

VNIVERSIDAD D SALAMANCA
DEPARTAMENTO DE ESTADÍSTICA



**CONTRIBUCIONES AL ANÁLISIS
MULTIVARIANTE NO LINEAL**

Guillermo Correa Londoño

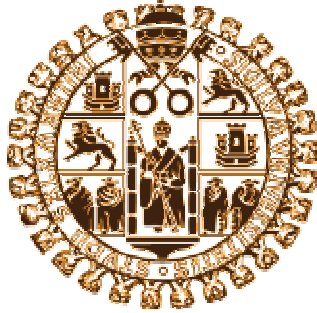
2008

CONTRIBUCIONES AL ANÁLISIS MULTIVARIANTE NO LINEAL

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

Guillermo Correa Londoño

Salamanca
2008



*Universidad de Salamanca
Departamento de Estadística*

M^a PURIFICACIÓN GALINDO VILLARDÓN

y

CARMELO ANTONIO ÁVILA ZARZA

*Profesores Titulares del Departamento de Estadística
de la Universidad de Salamanca*

CERTIFICAN: Que **Don Guillermo Antonio Correa Londoño**, ha realizado, en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo que, para optar al Grado de Doctor, presenta con el título: “*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 Agronómicas, han realizado en pro de mi formación.

Al **Grupo Santander** por la financiación de este programa doctoral a través de una de las becas *Universidad de Salamanca-Grupo Santander*, destinadas a estudiantes Iberoamericanos.

A la doctora **M^a. Purificación Galindo Villardón**, directora de este trabajo, por la cálida acogida que desde el primer momento me brindó, por su constante apoyo, tanto en la parte académica 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 técnicas al mismo, pero sobre todo por haberme hablado de fotografía, de política, de religión, de turismo, de iPhones y hasta de la Pasión según San Mateo; en resumen, por haberme brindado su amistad.

Al doctor **José Luis Vicente Villardón** por sus valiosos aportes en diversos aspectos matemáticos y computacionales.

Al cardiólogo **Maximiliano Diego Domínguez** por confiarme la base de datos para la aplicación práctica.

Al cardiólogo **Víctor Hugo Ramírez Castro**, quien además de haberme dado una valiosa asesoría en el campo técnico de este trabajo, me brindó su amistad y soporte en los momentos difíciles.

A todo el **Personal del Departamento de Estadística de la Universidad de Salamanca** por demostrar que la calidad académica no tiene que ser excluyente con la calidad humana.

A **Luz Estela Sánchez Herrera**, quien, gracias a su formación matemática, me ayudó a salir de numerosos escollos.

A mi esposa, **Luz Marina**, con quien he compartido tantos años de mi vida, cada uno más excitante que el anterior, por ser la gran motivación detrás 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 desarrolló este trabajo, por haberme brindado esos momentos tan maravillosos de esparcimiento, y también a los que por la distancia no pudieron hacerlo, por esperarme pacientemente.

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

Cuantifica, siempre que sea posible.

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

Carl Sagan

ÍNDICE

	Pág.
INTRODUCCIÓN	1
CAPÍTULO 1. ASPECTOS GENERALES	5
1.1 INTRODUCCIÓN	6
1.2 CONCEPTOS BÁSICOS	7
1.3 ALGUNAS TÉCNICAS MULTIVARIANTES DE ANÁLISIS DE DATOS	11
1.4 CAMBIOS DE ESCALA	15
1.5 CATEGORIZACIÓN DE VARIABLES	17
1.6 CODIFICACIÓN DE VARIABLES	18
1.7 CUANTIFICACIÓN DE VARIABLES.....	20
CAPÍTULO 2. SISTEMA GIF1	22
2.1 INTRODUCCIÓN	23
2.2 SISTEMA GIF1 DE ANÁLISIS MULTIVARIANTE NO LINEAL.....	24
2.3 ESCALAMIENTO ÓPTIMO	27
2.4 MÍNIMOS CUADRADOS ALTERNADOS	29
2.5 ESCALAMIENTO, TRANSFORMACIONES Y CUANTIFICACIONES.....	32
2.6 HOMALS.....	37
2.7 OVERALS	42
CAPÍTULO 3. EFECTO DE LA DIMENSIONALIDAD SOBRE LAS CUANTIFICACIONES	45
3.1 INTRODUCCIÓN	46
3.2 DIMENSIONALIDAD DE LAS SOLUCIONES.....	47
3.3 TIPOS DE CUANTIFICACIÓN	49
3.4 EFECTO DE LA DIMENSIONALIDAD SOBRE LAS CUANTIFICACIONES. 51	
3.5 ELECCIÓN DE LA DIMENSIONALIDAD EN PROBLEMAS DE CUANTIFICACIÓN	57

	Pág.
CAPÍTULO 4. CUANTIFICACIÓN ÓPTIMA	59
4.1 INTRODUCCIÓN	60
4.2 DEFINICIÓN GENERAL DE CUANTIFICACIÓN ÓPTIMA	61
4.3 ILUSTRACIÓN DE LA DEFINICIÓN GENERAL.....	62
4.4 DEFINICIÓN ESPECÍFICA DE CUANTIFICACIÓN ÓPTIMA	79
4.5 PLANTEAMIENTO ESPECÍFICO DE LA SOLUCIÓN	81
4.6 CONSIDERACIONES SOBRE LA PROPUESTA	84
4.7 ALGORITMO PARA LA ELECCIÓN DE LA CUANTIFICACIÓN ÓPTIMA... 85	
4.7.1 CAPTACIÓN DE INFORMACIÓN GENERAL	85
4.7.2 LECTURA DE INFORMACIÓN ESPECÍFICA	87
4.7.3 CÁLCULO DE LAS <i>INTERDISTANCIAS</i>	90
4.7.4 PRESENTACIÓN DE RESULTADOS	90
4.8 ILUSTRACIÓN DE LA PROPUESTA EN UNA APLICACIÓN PRÁCTICA	91
CAPÍTULO 5. PROPUESTA GENERALIZADA DE CUANTIFICACIÓN	
ÓPTIMA: CUANTIFICA	93
5.1 INTRODUCCIÓN	94
5.2 MARCO DE REFERENCIA DE LA CUANTIFICACIÓN ÓPTIMA.....	96
5.3 RUTINA CUANTIFICA.....	100
5.3.1 CAPTACIÓN DE LA INFORMACIÓN DE ENTRADA	102
5.3.2 LECTURA Y ADECUACIÓN DE LA BASE DE DATOS	108
5.3.3 OBTENCIÓN DE LA MATRIZ INICIAL DE CUANTIFICACIONES ..	110
5.3.4 SUBROUTINAS DE INICIALIZACIÓN.....	111
5.3.5 CÁLCULO DE LAS <i>INTERDISTANCIAS</i> INICIALES	119
5.3.6 CICLO DE MINIMIZACIÓN DE LAS <i>INTERDISTANCIAS</i>	119
5.3.7 PRESENTACIÓN DE RESULTADOS	126
5.4 COMPARACIÓN ENTRE CUANTIFICA Y EL SISTEMA GIFL.....	132

CAPÍTULO 6. APLICACIÓN DE LA PROPUESTA: CARDIOPATÍA ISQUÉMICA	137
6.1 INTRODUCCIÓN	138
6.2 GENERALIDADES DE LA CARDIOPATÍA ISQUÉMICA	140
6.3 IMPACTO DE LA CARDIOPATÍA ISQUÉMICA	142
6.4 BASE DE DATOS.....	143
6.5 DESCRIPCIÓN DE ALGUNAS VARIABLES	148
6.6 CUANTIFICACIÓN	153
6.7 CONSTRUCCIÓN DE LA REPRESENTACIÓN BIPLLOT.....	161
6.7.1 BASE TEÓRICA	161
6.7.2 RUTINA MARCADORES GCMP-BIPLLOT	170
6.7.3 RUTINA REPRESENTACIONES GCMP-BIPLLOT	171
6.8 RESULTADOS	177
6.9 SÍNTESIS DE LA PROPUESTA DE ANÁLISIS	186
6.10 COMPARACIÓN DE RESULTADOS CON LOS OBTENIDOS MEDIANTE TÉCNICAS DEL SISTEMA GIF1.....	191
CONCLUSIONES	198
BIBLIOGRAFÍA	202

INTRODUCCIÓN

Aunque las primeras pruebas documentales del interés por cuantificar lo cualitativo datan de principios del siglo XX (YULE, 1910, citado por YOUNG, 1981), es de suponer que tal inquietud nació casi a la par del reconocimiento de estas dos facetas de la información. En la década del 40, Guttman establece las bases de uno de los primeros métodos de cuantificación (GUTTMAN, 1941), el cual, además de haber gozado de amplia difusión, 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 así, la cuantificación es un tópico en el que siguen quedando cuestiones por resolver. En tal sentido, se ha desarrollado este trabajo como una aportación a dicha temática.

Hemos elaborado nuestras propuestas con base en el siguiente **plan general**. En el **Capítulo 1**, a forma de introducción, presentamos algunos conceptos básicos y definiciones necesarias para la adecuada comprensión de la problemática de la cuantificación, la cual se desarrolla con mayor profundidad en los capítulos subsiguientes. **El Capítulo 2** está centrado en el sistema Gifi de análisis multivariante no lineal, el cual constituye el cuerpo de doctrina más depurado sobre el análisis multivariante no lineal, uno de cuyos aspectos es la cuantificación.

En el **Capítulo 3** se pone en evidencia el hecho de que al usar las técnicas del sistema Gifi para la cuantificación de variables cualitativas, surgen diferentes conjuntos de cuantificaciones, sin que exista ningún criterio para decidir cuál de los posibles conjuntos de cuantificaciones es el más adecuado. En tal sentido, en este capítulo aparece la primera aportación de este trabajo, pues el problema anotado, al parecer, no había sido detectado o, por lo menos, no había sido discutido anteriormente.

El **Capítulo 4** ofrece una solución al problema planteado en el capítulo precedente, partiendo para ello de un criterio de cuantificación óptima independiente del modelo de análisis. Tanto el criterio de cuantificación óptima como el algoritmo que permite elegir el conjunto de cuantificaciones óptimas a partir de varios posibles conjuntos de cuantificaciones son desarrollos originales, constituyendo, por tanto, la segunda aportación de este trabajo.

En el **Capítulo 5** se desarrolla una generalización de la solución planteada en el capítulo precedente para la obtención de cuantificaciones óptimas. Dicha generalización hace posible obtener cuantificaciones óptimas directamente de los datos originales, sin que se requieran cuantificaciones previas generadas mediante técnicas del sistema Gifi ni de ningún otro sistema. Para tal efecto, se implementa un módulo computacional desarrollado en MATLAB, el cual asigna cuantificaciones óptimas, tomando en consideración 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 teórica que los sustenta son originales, este capítulo en su conjunto representa la mayor aportación de este trabajo.

En el **Capítulo 6** se ilustra el uso de la propuesta desarrollada en el capítulo precedente, a través de una aplicación práctica, consistente en el análisis de una base de datos de 6.965 pacientes con cardiopatía isquémica o sintomatología afín, con 53 variables medidas en diferentes escalas y con 53,12 % de información faltante. En este capítulo, se utiliza el método de cuantificación propuesto en el capítulo precedente como un paso de adecuación de la base de datos para su posterior análisis mediante una técnica lineal. En este caso, hemos utilizado la metodología de las representaciones Biplot como técnica de exploración final. Los resultados que se desprenden del análisis de esta base de datos son otra de las aportaciones de este trabajo.

Con el fin de llevar el análisis de la base de datos hasta el final, aprovechando toda la información disponible, en el **Capítulo 6** se desarrolla una propuesta para la elaboración de representaciones Biplot de matrices con datos faltantes. Esta propuesta constituye también una aportación de este trabajo, por la originalidad tanto de su desarrollo teórico como de las rutinas computacionales que automatizan su implementación. No obstante, hemos optado por mantener esta propuesta como un apartado del Capítulo 6, por concebirla como un complemento al núcleo central de este trabajo, que es la cuantificación óptima.

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 aplicación práctica del Capítulo 6.

Para facilitar el uso de las rutinas computacionales elaboradas en este trabajo, se ha creado una interfaz gráfica para cada una de ellas. A través de tales interfaces, el usuario puede ingresar la información de entrada, especificar las opciones de los análisis y acceder a las correspondientes ayudas. Asimismo, con el fin de integrar los diferentes módulos computacionales y de proporcionar un acceso sencillo a los mismos, se ha creado una interfaz de inicio común que permite derivar hacia cualquiera de los módulos. Luego, basta con realizar una instalación conjunta de todos los módulos y acceder a cualquiera de ellos a través del módulo general. Tanto el módulo de entrada como los módulos específicos cuentan con ayudas que van guiando al usuario a través del proceso.

Cerramos esta memoria presentando las conclusiones y las referencias bibliográficas que han servido de base y de inspiración para los desarrollos realizados en este trabajo.

CAPÍTULO 1

ASPECTOS GENERALES

1.1 INTRODUCCIÓN

Iniciamos **este capítulo** presentando algunos conceptos generales como el de análisis multivariante y sistemas multivariantes, así como una clasificación de estos últimos acorde con el rol desempeñado por las variables que los conforman y con las escalas de medición de estas. Definimos las escalas de medición numérica, ordinal y nominal. Seguidamente, presentamos una definición de sistemas multivariantes mixtos y homogéneos, con base en las escalas de medición de las variables que los conforman. Asimismo, con base en las escalas de medición, precisamos la definición de variables cuantitativas y cualitativas.

A continuación, presentamos una relación de las principales técnicas multivariantes clásicas de análisis de datos, ubicándolas en un esquema general, según el tipo de datos sobre los cuales sean aplicables (numéricos, ordinales, nominales) y el rol de las variables (técnicas de dependencia y de interdependencia).

Se expone luego el tópico de cambio de escala de variables, en sus dos formas posibles: debilitamiento de escala y fortalecimiento de escala. Se realiza una breve discusión sobre categorización y codificación de variables, cerrando el capítulo con lo tocante a su cuantificación.

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 análisis multivariante y, como tal, empezaremos definiendo algunas de sus características.

Los **sistemas multivariantes**, es decir, aquellos conformados por múltiples variables, pueden ser clasificados con base en diferentes criterios, a dos de los cuales haremos referencia: con base en el rol desempeñado por las variables en el análisis; y con base en las escalas de medición de las variables que los conforman. Mientras el primer aspecto depende de la técnica de análisis, pudiendo cambiar al usar diferentes técnicas para un mismo sistema, el segundo tiene que ver con características intrínsecas del sistema.

Según el **rol desempeñado por las variables** en el análisis, los sistemas multivariantes pueden ser **asimétricos** o **simétricos**. Se dice que un sistema multivariante es asimétrico 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 más variables. Por simplicidad, en los casos anteriores se habla de datos asimétricos o simétricos, respectivamente. A las técnicas de análisis multivariante que procesan la información 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 análogo de todas las variables, se dice que el sistema multivariante es simétrico. A las técnicas multivariantes que otorgan igual rol a las variables se les denomina **técnicas de interdependencia** (HAIR et ál., 1999).

Si bien es cierto que la definición del rol de las variables no puede ser arbitraria y que debe estar en consonancia con los objetivos del análisis, también 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 intrínseco inmutable, sino que dependa de los objetivos del análisis y de la técnica utilizada para alcanzar tales objetivos.

El otro criterio de tipificación 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 medición son convenciones que se adoptan para registrar una realidad. Obsérvese, por ejemplo, que existen múltiples propuestas o escalas para registrar una única realidad como la temperatura: Celsius, Fahrenheit, Kelvin, Rankine y Réaumur, entre otras. Luego, podemos definir escala de medición como *cualquier sistema que permita registrar una característica o atributo de un objeto de interés*.

Lógicamente, más allá de la naturaleza de la realidad que se pretenda registrar, las propiedades de una *variable observada* estarán determinadas por su escala de medición, es decir, por el sistema utilizado para el registro de la realidad subyacente.

Antes de realizar una caracterización de las principales escalas de medición, es importante anotar que las variables cuantitativas pueden, a su vez, considerarse **continuas** o **discretas**. Desde el punto de vista teórico, este concepto está ligado a la función de probabilidad asociada con cada variable aleatoria. En la práctica, 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 medición. Considerando que todas las variables continuas son discretizables, podemos obviar esta distinción 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 aparición de una serie de publicaciones de STEVENS (1946, 1951, 1958) dio lugar a una caracterización y clasificación de las escalas de medición, la cual ha sido ampliada, reducida o modificada posteriormente, en función de objetivos y enfoques particulares. Para nuestros fines, bastará con considerar los tres **tipos de escalas de medición** que se describen a continuación.

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 sí, será posible definir un infinito número de instantes. No obstante, la edad de un individuo medida en años cumplidos es una variable discreta.

Escala Numérica. El valor asociado con cada nivel de esta escala indica la cantidad o intensidad de la característica 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 además, la escala posee un valor de referencia cero, correspondiente a la ausencia de la característica medida, también será posible establecer relaciones de razón. Obsérvese que esta definición recoge lo que para otros fines serían las escalas de intervalo y de razón.

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 podría ocurrir eventualmente, no tendría que satisfacerse para todos los niveles; si así fuera, la escala no sería ordinal, sino numérica.

Escala Nominal. El valor asociado con cada uno de los niveles de la variable no es más que una etiqueta de identificación, sin otro valor de comparación 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 categorías exhaustivas y mutuamente excluyentes, esto es, las variables binarias o dicotómicas.

Existe una clasificación general de las **variables** en **cuantitativas** y **cualitativas**. Aunque estos términos se autodefinen y no demandan precisiones adicionales, vale la pena relacionarlos con las escalas de medición que acabamos de describir. Se dice que las variables numéricas 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 medición descritos los llamaremos simplemente *escalas de medición*. Además, cuando una variable se registre utilizando alguna de estas escalas de medición, se dirá que tal variable está en dicha escala o se utilizará el nombre de la escala como adjetivo, v. gr., variable numérica, variable ordinal o variable nominal.

Tomando como referente las tres escalas de medición descritas, definimos un **sistema multivariante mixto** como aquel en el que no todas las variables están en la misma escala de medición. Por simplicidad, puede hablarse también, en tales casos, de información mixta o datos mixtos. En contraposición, y manteniendo el presente contexto, un **sistema multivariante homogéneo** es aquel en el que todas las variables están en la misma escala de medición. En este caso podrá hablarse de datos homogéneos.

1.3 ALGUNAS TÉCNICAS MULTIVARIANTES DE ANÁLISIS DE DATOS

La escala o combinación de escalas de medición de las variables que conforman los sistemas multivariantes, así como el rol de las variables, determinan las técnicas disponibles para su análisis. La aplicación de una u otra dependerá, desde luego, de los objetivos trazados.

Aunque la principal aportación del presente estudio es precisamente un sistema integral, el cual, a partir de la asignación de cuantificaciones óptimas, hace posible analizar sistemas multivariantes mixtos usando prácticamente cualquier técnica multivariante, por el momento nos circunscribiremos al ámbito de aplicación de los principales métodos multivariantes en sus versiones clásicas². En la Figura 1.1 se presenta un esquema que considera la ubicación de tales métodos, teniendo en cuenta si el sistema multivariante es mixto u homogéneo y si los datos son simétricos o asimétricos.

