UNIVERSIDAD & SALAMANCA

DEPARTAMENTO DE ESTADÍSTICA

CONTRIBUCIONES AL ANÁLISIS MULTIVARIANTE NO LINEAL

Guillermo Correa Londoño

2008

CONTRIBUCIONES AL ANÁLISIS MULTIVARIANTE NO LINEAL

Memoria que, para optar al Grado de Doctor por el Departamento de Estadistica de la Universidad de Salamanca, presenta:

Guillermo Correa Londoño

Salamanca

2008

Universidad de Salamanca
Departamento de Estadistica

Mª PURIFICACIÓN GALINDO VILLARDÓN

y

CARMELO ANTONIO ÁVILA ZARZA

Profesores Titulares del Departamento de Estadistica de la Universidad de Salamanca

CERTIFICAN: Que **Don Guillermo Antonio Correa Londoño**, ha realizado, en el Departamento de Estadistica de la Universidad de Salamanca, bajo su direccion, el trabajo que, para optar al Grado de Doctor, presenta con el titulo: "***Contribuciones al Análisis Multivariante no Lineal***"; y para que conste, firman el presente certificado en Salamanca, en Septiembre de 2008.

A la memoria de mi madre, Luz Gabriela
A mi compañera de siempre, Luz Marina

AGRADECIMIENTOS

A la **Universidad Nacional de Colombia**, por el esfuerzo que, desde las instancias
administrativas hasta mis colegas del Departamento de Ciencias Agronomicas, han
realizado en pro de mi formacion.
Al **Grupo Santander** por la financiacion de este programa doctoral a traves de una de
las becas *Universidad de Salamanca-Grupo Santander*, destinadas a estudiantes
Iberoamericanos.
A la doctora **Mª. Purificación Galindo Villardón**, directora de este trabajo, por la
calida acogida que desde el primer momento me brindo, por su constante apoyo, tanto
en la parte academica como en la personal, y por no dejar que su talento opaque su
calidad humana.
Al doctor **Carmelo Ávila Zarza**, codirector de este trabajo, por todas sus aportaciones
tecnicas al mismo, pero sobre todo por haberme hablado de fotografia, de politica, de
religion, de turismo, de iPhones y hasta de la Pasion segun San Mateo; en resumen, por
haberme brindado su amistad.
Al doctor **José Luis Vicente Villardón** por sus valiosos aportes en diversos aspectos
matematicos y computacionales.
Al cardiologo **Maximiliano Diego Domínguez** por confiarme la base de datos para la
aplicacion practica.
Al cardiologo **Víctor Hugo Ramírez Castro**, quien ademas de haberme dado una
valiosa asesoria en el campo tecnico de este trabajo, me brindo su amistad y soporte en
los momentos dificiles.
A todo el **Personal del Departamento de Estadística de la Universidad de**

Salamanca por demostrar que la calidad academica no tiene que ser excluyente con la
calidad humana.

A **Luz Estela Sánchez Herrera**, quien, gracias a su formacion matematica, me ayudo a
salir de numerosos escollos.

A mi esposa, **Luz Marina**, con quien he compartido tantos anos de mi vida, cada uno
mas excitante que el anterior, por ser la gran motivacion detras de todo mi quehacer.

A **mi familia**, por haberme apoyado irrestrictamente en este proyecto.

A mis **amigos y amigas**. A los que compartieron conmigo durante el periodo en que se
desarrollo este trabajo, por haberme brindado esos momentos tan maravillosos de
esparcimiento, y tambien a los que por la distancia no pudieron hacerlo, por esperarme
pacientemente.

A mi madre, **Luz Gabriela**, quien, con su ejemplo, me enseno que la vida habia que
vivirla con buen talante, sin importar las adversidades. Su memoria siempre me
inspirara una sonrisa.

Cuantifica, siempre que sea posible.

Sexta herramienta para la detección de falacias, tomado de '*El Mundo y Sus Demonios*'.

Carl Sagan

Índice

Pag.

IntroducciónIntroducción

Aunque las primeras pruebas documentales del interes por cuantificar lo cualitativo
datan de principios del siglo XX (YULE, 1910, citado por YOUNG, 1981), es de
suponer que tal inquietud nacio casi a la par del reconocimiento de estas dos facetas de
la informacion. En la decada del 40, Guttman establece las bases de uno de los primeros
metodos de cuantificacion (GUTTMAN, 1941), el cual, ademas de haber gozado de

amplia difusion, ha sido uno de los de mayor continuidad, dando lugar a numerosos
desarrollos, hasta llegar a sistemas tan depurados como el propuesto por la escuela
holandesa de escalamiento de datos (GIFI 1981, 1990). Aun asi, la cuantificacion es un
topico en el que siguen quedando cuestiones por resolver. En tal sentido, se ha
desarrollado este trabajo como una aportacion a dicha tematica.
Hemos elaborado nuestras propuestas con base en el siguiente **plan general**. En el
Capítulo 1, a forma de introduccion, presentamos algunos conceptos basicos y
definiciones necesarias para la adecuada comprension de la problematica de la
cuantificacion, la cual se desarrolla con mayor profundidad en los capitulos
subsiguientes. **El Capítulo 2** esta centrado en el sistema Gifi de analisis multivariante
no lineal, el cual constituye el cuerpo de doctrina mas depurado sobre el analisis
multivariante no lineal, uno de cuyos aspectos es la cuantificacion.
En el **Capítulo 3** se pone en evidencia el hecho de que al usar las tecnicas del sistema
Gifi para la cuantificacion de variables cualitativas, surgen diferentes conjuntos de
cuantificaciones, sin que exista ningun criterio para decidir cual de los posibles
conjuntos de cuantificaciones es el mas adecuado. En tal sentido, en este capitulo
aparece la primera aportacion de este trabajo, pues el problema anotado, al parecer, no
habia sido detectado o, por lo menos, no habia sido discutido anteriormente.
El **Capítulo 4** ofrece una solucion al problema planteado en el capitulo precedente,
partiendo para ello de un criterio de cuantificacion optima independiente del modelo de
analisis. Tanto el criterio de cuantificacion optima como el algoritmo que permite elegir

el conjunto de cuantificaciones optimas a partir de varios posibles conjuntos de
cuantificaciones son desarrollos originales, constituyendo, por tanto, la segunda
aportacion de este trabajo.

En el **Capítulo 5** se desarrolla una generalizacion de la solucion planteada en el capitulo
precedente para la obtencion de cuantificaciones optimas. Dicha generalizacion hace
posible obtener cuantificaciones optimas directamente de los datos originales, sin que se
requieran cuantificaciones previas generadas mediante tecnicas del sistema Gifi ni de
ningun otro sistema. Para tal efecto, se implementa un modulo computacional
desarrollado en MATLAB, el cual asigna cuantificaciones optimas, tomando en
consideracion el nivel de escalamiento de cada una de las variables. **La rutina
computacional producto de los desarrollos teóricos de este capítulo constituye una
herramienta integral para la generación de cuantificaciones óptimas de un sistema
multivariante mixto**. Puesto que tanto los algoritmos que conforman la rutina como la
base teorica que los sustenta son originales, este capitulo en su conjunto representa la
mayor aportacion de este trabajo.
En el **Capítulo 6** se ilustra el uso de la propuesta desarrollada en el capitulo precedente,
a traves de una aplicacion practica, consistente en el analisis de una base de datos de
6.965 pacientes con cardiopatia isquemica o sintomatologia afin, con 53 variables
medidas en diferentes escalas y con 53,12 % de informacion faltante. En este capitulo,
se utiliza el metodo de cuantificacion propuesto en el capitulo precedente como un paso
de adecuacion de la base de datos para su posterior analisis mediante una tecnica lineal.

En este caso, hemos utilizado la metodologia de las representaciones Biplot como
tecnica de exploracion final. Los resultados que se desprenden del analisis de esta base
de datos son otra de las aportaciones de este trabajo.
Con el fin de llevar el analisis de la base de datos hasta el final, aprovechando toda la
informacion disponible, en el **Capítulo 6** se desarrolla una propuesta para la elaboracion
de representaciones Biplot de matrices con datos faltantes. Esta propuesta constituye
tambien una aportacion de este trabajo, por la originalidad tanto de su desarrollo teorico
como de las rutinas computacionales que automatizan su implementacion. No obstante,
hemos optado por mantener esta propuesta como un apartado del Capitulo 6, por
concebirla como un complemento al nucleo central de este trabajo, que es la
cuantificacion optima.
El uso combinado de la propuesta de cuantificación óptima desarrollada en el
Capítulo 5, con el de las representaciones Biplot de matrices con datos faltantes,

desarrollado en el Capítulo 6, permite explorar grandes sistemas multivariantes
mixtos con información faltante, aprovechando toda la información disponible, tal
y como se ilustra en la aplicacion practica del Capitulo 6.
Para facilitar el uso de las rutinas computacionales elaboradas en este trabajo, se ha
creado una interfaz grafica para cada una de ellas. A traves de tales interfaces, el usuario
puede ingresar la informacion de entrada, especificar las opciones de los analisis y
acceder a las correspondientes ayudas. Asimismo, con el fin de integrar los diferentes
modulos computacionales y de proporcionar un acceso sencillo a los mismos, se ha

creado una interfaz de inicio comun que permite derivar hacia cualquiera de los
modulos. Luego, basta con realizar una instalacion conjunta de todos los modulos y
acceder a cualquiera de ellos a traves del modulo general. Tanto el modulo de entrada
como los modulos especificos cuentan con ayudas que van guiando al usuario a traves
del proceso.
Cerramos esta memoria presentando las conclusiones y las referencias bibliograficas
que han servido de base y de inspiracion para los desarrollos realizados en este trabajo.

CAPÍTULO 1
Aspectos Generales

1.1 INTRODUCCIÓN

Iniciamos **este capítulo** presentando algunos conceptos generales como el de analisis
multivariante y sistemas multivariantes, asi como una clasificacion de estos ultimos
acorde con el rol desempenado por las variables que los conforman y con las escalas de
medicion de estas. Definimos las escalas de medicion numerica, ordinal y nominal.
Seguidamente, presentamos una definicion de sistemas multivariantes mixtos y
homogeneos, con base en las escalas de medicion de las variables que los conforman.
Asimismo, con base en las escalas de medicion, precisamos la definicion de variables
cuantitativas y cualitativas.
A continuacion, presentamos una relacion de las principales tecnicas multivariantes
clasicas de analisis de datos, ubicandolas en un esquema general, segun el tipo de datos
sobre los cuales sean aplicables (numericos, ordinales, nominales) y el rol de las

variables (tecnicas de dependencia y de interdependencia).
Se expone luego el topico de cambio de escala de variables, en sus dos formas posibles:
debilitamiento de escala y fortalecimiento de escala. Se realiza una breve discusion
sobre categorizacion y codificacion de variables, cerrando el capitulo con lo tocante a su
cuantificacion.

1.2 CONCEPTOS BÁSICOS

GIFI (1990) define el **análisis multivariante** como "*el estudio de sistemas de variables aleatorias correlacionadas o muestras aleatorias de tales sistemas*". En tal sentido, puede afirmarse que los sistemas multivariantes son el objeto de estudio directo del analisis multivariante y, como tal, empezaremos definiendo algunas de sus caracteristicas.

Los **sistemas multivariantes**, es decir, aquellos conformados por multiples variables, pueden ser clasificados con base en diferentes criterios, a dos de los cuales haremos referencia: con base en el rol desempenado por las variables en el analisis; y con base en las escalas de medicion de las variables que los conforman. Mientras el primer aspecto depende de la tecnica de analisis, pudiendo cambiar al usar diferentes tecnicas para un mismo sistema, el segundo tiene que ver con caracteristicas intrinsecas del sistema.

Segun el **rol desempeñado por las variables** en el analisis, los sistemas multivariantes pueden ser **asimétricos** o **simétricos**. Se dice que un sistema multivariante es asimetrico cuando es posible definir dos grupos de variables con roles claramente diferenciados: un grupo de variables independientes y otro grupo de variables dependientes, pudiendo cada grupo estar conformado por una o mas variables. Por simplicidad, en los casos

anteriores se habla de datos asimetricos o simetricos, respectivamente. A las tecnicas de
analisis multivariante que procesan la informacion teniendo en cuenta estos diferentes
roles de las variables se les denomina **técnicas de dependencia**. Por otra parte, cuando
no existen roles de dependencia-independencia entre variables, lo que implica un
manejo analogo de todas las todas las variables, se dice que el sistema multivariante es
simetrico. A las tecnicas multivariantes que otorgan igual rol a las variables se les
denomina **técnicas de interdependencia** (HAIR et al., 1999).
Si bien es cierto que la definicion del rol de las variables no puede ser arbitraria y que
debe estar en consonancia con los objetivos del analisis, tambien es cierto que un mismo
sistema multivariante puede analizarse con base en diferentes enfoques, lo que hace que
el rol de las variables no sea un aspecto intrinseco inmutable, sino que dependa de los
objetivos del analisis y de la tecnica utilizada para alcanzar tales objetivos.

El otro criterio de tipificacion de los sistemas multivariantes se basa en la **escala de
medición** de las variables que lo conforman. Tal y como anotan NUNNALLY y
BERNSTEIN (1995), las escalas de medicion son convenciones que se adoptan para
registrar una realidad. Observese, por ejemplo, que existen multiples propuestas o
escalas para registrar una unica realidad como la temperatura: Celsius, Fahrenheit,
Kelvin, Rankine y Reaumur, entre otras. Luego, podemos definir escala de medicion
como *cualquier sistema que permita registrar una caracteristica o atributo de un objeto
de interes*.
Logicamente, mas alla de la naturaleza de la realidad que se pretenda registrar, las

propiedades de una *variable observada* estaran determinadas por su escala de medicion,
es decir, por el sistema utilizado para el registro de la realidad subyacente.
Antes de realizar una caracterizacion de las principales escalas de medicion, es
importante anotar que las variables cuantitativas pueden, a su vez, considerarse
continuas o **discretas**. Desde el punto de vista teorico, este concepto esta ligado a la
funcion de probabilidad asociada con cada variable aleatoria. En la practica, sin
embargo, el sistema utilizado para el registro de la realidad subyacente, puede
determinar la naturaleza continua o discreta de la *variable observada*[1], motivo por el
cual estos conceptos a menudo aparecen traslapados y confundidos en las definiciones
de escalas de medicion. Considerando que todas las variables continuas son
discretizables, podemos obviar esta distincion y considerar todas las variables
cuantitativas como discretas, lo que nos permite hablar en todos los casos de los niveles
de la variable (los diferentes valores que puede tomar la variable), tal y como se hace en
las definiciones que presentamos seguidamente.
La aparicion de una serie de publicaciones de STEVENS (1946, 1951, 1958) dio lugar a
una caracterizacion y clasificacion de las escalas de medicion, la cual ha sido ampliada,
reducida o modificada posteriormente, en funcion de objetivos y enfoques particulares.
Para nuestros fines, bastara con considerar los tres **tipos de escalas de medición** que se
describen a continuacion.

[1] La edad de un individuo medida como el tiempo transcurrido desde el nacimiento es una variable
continua: entre cualquier par de instantes que se definan en el tiempo, por cercanos que se encuentren
entre si, sera posible definir un infinito numero de instantes. No obstante, la edad de un individuo medida

en anos cumplidos es una variable discreta.

Escala Numérica. El valor asociado con cada nivel de esta escala indica la cantidad o intensidad de la caracteristica medida. La distancia entre cualquier par de niveles adyacentes de una variable medida en esta escala es la misma. Esta propiedad permite establecer comparaciones entre cualquier par de intervalos en la escala. Si ademas, la escala posee un valor de referencia cero, correspondiente a la ausencia de la caracteristica medida, tambien sera posible establecer relaciones de razon. Observese que esta definicion recoge lo que para otros fines serian las escalas de intervalo y de razon.

Escala Ordinal. El valor asociado con cada uno de los niveles de esta escala representa un rango, lo que solo permite establecer comparaciones de orden entre los diferentes niveles (mayor que, menor que). En este caso, no es posible suponer que la distancia entre un nivel y sus niveles adyacentes superior e inferior sea la misma. Aunque esto podria ocurrir eventualmente, no tendria que satisfacerse para todos los niveles; si asi fuera, la escala no seria ordinal, sino numerica.

Escala Nominal. El valor asociado con cada uno de los niveles de la variable no es mas que una etiqueta de identificacion, sin otro valor de comparacion con otros niveles de la escala que el de igualdad o diferencia. Un caso particular de variables nominales es el de aquellas con solo dos categorias exhaustivas y mutuamente excluyentes, esto es, las variables binarias o dicotomicas.

Existe una clasificacion general de las **variables** en **cuantitativas** y **cualitativas**.

Aunque estos terminos se autodefinen y no demandan precisiones adicionales, vale la pena relacionarlos con las escalas de medicion que acabamos de describir. Se dice que las variables numericas son cuantitativas, mientras que tanto las ordinales como las nominales son cualitativas.

Con el fin de simplificar el lenguaje (y aun a riesgo de abusar un poco del mismo), a los tres tipos de escalas de medicion descritos los llamaremos simplemente *escalas de medicion*. Ademas, cuando una variable se registre utilizando alguna de estas escalas de medicion, se dira que tal variable esta en dicha escala o se utilizara el nombre de la escala como adjetivo, v. gr., variable numerica, variable ordinal o variable nominal.

Tomando como referente las tres escalas de medicion descritas, definimos un **sistema multivariante mixto** como aquel en el que no todas las variables estan en la misma escala de medicion. Por simplicidad, puede hablarse tambien, en tales casos, de informacion mixta o datos mixtos. En contraposicion, y manteniendo el presente contexto, un **sistema multivariante homogéneo** es aquel en el que todas las variables estan en la misma escala de medicion. En este caso podra hablarse de datos homogeneos.

1.3 ALGUNAS TÉCNICAS MULTIVARIANTES DE ANÁLISIS DE DATOS

La escala o combinacion de escalas de medicion de las variables que conforman los sistemas multivariantes, asi como el rol de las variables, determinan las tecnicas disponibles para su analisis. La aplicacion de una u otra dependera, desde luego, de los

objetivos trazados.
Aunque la principal aportacion del presente estudio es precisamente un sistema integral,
el cual, a partir de la asignacion de cuantificaciones optimas, hace posible analizar
sistemas multivariantes mixtos usando practicamente cualquier tecnica multivariante,
por el momento nos circunscribiremos al ambito de aplicacion de los principales
metodos multivariantes en sus versiones clasicas[2]. En la Figura 1.1 se presenta un
esquema que considera la ubicacion de tales metodos, teniendo en cuenta si el sistema
multivariante es mixto u homogeneo y si los datos son simetricos o asimetricos.
Observese que las posibles combinaciones de las tres escalas de medicion consideradas
definen una matriz 3 x 3, con los sistemas multivariantes homogeneos en la diagonal
principal y los sistemas multivariantes mixtos por fuera de la diagonal. Puesto que las
combinaciones por encima de la diagonal son las mismas que se dan por debajo de esta,
solamente se presenta la parte superior de dicha matriz.
Dentro de cada una de las celdas de la matriz definida por las combinaciones de las
escalas de medicion aparece, a su vez, una matriz 3 x 3, generada por las combinaciones
de los posibles roles de las variables: independiente, dependiente o sin rol definido, esto
es, interdependiente. Muchas de las celdas de estas submatrices constituyen lo que en
una tabla de contingencia se conoce como ceros estructurales, es decir, combinaciones
que, por definicion, son imposibles de obtener. Por ejemplo, no existe ninguna tecnica
en la que los dos grupos de variables sean independientes o los dos grupos
dependientes[3]. Tales celdas se han cruzado con una X.

[2] Nos referimos a las formas mas populares de las tecnicas, sin considerar las adaptaciones basadas en
cambios de escala. Tampoco incluimos las tecnicas de analisis de datos de tres vias.

3 Las tecnicas de interdependencias, es decir, aquellas en las que todas las variables juegan el mismo rol,
aparecen siempre en la celda central de las submatrices.

Numérica Ordinal Nominal

IND INT DEP IND INT DEP IND INT DEP

IND ARM AD
RLOG

INT
ACP
AF
ACOP
ACC
ACONG
BIPLOT

Numérica

DEP MANOVA

IND ACNSVO

INT EMNM

Ordinal

DEP

IND ACNS

INT
AC
ACM
BIPLOG

Nominal

DEP

AC: Analisis de Correspondencias; **ACM**: Analisis de Correspondencias Multiple; **ACNS**: Analisis de Correspondencias no Simetrico; **ACNSVO**: Analisis de Correspondencias no Simetrico con Variables Ordinales; **ACON**: Analisis de Conglomerados; **ACOP**: Analisis de Coordenadas Principales; **ACP**: Analisis de Componentes
Principales; **AD**: Analisis Discriminante; **AF**: Analisis Factorial; **ARM**: Analisis de Regresion Multiple; **BIPLOG**: Biplot Logistico; **BIPLOT**: Analisis Biplot; **EMNM**: Escalamiento Multidimensional no Metrico; **MANOVA**: Analisis de Varianza Multivariante; **RLOG**: Regresion Logistica.

Figura 1.1. Clasificacion de las principales tecnicas multivariantes clasicas, acorde con las escalas de medicion y la simetria de los datos.

En el caso de los sistemas multivariantes homogeneos, se ha deshabilitado tambien la
celda inferior izquierda, no porque tal combinacion sea imposible de obtener, sino

porque en estos sistemas dicha combinacion es exactamente igual a la de la celda superior derecha. Esto, desde luego, no es aplicable a los sistemas multivariantes mixtos.

Sistemas Multivariantes Numéricos. El analisis de un sistema multivariante homogeneo en el que todas las variables sean numericas puede realizarse con base en diferentes tecnicas, acorde con los objetivos del analisis y con el rol que desempenen las variables. Como tecnicas de interdependencia, es decir, tecnicas en las que se asigna el mismo rol a todas las variables, cabe destacar las siguientes: Analisis de Componentes Principales (PEARSON, 1901; HOTELLING, 1933), Analisis Factorial (SPEARMAN, 1904; HOTELLING, 1933), Analisis de Correlacion Canonica (HOTELLING, 1936), Analisis de Coordenadas Principales (GOWER, 1966), Analisis de Conglomerados (FISHER, 1958; SOKAL and SNEATH, 1963) y Analisis Biplot (GABRIEL, 1971; GALINDO, 1985, 1986). Como tecnica de dependencia, se destaca el Analisis de Regresion Multiple (PEARSON, 1896; BARTLETT, 1933).

Sistemas Multivariantes Ordinales. La tecnica mas popular de interdependencia es el Escalamiento Multidimensional no Metrico (SHEPARD, 1962). Recientemente, LOMBARDO et al. (2007) desarrollaron una extension del Analisis de Correspondencias no Simetrico, llamada Analisis de Correspondencias no Simetrico con Variables Ordinales, la cual permite relacionar dos variables ordinales con diferentes roles.

Sistemas Multivariantes Nominales. En sistemas multivariantes conformados por variables nominales, puede utilizarse el Analisis de Correspondencias (BENZECRI et al., 1973), si se trata de solo dos variables, o el Analisis de Correspondencias Multiples[4]

(BENZECRI et al., 1973; BENZECRI, 1977), cuando se tienen mas de dos variables.

Un caso particular de variables nominales es aquel en el que cada variable esta conformada por solo dos categorias (variables binarias). Para analizar este tipo de datos,

[4] El ACM ha recibido muy diversos nombres dependiendo del grupo de investigadores involucrado en su desarrollo (cf. § 2.6). Una de las tecnicas equivalentes, que nos interesa destacar, por formar parte del sistema de analisis al cual dedicaremos parte de este trabajo, es el Analisis de Homogeneidad mediante Minimos Cuadrados Alternados (HOMALS).

el grupo de investigacion de Analisis Multivariante de la Universidad de Salamanca desarrollo una variante del Analisis Biplot llamada Biplot Logistico (VICENTEVILLARDON et al., 2006), mediante el cual es posible obtener una representacion Biplot de multiples variables binarias. Las tres tecnicas mencionadas (AC, ACM y BIPLOG) son de interdependencia. LAURO y D' AMBRA (1984) desarrollaron una propuesta de Analisis de Correspondencias no Simetrico que es aplicable cuando las variables juegan diferentes roles.

Sistemas Multivariantes Mixtos. Las tecnicas mas populares de analisis de sistemas multivariantes mixtos son de interdependencia, es decir, que existen dos grupos claramente diferenciados, donde uno desempena el papel de variable independiente, y el otro, el papel de variable dependiente. Cuando se tiene una variable respuesta nominal que intenta predecirse en funcion de un grupo de variables numericas, se utiliza el Analisis Discriminante (FISHER, 1936) o la Regresion Logistica (VERHULST, 1838, 1845, 1847). Cuando las variables independientes son nominales y las variables

respuestas son numericas, se utiliza el Analisis de Varianza Multivariante (WILKS, 1932; ROY, 1957).

1.4 CAMBIOS DE ESCALA

Las escalas de medicion de las variables que conforman un sistema multivariante determinan el conjunto de tecnicas que pueden usarse para su analisis. La mayoria de metodos clasicos son aplicables a sistemas homogeneos o conformados por grupos de variables con combinaciones predeterminadas de escalas de medicion. Aunque en muchos casos existen generalizaciones que permiten aplicar las tecnicas de analisis en un ambito mas amplio de situaciones que aquellas para las cuales fueron desarrolladas, muchas de tales variantes no se encuentran implementadas en los paquetes estadisticos comerciales que suelen utilizar los analistas de datos. Por tal motivo, constantemente se realizan **procesos de cambio de escala**, bien sea fortaleciendolas o debilitandolas[5].

Aunque existen tecnicas formales de **fortalecimiento de escala**, en algunas de las cuales nos concentraremos en los siguientes capitulos, es habitual pretender que este proceso se da de manera espontanea. Al menos, esto es lo que implicitamente se supone cuando se usan variables medidas en una escala debil como si fueran de escala fuerte[6].

En tales casos, el analista invoca una supuesta robustez de los metodos estadisticos que no siempre esta claramente sustentada.

Puesto que no tiene mucho sentido abordar el topico de la robustez aisladamente de las tecnicas a las que se refiere, siendo necesario un analisis particular para cada una de ellas, no profundizaremos en el mismo. Bastara con decir que en la mayoria de los

casos, el fortalecimiento de escalas –en particular, cuando no media ningun proceso
formal– es un proceso menos ortodoxo que su respectivo debilitamiento.
Para aplicar una tecnica de analisis de datos de escala debil sobre datos de escala fuerte,
es necesario efectuar un proceso de **debilitamiento de escala**, el cual se logra
resumiendo informacion o simplemente ignorandola.

[5] La fortaleza de una escala de medicion esta determinada por la cantidad de informacion que conlleva.
Asi, la escala numerica es la mas fuerte, seguida por la ordinal, siendo la nominal la mas debil. El proceso
de pasar de la escala nominal a la ordinal o a la numerica es de fortalecimiento, mientras que el paso de la
escala numerica a la ordinal o a la nominal es de debilitamiento.

[6] En particular, suelen utilizarse variables ordinales como si fueran numericas.

Si se tiene, por ejemplo, una variable ordinal, con un numero conveniente de niveles (no
se estima necesario colapsar niveles) y se desea aplicar una tecnica de analisis de datos
nominales, se pasa de la escala ordinal a la nominal mediante el simple recurso de
ignorar la informacion de orden contenida en las etiquetas de los niveles. Puesto que, en
este caso, el usuario no tiene que realizar ninguna accion y es el algoritmo
computacional el que ignora la informacion de orden, este debilitamiento de escala a
menudo pasa desapercibido.
Para realizar el debilitamiento de escala de una variable numerica podria ser suficiente
con efectuar un proceso de discretizacion (si la variable original no era discreta).
Logicamente, para llegar a una verdadera escala nominal habria que agregar el paso de
ignorar la informacion de orden contenida en los niveles de la variable discretizada.
Al resumir variables numericas u ordinales mediante su agrupacion en intervalos

adyacentes se genera una escala ordinal. Basta, sin embargo, con agregar el paso
implicito de ignorar la informacion de orden para llegar a una escala nominal.
Observese que al realizar un proceso de debilitamiento de escala hacia la nominal, sin
importar los pasos explicitos o implicitos involucrados en el proceso (discretizacion,
colapsamiento de niveles, ignorar la informacion de orden), la variable queda
constituida finalmente por un conjunto de categorias. Es por ello que a este proceso se le
conoce comunmente como categorizacion.
Tras un proceso de categorizacion, a las categorias generadas puede asignarseles
etiquetas sin ninguna restriccion, ya que su unica funcion sera la de senalar la
membresia de un individuo a una categoria determinada. En tal sentido, es perfectamente viable asignar incluso etiquetas alfanumericas, a no ser, desde luego, que
por restricciones propias de alguna rutina computacional con base en la cual se pretenda
analizar los datos, se exijan etiquetas numericas. No obstante, incluso en tales casos,
debe tenerse en mente que los numeros utilizados para identificar a cada categoria no
conllevan ninguna informacion numerica y bien podrian intercambiarse entre categorias.

1.5 CATEGORIZACIÓN DE VARIABLES

Una de las formas de categorizar una variable es mediante un proceso analogo al
utilizado en la construccion de tablas o histogramas de frecuencia, es decir, generando
un numero determinado de intervalos adyacentes de igual amplitud (excepto,
posiblemente, por los intervalos extremos) y asignando cada una de las observaciones al
intervalo que le corresponda. Esta es, quiza, la forma mas utilizada de categorizar una

variable. No obstante, es la menos enfocada a minimizar la perdida de informacion
implicita en el proceso de debilitamiento de escala.
El aspecto fundamental para minimizar la perdida de informacion propia de cualquier
proceso de categorizacion es, en nuestro concepto, la busqueda de categorias
contrastantes entre si, pero con alta homogeneidad interna. Para tal efecto, se requiere la
participacion del especialista en el fenomeno estudiado, quien debera definir los limites
de las categorias, acorde con su percepcion multivariante del fenomeno.
Supongase, por ejemplo, que se esta analizando una base de datos multivariante de
pacientes con cardiopatia isquemica, una de cuyas variables es la edad medida en anos
cumplidos. Para categorizar esta variable, el cardiologo debera definir los rangos de
edad que, segun su criterio, marcan diferencias en el contexto general de esta
enfermedad, haciendo caso omiso de relaciones muy especificas que puedan existir
entre la edad y algunas variables particulares de este contexto. Obviamente, los
intervalos de edad que asi se definan no tienen por que ser de la misma amplitud.
Aunque esta segunda forma de categorizar variables es superior a la descrita
previamente, tanto en terminos de conservacion de la informacion como de las
principales relaciones entre variables -siempre que se cuente, desde luego, con el
acertado criterio de un especialista-, la primera sigue siendo la mas utilizada por su
mayor objetividad (y menor nivel de compromiso del usuario con la decision tomada),
pues todo lo que el analista debe definir es el numero de intervalos, con lo cual los
limites de las categorias se generan automaticamente.

1.6 CODIFICACIÓN DE VARIABLES

Tras la categorizacion, algunas tecnicas, como el Analisis de Correspondencias (AC), reexpresan las variables mediante una matriz indicadora, con tantas filas como observaciones y tantas columnas como categorias hayan resultado del proceso de categorizacion. El sistema mas usual es el de **codificación disyuntiva completa**, mediante el cual cada observacion (fila) tendra un uno en la columna correspondiente a la categoria a la que pertenece y cero en las demas columnas.

Como alternativa a la codificacion disyuntiva completa, CAZES (1990) propone un sistema de **codificación baricéntrica**, segun el cual, una observacion puede quedar caracterizada por valores diferentes de cero en mas de una columna de la matriz de codificacion. Ello permite conservar no solo la informacion de pertenencia de una observacion a una categoria determinada, sino tambien la informacion relativa a la distancia entre las observaciones.

Para la aplicacion de la codificacion baricentrica, en lugar de definir los limites de las categorias, se definen r pivotes, es decir, r valores t_1, t_2, …, t_r, tales que, $t_1 < t_2 < t_3 < \cdots < t_r$. Cuando los pivotes son equidistantes, constituyen las marcas de clase de unos intervalos implicitamente definidos por este sistema de codificacion.

Si x_{ij} denota la i-esima observacion de la j-esima variable, y se denotan sus correspondientes codificaciones para cada uno de los r pivotes como $k(i, j_1)$, $k(i, j_2)$, …, $k(i, j_r)$, podemos expresar las reglas del sistema de codificacion baricentrica asi:

$Si\ x_{ij} \leq t_1 \Rightarrow k(i, j_1)=1$

$k(i, j_s)=0, \ \forall s \neq 1$

$Si\ x_{ij} \geq t_r \Rightarrow k(i, j_r)=1$

$k(i, j_s)=0, \ \forall s \neq r$

$Si\ \ t_m \leq x_{ij} \leq t_{m+1} \Rightarrow k(i,\ j_m)=(\ t_{m+1}-\ x_{ij})/\ (\ t_{m+1}-\ t_m)$
$k(i,\ j_{m+1})=(\ x_{ij}\ -t_m)/\ (\ t_{m+1}-\ t_m)$
$k(i,\ j_s)=0,\ \ \forall\ (\ s\neq m\ \wedge\ s\neq m+1)$

Para ilustrar estos dos metodos de codificacion, supongase una variable medida en
escala numerica, que tiene, entre otros, los siguientes valores: 3, 5, 9, 10, 11, 24 y 26.
Supongase tambien que para realizar una codificacion disyuntiva completa se han
definido los intervalos [0, 10), [10, 20) y [20, 30). Y supongase ademas que para
realizar una codificacion baricentrica, se han definido los pivotes 5, 15 y 25. La Tabla
1.1 muestra la codificacion de cada uno de estos valores con base en cada uno de los dos
sistemas.

Codificación Disyuntiva Completa			**Codificación Baricéntrica**		
Intervalos			**Pivotes**		
Valores originales					
[0, 10)	**[10, 20)**	**[20, 30)**	**5**	**15**	**25**
3					
1	0	0	1	0	0
5					
1	0	0	1	0	0
9					
1	0	0	0,6	0,4	0
10					
0	1	0	0,5	0,5	0
11					
0	1	0	0,4	0,6	0
24					
0	0	1	0	0,1	0,9
26					
0	0	1	0	0	1

Tabla 1.1. Codificacion disyuntiva completa y codificacion baricentrica de un conjunto
de observaciones.

Es importante resaltar un par de aspectos relativos al sistema de codificacion
baricentrica. En primer lugar, que este solo es aplicable a la codificacion de variables
numericas, no siendo adecuado para codificar variables ordinales. En segundo lugar,

que este metodo de codificacion resulta util solamente si se van a aplicar tecnicas que
permitan operar sobre matrices indicadoras, v. gr., Analisis de Correspondencias; de
hecho, es en tal contexto en el que Cazes desarrolla su propuesta.
Finalmente, hay que destacar que si bien existe una relacion directa entre cada uno de
estos sistemas de codificacion y una categorizacion subyacente, estos no constituyen
metodos de categorizacion en si mismos. En tal sentido, ninguno de ellos libera al
analista de la responsabilidad de definir los limites de las categorias.

1.7 CUANTIFICACIÓN DE VARIABLES

Ante la necesidad de analizar datos cualitativos mediante las tecnicas disponibles,
muchas de las cuales han sido desarrolladas para datos cuantitativos, a menudo se
recurre al **fortalecimiento automático de la escala**, consistente en utilizar la
informacion ordinal como si fuera numerica. Los analistas que se ven obligados a ello,
justifican su proceder en la robustez de los metodos. Sin entrar en las particularidades
de ningun metodo especifico, nos limitaremos a senalar que, a no ser que exista un claro
sustento de la robustez en cuestion, esta practica no es recomendable.
La alternativa mas ortodoxa para la aplicacion de metodos multivariantes cuantitativos
sobre sistemas cualitativos o mixtos consiste en utilizar procesos formales, mediante los
cuales, los diferentes niveles de las variables cualitativas pasan de estar identificados
por etiquetas a tener asociado un valor numerico, motivo por el cual se les denomina
métodos de cuantificación.
En ocasiones, la cuantificacion no constituye un fin en si misma, sino que forma parte
integral de las tecnicas, apareciendo, a lo sumo, como un subproducto de estas. Por tal

motivo, a menudo se ignora el papel de la cuantificacion y se tiende a creer que tecnicas concebidas para datos cuantitativos funcionan igualmente bien al ser aplicadas sobre datos cualitativos.

Segun la percepcion que se tenga del proceso de aplicacion de tecnicas cuantitativas a datos mixtos o cualitativos, este puede abordarse desde dos opticas: la primera, segun la cual algunas tecnicas cuantitativas pueden aplicarse indistintamente a datos cuantitativos, mixtos o totalmente cualitativos; la segunda, que considera que las tecnicas cuantitativas siempre exigen datos cuantitativos. El primer punto de vista es correcto, siempre que medie una cuantificacion de las variables cualitativas, lo cual nos lleva al segundo escenario. Si bien esta es una cuestion mas semantica que practica, la traemos a colacion para hacer notar que la aplicacion de metodos cuantitativos sobre datos cualitativos o mixtos siempre conlleva una cuantificacion implicita o explicita de las variables cualitativas.

Aunque el interes por el topico de la cuantificacion se pone de manifiesto en trabajos tan antiguos como el de Yule (*An introduction to the theory of statistics*, 1910, citado por YOUNG, 1981), una de las primeras propuestas de amplia difusion fue la de GUTTMAN (1941), siendo ademas una de las que daria lugar a una linea de trabajo de mayor continuidad (cf. § 2.6). A principios de la decada de 1950, varios autores realizaron contribuciones sobre este tema (BURT, 1950, 1953; HAYASHI, 1950), las cuales fueron resumidas a finales de la decada por TORGERSON (1958). En 1976 se

publica una propuesta para analizar la estructura aditiva en datos cualitativos (De LEEUW et al., 1976). Paralelamente, aparecen aportaciones de las escuelas francesa, nipona, canadiense e inglesa (BENZECRI et al., 1973; BENZECRI, 1977; SAPORTA, 1975; SAITO, 1973; NISHISATO, 1980; MARDIA et al., 1979).

A principios de la decada de 1980, Young publica un trabajo titulado *Analisis Cuantitativo de Datos Cualitativos* (YOUNG, 1981), en el que hace referencia al proceso de cuantificacion como *escalamiento optimo*, termino introducido por BOCK (1960). En su trabajo, Young define el **escalamiento óptimo**, asi: "*Tecnica de analisis de datos que asigna valores numericos a las categorias, de manera que se maximice la relacion entre las observaciones y el modelo de analisis de datos, respetando el caracter de medicion de los datos*". Posteriormente, en este trabajo, retomaremos el concepto de escalamiento optimo y actualizaremos su definicion (cf. § 2.3).

Tambien en 1981, sale a la luz el libro *Analisis Multivariante no Lineal*, publicado por la escuela holandesa de sistema de escalamiento de datos, el cual es reeditado en 1990 (GIFI, 1981, 1990). Alli se exponen detalladamente los principios y aplicaciones del escalamiento optimo, con lo cual este trabajo se convierte en el referente por excelencia de lo relacionado con cuantificacion y metodos multivariantes no lineales. Por tal motivo, hemos dedicado el Capitulo 2 del presente trabajo a la presentacion de algunos aspectos del sistema Gifi de analisis multivariante no lineal, en particular, los relacionados con la cuantificacion de variables cualitativas.

CAPÍTULO 2

Sistema Gifi

2.1 INTRODUCCIÓN

En **este capítulo** presentaremos inicialmente el sistema Gifi de analisis multivariante no lineal en forma general. Luego, se definira la homogeneidad, en el contexto de informacion compartida por las variables del sistema, concepto central al sistema Gifi y totalmente diferente de la definicion de sistemas multivariantes homogeneos que se presento en el Capitulo 1, haciendo referencia a sistemas constituidos por variables medidas en la misma escala.

Se retoma, luego, la definicion de escalamiento optimo presentada en el Capitulo 1, cotejandola con la de Gifi. A partir de ambas definiciones, se elabora una definicion unificada de escalamiento optimo que recoge los elementos de las definiciones de YOUNG (1981) y GIFI (1990).

Seguidamente, se expone el metodo de los minimos cuadrados alternados, el algoritmo mas popular para resolver problemas de escalamiento optimo.

En el siguiente apartado se discute lo relativo al uso de transformaciones para la generacion de cuantificaciones, detallando las caracteristicas de las transformaciones utilizadas, las restricciones que cada una de ellas conlleva y las caracteristicas de las cuantificaciones generadas.

A continuacion, se presenta el HOMALS como una generalizacion de la funcion de perdida del Analisis de Componentes Principales. Finalmente se presenta el OVERALS, la tecnica mas general del sistema Gifi, que permite expresar las demas tecnicas como casos particulares de esta.

2.2 SISTEMA GIFI DE ANÁLISIS MULTIVARIANTE NO LINEAL

La publicacion en 1981 de la version en ingles del texto *Nonlinear Multivariate Analysis* (GIFI, 1981), por parte del grupo de investigacion en Teoria de Datos de la Facultad de Ciencias Sociales de la Universidad de Leiden, en Holanda, que para la ocasion escribio bajo el seudonimo de ***Albert Gifi***[7], marco un hito en la concepcion de las tecnicas multivariantes no lineales.

Este texto recoge y amplia los avances sobre metodos multivariantes no lineales que hasta la fecha se encontraban dispersos en publicaciones especializadas o que incluso eran de difusion limitada por formar parte de reportes internos de la Universidad de Leiden. Ademas de esta labor recopilatoria y de ampliacion, que de por si constituye ya un invaluable aporte, el gran merito del texto en cuestion consiste en la integracion que hace de los metodos multivariantes no lineales, mostrandolos como componentes de un sistema, lo que explica el hecho de que Jan de Leeuw se refiera posteriormente a este conjunto de tecnicas como **sistema Gifi de análisis multivariante no lineal** (De LEEUW, 1984), designacion que fue incorporada en la siguiente version del texto de Gifi (GIFI, 1990) y que ha seguido utilizandose desde entonces.

Luego, cuando hacemos referencia al sistema Gifi, ello no significa que estemos tratando exclusivamente con lo publicado por el grupo Gifi, sino que compartimos el enfoque sistemico usado por estos autores[8] para estudiar las tecnicas multivariantes no lineales.

