor x_{ij} representa la coordenada del estímulo i en ladimensión j.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

Una vez obtenida la matriz X se puede calcular la distancia existente entre dos estímulos cualesquiera i y j, simplemente aplicando la fórmula general de la distancia:

$$d_{ij} = \left\{ \sum_{k=1}^{p} (x_{ik} - x_{jk})^p \right\}^{1/p}$$

Donde:

- d_{ij} es la distancia entre los estímulos i y j.
- x_{ia} y x_{ja} son, respectivamente, las coordenadas de los estímulos i y j en la dimensión *a-ésima*.
- p es un valor que puede oscilar entre uno e infinito (en el caso de la distancia euclídeana su valor es 2)

Dadas las distancias d_{ij} se construye la matriz cuadrada de distancias D entre n estímulos.

$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{pmatrix}$$

La solución proporcionada por el MDS debe ser de tal modo que haya la máxima correspondencia entre la matriz de proximidades inicial Δ y la matriz de distancias obtenidas D.

5.3. MODELO DE ESCALAMIENTO MÉTRICO

Dada la matriz $\Delta = \delta_{ij}$ de (di) similaridades entre los pares de n objetos, se trata de encontrar n puntos $P_i(i = 1,2,..,n)$ en un espacio de dimensión mínima k, donde k es menor que n, con una matriz de coordenadas X y con distancias $D=(d_{ij})$, por tanto se busca hallar las coordenadas de estos puntos en un espacio de dimensión k (R^k, d).Para ello, el primer paso consiste en transformar las (di)similaridades δ_{ij} en distancias absolutas que cumplan la desigualdad triangular en R^n, $d_{ij} \leq d_{ir} + d_{rj}$.

Esto parte de la idea que las distancias son una función de las proximidades, es decir, $d_{ij} = f(\delta_{ij})$ la cual es de tipo lineal. A partir de D se construye una matriz $A=(a_{ij})$ de coeficientes de asociación entre objetos con elementos definidos como:

$$a_{ii} = -(\frac{1}{2})d_{ii}^2=0$$

$$a_{ij} = -(\frac{1}{2})d_{ij}^2$$

A continuación se construye una matriz B simétrica de productos escalares, cuyos elementos vienen dados por:

$$b_{ij} = a_{ij} - a_{i.} - a_{.j} + a_{..}$$

Donde:

$$a_{i.} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}$$

$$a_{.j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij}$$

$$a_{..} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}$$

$$b_{ij} = x_i' x_j$$

El producto de la matriz $B = (b_{ij})$ se puede expresar como:

$$B = X \Lambda X'$$

Donde$X = (x_1, \ldots, x_n)'$ es la matriz de coordenadas de cada uno de los *n* objetos en cada una de las *p* dimensiones. La matriz B es simétrica, definida semipositiva y de rango *p;* por lo que sus eigenvalores son no-negativos y *n-p* ceros eigenvalores cualquier factorización permite transformar B en $X \Lambda X'$

5.4. MODELO DE ESCALAMIENTO NO MÉTRICO

El objetivo del MDS no métrico tiene el mismo objetivo que el métrico; sin embargo el MDS no métrico tiene una relación menos rígida entre las proximidades y las distancias suponiendo una relación mónotona creciente entre ambas, es decir, si:

$$\delta_{ij} < \delta_{ki} \rightarrow d_{ij} \leq d_{ki}$$

En el MDS no métrico se desea a encontrar una configuración de puntos en un espacio de menor dimensión tal que la distancia Euclideana es $f(\delta_{ij})$, donde f es una función creciente.