Obsérvese que las posibles combinaciones de las tres escalas de medición consideradas definen una matriz 3 x 3, con los sistemas multivariantes homogéneos 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 medición 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 definición, son imposibles de obtener. Por ejemplo, no existe ninguna técnica en la que los dos grupos de variables sean independientes o los dos grupos dependientes³. Tales celdas se han cruzado con una X.

² Nos referimos a las formas más populares de las técnicas, sin considerar las adaptaciones basadas en cambios de escala. Tampoco incluimos las técnicas de análisis de datos de tres vías.

³ Las técnicas 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
Numérica	IND			ARM						AD RLOG
	INT	ACP AF ACOP ACC ACONG BIPLOT								
	DEP							MANOVA		
Ordinal	IND						ACNSVO			
	INT					EMNM				
	DEP									
Nominal	IND									ACNS
	INT								AC ACM BIPLOG	
	DEP									

AC: Análisis de Correspondencias; **ACM:** Análisis de Correspondencias Múltiple; **ACNS:** Análisis de Correspondencias no Simétrico; **ACNSVO:** Análisis de Correspondencias no Simétrico con Variables Ordinales; **ACON:** Análisis de Conglomerados; **ACOP:** Análisis de Coordenadas Principales; **ACP:** Análisis de Componentes Principales; **AD:** Análisis Discriminante; **AF:** Análisis Factorial; **ARM:** Análisis de Regresión Múltiple; **BILOG:** Biplot Logístico; **BIPLOT:** Análisis Biplot; **EMNM:** Escalamiento Multidimensional no Métrico; **MANOVA:** Análisis de Varianza Multivariante; **RLOG:** Regresión Logística.

Figura 1.1. Clasificación de las principales técnicas multivariantes clásicas, acorde con las escalas de medición y la simetría de los datos.

En el caso de los sistemas multivariantes homogéneos, se ha deshabilitado también la celda inferior izquierda, no porque tal combinación sea imposible de obtener, sino porque en estos sistemas dicha combinación 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 análisis de un sistema multivariante homogéneo en el que todas las variables sean numéricas puede realizarse con base en diferentes técnicas, acorde con los objetivos del análisis y con el rol que desempeñen las variables. Como técnicas de interdependencia, es decir, técnicas en las que se asigna el mismo rol a todas las variables, cabe destacar las siguientes: Análisis de Componentes Principales (PEARSON, 1901; HOTELLING, 1933), Análisis Factorial (SPEARMAN, 1904; HOTELLING, 1933), Análisis de Correlación Canónica (HOTELLING, 1936), Análisis de Coordenadas Principales (GOWER, 1966), Análisis de Conglomerados (FISHER, 1958; SOKAL and SNEATH, 1963) y Análisis Biplot (GABRIEL, 1971; GALINDO, 1985, 1986). Como técnica de dependencia, se destaca el Análisis de Regresión Múltiple (PEARSON, 1896; BARTLETT, 1933).

Sistemas Multivariantes Ordinales. La técnica más popular de interdependencia es el Escalamiento Multidimensional no Métrico (SHEPARD, 1962). Recientemente, LOMBARDO et ál. (2007) desarrollaron una extensión del Análisis de Correspondencias no Simétrico, llamada Análisis de Correspondencias no Simétrico 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 Análisis de Correspondencias (BENZÉCRI et ál., 1973), si se trata de solo dos variables, o el Análisis de Correspondencias Múltiples⁴ (BENZÉCRI et ál., 1973; BENZÉCRI, 1977), cuando se tienen más de dos variables. Un caso particular de variables nominales es aquel en el que cada variable está conformada por solo dos categorías (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 técnicas equivalentes, que nos interesa destacar, por formar parte del sistema de análisis al cual dedicaremos parte de este trabajo, es el Análisis de Homogeneidad mediante Mínimos Cuadrados Alternados (HOMALS).

el grupo de investigación de Análisis Multivariante de la Universidad de Salamanca desarrolló una variante del Análisis Biplot llamada Biplot Logístico (VICENTE-VILLARDÓN et ál., 2006), mediante el cual es posible obtener una representación Biplot de múltiples variables binarias. Las tres técnicas mencionadas (AC, ACM y BIPLOG) son de interdependencia. LAURO y D'AMBRA (1984) desarrollaron una propuesta de Análisis de Correspondencias no Simétrico que es aplicable cuando las variables juegan diferentes roles.

Sistemas Multivariantes Mixtos. Las técnicas más populares de análisis de sistemas multivariantes mixtos son de interdependencia, es decir, que existen dos grupos claramente diferenciados, donde uno desempeña el papel de variable independiente, y el otro, el papel de variable dependiente. Cuando se tiene una variable respuesta nominal que intenta predecirse en función de un grupo de variables numéricas, se utiliza el Análisis Discriminante (FISHER, 1936) o la Regresión Logística (VERHULST, 1838, 1845, 1847). Cuando las variables independientes son nominales y las variables respuestas son numéricas, se utiliza el Análisis de Varianza Multivariante (WILKS, 1932; ROY, 1957).

1.4 CAMBIOS DE ESCALA

Las escalas de medición de las variables que conforman un sistema multivariante determinan el conjunto de técnicas que pueden usarse para su análisis. La mayoría de métodos clásicos son aplicables a sistemas homogéneos o conformados por grupos de variables con combinaciones predeterminadas de escalas de medición. Aunque en muchos casos existen generalizaciones que permiten aplicar las técnicas de análisis en un ámbito más amplio de situaciones que aquellas para las cuales fueron desarrolladas, muchas de tales variantes no se encuentran implementadas en los paquetes estadísticos comerciales que suelen utilizar los analistas de datos. Por tal motivo, constantemente se realizan **procesos de cambio de escala**, bien sea fortaleciéndolas o debilitándolas⁵.

Aunque existen técnicas formales de **fortalecimiento de escala**, en algunas de las cuales nos concentraremos en los siguientes capítulos, es habitual pretender que este proceso se da de manera espontánea. Al menos, esto es lo que implícitamente se supone cuando se usan variables medidas en una escala débil como si fueran de escala fuerte⁶. En tales casos, el analista invoca una supuesta robustez de los métodos estadísticos que no siempre está claramente sustentada.

Puesto que no tiene mucho sentido abordar el tópico de la robustez aisladamente de las técnicas a las que se refiere, siendo necesario un análisis particular para cada una de ellas, no profundizaremos en el mismo. Bastará con decir que en la mayoría de los casos, el fortalecimiento de escalas —en particular, cuando no media ningún proceso formal— es un proceso menos ortodoxo que su respectivo debilitamiento.

Para aplicar una técnica de análisis de datos de escala débil sobre datos de escala fuerte, es necesario efectuar un proceso de **debilitamiento de escala**, el cual se logra resumiendo información o simplemente ignorándola.

⁵ La fortaleza de una escala de medición está determinada por la cantidad de información que conlleva. Así, la escala numérica es la más fuerte, seguida por la ordinal, siendo la nominal la más débil. El proceso de pasar de la escala nominal a la ordinal o a la numérica es de fortalecimiento, mientras que el paso de la escala numérica a la ordinal o a la nominal es de debilitamiento.

⁶ En particular, suelen utilizarse variables ordinales como si fueran numéricas.

Si se tiene, por ejemplo, una variable ordinal, con un número conveniente de niveles (no se estima necesario colapsar niveles) y se desea aplicar una técnica de análisis de datos nominales, se pasa de la escala ordinal a la nominal mediante el simple recurso de ignorar la información de orden contenida en las etiquetas de los niveles. Puesto que, en este caso, el usuario no tiene que realizar ninguna acción y es el algoritmo computacional el que ignora la información de orden, este debilitamiento de escala a menudo pasa desapercibido.

Para realizar el debilitamiento de escala de una variable numérica podría ser suficiente con efectuar un proceso de discretización (si la variable original no era discreta). Lógicamente, para llegar a una verdadera escala nominal habría que agregar el paso de ignorar la información de orden contenida en los niveles de la variable discretizada.

Al resumir variables numéricas u ordinales mediante su agrupación en intervalos adyacentes se genera una escala ordinal. Basta, sin embargo, con agregar el paso implícito de ignorar la información de orden para llegar a una escala nominal.

Obsérvese que al realizar un proceso de debilitamiento de escala hacia la nominal, sin importar los pasos explícitos o implícitos involucrados en el proceso (discretización, colapsamiento de niveles, ignorar la información de orden), la variable queda constituida finalmente por un conjunto de categorías. Es por ello que a este proceso se le conoce comúnmente como categorización.

Tras un proceso de categorización, a las categorías generadas puede asignárseles etiquetas sin ninguna restricción, ya que su única función será la de señalar la membresía de un individuo a una categoría determinada. En tal sentido, es perfectamente viable asignar incluso etiquetas alfanuméricas, 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 numéricas. No obstante, incluso en tales casos, debe tenerse en mente que los números utilizados para identificar a cada categoría no conllevan ninguna información numérica y bien podrían intercambiarse entre categorías.

1.5 CATEGORIZACIÓN DE VARIABLES

Una de las formas de categorizar una variable es mediante un proceso análogo al utilizado en la construcción de tablas o histogramas de frecuencia, es decir, generando un número 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, quizá, la forma más utilizada de categorizar una variable. No obstante, es la menos enfocada a minimizar la pérdida de información implícita en el proceso de debilitamiento de escala.

El aspecto fundamental para minimizar la pérdida de información propia de cualquier proceso de categorización es, en nuestro concepto, la búsqueda de categorías contrastantes entre sí, pero con alta homogeneidad interna. Para tal efecto, se requiere la participación del especialista en el fenómeno estudiado, quien deberá definir los límites de las categorías, acorde con su percepción multivariante del fenómeno.

Supóngase, por ejemplo, que se está analizando una base de datos multivariante de pacientes con cardiopatía isquémica, una de cuyas variables es la edad medida en años cumplidos. Para categorizar esta variable, el cardiólogo deberá definir los rangos de edad que, según su criterio, marcan diferencias en el contexto general de esta enfermedad, haciendo caso omiso de relaciones muy específicas que puedan existir entre la edad y algunas variables particulares de este contexto. Obviamente, los intervalos de edad que así se definan no tienen por qué ser de la misma amplitud.

Aunque esta segunda forma de categorizar variables es superior a la descrita previamente, tanto en términos de conservación de la información 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 más utilizada por su mayor objetividad (y menor nivel de compromiso del usuario con la decisión tomada), pues todo lo que el analista debe definir es el número de intervalos, con lo cual los límites de las categorías se generan automáticamente.

1.6 CODIFICACIÓN DE VARIABLES

Tras la categorización, algunas técnicas, como el Análisis de Correspondencias (AC), reexpresan las variables mediante una matriz indicadora, con tantas filas como observaciones y tantas columnas como categorías hayan resultado del proceso de categorización. El sistema más usual es el de **codificación disyuntiva completa**, mediante el cual cada observación (fila) tendrá un uno en la columna correspondiente a la categoría a la que pertenece y cero en las demás columnas.

Como alternativa a la codificación disyuntiva completa, CAZES (1990) propone un sistema de **codificación baricéntrica**, según el cual, una observación puede quedar caracterizada por valores diferentes de cero en más de una columna de la matriz de codificación. Ello permite conservar no solo la información de pertenencia de una observación a una categoría determinada, sino también la información relativa a la distancia entre las observaciones.

Para la aplicación de la codificación baricéntrica, en lugar de definir los límites de las categorías, se definen r pivotes, es decir, r valores t_1, t_2, \dots, t_r , tales que, $t_1 < t_2 < t_3 < \dots < t_r$. Cuando los pivotes son equidistantes, constituyen las marcas de clase de unos intervalos implícitamente definidos por este sistema de codificación.

Si x_{ij} denota la i -ésima observación de la j -ésima variable, y se denotan sus correspondientes codificaciones para cada uno de los r pivotes como $k(i, j_1), k(i, j_2), \dots, k(i, j_r)$, podemos expresar las reglas del sistema de codificación baricéntrica así:

$$\begin{aligned} \text{Si } x_{ij} \leq t_1 & \Rightarrow & k(i, j_1) = 1 \\ & & k(i, j_s) = 0, \quad \forall s \neq 1 \end{aligned}$$

$$\begin{aligned} \text{Si } x_{ij} \geq t_r & \Rightarrow & k(i, j_r) = 1 \\ & & k(i, j_s) = 0, \quad \forall s \neq r \end{aligned}$$

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

Para ilustrar estos dos métodos de codificación, supóngase una variable medida en escala numérica, que tiene, entre otros, los siguientes valores: 3, 5, 9, 10, 11, 24 y 26. Supóngase también que para realizar una codificación disyuntiva completa se han definido los intervalos [0, 10), [10, 20) y [20, 30). Y supóngase además que para realizar una codificación baricéntrica, se han definido los pivotes 5, 15 y 25. La Tabla 1.1 muestra la codificación de cada uno de estos valores con base en cada uno de los dos sistemas.

Valores originales	Codificación Disyuntiva Completa			Codificación Baricéntrica		
	Intervalos			Pivotes		
	[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. Codificación disyuntiva completa y codificación baricéntrica de un conjunto de observaciones.

Es importante resaltar un par de aspectos relativos al sistema de codificación baricéntrica. En primer lugar, que este solo es aplicable a la codificación de variables numéricas, no siendo adecuado para codificar variables ordinales. En segundo lugar, que este método de codificación resulta útil solamente si se van a aplicar técnicas que permitan operar sobre matrices indicadoras, v. gr., Análisis de Correspondencias; de hecho, es en tal contexto en el que Cazes desarrolla su propuesta.

Finalmente, hay que destacar que si bien existe una relación directa entre cada uno de estos sistemas de codificación y una categorización subyacente, estos no constituyen métodos de categorización en si mismos. En tal sentido, ninguno de ellos libera al analista de la responsabilidad de definir los límites de las categorías.

1.7 CUANTIFICACIÓN DE VARIABLES

Ante la necesidad de analizar datos cualitativos mediante las técnicas 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 información ordinal como si fuera numérica. Los analistas que se ven obligados a ello, justifican su proceder en la robustez de los métodos. Sin entrar en las particularidades de ningún método específico, nos limitaremos a señalar que, a no ser que exista un claro sustento de la robustez en cuestión, esta práctica no es recomendable.

La alternativa más ortodoxa para la aplicación de métodos 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 numérico, motivo por el cual se les denomina **métodos de cuantificación**.

En ocasiones, la cuantificación no constituye un fin en si misma, sino que forma parte integral de las técnicas, apareciendo, a lo sumo, como un subproducto de estas. Por tal motivo, a menudo se ignora el papel de la cuantificación y se tiende a creer que técnicas concebidas para datos cuantitativos funcionan igualmente bien al ser aplicadas sobre datos cualitativos.

Según la percepción que se tenga del proceso de aplicación de técnicas cuantitativas a datos mixtos o cualitativos, este puede abordarse desde dos ópticas: la primera, según la cual algunas técnicas cuantitativas pueden aplicarse indistintamente a datos cuantitativos, mixtos o totalmente cualitativos; la segunda, que considera que las técnicas cuantitativas siempre exigen datos cuantitativos. El primer punto de vista es correcto, siempre que medie una cuantificación de las variables cualitativas, lo cual nos lleva al segundo escenario. Si bien esta es una cuestión más semántica que práctica, la traemos a colación para hacer notar que la aplicación de métodos cuantitativos sobre datos cualitativos o mixtos siempre conlleva una cuantificación implícita o explícita de las variables cualitativas.

Aunque el interés por el tópico de la cuantificación 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 difusión fue la de GUTTMAN (1941), siendo además una de las que daría lugar a una línea de trabajo de mayor continuidad (cf. § 2.6). A principios de la década de 1950, varios autores realizaron contribuciones sobre este tema (BURT, 1950, 1953; HAYASHI, 1950), las cuales fueron resumidas a finales de la década por TORGERSON (1958). En 1976 se publica una propuesta para analizar la estructura aditiva en datos cualitativos (De LEEUW et ál., 1976). Paralelamente, aparecen aportaciones de las escuelas francesa, nipona, canadiense e inglesa (BENZÉCRI et ál., 1973; BENZÉCRI, 1977; SAPORTA, 1975; SAITO, 1973; NISHISATO, 1980; MARDIA et ál., 1979).

A principios de la década de 1980, Young publica un trabajo titulado *Análisis Cuantitativo de Datos Cualitativos* (YOUNG, 1981), en el que hace referencia al proceso de cuantificación como *escalamiento óptimo*, término introducido por BOCK (1960). En su trabajo, Young define el **escalamiento óptimo**, así: “*Técnica de análisis de datos que asigna valores numéricos a las categorías, de manera que se maximice la relación entre las observaciones y el modelo de análisis de datos, respetando el carácter de medición de los datos*”. Posteriormente, en este trabajo, retomaremos el concepto de escalamiento óptimo y actualizaremos su definición (cf. § 2.3).

También en 1981, sale a la luz el libro *Análisis Multivariante no Lineal*, publicado por la escuela holandesa de sistema de escalamiento de datos, el cual es reeditado en 1990 (GIFI, 1981, 1990). Allí se exponen detalladamente los principios y aplicaciones del escalamiento óptimo, con lo cual este trabajo se convierte en el referente por excelencia de lo relacionado con cuantificación y métodos multivariantes no lineales. Por tal motivo, hemos dedicado el Capítulo 2 del presente trabajo a la presentación de algunos aspectos del sistema Gifi de análisis multivariante no lineal, en particular, los relacionados con la cuantificación de variables cualitativas.

CAPÍTULO 2

SISTEMA GIFÍ

2.1 INTRODUCCIÓN

En **este capítulo** presentaremos inicialmente el sistema Gifi de análisis multivariante no lineal en forma general. Luego, se definirá la homogeneidad, en el contexto de información compartida por las variables del sistema, concepto central al sistema Gifi y totalmente diferente de la definición de sistemas multivariantes homogéneos que se presentó en el Capítulo 1, haciendo referencia a sistemas constituidos por variables medidas en la misma escala.

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

Seguidamente, se expone el método de los mínimos cuadrados alternados, el algoritmo más popular para resolver problemas de escalamiento óptimo.

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

A continuación, se presenta el HOMALS como una generalización de la función de pérdida del Análisis de Componentes Principales. Finalmente se presenta el OVERALS, la técnica más general del sistema Gifi, que permite expresar las demás técnicas como casos particulares de esta.

2.2 SISTEMA GIFI DE ANÁLISIS MULTIVARIANTE NO LINEAL

La publicación en 1981 de la versión en inglés del texto *Nonlinear Multivariate Analysis* (GIFI, 1981), por parte del grupo de investigación en Teoría de Datos de la Facultad de Ciencias Sociales de la Universidad de Leiden, en Holanda, que para la ocasión escribió bajo el seudónimo de **Albert Gifi**⁷, marcó un hito en la concepción de las técnicas multivariantes no lineales.