GIFI (1990) define el **análisis multivariante** como el estudio de sistemas de variables aleatorias correlacionadas o de muestras aleatorias de tales sistemas. Estos sistemas

siempre estaran referenciados en espacios multidimensionales, lo que dificulta su representacion y, por tanto, su interpretacion. Uno de los principales objetivos de

7 Son numerosos los investigadores que han realizado aportaciones a este proyecto. En el prefacio de la edicion de 1990 (GIFI, 1990), los editores destacan los siguientes miembros: Bert Bettonvil, Eeke van der Burg, John van de Geer, Willem Heiser, Jan de Leeuw, Jacqueline Meulman, Jan van Rijckevorsel, Ineke Stoop, Peter van der Heijden, Adrian Meester, Peter Neufeglise, Renee Verdegaal, Peter Verboon, Ivo van der Lans, Gerda van den Berg y Patrick Groenen.

8 En adelante, cuando hablemos de Gifi, no nos referiremos a los autores, sino al grupo. Por tanto, cuando *Gifi* sea el sujeto de una oracion, los verbos se conjugaran en la tercera persona del singular.

muchas tecnicas multivariantes es la reduccion de la dimensionalidad, con el fin de facilitar la representacion e interpretacion de tales sistemas, con la menor perdida posible de informacion.

En este contexto aparece el concepto de **homogeneidad**, que conlleva la idea de que diferentes variables comparten informacion o redundan en la medicion de ciertos aspectos (GIFI, 1990). A mayor informacion compartida, mayor sera la homogeneidad de ese grupo de variables y mas eficientes seran las tecnicas de reduccion de dimensionalidad. Notese que este es un concepto no tiene nada que ver con la definicion de sistemas multivariantes homogeneos (cf. § 1.2). Todas las alusiones a la homogeneidad que se hacen en este capitulo tienen que ver con la cantidad de informacion comun del sistema, tal y como acaba de presentarse.

Si se considera una matriz X con n individuos (u objetos, segun la terminologia utilizada en el sistema Gifi) en las filas y m variables en las columnas, es posible reducir

la dimensionalidad del sistema remplazando todas las columnas de *X* por un vector *y*, elegido de manera que contenga la mayor cantidad de informacion disponible en *X*. Esto equivale a remplazar todas las variables de la matriz *X* por una variable resumen. A *y* se le denomina **vector de puntuaciones de los individuos**.

Antes de analizar diferentes posibilidades para la eleccion de *y*, de manera que contenga la mayor cantidad de informacion posible, podria pensarse en la solucion mas basica; esto es, obtener *y* como el vector promedio de los vectores columna de la matriz *X*. GIFI (1990) muestra que esta solucion solo seria aceptable si todas las variables estuvieran altamente correlacionadas entre si y que, en general, es posible obtener una solucion que capte mas informacion.

La acepcion mas utilizada para el termino ***información*** es la que se refiere a variabilidad. Se busca, pues, reducir la dimensionalidad del sistema original, conservando la mayor cantidad posible de la variabilidad total del sistema.

En el metodo lineal clasico de reduccion de dimensionalidad por antonomasia, es decir, en el Analisis de Componentes Principales (ACP), el vector de puntuaciones de los individuos se obtiene con base en combinaciones lineales de las variables

(HOTELLING, 1933). Esto equivale a obtener el vector *y* como el promedio[9] de las variables previamente escaladas o transformadas linealmente, de manera que cada cual tenga una ponderacion acorde con la cantidad de informacion que comparte con las demas variables. Las transformaciones que garantizan que el vector resumen, o vector de puntuaciones de los individuos, recoja la maxima informacion posible, se obtienen

usualmente con base en la descomposicion en valores y vectores propios de la matriz
$X'X$.

9 Aunque es mas natural verlo como una suma que como un promedio, solo se trata de una cuestion de escala que se resuelve dividiendo o multiplicando por una constante.

2.3 ESCALAMIENTO ÓPTIMO

Puesto que en la practica, los sistemas multivariantes nunca tienen maxima homogeneidad (las variables pueden compartir informacion, pero no todas miden exactamente lo mismo), la reduccion de la dimensionalidad siempre conlleva perdida de informacion, la cual puede medirse a traves una **función de pérdida** (*loss function*). Las funciones mas usadas para tal efecto se construyen promediando las sumas de cuadrados de las diferencias entre el vector de puntuaciones de los individuos y cada una de las variables transformadas, asi:

$$\sigma(y, a) \equiv m^{-1}\sum_{j} SSQ(y - a_j x_j) \quad [2.1]$$

Donde:

$\sigma(y,a)$: Funcion de perdida en los parametros y y a.
m: Numero de variables.
SSQ: Suma de cuadrados.
y: Vector de puntuaciones de los individuos.
a_j: Ponderacion o transformacion aplicada a la j-esima variable.
x_j: j-esima columna de la matriz X $(n \times m)$.

Una alternativa a la descomposicion en valores y vectores propios de la matriz $X'X$ para la obtencion del vector de puntuaciones y consiste en minimizar la funcion de perdida [2.1]. Se denomina **escalamiento óptimo** a cualquier tecnica generadora de transformaciones que minimicen una funcion de perdida, es decir, que minimicen la perdida de informacion que se produce al reducir la dimensionalidad del sistema

multivariante, con base en un modelo dado.
Aunque esta definicion de escalamiento optimo, adaptada de GIFI (1990), tenga una
apariencia un tanto diferente a la de la definicion dada por YOUNG (1981) (cf. § 1.7),
ambas hacen referencia al mismo mecanismo, solo que enfatizando aspectos diferentes.
Mientras la definicion de Young esta enfocada en el topico de las cuantificaciones, la

definicion de Gifi se centra en la reduccion de dimensionalidad. No obstante, si
consideramos que las transformaciones a las que se hace referencia en la definicion de
Gifi constituyen una asignacion de valores numericos a las categorias, tal y como se
indica en la definicion de Young, podemos integrar ambos conceptos para presentar una
definicion acorde con nuestro proposito.

Definición 2.1 (Escalamiento Óptimo): Tecnica que, mediante el uso de
transformaciones, cuantifica las categorias de las variables de un sistema
multivariante, de manera que se minimice la perdida de informacion que
resulta al reducir la dimensionalidad del sistema cuantificado, con base en
un modelo de analisis dado.

Debe evitarse otorgar al apelativo *optimo* un significado mas alla del contexto al que
pertenece. Observese que este termino solo indica que las variables escaladas con base
en tal criterio pueden promediarse para pasar del espacio multidimensional a un espacio
de menor dimensionalidad, con minima perdida de informacion.
Hay que anotar, sin embargo, que aunque las transformaciones solo pueden entenderse
como optimas bajo tal acepcion y para el grupo particular de variables que conforma el

sistema multivariante, las variables transformadas son utilizadas frecuentemente en un
amplio rango de aplicaciones, mas alla de las tecnicas generadoras de las transformaciones.
En los dos apartados siguientes detallaremos los aspectos centrales de la definicion 2.1.
En el apartado 2.4 analizaremos el mecanismo que permite minimizar la perdida de
informacion resultante de reducir la dimensionalidad del sistema cuantificado, con base
en un modelo de analisis dado. En el apartado 2.5 analizaremos las transformaciones
usadas para la generacion de cuantificaciones.

2.4 MÍNIMOS CUADRADOS ALTERNADOS

Como parte integral de las tecnicas que conforman el sistema Gifi, el algoritmo
computacional que mas se ha utilizado para resolver problemas de escalamiento optimo
(minimizar funciones de perdida con base en un modelo de analisis dado) es el de los
minimos cuadrados alternados, que usualmente se denomina mediante la sigla **ALS**, por
sus iniciales en ingles: ***A****lternating* ***L****east* ***S****quares*. La principal ventaja de este
algoritmo es su generalidad, lo que lo hace aplicable en una amplia gama de situaciones.
YOUNG (1981) se refiere al conjunto de algoritmos computacionales utilizados para
resolver problemas de escalamiento optimo mediante minimos cuadrados alternados
como programas **ALSOS**, sigla correspondiente a la expresion ***A****lternating* ***L****east*
S*quares approach to* ***O****ptimal* ***S****caling*, que podria traducirse como Aproximacion
Minimo Cuadratica al Escalamiento Optimo.
Vale la pena anotar, sin embargo, que aunque no sean tan generales como el ALS,
existen algoritmos alternativos como el de los valores propios, o el de la mayorizacion

(De LEEUW, 1994, 2006; VERBOON and HEISER, 1994; HEISER, 1995; LANGE et al., 2000) que podrian ser mas eficientes que el ALS en algunas situaciones particulares
y que de hecho se utilizan en algunos programas especificos del sistema Gifi como el
ANACOR que realiza Analisis de Correspondencias y el ANAPROF, que realiza Analisis de Frecuencias de Perfil cuando el numero de individuos es mucho mas grande
que el numero de perfiles.
En general, las funciones de perdida tienen dos conjuntos de parametros: uno de ellos
relacionado con las puntuaciones de los individuos, y el otro, con las transformaciones
de las variables. El ALS es un algoritmo iterativo, es decir, que converge a la solucion a
traves de ciclos. Cada ciclo consta de varias etapas: en una se estiman las puntuaciones
optimas para unos valores dados de las transformaciones, y en otra se estiman las
transformaciones optimas para unas puntuaciones dadas. Estas etapas se alternan,
actualizando en cada una de ellas los valores de un conjunto de parametros con base en
los valores obtenidos en la etapa anterior para el otro conjunto. Esto genera una
secuencia decreciente de valores para la funcion de perdida, que converge a una
constante a medida que los parametros se estabilizan. El proceso se da por terminado

cuando la diferencia obtenida entre dos iteraciones sea menor que un criterio de parada
definido por el usuario (diferencia $\leq \varepsilon$).
Una de las versiones mas sencillas de una funcion de perdida es la presentada en [2.1],
que es la correspondiente al ACP. Notese que si en tal funcion se hiciera $y = 0$ y $a = 0$,
se obtendria una solucion trivial. Para evitarla se imponen restricciones sobre y o sobre
a, de manera que tengan norma unitaria.

A continuacion se muestran los pasos correspondientes al **algoritmo ALS con puntuaciones normalizadas**, para minimizar la funcion de perdida [2.1].

0. Elegir un vector arbitrario de ponderaciones, a, de tamano m y diferente de cero,
para iniciar el algoritmo.

1. Actualizar el vector de puntuaciones, y.
$y \leftarrow Xa / m$

2. Normalizar el vector de puntuaciones, y.
$y \; y(y' y)^{1/2} \leftarrow -$

3. Actualizar el vector de ponderaciones, a.
$a \leftarrow X' y$

4. Evaluar convergencia. Si el valor de la diferencia entre la funcion de perdida
obtenida en esta iteracion y la funcion de perdida obtenida en la iteracion
anterior satisface el criterio de parada elegido por el usuario (diferencia $\leq \varepsilon$), se
da fin al proceso. En caso contrario, se vuelve al paso 1.

El **algoritmo ALS con ponderaciones normalizadas** seria analogo:

0. Elegir un vector arbitrario de puntuaciones, y, de tamano n y diferente de cero,
para iniciar el algoritmo.

1. Actualizar el vector de ponderaciones, a.
$a \leftarrow X' y$

2. Normalizar el vector de ponderaciones, a.
$a \leftarrow a(a'a)^{-1/2}$

3. Actualizar el vector de puntuaciones, y.
$y \leftarrow Xa / m$

4. Evaluar convergencia. Si el valor de la diferencia entre la funcion de perdida
obtenida en esta iteracion y la funcion de perdida obtenida en la iteracion
anterior satisface el criterio de parada elegido por el usuario (diferencia $\leq \varepsilon$), se
da fin al proceso. En caso contrario, se vuelve al paso 1.

Una de las caracteristicas que le confiere mayor versatilidad al sistema Gifi es su

capacidad de manejar sistemas multivariantes mixtos, es decir, conformados por
variables con diferentes escalas de medicion. Tal y como indica YOUNG (1981),
siempre que exista un procedimiento de minimos cuadrados para el ajuste de un modelo
particular a datos cuantitativos, podra usarse escalamiento optimo mediante ALS para
obtener un ajuste analogo de dicho modelo, usando datos cualitativos o una
combinacion de variables con diferentes niveles de escalamiento.
A continuacion analizaremos el otro aspecto de la definicion 2.1, es decir, el relativo a
la generacion de cuantificaciones mediante el uso de transformaciones.

2.5 ESCALAMIENTO, TRANSFORMACIONES Y CUANTIFICACIONES

Las tecnicas del Sistema Gifi que aqui consideramos exigen que todas las variables sean
discretizadas o categorizadas antes de ser procesadas mediante escalamiento optimo.
Por tal razon, siempre podremos hablar de las *categorias* de las variables, sin importar
su escala de medicion.
Tal y como se indico en la definicion 2.1, al someter un sistema multivariante a un
proceso de escalamiento optimo, la cuantificacion de las categorias se genera mediante
transformacion de las variables. El tipo de transformaciones que se utilice para
cuantificar una variable especifica depende de la eleccion por parte del usuario del nivel
de escalamiento de dicha variable (*scaling level*).
Aunque los conceptos de **nivel de escalamiento** y escala de medicion (cf. § 1.2) se
encuentran estrechamente ligados, se diferencian por el hecho de que el nivel de
escalamiento lo define el usuario, mientras que la escala de medicion es una propiedad
inherente al sistema de registro utilizado para plasmar la realidad estudiada. Despues de

presentar los niveles de escalamiento volveremos sobre este punto.

Los **niveles de escalamiento** basicos[10] considerados en este trabajo son el **nominal**, el **ordinal** y el **numérico**. Cada nivel de escalamiento define una familia de transformaciones permisibles. Se utilizan transformaciones lineales para las variables escaladas a nivel numerico; transformaciones monotonamente ascendentes para las variables escaladas a nivel ordinal; y transformaciones isomorficas para las variables escaladas a nivel nominal (Van der BURG et al. 1994).

Una **transformación lineal** de una variable consiste en multiplicar cada uno de sus valores por una constante. Luego, los valores transformados seran proporcionales a los valores originales. En consecuencia, al representar los valores originales y los transformados en un plano cartesiano se forma una linea recta.

[10] Hablamos de niveles de escalamiento basicos, pues aunque en el Capitulo 5 definimos dos niveles de escalamiento adicionales, aquellos no son mas que variaciones de los niveles de escalamiento que aqui se presentan.

En contraste con las transformaciones lineales, cualquier transformacion que genere valores transformados no proporcionales a los originales sera una **transformación no lineal**. Si se aplica una transformacion no lineal y se representan los valores originales frente a los transformados en un plano cartesiano no se obtendra una linea recta[11].

Las **transformaciones monótonamente ascendentes**, que se aplican al escoger el nivel de escalamiento ordinal, forman parte de las transformaciones no lineales, y se caracterizan por el hecho de que los estadisticos de orden de la variable original coinciden con los estadisticos de orden de la variable transformada (excepto cuando hay

empates). Esto significa que el orden de la variable original se mantiene en la variable transformada. La funcion correspondiente es, por tanto, no decreciente.

Formando parte tambien de las transformaciones no lineales, se encuentran las **transformaciones isomórficas**, que se aplican al elegir el nivel de escalamiento nominal. Este tipo de transformacion es el que conlleva menos restricciones; lo unico que debe satisfacer es que a todas las observaciones correspondientes a una categoria se les asigne el mismo numero real, sin que tenga que satisfacerse ninguna relacion entre los valores asignados a diferentes categorias de una misma variable.

Las transformaciones descritas son funciones que, cuando se incorporan en un algoritmo ALS (o en cualquier otro algoritmo de minimizacion de funciones de perdida), asignan puntuaciones normalizadas a cada una de las categorias de las variables. A este proceso se le denomina ***cuantificación***, puesto que las puntuaciones asi asignadas a cada una de las categorias, a diferencia de las etiquetas originales, tienen propiedades metricas, lo cual posibilita analizar la base de datos transformada mediante tecnicas estandar del analisis multivariante, es decir, tecnicas lineales que exijan datos numericos.

11 GIFI (1990) clasifica las tecnicas del analisis multivariante en lineales, monotonas y no lineales con base en la invarianza de los resultados bajo los diferentes tipos de transformaciones uno a uno, aplicadas a las variables aleatorias. En general, se dice que una tecnica es no lineal si incluye transformaciones no lineales entre la bateria de posibles transformaciones, aun cuando en algunos casos particulares no se haga uso de las mismas.

Dado que las variables numericas tienen propiedades metricas aun antes de ser transformadas, lo mas correcto seria

denominar ***escalamiento*** al proceso de aplicar transformaciones sobre un sistema multivariante numerico, reservando el termino ***cuantificación*** para hacer referencia al resultado de aplicar transformaciones sobre variables ordinales y/o nominales. No obstante, esta es una diferenciacion netamente linguistica que no cambia en nada los resultados del proceso. Al aplicar transformaciones sobre sistemas multivariantes mixtos, usualmente se utiliza indistintamente cualquiera de los dos terminos, pudiendo hablarse tambien de ***escalamientos y/o cuantificaciones***.

La Figura 2.1 muestra el aspecto de tres graficos de cuantificaciones tipicos, obtenidos mediante transformaciones lineales, monotonamente ascendentes e isomorficas.

Es importante insistir en el hecho de que todas las variables transformadas mediante un programa ALSOS tienen propiedades metricas, cualquiera que sea el nivel de escalamiento elegido por el usuario. La unica funcion del nivel de escalamiento es informar al proceso sobre las relaciones iniciales entre los niveles o categorias de la variable, de manera que tales relaciones se transmitan a la variable transformada. Estas relaciones son analogas a las definidas para las escalas de medicion (cf. § 1.2).

Figura 2.1. Graficos de cuantificaciones tipicos para una transformacion (a) lineal, (b) monotonamente ascendente y (c) isomorfica.

a

b
c

En vista de la estrecha relacion existente entre la escala de medicion y el nivel de escalamiento, podria pensarse que el concepto de nivel de escalamiento es redundante y que bastaria con indicar la escala de medicion de cada variable. Esto, sin embargo, no es asi, debido a un par de aspectos que detallaremos a continuacion, los cuales hacen util el concepto de nivel de escalamiento y justifican su existencia como ente diferenciado de la escala de medicion.

En primer lugar, las tecnicas del sistema Gifi que aqui consideramos exigen que todas las variables sean discretizadas o categorizadas previa aplicacion del escalamiento optimo. Este paso de categorizacion puede conllevar perdida de fortaleza de la escala con relacion a la escala de medicion de la variable original. Podria suceder, por ejemplo, que al categorizar una variable con escala de medicion numerica como la edad, se obtuvieran categorias para las cuales no se satisficiera que la distancia entre cualquier par de niveles adyacentes de la variable categorizada fuera la misma. Luego, aunque la escala de medicion de la variable edad fuera numerica, su nivel de escalamiento seria ordinal.

En segundo lugar, aun cuando la categorizacion no alterase la escala de medicion de la variable original, el analista puede tener interes en realizar un escalamiento mas flexible. Si en el contexto de las escalas de medicion definiamos la fortaleza de la escala, acorde con la cantidad de informacion de cada una (cf. § 1.4, nota 5), hablando

de nivel de escalamiento, definimos la flexibilidad acorde con las restricciones que conlleva. En tal sentido, el nivel de escalamiento nominal es el mas flexible, seguido del escalamiento ordinal, siendo el escalamiento numerico el menos flexible o el que mas restricciones conlleva.

La utilizacion de un nivel de escalamiento mas flexible que aquel que corresponderia naturalmente a la escala de medicion de la variable puede resultar util para evaluar posibles relaciones no lineales de una variable con otras variables del sistema, en particular, cuando se cree que las distancias definidas por la escala de medicion pueden no estar muy acordes con el papel desempenado por la variable en el contexto estudiado.

Teniendo en cuenta los dos aspectos detallados anteriormente, consideramos perfectamente viable la utilizacion de un nivel de escalamiento que no corresponda con la escala de medicion de la variable. No obstante, mas alla de tales consideraciones, creemos que, en general, debe utilizarse el nivel de escalamiento que corresponda a la escala de medicion de la variable. En tal sentido, somos mas conservadores que MEULMAN et al. (2005), quienes afirman que "*no existen propiedades intrinsecas de la variable que predefinan automaticamente un optimo nivel de escalamiento que deba especificarse para la misma, pudiendo explorar los datos de cualquier forma que tenga sentido y que facilite la interpretacion*".

Puesto que el nivel de escalamiento definido por el usuario es el que en ultima instancia determina el procesamiento de cada variable, en adelante utilizaremos exclusivamente

este concepto, sin ocuparnos mas de la escala de medicion. Asi, a menos que
indiquemos lo contrario, cuando hablemos de variables numericas, ordinales o
nominales, nos estaremos refiriendo al nivel de escalamiento elegido para las mismas.

2.6 HOMALS

Por extension de la definicion de sistemas multivariantes mixtos presentada en § 1.2 y
con adaptacion al presente contexto, definimos los sistemas multivariantes mixtos como
aquellos en los que no todas las variables se escalan al mismo nivel. Puesto que una de
las fortalezas del Sistema Gifi es su capacidad para manejar datos mixtos, las funciones
de perdida mas generales no estan basadas en la matriz X, que contiene los valores de
las variables, como se mostro en la expresion [2.1], sino en una matriz indicadora
completa G, que tiene una columna para cada uno de los posibles valores de cada
variable[12]. A cada objeto se le registrara un 1 en la columna que corresponda al valor o
categoria a la que pertenece y 0 en las demas columnas de la correspondiente variable
(cf. § 1.6). Esto explica por que estas tecnicas exigen que todas las variables sean
categorizadas.

Una funcion de perdida mas general que [2.1], la cual permite obtener una matriz de
puntuaciones, Y (n x p), en lugar de un vector[13] esta dada por la expresion [2.2]:

$$\sigma(Y, Q) \equiv \sum_{j=1}^{m} SSQ\left(Y_j - G_j Q_j\right) \quad [2.2]$$

Donde:

$\sigma(Y,Q)$: Funcion de perdida en los parametros Y y Q.

m: Numero de variables.

SSQ: Suma de cuadrados.

Y: Matriz n x p de puntuaciones de los individuos.
G_j: Matriz indicadora correspondiente a la j-esima columna de la matriz X.
Q_j : Matriz de cuantificaciones de las categorias de la j-esima variable.
[12] Se tendra una submatriz G_j para cada una de las columnas de X, con tantas columnas como valores
asuma la correspondiente variable. La matriz G se conforma por yuxtaposicion del conjunto de matrices
G_j.
[13] Se reduce la dimensionalidad del sistema a p, con $p > 1$.

Para evitar la solucion trivial que se obtendria cuando las matrices Y y Q_j son cero, se
aplica alguna restriccion o normalizacion. La mas usual consiste en hacer que cada uno
de los vectores de puntuaciones tenga norma 1.
La expresion [2.2], con la restriccion de que todas las variables esten escaladas a nivel
nominal es la funcion de perdida del Analisis de Homogeneidad o HOMALS, acronimo
de la expresion inglesa ***HOM****ogeneity Analysis by* ***A****lternating* ***L****east* ***S****quares*.
TENENHAUS y YOUNG (1985) muestran que el **HOMALS es uno de los muchos enfoques bajo los cuales se ha desarrollado la técnica más popularmente conocida
como Análisis de Correspondencias Múltiples** (ACM). Esta tecnica es equivalente al
Escalamiento Optimo, Puntuacion Optima y Puntuacion Adecuada de la escuela
americana (*Optimal Scaling, Optimal Scoring, Appropriate Scoring*); al Escalamiento
Dual canadiense (*Dual Scaling*); al Analisis de Escalograma israeli (*Scalogram
Analysis*); al metodo de cuantificacion japones (*Quantification Method*) y, por supuesto,
al Analisis de Correspondencias Multiples de la escuela francesa (*Multiple
Correspondence Analysis*).
Estos autores (TENENHAUS and YOUNG, 1985) indican que en adicion a la influencia que los diferentes paises y lenguas han ejercido para presentar este metodo

con diferentes nombres, esta el enfoque seguido para su desarrollo, siendo posible
distinguir cuatro corrientes o enfoques principales: el enfoque de los Promedios
Reciprocos, el del Analisis de Componentes Principales, el del Analisis Canonico
Generalizado y el del Analisis de Varianza.

El **enfoque de los Promedios Recíprocos** (*Reciprocal Averaging*) fue propuesto por
Richardon y Kuder (1933, Citados por HORST, 1935) y FISHER (1940), al parecer de
forma independiente, e impulsado por HILL (1973), quien lo presento como un metodo
de ordenacion en ecologia. GREENACRE (2007) senala que los Promedios Reciprocos
constituye un algoritmo alternativo a la descomposicion en valores singulares (DVS),
para la obtencion de la solucion de un Analisis de Correspondencias (AC). Aunque este
algoritmo no tiene la elegancia matematica de la DVS resulta muy util para ilustrar las
propiedades del Analisis de Correspondencias.

El algoritmo de los Promedios Reciprocos parte de cualquier conjunto de valores
estandarizados como coordenadas iniciales de las categorias columna (podria iniciarse
con las categorias fila). Seguidamente se obtienen las coordenadas de las categorias fila
con base en los promedios ponderados de las correspondientes categorias fila. A
continuacion, se recalculan las coordenadas de las categorias columna con base en los
promedios ponderados de las categorias fila y se aplica un nuevo paso de
estandarizacion. El proceso continua de forma iterativa hasta que se alcance la
convergencia, generandose asi las coordenadas del primer eje principal. NISHISATO
(2004) presenta una ilustracion numerica detallada de este algoritmo.

El **enfoque del Análisis de Componentes Principales** esta basado en las ideas de
HORST (1935) y BURT (1950, 1953) y ha sido desarrollado ampliamente por la
escuela francesa (BENZECRI et al., 1973; BENZECRI, 1977; GREENACRE, 1984).
Se parte del hecho de que analogamente a la forma en que puede obtenerse la solucion
del Analisis de Componentes Principales (ACP) a partir de la Descomposicion en
Valores Singulares (DVS) de una matriz, X, de datos multivariantes, es posible obtener
la solucion del Analisis de Correspondencias a partir de la DVS simultanea de dos
matrices de valores positivos (una para las categorias fila y otra para las categorias
columna) que considere la metrica introducida por la asignacion de ponderaciones a los
perfiles y a las dimensiones, asi: $_{p}^{1/2}\,_{d}^{1/2}D\,X\,D$, donde, D_p y D_d son matrices diagonales
que contienen las ponderaciones para los perfiles y para las dimensiones
respectivamente. GREENACRE (1984) indica que en ausencia de ponderaciones, el
procedimiento anteriormente descrito genera la solucion del ACP.
El **enfoque del Análisis Canónico Generalizado** se remonta al trabajo de McKEON
(1966). LECLERC (1980) presenta una sintesis del mismo. Bajo este enfoque, la
solucion se obtiene mediante la maximizacion de la suma de correlaciones cuadraticas
entre las variables cuantificadas y las puntuaciones de los individuos.
El **enfoque del Análisis de Varianza** (*The Analysis of Variance Approach*) se basa en
las ideas de GUTTMAN (1941), quien presento una formulacion completa del
problema, asi como su solucion. Posteriormente, BOCK (1960) y LINGOES (1963)
implementaron las correspondientes rutinas computacionales. Este es el enfoque en el

que se circunscriben desarrollos posteriores como el Escalamiento Dual (NISHISATO, 1979, 1980; MARAUN et al., 2005) y los trabajos de De LEEUW (1973) y Van RIJCKEVORSEL y De LEEUW (1978) que dan origen al HOMALS. Su formulacion matematica es presentada por TENENHAUS y YOUNG (1985).

El nombre de este ultimo enfoque se justifica en el hecho de que Guttman planteaba descomponer la variabilidad total de las cuantificaciones de las categorias en una componente de variabilidad entre individuos (suma de cuadrados de las desviaciones entre la puntuacion del individuo y la media general) y otra de variabilidad dentro de cada individuo (suma de cuadrados de las desviaciones entre la puntuacion del individuo y las cuantificaciones de sus correspondientes categorias). Seguidamente, resolvia el problema, maximizando la razon entre variabilidad entre individuos y variabilidad dentro de individuos.

Una vez obtenidas las cuantificaciones para las diferentes categorias de cada una de las variables, Guttman plantea el problema dual de descomponer a su vez la variabilidad total de las puntuaciones de los individuos en dos componentes: una componente debida a la variabilidad de las puntuaciones de los individuos que comparten una categoria especifica (suma de cuadrados de las desviaciones entre la puntuacion de cada individuo y la puntuacion promedio de cada una de sus categorias) y otra debida a la variabilidad de la puntuacion promedio de tal categoria, con relacion a la puntuacion promedio general (suma de cuadrados de las desviaciones la puntuacion promedio de la categoria y la puntuacion promedio general). Seguidamente, resolvia el problema, maximizando la razon entre variabilidad entre puntuaciones promedio de las categorias y variabilidad total de las puntuaciones de los individuos.

Segun WARRENS et al. (2007), el metodo de Guttman, al que se refieren como
Escalamiento Optimo Clasico, permite una descomposicion multidimensional de los
datos, con la dimension estructural mas informativa apareciendo en primer lugar, luego
la segunda y asi sucesivamente hasta extraer exhaustivamente toda la informacion
disponible en los datos.
TENENHAUS y YOUNG (1985) concluyen que cualquiera de los anteriores enfoques
conducen a las mismas ecuaciones al ser usados sobre un conjunto particular de datos.

MICHAILIDIS y De LEEUW (1998) precisan que la solucion HOMALS es equivalente
a la del Analisis de Correspondencias Multiples cuando todos los marginales fila de la
matriz indicadora completa[14] son iguales al numero de variables, es decir, cuando todos
los individuos cuentan con observaciones en todas las variables (cuando no hay datos
faltantes). Desde el punto de vista practico, la principal diferencia entre el ACM frances
y el HOMALS radica en el hecho de que en adicion a las cuantificaciones de las
categorias de las variables, el HOMALS tambien genera puntuaciones para los
individuos, las cuales pueden representarse conjuntamente con las cuantificaciones de
las categorias.

[14] La matriz G que se conforma por yuxtaposicion del conjunto de matrices G_j (cf. § 2.6, nota 12).

2.7 OVERALS

En adicion a la generalizacion que el HOMALS hace con respecto al ACP, en cuanto a
la posibilidad de generar una matriz de cuantificaciones, en lugar de un vector, el

OVERALS incorpora dos niveles de generalizacion adicionales. Por una parte, permite
trabajar con variables escaladas a diferente nivel (numerico, ordinal o nominal) y, por
otra parte, considera la posibilidad de que las variables esten estructuradas en K grupos.
Por tal motivo, se le conoce tambien como Analisis de Correlacion Canonica no Lineal.
La funcion de perdida esta dada por la expresion [2.3]:

$$\sigma(Y, Q) = \sum_{k=1}^{K} SSQ\left(Y_{J_k} G_j Q_j\right) \quad [2.3]$$

Donde:
$\sigma(Y, Q_j)$: Funcion de perdida en los parametros Y y Q_j.
K: Numero de grupos.
SSQ: Suma de cuadrados.
Y : Matriz de puntuaciones de los individuos.
J_k: Conjunto de indices de las variables pertenecientes al k-esimo grupo (k=1,
2, ..., K).
G_j: Matriz indicadora correspondiente a la j-esima variable del k-esimo grupo.
Q_j: Matriz de cuantificaciones de las categorias de la j-esima variable del grupo
k-esimo, satisfaciendo las restricciones propias de cada nivel de escalamiento.
El algoritmo OVERALS utiliza minimos cuadrados alternados para minimizar la
funcion de perdida [2.3], admitiendo que los conjuntos puedan diferir tanto en el
numero de variables que los conforman como en la definicion de estas, esto es, en las
restricciones impuestas por el nivel de escalamiento de cada variable. El algoritmo es

tan general que no exige ni siquiera que los K conjuntos de variables sean exhaustivos
ni mutuamente excluyentes.
Para la minimizacion de la funcion de perdida [2.3] se requiere obtener simultaneamente el parametro relacionado con el modelo de correlacion canonica (Y) y
el parametro de escalamiento, que surge a partir de las transformaciones (Q_j). El
problema OVERALS se resuelve alternativamente para cada uno de los parametros,
manteniendo el otro constante. Dados unos valores para el parametro de escalamiento
(Q_j), puede obtenerse Y, la matriz con las puntuaciones de los objetos. En el siguiente
paso se actualiza Q_j, manteniendo Y constante. Estos dos pasos corresponden a un ciclo
de iteracion. La solucion para los diferentes parametros es discutida con detalle por Van
der BURG et al. (1994).
El OVERALS fue concebido por De LEEUW (1973), pero su presentacion formal aparece en el texto de GIFI (1981). En posteriores publicaciones de algunos de los
miembros del grupo Gifi aparecen detalles sobre el algoritmo utilizado para su
ejecucion (Van der BURG et al., 1988; Van der BURG et al., 1994).
El OVERALS constituye el modelo mas general del sistema Gifi, por cuanto permite
trabajar con variables escaladas a cualquier nivel y conformando cualquier numero de
grupos. En tal sentido, todos los demas modelos pueden verse como casos particulares
de este; de ahi la primera parte de su nombre (OVER). La segunda parte del acronimo
(ALS) corresponde al algoritmo utilizado para su solucion (***A****lternating* ***L****east* ***S****quares*).
Los modelos que incluye el OVERALS, y que pueden verse como casos particulares de
este, son PRINCALS (Analisis de Componentes Principales no Lineal), que surge

cuando cada grupo esta conformado por una sola variable, es decir, cuando no hay
agrupamiento. Si en este caso, todas las variables se escalan a nivel numerico, se tiene el
ACP; si todas las variables se escalan a nivel nominal, se tendria el HOMALS; si en el
caso anterior se tienen solo dos variables, se llega al ANACOR (Analisis de
Correspondencias) (MEULMAN et al., 2004). Adicionalmente, el modelo CORALS
(Analisis de Correlacion Canonica no Lineal) es un caso especial de OVERALS, que
aparece cuando K=2. Si en este caso todas las variables se escalan a nivel numerico, se

tiene el Analisis de Correlacion Canonica (ACC). Las anteriores relaciones se resumen
en la Figura 2.2.

Figura 2.2. Relacion entre el OVERALS y otras tecnicas del sistema Gifi.

CAPÍTULO 3

Efecto De La Dimensionalidad Sobre Las Cuantificaciones

3.1 INTRODUCCIÓN

Tras la breve presentacion en el Capitulo 2 del sistema Gifi de analisis multivariante no lineal y de su papel en la generacion de cuantificaciones, **en este capítulo** se mostrara como el modelo utilizado y, en particular, la dimensionalidad de la solucion afectan las cuantificaciones obtenidas.

Inicialmente se exponen los dos usos principales de las tecnicas del sistema Gifi: como tecnicas generadoras de cuantificaciones y como tecnicas de reduccion de dimensionalidad.

Tras definir las cuantificaciones simples y las cuantificaciones multiples, se ilustra mediante un ejemplo hipotetico de pacientes con cancer de laringe el efecto de la dimensionalidad de la solucion sobre las cuantificaciones simples.

3.2 DIMENSIONALIDAD DE LAS SOLUCIONES

Las tecnicas que conforman el Sistema Gifi buscan, en general, satisfacer dos objetivos:

1) Servir como primer paso para la obtencion de transformaciones y/o cuantificaciones que permitan la subsecuente aplicacion de tecnicas lineales.

2) Generar representaciones en subespacios de baja dimensionalidad, que resuman las relaciones existentes entre objetos, variables y/u objetos y variables.

Una cuestion omnipresente al usar tecnicas de reduccion de dimensionalidad de sistemas multivariantes es la eleccion de p, la dimension del subespacio de representacion. Puesto que las tecnicas de reduccion de dimensionalidad son tecnicas de resumen, es valido, en general, el criterio aplicable a cualquier resumen: que sea lo

menor posible, recogiendo lo mas esencial del todo resumido. No obstante, esta es una
recomendacion demasiado general, que debe precisarse con base en otras consideraciones.
GIFI (1990) argumenta que p=1 seria la eleccion natural cuando el enfasis del
investigador es hacia el primer objetivo; mientras que si el enfasis esta en la satisfaccion
del segundo objetivo, frecuentemente se elige p=2.
Cuando el objetivo es la cuantificacion como primer paso para el posterior uso de otras
tecnicas, consideramos, en contraste con lo anotado por GIFI (1990) sobre la 'natural'
eleccion de p=1, que **la elección de la dimensionalidad de la solución dista de ser
trivial, dado que las cuantificaciones obtenidas dependen de la dimensionalidad
elegida**.
Por otra parte, cuando el principal objetivo es generar representaciones bidimensionales,
el hecho de que p=2 sea una eleccion frecuente no significa que esta sea la 'mejor'
eleccion. En ocasiones puede ser deseable obtener un mayor numero de dimensiones
que permitan explorar relaciones pobremente reflejadas en el primer plano principal
(desde luego, a traves del analisis de los demas planos). A la incertidumbre que conlleva

la eleccion de p en cualquier tecnica de reduccion de dimensionalidad, habra que
anadirle el ingrediente de que algunas de las soluciones obtenidas a partir del sistema
Gifi para diferentes dimensionalidades no son anidadas [15], por lo que las
recomendaciones generales dadas para la eleccion de p podrian presentar aun mas
inconvenientes en este caso.
[15] Esto significa que las primeras dimensiones de un par de soluciones de diferente dimensionalidad no
coinciden.

3.3 TIPOS DE CUANTIFICACIÓN

Las cuantificaciones generadas mediante las tecnicas del Sistema Gifi que se basan en la minimizacion de la funcion de perdida del encuentro (*meet*), esto es, Analisis de Componentes Principales no Lineal (PRINCALS) y Analisis de Correlacion Canonica no Lineal (OVERALS), pueden ser multiples o simples, a eleccion del usuario (GIFI, 1990).

Una **cuantificación múltiple** significa que cada categoria de una variable recibe un numero p de cuantificaciones igual a la dimensionalidad de la solucion. En este caso, en que la relacion entre la dimensionalidad de la solucion y las cuantificaciones obtenidas es directa, se opta, como practica general, por un valor bajo de p, dado que se esta en el marco de tecnicas de *reduccion* de dimensionalidad.

Cuando se trabaja con **cuantificaciones simples**, cada categoria recibe una unica cuantificacion, sin importar el valor de p, hecho que puede enmascarar la relacion existente entre la dimensionalidad elegida y las cuantificaciones obtenidas y, por consiguiente, brindar una sensacion de tranquilidad al analista que opta por la ‘natural’ dimensionalidad p=1 para generar las cuantificaciones.

Los algoritmos PRINCALS y OVERALS calculan inicialmente p cuantificaciones para cada una de las variables (con p elegido por el usuario), sin importar si las variables han de recibir cuantificaciones simples o multiples (GIFI, 1990). Para las variables que el usuario defina como simples, el paso de las cuantificaciones multiples a las simples se obtiene imponiendo **restricciones de rango uno**; esto es, proyectando la matriz de cuantificaciones multiples de la correspondiente variable sobre su respectivo vector de

saturaciones, tal y como indican MICHAILIDIS y De LEEUW (1998). Para el efecto se
utiliza la expresion [3.1]:

$$y_j = Y_j c_j \quad [3.1]$$

Donde:

y_j: Cuantificaciones simples de la j-esima variable

Y_j: Matriz K_j x p, con las cuantificaciones multiples de la j-esima variable

c_j: Vector de saturaciones de la j-esima variable

La anterior expresion muestra que en la generacion de las cuantificaciones simples de
cada variable, participan todas las dimensiones de la solucion multiple en proporcion a
la magnitud de la saturacion de la variable en la correspondiente dimension. Luego,
cuando se opta por cuantificaciones simples basadas unicamente en la primera
dimension (p=1), se esta prescindiendo de la informacion contenida en las demas
dimensiones, lo cual afecta especialmente a las variables cuyas principales saturaciones
no estan en la primera dimension. En tal sentido, puede resultar deseable enriquecer las
cuantificaciones con aportes de otras dimensiones, de manera que no solo las variables
con altas saturaciones en la primera dimension se cuantifiquen adecuadamente.

3.4 EFECTO DE LA DIMENSIONALIDAD SOBRE LAS CUANTIFICACIONES

Para ilustrar el efecto de la dimensionalidad de la solucion sobre las cuantificaciones,
considerese un estudio hipotetico sobre calidad de vida en pacientes tratados por cancer

de laringe, en el que se evaluan, entre otras, las siguientes cuatro variables:

A: Dolor

a1: No hay dolor
a2: Dolor leve que no requiere medicamentos
a3: Dolor moderado controlable con analgesicos suaves
a4: Dolor severo solo controlable con codeina/morfina
a5: Dolor intenso que no se controla con ningun medicamento

B: Actividad

b1: Esta tan activo como siempre
b2: Ocasionalmente no puede mantener su anterior nivel de actividad
b3: Frecuentemente se siente cansado y mas lento para realizar actividades
b4: No sale de casa porque no tiene la fuerza necesaria
b5: No sale de casa, en donde habitualmente permanece en cama o sentado

C: Masticación

c1: Puede masticar tan bien como antes
c2: Puede comer alimentos solidos suaves, pero no puede masticar algunas comidas
c3: No puede masticar ni siquiera solidos suaves
c4: No tolera ningun alimento por via oral

D: Habla

d1: Puede hablar tan bien como antes
d2: Dificultad para decir algunas palabras, pero se entiende cuando habla por
telefono
d3: Solo la familia y amigos logran entenderle
d4: Nadie entiende cuando intenta hablar

Se simulo una muestra multivariante de 15 variables y 1.000 observaciones, la cual fue
procesada mediante Analisis de Componentes Principales Categorico (PRINCALS),
escalando cada una de las variables a nivel ordinal simple. Se obtuvieron soluciones con
p=1 y con p=2. A fin de ilustrar el aspecto de interes, manteniendo la presentacion lo
mas sencilla posible, solo se muestran los resultados de las cuatro variables
referenciadas.

En la Tabla 3.1 y en la Figura 3.1 se presentan las cuantificaciones para las categorias
de las cuatro variables en cuestion, basadas en las soluciones con p=1 y p=2.