El método más común es el algoritmo iterativo Shepard-Kruskal. En el primer paso se calcula la distancia Euclideana d_{ij} para una posición inicial los de puntos. En el segundo, se obtienen las disimilaridades d_{ij} que son las transformaciones de las proximidades por f, después se calcula el $Stress$ que Kruskal definió como:

$$Stress = \sqrt{\frac{\sum_{i,j}(f(\delta_{ij}) - d_{ij})^2}{\sum_{i,j} d_{ij}^2}}$$

También se suele utilizar una variante del *Stress* que se denomina *S-Stress*, definida como:

$$S - Stress = \sqrt{\frac{\sum_{i,j}\left(f(\delta_{ij})^2 - d_{ij}^2\right)^2}{\sum_{i,j} \quad (d_{ij}^2)^2}}$$

El siguiente paso basado en las diferencias entre $f(\delta_{ij})$y d_{ij}se define la nueva posición de puntos

$$x'_{ik} = x_{ik} + \frac{\propto}{n-1}\sum_{\substack{j=1 \\ j \neq i}}^{n}(1 - f(\delta_{ij})/d_{ij})/(x_{jk} - x_{ik})$$

Donde $\propto$ determina el ancho de iteración. En el quinto paso se ve el valor del Stress Kruskal (1964) sugiere las siguientes interpretaciones del *Stress*:

- 0.2 Pobre
- 0.1 Aceptable
- 0.05 Bueno

- 0.025 Muy bueno
- 0.0 Excelente

En caso de que el valor sea bajo se repité el proceso, en caso contrario se detiene

5.4. EJEMPLO EN *R*:

5.4.1 ESCALAMIENTO MÉTRICO

Para este ejemplo se utilizará la matriz de distancias del Estado de México, para ello se importan los datos, se puede utilizar la instrucción ***read.delim ("clipboard")***
en caso de tenerse la base en Excel (como es en este),primero se selecciona la tabla en Excel y se copia ,después en ***R*** se pone el nombre que se le va a dar a la tabla y la instrucción tal como se ve en el ejemplo en la siguiente imagen:

```
> datos<-read.delim("clipboard")
> datos
  Amecameca CuidaddeMéxico Naucalpan Tlanepantla Toluca ValledeBravo
1         0             70        78          88    130          217
2        70              0        15          16     70          157
3        78             15         0          10     70          157
4        88             16        10           0     80          167
5       130             70        70          80      0           87
6       217            157       157         167     87            0
```

Ejemplo 5.1

1-En primera instancia se obtiene la matriz de coordenadas, es decir *X*, con la función ***cmdscale( )***

```
> cmdscale(datos)
           [,1]       [,2]
[1,] -85.82420  36.684279
[2,] -27.51015  -5.341521
[3,] -25.79554 -14.350062
[4,] -31.78748 -32.997423
[5,]  41.33385   5.570202
[6,] 129.58351  10.434526
```

Ejemplo 5.2

Esta se gráfica con la instrucción ***plot( )*** (veáse ejemplo 1.10) y se le asignan los nombres a los municipios en la gráfica con la función ***text( )*** (veáse ejemplo 1.11).

```
> plot(cmdscale(datos),type='n')
> text(cmdscale(datos),labels=c("Amecameca","CuidaddeMéxico","Naucalpan","Tlanepantla","Toluca","Valle de Bravo"))
```

Ejemplo 5.3

Lo cual nos da la siguiente imagen:

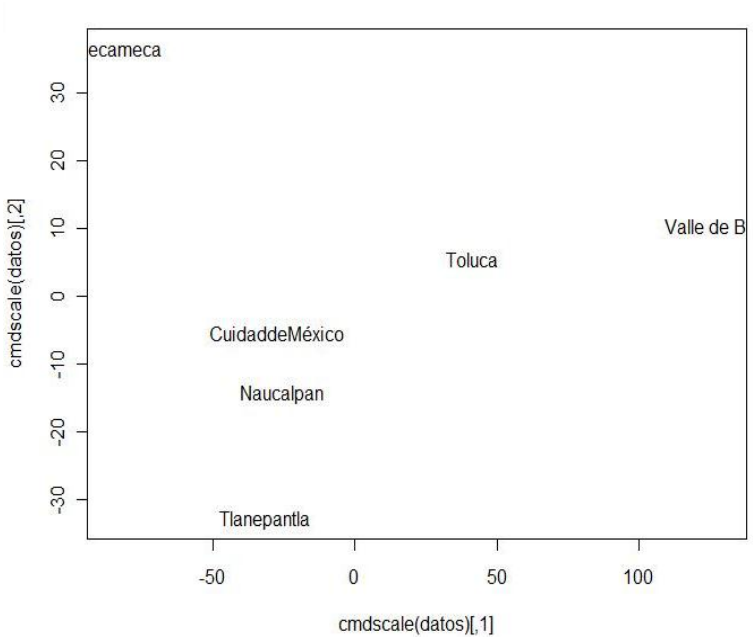

Figura 5.1

5.4.2 ESCALAMIENTO NO MÉTRICO

Para el escalamiento no métrico es necesario meter la matriz con la instrucción ***matrix ( )*** **,**poniendo primero los valores de la matriz con la instrucción ***c ( ) ,*** después el número de filas,de columnas, ***byrow( )*** (lee por filas) y los nombres de las filas y columnas con ***dimnames( )***

```
>mdat<-
matrix(c(0,70,78,88,130,217,70,0,15,16,70,157,78,15,0,10,70,157,88,16,10,0,80,167,130,70,70,80,0,87,217,157,157,167,87,0), nrow = 6, ncol=6, byrow=TRUE, dimnames = list(c("1", "2","3","4","5","6"), c("1", "2","3","4","5","6")))
```

```
> mdat
    1   2   3   4   5   6
1   0  70  78  88 130 217
2  70   0  15  16  70 157
3  78  15   0  10  70 157
4  88  16  10   0  80 167
5 130  70  70  80   0  87
6 217 157 157 167  87   0
```

Ejemplo 5.4

Ya teniendo la matriz ponemos la instrucción ***isoMDS( )*** del escalamiento no métrico también es necesario cargar la librería ***MASS*** (véase ejemplo 3.1)

```
> mdatnometrico<-isoMDS(mdat)
initial  value 1.750260
iter   5 value 0.203818
final  value 0.000000
converged
```

Ejemplo 5.5

Se verifica el stress agregándole al nombre ***$stress***

```
> mdatnometrico$stress
[1] 2.054285e-14
```

Ejemplo 5.6

Este valor se encuentra dentro del rango de excelente del criterio de Kruskal .Se gráfica con la instrucción ***plot( )*** (veáse ejemplo 1.10) y se le asignan los nombres a los municipios en la gráfica con la función ***text( )*** (veáse ejemplo 1.11) como se ve a continuación**.**

```
> plot(mdatnometrico$points,type="n")
> text(mdatnometrico$points,labels=c("Amecameca","Cd.de México","Naucalpan","Tlanenpantla","Toluca","ValledeBravo"))
```

Ejemplo 5.7

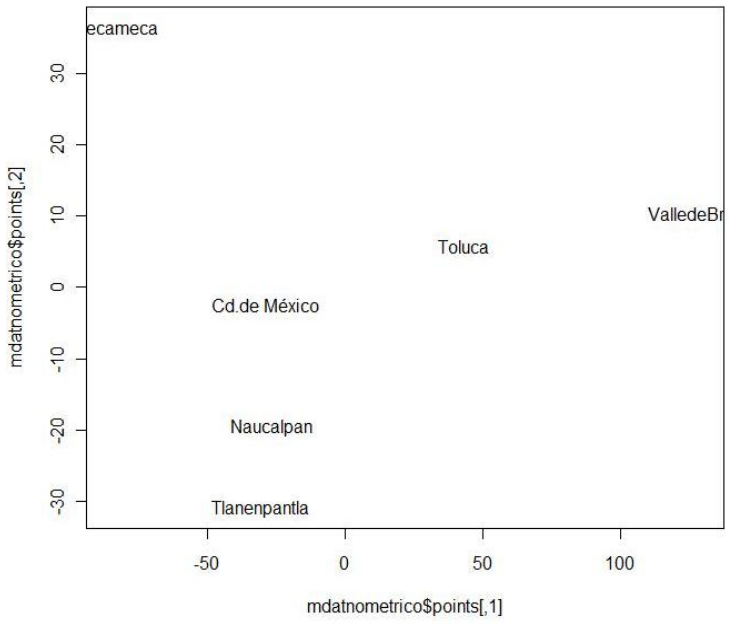

Figura 5.2

Ésta es otra forma de hacer de escalamiento no métrico, empleando ahora la instrucción ***sammon ( )***

```
> mdatsammon<-sammon(mdat)
Initial stress        : 0.01460
stress after  10 iters: 0.00095, magic = 0.500
stress after  20 iters: 0.00094, magic = 0.500
```