Este texto recoge y amplía los avances sobre métodos multivariantes no lineales que hasta la fecha se encontraban dispersos en publicaciones especializadas o que incluso eran de difusión limitada por formar parte de reportes internos de la Universidad de Leiden. Además de esta labor recopilatoria y de ampliación, que de por sí constituye ya un invaluable aporte, el gran mérito del texto en cuestión consiste en la integración que hace de los métodos multivariantes no lineales, mostrándolos como componentes de un sistema, lo que explica el hecho de que Jan de Leeuw se refiera posteriormente a este conjunto de técnicas como **sistema Gifi de análisis multivariante no lineal** (De LEEUW, 1984), designación que fue incorporada en la siguiente versión del texto de Gifi (GIFI, 1990) y que ha seguido utilizándose 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 sistémico usado por estos autores⁸ para estudiar las técnicas 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 estarán referenciados en espacios multidimensionales, lo que dificulta su representación y, por tanto, su interpretación. Uno de los principales objetivos de

⁷ Son numerosos los investigadores que han realizado aportaciones a este proyecto. En el prefacio de la edición 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, Renée 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 oración, los verbos se conjugarán en la tercera persona del singular.

muchas técnicas multivariantes es la reducción de la dimensionalidad, con el fin de facilitar la representación e interpretación de tales sistemas, con la menor pérdida posible de información.

En este contexto aparece el concepto de **homogeneidad**, que conlleva la idea de que diferentes variables comparten información o redundan en la medición de ciertos aspectos (GIFI, 1990). A mayor información compartida, mayor será la homogeneidad de ese grupo de variables y más eficientes serán las técnicas de reducción de dimensionalidad. Nótese que este concepto no tiene nada que ver con la definición de sistemas multivariantes homogéneos (cf. § 1.2). Todas las alusiones a la homogeneidad que se hacen en este capítulo tienen que ver con la cantidad de información común del sistema, tal y como acaba de presentarse.

Si se considera una matriz X con n individuos (u objetos, según la terminología utilizada en el sistema Gifi) en las filas y m variables en las columnas, es posible reducir la dimensionalidad del sistema reemplazando todas las columnas de X por un vector y , elegido de manera que contenga la mayor cantidad de información disponible en X . Esto equivale a reemplazar 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 elección de y , de manera que contenga la mayor cantidad de información posible, podría pensarse en la solución más básica; esto es, obtener y como el vector promedio de los vectores columna de la matriz X . GIFI (1990) muestra que esta solución solo sería aceptable si todas las variables estuvieran altamente correlacionadas entre sí y que, en general, es posible obtener una solución que capte más información.

La acepción más utilizada para el término **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 método lineal clásico de reducción de dimensionalidad por antonomasia, es decir, en el Análisis 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 promedio⁹ de las variables previamente escaladas o transformadas linealmente, de manera que cada cual tenga una ponderación acorde con la cantidad de información que comparte con las demás variables. Las transformaciones que garantizan que el vector resumen, o vector de puntuaciones de los individuos, recoja la máxima información posible, se obtienen usualmente con base en la descomposición en valores y vectores propios de la matriz $X'X$.

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

2.3 ESCALAMIENTO ÓPTIMO

Puesto que en la práctica, los sistemas multivariantes nunca tienen máxima homogeneidad (las variables pueden compartir información, pero no todas miden exactamente lo mismo), la reducción de la dimensionalidad siempre conlleva pérdida de información, la cual puede medirse a través una **función de pérdida** (*loss function*). Las funciones más 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, así:

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

Donde:

$\sigma(y, a)$: Función de pérdida en los parámetros y y a .

m : Número de variables.

SSQ : Suma de cuadrados.

y : Vector de puntuaciones de los individuos.

a_j : Ponderación o transformación aplicada a la j -ésima variable.

x_j : j -ésima columna de la matriz X ($n \times m$).

Una alternativa a la descomposición en valores y vectores propios de la matriz $X'X$ para la obtención del vector de puntuaciones y consiste en minimizar la función de pérdida [2.1]. Se denomina **escalamiento óptimo** a cualquier técnica generadora de transformaciones que minimicen una función de pérdida, es decir, que minimicen la pérdida de información que se produce al reducir la dimensionalidad del sistema multivariante, con base en un modelo dado.

Aunque esta definición de escalamiento óptimo, adaptada de GIFI (1990), tenga una apariencia un tanto diferente a la de la definición dada por YOUNG (1981) (cf. § 1.7), ambas hacen referencia al mismo mecanismo, solo que enfatizando aspectos diferentes. Mientras la definición de Young está enfocada en el tópico de las cuantificaciones, la

definición de Gifi se centra en la reducción de dimensionalidad. No obstante, si consideramos que las transformaciones a las que se hace referencia en la definición de Gifi constituyen una asignación de valores numéricos a las categorías, tal y como se indica en la definición de Young, podemos integrar ambos conceptos para presentar una definición acorde con nuestro propósito.

Definición 2.1 (Escalamiento Óptimo): Técnica que, mediante el uso de transformaciones, cuantifica las categorías de las variables de un sistema multivariante, de manera que se minimice la pérdida de información que resulta al reducir la dimensionalidad del sistema cuantificado, con base en un modelo de análisis dado.

Debe evitarse otorgar al apelativo *óptimo* un significado más allá del contexto al que pertenece. Obsérvese que este término 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 mínima pérdida de información.

Hay que anotar, sin embargo, que aunque las transformaciones solo pueden entenderse como óptimas bajo tal acepción y para el grupo particular de variables que conforma el sistema multivariante, las variables transformadas son utilizadas frecuentemente en un amplio rango de aplicaciones, más allá de las técnicas generadoras de las transformaciones.

En los dos apartados siguientes detallaremos los aspectos centrales de la definición 2.1. En el apartado 2.4 analizaremos el mecanismo que permite minimizar la pérdida de información resultante de reducir la dimensionalidad del sistema cuantificado, con base en un modelo de análisis dado. En el apartado 2.5 analizaremos las transformaciones usadas para la generación de cuantificaciones.

2.4 MÍNIMOS CUADRADOS ALTERNADOS

Como parte integral de las técnicas que conforman el sistema Gifi, el algoritmo computacional que más se ha utilizado para resolver problemas de escalamiento óptimo (minimizar funciones de pérdida con base en un modelo de análisis dado) es el de los mínimos cuadrados alternados, que usualmente se denomina mediante la sigla **ALS**, por sus iniciales en inglés: *Alternating Least Squares*. 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 óptimo mediante mínimos cuadrados alternados como programas **ALSOS**, sigla correspondiente a la expresión *Alternating Least Squares approach to Optimal Scaling*, que podría traducirse como Aproximación Mínimo Cuadrática al Escalamiento Óptimo.

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 mayorización (De LEEUW, 1994, 2006; VERBOON and HEISER, 1994; HEISER, 1995; LANGE et ál., 2000) que podrían ser más eficientes que el ALS en algunas situaciones particulares y que de hecho se utilizan en algunos programas específicos del sistema Gifi como el ANACOR que realiza Análisis de Correspondencias y el ANAPROF, que realiza Análisis de Frecuencias de Perfil cuando el número de individuos es mucho más grande que el número de perfiles.

En general, las funciones de pérdida tienen dos conjuntos de parámetros: 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 solución a través de ciclos. Cada ciclo consta de varias etapas: en una se estiman las puntuaciones óptimas para unos valores dados de las transformaciones, y en otra se estiman las transformaciones óptimas para unas puntuaciones dadas. Estas etapas se alternan, actualizando en cada una de ellas los valores de un conjunto de parámetros con base en los valores obtenidos en la etapa anterior para el otro conjunto. Esto genera una secuencia decreciente de valores para la función de pérdida, que converge a una constante a medida que los parámetros 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 $\leq \epsilon$).

Una de las versiones más sencillas de una función de pérdida es la presentada en [2.1], que es la correspondiente al ACP. Nótese que si en tal función se hiciera $y = 0$ y $a = 0$, se obtendría una solución trivial. Para evitarla se imponen restricciones sobre y o sobre a , de manera que tengan norma unitaria.

A continuación se muestran los pasos correspondientes al **algoritmo ALS con puntuaciones normalizadas**, para minimizar la función de pérdida [2.1].

0. Elegir un vector arbitrario de ponderaciones, a , de tamaño 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 .

$$y \leftarrow y(y'y)^{-1/2}$$

3. Actualizar el vector de ponderaciones, a .

$$a \leftarrow X'y$$

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

El algoritmo ALS con ponderaciones normalizadas sería análogo:

0. Elegir un vector arbitrario de puntuaciones, y , de tamaño 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 función de pérdida obtenida en esta iteración y la función de pérdida obtenida en la iteración anterior satisface el criterio de parada elegido por el usuario (diferencia $\leq \epsilon$), se da fin al proceso. En caso contrario, se vuelve al paso 1.

Una de las características 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 medición. Tal y como indica YOUNG (1981), siempre que exista un procedimiento de mínimos cuadrados para el ajuste de un modelo particular a datos cuantitativos, podrá usarse escalamiento óptimo mediante ALS para obtener un ajuste análogo de dicho modelo, usando datos cualitativos o una combinación de variables con diferentes niveles de escalamiento.

A continuación analizaremos el otro aspecto de la definición 2.1, es decir, el relativo a la generación de cuantificaciones mediante el uso de transformaciones.

2.5 ESCALAMIENTO, TRANSFORMACIONES Y CUANTIFICACIONES

Las técnicas del Sistema Gifi que aquí consideramos exigen que todas las variables sean discretizadas o categorizadas antes de ser procesadas mediante escalamiento óptimo. Por tal razón, siempre podremos hablar de las *categorías* de las variables, sin importar su escala de medición.

Tal y como se indicó en la definición 2.1, al someter un sistema multivariante a un proceso de escalamiento óptimo, la cuantificación de las categorías se genera mediante transformación de las variables. El tipo de transformaciones que se utilice para cuantificar una variable específica depende de la elección por parte del usuario del nivel de escalamiento de dicha variable (*scaling level*).

Aunque los conceptos de **nivel de escalamiento** y escala de medición (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 medición es una propiedad inherente al sistema de registro utilizado para plasmar la realidad estudiada. Después de presentar los niveles de escalamiento volveremos sobre este punto.

Los **niveles de escalamiento** básicos¹⁰ 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 numérico; transformaciones monótonamente ascendentes para las variables escaladas a nivel ordinal; y transformaciones isomórficas para las variables escaladas a nivel nominal (Van der BURG et ál. 1994).

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

¹⁰ Hablamos de niveles de escalamiento básicos, pues aunque en el Capítulo 5 definimos dos niveles de escalamiento adicionales, aquellos no son más que variaciones de los niveles de escalamiento que aquí se presentan.

En contraste con las transformaciones lineales, cualquier transformación que genere valores transformados no proporcionales a los originales será una **transformación no lineal**. Si se aplica una transformación no lineal y se representan los valores originales frente a los transformados en un plano cartesiano no se obtendrá una línea recta¹¹.

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 estadísticos de orden de la variable original coinciden con los estadísticos 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 función correspondiente es, por tanto, no decreciente.

Formando parte también de las transformaciones no lineales, se encuentran las **transformaciones isomórficas**, que se aplican al elegir el nivel de escalamiento nominal. Este tipo de transformación es el que conlleva menos restricciones; lo único que debe satisfacer es que a todas las observaciones correspondientes a una categoría se les asigne el mismo número real, sin que tenga que satisfacerse ninguna relación entre los valores asignados a diferentes categorías de una misma variable.

Las transformaciones descritas son funciones que, cuando se incorporan en un algoritmo ALS (o en cualquier otro algoritmo de minimización de funciones de pérdida), asignan puntuaciones normalizadas a cada una de las categorías de las variables. A este proceso se le denomina **cuantificación**, puesto que las puntuaciones así asignadas a cada una de las categorías, a diferencia de las etiquetas originales, tienen propiedades métricas, lo cual posibilita analizar la base de datos transformada mediante técnicas estándar del análisis multivariante, es decir, técnicas lineales que exijan datos numéricos.

¹¹ GIFÍ (1990) clasifica las técnicas del análisis multivariante en lineales, monótonas 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 técnica es no lineal si incluye transformaciones no lineales entre la batería de posibles transformaciones, aun cuando en algunos casos particulares no se haga uso de las mismas.

Dado que las variables numéricas tienen propiedades métricas aun antes de ser transformadas, lo más correcto sería denominar *escalamiento* al proceso de aplicar transformaciones sobre un sistema multivariante numérico, reservando el término *cuantificación* para hacer referencia al resultado de aplicar transformaciones sobre variables ordinales y/o nominales. No obstante, esta es una diferenciación netamente lingüística que no cambia en nada los resultados del proceso. Al aplicar transformaciones sobre sistemas multivariantes mixtos, usualmente se utiliza indistintamente cualquiera de los dos términos, pudiendo hablarse también de *escalamientos y/o cuantificaciones*.

La Figura 2.1 muestra el aspecto de tres gráficos de cuantificaciones típicos, obtenidos mediante transformaciones lineales, monótonamente ascendentes e isomórficas.

Es importante insistir en el hecho de que todas las variables transformadas mediante un programa ALSOS tienen propiedades métricas, cualquiera que sea el nivel de escalamiento elegido por el usuario. La única función del nivel de escalamiento es informar al proceso sobre las relaciones iniciales entre los

niveles o categorías de la variable, de manera que tales relaciones se transmitan a la variable transformada. Estas relaciones son análogas a las definidas para las escalas de medición (cf. § 1.2).

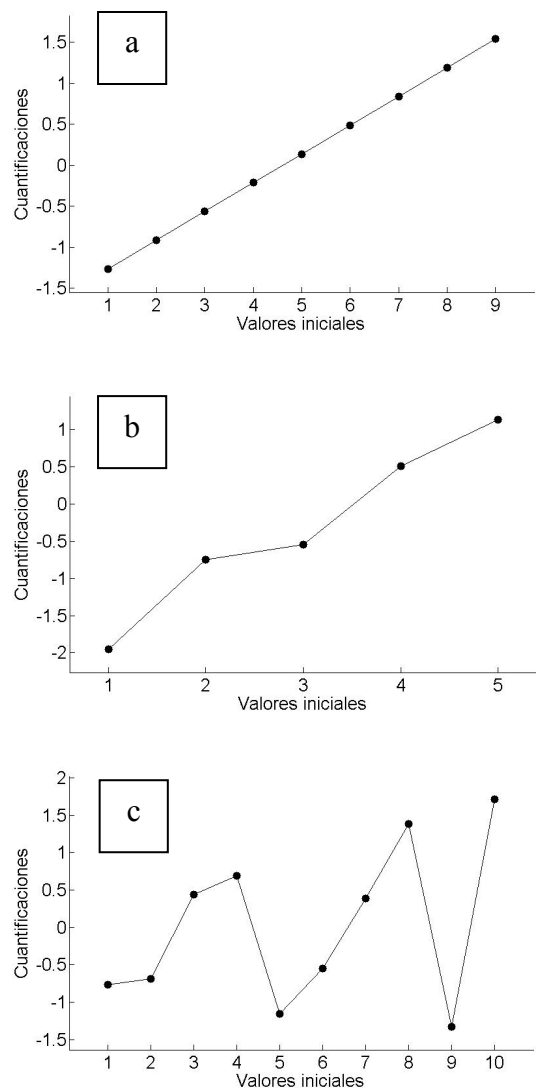


Figura 2.1. Gráficos de cuantificaciones típicos para una transformación (a) lineal, (b) monótonamente ascendente y (c) isomórfica.

En vista de la estrecha relación existente entre la escala de medición y el nivel de escalamiento, podría pensarse que el concepto de nivel de escalamiento es redundante y que bastaría con indicar la escala de medición de cada variable. Esto, sin embargo, no es así, debido a un par de aspectos que detallaremos a continuación, los cuales hacen útil el concepto de nivel de escalamiento y justifican su existencia como ente diferenciado de la escala de medición.

En primer lugar, las técnicas del sistema Gifi que aquí consideramos exigen que todas las variables sean discretizadas o categorizadas previa aplicación del escalamiento óptimo. Este paso de categorización puede conllevar pérdida de fortaleza de la escala con relación a la escala de medición de la variable original. Podría suceder, por ejemplo, que al categorizar una variable con escala de medición numérica como la edad, se obtuvieran categorías 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 medición de la variable edad fuera numérica, su nivel de escalamiento sería ordinal.

En segundo lugar, aun cuando la categorización no alterare la escala de medición de la variable original, el analista puede tener interés en realizar un escalamiento más flexible. Si en el contexto de las escalas de medición definíamos la fortaleza de la escala, acorde con la cantidad de información 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 más flexible, seguido del escalamiento ordinal, siendo el escalamiento numérico el menos flexible o el que más restricciones conlleva.

La utilización de un nivel de escalamiento más flexible que aquel que correspondería naturalmente a la escala de medición de la variable puede resultar útil 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 medición pueden no estar muy acordes con el papel desempeñado por la variable en el contexto estudiado.

Teniendo en cuenta los dos aspectos detallados anteriormente, consideramos perfectamente viable la utilización de un nivel de escalamiento que no corresponda con la escala de medición de la variable. No obstante, más allá de tales consideraciones, creemos que, en general, debe utilizarse el nivel de escalamiento que corresponda a la escala de medición de la variable. En tal sentido, somos más conservadores que MEULMAN et ál. (2005), quienes afirman que *“no existen propiedades intrínsecas de la variable que predefinan automáticamente un óptimo nivel de escalamiento que deba especificarse para la misma, pudiendo explorar los datos de cualquier forma que tenga sentido y que facilite la interpretación”*.

Puesto que el nivel de escalamiento definido por el usuario es el que en última instancia determina el procesamiento de cada variable, en adelante utilizaremos exclusivamente este concepto, sin ocuparnos más de la escala de medición. Así, a menos que indiquemos lo contrario, cuando hablemos de variables numéricas, ordinales o nominales, nos estaremos refiriendo al nivel de escalamiento elegido para las mismas.

2.6 HOMALS

Por extensión de la definición de sistemas multivariantes mixtos presentada en § 1.2 y con adaptación 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 pérdida más generales no están basadas en la matriz X , que contiene los valores de las variables, como se mostró en la expresión [2.1], sino en una matriz indicadora completa G , que tiene una columna para cada uno de los posibles valores de cada variable¹². A cada objeto se le registrará un 1 en la columna que corresponda al valor o categoría a la que pertenece y 0 en las demás columnas de la correspondiente variable (cf. § 1.6). Esto explica por qué estas técnicas exigen que todas las variables sean categorizadas.

Una función de pérdida más general que [2.1], la cual permite obtener una matriz de puntuaciones, Y ($n \times p$), en lugar de un vector¹³ está dada por la expresión [2.2]:

$$\sigma(Y, Q) \equiv m^{-1} \sum_j SSQ(Y - G_j Q_j) \quad [2.2]$$

Donde:

$\sigma(Y, Q)$: Función de pérdida en los parámetros Y y Q .

m : Número de variables.

SSQ : Suma de cuadrados.

Y : Matriz $n \times p$ de puntuaciones de los individuos.

G_j : Matriz indicadora correspondiente a la j -ésima columna de la matriz X .

Q_j : Matriz de cuantificaciones de las categorías de la j -ésima variable.