Variable	Categoría	Cuantificación p=1	Cuantificación p=2
A	a1	-2,86	-3,46
	a2	-1,33	-1,45
	a3	-0,07	-0,01
	a4	1,41	1,43
	a5	4,85	3,14
B	b1	-2,67	-3,09
	b2	-1,19	-1,32
	b3	-0,09	-0,01
	b4	1,26	1,31
	b5	3,74	2,58
C	c1	-0,10	-4,25
	c2	-0,10	-1,35
	c3	-0,10	0,64
	c4	9,95	1,78
D	d1	-1,97	-2,32
	d2	0,08	0,16
	d3	1,16	1,24
	d4	8,12	2,15

Tabla 3.1. Cuantificaciones para las categorias de las variables *A*, *B*, *C* y *D*, basadas en
soluciones PRINCALS con p=1 y p=2.

Figura 3.1. Cuantificaciones para las categorias de las variables *A*, *B*, *C* y *D*, basadas en
soluciones PRINCALS con p=1 y p=2.

A primera vista se evidencia que las dos soluciones (p=1 y p=2) generan diferentes
cuantificaciones. La magnitud de tales variaciones puede explicarse con ayuda de las
saturaciones que se presentan en la Tabla 3.2.

Variable	p=1 Dimensión 1	p=2 Dimensión 1	p=2 Dimensión 2
A	0,91		
		0,91	0,07
B	0,90		

0,92 0,06
C
0,01
-0,20 0,81
D
0,26
0,06 0,84

Tabla 3.2. Saturaciones para las variables *A*, *B*, *C* y *D*, basadas en soluciones PRINCALS con p=1 y p=2.

Tal y como ya se habia indicado, las soluciones PRINCALS no son anidadas, es decir que las correspondientes dimensiones de un par de soluciones de diferente

dimensionalidad no coinciden necesariamente (cf. § 3.2, nota 15). Ello explica las diferencias observadas entre las saturaciones de la primera dimension para las soluciones con p=1 y con p=2.

La Tabla 3.1 y la Figura 3.1 muestran que, en general, las cuantificaciones obtenidas con diferentes valores de p no son las mismas. En particular, al comparar las cuantificaciones obtenidas con base en la solucion con p=1 y la solucion con p=2, puede observarse que los cambios son mas drasticos cuanto menor es la saturacion de la variable en la primera dimension de la solucion con p=1.

Observese, por ejemplo, que las cuantificaciones de las categorias de las variables A y B son las mas estables. Esto es debido al bajo aporte que hace la segunda dimension a estas variables y a que estas variables tienen una alta saturacion en la primera dimension en la solucion con p=1.

En contraste, las cuantificaciones de las categorias de la variable C son las mas incongruentes, lo cual se explica por el hecho de la saturacion casi nula que tiene dicha variable sobre la primera dimension en la solucion con p=1. Notese, por ejemplo, que la

solucion basada en p=1 indica que las categorias c1, c2 y c3 de la variable C no se diferencian entre si, esto es, que para fines de caracterizacion de la calidad de vida en pacientes con cancer de laringe es lo mismo que el paciente pueda masticar tan bien como antes; que solo pueda masticar solidos suaves, o que no pueda masticar ni siquiera solidos suaves, diferenciandose del conjunto anterior solo el grupo de pacientes que no tolera ningun alimento por via oral. Por otra parte, la solucion basada en p=2 asigna diferentes cuantificaciones a cada una de las categorias de la variable *Masticacion*, lo que refleja de mejor forma el diferente papel jugado por cada una de estas condiciones en la calidad de vida del paciente.

Los cambios en las cuantificaciones de las categorias de la variable D, aunque son mucho mas drasticos que los exhibidos por las variables A y B, no lo son tanto como los observados para la variable C. Ello se explica por el hecho de que su saturacion en la primera dimension, en la solucion con p=1, no es tan baja como la de la variable C.

Estas incongruencias en las cuantificaciones se trasladan a cualquier tecnica basada en las mismas, generando diferentes resultados, segun se trabaje con las cuantificaciones basadas en p=1 o en las basadas en p=2, por mencionar solo los dos conjuntos de cuantificaciones obtenidos en este ejemplo.

Para el efecto, basta con observar la Tabla 3.3, en la que se muestran las matrices de correlaciones reducidas (solo para las cuatro variables de interes), entre las variables cuantificadas con base en p=1 y con base en p=2. Notese que la mayor incongruencia es

la de la correlacion lineal entre las variables C y D (*Masticacion* y *Habla*), las dos con
menor saturacion en la primera dimension y, por tanto, con cuantificaciones menos
coherentes.

	***p*=1**				***p*=2**			
	A	B	C	D	A	B	C	D
A	1,00	0,74	-0,02	0,11	1,00	0,74	-0,11	0,10
B	0,74	1,00	-0,00	0,09	0,74	1,00	-0,09	0,08
C	-0,02	-0,00	1,00	**0,10**	-0,11	-0,09	1,00	**0,54**
D	0,11	0,09	0,10	1,00	0,10	0,08	0,54	1,00

Tabla 3.3. Matrices de correlaciones reducidas para las variables A, B, C y D,
cuantificadas con base en PRINCALS con p=1 y p=2.

Para ilustrar con mayor detalle la forma en que las incongruencias en las cuantificaciones pueden trasladarse a otras tecnicas produciendo, a su vez, resultados
incongruentes, dependiendo de si se trabaja con uno u otro conjunto de cuantificaciones,
supongase que se desea construir un indice de calidad de vida de los pacientes con
cancer de laringe, basado en las cuatro variables descritas. Puesto que tras la
cuantificacion, todas las variables tienen propiedades metricas, puede utilizarse un
indice basado en la puntuacion de los individuos sobre la primera componente principal
de un ACP. Para facilitar su interpretacion, utilizaremos el inverso aditivo de la primera
componente principal, de manera que los valores altos indiquen alta calidad de vida y
viceversa.

En la Tabla 3.4 se muestran dos pacientes hipoteticos y sus respectivos indices de
calidad de vida (ICV), calculados como la inversa aditiva de la primera componente de
sendos ACP realizado sobre las cuatro variables cuantificadas a partir de PRINCALS
con p=1 y con p=2.

Paciente	**Combinación**

de categorías
ICV1
(basado en *p*=1)
ICV2
(basado en *p*=2)
Paciente 1 a2, b3, c2, d1 1,35 0,99
Paciente 2 a2, b2, c3, d4 0,27 1,89
Tabla 3.4. Indices de calidad de vida (ICV) para dos hipoteticos pacientes con cancer
de laringe, basados en sendos ACP para las cuantificaciones PRINCALS con *p*=1 y
p=2.

La Tabla 3.4 muestra que el indice de calidad de vida basado en la solucion con *p*=1 es
mayor para el paciente 1 que para el paciente 2. Contrario a este resultado, al calcular el
indice con base en la solucion con *p*=2, aparece que el paciente 2 goza de mayor calidad
de vida con relacion al paciente 1.

Este ejemplo no hace mas que poner en evidencia lo que puede deducirse del analisis de
la expresion [3.1]: **las soluciones basadas en diferentes valores de *p* generan
diferentes cuantificaciones**. Resulta claro ademas que la solucion obtenida con base en
p=1 solo resultaria adecuada si todas las variables tuvieran una alta saturacion sobre la
primera dimension. En tal sentido no consideramos que esta sea la solucion 'natural' y
estimamos conveniente evaluar las cuantificaciones basadas en otros valores de *p*, de
manera que no solo las variables con altas saturaciones sobre la primera dimension
queden adecuadamente cuantificadas.

3.5 ELECCIÓN DE LA DIMENSIONALIDAD EN PROBLEMAS DE CUANTIFICACIÓN

Tal y como acaba de ilustrarse en el ejemplo anterior (§ 3.4), las cuantificaciones
simples generadas mediante las tecnicas del sistema Gifi para un sistema multivariante

mixto dependen de la dimensionalidad de la solucion utilizada para su generacion.
Consideramos que este problema ha pasado desapercibido debido a que las tecnicas del sistema Gifi, ademas de generar cuantificaciones, sirven por si mismas para analizar los datos sin que se requiera utilizar las cuantificaciones generadas como datos de entrada de otras tecnicas. En tal sentido, se da la situacion que ya anotabamos en el Capitulo 1 (cf. § 1.7) cuando nos referiamos al hecho de que a menudo las cuantificaciones no constituyen un fin en si mismas, sino que forman parte integral de la tecnica, apareciendo como un subproducto del proceso o llegando incluso a ser totalmente ignoradas.
En tales casos, cuando el objetivo fundamental es la representacion en un espacio de baja dimensionalidad, el analista elige p con base en las consideraciones habituales o en alguna modificacion de las mismas (TIMMERMAN and KIERS, 2000; TUCKER, 1966, CATTELL, 1966; KAISER, 1960, HORN, 1965; VELICER, 1976; ZWICK and VELICER, 1986), sin preocuparse por las cuantificaciones generadas.
No existe, sin embargo, ninguna razon para creer que estas consideraciones puedan resultar adecuadas cuando el analista tiene interes en obtener el mejor conjunto de escalamientos y/o cuantificaciones simples de un grupo de variables, que le permitan analizarlas mediante tecnicas lineales. Por el contrario, consideramos inadecuado limitar la tecnica generadora de cuantificaciones con base en restricciones dirigidas a satisfacer otros objetivos.
Podemos afirmar, en terminos generales, que cuando se usan las tecnicas del sistema Gifi con el objetivo de generar escalamientos y/o cuantificaciones de las variables para su posterior uso en otras tecnicas lineales, debe utilizarse la dimensionalidad p que

genere cuantificaciones optimas.

Como ya se anoto anteriormente, el calificativo *optimo* que acompana los escalamientos
generados mediante las tecnicas del sistema Gifi solo debe entenderse como tal en el
sentido en que minimiza la perdida de informacion inherente a la reduccion de la
dimensionalidad del sistema original. Notese que si el objetivo no es reducir la
dimensionalidad, sino solamente obtener cuantificaciones que puedan utilizarse
posteriormente en otras tecnicas, podria ser mas conveniente partir de una definicion de
optimo acorde con tales objetivos.
En resumen, el problema en el que nos centraremos y cuya solucion constituye la base
de los desarrollos que se presentan en los dos siguientes capitulos, radica en el hecho de
que **al usar las técnicas del sistema Gifi para generar escalamientos y/o cuantificaciones simples para un sistema multivariante, se obtiene un conjunto de
posibles cuantificaciones diferentes entre sí, en función de la dimensionalidad de la
solución elegida**, no existiendo ningun criterio para establecer cual de los conjuntos de
cuantificaciones resultantes es el optimo.

CAPÍTULO 4
Cuantificación Óptima

4.1 INTRODUCCIÓN

Tras establecer en el capitulo anterior el problema de la generacion de diferentes
conjuntos de cuantificaciones cuando se utilizan diferentes dimensionalidades en las
soluciones de algunas de las tecnicas del sistema Gifi, **en este capítulo** planteamos una
solucion a este problema.

Empezamos presentando una definicion general de cuantificacion optima que no depende de ningun modelo de analisis. Esta definicion, asi como todos los desarrollos que se desprenden de la misma son aportes originales de este trabajo. Seguidamente, se utilizan un par de ejemplos hipoteticos para ilustrar la logica subyacente en la definicion, los cuales han sido cuidadosamente elegidos para recrear algunos de los escenarios mas comunes. Aunque es posible seguir la solucion planteada pasando directamente al siguiente apartado (§ 4.4), los razonamientos presentados en estos ejemplos ilustrativos constituyen en si mismos valiosos aportes, que no se encuentran en ningun otro lugar del texto. Invitamos, por tanto, a seguir el desarrollo de la propuesta tal y como esta presentado.

Basada en la definicion general de cuantificacion optima, se presenta luego una definicion especifica, la cual, aunque se representa en terminos matematicos, sigue sin depender de ningun modelo. Seguidamente, se desarrolla una propuesta especifica para resolver el problema planteado en el capitulo anterior. Esta propuesta consiste en elegir la solucion que mejor satisfaga la condicion de cuantificacion optima, acorde con los criterios presentados. Se discute el alcance de la propuesta y se bosqueja su implementacion algoritmica.

Finalmente, se ilustra como aplicar la solucion planteada, sin avanzar en el analisis contextualizado de los datos, pues en el Capitulo 6 analizaremos una aplicacion completa con base en el desarrollo general que se expone en el Capitulo 5.

4.2 DEFINICIÓN GENERAL DE CUANTIFICACIÓN ÓPTIMA

Tomando como escenario de referencia aquel en el que se usan las tecnicas del sistema
Gifi con el objetivo de obtener el mejor conjunto de cuantificaciones simples de un
sistema multivariante, y teniendo en cuenta que el calificativo optimo que acompana
tales escalamientos solo debe entenderse como tal en el sentido en que minimiza la
perdida de informacion que se produce al reducir la dimensionalidad del sistema
original con base en un modelo de analisis dado, presentaremos una definicion de
cuantificacion optima mas acorde con nuestros objetivos.

Definición 4.1 (Cuantificación Óptima, general): La cuantificacion optima de un sistema multivariante es la que asigna valores mas similares a
las categorias de las distintas variables que se asocian con mayor
frecuencia.

Notese que los algoritmos en los que se basan las tecnicas que estamos considerando
(PRINCALS, OVERALS y todos sus casos particulares) exigen que todas las variables
sean discretizadas o categorizadas, quedando constituida cada variable por un numero
determinado de *categorias*. Es por ello que la anterior definicion habla indistintamente
de *categorias*, sin que importe el nivel de escalamiento de las variables.
La definicion 4.1 constituye la base de todos los desarrollos posteriores. Seguidamente,
ilustraremos la idea subyacente a traves de un par de aplicaciones hipoteticas antes de
presentar su formulacion matematica.

4.3 ILUSTRACIÓN DE LA DEFINICIÓN GENERAL

EJEMPLO ILUSTRATIVO 1.

Considerese una base de datos hipotetica, constituida por la talla en centimetros de 25
varones y de 25 mujeres (Tabla 4.1).

Talla Sexo Talla Sexo Talla Sexo Talla Sexo Talla Sexo

174 V 163 V 173 V 169 V 172 V
168 M 171 V 175 V 170 V 168 V
172 V 161 M 176 V 161 M 163 M
165 M 174 V 174 V 164 M 166 M
152 M 165 M 175 V 160 M 163 M
166 V 163 M 174 V 166 M 167 V
168 M 162 M 178 V 164 M 180 V
169 M 163 M 155 M 183 V 168 V
173 V 168 M 166 M 172 M 177 V
176 V 171 M 164 M 171 V 162 M

Tabla 4.1. Talla en centimetros de un grupo de 25 varones y 25 mujeres.

Como primer paso de cualquier proceso de cuantificacion optima mediante las tecnicas que estamos considerando, todas las variables deben ser categorizadas[16]. Supongase que la variable talla se categoriza asi: (150-154), (155-159), …, (175-179), (180-184). La Tabla 4.2 muestra las frecuencias de las combinaciones de las categorias de las variables talla y sexo.

Ahora que las dos variables estan categorizadas, procede la cuantificacion de sus correspondientes categorias. Para tal efecto, no nos basaremos en la definicion de cuantificacion optima, sino que realizaremos una aproximacion intuitiva.

[16] En § 1.5 se discuten algunos criterios de categorizacion.

Sexo	
Talla	
V	**M**
150-154 0	1
155-159 0	1
160-164 1	12
165-169 5	9
170-174 11	2
175-179 6	0
180-184 2	0

Tabla 4.2. Tabla de contingencia para talla categorizada y sexo.

Teniendo en cuenta la naturaleza numerica de la variable talla, la cuantificacion logica de cada una de sus categorias estara dada por su punto medio o marca de clase. Este vector de cuantificaciones, o cualquier transformacion lineal del mismo, es el que mejor

recoge la informacion relativa a las interdistancias entre cada uno de los niveles o
categorias de la variable talla, respetando sus propiedades metricas. En aras de la
simplicidad, no realizaremos por el momento ninguna normalizacion y cuantificaremos
las categorias de la variable talla con base en el vector de las marcas de clase, esto es,
[152, 157, 162, 167, 172, 177, 182].
Puesto que el unico referente para la cuantificacion de las categorias de la variable sexo
es la variable talla (o la cuantificacion de sus categorias), la eleccion natural consiste en
cuantificar cada categoria con base en la talla promedio de los individuos en tal
categoria. En vista de que las categorias de la variable talla ya han sido cuantificadas,
cada una de las categorias de la variable sexo se cuantifica usando la siguiente
expresion:

$$qs_l = \frac{\sum_{k=1}^{K} f_{kl}\, qt_k}{\sum_{k=1}^{K} f_{kl}}, \quad l = V, M \qquad [4.1]$$

Donde:
qs_l: Cuantificacion de la categoria l de la variable sexo ($l = V, M$).
qt_k: Cuantificacion de la k-esima categoria de la variable talla (marca de clase).

f_{k1}: Frecuencia de aparicion conjunta de la k-esima categoria de la variable talla y la
categoria 1 de la variable sexo.
Luego, las cuantificaciones para las categorias V y M de la variable sexo, calculadas
mediante la expresion [4.1], son: q_{SV} = 172,6 y q_{SM} = 164.
Observese que este conjunto de cuantificaciones, que es sin duda el mas natural para
este ejemplo, se ha obtenido de manera absolutamente intuitiva, no habiendo forzado su
obtencion al cumplimiento de la definicion 4.1. No obstante, puede verificarse que el
razonamiento seguido conduce a un conjunto de cuantificaciones que satisface la
condicion especificada en la definicion de cuantificacion optima; esto es, asigna valores
mas similares a las categorias de las dos variables que se asocian con mayor frecuencia.
Asi, por ejemplo, al comparar la cuantificacion de la categoria V (q_{SV} = 172,6) con las
cuantificaciones de las categorias de la variable talla, se observa que a la que mas se
acerca es a la de la categoria 170-174 (q_{t5} = 172), siendo esta la categoria de la variable
talla a la que la categoria V se asocia con mayor frecuencia (f_{5V} = 11). De igual modo, al
comparar la cuantificacion de la categoria M (q_{SM} = 164) con las cuantificaciones de las
categorias de la variable talla, encontramos que a la que mas se acerca es a la de la
categoria 160-164 (q_{t3} = 162), siendo esta la categoria de la variable talla a la que la
categoria M se asocia con mayor frecuencia (f_{3M} = 12).
Logicamente, la similaridad a la que hace referencia la definicion de cuantificacion
optima puede entenderse en terminos de distancia (CUADRAS, 1991), lo que nos
permite avanzar en el analisis de su cumplimiento. En un sistema multivariante
cuantificado optimamente, las distancias entre las categorias cuantificadas de las

diferentes variables seran menores para aquellas categorias que se asocien con mayor
frecuencia. En tal sentido, cabe esperar que al sumar, sobre todas las observaciones, las
distancias entre las categorias cuantificadas optimamente, se obtenga una cifra inferior a
la que se obtendria con cualquier otro sistema de cuantificaciones.

Observando la expresion [4.1], resulta evidente que siempre que se cuantifique cada
categoria de la variable sexo con base en la media de la variable talla para tal categoria,
se obtendra una cuantificacion cercana a la de la(s) categorias(s) de la variable talla mas
frecuentes para la correspondiente categoria de sexo. De hecho, puede demostrarse que
de todas las posibles cuantificaciones, esta es la que minimiza la suma sobre todas las
observaciones de las distancias cuadraticas entre las cuantificaciones de las categorias
de la variable talla y las de las categorias de la variable sexo.

Supongase que qs_l es la cuantificacion buscada para una categoria determinada de la
variable sexo (l=V o l=M), qt_k es la cuantificacion de la k-esima categoria de la variable
talla y f_{kl} es la frecuencia de aparicion conjunta de las dos categorias consideradas. Si se
define D como la suma sobre todas las observaciones de las distancias al cuadrado entre
las cuantificaciones de las categorias de las dos variables en cuestion, tenemos:

$$D = \sum_{i=1}^{n} (qt - qs)^2 = \sum_{l'} \sum_{l} (qt_{l'} - qs_l)^2 f_{l'l}$$

$qt\ qs\ qt\ qs$

==

$=\Sigma - + \Sigma -$

$\left(\left(\right) \right) \left(\left(\right) \right)$ 2 2

, ,

1 1

K K

kl k l kl k l

k k

$f\ qt\ qs\ f\ qt\ qs$

==

$=\Sigma - + \Sigma -$

$\left(\right)$ 2 2 $\left(\right)$

, , , , , 2 2

1 1

2 2

K K

kl k kl k l kl l kl k kl k l kl l

k k

$f\ qt\ f\ qt\ qs\ f\ qs\ f\ qt\ f\ qt\ qs\ f\ qs$

==

$=\Sigma - + + \Sigma - +$

2 2 2 2

, , , , ,

1 1 1 1 1 1

2 2

K K K K K K

kl k l kl k l kl kl k l kl k l kl

k k k k k k

$f\ qt\ qs\ f\ qt\ qs\ f\ f\ qt\ qs\ f\ qt\ qs\ f$

======

$= - + + - +$

$\Sigma\ \Sigma\ \Sigma\ \Sigma\ \Sigma\ \Sigma$

La derivada de esta expresion, con respecto a qs_1 es:

1 1

2 2

K K

kl k l kl

l k k

dD

$f\ qt\ qs\ f$

dqs ==

$= - \Sigma + \Sigma$

Esta derivada se hace cero cuando:

1 1

$$\sum_{k}^{K} f_{kl} \, q_k \, q_{sl} = \sum_{k}^{K} f_{kl} \, q_k$$

$$q_{sl} = \frac{\sum_{k}^{K} f_{kl} \, q_k}{\sum_{k}^{K} f_{kl}}$$

Este resultado para q_{sl} es exactamente igual al obtenido mediante la expresion [4.1], es decir, a la media de las cuantificaciones de las categorias de la variable talla asociadas con la correspondiente categoria de la variable sexo. Luego, queda demostrado que la cuantificacion de una categoria cualquiera de la variable sexo obtenida como la media de las cuantificaciones de las categorias de la variable talla observadas para la correspondiente categoria de la variable sexo, minimiza la suma sobre todas las observaciones de las distancias cuadraticas entre las cuantificaciones de las categorias de la variable talla y las de las categorias de la variable sexo.

Notese ademas que la cuantificacion de las categorias de una variable con base en la cuantificacion promedio de las categorias de la otra variable sigue el mismo principio que el metodo de los Promedios Reciprocos (*Reciprocal Averaging*; cf. § 2.6).

Este ejemplo ilustra como **la cuantificación natural de las categorías de una variable nominal, basada en el promedio ponderado de las cuantificaciones de las otras**

categorías con las que se combine, además de corresponder con la solución de los
Promedios Recíprocos para el AC, satisface la condición de cuantificación óptima
especificada en la definición 4.1, asignando valores mas similares a las categorias de
las distintas variables que se asocian con mayor frecuencia.

EJEMPLO ILUSTRATIVO 2.

Supongase ahora que en la base de datos del ejemplo ilustrativo 1, ademas de la talla y
el sexo se cuenta con informacion sobre el peso de cada uno de los 50 individuos.

Podemos llevar estos datos al mismo escenario del ejemplo 1, condensando las dos
variables numericas (talla y peso) en una sola. A partir de alli, la cuantificacion para las
categorias de la variable sexo se realizaria de la misma forma indicada en el ejemplo 1,
por lo que no redundaremos en ello.

Para la condensacion de las variables numericas, podria pensarse en la simple
utilizacion de su media aritmetica. No obstante, podemos reducir la perdida de
informacion fruto de tal proceso, mediante un escalamiento o transformacion previa de
las variables. GIFI (1990) indica que dicha perdida de informacion se minimiza al
remplazar el sistema numerico original por su primera componente principal. Esta seria,
por tanto, la eleccion logica. La variable numerica condensada tendria la siguiente
forma:

$_{1\ 2}( )\ (\ Y_1 = v\ tallaest + v\ pesoest)$ [4.2]

Donde:

Y_1: Vector de la primera componente principal de las variables peso y talla.

v_1, v_2: Constantes.

tallaest: Vector de talla centrada y estandarizada.

pesoest: Vector de peso centrado y estandarizado.

Tal y como se desprende de la definicion 2.1 (§ 2.3), puesto que Y_1 resume las variables
talla y peso con minima perdida de informacion, Y_1 es el promedio de las variables talla
y peso escaladas optimamente. Se satisface, por tanto, la siguiente relacion:

$$1\ 2\ (\ )\ (\ )$$
$$2$$
$$tallaopt\ pesoopt$$
$$v\ tallaest\ v\ pesoest$$
$$+$$
$$+=$$

Donde, *tallaopt* y *pesoopt* corresponden a las variables talla y peso escaladas
optimamente, las cuales pueden expresarse en terminos de las variables centradas y
estandarizadas, asi:

$$tallaopt = 2v_1\ (tallaest)$$
$$_2\ pesoopt = 2v\ (\ pesoest)$$

Puesto que en el presente contexto todas las variables, sin importar su nivel de
escalamiento, son categorizadas o discretizadas, nos interesa obtener la cuantificacion
optima de cada categoria o nivel. Tomando como base los escalamientos optimos de las
variables talla y peso, podemos escribir la siguiente expresion general para las
cuantificaciones de cada categoria:

$$(\ ),\ 1,\ 2,\ldots,\ ;\ 1,\ 2,\ldots,\ _{jk\ j\ jk}\ qx = mv\ x\ k = K\ j = m \quad [4.3]$$

Donde:

qx_{jk}: Cuantificacion del k-esimo nivel o categoria de la j-esima variable.
m: Numero de variables.
v_j: Constante multiplicativa de la j-esima variable en la combinacion lineal
que genera la primera componente principal.
x_{jk}: k-esimo nivel centrado y estandarizado de la j-esima variable.

Consideremos, ahora, la perspectiva de PEARSON (1901) para la obtencion de la

primera componente principal de un conjunto de variables, segun la cual, la primera componente principal es la linea que minimiza la suma de proyecciones cuadraticas de los puntos de la nube multidimensional. En el caso de dos variables continuas, esto es facilmente visualizable como una linea que atraviesa el diagrama de dispersion en la direccion de su tendencia principal o maxima dispersion, haciendo minima la suma de proyecciones cuadraticas de los puntos a la misma, como se muestra en la Figura 4.1.

Figura 4.1. Primera componente principal construida a partir de la minimizacion de la suma de proyecciones cuadraticas de los puntos de la nube multidimensional.

Al trabajar con variables discretizadas, la visualizacion ya no es tan directa, puesto que el diagrama de dispersion consistira en una malla de puntos que representan cada una de las posibles combinaciones, con eventuales huecos correspondientes a combinaciones no presentes en la muestra. Podemos ayudarnos, sin embargo, de un diagrama de dispersion modificado, en el que el tamano de cada punto sea proporcional a su frecuencia.

Considerese la hipotetica tabla de contingencia para las variables talla y peso (Tabla 4.3) y su diagrama de dispersion modificado asociado (Figura 4.2). Para facilitar la visualizacion de la correspondencia entre la tabla de contingencia y el diagrama de dispersion, se ha invertido el orden usual de las filas en la tabla de contingencia, presentandolas desde los niveles mas altos de la variable talla hasta los mas bajos.

En este caso, a diferencia de la Figura 4.1, no es facil visualizar la tendencia principal

del diagrama de dispersion. No obstante, es claro que la direccion de linea que satisfaga
la condicion de minimizar la suma de proyecciones cuadraticas se vera afectada por los
puntos de mayor ponderacion (frecuencia), tendiendo a pasar cerca de estos.

Peso / **Talla**	**41-50**	**51-60**	**61-70**	**71-80**	**81-90**	**91-100**	**101-110**
180-184	1	0	0	1	0	0	0
175-179	0	2	2	0	1	0	1
170-174	1	2	3	1	4	1	1
165-169	1	4	4	3	1	1	0
160-164	0	4	5	0	2	1	1
155-159	0	0	0	1	0	0	0
150-154	0	1	0	0	0	0	0

Tabla 4.3. Tabla de contingencia para las variables categorizadas talla y peso.

Figura 4.2. Diagrama de dispersion modificado para peso y talla estandarizados, y
primera componente principal.

Existe una clara relacion entre la informacion comun de un sistema multivariante
numerico continuo y su primera componente principal. La primera componente
principal de un sistema multivariante es la combinacion lineal de las variables originales
que mejor recoge la informacion comun contenida en estas. A mayor informacion

comun compartida por el conjunto de variables, mas cercanas se encontraran estas de la
primera componente principal. Esta cercania, en el caso de variables continuas, puede
medirse en una representacion Biplot (GABRIEL, 1971; GALINDO, 1985, 1986; GOWER and HAND, 1996; YAN and KANG, 2003) en terminos del angulo formado entre la componente principal y las correspondientes variables.

Es posible generalizar el concepto de informacion comun y cercania a la primera

componente principal en sistemas multivariantes con variables discretizadas. En este caso, nos interesa discutir la cercania de combinaciones especificas de categorias de diferentes variables (puntos de la nube multidimensional) a la componente principal.

Observese que las categorias que con mayor frecuencia se presentan conjuntamente conforman puntos de alta ponderacion, los cuales tienden a estar cerca de la primera componente principal. Esto puede visualizarse mejor en la Figura 4.3, en la que se han rotado y reescalado los ejes, quedando referenciado cada uno de los puntos al plano de las componentes principales. Las coordenadas de los puntos con frecuencia cero se obtienen incluyendolos como puntos suplementarios en el Analisis de Componentes Principales.

Figura 4.3. Representacion de las observaciones de talla y peso discretizadas en el plano de sus componentes principales.

Analogamente a la forma en que la cercania de un grupo de variables continuas a la primera componente principal (en terminos de los angulos) indica que estas comparten informacion comun, la cercania de un punto a la primera componente principal indica que las categorias representadas por dicho punto tienen cuantificaciones similares. En el presente ejemplo, podemos verificarlo mediante el analisis de la Figura 4.3, donde puede observarse la distancia de cada punto a la primera componente principal; la Tabla 4.4, donde se han obtenido, mediante la expresion [4.3], las cuantificaciones para cada una de las categorias de las variables talla y peso; y la Figura 4.4, en la que se

representan simultaneamente sobre una linea escalada tales cuantificaciones.

Talla Peso
Categorías Cuantificación Categorías Cuantificación
150-154 -3,66 **41-50** -2,18
155-159 -2,54 **51-60** -1,30
160-164 -1,42 **61-70** -0,42
165-169 -0,29 **71-80** 0,46
170-174 0,83 **81-90** 1,34
175-179 1,96 **91-100** 2,22
180-184 3,08 **101-110** 3,10

Tabla 4.4. Cuantificaciones de las categorias de las variables talla y peso.

Figura 4.4. Dispersion conjunta de las categorias cuantificadas de talla y peso.

Notese que los puntos que representan las combinaciones de categorias con cuantificaciones mas similares ({t:180-184, p:101-110}, {t:160-164, p:51:60}, {t:165-169, p:61-70}) son los mas cercanos a la primera componente principal (Figura 4.3), mientras que los puntos que representan las combinaciones de categorias con cuantificaciones mas disimiles ({t:150-154, p:101-110}, {t:180-184, p:41-50}) son los mas alejados de la primera componente principal (Figura 4.3).

En general, **la distancia de un punto a la primera componente principal es proporcional a la diferencia en las cuantificaciones de las categorías representadas por dicho punto**. Mientras mas cercano este un punto a la primera componente principal, mas similares seran las cuantificaciones de las categorias representadas por dicho punto. En particular, cuando un punto se encuentre exactamente sobre la primera componente principal, la cuantificacion de las categorias coincidira. A continuacion demostraremos el cumplimiento de dicha condicion para el caso bidimensional.

Sean X_1 y X_2 variables centradas y estandarizadas, con la siguiente matriz de correlaciones:

1

1
r
R
r
⎣ ⎤
≡ ⎡ |
⎨ ⎦
Donde $r = corr\ (X1,\ X2)$.
Las componentes principales del sistema se obtienen con base en las soluciones de la
siguiente ecuacion:
1 1
2 2
1
1
$r\ v\ v$
$r\ v\ v$
λ
⎣ ⎤⎣ ⎤⎣ ⎤
⎡ |⎡ |= ⎡ |
⎨ ⎦⎨ ⎦⎨ ⎦
[4.4]
Donde λ es un valor propio de R y $( )_{1\ 2}\ v' = v\ v$ es su correspondiente vector propio.
Los valores propios, λ, se obtienen como soluciones de la siguiente ecuacion:

2 2 (1)
(1)
(1)
r
r
r
λ
λ
λ
−
= − −
−
$=1+\lambda_{2} - 2\lambda - r_{2}$
$= \lambda_{2} - 2\lambda + \left(1 - r_{2}\right)$

Se resuelve esta ecuacion de segundo grado en λ:

(2) 2₂ 4(1)(1 ₂)

2(1)

r λ

− − □} − −

=

2 4 ₂

2

r λ

□}

=

2 2

2

r λ

□}

=

λ=1□} r

Si $r > 0$, entonces $(1+r) > (1-r)$, por lo que $( )_1 \lambda = 1+r$ sera el primer valor propio y

$( )_2 \lambda = 1-r$ sera el segundo valor propio. Si $r < 0$, el orden de los valores propios sera

inverso, esto es, $( )_1 \lambda = 1-r$ y $( )_2 \lambda = 1+r$.

Suponiendo que $r > 0$ y remplazando el primer valor propio en la ecuacion [4.4], se

obtiene el primer vector propio, v_1:

$_1 ( )_1$

$_2 _2$

1

1

1

r v v

r

r v v

⎣⎤⎣⎤⎣⎤

⎧ ⎪⎧ ⎪= + ⎧ ⎪

⎨ ⎭⎨ ⎭⎨ ⎭

$( )_1 {}_2 {}_1$ $v + rv = 1+r$ v

$( )_1 {}_2 {}_2$ $rv + v = 1+r$ v

El desarrollo de cualquiera de las dos ecuaciones conduce al siguiente resultado:

$v_1 - (1+r)v_1 = -rv_2$

$v_1 - v_1 - rv_1 = -rv_2$

$-rv_1 = -rv_2$

$v_1 = v_2 = v$

$()_1 \therefore v' = v\ 1\ 1$

Siguiendo con el supuesto de que $r > 0$ y remplazando el segundo valor propio en la
ecuacion [4.4], se obtiene el segundo vector propio, v_2:

$_1 ( )\ _1$
$_{2\ 2}$
1
1
1
$r\ v\ v$
r
$r\ v\ v$

$$\begin{bmatrix} \\ \\ \end{bmatrix}\begin{bmatrix} \\ \\ \end{bmatrix}\begin{bmatrix} \\ \\ \end{bmatrix}$$
$$\begin{pmatrix} \\ \\ \end{pmatrix}\begin{pmatrix} \\ \\ \end{pmatrix} = -\begin{pmatrix} \\ \\ \end{pmatrix}$$

$()_{1\,2\,1}\ v + rv = 1 - r\ v$

$()_{1\,2\,2}\ rv + v = 1 - r\ v$

El desarrollo de cualquiera de estas dos ecuaciones conduce al siguiente resultado:

$()_{1\,1\,2}\ v - 1 - r\ v = -rv$

$v_1 - v_1 + rv_1 = -rv_2$

$rv_1 = -rv_2$

$v_1 = -v_2$

$()_2 \therefore v' = v\ 1\ -1$

En resumen, cuando r > 0, el primer vector propio es $()_1\ v' = v\ 1\ 1$, y el segundo vector
propio es $()_2\ v' = v\ 1\ -1$.
Por un razonamiento analogo, cuando r < 0, el primer vector propio es
$()_1\ v' = v\ 1\ -1$,
y el segundo vector propio es $()_2\ v' = v\ 1\ 1$.
Luego, las componentes principales de este sistema se expresan asi:

1 2
1 2
, si 0
, si 0 1
vX vX r
Y
vX vX r
⟩+>
=∫
⌠−<
[4.5]
1 2
1 2
, si 0
, si 0 2
vX vX r
Y
vX vX r
⟩−>
=∫
⌠+<
La distancia de un punto a la primera componente principal es el valor absoluto de su
coordenada sobre la segunda componente principal. Para un punto que represente la
k-esima categoria de X_1 y la *k'*-esima categoria de X_2, tal distancia se obtiene a partir de
la expresion [4.5], asi:
{ } 1 2 ,
1 2 ,
1 2 ,
, si 0
(, , CP1)
, si 0
k k
k k
k k
v x x r
d x x
v x x r
⟩|−>
=∫
+<|⌠
Con base en las expresiones [4.3] y [4.5], pueden obtenerse las cuantificaciones de la

k-esima categoria de X_1 y la k'-esima categoria de X_2. Tales cuantificaciones son 2_{vX1k} y $)2_{vX2k'}$, respectivamente. La diferencia entre tales cuantificaciones esta dada por la siguiente expresion:

1 2 '
1 2 '
1 2 '
2 , si 0
(,)
2 , si 0
k k
k k
k k
v x x r
d qx qx
v x x r
⟩| ->
=∫
+< | ∫

Asi **queda demostrado que la diferencia en la cuantificación de las categorías de dos variables es proporcional a la distancia a la primera componente principal del punto que representa la combinación de tales categorías**. En este caso, la diferencia de las cuantificaciones es el doble de la distancia a la primera componente principal.

Este factor de proporcionalidad, desde luego, no es absoluto y podra cambiar acorde con la normalizacion elegida para las cuantificaciones.

En resumen, **las categorías que se asocian con mayor frecuencia configuran puntos de alta ponderación que tienden a estar cercanos a la primera componente principal y, por consiguiente, sus correspondientes cuantificaciones serán similares**.

El hecho de que la cuantificacion mas similar en este ejemplo sea la exhibida por las categorias t:180-184 y p:101-110 (3,08 y 3,10, respectivamente), las cuales no se presentaron simultaneamente en ninguna de las 50 observaciones (frecuencia cero)

parece contradictorio con lo afirmado anteriormente. Esta situacion se explica, sin embargo, por la restriccion propia de los escalamientos lineales. Si bien es cierto que los puntos de mayor ponderacion son los que tienden a quedar mas cercanos a la primera componente principal, en una tecnica lineal como el ACP no es posible evitar que algunos puntos sin "meritos" suficientes tambien queden cercanos a esta. Para corregir esta situacion seria necesario asignar las cuantificaciones con base en una tecnica no lineal que impusiera menos restricciones.

Este ejemplo ilustra como **la cuantificación natural de un par de variables numéricas, para cuya obtención solo es adecuado utilizar transformaciones lineales, también verifica la condición establecida en la definición 4.1**; es decir, asigna valores mas similares a las categorias o niveles de las variables que coincidan con mayor frecuencia.

Finalmente, para cerrar este apartado, antes de pasar a la formulacion matematica de la cuantificacion optima, supongase que se tiene un sistema multivariante nominal y que se desea cuantificar cada una de sus categorias. La solucion clasica viene dada mediante la cuantificacion de estas a partir de las coordenadas de un Analisis de Correspondencias (AC), en caso de tener solo dos variables, o de un Analisis de Correspondencias Multiples (ACM) o un Analisis de Correspondencias Conjunto (*Joint Correspondence Analysis,* GREENACRE, 2007) [17], en caso de tener mas de dos variables.

[17] El Analisis de Correspondencias Conjunto (ACCj) consiste en un AC de una tabla Burt, en el que se ignoran las submatrices que conforman el bloque diagonal, teniendo en cuenta unicamente las submatrices por fuera de la diagonal. El ACCj hace una mejor generalizacion del AC que el ACM, en terminos de inercias totales, coincidiendo exactamente con el AC cuando se tienen dos variables.

Si bien es cierto que en estas tecnicas no es adecuado interpretar directamente la distancia entre los puntos categoria de diferentes variables, si es posible establecer una interpretacion conjunta de los mismos con relacion a los ejes principales de la representacion (GREENACRE, 1994). En el caso de una solucion unidimensional, que seria la de nuestro interes para la obtencion de cuantificaciones simples, todas las categorias estarian sobre una linea, siendo posible interpretar directamente las distancias entre las categorias de diferentes variables, tal y como lo ilustra GREENACRE (2007).

En general, las categorias asociadas con mayor frecuencia en la tabla de contingencia, apareceran mas cercanas sobre el primer eje factorial.

Asi, pues, **la cuantificación clásica para variables nominales, generada a partir de las coordenadas de un AC, ACM o ACCj, asigna valores más similares a las categorías de las diferentes variables que se asocian con mayor frecuencia**, satisfaciendo asi la condicion especificada en la definicion 4.1.

Habiendo ilustrado la logica subyacente en la definicion 4.1 y su satisfaccion en diferentes escenarios, presentamos seguidamente su formulacion matematica, con base en la cual desarrollaremos posteriormente una solucion al problema planteado en el Capitulo 3.

4.4 DEFINICIÓN ESPECÍFICA DE CUANTIFICCIÓN ÓPTIMA

Usando el paralelismo existente entre los conceptos de similaridad y distancia (CUADRAS, 1991), podemos reformular la definicion de cuantificacion optima de manera mas especifica, asi:

Definición 4.2 (Cuantificación Óptima, específica): La cuantificacion optima de un sistema multivariante es la que minimiza la suma sobre todas

las observaciones de las distancias cuadraticas entre las categorias
cuantificadas de las diferentes variables.

Para el caso general de m variables, esta condicion se expresa asi:

$$\phi(X_1, X_2, \ldots, X_m) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{j'=1}^{m} (q_{ij} - q_{ij'})^2 \quad [4.6]$$

Donde:

$\phi(X_1, X_2, \ldots, X_m)$: Cuantificacion optima de las variables $X_1, X_2, \ldots, X_m$.

n: Numero de observaciones (filas de la matriz).

m: Numero de variables.

q_{ij}: Cuantificacion de la categoria de la j-esima variable correspondiente a la i-esima observacion.

$q_{ij'}$: Cuantificacion de la categoria de la j'-esima variable correspondiente a la i-esima observacion.

Los dos sumatorios mas internos en la expresion [4.6] evaluan para una observacion particular (la i-esima) todas las distancias cuadraticas entre sus categorias cuantificadas, imponiendo la restriccion necesaria ($j' > j$) para que cada distancia sea considerada una

sola vez. El mismo proceso se realiza para cada una de las n observaciones (sumatorio mas externo).