Ejemplo 5.8

Se gráfica con la instrucción ***plot( )*** (veáse ejemplo 1.10) y se le asignan los nombres a los municipios en la gráfica con la función ***text( )*** (veáse ejemplo 1.11) como se presenta a continuación**.**

```
> plot(mdatsammon$points,type="n")
> text(mdatsammon$points,labels=c("Amecameca","Cd.de México","Naucalpan","Tlanenpantla","Toluca","ValledeBravo"))
```

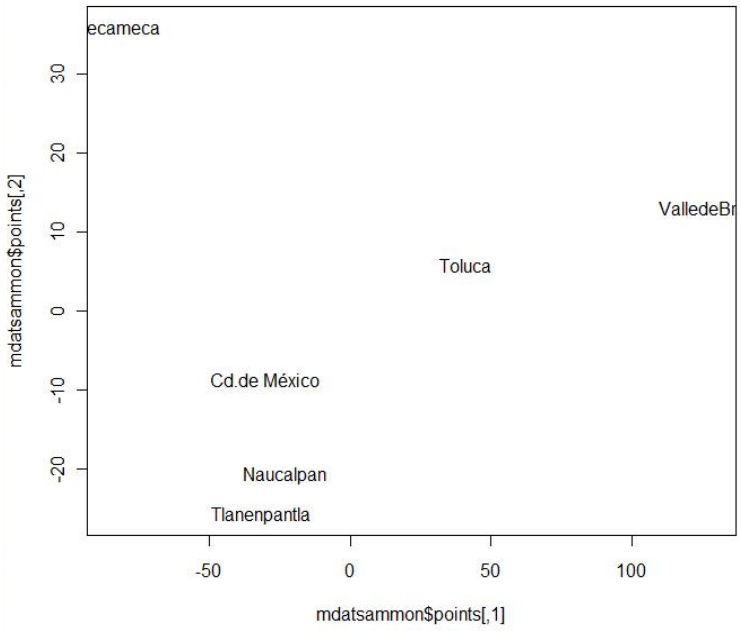

Figura 5.3

Se comparan los resultados con el mapa del Estado de México

Figura 5.4

CONCLUSIONES

A lo largo de este trabajo se expusieron los procedimientos de los temas de Análisis Multivariado para la herramienta estadística ***R***, los cuales se pueden realizar de manera sencilla, ya que las instrucciones no son complejas lo que es una de las ventajas para trabajar con ***R*** en este análisis.

Cabe subrayar que las funciones que se mencionaron en el trabajo no son las únicas que existen para los procedimientos, pero son utilizadas con mayor frecuencia.

Asimismo, se demostró que este software puede ser aplicado en el área de finanzas para la reducción, predicción y agrupamiento de información.

Si bien una desventaja de ***R*** respecto a otros paquetes estadísticos radica que en éstos no es necesario escribir instrucciones, solamente con dar click en una serie de botones se puede obtener el resultado deseado, en este caso la ventaja de ***R*** es que al momento de introducir las instrucciones se puede adquirir un mejor entendimiento de los procesos. Además ***R*** no tiene algún costo y es constantemente actualizado, lo cual beneficia a los usuarios, caso que en los softwares que no son libres es necesario esperar meses ó años para que se tenga una nueva versión. Otra ventaja de trabajar con éste es que es pueden conectar gestores de datos a éste como Oracle y SQL**,** los cuales en la actualidad son muy utilizados.

Un beneficio de haber trabajo con ***R*** para este análisis es la alta calidad de los gráficos ya que permite visualizar mejor la agrupación de los datos para una mejor interpretación de resultados.

En los métodos de Componentes Principales y Análisis Factorial ***R*** fue de gran utilidad para la reducción de información arrojando gráficas y valores que describían la información de las doce acciones en sólo cuatro componentes ó factores según el método.