¹² Se tendrá 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 yuxtaposición del conjunto de matrices G_j .

¹³ Se reduce la dimensionalidad del sistema a p , con $p > 1$.

Para evitar la solución trivial que se obtendría cuando las matrices Y y Q_j son cero, se aplica alguna restricción o normalización. La más usual consiste en hacer que cada uno de los vectores de puntuaciones tenga norma 1.

La expresión [2.2], con la restricción de que todas las variables estén escaladas a nivel nominal es la función de pérdida del Análisis de Homogeneidad o HOMALS, acrónimo de la expresión inglesa **HOM**ogeneity Analysis by **Alt**ernating **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 técnica es equivalente al Escalamiento Óptimo, Puntuación Óptima y Puntuación Adecuada de la escuela americana (*Optimal Scaling, Optimal Scoring, Appropriate Scoring*); al Escalamiento Dual canadiense (*Dual Scaling*); al Análisis de Escalograma israelí (*Scalogram Analysis*); al método de cuantificación japonés (*Quantification Method*) y, por supuesto, al Análisis de Correspondencias Múltiples de la escuela francesa (*Multiple Correspondence Analysis*).

Estos autores (TENENHAUS and YOUNG, 1985) indican que en adición a la influencia que los diferentes países y lenguas han ejercido para presentar este método con diferentes nombres, está el enfoque seguido para su desarrollo, siendo posible distinguir cuatro corrientes o enfoques principales: el enfoque de los Promedios Recíprocos, el del Análisis de Componentes Principales, el del Análisis Canónico Generalizado y el del Análisis de Varianza.

El **enfoque de los Promedios Recíprocos** (*Reciprocal Averaging*) fue propuesto por Richardson y Kuder (1933, Citados por HORST, 1935) y FISHER (1940), al parecer de forma independiente, e impulsado por HILL (1973), quien lo presentó como un método de ordenación en ecología. GREENACRE (2007) señala que los Promedios Recíprocos constituye un algoritmo alternativo a la descomposición en valores singulares (DVS), para la obtención de la solución de un Análisis de Correspondencias (AC). Aunque este algoritmo no tiene la elegancia matemática de la DVS resulta muy útil para ilustrar las propiedades del Análisis de Correspondencias.

El algoritmo de los Promedios Recíprocos parte de cualquier conjunto de valores estandarizados como coordenadas iniciales de las categorías columna (podría iniciarse con las categorías fila). Seguidamente se obtienen las coordenadas de las categorías fila con base en los promedios ponderados de las correspondientes categorías fila. A continuación, se recalculan las coordenadas de las categorías columna con base en los promedios ponderados de las categorías fila y se aplica un nuevo paso de estandarización. El proceso continúa de forma iterativa hasta que se alcance la convergencia, generándose así las coordenadas del primer eje principal. NISHISATO (2004) presenta una ilustración numérica detallada de este algoritmo.

El **enfoque del Análisis de Componentes Principales** está basado en las ideas de HORST (1935) y BURT (1950, 1953) y ha sido desarrollado ampliamente por la escuela francesa (BENZÉCRI et ál., 1973; BENZÉCRI, 1977; GREENACRE, 1984). Se parte del hecho de que análogamente a la forma en que puede obtenerse la solución del Análisis de Componentes Principales (ACP) a partir de la Descomposición en Valores Singulares (DVS) de una matriz, X , de datos multivariantes, es posible obtener la solución del Análisis de Correspondencias a partir de la DVS simultánea de dos matrices de valores positivos (una para las categorías fila y otra para las categorías columna) que considere la métrica introducida por la asignación de ponderaciones a los perfiles y a las dimensiones, así: $D_p^{1/2} (X) D_d^{1/2}$, 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 solución del ACP.

El **enfoque del Análisis Canónico Generalizado** se remonta al trabajo de McKEON (1966). LECLERC (1980) presenta una síntesis del mismo. Bajo este enfoque, la solución se obtiene mediante la maximización de la suma de correlaciones cuadráticas 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 presentó una formulación completa del problema, así como su solución. 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 ál., 2005) y los trabajos de De LEEUW (1973) y Van RIJCKEVORSEL y De LEEUW (1978) que dan origen al HOMALS. Su formulación matemática es presentada por TENENHAUS y YOUNG (1985).

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

Una vez obtenidas las cuantificaciones para las diferentes categorías 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 categoría específica (suma de cuadrados de las desviaciones entre la puntuación de cada individuo y la puntuación promedio de cada una de sus categorías) y otra debida a la variabilidad de la puntuación promedio de tal categoría, con relación a la puntuación promedio general (suma de cuadrados de las desviaciones la puntuación promedio de la categoría y la puntuación promedio general). Seguidamente, resolvía el problema, maximizando la razón entre variabilidad entre puntuaciones promedio de las categorías y variabilidad total de las puntuaciones de los individuos.

Según WARRENS et ál. (2007), el método de Guttman, al que se refieren como Escalamiento Óptimo Clásico, permite una descomposición multidimensional de los datos, con la dimensión estructural más informativa apareciendo en primer lugar, luego la segunda y así sucesivamente hasta extraer exhaustivamente toda la información 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 solución HOMALS es equivalente a la del Análisis de Correspondencias Múltiples cuando todos los marginales fila de la matriz indicadora completa¹⁴ son iguales al número 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 práctico, la principal diferencia entre el ACM francés y el HOMALS radica en el hecho de que en adición a las cuantificaciones de las categorías de las variables, el HOMALS también genera puntuaciones para los individuos, las cuales pueden representarse conjuntamente con las cuantificaciones de las categorías.

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

2.7 OVERALS

En adición a la generalización 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 generalización adicionales. Por una parte, permite trabajar con variables escaladas a diferente nivel (numérico, ordinal o nominal) y, por otra parte, considera la posibilidad de que las variables estén estructuradas en K grupos. Por tal motivo, se le conoce también como Análisis de Correlación Canónica no Lineal.

La función de pérdida está dada por la expresión [2.3]:

$$\sigma(Y, Q_j) \equiv K^{-1} \sum_k SSQ \left(Y - \sum_{j \in J_k} G_j Q_j \right) \quad [2.3]$$

Donde:

$\sigma(Y, Q_j)$: Función de pérdida en los parámetros Y y Q_j .

K : Número de grupos.

SSQ : Suma de cuadrados.

Y : Matriz de puntuaciones de los individuos.

J_k : Conjunto de índices de las variables pertenecientes al k -ésimo grupo ($k=1, 2, \dots, K$).

G_j : Matriz indicadora correspondiente a la j -ésima variable del k -ésimo grupo.

Q_j : Matriz de cuantificaciones de las categorías de la j -ésima variable del grupo k -ésimo, satisfaciendo las restricciones propias de cada nivel de escalamiento.

El algoritmo OVERALS utiliza mínimos cuadrados alternados para minimizar la función de pérdida [2.3], admitiendo que los conjuntos puedan diferir tanto en el número de variables que los conforman como en la definición 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 minimización de la función de pérdida [2.3] se requiere obtener simultáneamente el parámetro relacionado con el modelo de correlación canónica (Y) y el parámetro de escalamiento, que surge a partir de las transformaciones (Q_j). El problema OVERALS se resuelve alternativamente para cada uno de los parámetros, manteniendo el otro constante. Dados unos valores para el parámetro 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 iteración. La solución para los diferentes parámetros es discutida con detalle por Van der BURG et ál. (1994).

El OVERALS fue concebido por De LEEUW (1973), pero su presentación 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 ejecución (Van der BURG et ál., 1988; Van der BURG et ál., 1994).

El OVERALS constituye el modelo más general del sistema Gifi, por cuanto permite trabajar con variables escaladas a cualquier nivel y conformando cualquier número de grupos. En tal sentido, todos los demás modelos pueden verse como casos particulares de este; de ahí la primera parte de su nombre (OVER). La segunda parte del acrónimo (ALS) corresponde al algoritmo utilizado para su solución (*Alternating Least Squares*).

Los modelos que incluye el OVERALS, y que pueden verse como casos particulares de este, son PRINCALS (Análisis de Componentes Principales no Lineal), que surge cuando cada grupo está conformado por una sola variable, es decir, cuando no hay agrupamiento. Si en este caso, todas las variables se escalan a nivel numérico, se tiene el ACP; si todas las variables se escalan a nivel nominal, se tendría el HOMALS; si en el caso anterior se tienen solo dos variables, se llega al ANACOR (Análisis de Correspondencias) (MEULMAN et ál., 2004). Adicionalmente, el modelo CORALS (Análisis de Correlación Canónica no Lineal) es un caso especial de OVERALS, que aparece cuando $K=2$. Si en este caso todas las variables se escalan a nivel numérico, se

tiene el Análisis de Correlación Canónica (ACC). Las anteriores relaciones se resumen en la Figura 2.2.

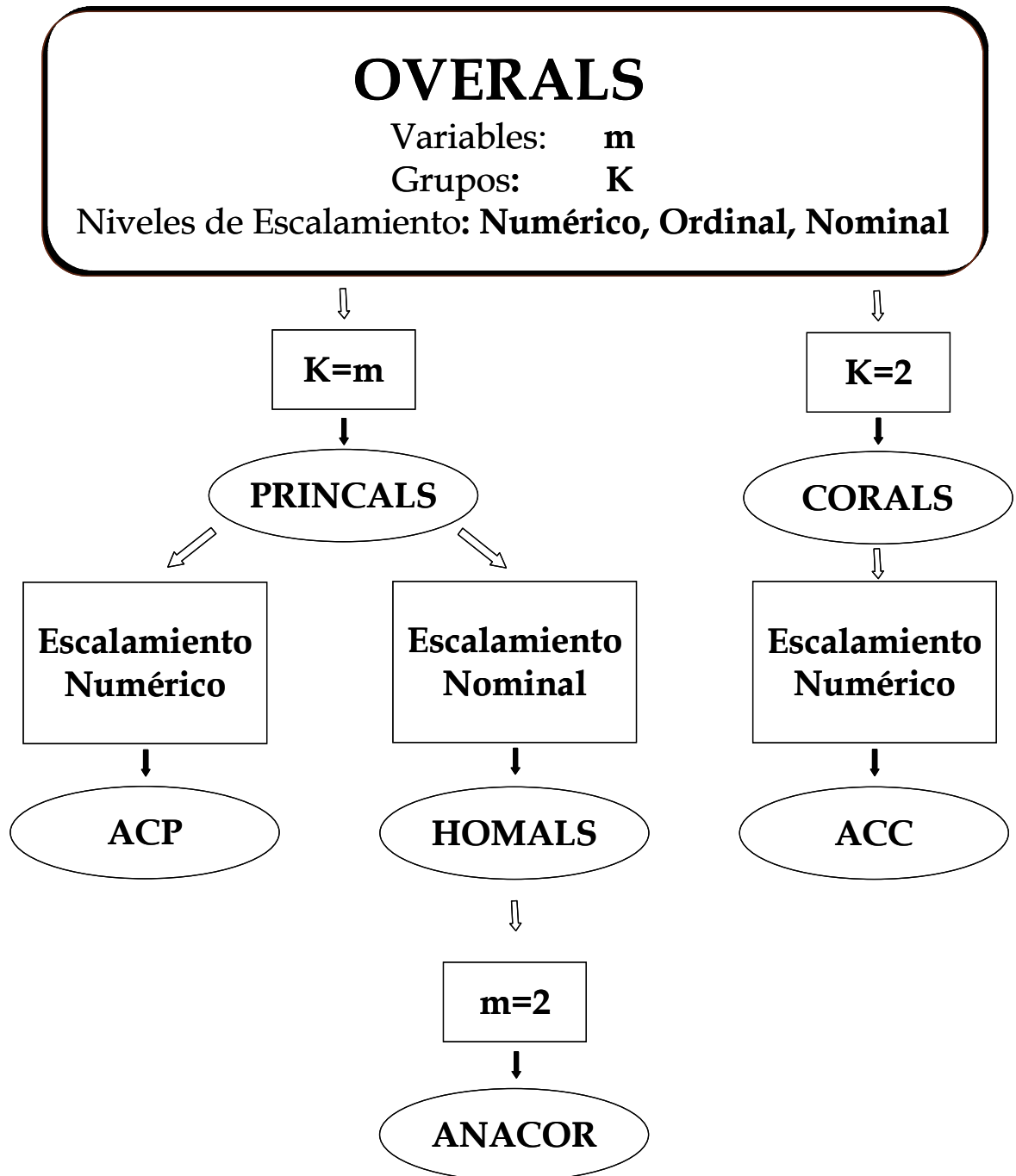


Figura 2.2. Relación entre el OVERALS y otras técnicas del sistema Gifi.

CAPÍTULO 3

EFECTO DE LA DIMENSIONALIDAD SOBRE LAS CUANTIFICACIONES

3.1 INTRODUCCIÓN

Tras la breve presentación en el Capítulo 2 del sistema Gifi de análisis multivariante no lineal y de su papel en la generación de cuantificaciones, **en este capítulo** se mostrará cómo el modelo utilizado y, en particular, la dimensionalidad de la solución afectan las cuantificaciones obtenidas.

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

Tras definir las cuantificaciones simples y las cuantificaciones múltiples, se ilustra mediante un ejemplo hipotético de pacientes con cáncer de laringe el efecto de la dimensionalidad de la solución sobre las cuantificaciones simples.

3.2 DIMENSIONALIDAD DE LAS SOLUCIONES

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

- 1) Servir como primer paso para la obtención de transformaciones y/o cuantificaciones que permitan la subsecuente aplicación de técnicas lineales.
- 2) Generar representaciones en subespacios de baja dimensionalidad que resuman las relaciones existentes entre objetos, variables y/u objetos y variables.

Una cuestión omnipresente al usar técnicas de reducción de dimensionalidad de sistemas multivariantes es la elección de p , la dimensión del subespacio de representación. Puesto que las técnicas de reducción de dimensionalidad son técnicas de resumen, es válido, en general, el criterio aplicable a cualquier resumen: que sea lo menor posible, recogiendo lo más esencial del todo resumido. No obstante, esta es una recomendación demasiado general, que debe precisarse con base en otras consideraciones.

GIFI (1990) argumenta que $p=1$ sería la elección natural cuando el énfasis del investigador es hacia el primer objetivo; mientras que si el énfasis está en la satisfacción del segundo objetivo, frecuentemente se elige $p=2$.

Cuando el objetivo es la cuantificación como primer paso para el posterior uso de otras técnicas, consideramos, en contraste con lo anotado por GIFI (1990) sobre la ‘natural’ elección 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 elección frecuente no significa que esta sea la ‘mejor’ elección. En ocasiones puede ser deseable obtener un mayor número de dimensiones que permitan explorar relaciones pobremente reflejadas en el primer plano principal (desde luego, a través del análisis de los demás planos). A la incertidumbre que conlleva

la elección de p en cualquier técnica de reducción de dimensionalidad, habrá que añadirle 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 elección de p podrían presentar aun más 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 técnicas del Sistema Gifi que se basan en la minimización de la función de pérdida del encuentro (*meet*), esto es, Análisis de Componentes Principales no Lineal (PRINCALS) y Análisis de Correlación Canónica no Lineal (OVERALS), pueden ser múltiples o simples, a elección del usuario (GIFI, 1990).

Una **cuantificación múltiple** significa que cada categoría de una variable recibe un número p de cuantificaciones igual a la dimensionalidad de la solución. En este caso, en que la relación entre la dimensionalidad de la solución y las cuantificaciones obtenidas es directa, se opta, como práctica general, por un valor bajo de p , dado que se está en el marco de técnicas de *reducción* de dimensionalidad.

Cuando se trabaja con **cuantificaciones simples**, cada categoría recibe una única cuantificación, sin importar el valor de p , hecho que puede enmascarar la relación existente entre la dimensionalidad elegida y las cuantificaciones obtenidas y, por consiguiente, brindar una sensación 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 múltiples (GIFI, 1990). Para las variables que el usuario defina como simples, el paso de las cuantificaciones múltiples a las simples se obtiene imponiendo **restricciones de rango uno**; esto es, proyectando la matriz de cuantificaciones múltiples 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 expresión [3.1]:

$$y_{j=} \frac{Y_j c_j}{c'_j c_j} \quad [3.1]$$

Donde:

y_j : Cuantificaciones simples de la j -ésima variable

Y_j : Matriz $K_j \times p$, con las cuantificaciones múltiples de la j -ésima variable

c_j : Vector de saturaciones de la j -ésima variable

La anterior expresión muestra que en la generación de las cuantificaciones simples de cada variable, participan todas las dimensiones de la solución múltiple en proporción a la magnitud de la saturación de la variable en la correspondiente dimensión. Luego, cuando se opta por cuantificaciones simples basadas únicamente en la primera dimensión ($p=1$), se está prescindiendo de la información contenida en las demás dimensiones, lo cual afecta especialmente a las variables cuyas principales saturaciones no están en la primera dimensión. 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 dimensión se cuantifiquen adecuadamente.

3.4 EFECTO DE LA DIMENSIONALIDAD SOBRE LAS CUANTIFICACIONES

Para ilustrar el efecto de la dimensionalidad de la solución sobre las cuantificaciones, considérese un estudio hipotético sobre calidad de vida en pacientes tratados por cáncer de laringe, en el que se evalúan, entre otras, las siguientes cuatro variables:

A: Dolor

- a1: No hay dolor
- a2: Dolor leve que no requiere medicamentos
- a3: Dolor moderado controlable con analgésicos suaves
- a4: Dolor severo solo controlable con codeína/morfina
- a5: Dolor intenso que no se controla con ningún medicamento

B: Actividad

- b1: Está tan activo como siempre
- b2: Ocasionalmente no puede mantener su anterior nivel de actividad
- b3: Frecuentemente se siente cansado y más 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 sólidos suaves, pero no puede masticar algunas comidas
- c3: No puede masticar ni siquiera sólidos suaves
- c4: No tolera ningún alimento por vía oral

D: Habla

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

Se simuló una muestra multivariante de 15 variables y 1.000 observaciones, la cual fue procesada mediante Análisis de Componentes Principales Categórico (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 interés, manteniendo la presentación lo más 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 categorías de las cuatro variables en cuestión, basadas en las soluciones con $p=1$ y $p=2$.

Variable	Categoría	Cuantificación	
		$p=1$	$p=2$
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 categorías de las variables *A*, *B*, *C* y *D*, basadas en soluciones con $p=1$ y $p=2$.