Con el fin de evitar soluciones triviales (que todas las cuantificaciones se concentren en un punto), es necesario establecer alguna **normalización**. En este trabajo normalizaremos el vector de cuantificaciones de cada una de las variables, de manera que tenga media cero y norma n, siendo n el numero de observaciones. Esta

normalizacion hace que las cuantificaciones obtenidas para los sistemas multivariantes
menos complejos, esto es, aquellos en los que todas las variables se escalan
numericamente, coincidan con las obtenidas mediante las tecnicas del sistema Gifi.

4.5 PLANTEAMIENTO ESPECÍFICO DE LA SOLUCIÓN

Es posible utilizar el criterio que acaba de presentarse (§ 4.4) para escoger las
cuantificaciones simples optimas de un sistema multivariante, a partir de un conjunto de
posibles cuantificaciones, basadas en diferentes dimensionalidades p, obtenidas
mediante PRINCALS, OVERALS o cualquiera de sus casos particulares. Bastara con
calcular para cada solucion la suma sobre todas las observaciones de las distancias al
cuadrado entre las correspondientes categorias cuantificadas, y elegir aquella solucion
para la cual dicha cifra sea minima. Con las definiciones 4.1 y 4.2 como referente, esta
sera la mejor solucion del conjunto, por cuanto sera la que mejor satisfaga que las
cuantificaciones de las categorias de las diferentes variables que se asocian con mayor
frecuencia esten mas cercanas entre si.

Para una observacion especifica, sea q_j el valor de la categoria cuantificada
correspondiente a la j-esima variable. Las interdistancias cuadraticas entre las categorias
cuantificadas de las diferentes variables esta dada por:

$$IDC(q_1, \ldots, q_m) = \sum_{j=1}^{m} \sum_{j'=1}^{m} (q_j - q_{j'})^2$$

( )2
,
1 , 1
1
2
m m
j j
j j
q q
= =
= ΣΣ −
(2 2)
, ,
1 , 1
1
2
2
m m
j j j j
j j
q q q q
= =
= ΣΣ − +
2 2
, ,
1 , 1 1 , 1 1 , 1
1
2
2
m m m m m m
j j j j
j j j j j j
q q q q
= = = = = =
⎡ ⎤
= ⎢ − + ⎥
⎣ ⎦
ΣΣ ΣΣ ΣΣ
2 2
, ,
1 1 , 1 , 1
1
2
2
m m m m
j j j j
j j j j
m q q q m q
= = = =
⎡ ⎤

= | − + |
□ ⌋
Σ Σ Σ Σ
Puesto que 2 2
’ ’
1 ’ 1 1 ’ 1
y
m m m m
j j j j
j j j j
q q q q
====
Σ =Σ Σ =Σ , podemos escribir esta expresion como:
2
2
1 1
1
2 2
2
m m
j j
j j
IDC m q q
==
⌊⌊⌋⌉
= | − ⌈ | |
|□⌊ ⌋ ⌊⌋
Σ Σ

82
2
2
1 1
m m
j j
j j
m q q
==
⌊⌋
= − ⌈ |
⌊ ⌋
Σ Σ
Y dado que la ‘varianza’, S_2, de la observacion en cuestion es:
( )
2
2
2 1 1
1
m m
j j
j j

$m\ q\ q$
S
$m\ m$
==
⎣ ⎤
− ⎡ ⎥
= ⎨ ⎬
−
$\Sigma\ \Sigma$
Podemos expresar la suma de interdistancias cuadraticas como $m(m-1)S_2$, lo cual facilita
los calculos numericos.
Luego, la suma de las interdistancias cuadraticas de todas las observaciones es
2
1
(1)
n
i
i
$m\ m\ S$
=
$-\Sigma$, donde 2
$_i S$ es la varianza de las diferentes categorias cuantificadas que conforman la i-esima observacion. Esta ‘varianza’ no debe confundirse con la varianza
de cada una de las variables, la cual, en virtud de la normalizacion impuesta, es igual a
n, el numero de observaciones.
Para fines de la comparacion que nos atane, podemos remplazar la constante $m(m-1)$ por
cualquier otra constante, pues no existe ningun interes particular en la magnitud de la
suma de las interdistancias cuadraticas; basta con saber cual es la solucion para la cual
dicha suma es minima. Con el fin de mantener los valores numericos bajos, el criterio
usado en este trabajo es el indicado mediante la expresion [4.7].
2
1
1
n
i
i
Interdistancias S

n

$= \sum$ [4.7]

Donde:

Interdistancias: Criterio escalado de suma de interdistancias cuadraticas.

n: Numero de observaciones (filas de la matriz).

$_i S^2$: 'Varianza' de la i-esima observacion.

Por simplicidad, en adelante nos referiremos a este criterio escalado de suma de interdistancias cuadráticas como *Interdistancias*.

No esta de mas insistir en el hecho de que la 'varianza' que aparece en la expresion [4.7] no es la varianza de las variables, sino la varianza de las observaciones.

La expresion [4.7] recoge la esencia de las definiciones 4.1 y 4.2 y es, por tanto, la que se implementa en el algoritmo computacional usado para resolver el problema planteado en el Capitulo 3, asi como en los algoritmos que sustentan la propuesta generalizada de cuantificacion optima que se desarrolla en el Capitulo 5.

4.6 CONSIDERACIONES SOBRE LA PROPUESTA

La presente propuesta permite escoger el conjunto optimo de cuantificaciones simples para un sistema multivariante a partir de una serie de conjuntos de cuantificaciones, basados todos ellos en diferentes reducciones de la dimensionalidad del sistema original. Para tal efecto, se pasa por la obtencion de la dimensionalidad p de la solucion que da lugar a las cuantificaciones optimas. Debe observarse, sin embargo, que la obtencion de la dimensionalidad p es solo un paso intermedio para la obtencion de las cuantificaciones optimas y no el objetivo de la propuesta. **Mediante esta propuesta no**

se pretende brindar una guía para la elección del número de ejes que deben retenerse en una técnica de reducción de dimensionalidad. Eventualmente, podria
usarse para tal efecto, pero seria necesario realizar una valoracion frente a otras
opciones.
Esta propuesta constituye un criterio de índole general para elegir la mejor cuantificación simple a partir de un conjunto de posibles cuantificaciones; el
aspecto de la dimensionalidad para la generacion del conjunto inicial de cuantificaciones es circunstancial. Este criterio podria usarse para elegir la mejor
cuantificacion simple de una serie de cuantificaciones generadas por cualquier medio,
sin importar si el proceso de generacion de las cuantificaciones involucra diferentes
dimensionalidades o no.
Es importante resaltar el hecho de que la solucion propuesta, al estar basada en las
definiciones 4.1 y 4.2 solo toma en consideracion las relaciones existentes entre los
datos, sin basarse en ningun modelo. Tal y como indica NISHISATO (2007), el uso de
un modelo puede ser bueno para el proposito de capturar un tipo especial de
informacion de los datos, pero en ausencia de un conocimiento total sobre los datos, un
modelo puede hacer que una gran cantidad de informacion contenida en los datos se
quede sin analizar. Asimismo, este procedimiento esta acorde con las ideas de Jean-Paul
Benzecri (citado por BLASIUS y GREENACRE, 2006), quien opinaba que los datos
son los ‘reyes’; no el modelo que uno pudiera querer proponer para estos.

4.7 ALGORITMO PARA LA ELECCIÓN DE LA CUANTIFICACIÓN ÓPTIMA

Para aplicar nuestra propuesta, hemos desarrollado la rutina **INTERDIS** en MATLAB, la

cual, partiendo de las cuantificaciones de diferentes soluciones OVERALS (o
cualquiera de sus casos particulares), calcula las *Interdistancias* y escoge aquella
solucion para la cual estas sean minimas.
Los pasos de la rutina son:
1. Captacion de informacion general.
2. Lectura de informacion especifica proveniente del SPSS.
3. Calculo de las *Interdistancias* para cada una de las soluciones.
4. Presentacion de resultados.
A continuacion se presentan los detalles basicos de cada uno de los pasos, omitiendo
aquellos que solo tengan interes desde el punto de vista computacional.
Por generalidad, nos referiremos al caso en que se tengan varias soluciones OVERALS.
Desde luego, la rutina tambien es aplicable si se tienen varias soluciones de cualquiera
de sus casos particulares (cf. § 2.7).
Para ejecutar la aplicacion basta con copiar todos sus componentes en la ruta de trabajo
de MATLAB (por defecto C:¥Archivos de programa¥Matlab¥work) y digitar "Interdis" en
la ventana de comandos de MATLAB.

4.7.1 CAPTACIÓN DE INFORMACIÓN GENERAL.

Al iniciar el programa aparece una interfaz grafica como la que se muestra en la Figura
4.5. Esta contiene tres casillas en las que el usuario debera suministrar la informacion
general que se detalla a continuacion:

1) Identificador de la base de datos.
2) Numero de soluciones que se van a comparar.
3) Identificador de valores perdidos.

Figura 4.5. Interfaz grafica de la rutina **INTERDIS**.

El identificador de la base de datos, al cual nos referiremos en adelante como '**Id**',
puede ser cualquier cadena alfanumerica (numeros, letras o combinacion de ambas).
Este identificador constituye la raiz de los nombres de los archivos de entrada y del

archivo de salida, tal y como se detalla en § 4.7.2 y § 4.7.4.

El numero de soluciones que se especifique en la segunda casilla debe coincidir con la dimensionalidad p de la maxima solucion. Asi, por ejemplo, si se indica que el numero de soluciones es 4, el sistema buscara las soluciones para p=1, p=2, p=3 y p=4. Si el usuario no modifica esta casilla, el sistema buscara por defecto 9 soluciones.

Si la base de datos tiene informacion faltante, es necesario utilizar un codigo numerico para identificar las celdas correspondientes. Por defecto, el sistema busca los valores iguales a -99 para marcarlos como datos perdidos. El usuario puede cambiar este valor si lo desea. En caso de que no haya datos perdidos, esta casilla no requiere ninguna modificacion.

4.7.2 LECTURA DE INFORMACIÓN ESPECÍFICA.

Previa ejecucion del programa, para comparar p soluciones, es necesario disponer p+2 archivos de entrada en formato ASCII (extension **txt**) en la ruta de trabajo de MATLAB.

Los archivos requeridos son:

1) Matriz de datos originales (1 archivo).
2) Niveles de cada variable (1 archivo)
3) Informacion para las cuantificaciones (1 archivo para cada solucion: p archivos)

La matriz de datos originales esta conformada por la parte numerica de la base de datos utilizada para la obtencion de las cuantificaciones en SPSS (no contiene titulos para las variables). En caso de que haya valores perdidos, estos deberan senalarse con -99 o con cualquier otro codigo numerico que el usuario senale en la tercera casilla de la interfaz grafica de captacion de informacion general.

La matriz de datos originales debe tener el nombre: **X+Id.txt**, donde '**Id**' es el

identificador de la base de datos suministrado por el usuario (cf. § 4.7.1). Si, por ejemplo, el identificador de la base de datos es '**luz**', la correspondiente matriz de datos originales debe estar disponible en la ruta de trabajo de MATLAB, con el nombre **Xluz.txt**.

El archivo con los niveles de cada variable debe tener el nombre generico **niv+Id.txt**, donde '**Id**' es el identificador de la base de datos suministrado por el usuario (cf. § 4.7.1). Si, por ejemplo, el identificador de la base de datos es '**luz**', la informacion sobre los niveles de las variables debera estar disponible en la ruta de trabajo de MATLAB, con el nombre **nivluz.txt**. La informacion sobre los niveles de cada variable forma parte de las salidas del SPSS, cuya obtencion y manejo se describen al final de este apartado.

Por cada solucion que se desee evaluar se requerira un archivo con informacion para el calculo de las cuantificaciones. Cada uno de estos archivos tendra el nombre generico: **C+Id+p+#.txt**, donde '**C**' es la letra que siempre dara comienzo al nombre de estos archivos; '**Id**' es el identificador de la base de datos (cf. § 4.7.1); '**p**' es un caracter que siempre ira en esa posicion, indicando la dimensionalidad de esa solucion; '**#**' es el numero correspondiente a la dimensionalidad p. Asi, si se quieren comparar las soluciones basadas en las dimensionalidades p=1, p=2 y p=3 para la base de datos con identificador '**luz**', en la ruta de trabajo de MATLAB deberan estar disponibles los archivos **Cluzp1.txt**, **Cluzp2.txt** y **Cluzp3.txt**.

Puesto que al usar el procedimiento OVERALS en SPSS, no es posible obtener las

cuantificaciones simples mas que por pantalla, para evitar su trascripcion manual, estas se calculan a partir de las coordenadas simples y las ponderaciones (*weights*), las cuales si pueden generarse como parte de los resultados del OVERALS. Para tal efecto, se utiliza la siguiente expresion.

$$q_j = (Y_j' w_j) / (w_j' w_j) \quad [4.8]$$

Donde,

q_j: Vector de cuantificaciones simples para la j-esima variable.
Y_j: Matriz de coordenadas simples para la j-esima variable.
w_j: Vector de ponderaciones (*weights*) para la j-esima variable.

Para que el SPSS genere un archivo de resultados con los niveles de cada variable y con la informacion necesaria para calcular las cuantificaciones mediante la expresion [4.8], debe incluirse la instruccion *MATRIX=OUT(nombrearchivo.sps)* en la correspondiente sintaxis del OVERALS. El archivo que se genera tiene un formato propio del SPSS (extension **sps**) y queda ubicado en la ruta de trabajo del SPSS.

Tras abrir el archivo de resultados en SPSS, es posible realizar todo el procedimiento de ordenamiento y copiado desde la misma aplicacion, o bien puede guardarse una copia con formato de Excel para realizar el ordenamiento y copiado en esta otra aplicacion.

En cualquier caso, debe tenerse en cuenta que el separador decimal de la informacion que se copie a los archivos ASCII que seran leidos luego por MATLAB es el punto, sin importar cual sea la configuracion regional del ordenador. Antes de copiar la

informacion sobre los niveles de las variables y los datos para el calculo de las
cuantificaciones, es necesario ordenar el archivo con base en la primera columna
(*ROWTYPE_*).

Obtención de la información para el archivo de niveles de las variables (niv+Id.txt)

Tras ordenar por tipo de fila (*ROWTYPE_*), se copia uno de los vectores de niveles (este
vector aparece en 4 ocasiones: al frente de CENTRO_, al frente de MCOOR_, al frente
ε de PCENTRO_ y al frente de SCOOR_) que aparece en la segunda columna (*LEVEL_*). Este vector debe tener

$$\sum_{j=1}^{m} niv_j$$

filas, siendo niv_j el numero de niveles de la $j$esima
variable y m el numero de variables. Esta informacion se pega en un archivo
ASCII que se guarda en la ruta de trabajo de MATLAB con el nombre generico **niv+Id.txt**.

Obtención de la información para el archivo de Cuantificaciones (C+Id+p+#.txt)

Tras ordenar por tipo de fila (*ROWTYPE_*), se copia toda la informacion correspondiente a las dimensiones de *SCOOR_* y *WEIGHT_*, tal y como quedan tras el
ordenamiento (se copia solo la informacion numerica correspondiente a cada una de las

p dimensiones). La informacion copiada debera tener

$$\sum_{j=1}^{m} niv_j$$

+ m filas y p columnas,
siendo niv_j el numero de niveles de la j-esima variable; m, el numero de variables, y p,

la dimensionalidad de la correspondiente solucion. Esta informacion se pega en un
archivo ASCII que se guarda con el nombre generico **C+Id+p+#.txt**.

4.7.3 CÁLCULO DE LAS INTERDISTANCIAS.

Una vez calculadas las cuantificaciones simples para cada variable con base en la
expresion [4.8], se construye para cada una de las *p* soluciones una matriz *Q* en la que
cada categoria de cada variable es remplazada por su correspondiente cuantificacion.
Seguidamente, se calculan para cada una de las *p* soluciones las *Interdistancias*, con
base en la expresion [4.7], y se determina cual es la solucion cuyas *Interdistancias* son
minimas.

4.7.4 PRESENTACIÓN DE RESULTADOS.

Finalmente, tras comparar los diferentes conjuntos de cuantificaciones, la aplicacion
INTERDIS, presenta un aviso por pantalla, indicando con cual de las soluciones
comparadas se obtienen las *Interdistancias* minimas.
Asimismo, se construye un grafico con las dimensionalidades de las diferentes
soluciones comparadas en la abcisa, y las *Interdistancias* en la ordenada.
De igual forma, genera un archivo en Excel, llamado **Q+Id.xls**, donde '**Id**' es el
identificador de la base de datos (cf. § 4.7.1), en el que se escribe la matriz de
cuantificaciones correspondiente a la solucion con Interdistancias minimas. En el caso
de la hipotetica base de datos '**luz**', este archivo tendra el nombre **Qluz.xls**. En caso de
que la base de datos original tenga datos perdidos (marcados con -99 o con cualquier
otro codigo numerico), estos apareceran como celdas vacias en la matriz de
cuantificaciones. Las cuantificaciones contenidas en dicho archivo pueden utilizarse
como entrada de cualquier tecnica lineal.

4.8 ILUSTRACIÓN DE LA PROPUESTA EN UNA APLICACIÓN PRÁCTICA

A continuacion se muestran los resultados de aplicar la rutina **INTERDIS** a una base de datos real conformada por 48 variables y 5.443 pacientes con cardiopatia isquemica, sobre la cual se realizo un analisis OVERALS. Los detalles sobre la base de datos y los resultados OVERALS aparecen en CORREA (2006). No profundizaremos en el analisis ni de la base de datos ni de los resultados, pues el unico objetivo en esta parte es ilustrar el funcionamiento de la rutina **INTERDIS**. En el Capitulo 6 se evaluara con detalle una aplicacion practica con una base de datos mayor, usando la generalizacion de esta propuesta que se desarrolla en el Capitulo 5.

Se obtuvieron 12 soluciones OVERALS (p=1 hasta p=12), cuyos resultados se dispusieron en 14 archivos, con base en las especificaciones dadas en § 4.7.2. El resultado grafico de la rutina **INTERDIS** se muestra en la Figura 4.6.

Figura 4.6. *Interdistancias* para soluciones OVERALS desde p=1 hasta p=12.

En este caso, la solucion OVERALS con p=6 es la que minimiza las *Interdistancias*, garantizando que las categorias de las variables mas frecuentemente asociadas sean las que reciban cuantificaciones mas similares. En consecuencia, esta seria la solucion mas recomendable si el objetivo del investigador fuera la generacion de cuantificaciones simples, para la subsiguiente aplicacion de tecnicas que exigieran datos numericos.

Vale la pena insistir en el ultimo aspecto anotado. Siempre que se utilicen calificativos como “optimo”, “mejor” o “mas recomendable” debe tenerse en mente el alcance de los mismos. En este capitulo se ha desarrollado una propuesta que garantiza la eleccion del conjunto de cuantificaciones en el que se den las mayores similitudes entre las

cuantificaciones de las categorias que se asocian con mayor frecuencia. El hecho de que
los conjuntos de cuantificaciones comparados sean el resultado de diferentes procesos
de reduccion de la dimensionalidad es circunstancial. El resultado obtenido no debe
interpretarse como una prueba en favor de la eleccion de la dimensionalidad en cuestion
para fines de representacion. El conjunto de cuantificaciones elegido puede utilizarse
como informacion de entrada de cualquier tecnica lineal. En particular, si se trata de una
tecnica de reduccion de dimensionalidad se tendran en cuenta otras consideraciones
para elegir el numero de ejes retenidos o la dimensionalidad de la representacion, la cual
no tiene que coincidir con la dimensionalidad de la solucion OVERALS que dio lugar a
las cuantificaciones.
En el siguiente capitulo se elabora una propuesta de cuantificacion basada en la
generalizacion de los criterios desarrollados en el presente capitulo.

CAPÍTULO 5
Propuesta Generalizada De Cuantificación Óptima: CUANTIFICA

5.1 INTRODUCCIÓN

En el capitulo anterior se establecio un criterio de cuantificacion optima que permite
elegir el mejor conjunto de cuantificaciones a partir de varias opciones generadas con
OVERALS o cualquiera de sus casos particulares. **En este capítulo**, presentamos una
propuesta generalizada basada en el mismo criterio, con base en la cual es posible

obtener directamente las cuantificaciones optimas sin que medie ningun proceso previo
de cuantificacion.
Inicialmente se establece un marco de referencia para el proceso de cuantificacion
optima, en el cual se incluyen aspectos tales como la normalizacion, el nivel de
escalamiento y los datos faltantes. Como novedad de este trabajo, incluimos los niveles
de escalamiento Numerico Flotante y Ordinal Flotante, los cuales no forman parte de
los niveles de escalamiento tradicionales.
Seguidamente se presenta la rutina **CUANTIFICA**, la cual esta conformada por 16
subrutinas o funciones que interactuan para leer una base de datos multivariante
convenientemente discretizada y asignar un esquema optimo de cuantificaciones
simples, acorde con las definiciones 4.1 y 4.2.
Se detallan las diferentes operaciones realizadas por la rutina **CUANTIFICA**, las cuales
comprenden 7 pasos principales, a saber:
1) Captacion de la informacion suministrada por el usuario.
2) Lectura y adecuacion de la base de datos.
3) Obtencion de una matriz inicial de cuantificaciones.
4) Inicializacion de los diferentes tipos de variables.
5) Calculo de las *Interdistancias* iniciales.
6) Ciclo de minimizacion de las *Interdistancias*, acorde con el nivel de escalamiento de
cada variable.
7) Generacion y presentacion de resultados.

Cerramos este capitulo comparando **CUANTIFICA** con los tres principales modulos de
escalamiento optimo del sistema Gifi de analisis multivariante no lineal, i. e.,
PRINCALS, HOMALS y OVERALS, cuando son usados como primer paso para la generacion de cuantificaciones.

5.2 MARCO DE REFERENCIA DE LA CUANTIFICACIÓN ÓPTIMA

En el capitulo anterior se ofrece una solucion especifica al problema planteado en el
Capitulo 3, esto es, obtener un conjunto de cuantificaciones simples optimas para un
sistema multivariante. Para tal efecto, se parte de la siguiente definicion: ***"La cuantificación óptima de un sistema multivariante es la que asigna valores más similares a las categorías de las distintas variables que se asocian con mayor frecuencia"***. Posteriormente, esta definicion se precisa asi: ***"La cuantificación óptima de un sistema multivariante es la que minimiza la suma sobre todas las observaciones de las distancias cuadráticas entre las categorías cuantificadas de las diferentes variables"***.

Luego, para elegir el mejor conjunto de cuantificaciones a partir de varios posibles
conjuntos, basta con tomar el que satisfaga el criterio anterior, lo cual puede evaluarse
usando cualquier transformacion lineal de la suma de distancias cuadraticas. En este
estudio se ha optado por la expresion [4.7], que no es mas que un criterio escalado de la
suma sobre todas las observaciones de la suma de interdistancias cuadraticas entre las
categorias cuantificadas. Por simplicidad, a tal criterio lo hemos denominado
Interdistancias.

Notese que la solucion presentada en el Capitulo 4, tiene dos limitaciones. Por una
parte, exige utilizar el OVERALS o cualquier otra tecnica de cuantificacion para
generar los conjuntos de cuantificaciones de partida. Por otra parte, si bien permite
obtener el mejor conjunto de cuantificaciones simples a partir de los conjuntos
disponibles, no garantiza que dicha solucion sea un optimo absoluto en cuanto a la

satisfaccion del criterio propuesto, y bien podrian obtenerse otros conjuntos con
Interdistancias menores.
Siendo consecuentes con la propuesta desarrollada en el Capitulo 4, resulta logico
buscar un esquema optimo absoluto de cuantificaciones simples, esto es, que satisfaga
la definicion 4.2. Para tal efecto, podria tomarse como punto de partida la mejor
solucion de un conjunto de posibles cuantificaciones, obtenida mediante el
procedimiento detallado en el Capitulo 4, o podria partirse directamente de la base de

datos original. Este ultimo enfoque ataca las dos limitaciones anotadas anteriormente y
es, por tanto, el que hemos seguido en este trabajo.
Antes de desarrollar la propuesta, consideraremos algunos aspectos involucrados en el
proceso de cuantificacion:
Normalización. Para evitar una solucion trivial, consistente en que todas las
cuantificaciones se colapsen en un punto (lo cual, obviamente, minimizaria las
Interdistancias), se requiere aplicar algun proceso de normalizacion. En este trabajo,
hemos elegido una normalizacion tal que cada una de las variables cuantificadas este
centrada en cero y tenga norma n , siendo n el numero de observaciones.
Nivel de Escalamiento. Para cada variable podra definirse cualquiera de los cinco
siguientes niveles de escalamiento: Numerico, Ordinal, Nominal, Numerico Flotante u
Ordinal Flotante.
 Numérico: Se aplican transformaciones lineales sobre los valores iniciales de
las categorias, de manera que las distancias entre las categorias transformadas
sean proporcionales a las distancias entre las categorias de partida. El grafico de

cuantificaciones de una variable escalada a nivel Numerico es una linea recta.

- **Ordinal**: Se aplican transformaciones monotonas, de manera que las categorias transformadas conserven el orden de las categorias iniciales. Luego, el correspondiente grafico de cuantificaciones sera no descendente.
- **Nominal**: Se aplican transformaciones isomorficas, tales que cada una de las categorias reciba una cuantificacion posiblemente diferente, sin mas restricciones que las impuestas por la normalizacion. Los graficos de cuantificaciones no siguen ningun patron especifico.
- **Numérico Flotante**: Todas las categorias, excepto una que se deja libre, se escalan a nivel Numerico.

- **Ordinal Flotante**: Todas las categorias, excepto una que se deja libre, se escalan a nivel Ordinal.

En adicion a la propuesta misma de cuantificacion optima basada en las definiciones 4.1 y 4.2 e implementada a traves de la rutina **CUANTIFICA**, otro de los aspectos novedosos de este trabajo consiste en la definicion e implementacion de dos niveles de escalamiento adicionales a los tradicionalmente usados: Numerico Flotante y Ordinal Flotante.

Cuando alguna variable posea una categoria que no siga el patron de las demas, tal como la que se presenta a menudo bajo la etiqueta "*no sabe/no responde*", pueden seguirse dos cursos de accion, dependiendo del tipo de estudio, de la homogeneidad de la categoria en cuestion y de los objetivos del investigador: manejar la informacion de dicha categoria como datos faltantes o declarar tal categoria como flotante.

Al declarar todas las observaciones de una categoria como datos faltantes, estas se

excluyen del proceso, lo que implica que no participan en la cuantificacion de las demas
categorias. Dicha categoria, desde luego, tampoco recibe ninguna cuantificacion.
La otra posibilidad consiste en asignarle una cuantificacion libre a la categoria en
cuestion, declarandola como flotante, en cuyo caso participa en la cuantificacion de las
categorias de las demas variables y recibe asimismo una cuantificacion, lo que permite
evaluar su relacion con las demas categorias.
En este ultimo caso, se utiliza uno de los dos tipos de escalamiento flotante (Numerico
Flotante u Ordinal Flotante), acorde con el nivel de escalamiento que se quiera asignar a
las demas categorias de la variable. El programa escala las primeras K-1 categorias de
la variable como Numericas u Ordinales, dependiendo del nivel de escalamiento
elegido, dejando libre (o flotante) la K-esima categoria.
Para las variables Nominales, no es necesario definir un tipo de escalamiento flotante
aunque existan categorias especiales del tipo "*no sabe/no responde*", dado que el
escalamiento Nominal asigna una cuantificacion libre a cada una de las categorias.

Datos Faltantes. En el presente contexto, siempre que se hable de datos faltantes,
valores perdidos u observaciones perdidas, se estara haciendo referencia a lecturas
especificas de una o mas variables que no existen en alguna fila especifica, es decir, a
celdas especificas de la matriz, sin que ello implique la perdida de toda la fila.
Bajo el enfoque teorico propuesto y su correspondiente implementacion algoritmica, la
perdida de datos en una fila, no impide que los datos restantes de la misma participen en

el proceso de minimizacion de las *Interdistancias*. **Para cualquier observación con lecturas en al menos dos variables será posible calcular la correspondiente componente de las *Interdistancias*, sin necesidad de utilizar ningún método de imputación**.

5.3 RUTINA CUANTIFICA

La rutina **CUANTIFICA** esta conformada por un conjunto de 16 subrutinas o funciones (**Inicio**, **Main**, **Base**, **Numini**, **Floini**, **Nomini**, **Nominit**, **Num**, **Ord**, **Nom**, **Numflo**, **Ordflo**, **Norma**, **Inter**, **Res** y **Gcuant**) que, en conjunto, leen una base de datos multivariante convenientemente discretizada y asignan un esquema optimo de cuantificaciones simples, acorde con las definiciones 4.1 y 4.2.

A partir de la informacion de entrada captada mediante una interfaz grafica de usuario que es controlada por la subrutina **Inicio**, la subrutina **Main** se encarga de llamar a las demas subrutinas y de ir conduciendo el flujo general del proceso. Cada una de las otras subrutinas realiza procesos especificos, tal y como se muestra en la Figura 5.1.

Figura 5.1. Conjunto de subrutinas que integran la aplicacion **CUANTIFICA**.

Base	Lee y adecua la base de datos
Norma	Normaliza cada una de las columnas de la matriz Q, de manera que queden centradas en cero y con norma igual a raiz de n
Numini	Genera un direccionamiento inicial optimo de las variables numericas y ordinales
Nomini	Genera una cuantificacion inicial de las variables flotantes, con base en las numericas y ordinales
Nominit	Generan cuantificaciones iniciales de las variables nominales
Floini	
Num	
Ord	
Nom	
Numflo	
Ordflo	
Inter	Calcula las *Interdistancias*, con base en la expresion [4.7]

A cada variable, acorde con su nivel de escalamiento, le asigna cuantificaciones que minimicen las *Interdistancias*
Res
Gcuant
Res: Genera cuantificaciones por pantalla y las guarda.
Gcuant: Construye un grafico de cuantificaciones para cada variable.
Main
Inicio

Aunque desde el punto de vista tecnico de la programacion, cada una de estas secciones es una funcion (tienen argumentos de entrada y de salida), nos referiremos a ellas como subrutinas. Reservaremos el termino rutina para referirnos al proceso general englobado por **CUANTIFICA.**

Los pasos generales de la rutina **CUANTIFICA** son:

1. Captacion de la informacion suministrada por el usuario.
2. Lectura y adecuacion de la base de datos.
3. Obtencion de una matriz inicial de cuantificaciones.
4. Inicializacion de los diferentes tipos de variables.
5. Calculo de las *Interdistancias* iniciales.
6. Ciclo de minimizacion de las *Interdistancias*, acorde con el nivel de escalamiento de cada variable.
7. Generacion y presentacion de resultados.

A continuacion se detallan los aspectos fundamentales del proceso, omitiendo aquellos que solo tengan interes desde el punto de vista computacional.

Los pasos 1 y 7, es decir, los que exponen lo relacionado con la entrada de la informacion y los resultados generados por la aplicacion, son los que mas conciernen al usuario final, a quien le bastaria con revisar estos dos apartados para realizar un proceso de cuantificacion optima. Las ayudas del programa estan conformadas por una version modificada de estos dos apartados. En los pasos 2 a 6 se encuentra la fundamentacion teorica del proceso, expuesta en forma lineal, acorde con el flujo algoritmico.

5.3.1 CAPTACIÓN DE LA INFORMACIÓN DE ENTRADA.

La interaccion entre el usuario y la aplicacion se realiza a traves de la interfaz grafica
que se muestra en la Figura 5.2. Esta interfaz es controlada por la subrutina **Inicio**.
Luego, para ejecutar la rutina **CUANTIFICA**, bastara con copiar todos los componentes de
la misma en la ruta de trabajo de MATLAB (por defecto C:¥Archivos de programa¥Matlab¥work), digitar "QInicio"[18] en la ventana de comandos de MATLAB,
pulsar el boton "*! Leer Archivo Excel !*" y abrir el archivo que contiene la informacion
de entrada.

Figura 5.2. Interfaz grafica de usuario de la aplicacion **CUANTIFICA**.

[18] Para facilitar la organizacion de los archivos, todos ellos se han nombrado anteponiendo la letra Q al
nombre de la subrutina, de modo que queden juntos al realizar un ordenamiento alfabetico.

La interfaz grafica de la aplicacion **CUANTIFICA** tiene cuatro componentes con los que el
usuario puede interactuar: Opciones de la Base de Datos, Lectura del Archivo,
Resultados y Ayuda.
En el panel ubicado en la parte izquierda de la ventana, aparecen las especificaciones
que debe satisfacer la base de datos. El usuario puede interactuar con algunos de los
componentes de dicho panel, tal y como detallamos a continuacion.

BASE DE DATOS

La base de datos debe estar organizada en un archivo de Excel con extension **xls**, el cual
ha de contener toda la informacion necesaria para realizar la cuantificacion optima. La
primera fila debe estar conformada por los nombres de las variables; la segunda por el
nivel de escalamiento; la tercera puede contener opcionalmente un vector de filtrado; en
las subsiguientes filas deben disponerse los datos que se usaran como base para la

cuantificacion. Si el archivo Excel tiene varias hojas, la base de datos debera ubicarse en
la primera de ellas, es decir, en la ubicada mas hacia la izquierda, sin importar los
nombres que las diferentes hojas puedan tener.

NOMBRES DE LAS VARIABLES

Deben estar localizados en la primera fila de la base de datos. Esta condicion no es
susceptible de modificacion por parte del usuario. Los nombres pueden contener
espacios y/o caracteres especiales. La unica restriccion consiste en que no esten
conformados unicamente por numeros. En ocasiones, el sistema tampoco es capaz de
leer los nombres cuando comienzan por un numero, generando un error. Por tanto, se
desaconseja dicha practica.

NIVEL DE ESCALAMIENTO

En la segunda fila del archivo debe especificarse el nivel de escalamiento de cada una
de las variables. Esta tambien es una condicion que no permite modificacion por parte
del usuario. Los niveles de escalamiento se eligen teniendo en cuenta las definiciones
presentadas en § 5.2. Para el efecto, se utiliza un numero entre 1 y 5, asi:

1. Numerico
2. Ordinal
3. Nominal
4. Numerico Flotante
5. Ordinal Flotante

FILTRADO DE VARIABLES

La tercera fila puede contener un vector de filtrado de variables, el cual permite utilizar
un subconjunto de las variables presentes en la base de datos, sin necesidad de realizar
copias o modificaciones de la misma. Si el usuario desactiva esta opcion que viene por

defecto, el sistema leera los datos a partir de la tercera fila y cuantificara todas las
variables contenidas en la base de datos.
En caso de que la tercera fila incluya un vector de filtrado, este debe contener unos o
ceros, asi:
1. SI se incluye la variable en el analisis.
0. NO se incluye la variable en el analisis
Notese que para realizar un analisis con todas las variables, aun cuando se haya incluido
un vector de filtrado, basta con usar *unos* (1) en todas las celdas de la tercera fila.
Cuando se incluye un vector de filtrado, los datos se dispondran a partir la cuarta fila.

DATOS (ETIQUETAS DE LAS CATEGORÍAS)

Todas las variables tienen que estar categorizadas o discretizadas, sin que exista
ninguna restriccion para la asignacion de etiquetas o identificadores a las categorias,
mas alla de que sean numeros enteros y de la conservacion de la ordinalidad o la
proporcionalidad de distancias para las categorias de las variables escaladas como
Ordinales o Numericas, respectivamente.
Estas etiquetas o identificadores de las categorias constituyen los datos de entrada del
proceso de cuantificacion optima, y deben estar ubicadas a partir de la cuarta o tercera
fila, dependiendo de si se ha incluido un vector de filtrado o no, respectivamente. La
aparicion de valores numericos no enteros en esta region de la base de datos sera
interpretada como un error involuntario del usuario, producido, quiza, por el formato de
visualizacion de Excel, por lo que la aplicacion procedera a su redondeo al entero mas
proximo.
Si se tiene, por ejemplo, una base de datos que, entre otras, incluya las variables *Sexo*,

Profesion, *Edad* e *Ingresos*, escaladas a niveles Nominal (3), Nominal (3), Numerico
(1) y Ordinal Flotante (5), respectivamente, y se desea excluir *Profesion* del proceso de
cuantificacion optima (filtro=0), la base de datos tendra el aspecto que se muestra en la
Figura 5.3.

Figura 5.3. Estructura en Excel del archivo de entrada para la rutina **CUANTIFICA**.

Nombres de las variables
Nivel de Escalamiento
Filtrado
Datos

Cuando se utiliza alguno de los dos niveles de escalamiento flotante, i. e., Numerico
Flotante u Ordinal Flotante, la categoria flotante debera identificarse con la etiqueta de
mayor valor de la correspondiente variable. Considerese, por ejemplo, la variable
Ingresos, escalada a nivel Ordinal Flotante, conformada por las siguientes categorias:

- No responde
- Menos de 600 euros
- Entre 600 y 1.500 euros
- Entre 1.500 y 3.000 euros
- Mas de 3.000 euros

Para su identificacion en la base de datos, se utilizarian respectivamente las etiquetas 5,
1, 2, 3 y 4. Alternativamente, podria utilizarse cualquiera de los siguientes conjuntos de
etiquetas, sin que ello afecte el resultado: 4, 0, 1, 2, 3 o 4000, 500, 600, 1500, 3000.
Notese que cualquiera de los conjuntos de etiquetas propuesto respeta la ordinalidad de
la parte no flotante (categorias 2 a 5) y asigna la etiqueta de mayor valor a la categoria
flotante (*No responde*).
Debe observarse, sin embargo, que los anteriores conjuntos de etiquetas propuestos solo

son equivalentes si se utiliza el nivel de escalamiento Ordinal Flotante. En caso de
utilizarse el nivel de escalamiento Numerico Flotante, solo los dos primeros conjuntos
de etiquetas propuestos serian equivalentes entre si, difiriendo del tercer conjunto y
generando, por tanto, diferentes resultados.
Aunque las listas de referencia (por ejemplo, los cuestionarios) usualmente se organizan
de manera que la categoria *No responde* sea la ultima de su grupo, en este caso la hemos
presentado intencionalmente encabezando el grupo de categorias, para enfatizar el
hecho de que, sin importar su posicion en la lista de referencia, la categoria *No responde*
siempre se codificara con la etiqueta de mayor valor de su grupo de categorias.

DATOS FALTANTES

Cuando falta la informacion de algunas celdas, basta con dejarlas vacias o llenarlas con
cualquier caracter o grupo de caracteres no numericos. Es posible, incluso, combinar
estas dos posibilidades: dejar celdas vacias y utilizar caracteres no numericos en otras,
los cuales no tienen que coincidir entre si. Todos los valores no numericos o ausentes
que se encuentren de la cuarta fila en adelante seran manejados por el programa como
datos perdidos.
En ningun caso debe utilizarse el cero (0) para senalar valores faltantes, pues este es un
valor numerico que no se interpretaria como marcador de valor perdido, sino como la
etiqueta de una categoria.
Aunque la importacion de la base de datos es un proceso que normalmente se realiza sin
problemas y de manera bastante rapida, en algunas ocasiones, la presencia de muchas

celdas no numericas (datos perdidos) puede afectar negativamente el desempeno de la
funcion de importacion. Para resolver esta situacion, en caso de que se presente, basta
con identificar todas las observaciones perdidas de la base de datos con el valor
numerico -99. Este es el unico valor numerico que siempre sera interpretado por el
programa como marcador de datos faltantes; cualquier otro valor numerico es manejado
como la etiqueta de una categoria.
En adicion al boton *! Leer Archivo de Excel !*, la interfaz grafica permite elegir los
resultados que se desean generar. Por defecto, todos aparecen marcados. En § 5.3.7 se
describe en que consiste cada uno de ellos.
El otro punto de interaccion del usuario con la aplicacion es a traves del boton *Ayuda*, el
cual le brinda acceso a una version modificada de lo descrito en este apartado (§ 5.3.1),
complementada con la descripcion de los resultados, la cual presentamos mas adelante
en esta memoria (§ 5.3.7).
La comprension de lo expuesto hasta aqui, en lo que tiene que ver con la rutina
CUANTIFICA, es suficiente para su utilizacion por parte del usuario final, quien
interactua con la misma unicamente a traves de la interfaz grafica. A continuacion

describimos el flujo general de la aplicacion, asi como los procesos especificos
realizados por cada una de las demas subrutinas que la conforman.

5.3.2 LECTURA Y ADECUACIÓN DE LA BASE DE DATOS.

Una vez el usuario abre la base de datos utilizando la interfaz grafica, la informacion es
transmitida a la subrutina **Main**, la cual orquesta en adelante el proceso de
cuantificacion. Inicialmente, la subrutina **Main**, llama a la subrutina **Base**, la cual se

encarga de verificar que todas las especificaciones de la base de datos sean correctas,
generando mensajes especificos de error y deteniendo el proceso, cuando las
especificaciones sean incorrectas, o realizando las adecuaciones del caso cuando ello
sea plausible.
Inicialmente, se verifica que se cuente con suficiente informacion para realizar el
analisis, esto es, que la base de datos tenga como minimo dos observaciones y dos
variables (aunque una cuantificacion para este caso extremo no resultaria demasiado
interesante). De no ser asi, el sistema genera el correspondiente mensaje de error y la
aplicacion se detiene.
A continuacion, se verifica que los niveles de escalamiento hayan sido definidos
adecuadamente, esto es, que en cada una de las celdas de la segunda fila de la base de
datos se hayan utilizado valores enteros entre 1 y 5. De no ser asi, se informa sobre
dicho error y se detiene la aplicacion.
Seguidamente se verifica lo relacionado con el vector de filtrado. En caso de que el
mismo no se haya definido adecuadamente, se genera un aviso de error y se detiene la
aplicacion. Asimismo, si el vector de filtrado se definio adecuadamente, el sistema
informa sobre su estado indicando cuantas variables han sido filtradas o si todas han
sido incluidas en el analisis.
Se verifica luego que todas las etiquetas de las categorias sean numeros enteros. De no
ser asi, se redondean al entero mas proximo.