En el caso del Análisis Discriminante, ***R*** mostró los valores de los ejes discriminantes, los coeficientes de los discriminantes lineales y algunos atributos de los grupos como las

probabilidades *a priori* y sus medias. Igualmente al agregarle información de nuevas fechas hizo la predicción a que semana pertenecían.

En el apartado de Conglomerados, ***R*** pudo agrupar las fechas mediante diferentes los métodos y hacer una comparación de sus resultados entre estos métodos y graficar la información de dos formas dendograma y bandera.

Para el Escalamiento Multidimensional ***R*** arrojó para los métodos métrico y no métrico sus respectivas gráficas lo cual mostró la cercanía ó lejanía de lugares lo que puede permite resolver algún problema como es el de minimizar tiempos rutas.

Cabe destacar que durante el proceso de investigación se descubrió que no existen muchas fuentes en español por lo que a veces hay que hacer un esfuerzo de traducción; las que se encontraron en español muchas son de España, las cuales tiene ejemplos de otras áreas que no son afines a la carrera. Al igual que la información se encuentran dispersa por lo que se realizó un esfuerzo de recopilación.

También en el transcurso del tiempo en que se elaboró esta tesina se descubrió que es importante para cualquier actuario que se tenga noción de algún paquete estadístico ya que algún día le pueda facilitar el quehacer cotidiano.

ANEXO 1

Tabla de acciones(Yahoo Finanzas México)[16]

Día	Axtel	Cemex	Geo	Soriana	Compartamos	Gruma	Gfamsa	Kfol	Bachoco	Cmr	Liverpool	Maseca
31/08/2011	6	6.69	23.71	30.98	74	23.17	13.62	121.21	22.77	3.05	98	13.95
30/08/2011	5.93	6.62	23.7	30.01	74.01	23.11	13.2	120.23	22.5	3.05	98	12.62
29/08/2011	5.8	6.34	23.25	30	80.25	22.23	13.25	117.11	22.1	3.05	94.99	12.62
26/08/2011	5.69	6.37	23.56	29.7	73.2	21.55	13.13	119.82	21.85	3.05	94	12.62
25/08/2011	5.87	6.57	24.02	30.94	81	22.89	13.8	120.39	21.79	3.05	93.99	12.62
24/08/2011	5.82	6.4	23.85	30.89	80	22.98	13.5	119.44	21.93	3.2	94	12.65
23/08/2011	5.6	6.13	23.42	29.01	73.07	23	12.85	116.8	22.6	3.06	94.45	12.7
22/08/2011	5.58	6.28	23.63	28.46	73.07	22	12.81	115.72	22.7	3.05	93	13.95
19/08/2011	5.66	6.29	24.32	28.2	73.05	22.54	13.01	117.55	23	3.05	93	13.95
18/08/2011	5.75	6.6	24.5	28.25	73	22.25	13.91	118.42	22.8	2.8	92.2	12.57
17/08/2011	5.66	7.07	25.5	28.94	70	22	14.45	117.98	22.75	2.9	95.01	14
16/08/2011	5.4	6.66	24.9	28.61	68.05	22.62	14.65	113.17	21.6	3.2	94.99	13.96
15/08/2011	5.23	6.49	25.35	28.75	68.02	23	14.5	112.99	21.39	3.2	94.53	12.8
12/08/2011	5.3	6.52	26.52	29.19	73	23	14.4	111.94	21.3	3.05	94.85	13.96
11/08/2011	5.14	6.26	25.3	28.06	65	22.21	13.38	108.96	21	3	92.29	13.92
10/08/2011	5.1	6.38	24.74	28.2	70	21.38	13	107.4	21.08	3.95	92.28	13.9
09/08/2011	5.27	6.37	23.86	29	60.55	20.8	12.8	104.61	21.2	3	94	13.9
08/08/2011	5.8	6.8	22.11	30.51	72.5	22.1	12.93	108.25	22.79	3	94	13.85
05/08/2011	6	7.01	21.53	30.39	69.38	22.67	13.26	110.41	23	3.2	95.22	13.93
04/08/2011	6.55	7.3	22.3	31.77	77.2	24.32	14.09	112.9	23	3.2	99.32	13.94
03/08/2011	6.6	7.55	22.55	32.15	78	25	14.47	114.52	23.2	3.35	96	13.29
02/08/2011	6.9	8	23.38	33	77.5	25.2	14.86	115.55	23.05	3.35	94.46	13.79
01/08/2011	6.85	8.4	24	32.55	77.2	25	15.6	114.22	23	3.35	95	13.94