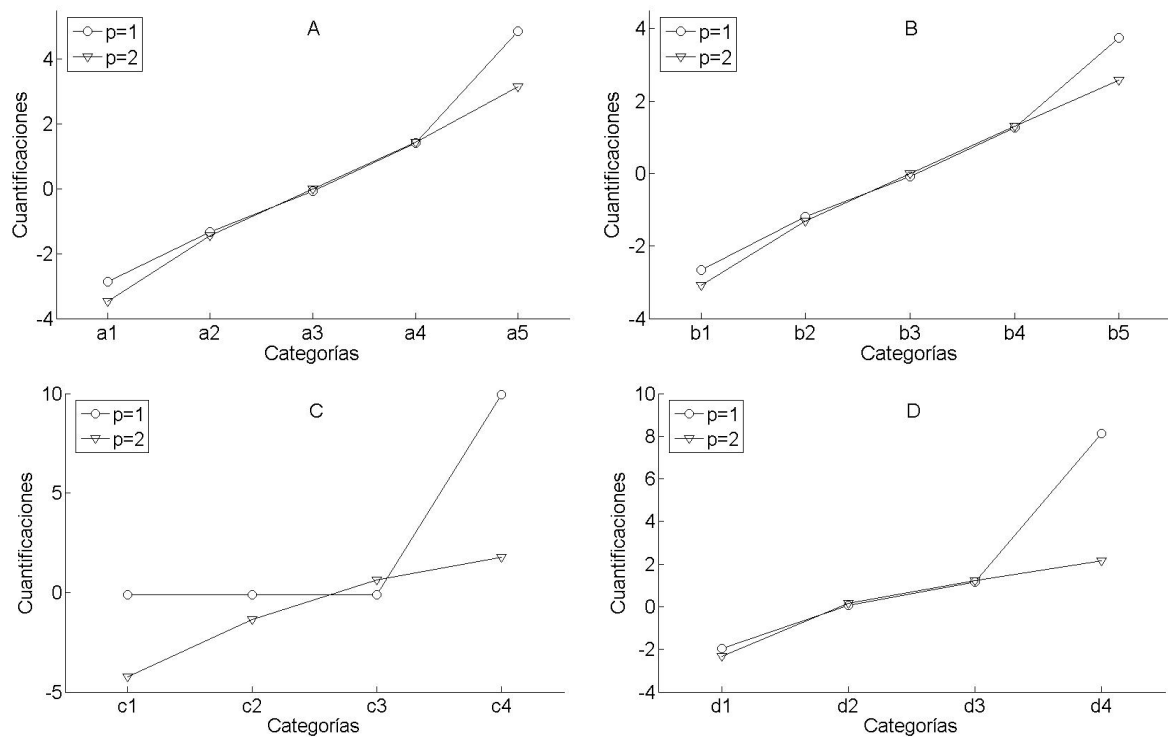


Figura 3.1. Cuantificaciones para las categorías de las variables *A*, *B*, *C* y *D*, basadas en soluciones con $p=1$ y $p=2$.

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	$p=1$	$p=2$	
	Dimensión 1	Dimensión 1	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 con $p=1$ y $p=2$.

Tal y como ya se había 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 dimensión 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 solución con $p=1$ y la solución con $p=2$, puede observarse que los cambios son más drásticos cuanto menor es la saturación de la variable en la primera dimensión de la solución con $p=1$.

Obsérvese, por ejemplo, que las cuantificaciones de las categorías de las variables A y B son las más estables. Esto es debido al bajo aporte que hace la segunda dimensión a estas variables y a que estas variables tienen una alta saturación en la primera dimensión en la solución con $p=1$.

En contraste, las cuantificaciones de las categorías de la variable C son las más incongruentes, lo cual se explica por el hecho de la saturación casi nula que tiene dicha variable sobre la primera dimensión en la solución con $p=1$. Nótese, por ejemplo, que la solución basada en $p=1$ indica que las categorías c1, c2 y c3 de la variable C no se diferencian entre sí, esto es, que para fines de caracterización de la calidad de vida en pacientes con cáncer de laringe es lo mismo que el paciente pueda masticar tan bien como antes; que solo pueda masticar sólidos suaves, o que no pueda masticar ni siquiera sólidos suaves, diferenciándose del conjunto anterior solo el grupo de pacientes que no tolera ningún alimento por vía oral. Por otra parte, la solución basada en $p=2$ asigna diferentes cuantificaciones a cada una de las categorías de la variable *Masticación*, 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 categorías de la variable D, aunque son mucho más drásticos 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 saturación en la primera dimensión, en la solución con $p=1$, no es tan baja como la de la variable C.

Estas incongruencias en las cuantificaciones se trasladan a cualquier técnica basada en las mismas, generando diferentes resultados, según 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 interés), entre las variables cuantificadas con base en $p=1$ y con base en $p=2$. Nótese que la mayor incongruencia es la de la correlación lineal entre las variables C y D (*Masticación* y *Habla*), las dos con menor saturación en la primera dimensión 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 $p=1$ y $p=2$.

Para ilustrar con mayor detalle la forma en que las incongruencias en las cuantificaciones pueden trasladarse a otras técnicas produciendo, a su vez, resultados incongruentes, dependiendo de si se trabaja con uno u otro conjunto de cuantificaciones, supóngase que se desea construir un índice de calidad de vida de los pacientes con cáncer de laringe, basado en las cuatro variables descritas. Puesto que tras la cuantificación, todas las variables tienen propiedades métricas, puede utilizarse un índice basado en la puntuación de los individuos sobre la primera componente principal de un ACP. Para facilitar su interpretación, 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 hipotéticos y sus respectivos índices 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. Índices de calidad de vida (ICV) para dos hipotéticos pacientes con cáncer de laringe, basados en sendos ACP para las cuantificaciones PRINCALS con $p=1$ y $p=2$.

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

Este ejemplo no hace más que poner en evidencia lo que puede deducirse del análisis de la expresión [3.1]: **las soluciones basadas en diferentes valores de p generan diferentes cuantificaciones**. Resulta claro además que la solución obtenida con base en $p=1$ solo resultaría adecuada si todas las variables tuvieran una alta saturación sobre la primera dimensión. En tal sentido no consideramos que esta sea la solución ‘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 dimensión 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 técnicas del sistema Gifi para un sistema multivariante mixto dependen de la dimensionalidad de la solución utilizada para su generación.

Consideramos que este problema ha pasado desapercibido debido a que las técnicas del sistema Gifi, además de generar cuantificaciones, sirven por sí mismas para analizar los datos sin que se requiera utilizar las cuantificaciones generadas como datos de entrada de otras técnicas. En tal sentido, se da la situación que ya anotábamos en el Capítulo 1 (cf. § 1.7) cuando nos referíamos al hecho de que a menudo las cuantificaciones no constituyen un fin en sí mismas, sino que forman parte integral de la técnica, apareciendo como un subproducto del proceso o llegando incluso a ser totalmente ignoradas.

En tales casos, cuando el objetivo fundamental es la representación en un espacio de baja dimensionalidad, el analista elige p con base en las consideraciones habituales o en alguna modificación 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 razón para creer que estas consideraciones puedan resultar adecuadas cuando el analista tiene interés en obtener el mejor conjunto de escalamientos y/o cuantificaciones simples de un grupo de variables, que le permitan analizarlas mediante técnicas lineales. Por el contrario, consideramos inadecuado limitar la técnica generadora de cuantificaciones con base en restricciones dirigidas a satisfacer otros objetivos.

Podemos afirmar, en términos generales, que cuando se usan las técnicas del sistema Gifi con el objetivo de generar escalamientos y/o cuantificaciones de las variables para su posterior uso en otras técnicas lineales, debe utilizarse la dimensionalidad p que genere cuantificaciones óptimas.

Como ya se anotó anteriormente, el calificativo *óptimo* que acompaña los escalamientos generados mediante las técnicas del sistema Gifi solo debe entenderse como tal en el sentido en que minimiza la pérdida de información inherente a la reducción de la dimensionalidad del sistema original. Nótese que si el objetivo no es reducir la dimensionalidad, sino solamente obtener cuantificaciones que puedan utilizarse posteriormente en otras técnicas, podría ser más conveniente partir de una definición de *óptimo* acorde con tales objetivos.

En resumen, el problema en el que nos centraremos y cuya solución constituye la base de los desarrollos que se presentan en los dos siguientes capítulos, 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 ningún criterio para establecer cuál de los conjuntos de cuantificaciones resultantes es el óptimo.

CAPÍTULO 4

CUANTIFICACIÓN ÓPTIMA

4.1 INTRODUCCIÓN

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

Empezamos presentando una definición general de cuantificación óptima que no depende de ningún modelo de análisis. Esta definición, así como todos los desarrollos que se desprenden de la misma son aportes originales de este trabajo. Seguidamente, se utilizan un par de ejemplos hipotéticos para ilustrar la lógica subyacente en la definición, los cuales han sido cuidadosamente elegidos para recrear algunos de los escenarios más comunes. Aunque es posible seguir la solución 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 ningún otro lugar del texto. Invitamos, por tanto, a seguir el desarrollo de la propuesta tal y como está presentado.

Basada en la definición general de cuantificación óptima, se presenta luego una definición específica, la cual, aunque se representa en términos matemáticos, sigue sin depender de ningún modelo. Seguidamente, se desarrolla una propuesta específica para resolver el problema planteado en el capítulo anterior. Esta propuesta consiste en elegir la solución que mejor satisfaga la condición de cuantificación óptima, acorde con los criterios presentados. Se discute el alcance de la propuesta y se bosqueja su implementación algorítmica.

Finalmente, se ilustra cómo aplicar la solución planteada, sin avanzar en el análisis contextualizado de los datos, pues en el Capítulo 6 analizaremos una aplicación completa con base en el desarrollo general que se expone en el Capítulo 5.

4.2 DEFINICIÓN GENERAL DE CUANTIFICACIÓN ÓPTIMA

Tomando como escenario de referencia aquel en el que se usan las técnicas 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 óptimo que acompaña tales escalamientos solo debe entenderse como tal en el sentido en que minimiza la pérdida de información que se produce al reducir la dimensionalidad del sistema original con base en un modelo de análisis dado, presentaremos una definición de cuantificación óptima más acorde con nuestros objetivos.

Definición 4.1 (Cuantificación Óptima, general): 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.

Nótese que los algoritmos en los que se basan las técnicas 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 número determinado de *categorías*. Es por ello que la anterior definición habla indistintamente de *categorías*, sin que importe el nivel de escalamiento de las variables.

La definición 4.1 constituye la base de todos los desarrollos posteriores. Seguidamente, ilustraremos la idea subyacente a través de un par de aplicaciones hipotéticas antes de presentar su formulación matemática.

4.3 ILUSTRACIÓN DE LA DEFINICIÓN GENERAL

EJEMPLO ILUSTRATIVO 1.

Considérese una base de datos hipotética, constituida por la talla en centímetros 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 centímetros de un grupo de 25 varones y 25 mujeres.

Como primer paso de cualquier proceso de cuantificación óptima mediante las técnicas que estamos considerando, todas las variables deben ser categorizadas¹⁶. Supóngase que la variable talla se categoriza así: (150-154), (155-159), ..., (175-179), (180-184). La Tabla 4.2 muestra las frecuencias de las combinaciones de las categorías de las variables talla y sexo.

Ahora que las dos variables están categorizadas, procede la cuantificación de sus correspondientes categorías. Para tal efecto, no nos basaremos en la definición de cuantificación óptima, sino que realizaremos una aproximación intuitiva.

¹⁶ En § 1.5 se discuten algunos criterios de categorización.

Talla	Sexo	
	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 numérica de la variable talla, la cuantificación lógica de cada una de sus categorías estará dada por su punto medio o marca de clase. Este vector de cuantificaciones, o cualquier transformación lineal del mismo, es el que mejor recoge la información relativa a las interdistancias entre cada uno de los niveles o categorías de la variable talla, respetando sus propiedades métricas. En aras de la simplicidad, no realizaremos por el momento ninguna normalización y cuantificaremos las categorías 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 único referente para la cuantificación de las categorías de la variable sexo es la variable talla (o la cuantificación de sus categorías), la elección natural consiste en cuantificar cada categoría con base en la talla promedio de los individuos en tal categoría. En vista de que las categorías de la variable talla ya han sido cuantificadas, cada una de las categorías de la variable sexo se cuantifica usando la siguiente expresión:

$$q_{S_l} = \frac{\sum_{k=1}^K f_{kl} q_{t_k}}{\sum_{k=1}^K f_{kl}}, \quad l = V, M \quad [4.1]$$

Donde:

qs_l : Cuantificación de la categoría l de la variable sexo ($l = V, M$).

qt_k : Cuantificación de la k -ésima categoría de la variable talla (marca de clase).

f_{kl} : Frecuencia de aparición conjunta de la k -ésima categoría de la variable talla y la categoría l de la variable sexo.

Luego, las cuantificaciones para las categorías V y M de la variable sexo, calculadas mediante la expresión [4.1], son: $qs_V = 172,6$ y $qs_M = 164$.

Obsérvese que este conjunto de cuantificaciones, que es sin duda el más natural para este ejemplo, se ha obtenido de manera absolutamente intuitiva, no habiendo forzado su obtención al cumplimiento de la definición 4.1. No obstante, puede verificarse que el razonamiento seguido conduce a un conjunto de cuantificaciones que satisface la condición especificada en la definición de cuantificación óptima; esto es, asigna valores más similares a las categorías de las dos variables que se asocian con mayor frecuencia.

Así, por ejemplo, al comparar la cuantificación de la categoría V ($qs_V = 172,6$) con las cuantificaciones de las categorías de la variable talla, se observa que a la que más se acerca es a la de la categoría 170-174 ($qt_5 = 172$), siendo esta la categoría de la variable talla a la que la categoría V se asocia con mayor frecuencia ($f_{5V} = 11$). De igual modo, al comparar la cuantificación de la categoría M ($qs_M = 164$) con las cuantificaciones de las categorías de la variable talla, encontramos que a la que más se acerca es a la de la categoría 160-164 ($qt_3 = 162$), siendo esta la categoría de la variable talla a la que la categoría M se asocia con mayor frecuencia ($f_{3M} = 12$).

Lógicamente, la similitud a la que hace referencia la definición de cuantificación óptima puede entenderse en términos de distancia (CUADRAS, 1991), lo que nos permite avanzar en el análisis de su cumplimiento. En un sistema multivariante cuantificado óptimamente, las distancias entre las categorías cuantificadas de las diferentes variables serán menores para aquellas categorías que se asocien con mayor frecuencia. En tal sentido, cabe esperar que al sumar, sobre todas las observaciones, las distancias entre las categorías cuantificadas óptimamente, se obtenga una cifra inferior a la que se obtendría con cualquier otro sistema de cuantificaciones.

Observando la expresión [4.1], resulta evidente que siempre que se cuantifique cada categoría de la variable sexo con base en la media de la variable talla para tal categoría, se obtendrá una cuantificación cercana a la de la(s) categorías(s) de la variable talla más frecuentes para la correspondiente categoría 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 cuadráticas entre las cuantificaciones de las categorías de la variable talla y las de las categorías de la variable sexo.

Supóngase que qs_l es la cuantificación buscada para una categoría determinada de la variable sexo ($l=V$ o $l=M$), qt_k es la cuantificación de la k -ésima categoría de la variable talla y f_{kl} es la frecuencia de aparición conjunta de las dos categorías consideradas. Si se define D como la suma sobre todas las observaciones de las distancias al cuadrado entre las cuantificaciones de las categorías de las dos variables en cuestión, tenemos:

$$\begin{aligned}
 D &= \sum_{i=1}^n (qt - qs)^2 \\
 &= \sum_{l=l} (qt - qs_l)^2 + \sum_{l=l'} (qt - qs_{l'})^2 \\
 &= \sum_{k=1}^K (f_{kl} (qt_k - qs_l)^2) + \sum_{k=1}^K (f_{kl'} (qt_k - qs_{l'})^2) \\
 &= \sum_{k=1}^K (f_{kl} qt_k^2 - 2f_{kl} qt_k qs_l + f_{kl} qs_l^2) + \sum_{k=1}^K (f_{kl'} qt_k^2 - 2f_{kl'} qt_k qs_{l'} + f_{kl'} qs_{l'}^2) \\
 &= \left(\sum_{k=1}^K f_{kl} qt_k^2 - 2qs_l \sum_{k=1}^K f_{kl} qt_k + qs_l^2 \sum_{k=1}^K f_{kl} \right) + \left(\sum_{k=1}^K f_{kl'} qt_k^2 - 2qs_{l'} \sum_{k=1}^K f_{kl'} qt_k + qs_{l'}^2 \sum_{k=1}^K f_{kl'} \right)
 \end{aligned}$$

La derivada de esta expresión, con respecto a qs_l es:

$$\frac{dD}{dq_{s_l}} = -2 \sum_{k=1}^K f_{kl} qt_k + 2qs_l \sum_{k=1}^K f_{kl}$$

Esta derivada se hace cero cuando:

$$\sum_{k=1}^K f_{kl} q t_k = q s_l \sum_{k=1}^K f_{kl}$$

$$\Rightarrow q s_l = \frac{\sum_{k=1}^K f_{kl} q t_k}{\sum_{k=1}^K f_{kl}}$$

Este resultado para $q s_l$ es exactamente igual al obtenido mediante la expresión [4.1], es decir, a la media de las cuantificaciones de las categorías de la variable talla asociadas con la correspondiente categoría de la variable sexo. Luego, queda demostrado que la cuantificación de una categoría cualquiera de la variable sexo obtenida como la media de las cuantificaciones de las categorías de la variable talla observadas para la correspondiente categoría de la variable sexo, minimiza la suma sobre todas las observaciones de las distancias cuadráticas entre las cuantificaciones de las categorías de la variable talla y las de las categorías de la variable sexo.

Nótese además que la cuantificación de las categorías de una variable con base en la cuantificación promedio de las categorías de la otra variable sigue el mismo principio que el método de los Promedios Recíprocos (*Reciprocal Averaging*; cf. § 2.6).

Este ejemplo ilustra cómo **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 más similares a las categorías de las distintas variables que se asocian con mayor frecuencia.

EJEMPLO ILUSTRATIVO 2.

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

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

Para la condensación de las variables numéricas, podría pensarse en la simple utilización de su media aritmética. No obstante, podemos reducir la pérdida de información fruto de tal proceso, mediante un escalamiento o transformación previa de las variables. GIFÍ (1990) indica que dicha pérdida de información se minimiza al remplazar el sistema numérico original por su primera componente principal. Esta sería, por tanto, la elección lógica. La variable numérica condensada tendría la siguiente forma:

$$Y_1 = v_1 (tallaest) + v_2 (pesoest) \quad [4.2]$$

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 definición 2.1 (§ 2.3), puesto que Y_1 resume las variables talla y peso con mínima pérdida de información, Y_1 es el promedio de las variables talla y peso escaladas óptimamente. Se satisface, por tanto, la siguiente relación:

$$v_1 (tallaest) + v_2 (pesoest) = \frac{tallaopt + pesoopt}{2}$$

Donde, $tallaopt$ y $pesoopt$ corresponden a las variables talla y peso escaladas óptimamente, las cuales pueden expresarse en términos de las variables centradas y estandarizadas, así:

$$talla_{opt} = 2v_1 (talla_{est})$$

$$peso_{opt} = 2v_2 (peso_{est})$$

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

$$qx_{jk} = mv_j(x_{jk}), \quad k = 1, 2, \dots, K; \quad j = 1, 2, \dots, m \quad [4.3]$$

Donde:

- qx_{jk} : Cuantificación del k -ésimo nivel o categoría de la j -ésima variable.
- m : Número de variables.
- v_j : Constante multiplicativa de la j -ésima variable en la combinación lineal que genera la primera componente principal.
- x_{jk} : k -ésimo nivel centrado y estandarizado de la j -ésima variable.