Se realiza a continuacion la marcacion de valores perdidos, informando por pantalla
cuantos valores perdidos contiene la base de datos y lo que ello representa en terminos

porcentuales.
Puesto que ninguno de los pasos de la rutina **CUANTIFICA** esta basado en operaciones
que exijan matrices completas, no es necesario utilizar ninguna tecnica de imputacion
de datos. Esta caracteristica hace que el procedimiento sea transparente, por cuanto esta
basado unicamente en la informacion disponible, sin que se requiera forzar la
estimacion de categorias desconocidas mediante tecnicas mas o menos subjetivas y sin
que la ausencia de valores especificos sea un impedimento para la utilizacion de los
demas valores, usandose siempre el 100 % de la informacion susceptible de generar
Interdistancias, esto es, todas la filas con al menos dos observaciones.
Logicamente, las filas con menos de dos observaciones ni pueden usarse en el calculo
de las *Interdistancias* (cf. expresion [4.7]) ni contienen ninguna informacion
multivariante. Se procede, por tanto, a retirarlas de la base de datos.
Puesto que los proceso de filtrado y eliminacion de filas con menos de dos lecturas
reducen el tamano de la base de datos, la subrutina verifica nuevamente que la base de
datos final contenga la informacion minima para el analisis, esto es, dos filas y dos
columnas. De no ser asi, se informa que no se cuenta con informacion suficiente para
realizar la cuantificacion y la aplicacion se detiene.
Aun cuando la reduccion de la base de datos no sea tan drastica como para impedir su
uso, cualquier reduccion hace que la matriz de cuantificaciones que se genera como
parte de los resultados de la rutina (§ 5.3.7) no coincida celda a celda con la base de
datos original. Para evitar las posibles confusiones que esto puede ocasionar, en tales
casos, la rutina devuelve una copia de la base de datos reducida, es decir, la que

efectivamente se utiliza en el proceso de cuantificacion. Dicha base de datos reducida es
copiada en una hoja del archivo Excel que contiene la base de datos original, con el
nombre "***BD* red**", donde ***BD*** es el nombre de la hoja que contiene la base de datos
original. Asi, por ejemplo, si la base de datos original aparece en una hoja llamada **Luz**,
la base de datos reducida se copiara en una hoja llamada **Luz red**.

Finalmente, tras haber realizado todas las verificaciones y adecuaciones del caso, se
genera un aviso informativo de la finalizacion del proceso de lectura y adecuacion de la
base de datos, indicando el numero de filas y de columnas que seran utilizadas para la
cuantificacion.

5.3.3 OBTENCIÓN DE LA MATRIZ INICIAL DE CUANTIFICACIONES.

Una vez leida la base de datos y sus caracteristicas (niveles de escalamiento y
etiquetas), la subrutina **Base** devuelve el control a la subrutina **Main**, la cual se encarga
de llamar a la subrutina **Norma** para generar una matriz inicial de cuantificaciones, Q, a
partir de la normalizacion por columnas de la matriz de datos, X.

La normalizacion por columnas de una matriz es un proceso recurrente a lo largo de la
rutina **CUANTIFICA**. En todos los casos, se realiza con base en la subrutina **Norma**,
haciendo que cada columna quede con media cero y norma n , siendo n, el numero de
observaciones (numero de filas).

Segun se trate de una columna completa o de una columna con datos faltantes, el
algoritmo implementa la expresion [5.1] o la expresion [5.2], respectivamente.

Si se trata de una columna completa, es decir, sin datos perdidos:

$$q(\ ) = n(\ ),\ q(\ )$$

$q\ n$
$q\ q\ q\ q$
$-$
$=$
$--$
[5.1]
Donde:
q_n: Columna normalizada de la matriz Q.
q: Columna de la matriz Q antes de la normalizacion.
q : Media de la columna q.
n: Numero de filas de la matriz Q.

Si se trata de una columna con datos perdidos:
$( )$
$( )^2$
i
n
i
q
$q\ q$
$q\ n$
$q\ q$
$\in$
$-$
$=$
$\Sigma -$
$\mathbb{R}$
[5.2]
Donde:
q_n: Columna 'normalizada' de la matriz Q.
q: Columna de la matriz Q antes de la normalizacion.
q : Media de la columna q, omitiendo los valores perdidos.
q_i: i-esima observacion de la columna q, con i evaluado sobre el conjunto de
valores observados (valores no perdidos).
n: Numero de filas de la matriz Q.
El entrecomillado en 'normalizada' indica que, puesto que la columna en cuestion tiene
datos perdidos, no es posible obtener la norma mediante la expresion usual: $q = q\,'\,q$,
sino que se obtiene una norma generalizada mediante la expresion: 2
i
i

q

$q\ q$

$\in$

$= \sum$

$\mathbb{R}$

.

Notese que ambas expresiones coinciden cuando la columna esta completa.

5.3.4 SUBRUTINAS DE INICIALIZACIÓN.

Antes de aplicar el ciclo general de minimizacion de las *Interdistancias*, se llevan a cabo una serie de precuantificaciones por grupos de variables, acorde con sus niveles de escalamiento, con lo cual se obtiene una convergencia mas rapida hacia las *Interdistancias* minimas.

SUBRUTINA "Numini"

La cuantificacion optima de variables escaladas a nivel Numerico se obtiene, tal y como se ha indicado anteriormente (cf. § 2.5), mediante transformaciones lineales, lo cual, para el caso de variables centradas equivale a multiplicarlas por una constante.

Si se consideran aisladamente dos variables que se correlacionen positivamente, sus correspondientes cuantificaciones se obtienen utilizando constantes del mismo signo.

Por el contrario, si se consideran aisladamente dos variables correlacionadas negativamente, sus correspondientes cuantificaciones se obtienen utilizando constantes de diferente signo.

Este hecho se explica a la luz de la logica de las cuantificaciones optimas expuesta en el capitulo anterior, en particular, con base en la definicion 4.1. Observese, por ejemplo, que si un par de variables numericas se correlacionan negativamente las categorias bajas de la primera tiende a asociarse con las altas categorias de la segunda y viceversa.

Luego, las cuantificaciones de las categorias inferiores de la primera variable deberan estar cercanas a las cuantificaciones de las categorias superiores de la segunda variable y viceversa. Esto, desde luego, solo se logra utilizando constantes de signo contrario para las correspondientes transformaciones. Un razonamiento analogo se sigue para las variables correlacionadas positivamente.

Notese que en cualquiera de los dos casos anteriores, bien sea que las variables originales se correlacionen positiva o negativamente, las variables transformadas se correlacionaran positivamente. Esto sera asi siempre que se consideren solo dos variables numericas o que las correlaciones sean transitivas[19].

En un sistema optimamente cuantificado, donde, por definicion, se han asignado cuantificaciones similares a las categorias que se asocian con mayor frecuencia, todas las correlaciones tienden a ser positivas. No obstante, en la practica se manejan bases de

[19] Un sistema de correlaciones es transitivo si se satisfacen las siguientes condiciones para cualquier terna de variables:

A corr(+) B y B corr(+) C ⇒ A corr(+) C

A corr(+) B y B corr(-) C ⇒ A corr(-) C

A corr(-) B y B corr(-) C ⇒ A corr(+) C

datos con un elevado numero de variables que interactuan de forma compleja, impidiendo que esto siempre sea asi, tal y como se ilustra a continuacion.

Supongase, por ejemplo, que en un estudio sobre los habitos de un colectivo estudiantil se incluyen, entre otras, las variables 1: '*horas por semana dedicadas a realizar algun deporte*', 2: '*numero de cervezas consumidas por semana*' y 3: '*horas por semana dedicadas a la lectura*'. Por brevedad, nos referiremos a estas variables como deporte

(D), cerveza (C) y lectura (L), respectivamente. Supongase ademas que la matriz de
correlaciones de estas tres variables, antes de ser transformadas es la que se muestra a
continuacion.
1,00 0, 40 0, 23
0, 40 1,00 0,11
0,23 0,11 1,00
D C L
D
C
L
⌊ − ⌉
|| − ||
|□ |⌋
Las variables deporte y lectura se correlacionan positivamente, por lo que podrian
transformarse con sendas constantes positivas, manteniendose la correlacion positiva
entre las correspondientes variables transformadas. La variable cerveza se correlaciona
negativamente con deporte, pero positivamente con lectura. Luego, no hay forma de
obtener una transformacion para cerveza que se correlacione positivamente con las
transformaciones de deporte y de lectura. Si se transforma mediante una constante
positiva, tendra correlacion positiva con lectura, pero negativa con deporte; si se
transforma mediante una constante negativa, tendra correlacion positiva con deporte,
pero negativa con lectura.
Como se indico anteriormente, en un sistema optimamente cuantificado, todas las
correlaciones tienden a ser positivas, pero dado que ello no siempre es posible, se
obtienen transformaciones que maximicen la suma de correlaciones.
Es importante anotar que la combinacion de restricciones para las variables escaladas
numericamente (linealidad y normalizacion) hace posible obtener las cuantificaciones

cuasi definitivas desde el primer paso, donde se obtiene la matriz de cuantificaciones
inicial mediante normalizacion, pudiendo estas sufrir cambios posteriores solamente en
su direccion, es decir que bastara con multiplicar algunas de ellas por menos uno para
obtener las cuantificaciones definitivas.
Al multiplicar las cuantificaciones de una variable por menos uno, se conserva la
magnitud de las correlaciones con las demas variables, cambiando unicamente su signo.
En general, si la suma de correlaciones de una variable con las demas es mayor al
cambiar el signo que la suma de correlaciones existente antes de cambiarlo, dicho
cambio estara justificado.
Tras la obtencion de la matriz inicial de cuantificaciones, la subrutina **Numini** realiza un
paso de direccionamiento de las variables Numericas y Ordinales, para la maximizacion
de la suma sus correlaciones, tratando a las variables Ordinales como si fueran
Numericas. Aqui se incluyen tambien las variables con nivel de escalamiento Numerico
Flotante y Ordinal Flotante, ignorando durante el proceso las categorias flotantes. Este
paso genera una estructura estable que permite obtener mejores estimaciones iniciales
para las variables Nominales y, consecuentemente, una convergencia mas rapida.
Puesto que el sistema de correlaciones usualmente es complejo y no transitivo, la
subrutina **Numini** debe utilizar un mecanismo que permita encontrar rapidamente la
maxima matriz de correlaciones. Para tal fin, se mide el efecto que tiene el cambio de
signo de cada una de las variables, manteniendo las demas variables constantes y al

final se procede con el cambio que produzca el mayor incremento de la suma total de
correlaciones. Seguidamente se realiza otro ciclo donde se evalua nuevamente el cambio
en las correlaciones totales producido por el cambio de direccion de cada una de las
variables, y se sigue asi hasta que no sea posible un mayor incremento en la suma de
correlaciones. La Tabla 5.1 ilustra los pasos de este proceso para el ejemplo planteado.
Se parte de una suma de correlaciones de 2,90, que es la correspondiente a las variables
no transformadas. En el primer ciclo, al cambiar el sentido de deporte (D), la suma de
correlaciones es 3,56; cambiando el sentido de cerveza (C), 4,04; cambiando el sentido
de lectura (L), 1,50. Por tanto, se cambia el sentido de cerveza, multiplicando sus
cuantificaciones por menos uno, con lo que la nueva suma de correlaciones es 4,04. En

el segundo ciclo, al cambiar el sentido de deporte, la suma de correlaciones es 1,50;
cambiando el sentido de cerveza, 2,90; cambiando el sentido de lectura, 3,56. Dado que
ningun cambio logra incrementar la suma de correlaciones, el proceso concluye,
quedando las variables deporte y lectura con su orientacion inicial y cambiandose la
orientacion de cerveza.

Pasos	**Componentes de la matriz de correlaciones**	**Correlación total**
r total inicial	1 – 0,40 + 0,23 – 0,40 + 1 + 0,11 + 0,23 + 0,11 + 1	2,90
Ciclo 1 –D	1 + 0,40 – 0,23 + 0,40 + 1 + 0,11 – 0,23 + 0,11 + 1	3,56
Ciclo 1 –C	1 + 0,40 + 0,23 + 0,40 + 1 – 0,11 + 0,23 – 0,11 + 1	**4,04**
Ciclo 1 –L	1 – 0,40 – 0,23 – 0,40 + 1 – 0,11 – 0,23 – 0,11 + 1	1,50
Ciclo 2 –D	1 – 0,40 – 0,23 – 0,40 + 1 – 0,11 – 0,23 – 0,11 + 1	1,50
Ciclo 2 –C	1 – 0,40 + 0,23 – 0,40 + 1 + 0,11 + 0,23 + 0,11 + 1	2,90
Ciclo 2 –L	1 + 0,40 – 0,23 + 0,40 + 1 + 0,11 – 0,23 + 0,11 + 1	3,56

Tabla 5.1. Ciclos de evolucion de la correlacion total.

Los correspondientes graficos de cuantificaciones se muestran en la Figura 5.4.

Figura 5.4. Graficos de cuantificaciones para deportes, cerveza y lectura.

Si todas las variables del sistema multivariante son Numericas y/u Ordinales, las direcciones obtenidas mediante este procedimiento se conservan. No obstante, cuando hay variables Nominales, algunas direcciones eventualmente varian durante el proceso.

Aunque para fines de minimizar las *Interdistancias* se permitira el libre cambio de direccion de las variables Numericas y Ordinales (asi como de la parte no flotante de las variables Numericas Flotantes y Ordinales Flotantes), la direccion original de estas

variables se recuperara antes de presentar los resultados finales, mediante la multiplicacion de las cuantificaciones por -1, de manera que las cuantificaciones de estas variables se correlacionen positivamente con las etiquetas originales y, lo mas importante, que las direcciones de las correlaciones iniciales entre estas variables se conserven, lo cual facilitara las interpretaciones de los resultados de otras tecnicas basados en tales cuantificaciones.

Luego, los graficos finales de cuantificaciones para las variables Numericas u Ordinales no tendran nunca el aspecto del grafico de cuantificaciones de cerveza que aparece en la Figura 5.2 (pendiente negativa). Hemos incluido este grafico solo para ilustrar el manejo realizado por la subrutina **Numini** en esta parte del proceso.

SUBRUTINA "Floini"

Si se tienen variables con categorias flotantes, mediante esta subrutina se obtiene una cuantificacion inicial de las mismas, con base en las variables Numericas y Ordinales, tratando a estas ultimas como Numericas. Las categorias fijas de las variables con

escalamiento Numerico Flotante y Ordinal Flotante tambien son usadas como base para
la cuantificacion inicial de las categorias flotantes.
Las categorias flotantes se cuantifican como si formaran parte de variables con
escalamiento Nominal. Su cuantificacion se obtiene, por tanto, como el promedio de las
demas cuantificaciones, tal y como se detalla en el apartado correspondiente a la
cuantificacion de variables Nominales (subrutina **Nom**).
Si hay mas de una variable con categorias flotantes, se evalua inicialmente el cambio
producido en la matriz de correlaciones al cuantificar las categorias flotantes,
cuantificando inicialmente la que genere la mayor suma de correlaciones, y asi
sucesivamente hasta cuantificarlas todas.

Si bien es cierto que el concepto de correlacion solo tiene sentido en variables
Numericas u Ordinales, su uso con las variables transformadas es totalmente valido
puesto que estas tienen propiedades metricas.

INICIALIZACIÓN DE VARIABLES NOMINALES

Se distinguen dos casos: cuando ademas de las variables Nominales hay variables con
otros niveles de escalamiento (Subrutina **Nomini**), y cuando todas las variables son
Nominales (Subrutina **Nominit**).

SUBRUTINA "Nomini"

Si se tiene una mezcla de escalamiento Nominal con otros niveles de escalamiento, se
realiza una categorizacion inicial de las variables Nominales con base en las demas
variables, mediante un procedimiento analogo al usado para la cuantificacion de las
categorias flotantes (Subrutina **Floini**).
Partiendo de una estructura estable proporcionada por las variables Numericas, las

Ordinales y las Flotantes, se introducen las Nominales una a una en el orden que
proporcione para cada paso la mayor suma de correlaciones. Las variables introducidas
en cada paso se usan junto con las demas para la cuantificacion de las siguientes
variables Nominales, hasta cuantificarlas todas.
No esta de mas insistir en el hecho de que no obstante que el concepto de correlacion
solo tiene sentido en variables Numericas u Ordinales, su uso con variables
transformadas, aun de variables Nominales, es totalmente valido, pues todas las
variables transformadas tienen propiedades metricas.

SUBRUTINA "Nominit"

Se utiliza cuando todas las variables estan escaladas a nivel Nominal. En este caso,
puesto que no existe una estructura de partida conformada por las variables Numericas-
Ordinales, se crea una estructura estable a partir de las dos variables que mas se
relacionen entre si.
Si se considera aisladamente cualquier variable y se desea cuantificarla a partir de una
unica variable proveniente de un conjunto, logicamente la mejor cuantificacion sera la
obtenida con base en la variable con la cual este mas relacionada.
Aunque hemos insistido en la validez de utilizar el coeficiente de correlacion como
medida de asociacion entre variables transformadas, debe notarse que en este primer
paso, las variables no han recibido aun ninguna cuantificacion, por lo que son nominales
y no tendria ningun sentido usar el coeficiente de correlacion. Por consiguiente, en esta
parte se evaluan las asociaciones mediante el estadistico Ji cuadrado, el cual se calcula
para todos los posibles pares de variables.

Puesto que el numero de categorias de las variables puede diferir, los estadisticos Ji cuadrado no son directamente comparables entre si. Lo mas adecuado, desde el punto de vista teorico, seria utilizar los correspondientes p-valores, pero esta opcion no resulta viable, puesto que, al nivel de precision de MATLAB, estos se igualan a cero con mucha rapidez. Por tanto, se ha utilizado para fines de comparacion un valor ajustado del estadistico Ji cuadrado, obtenido como lo razon entre el estadistico y sus grados de libertad.

Una vez elegido el par de variables cuyo valor del Ji cuadrado ajustado sea maximo, se realiza un ciclo donde, omitiendo todas las demas variables, se cuantifica la primera variable del par con base en la segunda. Seguidamente, se cuantifica la segunda con base en la primera, y se mantiene el ciclo hasta que se estabilicen las cuantificaciones de ambas variables. A continuacion se incluyen sucesivamente las demas variables, evaluando en cada paso las correlaciones totales y eligiendo la variable cuya inclusion de lugar a la mayor correlacion total, hasta incluirlas todas.

5.3.5 CÁLCULO DE LAS *INTERDISTANCIAS* INICIALES.

La rutina **CUANTIFICA** genera un esquema optimo de cuantificaciones simples, acorde con las definiciones 4.1 y 4.2, lo cual se logra mediante la minimizacion de las *Interdistancias*. Estas se calculan mediante la subrutina **Inter**, la cual implementa la expresion [4.7].

Es importante anotar que el calculo de las *Interdistancias* no se ve afectado por los datos faltantes, siendo posible calcularlas en todos los casos, con la participacion del 100 % de las observaciones multivariantes.

5.3.6 CICLO DE MINIMIZACIÓN DE LAS *INTERDISTANCIAS*.

Esta es la parte central de la rutina **CUANTIFICA**. Se desarrolla mediante una serie de ciclos, en cada uno de los cuales se recuantifica cada una de las variables con base en todas las demas, aplicando un paso de normalizacion tras cada recuantificacion.

La cuantificacion de cada variable se realiza teniendo en cuenta las particularidades de su nivel de escalamiento y buscando minimizar las *Interdistancias*. Para el efecto se han elaborado cinco subrutinas correspondientes a los niveles de escalamiento definidos anteriormente, asi: **Num** (Numerico), **Ord** (Ordinal), **Nom** (Nominal), **Numflo** (Numerico Flotante) y **Ordflo** (Ordinal Flotante).

Un ciclo completo consta de la recuantificacion de cada una de las variables con base en las restantes, eligiendolas en un orden aleatorio, con el fin de evitar posibles sesgos.

Tras la recuantificacion de cada variable se aplica un paso de normalizacion y se calculan las *Interdistancias*, verificando que sean menores que las anteriores.

Si tras la ejecucion de un ciclo se observa alguna disminucion en las *Interdistancias*, se procede con otro ciclo completo. El proceso solo se detiene cuando ninguna de las recuantificaciones logra generar *Interdistancias* menores, lo que significa que se han alcanzado las *Interdistancias* minimas y, en consecuencia, las cuantificaciones optimas.

A continuacion se describe el funcionamiento de cada una de las subrutinas correspondientes a los 5 niveles de escalamiento considerados en el apartado 5.2.

SUBRUTINA “Num” (ESCALAMIENTO NUMÉRICO)

Como se indico anteriormente, la rigidez del escalamiento Numerico se convierte en

una ventaja a la hora de generar cuantificaciones. Si todas las variables fueran
Numericas, tras la normalizacion de la matriz original de datos, bastaria con verificar
los direccionamientos optimos de cada variable que fueron generados por la subrutina
Numini, cambiando el signo de las cuantificaciones donde fuera necesario, para tener
las cuantificaciones optimas.
En la practica, cuando se tienen variables con diferentes niveles de escalamiento, las
cuantificaciones eventualmente pueden cambiar con relacion a las cuantificaciones
iniciales generadas por la subrutina **Numini**, pero el unico cambio posible es de
direccion.
La subrutina **Num** calcula las *Interdistancias* obtenidas al cambiar la direccion de la
variable, es decir, al multiplicar sus cuantificaciones por -1, y las compara con las
Interdistancias de referencia. Si las *Interdistancias* obtenidas son menores, se invertira
la direccion de la variable; en caso contrario, se mantendra en su estado inicial.

SUBRUTINA "Nom" (ESCALAMIENTO NOMINAL)

El escalamiento Nominal se logra mediante la aplicacion de transformaciones
isomorficas, que son las menos restrictivas y las mas generales del conjunto de
transformaciones aplicadas. Todas las demas transformaciones pueden derivarse como
casos particulares de estas.
Una transformacion isomorfica consiste en asignar la misma cuantificacion a todas las
observaciones de una categoria, sin que tenga que satisfacerse ninguna relacion
particular entre las cuantificaciones de diferentes categorias. En tal sentido, la

cuantificacion de todas las categorias es libre. Es por ello que no hace falta definir un
nivel de escalamiento Nominal Flotante cuando se tiene una categoria especial, del tipo
“*no sabe/no responde*”, en una variable escalada a nivel Nominal. Basta con elegir el
nivel de escalamiento Nominal, incluyendo la categoria especial como una mas de las
categorias del conjunto, sin ser necesario siquiera que se le asigne la etiqueta de
maximo valor.
Puesto que el escalamiento Nominal no impone ningun tipo de restriccion para la
cuantificacion de una categoria, siendo esta independiente de la cuantificacion de las
demas categorias, el proceso de cuantificacion de las diferentes categorias de una
variable escalada a nivel Nominal se lleva a cabo, asimismo, de manera independiente,
mediante la asignacion de cuantificaciones optimas a cada una de las categorias, esto es,
mediante la asignacion de cuantificaciones que minimicen las *Interdistancias*.
Aunque la expresion [4.7] define las *Interdistancias*, como una suma sobre las n
observaciones de la muestra, estas pueden obtenerse equivalentemente sumando las
Interdistancias parciales de cada una de las K categorias de una variable cualquiera,
como se indica en la expresion [5.3].

$$Interdistancias\ S_{ik}^{2} = \sum_{k}^{K} \frac{1}{n_k} \sum_{i}^{n_k} \sum_{1}^{1} = \sum\sum \quad [5.3]$$

Donde:
Interdistancias: Criterio escalado de suma de interdistancias cuadraticas.
n: Numero de observaciones (filas de la matriz).

K: Numero de categorias de la variable.
n_k: Numero de observaciones de la k-esima categoria, donde
1
K
$k\ miss$
k
$n\ n\ n$
=
$\sum$ + =, con n_{miss}, el numero de observaciones perdidas.
2
$_{ik}S$: 'Varianza' de la i-esima observacion, de la k-esima categoria.

Puesto que la cuantificacion de cada categoria es independiente de la de las demas, es
posible minimizar la expresion [5.3] minimizando cada uno de sus K sumandos o
Interdistancias parciales, esto es, 2
1
, 1, 2, ,
n_k
ik
i
$S\ k\ K$
=
$\sum$ = ... Para tal efecto, resulta
conveniente recurrir a la expresion original de la suma de interdistancias cuadraticas, tal
y como se presenta en el apartado 4.4 (cf. expresion [4.6]). La suma de interdistancias
cuadraticas para una categoria especifica se denota asi:
( ) 1 2
,
1 1 '
, t. q.
k
ij
$n\ m\ m$
$ij\ ij$
$i\ j\ j\ j$
$q\ q\ j\ q$
–
==>
$\sum\sum\sum$ – $\in \mathbb{R}$ [5.4]
Donde:
n_k: Numero de observaciones (filas) de la categoria evaluada.
m: Numero de variables.
q_{ij}: Cuantificacion de la i-esima categoria de la j-esima variable.

$q_{ij'}$: Cuantificacion de la i-esima categoria de la j'-esima variable.
La restriccion t. q. $_{ij} j\ q \in \mathbb{R}$ indica que esta suma solo se hace sobre los valores
efectivamente observados, ignorando las observaciones perdidas.
Supongamos, ahora, sin perdida de generalidad, que la variable que estamos
cuantificando es la m-esima. Podemos particionar la expresion [5.4] en dos
componentes: uno debido a las distancias entre las primeras m-1 variables (aquellas con
base en las cuales estamos cuantificando la variable de interes) y otro debido a las
distancias entre la cuantificacion de interes (c) y las cuantificaciones de las restantes
m-1 variables, asi.

$$\sum_{'}\sum_{'}(\ )(\ )_{1\,2\,2\,1\,2\,1\,2}$$

$$\sum_{1\,1'}\sum_{1\,1'}\sum_{1\,1}$$

( ) , t. q.

$$\sum_{k\,k\,k}\sum_{ij}$$

$$\sum_{n\ m\ m\ n\ m\ m\ n\ m}\sum_{ij\ ij\ ij\ ij\ ij}$$

$$\sum_{i\ j\ j\ j\ i\ j\ j\ j\ i\ j}$$

$$q\ q\ q\ q\ q\ c\ j\ q$$

$$\sum_{----}$$

$$\sum_{==>==>==}$$

$$\sum\sum\sum - = \sum\sum\sum - + \sum\sum - \in \mathbb{R}$$

La derivada de esta expresion con respecto a c es:

$$\sum(\ )$$

$$\sum(\ )$$

$$\sum_{1\,2}\sum_{'\ 1}\sum_{1\,1'}\sum_{1\,1}$$

$$\sum_{0\ 2\ 2}\sum_{k}\sum_{k}$$

$$\sum_{n\ m\ m}\sum_{ij\ ij\ n\ m}\sum_{i\ j\ j\ j}\sum_{ij}\sum_{i\ j}$$

$$q\ q$$

$$q\ c$$

$$c$$

δ
δ
−
−
==>
==
−
=+−+
ΣΣΣ
ΣΣ
( )
1 1
1 1 1 1
2 2
nk m nk m
ij
i j i j
c q
−−
====
=ΣΣ −ΣΣ
1
1 1
2 (1) 2
nk m
k ij
i j
n m c q
−
==
=−−ΣΣ
Esta derivada se hace cero cuando:
1
1 1
(1)
nk m
ij
i j
k
k
q
c Q
n m
−
==≡
−
ΣΣ
, con kQ , la media de la *k*-esima categoria. [5.5]
Luego, la cuantificacion de una categoria que minimiza la correspondiente componente

de las *Interdistancias* es la que se obtiene promediando las cuantificaciones de las
categorias de las demas variables que aparecen conjuntamente con la categoria en
cuestion.

SUBRUTINA "Ord" (ESCALAMIENTO ORDINAL)

La cuantificacion optima de las variables escaladas a nivel Ordinal se realiza asignando
cuantificaciones a cada una de sus categorias, a traves del mismo procedimiento
utilizado por la subrutina **Nom**, esto es, promediando las cuantificaciones de las demas
variables en el correspondiente grupo de observaciones. Seguidamente, se aplica un
paso de restriccion, en el que se verifica la ordinalidad de la cuantificacion y se aplican
los correctivos necesarios. Esto respalda la afirmacion de que todas las demas
transformaciones pueden derivarse como casos particulares de la transformacion
isomorfica. En este caso, se trata de una transformacion isomorfica con restricciones.

Puesto que en el apartado correspondiente a la subrutina **Nom** se ha detallado como
obtener una cuantificacion optima irrestricta, mediante la aplicacion de
transformaciones isomorficas, no redundaremos en ello y nos concentraremos en la
restriccion por ordinalidad.

La ordinalidad puede ser ascendente o descendente. Para facilitar la aplicacion de la
restriccion por ordinalidad, se hace que todas las ordinalidades sean ascendentes,
modificando temporalmente las descendentes. Para tal efecto, se calcula la correlacion
entre las etiquetas originales de las categorias y las cuantificaciones. En caso de resultar
negativa, se multiplican todas las cuantificaciones por menos uno antes de aplicar la

restriccion por ordinalidad y posteriormente vuelven a llevarse al estado inicial
multiplicandolas nuevamente por menos uno.
Una vez que las cuantificaciones esten correlacionadas positivamente con las
correspondientes etiquetas de sus categorias, se verifica que se satisfaga una relacion de
ordinalidad ascendente entre las mismas, esto es, que $q_1 \leq q_2 \leq ... \leq q_k \leq ... \leq q_K$, donde
q_k es la cuantificacion de la k-esima categoria.
Si se encuentran dos cuantificaciones q_k y $q_{k'}$, con $k < k'$, tales que $q_k > q_{k'}$, sera
necesario igualar ambas cuantificaciones, asi como todas las que se encuentren en
medio de estas ($k < c < k'$). La verificacion se realiza comparando inicialmente el par
mas externo y se va avanzando hacia los pares mas internos, asi:
$q_1 q_1 q_1 q_1 q_2 q_2 q_2 q_1 q_2 q_3 q_2 q_2 q_1 q_1, q, ..., q, q, q, ..., q, ..., q, q, KKKKKKKKKKK q q q q - q q q q q q - q q q - q q - q - q - q <<<<<<<<<<$
Si en todas las comparaciones se satisfacen las relaciones de ordinalidad, se tendran ya
las cuantificaciones optimas, sin necesidad de ninguna modificacion mas alla de
regresar a ordinalidad descendente las cuantificaciones que hayan sido previamente
modificadas[20].

[20] En los casos en que no hay empates en las cuantificaciones de categorias adyacentes, las
cuantificaciones para una variable escalada a nivel Ordinal coinciden con las obtenidas para la misma
variable escalada a nivel Nominal, excepto, posiblemente, por la direccion.

En caso de que al realizar alguna de las comparaciones no se satisfaga la relacion de
ordinalidad, sera necesario igualar las cuantificaciones del par en cuestion asi como las
de todas las categorias intermedias, en caso de que existan.
Puesto que hay infinitas formas de realizar tal procedimiento[21], la aplicacion de esta
restriccion no es una cuestion trivial. De todas las posibles formas de aplicar dicha

restriccion, nos interesa la que produzca las menores *Interdistancias*. Puesto que cualquiera que sea la restriccion aplicada, las categorias sometidas a la misma quedaran con la misma cuantificacion, podemos ver este proceso como un colapsamiento de categorias. Para la cuantificacion optima de esa nueva categoria, conformada por las categorias colapsadas, nos basamos en el resultado obtenido al buscar la cuantificacion que minimizara para una categoria especifica, la correspondiente componente de las *Interdistancias*, (cf. Subrutina **Nom**, expresion [5.5]). Asi, pues, puede afirmarse que la cuantificacion optima de la categoria colapsada estara dada por la cuantificacion promedio de las categorias de las demas variables en el grupo colapsado.

Es importante anotar que el hecho de que un par de categorias o grupo de categorias se colapsen para fines de cuantificacion en un paso especifico, no implica que tengan que permanecer colapsadas en los ciclos subsiguientes. En cada nuevo ciclo se realizaran todos los pasos aqui descritos, partiendo de una cuantificacion Nominal y aplicando las restricciones por ordinalidad que sean necesarias en ese paso especifico.

SUBRUTINA "Numflo" (ESCALAMIENTO NUMÉRICO FLOTANTE)

Esta subrutina genera cuantificaciones que minimizan las *Interdistancias* para una variable escalada a nivel Numerico Flotante. Es decir que las primeras $K-1$ categorias se escalan a nivel Numerico y la K-esima categoria se escala a nivel Nominal.

21 Haciendo ambas cuantificaciones iguales a la mayor de ellas, haciendo ambas cuantificaciones iguales a la menor, promediando ambas cuantificaciones o igualandolas a cualquier otro punto ubicado entre ambas.

En este caso se obtiene la cuantificacion optima para la categoria flotante, con base en el

promedio de las categorias de las demas variables en el correspondiente grupo, acorde
con el resultado [5.5]. Para la parte Numerica de la variable, es decir, las primeras K-1
categorias, se evaluan sus dos posibles configuraciones, es decir, la que traiga del paso
anterior y esa misma configuracion multiplicada por menos uno. Tras el paso de
normalizacion, se elige aquella configuracion que de lugar a las menores *Interdistancias*.

SUBRUTINA "Ordflo" (ESCALAMIENTO ORDINAL FLOTANTE)

Esta subrutina genera cuantificaciones optimas para una variable escalada a nivel
Ordinal Flotante, es decir, que las primeras K-1 categorias se escalan a nivel Ordinal y
la K-esima categoria se escala a nivel Nominal.
La parte Ordinal de la variable se cuantifica con base en el procedimiento descrito en la
subrutina **Ord**, mientras que la K-esima categoria se escala a nivel Nominal con base en
el procedimiento descrito en la subrutina **Nom**.

5.3.7 PRESENTACIÓN DE RESULTADOS.

El procedimiento descrito anteriormente (§ 5.3.1 hasta § 5.3.6) genera escalamientos y/o
cuantificaciones optimas para cada una de las variables del conjunto multivariante,
respetando las restricciones propias de cada nivel de escalamiento.
Tal y como se indico en los parrafos finales de la descripcion de la subrutina **Numini**
(§ 5.3.4), antes de presentar los resultados, se realizan los cambios de direccion
necesarios para que las cuantificaciones de las variables Numericas y Ordinales (asi
como de las Numericas Flotantes y Ordinales Flotantes, sin incluir la categoria flotante),
se correlacionen positivamente con las etiquetas originales de las categorias. Esto se
obtiene multiplicando por menos uno las cuantificaciones de las variables en cuestion.

El objetivo principal de esta modificacion es hacer que se conserven las direcciones de las correlaciones iniciales entre tales variables, lo cual hace mas facil y directa la interpretacion de los resultados de otras tecnicas que se basen en dichas cuantificaciones. Por tanto, todas las salidas finales del programa incorporan esta modificacion.

Todos los resultados alfanumericos, tanto los que aparecen por pantalla como los que se guardan en archivo son generados por la subrutina **Res**. Los resultados graficos, esto es, los graficos de cuantificaciones son generados mediante la subrutina **Gcuant**.

SUBRUTINA "Res" (RESULTADOS ALFANUMÉRICOS)

Ademas de la Tabla de Cuantificaciones que se muestra en la ventana de comandos de MATLAB (y desde donde puede copiarse si se desea), la subrutina **Res** genera dos hojas adicionales en el archivo que contiene la base de datos de entrada: TC, que contiene la **Tabla de Cuantificaciones** y MC con la **Matriz de Cuantificaciones**.

Tabla de Cuantificaciones. Consiste en un arreglo tabular en el que se muestra para cada variable la cuantificacion optima de cada una de sus categorias. Alli se indica tambien el nivel de escalamiento con base en el cual se ha obtenido la cuantificacion optima de cada variable.

La Tabla de Cuantificaciones que se muestra por pantalla y la que se guarda en la hoja TC solo difieren en formato, pero contienen exactamente la misma informacion. En la Figura 5.5 se muestra el aspecto general de tales tablas para la misma seccion de un conjunto de cuantificaciones.

Matriz de Cuantificaciones. Se trata de una matriz con la misma estructura que la matriz de datos utilizada para la obtencion de las cuantificaciones[22], en la que se han

[22] En caso de que no se haya generado una matriz de datos reducida (cf. § 5.3.2), la comparacion con la

base de datos original se realiza omitiendo las filas de escalamiento y de filtrado. En los casos en que si se
haya generado una matriz de datos reducida, las estructuras de ambas matrices coinciden celda a celda.

remplazado las etiquetas iniciales de las categorias por sus correspondientes
cuantificaciones optimas. Todas las variables de esta matriz gozan de propiedades
metricas, lo que hace posible analizarlas mediante procedimientos multivariantes
lineales.

Figura 5.5. Aspecto general de la Tabla de Cuantificaciones, mostrada por pantalla
(izquierda) y guardada en la hoja TC (derecha).

Puesto que el procedimiento de cuantificacion optima no utiliza ningun metodo de
imputacion para los valores faltantes, la matriz cuantificada tendra exactamente los
mismos valores faltantes que la matriz original. En esta, los valores perdidos aparecen
como espacios en blanco.

Consideraciones sobre el archivo Excel.

Aunque el proceso de adicion de las hojas de TC y MC no altera la base de datos
original, localizada en la primera hoja del archivo, siempre es recomendable mantener
respaldos de la informacion.

Debe verificarse que el archivo no contenga hojas llamadas TC ni MC, pues en tal caso
serian sobrescritas con la informacion de la Tabla de Cuantificaciones y la Matriz de
Cuantificaciones.

Asimismo, si se realizan diferentes ejecuciones de **CUANTIFICA**, cambiando las
especificaciones de filtrado, en particular, excluyendo variables del analisis, se
recomienda limpiar el contenido de las hojas TC y MC o eliminarlas del archivo, pues

la rutina de escritura de resultados solo sobrescribe el rango especifico de salida de la
informacion, dejando inalterado el resto de la hoja, lo que puede dar lugar a que se
cometan errores con el posterior manejo de la informacion.
Supongase, por ejemplo, que se inicialmente se realiza un analisis con 15 variables y
que posteriormente, tras filtrar 5 de las variables, se realiza un nuevo proceso de
cuantificacion. La rutina de resultados solamente sobrescribira en el rango
correspondiente a las 10 primeras variables. Luego, la Tabla de Cuantificaciones de la
nueva ejecucion quedara sobrescrita en el rango de filas que corresponda a las 10
variables incluidas en la ejecucion, pero esta informacion estara seguida por un grupo de
filas remanentes correspondientes a las cuantificaciones de las ultimas cinco variables
de la ejecucion anterior. Asimismo, la Matriz de Cuantificaciones seguira teniendo 15
columnas: las 10 primeras correspondientes a las variables incluidas en el ultimo
analisis y las 5 siguientes correspondientes a las ultimas 5 variables de la ejecucion
anterior.
Desde luego, tales precauciones no seran necesarias cuando no se modifique el vector
de filtrado o cuando la modificacion consista en agregar variables al analisis.
Finalmente, tras realizar modificaciones sobre el archivo de Excel, este debera cerrarse
para permitir su sobrescritura.

SUBRUTINA “Gcuant” (RESULTADOS GRÁFICOS)

Los resultados graficos del proceso de cuantificacion optima generado por la rutina
CUANTIFICA, consisten en un **Gráfico de Cuantificaciones** para cada una de las
variables involucradas en el proceso, ubicando las etiquetas originales de las categorias

en la abcisa y sus correspondientes cuantificaciones optimas en la ordenada.

Para las variables escaladas a nivel Numerico, el grafico sera una linea recta con pendiente positiva; si la variable se escala a nivel Ordinal, se observaran relaciones monotonas no descendentes y si la variable se escala a nivel Nominal, el grafico puede asumir cualquier forma. Para las variables con nivel de escalamiento flotante se satisfara la relacion base entre las primeras K-1 categorias; la ultima categoria sera flotante. Las Figura 5.6, 5.7, 5.8, 5.9 y 5.10 muestran graficos de cuantificaciones tipicos para variables escaladas a nivel Numerico, Ordinal, Nominal, Numerico Flotante y Ordinal Flotante, correspondientemente.

Figura 5.6. Linea recta, tipica de los graficos de cuantificaciones de variables escaladas a nivel Numerico.

Figura 5.7. Funcion no decreciente, tipica de los graficos de cuantificaciones de variables escaladas a nivel Ordinal.

Figura 5.8. Grafico sin ningun patron definido, tipico de las cuantificaciones de variables escaladas a nivel Nominal.

Figura 5.9. Linea recta para las primeras K-1 categorias, tipica de los graficos de cuantificaciones de variables escaladas a nivel Numerico Flotante.

Figura 5.10. Funcion no decreciente para las primeras K-1 categorias, tipica de los graficos de cuantificaciones de variables escaladas a nivel Ordinal Flotante.

5.4 COMPARACIÓN ENTRE CUANTIFICA Y EL SISTEMA GIFI

Ya hemos indicado que las tecnicas del sistema Gifi de analisis multivariante no lineal pueden usarse para la obtencion de cuantificaciones que permitan la subsecuente aplicacion de tecnicas lineales; o para generar representaciones graficas en subespacios de baja dimensionalidad, que resuman las relaciones entre objetos, variables y/u objetos y variables. (cf. § 3.2).

Puesto que el producto unico de la rutina **CUANTIFICA** es la cuantificacion, compararemos nuestra propuesta con las alternativas del sistema Gifi cuando son utilizadas como primer paso para la obtencion de cuantificaciones. Posteriormente, en § 6.10, contrastaremos la segunda faceta de las tecnicas del sistema Gifi (generacion de representaciones graficas) con nuestra propuesta usada conjuntamente con las tecnicas de representacion Biplot (GABRIEL, 1971).

La Tabla 5.2 sintetiza los aspectos de comparacion entre **CUANTIFICA** y las principales tecnicas del sistema Gifi, cuando son utilizadas para la obtencion de cuantificaciones.

Aspecto	CUANTIFICA	PRINCALS	HOMALS	OVERALS
Manejo simetrico de todas las variables	Si	Si	Si	No
Posibilidad de agrupacion	No	No	No	Si
Puede usar toda la informacion aun cuando hay datos faltantes	Si	Si	Si	No
Flexibilidad en niveles de escalamiento	Si	Si	No	Si
Escalamiento Numerico Flotante u Ordinal Flotante	Si	No	No	No
Cuantificaciones unicas	Si	No	No	No
Interdistancias minimas	Si	No	No	No

Tabla 5.2. Comparacion entre **CUANTIFICA** y las principales tecnicas del sistema Gifi.