Tabla de distancias del Estado de México[17]

Lugar	Amecameca	Cuidad de México	Naucalpan	Tlanepantla	Toluca	Valle de Bravo
Amecameca	0	70	78	88	130	217
CuidaddeMéxico	70	0	15	16	70	157
Naucalpan	78	15	0	10	70	157
Tlanepantla	88	16	10	0	80	167
Toluca	130	70	70	80	0	87
ValledeBravo	217	157	157	167	87	0

[16]Esta base de datos fue construida con la información encontrada en Yahoo Finanzas México.http://mx.finance.yahoo.com/.revisado el 14 de Septiembre de 2011

[17]http://www.visitehotelesdemexico.com/Mexico/Estado_de_Mexico/distancias.aspx

FUENTES CONSULTADAS

BIBLIOGRÁFICAS

Cuadras Avellana, Carlos, *Métodos de Análisis Multivariante.* Barcelona,Universidad de Barcelona, 1981

Ferrán, M., *SPSS para Windows, Programación y Análisis Estadístico,*España, McGraw Hill, 1996.

Álvarez Caceres, Rafael, *Estadística multivariante y no paramétrica con SPSS,* Primera edición,Madrid,Díaz de Santos, 1995.

Rodríguez Jaime, María José ,*Modelo sociodemográficos:Atlas social de la cuidad de Alicate,* España, Biblioteca Miguel de Cervantes, 2001.

Wolfgang, Härdle, *Applied Multivariate Statistical Analysis,* Berlín, Springer, 2007.

Rivas, Teresa, *Relación entre escalamiento multidimensional métrico y análisis multivariante* 2, Malaga , Psicothema, 1991, Vol. 3.

Pérez Carbonel, Amparo, *Modelos complementarios al Análisis Factorial en la construcción de escalas ordinales: un ejemplo aplicado a la medida del Clima Social,* 354, España ,Ministrio de Educación ,Cultura y Deporte, 2011.

Hair ,Joseph, *Anáisis Multivariante,* España, Prentice Hall, 2010.

Alvin, Rencher, *Methods of Multivariate,* New York , Wiley-Interscience, 2002.

J.P Marque de Sá, *Applied Statistics Using SPSS,STATISTICA,MATLAB and R* ,New York, Springer, 2007.

Everitt, Brian.*An Introduccion to Applied Multivariate Analysis with R,* New York, Springer, 2011.

Peña,Daniel.*Análisis de datos multivariantes,* España , McGraw-Hill, 2002.

S., Everitt Brain.*A handbook Statisticla Analyses,*London, CRC Taylor & Francis group, 2010.

TESIS

Polo Miranda,Carlos,*Estadística multivariante*, México, UNAM, 2010.

López Gonzaléz,Fabiola,*Lenguaje R: Un complemento libre para las asignaturas de estadística*,México, UNAM,2008.

SITIOS DE INTERNET

http://allman.rhon.itam.mx/~lnieto/index_archivos/Modulo61.pdf. Revisado en 13 de mayo de 2011.

http://www.5campus.com/leccion/discri.Revisado el 14 de mayo de 2011

http://www.uv.es/asepuma/XII/comunica/bernal_martinez_sanchez.pdf. Revisado el 25 de mayo del 2011.

http://halweb.uc3m.es/esp/Personal/pers

http://halweb.uc3m.es/esp/Personal/personas/agrane/ficheros_docencia/MULTIVARIANT/tema8.pdf. revisado 23 de julio de 2011.

.http://ocw.uc3m.es/estadistica/aprendizaje-del-software-estadistico-r-un-entorno-para-simulacion-y-computacion-estadistica/clasificacion-de-datos-multivariantes-analisis-discriminant.revisado el 01 de mayo de 2011

The R Project for Statistical Computing .http://www.r-project.org/ revisado el 01 de julio del 2012.