Consideremos, ahora, la perspectiva de PEARSON (1901) para la obtención de la primera componente principal de un conjunto de variables, según la cual, la primera componente principal es la línea que minimiza la suma de proyecciones cuadráticas de los puntos de la nube multidimensional. En el caso de dos variables continuas, esto es fácilmente visualizable como una línea que atraviesa el diagrama de dispersión en la dirección de su tendencia principal o máxima dispersión, haciendo mínima la suma de proyecciones cuadráticas de los puntos a la misma, como se muestra en la Figura 4.1.

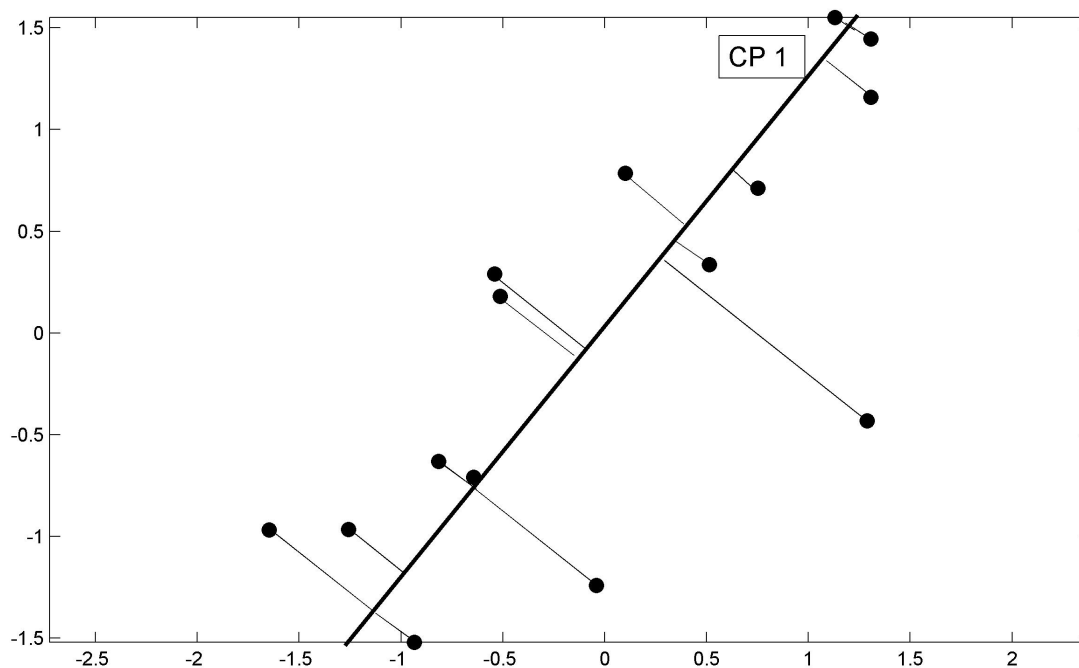


Figura 4.1. Primera componente principal construida a partir de la minimización de la suma de proyecciones cuadráticas de los puntos de la nube multidimensional.

Al trabajar con variables discretizadas, la visualización ya no es tan directa, puesto que el diagrama de dispersión consistirá 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 dispersión modificado, en el que el tamaño de cada punto sea proporcional a su frecuencia.

Considérese la hipotética tabla de contingencia para las variables talla y peso (Tabla 4.3) y su diagrama de dispersión modificado asociado (Figura 4.2). Para facilitar la visualización de la correspondencia entre la tabla de contingencia y el diagrama de dispersión, se ha invertido el orden usual de las filas en la tabla de contingencia, presentándolas desde los niveles más altos de la variable talla hasta los más bajos.

En este caso, a diferencia de la Figura 4.1, no es fácil visualizar la tendencia principal del diagrama de dispersión. No obstante, es claro que la dirección de línea que satisfaga la condición de minimizar la suma de proyecciones cuadráticas se verá afectada por los puntos de mayor ponderación (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.

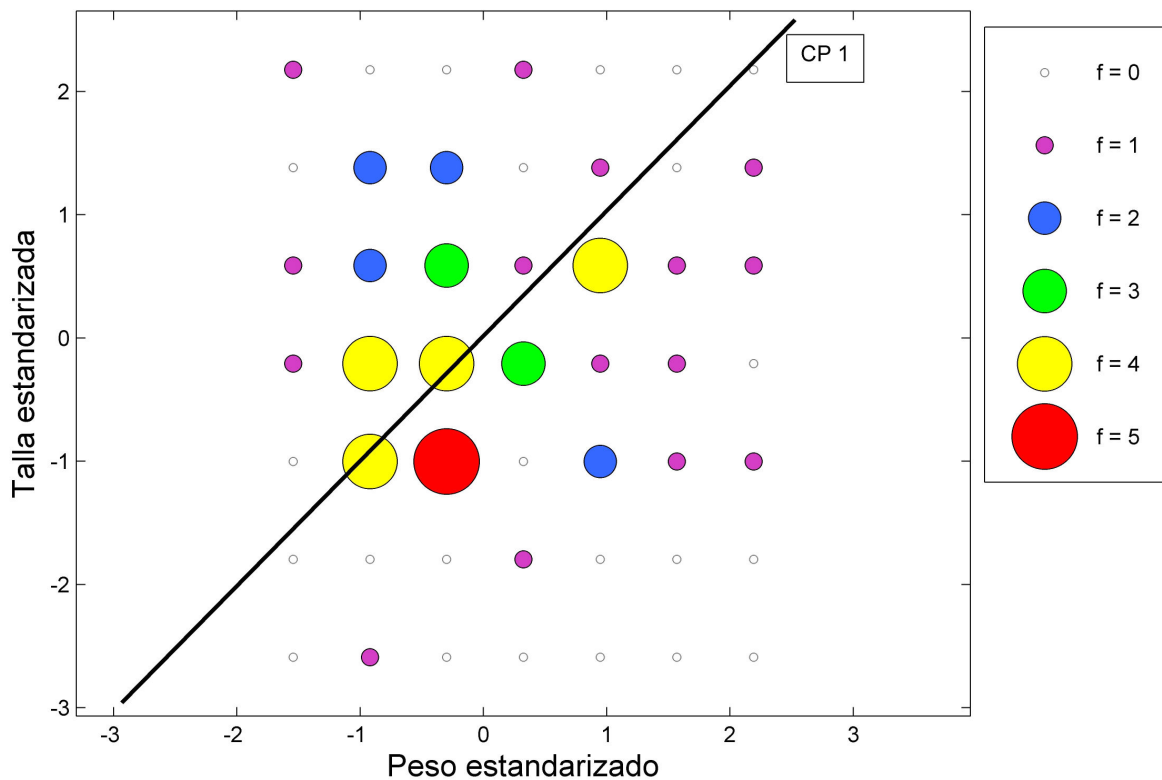


Figura 4.2. Diagrama de dispersión modificado para peso y talla estandarizados, y primera componente principal.

Existe una clara relación entre la información común de un sistema multivariante numérico continuo y su primera componente principal. La primera componente principal de un sistema multivariante es la combinación lineal de las variables originales que mejor recoge la información común contenida en estas. A mayor información

común compartida por el conjunto de variables, más cercanas se encontrarán éstas de la primera componente principal. Esta cercanía, en el caso de variables continuas, puede medirse en una representación Biplot (GABRIEL, 1971; GALINDO, 1985, 1986; GOWER and HAND, 1996; YAN and KANG, 2003) en términos del ángulo formado entre la componente principal y las correspondientes variables.

Es posible generalizar el concepto de información común y cercanía a la primera componente principal en sistemas multivariantes con variables discretizadas. En este caso, nos interesa discutir la cercanía de combinaciones específicas de categorías de diferentes variables (puntos de la nube multidimensional) a la componente principal.

Obsérvese que las categorías que con mayor frecuencia se presentan conjuntamente conforman puntos de alta ponderación, 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 incluyéndolos como puntos suplementarios en el Análisis de Componentes Principales.

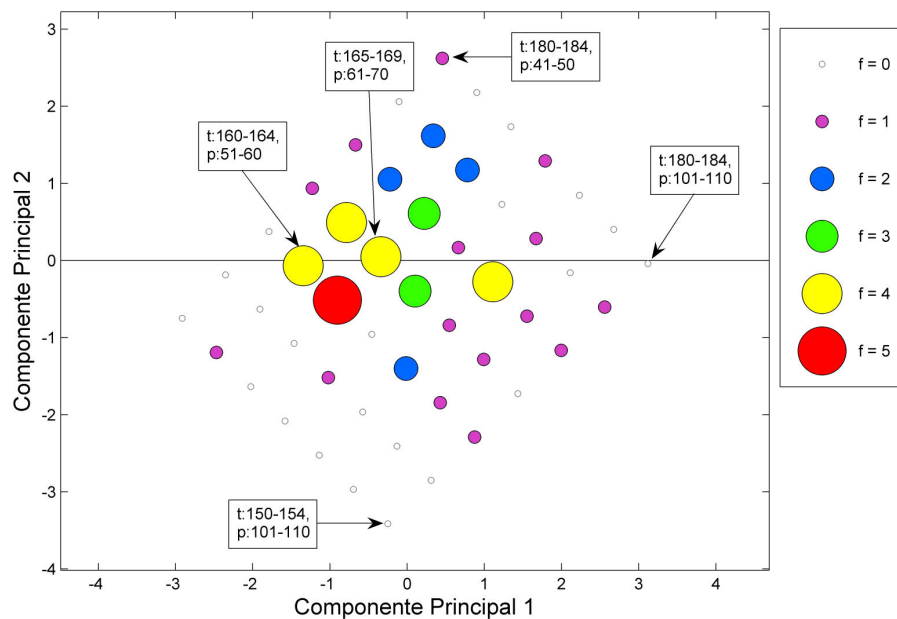


Figura 4.3. Representación de las observaciones de talla y peso discretizadas en el plano de sus componentes principales.

Análogamente a la forma en que la cercanía de un grupo de variables continuas a la primera componente principal (en términos de los ángulos) indica que estas comparten información común, la cercanía de un punto a la primera componente principal indica que las categorías representadas por dicho punto tienen cuantificaciones similares. En el presente ejemplo, podemos verificarlo mediante el análisis 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 expresión [4.3], las cuantificaciones para cada una de las categorías de las variables talla y peso; y la Figura 4.4, en la que se representan simultáneamente sobre una línea 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 categorías de las variables talla y peso.

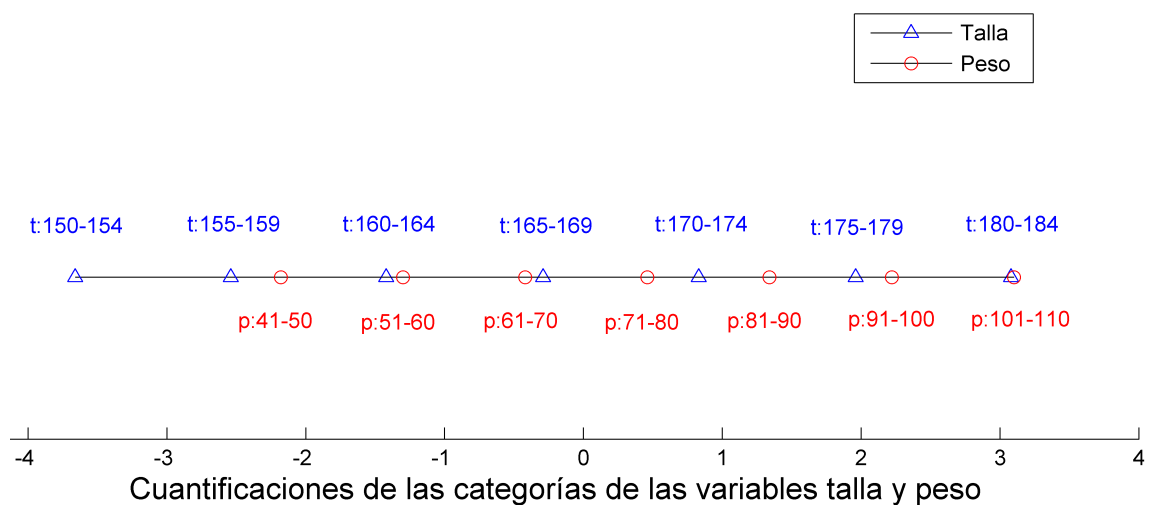


Figura 4.4. Dispersión conjunta de las categorías cuantificadas de talla y peso.

Nótese que los puntos que representan las combinaciones de categorías con cuantificaciones más similares ($\{t:180-184, p:101-110\}$, $\{t:160-164, p:51-60\}$, $\{t:165-169, p:61-70\}$) son los más cercanos a la primera componente principal (Figura 4.3), mientras que los puntos que representan las combinaciones de categorías con cuantificaciones más disímiles ($\{t:150-154, p:101-110\}$, $\{t:180-184, p:41-50\}$) son los más 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 más cercano esté un punto a la primera componente principal, más similares serán las cuantificaciones de las categorías representadas por dicho punto. En particular, cuando un punto se encuentre exactamente sobre la primera componente principal, la cuantificación de las categorías coincidirá. A continuación demostraremos el cumplimiento de dicha condición para el caso bidimensional.

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

$$R \equiv \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$$

Donde $r = \text{corr}(X_1, X_2)$.

Las componentes principales del sistema se obtienen con base en las soluciones de la siguiente ecuación:

$$\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \lambda \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \quad [4.4]$$

Donde λ es un valor propio de R y $\underline{v}' = (v_1 \ v_2)$ es su correspondiente vector propio.

Los valores propios, λ , se obtienen como soluciones de la siguiente ecuación:

$$\begin{aligned}
 \begin{vmatrix} (1-\lambda) & r \\ r & (1-\lambda) \end{vmatrix} &= (1-\lambda)^2 - r^2 \\
 &= 1 + \lambda^2 - 2\lambda - r^2 \\
 &= \lambda^2 - 2\lambda + (1-r^2)
 \end{aligned}$$

Se resuelve esta ecuación de segundo grado en λ :

$$\lambda = \frac{-(-2) \pm \sqrt{2^2 - 4(1)(1-r^2)}}{2(1)}$$

$$\lambda = \frac{2 \pm \sqrt{4r^2}}{2}$$

$$\lambda = \frac{2 \pm 2r}{2}$$

$$\lambda = 1 \pm r$$

Si $r > 0$, entonces $(1+r) > (1-r)$, por lo que $\lambda_1 = (1+r)$ será el primer valor propio y $\lambda_2 = (1-r)$ será el segundo valor propio. Si $r < 0$, el orden de los valores propios será inverso, esto es, $\lambda_1 = (1-r)$ y $\lambda_2 = (1+r)$.

Suponiendo que $r > 0$ y reemplazando el primer valor propio en la ecuación [4.4], se obtiene el primer vector propio, \underline{v}_1 :

$$\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = (1+r) \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

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

$$rv_1 + v_2 = (1+r)v_2$$

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

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

$$v_1 - v_1 - rv_1 = -rv_2$$

$$-rv_1 = -rv_2$$

$$v_1 = v_2 = v$$

$$\therefore \underline{v}_1' = v(1 \ 1)$$

Siguiendo con el supuesto de que $r > 0$ y remplazando el segundo valor propio en la ecuación [4.4], se obtiene el segundo vector propio, \underline{v}_2 :

$$\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = (1-r) \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

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

$$rv_1 + v_2 = (1-r)v_2$$

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

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

$$v_1 - v_1 + rv_1 = -rv_2$$

$$rv_1 = -rv_2$$

$$v_1 = -v_2$$

$$\therefore \underline{v}_2' = v(1 \ -1)$$

En resumen, cuando $r > 0$, el primer vector propio es $\underline{v}_1' = v(1 \ 1)$, y el segundo vector propio es $\underline{v}_2' = v(1 \ -1)$.

Por un razonamiento análogo, cuando $r < 0$, el primer vector propio es $\underline{v}_1' = v(1 \ -1)$, y el segundo vector propio es $\underline{v}_2' = v(1 \ 1)$.

Luego, las componentes principales de este sistema se expresan así:

$$Y_1 = \begin{cases} vX_1 + vX_2, & \text{si } r > 0 \\ vX_1 - vX_2, & \text{si } r < 0 \end{cases} \quad [4.5]$$

$$Y_2 = \begin{cases} vX_1 - vX_2, & \text{si } r > 0 \\ vX_1 + vX_2, & \text{si } r < 0 \end{cases}$$

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 -ésima categoría de X_1 y la k' -ésima categoría de X_2 , tal distancia se obtiene a partir de la expresión [4.5], así:

$$d(\{x_{1k}, x_{2k'}\}, \text{CP1}) = \begin{cases} v|x_{1k} - x_{2k'}|, & \text{si } r > 0 \\ v|x_{1k} + x_{2k'}|, & \text{si } r < 0 \end{cases}$$

Con base en las expresiones [4.3] y [4.5], pueden obtenerse las cuantificaciones de la k -ésima categoría de X_1 y la k' -ésima categoría de X_2 . Tales cuantificaciones son $2vx_{1k}$ y $\pm 2vx_{2k'}$, respectivamente. La diferencia entre tales cuantificaciones está dada por la siguiente expresión:

$$d(qx_{1k}, qx_{2k'}) = \begin{cases} 2v|x_{1k} - x_{2k'}|, & \text{si } r > 0 \\ 2v|x_{1k} + x_{2k'}|, & \text{si } r < 0 \end{cases}$$

Así 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 podrá cambiar acorde con la normalización 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 cuantificación más similar en este ejemplo sea la exhibida por las categorías t:180-184 y p:101-110 (3,08 y 3,10, respectivamente), las cuales no se presentaron simultáneamente en ninguna de las 50 observaciones (frecuencia cero) parece contradictorio con lo afirmado anteriormente. Esta situación se explica, sin embargo, por la restricción propia de los escalamientos lineales. Si bien es cierto que los puntos de mayor ponderación son los que tienden a quedar más cercanos a la primera componente principal, en una técnica lineal como el ACP no es posible evitar que algunos puntos sin “meritos” suficientes también queden cercanos a esta. Para corregir esta situación sería necesario asignar las cuantificaciones con base en una técnica no lineal que impusiera menos restricciones.

Este ejemplo ilustra cómo **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 más similares a las categorías o niveles de las variables que coincidan con mayor frecuencia.

Finalmente, para cerrar este apartado, antes de pasar a la formulación matemática de la cuantificación óptima, supóngase que se tiene un sistema multivariante nominal y que se desea cuantificar cada una de sus categorías. La solución clásica viene dada mediante la cuantificación de estas a partir de las coordenadas de un Análisis de Correspondencias (AC), en caso de tener solo dos variables, o de un Análisis de Correspondencias Múltiples (ACM) o un Análisis de Correspondencias Conjunto (*Joint Correspondence Analysis*, GREENACRE, 2007)¹⁷, en caso de tener más de dos variables.

¹⁷ El Análisis 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 únicamente las submatrices por fuera de la diagonal. El ACCj hace una mejor generalización del AC que el ACM, en términos de inercias totales, coincidiendo exactamente con el AC cuando se tienen dos variables.