Aunque en terminos teoricos las tecnicas del sistema Gifi estan interrelacionadas entre

si, de modo que algunas pueden concebirse como casos particulares de otras mas
generales (cf. § 2.7), su implementacion algoritmica conlleva algunas diferencias

practicas. Es por esta razon que aunque PRINCALS y HOMALS pueden considerarse,
desde el punto de vista teorico, casos particulares de OVERALS, los hemos incluido en
la Tabla 5.2, pues asi recogemos los tres modulos de escalamiento optimo del sistema
Gifi con sus correspondientes particularidades.
La Tabla 5.2 muestra que mientras **CUANTIFICA**, PRINCALS y HOMALS procesan todas las variables en igualdad de condiciones, OVERALS las analiza por grupos. Esta
caracteristica de OVERALS, ademas de implicar un manejo diferente para las variables,
exige observaciones completas por grupos, es decir, que para que una observacion sea
utilizada en el calculo de las cuantificaciones de un grupo, debe contar con registros en
todas las variables del correspondiente grupo. Puesto que ni **CUANTIFICA**, ni
PRINCALS, ni HOMALS manejan grupos, pueden utilizar toda la informacion disponible, aun cuando haya datos faltantes.
En terminos de escalamiento, HOMALS es la tecnica mas restrictiva, puesto usa el nivel
de escalamiento Nominal para todas las variables. Este es solo uno de los posibles
niveles de escalamiento que puede usar **CUANTIFICA**, asi como las demas tecnicas
comparadas del sistema Gifi. En tal sentido, HOMALS puede verse como un caso
particular de PRINCALS o de OVERALS.
Es necesario discutir varios aspectos relacionados con los niveles de escalamiento
flotante. En primer lugar, debe tenerse presente que tales niveles de escalamiento solo
tienen sentido cuando se desea escalar las restantes categorias de la variable con base en

el esquema del escalamiento Numerico o del escalamiento Ordinal, pues el escalamiento
Nominal implica que todas sus categorias son libres o flotantes (cf. § 5.2). Esto excluye
la posibilidad de un nivel de escalamiento flotante en HOMALS.
PRINCALS posee un mecanismo que, bajo ciertas condiciones, procesa los datos de
una categoria de forma analoga a como lo hace **CUANTIFICA** con las categorias
flotantes, es decir, que realiza una cuantificacion libre para dicha categoria, respetando
para las demas categorias de la variable el nivel de escalamiento definido (bien sea
Numerico u Ordinal).

La diferencia entre el mecanismo utilizado por PRINCALS y la forma en que **CUANTIFICA** procesa la informacion cuando se define un nivel de escalamiento
flotante, tiene su origen en la concepcion de los niveles de escalamiento flotante.
Durante el desarrollo de **CUANTIFICA** hemos considerado la posibilidad de que la
variable posea categorias que no sigan el patron general de las demas, tales como las
que surgen cuando las encuestas incluyen un item con la etiqueta "*no sabe/no responde*". Para contrastar esta categoria con las demas categorias de la misma variable
asi como con las categorias de las demas variables, debera asignarsele una
cuantificacion, la cual, desde luego, debera ser libre o flotante, por no conocerse su
rango en relacion con las demas categorias de la variable de la cual forma parte (cf.
§ 5.2). El uso de este nivel de escalamiento no excluye la posibilidad de que la variable
en cuestion ademas contenga datos faltantes o perdidos, los cuales, en la mayoria de
ocasiones no tiene sentido agrupar en una categoria para fines de cuantificacion.

PRINCALS, por su parte, no permite manejar independientemente las dos situaciones
anotadas anteriormente. Cuando se usa PRINCALS hay que elegir una de dos opciones:
imputar los datos considerandolos como una categoria extra o no considerarlos en el
analisis. PRINCALS no permite manejar parte de la informacion como una categoria
extra (categoria flotante) y otra parte como datos faltantes. Esta diferencia de manejo en
relacion con **CUANTIFICA** proviene, como ya lo hemos indicado, de la concepcion
misma de ambos mecanismos. En las tecnicas del sistema Gifi no se plantea la
posibilidad de que exista una categoria extra que, sin ser informacion faltante, no siga el
patron de las demas categorias. El proceso senalado se concibe solamente como una
forma de manejar la informacion faltante. Es por ello que en las tecnicas del sistema
Gifi este proceso no forma parte de los niveles de escalamiento, impidiendo, por tanto,
el manejo simultaneo de una categoria extra o flotante y de datos faltantes por otras
causas.
En resumen, aunque PRINCALS incluye un mecanismo para manejo de informacion
perdida, que, en ausencia de informacion perdida, puede servir para realizar una
cuantificacion equivalente a la que realiza **CUANTIFICA** al escalar una variable a nivel
flotante, este mecanismo no puede utilizarse cuando hay datos perdidos. Podemos

establecer, por tanto, que las tecnicas del sistema Gifi no consideran los niveles de
escalamiento flotante.
Esta opcion de manejo de informacion faltante, implementada en PRINCALS, permite

ejemplificar lo anotado anteriormente sobre tecnicas del sistema Gifi que, siendo teoricamente iguales, pueden exhibir diferencias practicas. En terminos teoricos, la solucion PRINCALS es equivalente a la solucion OVERALS asignando una variable a cada grupo, esto es, un OVERALS sin agrupamiento. En la practica, sin embargo, si se utiliza el modulo OVERALS con una variable por grupo, para llegar a la solucion PRINCALS, no es posible manejar los datos faltantes como categoria extra, tal y como puede hacerse cuando se usa el modulo PRINCALS.

Tal y como se ha indicado y ejemplificado anteriormente, las tecnicas del sistema Gifi generan varios posibles conjuntos de cuantificaciones, en funcion de la dimensionalidad elegida (cf. Capitulo 3). La deteccion de esta situacion y la busqueda de la mejor solucion constituyeron la motivacion inicial de la presente investigacion. Aunque en el Capitulo 4 presentamos un criterio que permite seleccionar el mejor conjunto de cuantificaciones, partiendo de varios posibles conjuntos, nuestra mayor aportacion consiste en la implementacion del sistema de cuantificacion optima **CUANTIFICA**, el cual no involucra el concepto de dimensionalidad, proporcionando siempre un unico conjunto de cuantificaciones optimas.

En el Capitulo 4 definimos y sustentamos que la cuantificacion optima de un sistema multivariante es la que asigna valores mas similares a las categorias de las distintas variables que se asocian con mayor frecuencia. Una forma objetiva de medir la satisfaccion de tal condicion es mediante las *Interdistancias*, criterio escalado de la suma sobre todas las observaciones de las distancias cuadraticas entre las categorias

cuantificadas de las diferentes variables. Un sistema de cuantificacion optima debera
tener, por tanto, *Interdistancias* minimas. La implementacion algoritmica de
CUANTIFICA esta basada justamente en la satisfaccion de tal condicion.
Como sintesis de las anteriores comparaciones, la tecnica del sistema Gifi que tiene
mayor similitud con **CUANTIFICA** es PRINCALS, puesto que, al igual que

CUANTIFICA, procesa todas las variables en igualdad de condiciones, usa toda la
informacion disponible aun cuando la matriz tenga datos faltantes y permite usar los
niveles de escalamiento Numerico, Ordinal y Nominal.
Notese que los aspectos que marcan la diferencia entre **CUANTIFICA** y PRINCALS son
precisamente aquellos que le confieren superioridad a **CUANTIFICA** frente al conjunto
de tecnicas del sistema Gifi y que resumimos a continuacion:
1) **CUANTIFICA** brinda la posibilidad de realizar escalamientos flotantes con
independencia de si hay o no hay datos faltantes.
2) **CUANTIFICA** genera cuantificaciones unicas, al no basar su solucion en el
concepto de dimensionalidad.
3) **CUANTIFICA** genera cuantificaciones que minimizan las *Interdistancias*, con lo
cual, las categorias mas frecuentemente asociadas reciben cuantificaciones mas
similares.

CAPÍTULO 6
Aplicación De La Propuesta: Cardiopatía Isquémica

6.1 INTRODUCCIÓN

En este capítulo se analiza una base de datos de pacientes con cardiopatía

isquémica o sintomatologia relacionada con la misma, procedente de la Unidad de
Cardiologia Nuclear y de Ergometria del Hospital Universitario de Salamanca. Aunque
la base de datos esta estructurada en una matriz de 6.965 pacientes x 53 variables, el
53,12 % de las lecturas de dicha matriz no estan presentes, lo cual representa un desafio
tanto para la cuantificacion como para el posterior analisis, pues en general es mas
directo trabajar con matrices completas que con matrices que tienen celdas vacias.
**Teniendo en cuenta que el escenario de bases de datos con información faltante es
quizá el más frecuente en las ciencias aplicadas, en este capítulo presentamos, a
través del análisis de la base de datos que nos concierne, un método integral de
análisis que puede ser aplicado en situaciones análogas**.
Es importante senalar que la presente base de datos, a pesar del alto porcentaje de
informacion faltante, sigue contando con gran cantidad de informacion. Notese que las
lecturas efectivamente registradas ascienden a 173.042 observaciones, cifra nada
despreciable.
El sistema multivariante objeto de este estudio esta conformado por variables con
diferentes escalas de medicion, es decir, que se trata de un sistema multivariante mixto
(cf. § 1.2). Por tanto, **inicialmente se realiza una cuantificación óptima de las
variables con base en la propuesta desarrollada en este trabajo**, utilizando la rutina
CUANTIFICA, descrita en el Capitulo 5. Las variables optimamente cuantificadas tienen
propiedades numericas, por lo que seguidamente pueden ser analizadas mediante
cualquier tecnica lineal.
Puesto que analizamos una base de datos con tanta informacion, en particular en lo que

al numero de variables se refiere, nuestro interes se centra en explorar las principales
relaciones, mas que en contrastar hipotesis especificas. Para el efecto nos apoyamos en
los **Métodos Biplot** (GABRIEL, 1971), los cuales permiten resumir y representar los
principales patrones de asociacion entre variables, individuos, asi como entre variables e
individuos, partiendo de una matriz numerica de individuos *x* variables.

Dado que la rutina **CUANTIFICA** no requiere utilizar ningun metodo de imputacion en el
evento de datos faltantes, la matriz de variables optimamente escaladas carece,
asimismo, del 53,12 % de la informacion. Puesto que las representaciones Biplot, en su
forma clasica se basan en la descomposicion en valores singulares de la matriz que se
desea representar, la cual no puede obtenerse para una matriz incompleta, es necesario
construir una representacion Biplot alternativa. **Detallamos el razonamiento y el
procedimiento seguido para construir una representación Biplot acorde con nuestros objetivos, a partir de la matriz de variables óptimamente cuantificadas
con información faltante**.
Aunque **tanto la propuesta para la obtención de representaciones Biplot a partir de
matrices con información faltante como los algoritmos computacionales utilizados
para el efecto son también aportaciones originales de este trabajo**, hemos preferido
mantenerlos como un apartado de este capitulo de la aplicacion practica, mas que en los
capitulos de desarrollo teorico, por concebirlos como un complemento a la tematica
central de este trabajo que es la cuantificacion optima.
Para esta base de datos, en particular, una vez obtenidas las cuantificaciones optimas,
hemos analizado los patrones de asociacion mas sobresalientes a traves de

representaciones GCMP-Biplot de las primeras cuatro dimensiones factoriales.

Cerramos este capitulo con la **comparación de los resultados obtenidos mediante la aplicación de las técnicas propuestas y los que se obtienen al usar técnicas del sistema Gifi**.

6.2 GENERALIDADES DE LA CARDIOPATÍA ISQUÉMICA

El sistema cardiovascular o aparato cardiovascular[23] esta conformado por el corazon, los vasos sanguineos y la sangre. Es el encargado de transportar nutrientes, hormonas y gases hacia todas las celulas del cuerpo y de recoger los desechos metabolicos que alli se generan. Tambien desempena un importante papel en la estabilizacion del pH celular y la temperatura corporal (GANONG, 2005).

Existen diversos factores que pueden causar una disminucion transitoria o permanente del riego sanguineo a un tejido, con la consecuente disminucion del aporte de oxigeno al mismo. Al sufrimiento celular resultante de esta situacion se le denomina isquemia. La ausencia de oxigeno puede causar danos o disfuncion del tejido o ser suficientemente severa como para causar su muerte, siendo diferente el nivel de tolerancia de cada tejido. Entre los tejidos especialmente sensibles a la falta de oxigeno se encuentran los del rinon, el cerebro y el corazon (GUYTON y HALL, 2001).

El corazon es el organo que se encarga de bombear la sangre a traves de los vasos sanguineos, mediante contracciones ritmicas repetidas. Esta conformado esencialmente por un tejido muscular denominado miocardio y, en menor proporcion, por tejido conectivo y fibroso (tejido de sosten, valvulas). Las arterias coronarias que rodean el corazon son las encargadas de su irrigacion sanguinea (GANONG, 2005).

En las arterias sanas, las paredes internas son suaves y de grosor uniforme, permitiendo
que los globulos rojos y demas sustancias que viajan en la sangre fluyan libremente
hacia el corazon. No obstante, con frecuencia se generan depositos grasos llamados
placas de ateroma, conformados por colesterol, compuestos grasos, calcio y una
sustancia coagulante denominada fibrina, los cuales pueden endurecerse y fijarse a las
paredes, en un proceso llamado arteriosclerosis, lo que produce el estrechamiento de la
arteria (estenosis). **A la arteriosclerosis de las arterias coronarias se le llama
enfermedad coronaria o cardiopatía isquémica** (BRAUNWALD et al., 1997).

[23] En anatomia, la denominacion de *sistema* corresponde a un conjunto de organos formados predominantemente por el mismo tipo de tejidos (como en el caso del sistema nervioso). Puesto que los
organos que contribuyen a la funcion cardiovascular estan formados por diferentes tejidos (miocardio,
endotelio, sangre), desde este punto de vista seria mas adecuada la denominacion *aparato* cardiovascular.
No obstante, si se acude a la definicion mas general de sistema: "*conjunto de cosas que relacionadas
entre si contribuyen a determinado objeto*", puede validarse el uso del termino *sistema.*

Entre los principales factores de riesgo involucrados en el desarrollo de la
arteriosclerosis, se cuentan la predisposicion genetica, la hipertension arterial, la
diabetes mellitus, la dislipidemia, el sobrepeso/obesidad, el tabaquismo, el sedentarismo
y algunos trastornos metabolicos como el hipertiroidismo o los sindromes post
menopausia en mujeres (SELWYN et al., 1997).
El espesor de la placa de ateroma puede ser suficiente para disminuir o bloquear el flujo
sanguineo al miocardio, lo que puede producir un dano cardiaco, tecnicamente llamado
infarto de miocardio. No obstante, la mayoria de infartos del miocardio se producen

cuando la cubierta calcificada de la placa se rompe, dejando expuesto el nucleo graso al torrente sanguineo, lo que hace que la sangre se coagule y forme trombos, los cuales pueden bloquear el flujo sanguineo al corazon (BRAUNWALD et al., 1997).

Por otra parte, las paredes de las arterias coronarias tienen musculo, el cual puede sufrir espasmos que producen un mayor estrechamiento del vaso en una zona determinada.

Los espasmos coronarios pueden ocurrir sin causa aparente, pero tambien por exposicion al frio, por fuerte estres emocional o por el uso de sustancias psicoactivas (BRAUNWALD et al., 1997).

Los principales sindromes de la cardiopatia isquemica son la angina de pecho (dolor toracico y sensacion de opresion aguda y sofocante), el infarto de miocardio y la muerte subita[24].

[24] El termino *infarto de miocardio* se refiere a la muerte de tejido muscular cardiaco, sin que ocurra necesariamente la muerte del paciente. La mayoria de paradas cardiacas que conducen a la *muerte subita* ocurren cuando los impulsos electricos se vuelven rapidos (taquicardia ventricular) o caoticos (fibrilacion ventricular) o ambos. Este ritmo cardiaco irregular (arritmia) hace que el corazon deje de latir subitamente, produciendose la muerte del paciente. Luego, aunque un infarto de miocardio puede causar una parada cardiaca y la muerte del paciente, los terminos *muerte por infarto de miocardio* y *muerte subita* no son sinonimos.

6.3 IMPACTO DE LA CARDIOPATÍA ISQUÉMICA

La Organizacion Mundial de la Salud senala en su informe de 2004 sobre la salud en el mundo (OMS, 2004) que la cardiopatia isquemica es la principal causa de muerte a nivel mundial, cobrando 7,2 millones de vidas cada ano, lo que representa el 12,6 % del total de muertes, siendo esta cifra 2,5 veces mayor que la de las muertes causadas por

VIH/SIDA y 6 veces mayor que la de las muertes por canceres de traquea, bronquios y
pulmon. En proyecciones realizadas por la Organizacion Mundial de la Salud
(MATHERS and LONCAR, 2005) se estima que para el ano 2030 esta enfermedad seguira siendo el primer factor de mortalidad a nivel mundial (13,1 %).
Segun el Ministerio de Sanidad y Consumo[25], durante 2005 se produjeron en Espana un
total de 39.313 defunciones (22.188 en hombres y 17.125 en mujeres) por cardiopatia
isquemica. La tasa de mortalidad por 100.000 habitantes fue de 103,84 para hombres y
de 77,73 para mujeres, representando la primera causa de muerte en hombres y la
segunda en mujeres, despues de la enfermedad cerebrovascular.
OLIVA et al. (2004) estiman que la perdida de productividad laboral en Espana por
mortalidad prematura a causa de cardiopatia isquemica asciende a 460,45 millones de
euros anuales. Las perdidas por incapacidad laboral temporal se estiman en 187,06
millones de euros anuales, mientras que la incapacidad permanente genera perdidas que
estan entre 431 y 488,5 millones de euros anuales. Luego, las perdidas totales derivadas
de la cardiopatia isquemica oscilarian entre 1.078 y 1.136 millones de euros anuales[26].
Notese que en estas estimaciones no se incluyen los costes directos de la prevencion y el
tratamiento de la enfermedad, sino solamente los costes indirectos de la misma, los
cuales muchas veces son ignorados.

[25] Series 1981-2005: Mortalidad por causa de muerte, Espana y comunidades autonomas. En el sitio:
http://www.msc.es/estadEstudios/estadisticas/estadisticas/estMinisterio/mortalidad/seriesTablas.htm,
visitado el 24 de julio de 2008.

[26] Estimaciones para el ano 2003.

6.4 BASE DE DATOS

La Unidad de Cardiologia Nuclear y de Ergometria del Hospital Universitario de
Salamanca evalua anualmente numerosos pacientes 27 remitidos por cardiopatia
isquemica diagnosticada o por sospecha de la misma. El historial completo de cada
paciente, incluyendo antecedentes personales y familiares, evaluacion general,
resultados de pruebas y diagnosticos, asi como el seguimiento se encuentra
sistematizado, contandose actualmente con registros que datan desde 1988 hasta la
fecha.
Aunque se han realizado numerosos estudios enfocados en hipotesis especificas
(DIEGO DOMINGUEZ et al., 2005; DIEGO DOMINGUEZ et al., 2006; MARTINMOREIRAS
et al., 2005; RAMIREZ CASTRO et al., 2006; RUANO et al., 2005a;
RUANO et al., 2005b), la magnitud de la base de datos, asi como su complejidad, dada
por la mezcla de variables que la conforman, hacen dificil realizar un proceso
exploratorio general que permita usar simultaneamente gran parte de la informacion
disponible y resumir los patrones de asociacion mas caracteristicos entre las variables,
entre individuos, asi como entre individuos y variables.
Mediante el presente estudio se pretende hacer una aportacion en tal sentido. Para tal
efecto, partimos de una base de datos de 6.965 pacientes y 53 variables, en la cual el
53,12 % de las lecturas estan ausentes, lo cual deja un total de 173.042 observaciones
efectivas.
Las bases de datos con informacion incompleta surgen en la practica quiza con mas
frecuencia que las bases de datos completas. Muchas de las bases de datos “completas”
usadas en los analisis no son mas que versiones reducidas de bases de datos mayores en
las cuales se han eliminado las filas con observaciones faltantes.

Aunque esta es una practica rebatible, eventualmente podria ser viable. Sin embargo, en
algunas situaciones, como la presente, esta practica es totalmente inviable. Si
eliminaramos todas las filas con al menos una observacion faltante, pretendiendo
[27] Cada ano se realizan alrededor de 2.000 pruebas de esfuerzo y 1.000 SPECT.

mantener las 53 variables, nos quedariamos con cero pacientes. Si se eliminan
sucesivamente las variables con menos lecturas, obtendriamos 107 pacientes con
lecturas completas en 48 variables. Si seguimos reduciendo el numero de variables, con
el fin de incluir un mayor numero de pacientes, tendriamos que trabajar con solamente
39 variables para poder tener 1.054 pacientes con observaciones en todas las variables.
Esto, desde luego, iria en detrimento del analisis multivariante que pretendemos
realizar.
Otra alternativa para el manejo de bases de datos incompletas consiste en utilizar algun
metodo de imputacion que permita estimar los valores faltantes con base en la
informacion observada. Si bien, esto facilitaria bastante todos los procesos
subsiguientes, pues, en general, es mas facil trabajar con matrices completas que con
matrices en las que falte informacion, los resultados obtenidos con base en las matrices
imputadas se verian afectados por el metodo de imputacion utilizado.
Antes de plantear algun eventual metodo de imputacion, hay que considerar que las
matrices incompletas pueden tener diversos origenes. Puede tratarse de matrices en las
que las celdas vacias son producto de informacion efectivamente generada, que por
algun motivo no llego a la matriz final o fue retirada de esta. En este caso, hablamos de

datos perdidos. En contraste, con este conjunto de situaciones, las matrices pueden ser
incompletas desde su origen, por no haberse generado nunca el elemento correspondiente a las celdas vacias. Utilizaremos el termino **datos faltantes** para definir
esta segunda situacion.
En el primer caso, puede ser viable y hasta recomendable utilizar algun metodo de
imputacion que permita estimar de la mejor manera posible la informacion perdida. En
el segundo caso, ello no tendria sentido y **cualquier intento por estimar la información que nunca ha existido no haría más que introducir ruido**.
Las matrices con informacion faltante -mas no perdida- son comunes en bases de
datos provenientes de la practica clinica, donde cada paciente puede ser sometido a
diferentes tipos de pruebas. El escenario del presente analisis se corresponde con dicha
situacion. En este caso, la base de datos de la Unidad de Cardiologia Nuclear y de
Ergometria incluye diferentes pruebas diagnosticas, v. gr., SPECT, ecocardiografia y

coronariografia. Algunos pacientes pueden ser sometidos a las tres pruebas, mientras
que otros son evaluados mediante dos o solo una de ellas (cf. § 6.5). Luego, si un
paciente especifico solo es evaluado mediante SPECT, no tiene ningun sentido tratar de
estimar los resultados de la ecocardiografia y de la coronariografia, pues no son datos
perdidos, sino datos que nunca han existido.
En § 6.7 tomaremos en consideracion este hecho para elaborar nuestra propuesta de
representaciones Biplot. Entretanto, vale la pena recordar que nuestra propuesta de
cuantificacion, implementada en la rutina **CUANTIFICA**, utiliza el 100 % de la
informacion disponible, sin hacer uso de ningun metodo de imputacion (cf. § 5.2). Esto

implica que la matriz cuantificada tiene exactamente la misma estructura de datos
faltantes que la matriz original.
Mas alla de la indiscutible importancia que desde el punto de vista clinico tiene el
estudio de la cardiopatia isquemica, la base de datos utilizada en el presente estudio
resulta optima desde el punto de vista tecnico para ilustrar la funcionalidad de la rutina
CUANTIFICA como herramientas de cuantificacion en condiciones extremas. Por una
parte, tal y como acabamos de indicar, el hecho de tener una base de datos incompleta,
exige en general manejos mas complejos que los que podrian realizarse sobre las
matrices completas correspondientes a bases de datos sin informacion faltante. Por otra
parte, el considerable tamano de la base de datos, tanto en numero de pacientes como en
numero de variables, aunado a las diferentes escalas de medicion de las variables que
conforman el sistema multivariante ponen a prueba los diferentes componentes de la
rutina **CUANTIFICA**.
A continuacion nombramos las 53 variables que conforman la base de datos:
BRI: Bloqueo de rama izquierda del haz de hiss.
CI: Cardiopatia isquemica en el diagnostico preliminar.
EA: Estado actual.
DM: Diabetes mellitus.
HTA: Hipertension arterial sistemica.
Edad: Edad.
IMC: Indice de masa corporal.
Sexo: Sexo.

SSS manual DA: Score manual en estres de defectos en territorio de la arteria descendente anterior.
SSS manual Cx: Score manual en estres de defectos en territorio de la arteria circunfleja.
SSS manual CD: Score manual en estres de defectos en territorio de la arteria coronaria derecha.
SSS manual Total: Score manual total de defectos en estres.
SSS auto DA: Score automatico en estres de defectos en territorio de la arteria descendente

anterior.
SSS auto Cx: Score automatico en estres de defectos en territorio de la arteria circunfleja.
SSS auto CD: Score automatico en estres de defectos en territorio de la arteria coronaria derecha.
SSS auto Total: Score automatico total de defectos en estres.
SRS manual DA: Score manual en reposo de defectos en territorio de la arteria descendente anterior.
SRS manual Cx: Score manual en reposo de defectos en territorio de la arteria circunfleja.
SRS manual CD: Score manual en reposo de defectos en territorio de la arteria coronaria derecha.
SRS manual Total: Score manual total de defectos en reposo.
SDS manual DA: Score manual diferencia de defectos en territorio de la arteria descendente anterior.
SDS manual Cx: Score manual diferencia de defectos en territorio de la arteria circunfleja.
SDS manual CD: Score manual diferencia de defectos en territorio de la arteria coronaria derecha.
SDS manual Total: Score manual total de defectos por diferencia.
SDS auto DA: Score automatico diferencia de defectos en territorio de la arteria descendente
anterior.
SDS auto Cx: Score automatico diferencia de defectos en territorio de la arteria circunfleja.
SDS auto CD: Score automatico diferencia de defectos en territorio de la arteria coronaria derecha.
SDS auto Total: Score automatico total de defectos por diferencia.
NSE: Numero de segmentos afectados en estres.
NSR: Numero de segmentos afectados en reposo.
Clínica: Resultado clinico de la prueba de estres para angina de pecho.
Eléctrica: Resultado electrico de la prueba de esfuerzo.
Necrosis: Necrosis.
Isquemia: Isquemia.
Di.VIE: Dilatacion de ventriculo izquierdo al estres.
Di.VIR: Dilatacion de ventriculo izquierdo en reposo.
Di.VID: Dilatacion transitoria del ventriculo izquierdo, obtenida por diferencia.
VTDE: Volumen telediastolico estandarizado de ventriculo izquierdo en estres.
VTDR: Volumen telediastolico estandarizado de ventriculo izquierdo en reposo.
VTSE: Volumen telesistolico estandarizado de ventriculo izquierdo en estres.
VTSR: Volumen telesistolico estandarizado de ventriculo izquierdo en reposo.
FEE: Fraccion de eyeccion en estres calculada por SPECT.
FER: Fraccion de eyeccion en reposo calculada por SPECT.
FE eco: Fraccion de eyeccion calculada por ecocardiografia.
FE hemo: Fraccion de eyeccion calculada por hemodinamia.
CF: Clase funcional.

Estenosis DA: Porcentaje de estenosis en arteria descendente anterior.

Estenosis Cx: Porcentaje de estenosis en arteria circunfleja.
Estenosis CD: Porcentaje de estenosis en arteria coronaria derecha.
AI: Angina Inestable durante el seguimiento.
CRC: Necesidad de cirugia de revascularizacion coronaria durante el seguimiento.
Infarto: Infarto de Miocardio durante el seguimiento.
Vasos hemo: Numero de vasos afectados, evaluados por hemodinamia.

6.5 DESCRIPCIÓN DE ALGUNAS VARIABLES

Con el fin de facilitar la interpretacion de los resultados que se presentan en § 6.8, a continuacion se detalla la forma de obtencion y el significado de algunas de las variables que conforman el sistema multivariante que es objeto de este estudio.

La sangre es la encargada de transportar nutrientes, hormonas y gases hacia todas las celulas del cuerpo y de recoger los desechos metabolicos que alli se generan. Existen diversos factores que pueden causar una disminucion transitoria o permanente del riego sanguineo a un tejido, con la consecuente disminucion del aporte de oxigeno al mismo.

Al sufrimiento celular resultante de esta situacion se le denomina **Isquemia**.

En general, la **Isquemia** miocardica se produce a causa de un desequilibrio entre la demanda y el aporte de oxigeno al musculo cardiaco. Esto puede suceder ante dos eventos: en primer lugar, por un incremento de los requerimientos de oxigeno del miocardio (por ejemplo, al realizar ejercicio) que no vaya acompanado de un incremento del aporte de oxigeno por la sangre (en el caso de obstrucciones parciales de las arterias coronarias [28]) o, en segundo lugar, por una disminucion del aporte en condiciones de reposo, por ejemplo, cuando un trombo obstruye completamente y en forma aguda las arterias coronarias, aun cuando los requerimientos del musculo cardiaco sean minimos.

La **Isquemia** puede tener efectos reversibles o ser suficientemente grave como para causar la muerte del tejido, evento conocido como **Necrosis**, el cual es de caracter irreversible.

Existen una serie de pruebas diagnosticas que permiten caracterizar la gravedad y las particularidades de la cardiopatia isquemica. El sistema multivariante que es objeto de este estudio incluye resultados de SPECT, de coronariografias y de ecocardiografias.

El **SPECT** (***S****ingle* ***P****hoton* ***E****mission* ***C****omputed* ***T****omography*) es una tecnica que permite generar imagenes tridimensionales del corazon a partir de la deteccion de la

[28] A este evento se le denomina disminucion de la reserva vasodilatadora coronaria.

radiacion gamma emitida por un trazador radioactivo administrado previamente al paciente[29]. La comparacion de los grados de captacion miocardica bajo condiciones de estres cardiaco y de reposo permite diagnosticar condiciones de normalidad, isquemia, y/o necrosis, entre otras.

El Gated SPECT es una tecnica de punta que incorpora mejoras en relacion con el SPECT estandar tanto en la forma en que la gammacamara realiza el barrido como en los algoritmos utilizados para la reconstruccion de las imagenes, lo cual permite obtener imagenes de diferentes etapas del ciclo cardiaco, incrementando la informacion clinica que de estas puede extraerse. En la Figura 6.1 se muestra algunas imagenes tipicas de este tipo de pruebas.

Figura 6.1. Imagenes tipicas de un Gated SPECT.

[29] Suele utilizarse Tc-99m MIBI, Talio 201 y Tc-99m Tetrafosmin.

Como resultado del procesamiento del SPECT aparecen multiples indicadores

relacionados con la perfusion miocardica. Estos indicadores pueden ser generados de
forma automatica por el sistema o pueden surgir del analisis de las imagenes por parte
del especialista. En esta memoria, al primer conjunto de resultados los hemos
denominado **automáticos** y al segundo conjunto **manuales**.
Entre los resultados del SPECT, destacan los denominados con el nombre generico de
Scores, los cuales son sumas de puntuaciones de defectos en una region irrigada por un
vaso concreto, con lo cual reflejan conjuntamente la intensidad y la extension de los
defectos. Su medicion puede realizarse en estres (*Summed Stress Score:* **SSS**), en reposo
(*Summed Rest Score:* **SRS**) o pueden obtenerse como la diferencia entre el puntaje en
estres y el puntaje en reposo (*Summed Difference Score:* **SDS = SSS – SRS**). El estres
cardiaco puede alcanzarse mediante la realizacion de esfuerzo fisico en una cinta sinfin,
mediante la aplicacion de farmacos como el dipiridamol y la dobutamina o combinando
ambos metodos.
Las arterias coronarias, que rodean el miocardio formando una corona, son las
encargadas de su irrigacion. La **arteria descendente anterior** (DA), que es la principal
arteria coronaria, irriga la cara anterior y la punta del ventriculo izquierdo, asi como el
tabique interventricular; la **arteria coronaria derecha** (CD) es la responsable de irrigar
el ventriculo derecho y la region inferior y posterior del ventriculo izquierdo; la **arteria
circunfleja** (Cx) irriga la cara lateral del ventriculo izquierdo.
Esta diferenciacion genera el concepto de territorios irrigados por cada una de las
arterias coronarias. No obstante, es importante anotar que se trata de un concepto
esquematico y relativamente amplio, dada la gran variabilidad en la delimitacion de los

territorios irrigados por cada rama coronaria entre pacientes y la existencia de
circulacion cruzada entre territorios.[30]

[30] Como ejemplo de esta variabilidad, esta el origen de la arteria descendente posterior, la cual irriga la
cara diafragmatica del ventriculo izquierdo. El origen de esta arteria define el concepto de "dominancia
coronaria". Aproximadamente en un 70 % de los pacientes, se origina en la CD (dominancia derecha); en
un 30 %, aproximadamente, en la Cx (dominancia izquierda); en algunos pocos casos puede originarse en
ambas (dominancia balanceada).

Teniendo en cuenta el territorio analizado (de arteria descendente anterior, de arteria
coronaria derecha o de arteria circunfleja), las condiciones de la prueba (estres o reposo)
y la forma en que se genera la lectura (automatica o manual), surgen diferentes
puntuaciones o scores que evaluan conjuntamente la extension y la intensidad de los
defectos en dicho territorio. Asi, por ejemplo, la variable **SSS auto Cx** indica la
extension e intensidad de defectos en territorio de la arteria circunfleja, evaluado de
manera automatica bajo condiciones de estres cardiaco. Los scores totales son la suma
de los defectos en los tres territorios considerados.

Como resultado del SPECT, tambien se calculan los volumenes telesistolico y
telediastolico del ventriculo izquierdo, en condiciones de estres o de reposo (**VTSE**,
VTDE, **VTSR**, **VTDR**). Estos son los volumenes que presenta el ventriculo izquierdo
hacia el final de la sistole (contraccion) y de la diastole (relajacion), respectivamente.

La evaluacion de los anteriores volumenes da lugar al concepto de dilatacion del
ventriculo izquierdo en condiciones de estres o de reposo (**Di.VIE**, **Di.VIR**). Asimismo,
surge el concepto de dilatacion transitoria del ventriculo izquierdo, obtenida por

diferencia entre la dilatacion en estres y la dilatacion en reposo (**Di.VID**).
Como resultado del SPECT tambien puede obtenerse el numero de segmentos afectados
en estres o en reposo (**NSE**, **NSR**), como indicador general de afeccion cardiaca sin
diferenciar territorios.
La **fracción de eyección** es el predictor mas importante de morbimortalidad de origen
cardiaco. Esta variable expresa en terminos porcentuales la disminucion del volumen
del ventriculo izquierdo en sistole (contraccion), con respecto a la diastole
(relajacion).[31] Asi, por ejemplo, una fraccion de eyeccion del 50 % significa que el
volumen del ventriculo izquierdo en maxima contraccion (al final de la sistole) es la
mitad de su volumen en maxima relajacion (al final de la diastole). Un corazon con poca
capacidad contractil (fraccion de eyeccion baja) tiene poca capacidad de bombeo. En
estos casos se habla de insuficiencia cardiaca. La **fracción de eyección** es calculada de
[31]

$$FE\ x = \frac{Vol\ VI\ Diastole - Vol\ VI\ Sistole}{Vol\ VI\ Diastole} \cdot 100$$

forma automatica por el SPECT, bien sea en condiciones de estres o de reposo (**FEE**,
FER).
Adicionalmente al SPECT, en ocasiones se requieren pruebas de tipo invasivo o
hemodinamicas, conocidas con el nombre generico de **coronariografías** o
cateterismos

cardiacos. Estas pruebas se realizan mediante la insercion de un cateter a traves de la
arteria femoral (region inguinal) o radial (brazo) hasta las arterias coronarias, en las que
se inyecta un medio de contraste que facilita la visualizacion de imagenes en las que se
pueden apreciar posibles estrechamientos del paso de sangre u oclusiones completas.
A partir de los cateterismos tambien puede evaluarse la fraccion de eyeccion. En este
caso, se denomina fraccion de eyeccion calculada por hemodinamia (**FE hemo**). La
fraccion de eyeccion tambien puede obtenerse de forma incruenta usando ultrasonidos,
mediante ecocardiografia (**FE eco**).
La coronarigrafia tambien permite determinar el grado de estenosis o estrechamiento de
las arterias coronarias (**Estenosis DA**, **Estenosis Cx**, **Estenosis CD**) y el numero de
vasos afectados (**Vasos hemo**). Esta ultima variable contabiliza el numero de vasos con
algun grado de estenosis, sin calificar la intensidad de la misma.
**Esta gama de pruebas, de condiciones en las que pueden ser realizadas y los
resultados que pueden obtenerse de cada una de ellas explican la cantidad de
información faltante en la base de datos, pues no todos los pacientes son sometidos
al mismo tipo de pruebas ni bajo las mismas condiciones.**

6.6 CUANTIFICACIÓN
El sistema multivariante que es objeto de este estudio en este capitulo es de tipo mixto,
puesto que esta conformado por variables con diferentes escalas de medicion, v. gr.,
Numerica, Ordinal y Nominal (cf. § 1.2).
Nuestra propuesta de analisis consiste en generar cuantificaciones optimas con base en
los desarrollos del Capitulo 5. Seguidamente, aprovechando las propiedades metricas de

las cuantificaciones asi obtenidas, construiremos una representacion conjunta de los
pacientes y las variables, utilizando un metodo lineal como es el de las representaciones
Biplot (GABRIEL, 1971).
Para realizar el proceso de cuantificacion, todas las variables, sin importar su nivel de
escalamiento, deben someterse a un proceso de categorizacion o discretizacion
(cf. § 5.3.1). A continuacion se presenta la categorizacion utilizada para las variables
que conforman el sistema multivariante objeto del presente estudio:

BRI: No, Si.
CI: No, Si.
EA: Asintomatico, Sintomatico.
DM: No, Si.
HTA: No, Si.
Edad: <30, 30-55, 56-80, >80.
IMC: <25 (Normal), 25-30 (Sobrepeso), >30 (Obesidad).
Sexo: Femenino, Varon.
SSS manual DA: <4, 4-8, 9-13, >13.
SSS manual Cx: <4, 4-8, 9-13, >13.
SSS manual CD: <4, 4-8, 9-13, >13.
SSS manual Total: <4, 4-8, 9-13, >13.
SSS auto DA: <25, 25-49,99, 50-75, >75.
SSS auto Cx: <25, 25-49,99, 50-75, >75.
SSS auto CD: <25, 25-49,99, 50-75, >75.
SSS auto Total: <25, 25-49,99, 50-75, >75.
SRS manual DA: <4, 4-8, 9-13, >13.
SRS manual Cx: <4, 4-8, 9-13, >13.
SRS manual CD: <4, 4-8, 9-13, >13.
SRS manual Total: <4, 4-8, 9-13, >13.

SDS manual DA: <4, 4-8, 9-13, >13.
SDS manual Cx: <4, 4-8, 9-13, >13.
SDS manual CD: <4, 4-8, 9-13, >13.
SDS manual Total: <4, 4-8, ≥9.
SDS auto DA: <25, 25-49,99, 50-75, >75.
SDS auto Cx: <25, 25-49,99, 50-75, >75.
SDS auto CD: <25, 25-49,99, 50-75, >75.
SDS auto Total: <25, 25-49,99, 50-75, >75.
NSE: 0, 1, 2, 3, 4, 5, 6-9, 10-13, ≥14.
NSR: 0, 1, 2, 3, 4, 5, 6-9, 10-13, ≥14.
Clínica: Negativa, Positiva, Dudosa.
Eléctrica: Negativa, Positiva, Dudosa, BRI previo, no valorable, elevacion ST.
Necrosis: No, Si.

Isquemia: No, Si.
Di.VIE: No, Si.
Di.VIR: No, Si.
Di.VID: No, Si.
VTDE: <30, 30-59,99, 60-80, >80.
VTDR: <30, 30-59,99, 60-80, >80.
VTSE: <30, 30-59,99, 60-80, >80.
VTSR: <30, 30-59,99, 60-80, >80.
FEE: <20, 21-40, 41-50, >50
FER: <20, 21-40, 41-50, >50
FE eco: <20, 21-40, 41-50, >50
FE hemo: <20, 21-40, 41-50, >50
CF: I, II, III, IV.
Estenosis DA: <50, 50-70, 70,01-99,99, 100
Estenosis Cx: <50, 50-70, 70,01-99,99, 100
Estenosis CD: <50, 50-70, 70,01-99,99, 100
AI: No, Si.
CRC: No, Si.
Infarto: No, Si.
Vasos hemo: 0, 1, 2, 3.

En § 2.5 hemos establecido un paralelo entre la escala de medicion de una variable y su nivel de escalamiento. Tal y como se indica al final de dicho apartado, la utilizacion de un nivel de escalamiento mas flexible que aquel que corresponderia naturalmente a la escala de medicion de una variable puede resultar util para evaluar posibles relaciones

no lineales de una variable con otras variables del sistema. Con base en tal consideracion y teniendo en cuenta lo restrictivo que resulta el nivel de escalamiento Numerico, en todos los casos hemos utilizado los niveles de escalamiento Ordinal o Nominal. Las variables **Sexo**, **Clínica** y **Eléctrica**, se escalaron a nivel Nominal; las demas variables se escalaron a nivel Ordinal.

El nivel de escalamiento utilizado para las variables binarias es irrelevante, obteniendose siempre las mismas cuantificaciones o cuantificaciones con signo

contrario, cualquiera que sea el nivel de escalamiento elegido. La unica razon por la que
hemos elegido el nivel de escalamiento Ordinal en tales casos, en lugar del Nominal, es
para asegurarnos de obtener cuantificaciones negativas para las categorias de ausencia
(No, Asintomatico) y positivas para las categorias de presencia (Si, Sintomatico), lo
cual facilitara las interpretaciones en las representaciones Biplot subsiguientes.
En la Figura 6.2 se muestra el aspecto de un sector de la base de datos preparada para su
procesamiento mediante la aplicacion **CUANTIFICA**.

Figura 6.2. Preparacion de la base de datos para su procesamiento mediante la
aplicacion **CUANTIFICA**.

En la primera fila de la base de datos aparece el nombre de la variable; en la segunda
fila, se indica el nivel de escalamiento de cada variable, usando para el efecto numeros
entre 1 y 5, correspondientes a los niveles de escalamiento Numerico, Ordinal,
Nominal, Numerico Flotante y Ordinal Flotante (cf. § 5.3.1); en la tercera fila, se
incluye un vector de filtrado, el cual consta en este caso de unos, indicando que todas
las variables se incluyen en el proceso de cuantificacion optima (cf. § 5.3.1). Entre la

cuarta y la ultima fila se disponen las etiquetas de cada una de las categorias, las cuales
deben estar conformadas por numeros enteros que respeten la ordinalidad de las
categorias representadas, cuando se haya elegido el nivel de escalamiento Ordinal, tal y
como se aprecia al comparar los paneles de la Figura 6.2.
Como resultado del escalamiento optimo a traves de la aplicacion **CUANTIFICA**, se
genera la Tabla de Cuantificaciones, la Matriz de Cuantificaciones, asi como un grafico

de cuantificaciones para cada una de las variables.
No obstante que en este capitulo, la cuantificacion no es el fin ultimo de nuestro
analisis, sino que constituye un paso de adaptacion del sistema multivariante para su
posterior analisis con base en tecnicas lineales, resaltaremos algunos aspectos
relacionados con los resultados obtenidos al procesar la base de datos objeto de este
estudio mediante la aplicacion **CUANTIFICA**. Para una resena general sobre los
resultados que genera la aplicacion, vease § 5.3.7.
Vale la pena resaltar que en solo 2 de de los 50 conjuntos de cuantificaciones generados
para las variables escaladas a nivel Ordinal se obtuvieron empates: en **SRS manual Cx**
y en **SDR manual Cx**. En ambos casos se obtuvo la misma cuantificacion para los
niveles *'9-13'* y *'>13'*, tal y como se ilustra en la Figura 6.3 para la primera de estas
variables.
El hecho de obtener la misma cuantificacion para dos categorias adyacentes de una
variable escalada a nivel Ordinal significa, en general, que no es posible diferenciar el
comportamiento de las dos categorias en cuestion, en cuanto a sus relaciones con las
categorias de las demas variables, bajo las restricciones impuestas por el escalamiento
Ordinal. En particular, en este caso, podemos afirmar que, en la muestra analizada, no
se observo diferencia entre el grupo de pacientes con **SRS manual Cx** entre 9 y 13 y el
grupo de pacientes con **SRS manual Cx** mayor de 13, en lo relativo a su forma de
relacionarse con las demas variables. Desde luego, lo mismo puede decirse en referencia
a la variable **SDS manual Cx**.

Figura 6.3. Grafico de cuantificaciones para "Score manual en reposo de defectos en territorio de la arteria circunfleja".

Si bien el hecho de obtener la misma cuantificacion en categorias adyacentes de variables escaladas a nivel Ordinal no representa en si un problema, vale la pena reflexionar al respecto cuando se presente esta situacion, pues ello puede ser reflejo de diferentes circunstancias. Por una parte, podria estar indicando una sobrecategorizacion de la variable en cuestion, tratando de diferenciar niveles que, para fines practicos, tienen el mismo comportamiento. Por otra parte podria estar indicando que la relacion entre los niveles de la variable no es Ordinal, o al menos que no sigue el orden esperado.

En el caso de sobrecategorizacion de una variable no se requiere tomar ninguna medida adicional. Dependiendo de cual sea el objetivo final del investigador puede ser suficiente con tomar nota de ello e indicarlo en el informe. Si el analista lo desea, puede colapsar los niveles que reciben una misma cuantificacion en un solo nivel, pero ello no alteraria los resultados.

Si se trata de una variable para la cual la ordinalidad de sus categorias no sea un hecho claramente establecido, pudiendo, quiza, no existir, podria replantearse el nivel de escalamiento de la misma, usando el escalamiento Nominal, con base en el cual se

aplicarian cuantificaciones libres de restricciones que pudieran reflejar mejor las relaciones de las categorias de dicha variable con las categorias de las demas variables.

En el presente caso, en el que la ordinalidad de las categorias de las dos variables en

cuestion es un hecho indiscutible, no es necesario realizar ninguna accion adicional, no
requiriendose ninguna modificacion para utilizar estos resultados en la siguiente etapa.
Para satisfacer el objetivo final del presente capitulo, consistente en analizar las
principales relaciones entre individuos, variables y variables e individuos a traves de
representaciones Biplot, la Matriz de Cuantificaciones constituye el resultado mas
importante de la aplicacion **CUANTIFICA**. Esta matriz tiene la misma estructura que la
matriz original, es decir, 6.965 filas y 53 columnas, sin contabilizar, desde luego, la fila
que contiene los nombres de las variables, ni las filas con los codigos de nivel de
escalamiento y filtrado. A diferencia de la matriz de entrada, que tiene un aspecto como
el que se muestra en el panel derecho de la Figura 6.2, los valores numericos contenidos
en la Matriz de Cuantificaciones tienen propiedades metricas, lo cual permite su uso
como informacion de entrada de cualquier tecnica lineal. En particular, en este capitulo,
utilizaremos la Matriz de Cuantificaciones como informacion de entrada para la
obtencion de representaciones Biplot.
En la Figura 6.4 presentamos un esquema de las diferentes adaptaciones y
transformaciones a las que se somete la base de datos, desde su estado original hasta
llegar a la Matriz de Cuantificaciones, la cual constituye el resultado final de la
aplicacion **CUANTIFICA** y que se utiliza como informacion de entrada para las
representaciones Biplot.
En el panel (a) de la Figura 6.4 se presenta un sector hipotetico de la base de datos
original, dejando vacias las celdas cuando no se haya registrado la informacion de una
variable para un paciente especifico. El panel (b) corresponde a la base de datos

categorizada, en la cual se remplaza cada uno de los valores originales por la
denominacion de la categoria a la cual pertenece (cf. § 1.5). En el panel (c) se remplaza
cada una de las denominaciones de las diferentes categorias por una etiqueta numerica,
respetando la ordinalidad cuando sea del caso (cf. § 5.3.1). En el panel (d) se muestra el

correspondiente sector de la Matriz de Cuantificaciones generada por la aplicacion
CUANTIFICA.

Figura 6.4. Adaptaciones y transformaciones de la base de datos: (a) base de datos
original; (b) base de datos categorizada; (c) base de datos etiquetada; (d) Matriz de
Cuantificaciones.

Para fines practicos no se requiere construir la base de datos correspondiente al panel
(b) de la Figura 6.4. Normalmente se pasa directamente de la base de datos original
(panel (a)) a la base de datos categorizada y adaptada para lectura mediante la
aplicacion **CUANTIFICA** (panel (c)), siendo suficiente con tomar nota de la denominacion
(a): BD Original
(b): BD Categorizada
(c): BD Etiquetada
(d): M. Cuantificaciones

CATEGORIZACIÓN
ETIQUETADO
CUANTIFICACIÓN

correspondiente a cada categoria, para fines de la posterior interpretacion y descripcion
de resultados.

Notese que en los diferentes pasos se trabaja con la informacion disponible,
manteniendo inalterada la informacion faltante, la cual sigue estando ausente en la

Matriz de Cuantificaciones. Aunque hemos recomendado utilizar el codigo numerico -99 como marcador de las celdas faltantes en la base de datos que es leida por
la aplicacion **CUANTIFICA** (cf. § 5.3.1), en la Figura 6.4 hemos dejado tales celdas en
blanco para facilitar la visualizacion de su correspondencia con las celdas de los paneles
adyacentes.
Insistimos en el hecho de que la Matriz de Cuantificaciones no solo tiene la misma
estructura que la matriz de entrada en lo relativo al numero de filas y numero de
columnas, sino tambien en lo que tiene que ver con informacion faltante. Es decir que,
para el presente caso, la Matriz de Cuantificaciones carece del mismo 53,12 % de
informacion que la matriz inicial.

**Puesto que las representaciones Biplot, con base en las cuales pretendemos realizar
nuestro análisis final, exigen matrices completas, será necesario realizar una adaptación que cubra el evento de matrices con información faltante. El siguiente
apartado constituye nuestra aportación al respecto.**

6.7 CONSTRUCCIÓN DE LA REPRESENTACIÓN BIPLOT

6.7.1 BASE TEÓRICA.

La representacion Biplot de una matriz X, propuesta por GABRIEL (1971), se basa en la
obtencion de marcadores vectoriales a para las filas de X, y b para sus columnas, de
manera que el producto escalar de los marcadores de la i-esima fila y la j-esima
columna, $a_i' b_j$, reproduzca el elemento x_{ij} de la matriz X. Los marcadores vectoriales
permiten, asimismo, la **representación simultánea de las filas** (en este caso pacientes)
y las columnas (generalmente, asi como en este caso, variables) de la matriz X en un
subespacio de dimension reducida; de ahi, la particula '*Bi*' que encabeza el nombre de
la representacion.

Para obtener la representacion completa de una matriz de rango r se requeriria un espacio r-dimensional, lo cual no resulta practico para matrices de rango mayor de dos.
Es posible, sin embargo, obtener una representacion aproximada de cualquier matriz por medio de la representacion grafica de su correspondiente matriz de aproximacion a bajo rango.
Aunque la aproximacion de rango dos es la mas popular, por ser la que permite obtener la mejor representacion bidimensional de los principales patrones de correlacion y asociacion del sistema multivariante, no es la unica aproximacion posible. Si se usa una matriz de rango p para aproximar una matriz de rango r (con $p < r$), pueden obtenerse $p!\ 2 * (p-2)!$ representaciones Biplot. Algunas situaciones practicas pueden exigir el uso de aproximaciones de rango mayor que dos.
Los marcadores vectoriales para las filas y para las columnas de $X_{(n \times m)}$ se organizan en sendas matrices $A_{(n \times p)}$ y $B_{(m \times p)}$, siendo n el numero de filas, m el numero de columnas y p el rango de la matriz. Las filas de la matriz A contienen los marcadores vectoriales de las correspondientes filas de la matriz X; las filas de la matriz B contienen los marcadores vectoriales de las correspondientes columnas de la matriz X; luego, $AB' = X$.

Existen infinitas alternativas para elegir las matrices de factorizacion A y B con los marcadores vectoriales para las filas y las columnas de la matriz X. Las mas populares se basan en la descomposicion en valores singulares de X ($X = U\Sigma V'$), puesto que esta metodologia hace posible obtener de manera simultanea la aproximacion a bajo rango de X y los marcadores vectoriales basados en tal aproximacion.

Para obtener la aproximacion de rango p de X, basta con escoger las primeras p columnas de U (U_p, i. e., los primeros p valores propios de XX'), las primeras p filas y p columnas de Σ ($_p\Sigma$, i. e., la matriz diagonal con los primeros p valores singulares de X), y las primeras p columnas de V (V_p, i. e., los primeros p valores propios de $X'X$).

$()'$ $_{p\,p\,p\,p}X = U\,\Sigma\,V$ es la aproximacion de rango p de X (HOUSEHOLDER and YOUNG, 1938).

Para simplificar la notacion, supongamos que X representa la aproximacion a bajo rango de la matriz original, con lo cual podemos prescindir del subindice p. Las matrices A y B que contienen los marcadores vectoriales de X se obtienen a partir de alguna de las siguientes expresiones:

$A = U$; $B = V\Sigma$ (GH-Biplot o CMP-Biplot)

$A = U\Sigma$; $B = V$ (JK-Biplot o RMP-Biplot)

$A\ U$; $B\ V_1 = \Sigma_\alpha = \Sigma_{-\alpha}$, con $\alpha \in [0,1]$ (Si $\alpha = 0,5$, Biplot Simetrico)

Como consecuencia de la eleccion de los marcadores vectoriales, la representacion de las filas o de las columnas puede resultar favorecida en relacion con la representacion del otro conjunto.

El CMP-Biplot (*Column Metric Preserving Biplot*) se caracteriza por representar las columnas con la maxima calidad posible, representando las filas en forma estandar. En el RMP-Biplot (*Row Metric Preserving Biplot*), las filas aparecen con maxima calidad

de representacion, quedando las columnas representadas en forma estandar. Es posible elegir cualquier configuracion intermedia entre las dos anteriores, utilizando un valor de α entre cero y uno. Si $\alpha = 0$, se obtiene el CMP-Biplot; si $\alpha = 1$, se obtiene el RMPBiplot; si $\alpha = 0,5$ se tiene un Biplot Simetrico, en el que las filas y las columnas

aparecen con igual calidad de representacion.
En adicion a estas opciones clasicas, existe una forma de elegir los marcadores tal que
se maximice la calidad de representacion tanto para filas como para columnas. Dicha
solucion, propuesta por GALINDO (1985, 1986), consiste en elegir $A = U\Sigma$ y $B = V\Sigma$.
Notese que esta solucion es equivalente a tomar los marcadores columna, o H, del
GH-Biplot, en el cual las columnas gozan de la maxima calidad de representacion, y los
marcadores fila, o J, del JK-Biplot, en el cual las filas tienen la maxima calidad de
representacion. Por esta razon, a la representacion asi obtenida se le denomina
HJ-Biplot o RCMP-Biplot (*Row Column Metric Preserving Biplot*). Como contraparte a
la maxima calidad de representacion para filas y columnas brindada por esta
representacion, se sacrifica la propiedad de que el producto interno del i-esimo
marcador fila y el j-esimo marcador columna reproduzca el elemento de la celda x_{ij}, el
cual solo seria estimado en este caso. Notese que, bajo esta seleccion de marcadores,
$AB' = U\Sigma\Sigma V' = U\Sigma_2 V' \approx X$.
En el presente caso, donde no existen grupos naturales de pacientes, diferenciandose
estos solamente por los valores de las variables, estamos particularmente interesados en
representar la estructura de correlaciones de las variables con la maxima calidad
posible. Debe tenerse en cuenta, sin embargo, el ingrediente adicional que aparece al
intentar construir una representacion Biplot para una matriz con datos faltantes, lo cual
no puede hacerse mediante la aplicacion directa de ninguna de las expresiones clasicas
para la obtencion de marcadores, pues todas ellas estan basadas en la descomposicion en

valores singulares de la matriz, la cual no puede realizarse para matrices con datos
faltantes.
Una opcion viable en otras circunstancias consistiria en utilizar algun metodo para
imputar la informacion faltante, tras lo cual podrian obtenerse los marcadores
vectoriales mediante la descomposicion en valores singulares. No obstante, tal y como

fue senalado anteriormente (cf. § 6.4), en el presente caso, la informacion faltante no es
informacion perdida, sino informacion que nunca fue medida debido a que al paciente
no se le realizo esa prueba particular. Luego, cualquier intento por estimar esa
informacion que nunca ha existido no haria mas que introducir ruido.
Nuestra propuesta para generar la representacion Biplot de una matriz X con datos
faltantes —mas no perdidos— consiste en obtener los marcadores columna de una
hipotetica matriz $*X$ con estructura de correlaciones igual a la de la matriz X. Una vez se
tengan tales marcadores columna, se usan regresiones para la obtencion de los
marcadores fila de la matriz X.[32]
Para el efecto, se parte del calculo de las correlaciones por pares entre las variables de la
matriz incompleta, X. La correlacion de cada par de variables se calcula con base en
toda la informacion disponible para ese par, sin importar si en esa fila hay informacion
disponible o no para las demas variables. Considerese la informacion de la Tabla 6.1.

Observación	Variable 1	Variable 2	Variable 3
1	X_{11}	.	X_{13}
2	.	.	.
3	.	X_{32}	X_{33}
4	X_{41}	X_{42}	X_{43}
5	X_{51}	X_{52}	.

6 X_{61} . X_{63}

7 X_{71} X_{72} X_{73}

8 . . X_{83}

Tabla 6.1. Matriz con informacion faltante.

32 En adelante, en este apartado, denotaremos por X a la matriz de cuantificaciones generada por la aplicacion **CUANTIFICA** al realizar un proceso de cuantificacion optima sobre un sistema multivariante con datos faltantes. Si bien, esta matriz es un eslabon intermedio del proceso de analisis que se desarrolla en este capitulo, en este apartado nos referiremos a ella como la matriz de datos original.

Asi, en el calculo de r_{12} participarian las observaciones 4, 5 y 7, en r_{13}, las observaciones 1, 4, 6 y 7, y en r_{23}, las observaciones 3, 4 y 7.

Todos los pares de correlaciones calculados se organizan en una matriz de correlaciones, R, asi:

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1m} \\ r_{21} & 1 & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & 1 \end{pmatrix}$$

Esta matriz de correlaciones, R, es igual a $^*X' \, ^*X$, donde $^*X$ es una hipotetica matriz completa (sin datos faltantes), centrada y estandarizada, que tiene el mismo tamano y la misma estructura de correlaciones que la matriz incompleta original, X. Por tanto, los marcadores columna de las matrices $^*X$ y X coincidiran. Tales marcadores columna se obtienen con base en la descomposicion en valores y vectores propios de $^*X' \, ^*X$

$\left(X_*' X_* = V\Lambda V' \right)$.
En consecuencia con lo indicado anteriormente acerca de nuestro interes por generar
una representacion Biplot en la que las variables tengan maxima calidad de
representacion, y teniendo en cuenta que la descomposicion en valores singulares de la
hipotetica matriz $_*X$ tiene la forma:
$X_* = U\Lambda^{\frac{1}{2}} V'$, elegimos la matriz de marcadores
columna como
$B = V\Lambda^{\frac{1}{2}}$, eleccion correspondiente a los marcadores columna de un
CMP-Biplot. Notese que Λ es la matriz diagonal con los valores propios de la matriz
$X_*' X_*$, mientras que
$\Lambda^{\frac{1}{2}} \equiv \Sigma$ es la matriz diagonal con los valores singulares de $_*X$.
Puesto que la hipotetica matriz $_*X$ tiene una estructura de correlaciones igual a la de X,
los marcadores columna obtenidos para $_*X$ son igualmente validos para X. Esta eleccion
de marcadores bastaria para obtener una representacion de las columnas de la matriz X,
denominada H-Plot (GABRIEL, 1981, JOBSON, 1992). No obstante, es posible obtener
tambien marcadores fila para generar una representacion Biplot de X.

Para tal efecto, consideramos una hipotetica descomposicion en valores singulares de X,
asi
$X = U\Lambda^{\frac{1}{2}} V'$. Observese que la parte correspondiente a
$\Lambda^{\frac{1}{2}} V'$ es conocida, siendo
precisamente la que se obtuvo mediante la descomposicion en valores y vectores
propios de R. Si se conociera la matriz U, sus filas servirian de marcadores fila para las
correspondientes filas de la matriz X, lo que permitiria, en conjuncion con los

marcadores columna ya definidos, elaborar una representacion Biplot para la matriz X.

Si X fuera una matriz completa, podriamos obtener la matriz U asi:

$$X = U\Lambda^{1/2}V' = UB'$$

$$XB = UB'B$$

$$(XB)(B'B)^{-1} = UB'B(B'B)^{-1}$$

$$U = XB(B'B)^{-1}$$

Aunque los datos faltantes en X impiden el uso directo de la anterior expresion para calcular la matriz U que contendria los marcadores fila correspondientes a un CMP-Biplot, es posible estimar cada una de las filas de la matriz U de forma independiente, utilizando una ponderacion cero para los valores faltantes.

La i-esima fila de U, sin incluir aun la ponderacion cero, se calcularia asi:

$$u_i' = x_i'B(B'B)^{-1}$$

Al trasponer esta expresion surge la familiar ecuacion para la estimacion del vector de parametros β en regresion multiple:

$$u_i = (B'B)^{-1}B'x_i$$

Luego, cada marcador fila es igual al vector de parametros de una regresion multiple de la correspondiente fila de la matriz original sobre los marcadores columna. Notese, sin embargo, que a pesar de la analogia con el modelo de regresion, la anterior expresion no involucra un termino de error. Por tanto, el producto entre la matriz de marcadores

columna y un marcador fila obtenido mediante este procedimiento reproduciria exactamente la correspondiente fila de la matriz. [33]

Volviendo al hecho de que el vector x_i puede tener celdas vacias, y siguiendo la analogia con el modelo de regresion, u_i se calcula como el vector de parametros de una

regresion en la que se asigna ponderacion cero a los valores faltantes. Para tal efecto, se
utiliza una matriz diagonal W_i, con unos (1) en las posiciones correspondientes a los
valores observados y ceros (0) en las posiciones correspondientes a las celdas vacias.
La regresion ponderada tiene la siguiente forma:
$(')_1 ' _{iiii} u\ B\ WB\ B\ W\ x = -$
Dado que la incorporacion de ponderaciones nulas a la matriz $B' B$ puede generar
problemas de singularidad en la matriz ponderada $' _i B\ WB$, impidiendo la obtencion de
su inversa, es necesario agregar una pequena cantidad a su diagonal para corregir esta
situacion sin alterar significativamente los resultados. Este mecanismo es analogo al de
la regresion *ridge* y es mencionado por GABRIEL et al. (1998), quienes tambien lo
utilizan para sortear problemas de singularidad. En este trabajo obtuvimos resultados
satisfactorios sumando 1x10-9 a la diagonal de la matriz en cuestion.
La regresion ponderada para la obtencion del i-esimo marcador fila, incluyendo la
matriz diagonal K que rompe la singularidad de $' _i B\ WB$, tiene la siguiente forma:
$(')_1 ' _{iiii} u\ B\ W\ B\ K\ B\ W\ x = +-$
En resumen, la representacion Biplot que proponemos utiliza los siguientes marcadores
vectoriales:
1
$B = V\Lambda_2$; $[\ ]_1$
$_{1\,2}$, , , , , ', $(')'$, 1, , $_{in\,iiii} A\ a\ a\ a\ a\ a\ B\ WB\ K\ B\ W\ x\ i\ n = \cdots\cdots$
$= + - = \cdots$

[33] Tal producto equivale a $X\beta$ en el modelo de regresion $y = X\beta$.

Donde:
B: Matriz de marcadores columna para la matriz X.
A: Matriz de marcadores fila para la matriz X.
V: Matriz de vectores propios de la matriz R.
1

Λ_2 : Matriz diagonal de la raiz cuadrada de los valores propios de R.
a_i: Marcador fila correspondiente al i-esimo individuo.
W_i: Matriz diagonal indicadora de observaciones faltantes para el i-esimo individuo.
K: Matriz diagonal con una pequena constante ($1x10^{-9}$).
x_i: i-esima fila de la matriz original, en la que se remplazan los valores faltantes por
un valor numerico arbitrario.
La representacion Biplot obtenida con base en estos marcadores tiene las siguientes
propiedades:
1) La longitud de los vectores que representan a las variables estiman la desviacion
estandar de las correspondientes variables.
2) Los cosenos de los angulos entre los vectores que representan a las variables,
aproximan la correlacion entre las correspondientes variables representadas.
3) La proyeccion del i-esimo marcador fila sobre el j-esimo marcador columna estima
el elemento x_{ij} de la matriz X.
Es importante resaltar que esta propuesta se ha desarrollado teniendo en cuenta la
particularidad de la presente matriz, esto es, que hay datos faltantes, mas no perdidos.
En tal sentido, hemos centrado nuestro interes en reproducir lo mas fielmente posible la
estructura de correlaciones observada, asi como los datos observados. La utilizacion de
los marcadores vectoriales propuestos satisface ambos objetivos.
El hecho de que las correlaciones originales solo sean estimadas en la representacion
Biplot en lugar de reproducidas no es producto de la eleccion de marcadores, sino de la

reduccion de dimensionalidad. Notese que los marcadores columna son equivalente a
los de un CMP-Biplot para una matriz completa. En tal sentido, las relaciones entre

variables se reproducen con la maxima calidad posible para dicho subespacio de
representacion.
De igual forma, el hecho de que los valores observados de la matriz original solamente
sean estimados en lugar de ser reproducidos, tampoco es consecuencia de la forma en
que se calculan los marcadores, sino de la reduccion de dimensionalidad. El producto
interno entre los marcadores fila y columna sin reduccion de dimensionalidad
reproducen casi exactamente el dato observado. Tal 'reproduccion' solo presenta una
muy leve discrepancia con el dato original, a causa de la constante K (1x10-9) sumada a
la matriz $'_{i}BWB$ para sortear el problema de la singularidad.
Es importante anotar que el producto interno de un marcador fila y un marcador
columna, o equivalentemente la proyeccion de un marcador fila sobre un marcador
columna, solamente debe usarse para estimar los valores observados en X, no teniendo
ningun sentido utilizarlo como estimacion de valores que nunca han existido.
Es notable la similitud entre nuestra propuesta de representacion para una matriz
incompleta y el CMP-Biplot de GABRIEL (1971). Los marcadores columna utilizados
en nuestra propuesta de representacion son equivalentes a los del CMP-Biplot de una
matriz completa, las propiedades de la representacion que proponemos son analogas a
las del CMP-Biplot y la aplicacion de nuestra propuesta a una matriz sin datos faltantes
genera un CMP-Biplot. No obstante, no podemos decir, en sentido estricto, que nuestra
representacion sea un CMP-Biplot, al menos no si consideramos que este es un termino
ya acunado para referirse a la representacion Biplot de matrices completas.

Consecuentemente, utilizaremos las siglas **GCMP-Biplot (*Generalized Column Metric Preserving Biplot*)** para referirnos a nuestra propuesta de representacion. Con esta denominacion damos el merito correspondiente a la representacion que sirve de base a nuestra propuesta, a la vez que enfatizamos que no se trata de la representacion Biplot de una matriz completa, sino de una representacion generalizada para matrices con datos faltantes.

6.7.2 RUTINA MARCADORES GCMP-BIPLOT.

Hemos elaborado una rutina computacional en MATLAB, que sistematiza el proceso de generacion de los marcadores GCMP-Biplot descritos en § 6.7.1. Tras copiar todos los componentes de la rutina en la ruta de trabajo de MATLAB y digitar "MGCMPBInicio", aparece una interfaz grafica como la que se muestra en la Figura 6.5.

Figura 6.5. Interfaz grafica de la rutina **MARCADORES GCMP-BIPLOT**.

La informacion de entrada para esta aplicacion consiste en una matriz de datos incompletos, que bien puede estar encabezada en la primera fila con etiquetas para las variables o puede estar conformada unicamente por la parte numerica. Puede utilizarse el codigo numerico -99 para identificar las celdas con informacion faltante o bien pueden dejarse vacias. La matriz de datos se leera desde un archivo Excel.

Dado que esta rutina se ha elaborado como parte integral del proceso de analisis propuesto en este trabajo, la informacion de entrada se buscara por defecto en una hoja llamada MC dentro del archivo de Excel especificado, es decir, en la hoja que contenga

la matriz de cuantificaciones generada por la rutina **CUANTIFICA** (cf. § 5.3.7). No obstante, el usuario pude eliminar la marca de verificacion de la casilla *Los datos estan en la hoja "MC"*, con lo cual aparecera una ventana adicional que permitira escoger la hoja que contenga la matriz de entrada.

Tras procesar la matriz con datos faltantes, la rutina genera tres matrices, las cuales son copiadas en sendas hojas de calculo en el mismo archivo Excel que contiene la informacion de entrada, asi: F: Matriz de marcadores fila; C: Matriz de marcadores columna; D: Matriz diagonal con los valores propios de R (cf. § 6.7.1). Aunque las matrices F y C, que contienen los marcadores fila y columna, respectivamente, son suficientes para elaborar la representacion GCMP-Biplot, en las salidas se incluyen tambien los valores propios de R, los cuales son utilizados en el calculo de las calidades de representacion.

En ocasiones, MATLAB puede ser incapaz de importar la matriz por insuficiente memoria. Esta situacion puede resolverse trabajando sobre una copia del archivo Excel en la que se hayan eliminado todas las hojas innecesarias, dejando unicamente la hoja que contiene los datos de entrada necesarios para este modulo, esto es, MC.

El archivo Excel debera estar cerrado para permitir la escritura de las hojas con la informacion de salida. Si el programa se ejecuta en varias ocasiones, las hojas F, C y D seran sobrescritas.

6.7.3 RUTINA REPRESENTACIONES GCMP-BIPLOT.

A partir de los marcadores GCMP-Biplot generados con base en la rutina descrita en § 6.7.2, es posible construir representaciones GCMP-Biplot para diferentes dimensiones y con diferentes especificaciones. Para tal efecto, se ha elaborado la rutina **REPRESENTACIONES GCMP-BIPLOT** en MATLAB. Su instalacion se realiza

copiando todos sus componentes en la ruta de trabajo de MATLAB. Para su ejecucion, se
digita "RGCMPBInicio", con lo cual aparece una interfaz grafica como la que se
muestra en la Figura 6.6.

Figura 6.6. Interfaz grafica de la rutina **REPRESENTACIONES GCMP-BIPLOT**.

Al presionar el boton *! Leer Archivo de Excel !*, se abre una ventana que permite
escoger el archivo que contiene la informacion necesaria para la elaboracion de la
Representacion GCMP-Biplot, no siendo necesario que el mismo se encuentre localizado en la ruta de trabajo de MATLAB.

Los datos de entrada del programa estan conformados por las tres hojas generadas por el
modulo **MARCADORES GCMP-BIPLOT**, es decir, los Marcadores Fila, los Marcadores Columna y la Matriz Diagonal con los valores propios de la matriz de
correlaciones basada en la matriz incompleta. Esta informacion debe estar dispuesta en

sendas hojas del archivo Excel, nombradas F, C y D, respectivamente, tal y como son
generadas por la rutina **MARCADORES GCMP-BIPLOT**.

Aunque el programa tambien necesita leer la informacion de las etiquetas de las
variables y de las filas, en caso de que esta ultima exista, el modulo ha sido disenado
para darle continuidad al proceso, evitando la intervencion del usuario siempre que ello
sea posible. En tal sentido, el modulo **REPRESENTACIONES GCMP-BIPLOT** lee
las etiquetas de la primera hoja del archivo (la ubicada mas hacia la izquierda), es decir,
la que se utilizo como informacion de entrada del modulo **CUANTIFICA**.

Luego, aunque no se requiere ninguna modificacion por parte del usuario, es necesario

que la hoja con los datos utilizados en la cuantificacion este presente y que ocupe la primera posicion. En caso de que el modulo **CUANTIFICA** haya generado una matriz de datos reducida (cf. § 5.3.2), esta debera moverse a la primera posicion. Asimismo, el programa utiliza la informacion contenida en la matriz de datos para la conformacion de grupos de individuos por niveles de una variable, cuando se activa la opcion "*Pintar por niveles de variable...*" que aparece en la ventana de opciones de la representacion (Figura 6.7).

No esta de mas insistir en el hecho de que los procesos han sido automatizados de tal forma que se logre minimizar la intervencion del usuario. Luego, en caso de que el modulo **CUANTIFICA** no haya creado una matriz de datos reducida, no hace falta realizar ninguna modificacion a la base de datos; no siendo necesario ni siquiera retirar las filas correspondientes al nivel de escalamiento y al filtrado, en caso de que tal vector exista.

Asi, pues, el archivo de entrada que sera procesado por la rutina **REPRESENTACIONES GCMP-BIPLOT** debera contar con 4 hojas obligatorias: 1) Datos utilizados para la Cuantificacion, con los nombres de las variables en la primera fila. Esta informacion tiene que estar ubicada en la primera hoja del archivo. 2) F: Matriz de marcadores fila. 3) C: Matriz de marcadores columna. 4) D: Matriz diagonal con los valores propios de R. Solamente la hoja con los datos originales podra tener

datos faltantes; las matrices en las hojas F, C y D tienen que estar completas, tal y como las genera el modulo **MARCADORES GCMP-BIPLOT**.

Una vez leido el archivo Excel que contiene la informacion basica para construir la

representacion GCMP-Biplot, aparece una ventana como la de la Figura 6.7, en la que
pueden precisarse las opciones particulares de la representacion.

Figura 6.7. Interfaz grafica para las opciones de la **REPRESENTACIÓN GCMP-BIPLOT**.

DIMENSIONES

Dentro de las posibles dimensiones de representacion (cualquier dimension menor o
igual que el rango de la matriz completa), se elige la de la abcisa y la de la ordenada.

PINTAR POR NIVELES DE UNA VARIABLE

Esta opcion permite diferenciar a los individuos por colores con base en los niveles de
una variable que se elige en una tercera ventana. Las representaciones obtenidas tienen
el aspecto que se aprecia en las Figuras 6.9 y 6.10, en el apartado de resultados (§ 6.8)

RESCALAMIENTO DE COORDENADAS

Consiste en multiplicar los marcadores fila por una constante, dividiendo los
marcadores columna por la misma constante, lo cual tiene el efecto de expandir
(constante mayor que 1) o contraer (constante menor que 1) la informacion representada, lo que optimiza la visualizacion conjunta de los marcadores fila y los
marcadores columna. Este artilugio no altera las propiedades de la representacion
GCMP-Biplot.

GRAFICAR ETIQUETAS DE FILAS

Esta opcion hace que cada individuo aparezca identificado con una etiqueta en la
representacion GCMP-Biplot. Es posible seleccionar esta opcion incluso si no se
suministraron etiquetas para los individuos; en tal caso, el programa las genera
automaticamente.

GRAFICAR ETIQUETAS DE COLUMNAS

Al activar esta opcion, cada variable aparece identificada con su nombre en la
representacion GCMP-Biplot.

CALCULAR CONTRIBUCIONES

Se calculan los siguientes indicadores:
1) Calidad de Representacion Global.
2) Contribuciones Relativas del Elemento Fila al Factor.
3) Contribuciones Relativas del Elemento Columna al Factor.
4) Contribucion Relativa del Factor al Elemento Fila (Calidad de Representacion
de los Individuos).
5) Contribucion Relativa del Factor al Elemento Columna (Calidad de Representacion de las Variables)
Cada conjunto de indicadores se almacena en una hoja en el mismo archivo de entrada.

6.8 RESULTADOS

Tras generar la Matriz de Cuantificaciones, a partir del procedimiento descrito en § 6.6
y obtener sus correspondientes marcadores GCMP-Biplot con base en el procedimiento
descrito en § 6.7.1 y § 6.7.2, se construyen representaciones GCMP-Biplot para los
planos 1-2 y 3-4, mediante el modulo descrito en § 6.7.3. En la Figura 6.8 se muestra el
GCMP-Biplot original del primer plano factorial, incluyendo todas las variables y todas
las observaciones.

Figura 6.8. Representacion GCMP-Biplot del primer plano factorial.

En el Biplot de la Figura 6.8, cada paciente aparece representado por un punto y cada
variable por una linea que va del centro hacia la periferia. Luego, la representacion en

cuestion contiene 6.965 puntos y 53 lineas. El primer eje factorial recoge el 25,88 % de
la varianza total, y el segundo eje, el 9,33 %. Luego, el plano tiene una calidad global de
representacion de 35,21 %.

Para el tipo de marcadores vectoriales elegidos (cf. § 6.7), la longitud al cuadrado de cada linea aproxima la varianza de la variable representada. No obstante, puesto que cada variable ha sido centrada y estandarizada (varianza uno), el cuadrado de la longitud de una linea puede interpretarse como calidad de representacion de la variable representada en el correspondiente plano. Por tanto, las variables cuyas correspondientes lineas exhiban una longitud muy corta en un plano determinado, no se encuentran bien representadas en el mismo, por lo que se retiran, permitiendo la mejor visualizacion de las variables que gozan de una aceptable calidad de representacion.

El coseno del angulo formado por dos lineas proporciona una estimacion de la correlacion entre las dos variables representadas. Luego, un angulo muy pequeno entre dos variables se interpreta como una alta correlacion positiva entre tales variables.

Analogamente, los angulos obtusos se interpretan como correlacion negativa, siendo esta de mayor magnitud, cuanto mayor sea el angulo. Un angulo llano o de 180 grados estima una correlacion de -1.

El producto escalar entre un marcador fila y un marcador columna proporciona una estimacion de la correspondiente celda de la matriz (cf. § 6.7). En la representacion grafica, esto puede visualizarse como la proyeccion ortogonal de un punto sobre una linea, teniendo en cuenta que es posible prolongar una linea tanto como sea necesario en cualquier direccion para realizar dicha proyeccion. En general, se estima que los pacientes cuyas proyecciones ortogonales estan cerca del final de la linea (sitio donde aparece el nombre de cada variable) tienen altos valores de la variable en cuestion y viceversa.

Al simplificar la representacion Biplot de la Figura 6.8, no mostrando las variables con
baja calidad de representacion en dicho plano, se observan mejor los patrones de
correlacion entre las variables con buena calidad de representacion.

Tambien es posible visualizar lo bien representada que esta una variable en un plano,
pintando de diferentes colores los puntos correspondientes a cada una de sus diferentes
categorias. Si la variable no se encuentra bien representada, la separacion entre los
individuos que conforman las diferentes categorias no sera muy clara. Por el contrario,
una clara separacion entre categorias indica una buena calidad de representacion de la
variable en dicho plano.
La Figura 6.9 es una version modificada del Biplot de la Figura 6.8, en la que se han
ocultado las variables con baja calidad de representacion, se han eliminado los nombres
de las variables y se han pintando los puntos que representan a los pacientes, acorde con
la variable **Necrosis**: verde para ausencia de Necrosis; rojo para presencia de Necrosis.

Figura 6.9. GCMP-Biplot del primer plano factorial con pacientes discriminados por
Necrosis.

La marcacion de los pacientes se realiza con base en el registro de la variable **Necrosis**
en la base de datos original. Este es el motivo por el cual la Figura 6.9 tiene menor
densidad que la Figura 6.8 en cuanto a numero de pacientes se refiere, pues mientras en
la Figura 6.8 aparecen los 6.965 pacientes, en la Figura 6.9 solo aparecen aquellos
pacientes para los cuales se conocia su lectura de la variable **Necrosis**.

En la Figura 6.10 se realiza una separacion analoga a la de la Figura 6.9, pero en este caso, con base en la variable **Isquemia**, resaltando con verde la ausencia de Isquemia y con rojo la presencia de Isquemia.

Figura 6.10. GCMP-Biplot del primer plano factorial con pacientes discriminados por Isquemia.

Con base en el analisis de las Figuras 6.9 y 6.10, podemos afirmar que **el primer plano factorial de la representación GCMP-Biplot está configurado alrededor de las variables Necrosis e Isquemia**. Esta es una afirmacion mas clinica que estadistica,

pues si bien es cierto que al analizar las contribuciones relativas de los elementos columna a cada uno de los factores, pueden encontrarse otras variables con mayor efecto relativo en la conformacion de los ejes, todas las demas variables con alta calidad de representacion en el plano 1-2 son indices que de una u otra forma reflejan aspectos puntuales de una o ambas de las dos condiciones fundamentales que caracterizan la cardiopatia isquemica: **Necrosis** e **Isquemia**. Puede decirse en cierto modo que en la poblacion de pacientes que conforma la base de datos analizada, donde la mayoria de variables son resultados de SPECT, la **Necrosis** y la **Isquemia** son los factores basicos, mientras que las otras variables son la expresion de tales factores.

La Figura 6.11 es una version modificada de la representacion GCMP-Biplot de la Figura 6.8, en la que no se muestran los marcadores fila (por lo cual, tecnicamente hablando, se trata de un H-plot; cf. § 6.7.1) y se resaltan las principales relaciones entre las variables con alta calidad de representacion en el plano 1-2.

Figura 6.11. H-plot de las principales relaciones entre las variables en el plano 1-2.

En la Figura 6.11 se han utilizado colores para marcar grupos de variables, lo cual facilita su diferenciacion. Asi, se han resaltado con rojo las dos variables alrededor de las cuales gira la discusion, esto es, **Necrosis** e **Isquemia**, y ademas se ha agregado una cabeza de flecha en la direccion que apunta hacia sus correspondientes presencias. Se ha utilizado tambien el color rojo para marcar el grupo de vectores que representan las fracciones de eyeccion, medidas en diferentes condiciones y por diferentes metodos. Se han pintado con violeta los volumenes telediastolico y telesistolico del ventriculo izquierdo. Se ha usado el azul encendido para senalar las diferentes medidas de dilatacion del ventriculo izquierdo, asi como el numero de vasos afectados, evaluados por hemodinamia. Se han pintado con verde las diferentes componentes del score manual en reposo, de rosa las componentes de los scores en estres, y de cian los scores diferencia. Se han senalado con marron el numero de segmentos afectados, tanto en estres como en reposo.

En primer lugar, se destaca la conformacion de grupos de variables, algunas asociadas principalmente con **Necrosis**, otras con **Isquemia** y algunas otras que se asocian con ambas. Entre las variables asociadas principalmente con **Necrosis** estan las fracciones de eyeccion, los volumenes del ventriculo izquierdo asi como su dilatacion en estres y en reposo, el numero de segmentos afectados en reposo y los scores en reposo.

Las variables asociadas principalmente con **Isquemia** son los scores diferencia, y la dilatacion del ventriculo izquierdo por diferencia.

Aportando informacion sobre ambos factores estan los scores en estres, el numero de
segmentos afectados en estres, asi como el numero de vasos afectados.
Se aprecia una estrecha relacion entre **Necrosis** y los scores en reposo (**SRS**, pintados
de verde). Con el objetivo de mantener el grafico despejado, unicamente hemos
etiquetado el score manual total de defectos en reposo (**SRS manual Total**), el cual
como puede observarse esta cerca de la bisectriz del angulo formado por los dos
marcadores mas externos del score en reposo, lo que es de esperarse por ser este la suma
de los scores de cada una de las arterias.

Como habiamos indicado anteriormente, la **Necrosis** es una condicion permanente, la
cual no se asocia con un desequilibrio puntual entre la demanda y el aporte del oxigeno
en el momento de la medicion. Este es el motivo de que los scores en reposo (**SRS**) sean
los mas apropiados para evidenciar esta situacion. Lo mismo puede decirse del numero
de segmentos afectados en reposo (**NSR**), los cuales, como puede observarse, se
encuentran estrechamente asociados tanto con los scores en reposo (**SRS**) como con la
Necrosis.