Si bien es cierto que en estas técnicas no es adecuado interpretar directamente la distancia entre los puntos categoría de diferentes variables, sí es posible establecer una interpretación conjunta de los mismos con relación a los ejes principales de la representación (GREENACRE, 1994). En el caso de una solución unidimensional, que sería la de nuestro interés para la obtención de cuantificaciones simples, todas las categorías estarían sobre una línea, siendo posible interpretar directamente las distancias entre las categorías de diferentes variables, tal y como lo ilustra GREENACRE (2007). En general, las categorías asociadas con mayor frecuencia en la tabla de contingencia, aparecerán más cercanas sobre el primer eje factorial.

Así, 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 así la condición especificada en la definición 4.1.

Habiendo ilustrado la lógica subyacente en la definición 4.1 y su satisfacción en diferentes escenarios, presentamos seguidamente su formulación matemática, con base en la cual desarrollaremos posteriormente una solución al problema planteado en el Capítulo 3.

4.4 DEFINICIÓN ESPECÍFICA DE CUANTIFICCIÓN ÓPTIMA

Usando el paralelismo existente entre los conceptos de similaridad y distancia (CUADRAS, 1991), podemos reformular la definición de cuantificación óptima de manera más específica, así:

Definición 4.2 (Cuantificación Óptima, específica): 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.

Para el caso general de m variables, esta condición se expresa así:

$$QO(Q_{X_1}, Q_{X_2}, \dots, Q_{X_m}) \equiv \min \sum_{i=1}^n \sum_{j=1}^{m-1} \sum_{j'>j}^m (q_{ij} - q_{ij'})^2 \quad [4.6]$$

Donde:

$QO(Q_{X_1}, Q_{X_2}, \dots, Q_{X_m})$: Cuantificación óptima para las categorías de las variables X_1, X_2, \dots, X_m .

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

q_{ij} : Cuantificación de la categoría de la j -ésima variable correspondiente a la i -ésima observación.

$q_{ij'}$: Cuantificación de la categoría de la j' -ésima variable correspondiente a la i -ésima observación.

Los dos sumatorios más internos en la expresión [4.6] evalúan para una observación particular (la i -ésima) todas las distancias cuadráticas entre sus categorías cuantificadas, imponiendo la restricción 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 más 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 \sqrt{n} , siendo n el número de observaciones. Esta normalización hace que las cuantificaciones obtenidas para los sistemas multivariantes menos complejos, esto es, aquellos en los que todas las variables se escalan numéricamente, coincidan con las obtenidas mediante las técnicas 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 óptimas 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. Bastará con calcular para cada solución la suma sobre todas las observaciones de las distancias al cuadrado entre las correspondientes categorías cuantificadas, y elegir aquella solución para la cual dicha cifra sea mínima. Con las definiciones 4.1 y 4.2 como referente, esta será la mejor solución del conjunto, por cuanto será la que mejor satisfaga que las cuantificaciones de las categorías de las diferentes variables que se asocian con mayor frecuencia estén más cercanas entre sí.

Para una observación específica, sea q_j el valor de la categoría cuantificada correspondiente a la j -ésima variable. Las interdistancias cuadráticas entre las categorías cuantificadas de las diferentes variables está dada por:

$$\begin{aligned}
 IDC &= \sum_{j=1}^{m-1} \sum_{j'>j}^m (q_j - q_{j'})^2 \\
 &= \frac{1}{2} \sum_{j=1}^m \sum_{j'=1}^m (q_j - q_{j'})^2 \\
 &= \frac{1}{2} \sum_{j=1}^m \sum_{j'=1}^m (q_j^2 - 2q_j q_{j'} + q_{j'}^2) \\
 &= \frac{1}{2} \left[\sum_{j=1}^m \sum_{j'=1}^m q_j^2 - 2 \sum_{j=1}^m \sum_{j'=1}^m q_j q_{j'} + \sum_{j=1}^m \sum_{j'=1}^m q_{j'}^2 \right] \\
 &= \frac{1}{2} \left[m \sum_{j=1}^m q_j^2 - 2 \sum_{j=1}^m q_j \sum_{j'=1}^m q_{j'} + m \sum_{j'=1}^m q_{j'}^2 \right]
 \end{aligned}$$

Puesto que $\sum_{j=1}^m q_j = \sum_{j'=1}^m q_{j'}$ y $\sum_{j=1}^m q_j^2 = \sum_{j'=1}^m q_{j'}^2$, podemos escribir esta expresión como:

$$IDC = \frac{1}{2} \left[2m \sum_{j=1}^m q_j^2 - 2 \left(\sum_{j=1}^m q_j \right)^2 \right]$$

$$= m \sum_{j=1}^m q_j^2 - \left(\sum_{j=1}^m q_j \right)^2$$

Y dado que la ‘varianza’, S^2 , de la observación en cuestión es:

$$S^2 = \frac{m \sum_{j=1}^m q_j^2 - \left(\sum_{j=1}^m q_j \right)^2}{m(m-1)}$$

Podemos expresar la suma de interdistancias cuadráticas como $m(m-1)S^2$, lo cual facilita los cálculos numéricos.

Luego, la suma de las interdistancias cuadráticas de todas las observaciones es $m(m-1) \sum_{i=1}^n S_i^2$, donde S_i^2 es la varianza de las diferentes categorías cuantificadas que conforman la i -ésima observación. Esta ‘varianza’ no debe confundirse con la varianza de cada una de las variables, la cual, en virtud de la normalización impuesta, es igual a n , el número de observaciones.

Para fines de la comparación que nos atañe, podemos remplazar la constante $m(m-1)$ por cualquier otra constante, pues no existe ningún interés particular en la magnitud de la suma de las interdistancias cuadráticas; basta con saber cuál es la solución para la cual dicha suma es mínima. Con el fin de mantener los valores numéricos bajos, el criterio usado en este trabajo es el indicado mediante la expresión [4.7].

$$\text{Interdistancias} = \frac{1}{n} \sum_{i=1}^n S_i^2 \quad [4.7]$$

Donde:

Interdistancias: Criterio escalado de suma de interdistancias cuadráticas.

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

S_i^2 : ‘Varianza’ de la i -ésima observación.

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

No está de más insistir en el hecho de que la ‘varianza’ que aparece en la expresión [4.7] no es la varianza de las variables, sino la varianza de las observaciones.

La expresión [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 Capítulo 3, así como en los algoritmos que sustentan la propuesta generalizada de cuantificación óptima que se desarrolla en el Capítulo 5.

4.6 CONSIDERACIONES SOBRE LA PROPUESTA

La presente propuesta permite escoger el conjunto óptimo 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 obtención de la dimensionalidad p de la solución que da lugar a las cuantificaciones óptimas. Debe observarse, sin embargo, que la obtención de la dimensionalidad p es solo un paso intermedio para la obtención de las cuantificaciones óptimas 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, podría usarse para tal efecto, pero sería necesario realizar una valoración 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 generación del conjunto inicial de cuantificaciones es circunstancial. Este criterio podría usarse para elegir la mejor cuantificación simple de una serie de cuantificaciones generadas por cualquier medio, sin importar si el proceso de generación de las cuantificaciones involucra diferentes dimensionalidades o no.

Es importante resaltar el hecho de que la solución propuesta, al estar basada en las definiciones 4.1 y 4.2 solo toma en consideración las relaciones existentes entre los datos, sin basarse en ningún modelo. Tal y como indica NISHISATO (2007), el uso de un modelo puede ser bueno para el propósito de capturar un tipo especial de información de los datos, pero en ausencia de un conocimiento total sobre los datos, un modelo puede hacer que una gran cantidad de información contenida en los datos se quede sin analizar. Asimismo, este procedimiento está acorde con las ideas de Jean-Paul Benzécri (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 solución para la cual estas sean mínimas.

Los pasos de la rutina son:

1. Captación de información general.
2. Lectura de información específica proveniente del SPSS.
3. Cálculo de las *Interdistancias* para cada una de las soluciones.
4. Presentación de resultados.

A continuación se presentan los detalles básicos de cada uno de los pasos, omitiendo aquellos que solo tengan interés desde el punto de vista computacional.

Por generalidad, nos referiremos al caso en que se tengan varias soluciones OVERALS. Desde luego, la rutina también es aplicable si se tienen varias soluciones de cualquiera de sus casos particulares (cf. § 2.7).

Para ejecutar la aplicación 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 gráfica como la que se muestra en la Figura 4.5. Esta contiene tres casillas en las que el usuario deberá suministrar la información general que se detalla a continuación:

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

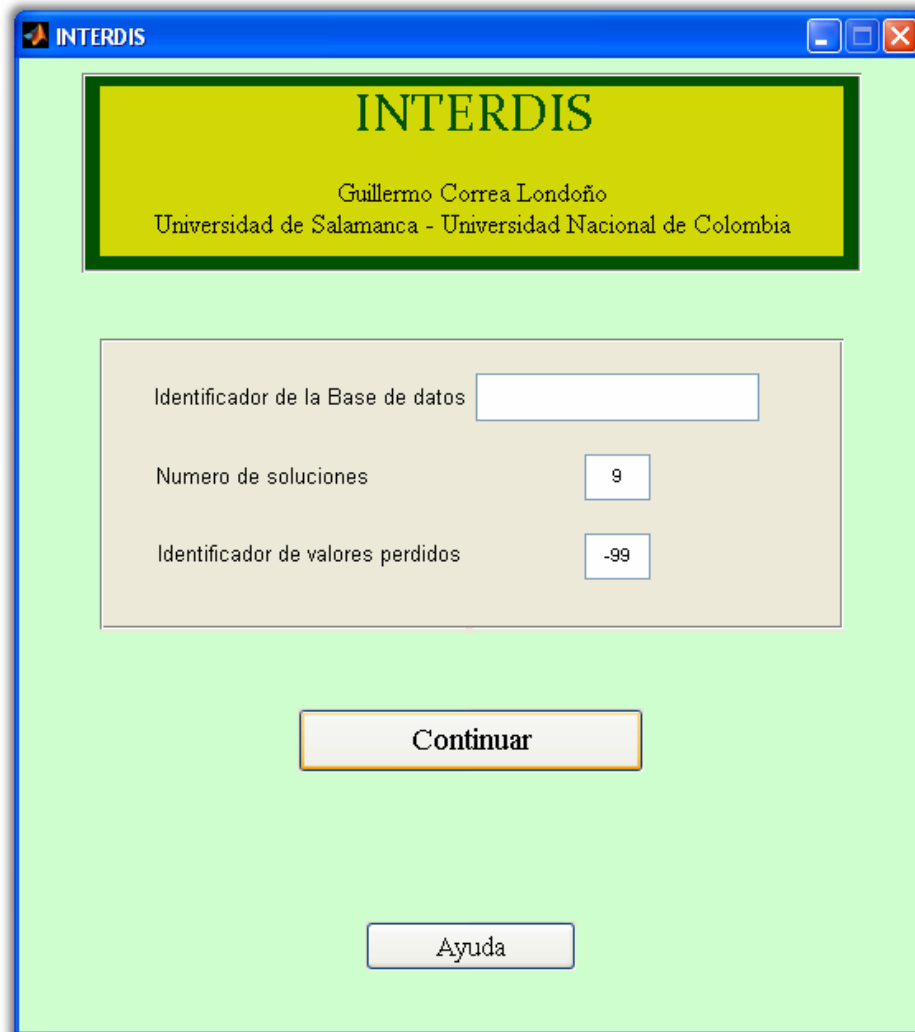


Figura 4.5. Interfaz gráfica de la rutina **INTERDIS**.

El identificador de la base de datos, al cual nos referiremos en adelante como '**Id**', puede ser cualquier cadena alfanumérica (números, letras o combinación de ambas). Este identificador constituye la raíz 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 número de soluciones que se especifique en la segunda casilla debe coincidir con la dimensionalidad p de la máxima solución. Así, por ejemplo, si se indica que el número de soluciones es 4, el sistema buscará las soluciones para $p=1$, $p=2$, $p=3$ y $p=4$. Si el usuario no modifica esta casilla, el sistema buscará por defecto 9 soluciones.

Si la base de datos tiene información faltante, es necesario utilizar un código numérico 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 modificación.

4.7.2 LECTURA DE INFORMACIÓN ESPECÍFICA.

Previa ejecución del programa, para comparar p soluciones, es necesario disponer $p+2$ archivos de entrada en formato ASCII (extensión **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) Información para las cuantificaciones (1 archivo para cada solución: p archivos)

La matriz de datos originales está conformada por la parte numérica de la base de datos utilizada para la obtención de las cuantificaciones en SPSS (no contiene títulos para las variables). En caso de que haya valores perdidos, estos deberán señalarse con -99 o con cualquier otro código numérico que el usuario señale en la tercera casilla de la interfaz gráfica de captación de información 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 genérico **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 información sobre los niveles de las variables deberá estar disponible en la ruta de trabajo de MATLAB, con el nombre **nivluz.txt**. La información sobre los niveles de cada variable forma parte de las salidas del SPSS, cuya obtención y manejo se describen al final de este apartado.

Por cada solución que se desee evaluar se requerirá un archivo con información para el cálculo de las cuantificaciones. Cada uno de estos archivos tendrá el nombre genérico: **C+Id+p+#.txt**, donde ‘**C**’ es la letra que siempre dará comienzo al nombre de estos archivos; ‘**Id**’ es el identificador de la base de datos (cf. § 4.7.1); ‘**p**’ es un carácter que siempre irá en esa posición, indicando la dimensionalidad de esa solución; ‘**#**’ es el número correspondiente a la dimensionalidad p . Así, 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 deberán 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 más que por pantalla, para evitar su transcripción manual, estas se calculan a partir de las coordenadas simples y las ponderaciones (*weights*), las cuales sí pueden generarse como parte de los resultados del OVERALS. Para tal efecto, se utiliza la siguiente expresión.

$$q_j = \frac{Y_j w_j}{w_j' w_j} \quad [4.8]$$

Donde,

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

Para que el SPSS genere un archivo de resultados con los niveles de cada variable y con la información necesaria para calcular las cuantificaciones mediante la expresión [4.8], debe incluirse la instrucción *MATRIX=OUT(nombrearchivo.sps)* en la correspondiente sintaxis del OVERALS. El archivo que se genera tiene un formato propio del SPSS (extensión **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 aplicación, o bien puede guardarse una copia con formato de Excel para realizar el ordenamiento y copiado en esta otra aplicación. En cualquier caso, debe tenerse en cuenta que el separador decimal de la información que se copie a los archivos ASCII que serán leídos luego por MATLAB es el punto, sin importar cuál sea la configuración regional del ordenador. Antes de copiar la información sobre los niveles de las variables y los datos para el cálculo 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 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 número de niveles de la j -

ésima variable y m el número de variables. Esta información se pega en un archivo ASCII que se guarda en la ruta de trabajo de MATLAB con el nombre genérico **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 información correspondiente a las dimensiones de *SCOOR_* y *WEIGHT_*, tal y como quedan tras el ordenamiento (se copia solo la información numérica correspondiente a cada una de las

p dimensiones). La información copiada deberá tener $\sum_{j=1}^m niv_j + m$ filas y p columnas, siendo niv_j el número de niveles de la j -ésima variable; m , el número de variables, y p , la dimensionalidad de la correspondiente solución. Esta información se pega en un archivo ASCII que se guarda con el nombre genérico **C+Id+p+#.txt**.

4.7.3 CÁLCULO DE LAS INTERDISTANCIAS.

Una vez calculadas las cuantificaciones simples para cada variable con base en la expresión [4.8], se construye para cada una de las p soluciones una matriz Q en la que cada categoría de cada variable es remplazada por su correspondiente cuantificación. Seguidamente, se calculan para cada una de las p soluciones las *Interdistancias*, con base en la expresión [4.7], y se determina cuál es la solución cuyas *Interdistancias* son mínimas.

4.7.4 PRESENTACIÓN DE RESULTADOS.

Finalmente, tras comparar los diferentes conjuntos de cuantificaciones, la aplicación **INTERDIS**, presenta un aviso por pantalla, indicando con cuál de las soluciones comparadas se obtienen las *Interdistancias* mínimas.

Asimismo, se construye un gráfico con las dimensionalidades de las diferentes soluciones comparadas en la abscisa, 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 solución con *Interdistancias* mínimas. En el caso de la hipotética base de datos 'luz', este archivo tendrá el nombre **Qluz.xls**. En caso de que la base de datos original tenga datos perdidos (marcados con -99 o con cualquier otro código numérico), estos aparecerán como celdas vacías en la matriz de cuantificaciones. Las cuantificaciones contenidas en dicho archivo pueden utilizarse como entrada de cualquier técnica lineal.

4.8 ILUSTRACIÓN DE LA PROPUESTA EN UNA APLICACIÓN PRÁCTICA

A continuación se muestran los resultados de aplicar la rutina **INTERDIS** a una base de datos real conformada por 48 variables y 5.443 pacientes con cardiopatía isquémica, sobre la cual se realizó un análisis OVERALS. Los detalles sobre la base de datos y los resultados OVERALS aparecen en CORREA (2006). No profundizaremos en el análisis ni de la base de datos ni de los resultados, pues el único objetivo en esta parte es ilustrar el funcionamiento de la rutina **INTERDIS**. En el Capítulo 6 se evaluará con detalle una aplicación práctica con una base de datos mayor, usando la generalización de esta propuesta que se desarrolla en el Capítulo 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 gráfico de la rutina **INTERDIS** se muestra en la Figura 4.6.

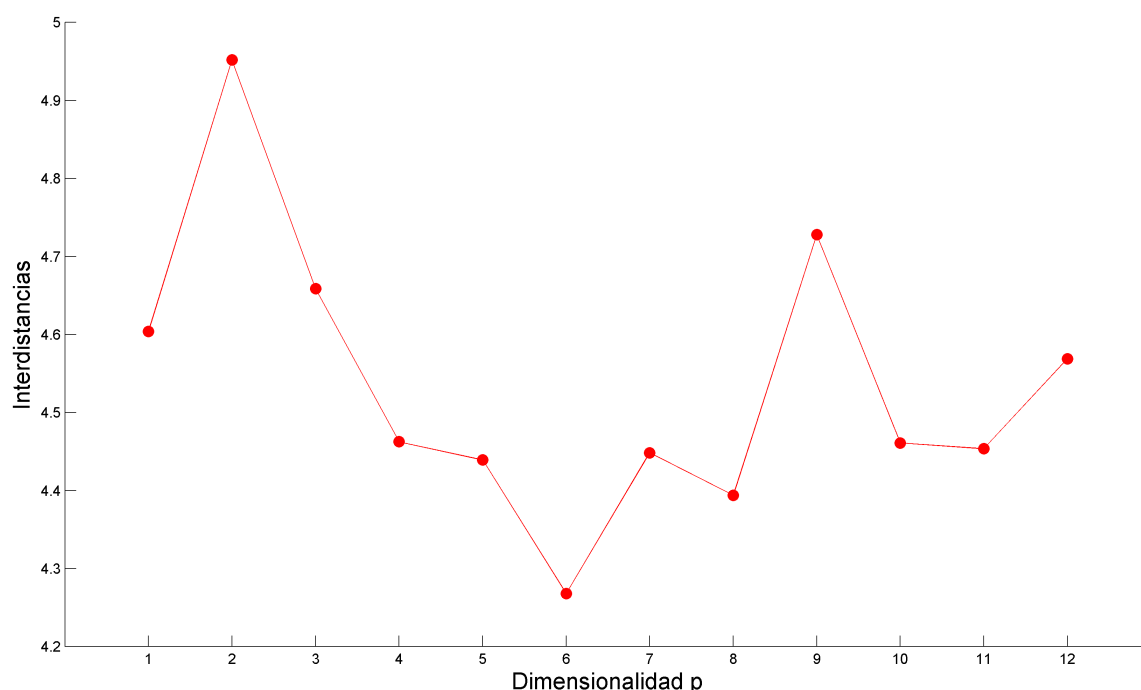


Figura 4.6. *Interdistancias* para soluciones OVERALS desde $p=1$ hasta $p=12$.