La Figura 6.11 muestra que la dilatacion del ventriculo izquierdo es una manifestacion
que se asocia frecuentemente con **Necrosis**. En particular, es llamativa la fuerte relacion
directa entre **Necrosis** y dilatacion del ventriculo izquierdo en reposo (**Di.VIR**), asi
como la relacion directa de los volumenes telediastolicos y telesistolicos del ventriculo
izquierdo (**VTDE**, **VTDR**, **VTSE**, **VTSR**) con las anteriores variables.
La dilatacion del ventriculo izquierdo trae como consecuencia una disminucion en la

capacidad de bombeo del corazon. Esta capacidad de bombeo es medida por la fraccion
de eyeccion. Esto explica la fuerte correlacion negativa observada entre las diferentes
medidas de la fraccion de eyeccion y los parametros de dilatacion del ventriculo
izquierdo (**Di.VIR**, **VTSE**, **VTDE**, **VTDR** y **VTSR**).
Mas alla de esta relacion general, vale la pena resaltar que la fraccion de eyeccion en
reposo (**FER**) esta particularmente asociada con los volumenes telediastolico y
telesistolico del ventriculo izquierdo en reposo (**VTDR** y **VTSR**), exhibiendo una
correlacion negativa perfecta con la primera de estas variables. Asimismo, la fraccion de
eyeccion en estres (**FEE**) esta particularmente asociada con los volumenes telesistolico
y telediastolico del ventriculo izquierdo en estres (**VTSE** y **VTDE**).
En cuanto a la **Isquemia**, se destaca su estrecha relacion con los scores diferencia, tanto
manuales como automaticos (**SDS manual**, **SDS auto**), asi como con la dilatacion
transitoria del ventriculo izquierdo, obtenida por diferencia (**Di.VID**).
Tal y como anotamos anteriormente, en estado de reposo se reflejan basicamente las
afecciones permanentes producto de la **Necrosis**. En los resultados obtenidos por
diferencia entre el estres y el reposo (**SDS** y **Di.VID**) se esta sustrayendo el efecto de la

Necrosis, aislando, por tanto, el efecto de la **Isquemia**. Esto explica la estrecha relacion
observada entre la **Isquemia**, los scores diferencia (**SDS**) y la dilatacion del ventriculo
izquierdo por diferencia (**Di.VID**).
Como es de esperarse, los scores totales, tanto manuales como automaticos (**SDS
manual Total**, **SDS auto Total**), tambien se localizan en este caso cerca de la bisectriz

de las dos lineas con mayor angulo interno, que representan los scores diferencia en
territorios especificos de cada una de las arterias coronarias.
Para analizar las variables que exhiben aproximadamente la misma correlacion con
Necrosis que con **Isquemia**, debe considerarse que el desempeno cardiaco en
condiciones de estres se ve afectado tanto por lesiones permanentes del miocardio
(**Necrosis**), como por condiciones puntuales propias del desequilibrio entre la demanda
y el aporte de oxigeno (**Isquemia**). Esto explica el hecho de que tanto el numero de
segmentos afectados en estres (**NSE**) como los scores en estres (**SSS manual** y **SSS
auto**) aparezcan como indicadores compromiso entre **Necrosis** e **Isquemia**.
La dilatacion del ventriculo izquierdo en estres (**Di.VIE**), a pesar de estar logicamente
asociada con la dilatacion permanente del ventriculo izquierdo (**Di.VIR**), incluye una
componente adicional que es el estres cardiaco, lo que la aproxima a las variables
medidas en condiciones de estres.
Puesto que el numero de vasos afectados, medido por hemodinamia (**Vasos hemo**) solo
contabiliza el numero de vasos con algun tipo de estenosis o estrechamiento, sin
calificar la severidad del mismo, aparece tambien como una variable compromiso entre
Necrosis e **Isquemia**. Cuando el estrechamiento es total, muy probablemente hay
Necrosis; en estrechamientos parciales, puede haber solamente **Isquemia**; y en ausencia
de vasos afectados, lo mas probable es que no haya **Necrosis** ni **Isquemia**.
En la Figura 6.12 se muestra el plano 3-4 de la representacion H-plot, el cual recoge el
10 % de la variabilidad total. En este plano se diferencian los territorios irrigados por
cada una de las tres arterias coronarias. Puede observarse, por ejemplo, una correlacion

positiva entre todos los indices o resultados correspondientes al territorio de la arteria
descendente anterior. Se observa tambien como dentro de cada territorio, las pruebas

manuales tienden a estar correlacionadas entre si al igual que las pruebas automaticas lo
estan entre ellas.

Figura 6.12. H-plot de las principales relaciones entre variables en el plano 3-4.

6.9 SÍNTESIS DE LA PROPUESTA DE ANÁLISIS

Los resultados ilustrados mediante la aplicacion practica de este capitulo, donde se han
logrado separar los principales patrones de asociacion de un sistema multivariante mixto
de gran tamano, con mas de la mitad de datos faltantes, nos permiten cerrar este trabajo
expresando nuestra satisfaccion por el desempeno general de todas las tecnicas aqui
desarrolladas.

Es importante resaltar que el hecho de que las cuantificaciones generadas mediante la
aplicacion **CUANTIFICA** sean unicas y no esten basadas en una solucion multidimensional -como si lo estan las generadas por las tecnicas del sistema Gifi-
no implica una reduccion de la dimensionalidad del sistema cuantificado. La Matriz de
Cuantificaciones no solo tiene propiedades metricas, sino que conserva todas las
relaciones multidimensionales existentes en la base de datos original, las cuales pueden
recuperarse y resumirse mediante alguna tecnica lineal de reduccion de la dimensionalidad, tal y como hemos ilustrado en este capitulo, mediante la obtencion de
representaciones GCMP-Biplot de la Matriz de Cuantificaciones.

Los modulos desarrollados como parte de este trabajo conforman una herramienta

integral para el analisis de bases de datos multivariantes con informacion faltante. Es
posible realizar un analisis analogo al presentado en este capitulo, con base en los
siguientes pasos:

4) **Cuantificación**. Modulo **CUANTIFICA** (Capitulo 5).
5) **Obtención de Marcadores GCMP-Biplot**. Modulo **MARCADORES GCMP-BIPLOT**. § 6.7.1 y § 6.7.2.
6) **Generación de Representaciones GCMP-Biplot**. Modulo **REPRESENTACIONES GCMP-BIPLOT**. § 6.7.3

Con el animo de estimular el uso de las aplicaciones desarrolladas, como una
herramienta integral para el analisis de bases de datos multivariantes mixtas con datos
faltantes (o completas), se ha dispuesto una interfaz grafica para unificar el acceso a los

diferentes modulos (Figura 6.13). Para ingresar a la misma, se digita "*Modulos*"
(Atencion: "Modulos" sin tilde).

Figura 6.13. Interfaz de integracion de modulos.

Notese que la mayor parte de los procesos han sido automatizados, por lo cual, tras
realizar las adecuaciones necesarias para la obtencion de las cuantificaciones (cf. § 6.6,
Figura 6.4), no hace falta realizar mas modificaciones "manuales" sobre los datos,
siendo posible tener toda la informacion en un solo archivo de Excel, el cual sirve tanto
de archivo de entrada como de archivo de salida para los diferentes modulos. La Figura
6.14 resume el proceso.

Figura 6.14. Resumen de los modulos de procesamiento de datos multivariantes
mixtos.

En resumen, los pasos para el analisis de una base de datos multivariante mixta con

informacion faltante, los modulos computacionales utilizados en cada uno de ellos, asi
como la informacion de entrada y salida, son:
1) Adecuar la base de datos para su cuantificacion. (§ 5.3.1; § 6.6, Figura 6.4). En el
esquema de la Figura 6.14, a los datos de entrada, usados para la cuantificacion,
los hemos denominado X.
2) Realizar la cuantificacion optima de la base de datos (X), mediante el modulo
CUANTIFICA. Como resultados, se generan la Tabla de Cuantificaciones (TC) y
la Matriz de Cuantificaciones (MC).
F, C, D
MC
CRG, CREFF, CRECF,
CRFEF, CRFEC
X, F, C, D
X
TC, MC
F, C, D
CRG, CREFF, CRECF, CRFEF, CRFEC
X
TC, MC

3) Usar la Matriz de Cuantificaciones (MC) como informacion de entrada del
modulo **MARCADORES GCMP-BIPLOT** para generar los marcadores fila (F), los marcadores columna (C) y la matriz diagonal con los vectores propios de
la matriz de correlaciones de MC (D).
4) Usar la base de datos original (X), y los marcadores GCMP-Biplot (F, C y D)
como informacion de entrada del modulo **REPRESENTACIONES GCMP-BIPLOT**, el cual, ademas de las representaciones, genera los indices de
contribuciones y calidad de representacion (CRG, CREFF, CRECF, CRFEF, CRFEC).
Los nombres de las salidas de un modulo corresponden con los nombres de las entradas
del siguiente, de manera que pueda realizarse un proceso continuo, sin necesidad de que

el usuario tenga que renombrar ninguna de las hojas del archivo, ni especificar en cada
paso en cuales hojas se encuentra la informacion. El unico caso en el que se requiere
intervencion del usuario, es cuando se genera una base de datos reducida (cf. § 5.3.2 y
§ 6.7.3). En tales casos, la hoja *X red* debera moverse a la primera posicion del archivo
antes de ejecutar el modulo **REPRESENTACIONES GCMP-BIPLOT**.
Desde el punto de vista de la continuidad del proceso, habria sido posible omitir el
modulo **MARCADORES GCMP-BIPLOT**, incorporandolo en el modulo
REPRESENTACIONES GCMP-BIPLOT, haciendo que el calculo de F, C y D, fuera
un proceso interno de este ultimo modulo. No obstante, hemos detectado que al trabajar
con grandes bases de datos como la utilizada en este capitulo, el calculo de los
marcadores GCMP-Biplot puede ser un proceso altamente demandante de recursos
computacionales, que exige un tiempo de procesamiento considerable. En tal sentido,
hemos encontrado mas practico calcular los marcadores GCMP-Biplot en una sola
ocasion, dejandolos disponibles en el archivo de datos para la posterior obtencion de
representaciones GCMP-Biplot.
Finalmente, es importante resaltar que, si bien una de las fortalezas del sistema
propuesto es el manejo de matrices con informacion faltante, todos los procesos
funcionan igualmente bien cuando se cuenta con informacion completa. Al aplicar todo

el proceso descrito en este apartado sobre una base de datos completa, se obtiene una
representacion CMP-Biplot (cf. § 6.7.1).
Asimismo, cuando se parte de bases de datos completas, la Matriz de Cuantificaciones

(MC) generada mediante la aplicacion **CUANTIFICA** sera, a su vez, completa, ademas de
numerica. Por tanto, es posible utilizarla para obtener cualquier otra representacion
Biplot o como informacion de entrada de algun otro procedimiento lineal, acorde con
los objetivos.

6.10 COMPARACIÓN DE RESULTADOS CON LOS OBTENIDOS MEDIANTE TÉCNICAS DEL SISTEMA GIFI

Antes de presentar las conclusiones de este trabajo, resulta obligatoria la comparacion
entre las tecnicas de analisis aqui propuestas y las tecnicas de analisis enmarcadas en el
sistema Gifi de analisis multivariante no lineal. En § 5.4 hemos realizado una
comparacion general entre **CUANTIFICA** y los tres principales modulos de escalamiento optimo del sistema Gifi de analisis multivariante no lineal, i. e.,
PRINCALS, HOMALS y OVERALS, cuando son usados como primer paso para la generacion de cuantificaciones.
En este apartado, realizaremos una comparacion particular entre las tecnicas usadas para
el analisis de la aplicacion practica de este capitulo (cuantificacion optima, usando
CUANTIFICA + representaciones GCMP-Biplot) y las tecnicas del sistema Gifi cuando
son usadas para generar representaciones graficas en subespacios de baja dimensionalidad. Para tal efecto, hemos elegido PRINCALS, pues tal y como se indica
en § 5.4, es la tecnica mas similar a **CUANTIFICA**, difiriendo de ella solamente en los
aspectos que le confieren superioridad a **CUANTIFICA** frente a las tecnicas del sistema
Gifi.
Puesto que en el analisis realizado en este capitulo, los principales patrones de
asociacion se manifiestan en el plano 1-2 de la representacion GCMP-Biplot, hemos

elegido inicialmente la solucion PRINCALS con p=2 para fines de tal comparacion
(Figura 6.15).
Antes de analizar los resultados graficos de PRINCALS, es pertinente aclarar que
aunque la representacion de la Figura 6.15 es denominada biplot en las salidas de SPSS,
y de hecho lo es en el sentido general de la palabra, en cuanto representa conjuntamente
observaciones y variables, no se trata de un biplot construido con base en la teoria de
GABRIEL (1971) (cf. § 6.7.1), es decir que no se trata de la representacion grafica
conjunta de los marcadores fila y los marcadores columna de una matriz, obtenidos a
partir de la descomposicion factorial de su aproximacion a bajo rango.

El biplot que se presenta como resultado de las tecnicas del sistema Gifi es una
representacion en la que se superponen las puntuaciones de los individuos, resultado de
minimizar la correspondiente funcion de perdida (cf. § 2.3 y 2.4), y los vectores de
saturaciones de las variables, debidamente reescalados.

Figura 6.15. Plano biplot (puntuaciones + saturaciones) correspondiente a la solucion
PRINCALS, con p=2.

Puesto que la representacion original que se muestra en la Figura 6.15 no resulta
adecuada para extraer informacion de la misma, hemos utilizado las coordenadas para
construir una representacion equivalente usando MATLAB (Figura 6.16 (a)). Desde
luego, la baja calidad de las representaciones graficas no es un defecto que deba
atribuirsele a las tecnicas del sistema Gifi, sino al programa en el que estan
implementadas.

Figura 6.16. (a) Plano biplot (puntuaciones + saturaciones) correspondiente a la solucion PRINCALS, con p=2 y (b) representacion GCMP-Biplot de la matriz de cuantificaciones generada mediante **CUANTIFICA**.

b

a

La Figura 6.16 permite contrastar el biplot (puntuaciones + saturaciones) resultante de la solucion PRINCALS con p=2 y el GCMP-Biplot obtenido mediante las tecnicas obtenidas en este trabajo (cf. Figura 6.8). Al comparar los componentes (a) y (b) de la Figura 6.16, resulta evidente que la estructura generada como resultado del analisis PRINCALS con p=2 no coincide con la generada mediante nuestra propuesta. Esto puede verse con mayor claridad en la Figura 6.17, version modificada de la Figura 6.16, en la cual hemos omitido los puntos que representan a los pacientes y hemos resaltando las relaciones mas sobresalientes entre variables.

La Figura 6.17 (a) permite apreciar que en la solucion PRINCALS, casi todas las variables tienen su principal carga sobre la primera dimension, mientras que las variables hipertension arterial sistemica (**HTA**), cardiopatia isquemica en el diagnostico inicial (**CI**) y **Sexo** son las unicas con altas cargas en la segunda dimension.

Excepto por la correlacion negativa entre las fracciones de eyeccion (**FEE**, **FER**, **FE hemo** y **FE eco**) y las demas variables, no se observa ninguno de los demas patrones detectados y analizados al usar las tecnicas propuestas en este trabajo. En particular, no se observa la diferenciacion entre **Necrosis** e **Isquemia**, ni sus relaciones con los

correspondientes resultados del SPECT, principal patron de la Figura 6.17 (b) y
alrededor del cual giro nuestro analisis (cf. § 6.8).
Los principales patrones de asociacion exhibidos en la Figura 6.17 (a) pueden resumirse
de la siguiente forma. La primera dimension esta dominada por la correlacion negativa
entre las medidas de fraccion de eyeccion (**FEE**, **FER**, **FE hemo** y **FE eco**) y todos los
demas resultados de SPECT, ecocardiografia y coronariografia que de una u otra forma
miden la gravedad de la cardiopatia isquemica. La segunda dimension esta dominada
por las relaciones entre **Sexo**, hipertension arterial sistemica (**HTA**) y cardiopatia
isquemica en el diagnostico inicial (**CI**), indicando que la prevalencia de hipertension
arterial es mayor en mujeres que en hombres y que la cardiopatia isquemica en el
diagnostico inicial es mas frecuente en pacientes con hipertension arterial sistemica que
en pacientes con tension arterial normal.
Estos resultados, si bien son correctos, son relativamente evidentes y faciles de predecir
por un cardiologo sin que se requiera para ello un analisis multivariante. A partir de esta

solucion PRINCALS no es posible detectar los matices y detalles que exhibe nuestra
propuesta y que hemos discutido ampliamente en el apartado de resultados de este
capitulo (cf. § 6.8).

Figura 6.17. (a) Saturaciones correspondiente a la solucion PRINCALS, con p=2 y (b)
H-plot de la matriz de cuantificaciones generada mediante **CUANTIFICA**.

b

a

Tal y como hemos anotado a lo largo de toda la memoria, las soluciones del sistema Gifi cambian en funcion de la dimensionalidad elegida (cf. Capitulo 3). Resulta, por tanto, logico evaluar otras soluciones PRINCALS de mayor dimensionalidad. Para el efecto, obtuvimos adicionalmente las soluciones PRINCALS desde p=3 hasta p=10. Se observo que, si bien, las soluciones efectivamente variaban, tal cambio no modificaba sustancialmente la estructura presentada las Figuras 6.16 (a) y 6.17 (a). Asimismo, revisamos otros planos mas alla del 1-2 en busqueda de los patrones que encontramos en nuestro analisis, pero estos no aparecieron.

La Figura 6.18 resulta suficiente para ilustrar esta busqueda. Alli se presenta el plano 3-4 del biplot (puntuaciones + saturaciones) correspondiente a la solucion PRINCALS con p=10. Como puede observarse, las relaciones exhibidas no tienen nada que ver con las encontradas mediante el uso de nuestra propuesta, y algunas de estas relaciones son incluso incoherentes, tal como la correlacion positiva entre la fraccion de eyeccion (**FE hemo**), el numero de vasos afectados (**Vasos hemo**) y la estenosis de las arterias coronarias, que aparece reflejada en la parte inferior derecha de la representacion.

Figura 6.18. Plano 3-4 del biplot (puntuaciones + saturaciones) correspondiente a la solucion PRINCALS, con p=10.

Finalmente, aunque se sabe que, por construccion, las cuantificaciones generadas mediante **CUANTIFICA** minimizan las *Interdistancias*, esto es, el criterio escalado de la suma sobre todas las observaciones de las distancias cuadraticas entre las categorias

cuantificadas de las diferentes variables (cf. expresion [4.7] y § 5.2), cerraremos esta
comparacion mostrando que tal condicion se satisface para este ejemplo especifico. En
la Figura 6.19 se muestran las *Interdistancias* calculadas para la solucion
CUANTIFICA, asi como las correspondientes a las soluciones PRINCALS desde p=1
hasta p=10. Como era de esperarse, las *Interdistancias* para la solucion **CUANTIFICA**
son menores que cualquiera de las otras *Interdistancias*, lo que significa que la
cuantificacion de **CUANTIFICA** es la que asigna valores mas similares a las categorias
de las distintas variables que se asocian con mayor frecuencia.

Figura 6.19. *Interdistancias* para **CUANTIFICA** y para soluciones PRINCALS desde
p=1 hasta p=10.

En vista de los resultados de las anteriores comparaciones, y teniendo en cuenta la
sustentacion teorica presentada, asi como la existencia de un soporte informatico para
nuestra propuesta, nos permitimos presentarla como alternativa a las tecnicas del
sistema Gifi para el analisis de sistemas multivariantes mixtos.

ConclusioneConclusiones

1. En este trabajo, hemos desarrollado un criterio de cuantificacion optima, el cual
se basa en la asignacion de valores mas similares a las categorias de las distintas
variables que se asocien con mayor frecuencia. Tales cuantificaciones optimas
se obtienen al hacer minima la suma sobre todas las observaciones de las distancias cuadraticas entre las categorias cuantificadas de las diferentes
variables.
2. Dado que las cuantificaciones generadas mediante el criterio de escalamiento

optimo propio de las tecnicas del Sistema Gifi no son unicas, sino que, por el
contrario, dependen de la dimensionalidad de la solucion elegida, debe utilizarse
algun criterio que permita elegir el mejor conjunto de cuantificaciones. Para tal
efecto, hemos utilizado el criterio de cuantificacion optima desarrollado en este
trabajo. La rutina computacional **INTERDIS** automatiza dicha eleccion.

3. Se ha elaborado una rutina computacional llamada **CUANTIFICA**, la cual implementa el criterio de cuantificacion optima desarrollado en este trabajo y
automatiza su obtencion. Mediante esta aplicacion es posible asignar cuantificaciones optimas a un sistema multivariante, tomando en consideracion
las restricciones propias del nivel de escalamiento de cada variable y sin que se
requieran las cuantificaciones de ningun otro sistema como punto de partida. En
adicion a los niveles de escalamiento Numerico, Ordinal y Nominal, tradicionalmente usados en las tecnicas del Sistema Gifi, en este trabajo incorporamos los niveles de escalamiento Numerico Flotante y Ordinal Flotante.

4. El hecho de que las cuantificaciones generadas mediante la aplicacion **CUANTIFICA** sean unicas y no esten basadas en una solucion multidimensional
-como si lo estan las generadas por las tecnicas del sistema Gifi- no implica
una reduccion de la dimensionalidad del sistema cuantificado. La matriz de
cuantificaciones no solo tiene propiedades metricas, sino que conserva todas las
relaciones multidimensionales existentes en el sistema original, las cuales
pueden recuperarse y resumirse mediante alguna tecnica lineal de reduccion de
dimensionalidad.

5. Se ha desarrollado una propuesta para generar la representacion Biplot de una

matriz incompleta, denominada GCMP-Biplot. Inicialmente se obtienen los marcadores columna de una matriz completa con estructura de correlaciones igual a la de la matriz objeto; a partir de estos marcadores y de la matriz original,
se despejan los marcadores fila utilizando ponderaciones cero para los datos
faltantes. Asimismo, se han desarrollado los algoritmos computacionales que
facilitan la automatizacion de dicho proceso.

6. La tecnica de cuantificacion desarrollada en este trabajo, implementada en la
rutina computacional **CUANTIFICA**, en combinacion con la propuesta de representaciones GCMP-Biplot, cuyos desarrollos teorico y algoritmico tambien
constituyen aportaciones de este trabajo, conforman una herramienta integral
que resulta efectiva para analizar grandes sistemas multivariantes mixtos con
informacion faltante y, desde luego, tambien para sistemas multivariantes completos. En particular, como aplicacion practica, en este trabajo se analizo un
sistema multivariante conformado por 6.965 pacientes con cardiopatia isquemica
o sintomas de la misma y 53 variables, con 53,12 % de informacion faltante. Las
tecnicas propuestas permitieron detectar los principales patrones de asociacion.

7. El primer plano principal de la representacion GCMP-Biplot, para la matriz de
cuantificaciones incompleta generada mediante la aplicacion **CUANTIFICA**, permitio reconocer que las principales asociaciones en la poblacion estudiada
giran en torno a las variables Necrosis e Isquemia, como factores basicos de la
cardiopatia isquemica. Se pudo establecer que la Necrosis, como defecto permanente, se asocia particularmente con las pruebas realizadas en reposo, asi
como con las diversas mediciones de dilatacion, volumen y fraccion de eyeccion
del ventriculo izquierdo. La Isquemia, como defecto transitorio, se asocia

especialmente con los *scores* diferencia. Los resultados de las pruebas realizadas
bajo condiciones de estres cardiaco, constituyen un indice general de los defectos de Necrosis e Isquemia.

Conclusiones

201

8. Los patrones develados mediante el uso de las tecnicas propuestas en este
trabajo no fueron detectados al realizar un analisis analogo mediante las tecnicas
del sistema Gifi de analisis multivariante no lineal, en particular, al evaluar
soluciones PRINCALS con dimensionalidades desde p=2 hasta p=10.

BibliografíBibliografía

Bibliografía

203

BARTLETT, M. S. (1933). On the theory of statistical regression. *Proc. Roy. Soc. Edinburgh*, 53:260-283.

BENZECRI, J. P. (1977). Sur l'analyse des tableaux binaires associes a une
correspondance multiple. *Les cahiers de l'analyse des donnees*, 2:55-71.

BENZECRI, J. P. et al. (1973). L'analyse des donnees : L'analyse des correspondances.
Paris: Dunod.

BLASIUS, J. and GREENACRE, M. (2006). Correspondence analysis and related methods in practice. **In:** Michael Greenacre and Jorg Blasius (Ed.), *Multiple
correspondence analysis and related methods*. London: Chapman & Hall.
p. 3-40.

BOCK, R. D. (1960). Methods and applications of optimal scaling. Psychometric
Laboratory Report No 25, University of North Carolina.

BRAUNWALD, E.; ZIPES, D. P. and LIBBY, P. (Eds.) (1997). Braunwald's Heart
disease: A textbook of Cardiovascular Medicine. 5th ed. Philadelphia: WB Saunders Company. 1943 p.

BURT, C. (1950). The factorial analysis of qualitative data. *British Journal of
Psychology*, 3:166-185.

BURT, C. (1953). Scale analysis and factor analysis. *British Journal of Statistical*
Psychology, 6:5-23.
CATTELL, R. B. (1966). The meaning and strategic use of factor analysis. **In:** R. B.
Cattell (Ed.), *Handbook of multivariate experimental psychology*. Chicago: Rand McNally. p. 174-243.
CAZES, P. (1990). Codage d' une variable continue en vue de l' analyse des
correspondances. *Rev. Statistique Appliquee*. 38(3):35-51.

CORREA, G. (2006). Caracterizacion multivariante no lineal de la cardiopatia
isquemica. Memoria Interna, Departamento de Estadistica, Universidad de Salamanca, Espana.
CUADRAS, C. M. (1991). Metodos de analisis multivariante. 2a ed. Barcelona:
Universidad de Barcelona. 644 p.
De LEEUW, J. (1973). Canonical analysis of categorical data. Leiden (The Netherlands): University of Leiden.
De LEEUW, J. (1984). The Gifi-system of non-linear multivariate analysis. **In:** E.
Diday, M. Jambu, L. Lebart, J. Pages, and R. Tomassone (Ed.), *Data Analysis*
and Informatics, Vol. III. Amsterdam (North-Holland): Diday et al.
De LEEUW, J. (1994). Block relaxation methods in statistics. **In:** H. H. Bock, W.
Lenski and M. M. Richter (Ed.), *Information Systems and Data Analysis*. Berlin:
Springer. p. 308-325.
De LEEUW, J. (2006). Nonlinear principal component analysis and related techniques.
In: Michael Greenacre and Jorg Blasius (Ed.), *Multiple correspondence analysis*
and related methods. London: Chapman & Hall. p. 107-133.
De LEEUW, J.; YOUNG, F. W. and TAKANE, Y. (1976). Additive structures in qualitative data: an alternating least squares method with optimal scaling
features. *Psychometrika*, 41(4):471-503.
DIEGO DOMINGUEZ, M.; RUANO, R.; RAMIREZ, V. H.; SANTOS RODRIGUEZ,

I.; MARTIN de ARRIBA, A.; GARCIA TALAVERA, J. R.; MARTIN LUENGO, C. (2005). Stress-Induced Systolic Left Ventricular Dysfunction and Coronary Artery Disease Severity in Patients with Pattern of Ischemia in Gated-Myocardial SPECT. **En:** Congreso Europeo de Cardiologia, Estocolomo (Suecia).

DIEGO DOMINGUEZ, M.; RAMIREZ, V. H.; RUANO, R.; SANTOS RODRIGUEZ, I.; MARTIN de ARRIBA, A.; Del CAMPO, F; GARCIA TALAVERA, P.; MARTIN LUENGO, C. (2006). Valor predictivo del comportamiento de la fraccion de eyeccion por Gated-SPECT para eventos y gravedad de la enfermedad coronaria. **En:** XII reunion de Cardiologia Nuclear y Actualizacion del Diagnostico por Imagen. VII Simposio Iberoamericano de Cardiologia Nuclear, Madrid.

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179-188.

FISHER, R. A. (1940). The precision of discriminant functions. *Annals of Eugenics*, 10:422-429.

FISHER, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53(284):789-798.

GALINDO, M. P. (1985). Contribuciones a la representacion simultanea de datos multidimensionales. Tesis doctoral. Universidad de Salamanca.

GALINDO, M. P. (1986). Una alternativa de representacion simultanea: HJ-biplot. *Questio*, 10(1):13-23.

GABRIEL, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58:453-467.

GABRIEL, K. R. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis. **In:** V. Barnett (Ed.). Interpreting multivariate data. London: Wiley. p. 147-173.

GABRIEL, R. K.; GALINDO, M. P.; VICENTE-VILLARDON, J. L. (1998). Use of

biplots to diagnose independence models in three-way contingency tables. **In:** Jorg Blasius and Michael Greenacre (Ed.), *Visualization of categorical data*. San Diego: Academic Press. p. 391-404.

GANONG, W. F. (2005). Fisiologia Medica. 20 ed. Mexico, D. F.: Manual Moderno. 900 p.

GIFI, A. (1981). Nonlinear multivariate analysis. Leiden (The Netherlands): University of Leiden.

GIFI, A. (1990). Nonlinear multivariate analysis. Chichester (England): John Wiley & Sons. 579 p.

GOWER, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53:325-338.

GOWER, J. C. and HAND, D. J. (1996). Biplots. London: Chapman & Hall. 277 p.

GREENACRE, M. J. (1984). Theory and applications of correspondence analysis. Orlando (Florida): Academic Press. 364 p.

GREENACRE, M. (1994). Correspondence analysis and its interpretation. **In:** Michael Greenacre and Jorg Blasius (Ed.), *Correspondence analysis in the social sciences*. London: Academic Press. p. 3-22.

GREENACRE, M. (2007). Correspondence analysis in practice. Boca Raton (Florida): Chapman & Hall. 280 p.

GUTTMAN, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. **In:** P. Horst (Ed.), *The Prediction of Personal Adjustment*. New York: Social Science Research Council. p. 319-348.

GUYTON, A. C. y HALL, J. E. (2001). Manual de Fisiologia Medica. 10a ed. Madrid: McGraw Hill/Interamericana de Espana. 800 p.

HAIR, J. F; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. (1999). Analisis Multivariante. 5 ed. Madrid: Prentice-Hall. 832 p.

HAYASHI, C. (1950). On the quantification of qualitative data from the mathematicostatistical
point of view. *Annals of the Institute of Statistical Mathematics*, 2(1):35-47.
HEISER, W. J. (1995). Convergent computing by iterative majorization: theory and
applications in multidimensional data analysis. **In:** W. J. Krzanowski (Ed.),
Recent Advances in Descriptive Multivariate Analysis. Oxford: Clarendon Press.
p. 157-189.
HILL, M. O. (1973). Reciprocal averaging: an eigenvector method of ordination.
Journal of Ecology. 61:237-249.
HORN, J. L. (1965). A rationale test for the number of factors in factor analysis.
Psychometrika, 30:179-185.
HORST, P. (1935). Measuring complex attitudes. *Journal of Social Psychology*,
6:369-374.
HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal
components. *Journal of Educational Psychology*, 24(6):417-441.
HOTELLING, H. (1936). Relations between two sets of variates. *Biometrika*, 28:321-327.
HOUSEHOLDER, A. S. and YOUNG, G. (1938). Matrix approximation and latent roots. *American mathematical monthly*, 45:165-171.
JOBSON, J. D. (1992). Applied multivariate data analysis. Volume II. Categorical and
multivariate methods. New York: Springer. 731 p.
KAISER, H. F. (1960). The application of electronic computers to factor analysis.
Educational and Psychological Measurement, 20:141-151.

LANGE, K.; HUNTER, D. R. and YANG, I. (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9:1-20.
LAURO, N. C. et D' AMBRA, L. (1984). l' Analyse non symetrique des

correspondances. **In:** E. Diday, M. Jambu, L. Lebart, J. Pages, and R. Tomassone (Eds.), *Data Analysis and Informatics, Vol. III*. Amsterdam (North-
Holland): Diday et al. p. 433-446.
LECLERC, A. (1980). Quelques proprietes optimales en analyse de donnees en terme
de correlation entre variables. *Mathematique et Sciences Humaines*, 18:51-67.
LINGOES, J. C. (1963). Multivariate analysis of contingencies: An IBM 7090 program
for analyzing metric/nonmetric or linear/non-linear data. [Computer program].
Ann Arbor, MI: University of Michigan Computing Center. (*Computational Report*, 2:1-24.)
LOMBARDO, R; BEH, E. J. and D' AMBRA, L. (2007). Non-symmetric correspondence analysis with ordinal variables using orthogonal polynomials.
Computational Statistics & Data Analysis, 52:566-577.
MARAUN, M. D.; SLANEY, K. and JALAVA, J. (2005). Dual scaling for the analysis
of categorical data. *Journal of Personality Assessment*, 85(2):209-217.
MARDIA, K. V.; KENT, J. T. and BIBBY, J. M. (1979). Multivariate analysis.
London: Academic Press. 521 p.
MARTIN-MOREIRAS, I.; CRUZ, M. D.; MARTIN-HERRERO, F.; LEON, V.; SANCHEZ, M.; RAMIREZ, V. H.; MARTIN-LUENGO, C. (2005). Clinical significance of NT-proBNP Levels in a Diagnosis Exercise Testing. **En:** Scientific Sessions of the American Heart Association. Dallas (Texas).
MATHERS, C. D. AND LONCAR, D. (2005). Updated projections of global mortality
and burden of disease, 2002-2030: data sources, methods and results. Ginebra:
World Health Organization. 128 p.

McKEON, J. J. (1966). Canonical analysis: Some relations between canonical
correlation, factor analysis, discriminant function analysis and scaling theory.
[Monograph No. 13]. *Psychometrika*.

MEULMAN, J. J.; HEISER, W. J; SPSS Inc. (2005). SPSS Categories 14.0. Chicago:
SPSS, 389 p.
MEULMAN, J. J.; Van Der KOOIJ, A. J.; HEISER, W. J. (2004). Principal components
analysis with nonlinear optimal scaling transformations for ordinal and nominal
data. **In:** David Kaplan (Ed.), *The sage handbook of quantitative methodology
for the social sciences*. Thousand Oaks (California): Sage Publications. p. 49-70.
MICHAILIDIS, G. and De LEEUW, J. (1998). The Gifi System of descriptive multivariate analysis. *Statistical Science*, 13(4):307-336.
NISHISATO, S. (1979). Dual Scaling and its variants. *New Directions for Testing and
Measurement*, 4:1-12.
NISHISATO, S. (1980). Analysis of categorical data: Dual scaling and its applications.
Toronto: University of Toronto Press. 276 p.
NISHISATO, S. (2004). Dual scaling. **In:** David Kaplan (Ed.), *The sage handbook of
quantitative methodology for the social sciences*. Thousand Oaks (California):
Sage Publications. p. 3-24.
NISHISATO, S. (2007). Multidimensional nonlinear descriptive analysis. Boca Raton
(Florida): Chapman & Hall. 312 p.
NUNNALLY, J. C. y BERNSTEIN, I. J. (1995). Teoria Psicometrica. 3 ed. Mexico:
McGraw-Hill. 843 p.

OLIVA MORENO, J.; LOBO ALEU, F.; LOPEZ BASTIDA, J.; DUQUE GONZALEZ, B.; OSUNA GUERRERO, R. (2004). Costes no sanitarios ocasionados por las enfermedades isquemicas del corazon en Espana. C*uadernos economicos de I.C.E.*, 67:263-298.
OMS. (2004). Informe sobre la salud en el mundo 2004: cambiemos el rumbo de la
historia. Ginebra: Organizacion Mundial de la Salud. 182 p.
PEARSON, K. (1901). On lines and planes of closest fit to systems of points in space.

Philosophical Magazine and Journal of Science, Ser. 6, 2(11): 559-572.
PEARSON, K. (1896). Mathematical contributions to the theory of evolution. III.
Regression, heredity and panmixia. *Philosophical Transactions of the Royal*
Society of London, Ser. A, 187: 253-318.
RAMIREZ CASTRO, V. H.; DIEGO DOMINGUEZ, M.; RUIZ OLGADO, M; SANCHEZ FLORES, M; SANTOS RODRIGUEZ, I; CASCON BUENO, M.; RODRIGUEZ COLLADO, J. MARTIN LUENGO, C. (2006). Valor y significado clinico de las imagenes de perfusion miocardica en pacientes con
angina de pecho y arterias coronarias sin lesiones angiograficas significativas.
En: XI Congreso de la Sociedad Castellano-Leonesa de Cardiologia (SOCALEC), Valladolid.
ROY, S. N. (1957). Some aspects of multivariate analysis. New York: John Wiley &
Sons. 214 p.
RUANO, R.; DIEGO DOMINGUEZ, M.; MARTIN de ARRIBA, A; RAMIREZ, V. H.; MARTIN LUENGO, C.; GARCIA-TALAVERA, J. R. (2005a). Gated myocardial SPECT in patients with ischemia: relationship between systolic ventricular dysfunction and coronary artery disease severity. **En:** Congreso de la
Asociacion Europea de Medicina Nuclear. Estambul (Turquia), octubre de 2005.

RUANO, R.; RAMIREZ, V. H.; MARTIN de ARRIBA, A; DIEGO DOMINGUEZ, M.; TAMAYO, P.; LEON, V. GARCIA-TALAVERA, J. (2005b). Seguimiento a largo plazo de pacientes con dolor toracico e isquemia demostrada mediante
Gated-SPECT de perfusion miocardica y coronarias angiograficamente normales. **En:** Congreso Espanol de Medicina Nuclear, Madrid, 2005.
SAITO, T. (1973). Quantification of categorical data by using the generalized variance.
Soken Kiyo, Nippon UNIVAC Sogo Kenkyu-sho, Inc., p. 61-80.
SAPORTA, G. (1975). Liaisons entre plusieurs ensembles de variables et codages de
donnees qualitatives. These de Doctorat de 3eme cycle, Paris.
SELWYN, A. P.; KINLAY, S; CREAGER, M.; LIBBY, P.; GANZ, P. (1997). Cell

dysfunction in atherosclerosis and the ischemic manifestations of coronary
artery disease. *American Journal of Cardiology*, 79:17-23.
SHEPARD, R. N. (1962). The analysis of proximities: Multidimensional scaling with
an unknown distance function (I & II). *Psychometrika*, 27:125-139, 219-246.
SOKAL, R. R. and SNEATH, P. H. A. (1963). Principles of numerical taxonomy. San
Francisco: Freeman. 359 p.
SPEARMAN, C. (1904). General intelligence objectively determined and measured.
A*merican Journal of Psychology*, 15:201-293.
STEVENS, S. S. (1946). On the theory of scales of measurement. *Science*. 103:677-680.
STEVENS, S. S. (1951). Mathematics, measurement, and psychophysics. **In:** S. S.
Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley. p. 1-49.
STEVENS, S. S. (1958). Problems and methods of psychophysics. *Psychological
Bulletin*. 55:177-196.

TENENHAUS, M. and YOUNG, F. W. (1985). An analysis and synthesis of multiple
correspondence analysis, optimal scaling, dual scaling, homogeneity analysis
and other methods for quantifying categorical multivariate data. *Psychometrika*,
50(1):91-119.
TIMMERMAN, M. E. and KIERS, H. A. L. (2000). Three-mode principal components
analysis: choosing the number of components and sensitivity to local optima.
British journal of mathematical and statistical psychology, 53:1-16.
TORGERSON, W. S. (1958). Theory and methods of scaling. New York: Wiley. 460 p.
TUCKER, L. R. (1966). Some mathematical notes on three-mode factor analysis.
Psychometrika, 31:279-311.

Van der BURG, E.; De LEEUW, J.; DIJKSTERHUIS, G. (1994). OVERALS: Nonlinear canonical correlation with k sets of variables. *Computational Statistics & Data Analysis*, 18:141-163.
Van der BURG, E.; De LEEUW, J. and VERDEGAAL, R. (1988). Homogeneity analysis with k sets of variables: an alternating least squares method with
optimal scaling features. *Psychometrika*, 53:177-197.
Van RIJCKEVORSEL, J. and De LEEUW, J. (1978). An outline to HOMALS-I. (Research Bulletin RB 002-78). Leiden: University of Leiden. 95 p.
VELICER, W. F. (1976). Determining the number of components from the matrix of
partial correlations. *Psychometrika*, 41:321-327.
VERBOON, P. and HEISER, W. J. (1994). Resistant lower rank approximation of
matrices by iterative majorization. *Computational statistics & data analysis*,
18:457-467.
VERHULST, P. F. (1838). Notice sur la loi que la population suit dans son accroissement. *Correspondance Mathematique et Physique*. 10:113-121.

VERHULST, P. F. (1845). Recherches mathematiques sur la loi d'accroissement de la
population. *Nouveaux Memoires de l'Academie Royale des Sciences, des lettres
et des Beaux-Arts de Belguique,* 18:1-38.
VERHULST, P. F. (1847). Deuxieme memoire sur la loi d'accroissement de la
population. *Nouveaux Memoires de l'Academie Royale des Sciences, des lettres
et des Beaux-Arts de Belguique,* 20:1-32.
VICENTE-VILLARDON, J. L.; GALINDO-VILLARDON, M. P. and BLAZQUEZZABALLOS, A. (2006). Logistic biplots. **In:** Michael Greenacre and Jorg Blasius (Ed.), *Multiple correspondence analysis and related methods*. London:
Chapman & Hall. p. 503-521.
WARRENS, M. J.; De GRUIJTER, D. N. M. and HEISER, W. (2007). A systematic comparison between classical optimal scaling and the two-parameter IRT model.
Applied Psychological Measurements, 31(2):106-120.

WILKS, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*,
24:471-494.
YAN, W. and KANG, M. S. (2003). GGE biplot analysis: A graphical tool for breeders,
geneticists, and agronomists. Boca Raton (Florida): CRC Press. 271 p.
YOUNG, F. W. (1981). Quantitative analysis of qualitative data. *Psychometrika*,
46(4):357-388.
ZWICK, W. R. and VELICER, W. F. (1986). Comparison of five rules for determining
the number of components to retain. *Psychological bulletin*, 99(3):432-442.