En este caso, la solución OVERALS con $p=6$ es la que minimiza las *Interdistancias*, garantizando que las categorías de las variables más frecuentemente asociadas sean las que reciban cuantificaciones más similares. En consecuencia, esta sería la solución más recomendable si el objetivo del investigador fuera la generación de cuantificaciones simples, para la subsiguiente aplicación de técnicas que exigieran datos numéricos.

Vale la pena insistir en el último aspecto anotado. Siempre que se utilicen calificativos como “óptimo”, “mejor” o “más recomendable” debe tenerse en mente el alcance de los mismos. En este capítulo se ha desarrollado una propuesta que garantiza la elección del conjunto de cuantificaciones en el que se den las mayores similitudes entre las cuantificaciones de las categorías que se asocian con mayor frecuencia. El hecho de que los conjuntos de cuantificaciones comparados sean el resultado de diferentes procesos de reducción de la dimensionalidad es circunstancial. El resultado obtenido no debe interpretarse como una prueba en favor de la elección de la dimensionalidad en cuestión para fines de representación. El conjunto de cuantificaciones elegido puede utilizarse como información de entrada de cualquier técnica lineal. En particular, si se trata de una técnica de reducción de dimensionalidad se tendrán en cuenta otras consideraciones para elegir el número de ejes retenidos o la dimensionalidad de la representación, la cual no tiene que coincidir con la dimensionalidad de la solución OVERALS que dio lugar a las cuantificaciones.

En el siguiente capítulo se elabora una propuesta de cuantificación basada en la generalización de los criterios desarrollados en el presente capítulo.

CAPÍTULO 5

PROPUESTA GENERALIZADA DE CUANTIFICACIÓN ÓPTIMA: CUANTIFICA

5.1 INTRODUCCIÓN

En el capítulo anterior se estableció un criterio de cuantificación óptima 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 óptimas sin que medie ningún proceso previo de cuantificación.

Inicialmente se establece un marco de referencia para el proceso de cuantificación óptima, en el cual se incluyen aspectos tales como la normalización, el nivel de escalamiento y los datos faltantes. Como novedad de este trabajo, incluimos los niveles de escalamiento Numérico Flotante y Ordinal Flotante, los cuales no forman parte de los niveles de escalamiento tradicionales.

Seguidamente se presenta la rutina **CUANTIFICA**, la cual está conformada por 16 subrutinas o funciones que interactúan para leer una base de datos multivariante convenientemente discretizada y asignar un esquema óptimo 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) Captación de la información suministrada por el usuario.
- 2) Lectura y adecuación de la base de datos.
- 3) Obtención de una matriz inicial de cuantificaciones.
- 4) Inicialización de los diferentes tipos de variables.
- 5) Cálculo de las *Interdistancias* iniciales.
- 6) Ciclo de minimización de las *Interdistancias*, acorde con el nivel de escalamiento de cada variable.
- 7) Generación y presentación de resultados.

Cerramos este capítulo comparando **CUANTIFICA** con los tres principales módulos de escalamiento óptimo del sistema Gifi de análisis multivariante no lineal, i. e., **PRINCALS**, **HOMALS** y **OVERALS**, cuando son usados como primer paso para la generación de cuantificaciones.

5.2 MARCO DE REFERENCIA DE LA CUANTIFICACIÓN ÓPTIMA

En el capítulo anterior se ofrece una solución específica al problema planteado en el Capítulo 3, esto es, obtener un conjunto de cuantificaciones simples óptimas para un sistema multivariante. Para tal efecto, se parte de la siguiente definición: “*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 definición se precisa así: “*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 transformación lineal de la suma de distancias cuadráticas. En este estudio se ha optado por la expresión [4.7], que no es más que un criterio escalado de la suma sobre todas las observaciones de la suma de interdistancias cuadráticas entre las categorías cuantificadas. Por simplicidad, a tal criterio lo hemos denominado *Interdistancias*.

Nótese que la solución presentada en el Capítulo 4, tiene dos limitaciones. Por una parte, exige utilizar el OVERALS o cualquier otra técnica de cuantificación 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 solución sea un óptimo absoluto en cuanto a la satisfacción del criterio propuesto, y bien podrían obtenerse otros conjuntos con *Interdistancias* menores.

Siendo consecuentes con la propuesta desarrollada en el Capítulo 4, resulta lógico buscar un esquema óptimo absoluto de cuantificaciones simples, esto es, que satisfaga la definición 4.2. Para tal efecto, podría tomarse como punto de partida la mejor solución de un conjunto de posibles cuantificaciones, obtenida mediante el procedimiento detallado en el Capítulo 4, o podría partirse directamente de la base de

datos original. Este último 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 cuantificación:

Normalización. Para evitar una solución trivial, consistente en que todas las cuantificaciones se colapsen en un punto (lo cual, obviamente, minimizaría las *Interdistancias*), se requiere aplicar algún proceso de normalización. En este trabajo, hemos elegido una normalización tal que cada una de las variables cuantificadas esté centrada en cero y tenga norma \sqrt{n} , siendo n el número de observaciones.

Nivel de Escalamiento. Para cada variable podrá definirse cualquiera de los cinco siguientes niveles de escalamiento: Numérico, Ordinal, Nominal, Numérico Flotante u Ordinal Flotante.

- **Numérico:** Se aplican transformaciones lineales sobre los valores iniciales de las categorías, de manera que las distancias entre las categorías transformadas sean proporcionales a las distancias entre las categorías de partida. El gráfico de cuantificaciones de una variable escalada a nivel Numérico es una línea recta.
- **Ordinal:** Se aplican transformaciones monótonas, de manera que las categorías transformadas conserven el orden de las categorías iniciales. Luego, el correspondiente gráfico de cuantificaciones será no descendente.
- **Nominal:** Se aplican transformaciones isomórficas, tales que cada una de las categorías reciba una cuantificación posiblemente diferente, sin más restricciones que las impuestas por la normalización. Los gráficos de cuantificaciones no siguen ningún patrón específico.
- **Numérico Flotante:** Todas las categorías, excepto una que se deja libre, se escalan a nivel Numérico.

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

En adición a la propuesta misma de cuantificación óptima basada en las definiciones 4.1 y 4.2 e implementada a través de la rutina **CUANTIFICA**, otro de los aspectos novedosos de este trabajo consiste en la definición e implementación de dos niveles de escalamiento adicionales a los tradicionalmente usados: Numérico Flotante y Ordinal Flotante.

Cuando alguna variable posea una categoría que no siga el patrón de las demás, tal como la que se presenta a menudo bajo la etiqueta “*no sabe/no responde*”, pueden seguirse dos cursos de acción, dependiendo del tipo de estudio, de la homogeneidad de la categoría en cuestión y de los objetivos del investigador: manejar la información de dicha categoría como datos faltantes o declarar tal categoría como flotante.

Al declarar todas las observaciones de una categoría como datos faltantes, estas se excluyen del proceso, lo que implica que no participan en la cuantificación de las demás categorías. Dicha categoría, desde luego, tampoco recibe ninguna cuantificación.

La otra posibilidad consiste en asignarle una cuantificación libre a la categoría en cuestión, declarándola como flotante, en cuyo caso participa en la cuantificación de las categorías de las demás variables y recibe asimismo una cuantificación, lo que permite evaluar su relación con las demás categorías.

En este último caso, se utiliza uno de los dos tipos de escalamiento flotante (Numérico Flotante u Ordinal Flotante), acorde con el nivel de escalamiento que se quiera asignar a las demás categorías de la variable. El programa escala las primeras $K-1$ categorías de la variable como Numéricas u Ordinales, dependiendo del nivel de escalamiento elegido, dejando libre (o flotante) la K -ésima categoría.

Para las variables Nominales, no es necesario definir un tipo de escalamiento flotante aunque existan categorías especiales del tipo “*no sabe/no responde*”, dado que el escalamiento Nominal asigna una cuantificación libre a cada una de las categorías.

Datos Faltantes. En el presente contexto, siempre que se hable de datos faltantes, valores perdidos u observaciones perdidas, se estará haciendo referencia a lecturas específicas de una o más variables que no existen en alguna fila específica, es decir, a celdas específicas de la matriz, sin que ello implique la pérdida de toda la fila.

Bajo el enfoque teórico propuesto y su correspondiente implementación algorítmica, la pérdida de datos en una fila, no impide que los datos restantes de la misma participen en el proceso de minimización 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 está 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 óptimo de cuantificaciones simples, acorde con las definiciones 4.1 y 4.2.

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

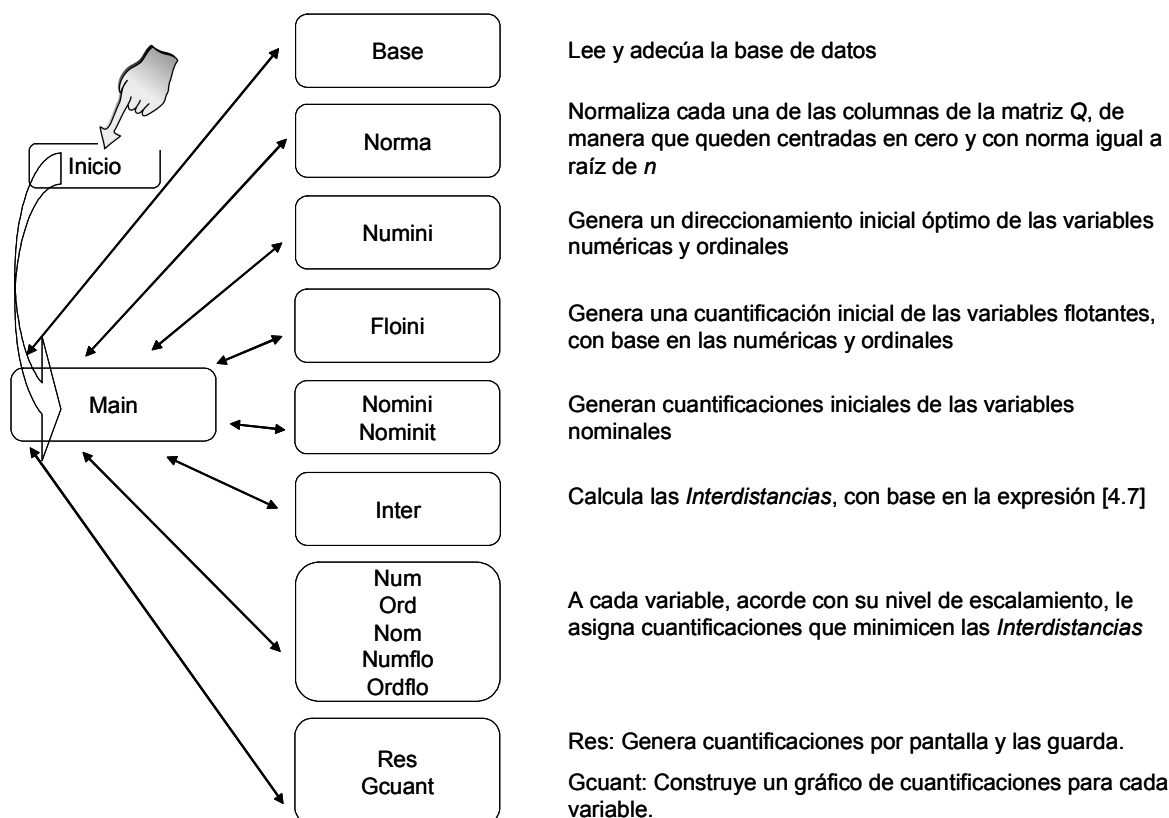


Figura 5.1. Conjunto de subrutinas que integran la aplicación CUANTIFICA.

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

Los pasos generales de la rutina CUANTIFICA son:

1. Captación de la información suministrada por el usuario.
2. Lectura y adecuación de la base de datos.
3. Obtención de una matriz inicial de cuantificaciones.
4. Inicialización de los diferentes tipos de variables.
5. Cálculo de las *Interdistancias* iniciales.
6. Ciclo de minimización de las *Interdistancias*, acorde con el nivel de escalamiento de cada variable.
7. Generación y presentación de resultados.

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

Los pasos 1 y 7, es decir, los que exponen lo relacionado con la entrada de la información y los resultados generados por la aplicación, son los que más conciernen al usuario final, a quien le bastaría con revisar estos dos apartados para realizar un proceso de cuantificación óptima. Las ayudas del programa están conformadas por una versión modificada de estos dos apartados. En los pasos 2 a 6 se encuentra la fundamentación teórica del proceso, expuesta en forma lineal, acorde con el flujo algorítmico.

5.3.1 CAPTACIÓN DE LA INFORMACIÓN DE ENTRADA.

La interacción entre el usuario y la aplicación se realiza a través de la interfaz gráfica que se muestra en la Figura 5.2. Esta interfaz es controlada por la subrutina **Inicio**. Luego, para ejecutar la rutina **CUANTIFICA**, bastará 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 botón “*¡ Leer Archivo Excel ¡*” y abrir el archivo que contiene la información de entrada.

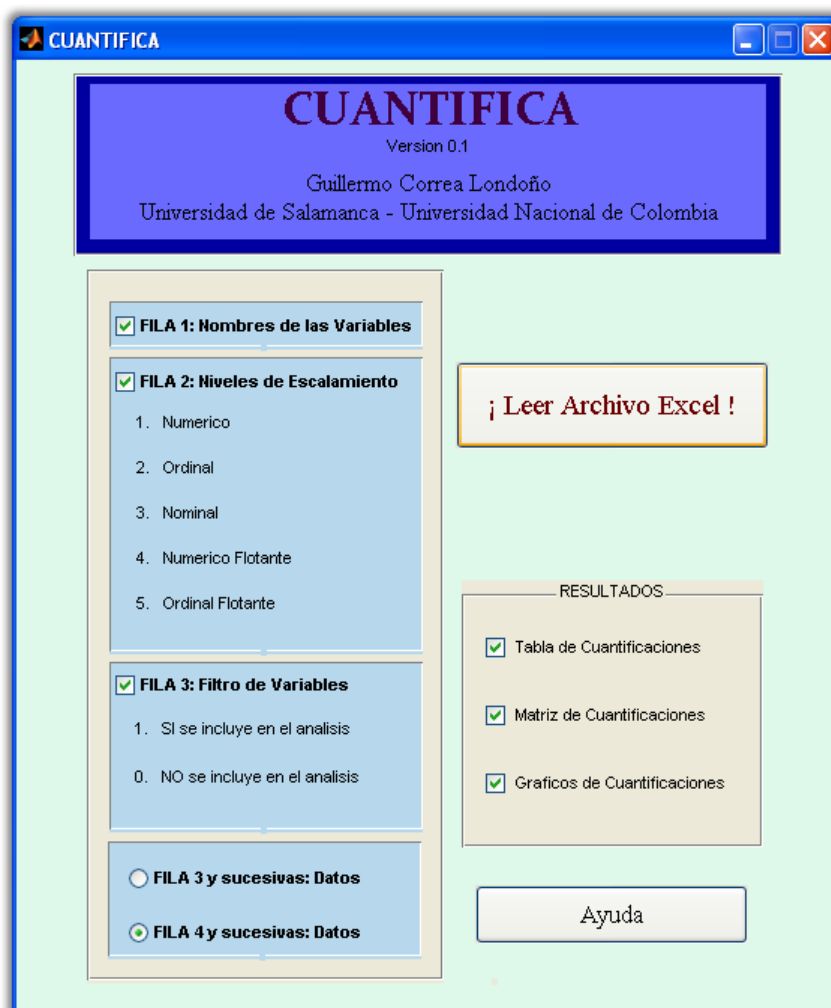


Figura 5.2. Interfaz gráfica de usuario de la aplicación **CUANTIFICA**.

¹⁸ Para facilitar la organización 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 alfabético.

La interfaz gráfica de la aplicación 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 continuación.

BASE DE DATOS

La base de datos debe estar organizada en un archivo de Excel con extensión **xls**, el cual ha de contener toda la información necesaria para realizar la cuantificación óptima. 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 usarán como base para la cuantificación. Si el archivo Excel tiene varias hojas, la base de datos deberá ubicarse en la primera de ellas, es decir, en la ubicada más 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 condición no es susceptible de modificación por parte del usuario. Los nombres pueden contener espacios y/o caracteres especiales. La única restricción consiste en que no estén conformados únicamente por números. En ocasiones, el sistema tampoco es capaz de leer los nombres cuando comienzan por un número, generando un error. Por tanto, se desaconseja dicha práctica.

NIVEL DE ESCALAMIENTO

En la segunda fila del archivo debe especificarse el nivel de escalamiento de cada una de las variables. Esta también es una condición que no permite modificación 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 número entre 1 y 5, así:

1. Numérico
2. Ordinal
3. Nominal
4. Numérico 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 opción que viene por defecto, el sistema leerá los datos a partir de la tercera fila y cuantificará 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, así:

1. SÍ se incluye la variable en el análisis.
0. NO se incluye la variable en el análisis

Nótese que para realizar un análisis 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 dispondrán a partir la cuarta fila.

DATOS (ETIQUETAS DE LAS CATEGORÍAS)

Todas las variables tienen que estar categorizadas o discretizadas, sin que exista ninguna restricción para la asignación de etiquetas o identificadores a las categorías, más allá de que sean números enteros y de la conservación de la ordinalidad o la proporcionalidad de distancias para las categorías de las variables escaladas como Ordinales o Numéricas, respectivamente.

Estas etiquetas o identificadores de las categorías constituyen los datos de entrada del proceso de cuantificación óptima, 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 aparición de valores numéricos no enteros en esta región de la base de datos será interpretada como un error involuntario del usuario, producido, quizá, por el formato de visualización de Excel, por lo que la aplicación procederá a su redondeo al entero más próximo.

Si se tiene, por ejemplo, una base de datos que, entre otras, incluya las variables *Sexo*, *Profesión*, *Edad* e *Ingresos*, escaladas a niveles Nominal (3), Nominal (3), Numérico (1) y Ordinal Flotante (5), respectivamente, y se desea excluir *Profesión* del proceso de cuantificación óptima (filtro=0), la base de datos tendrá el aspecto que se muestra en la Figura 5.3.

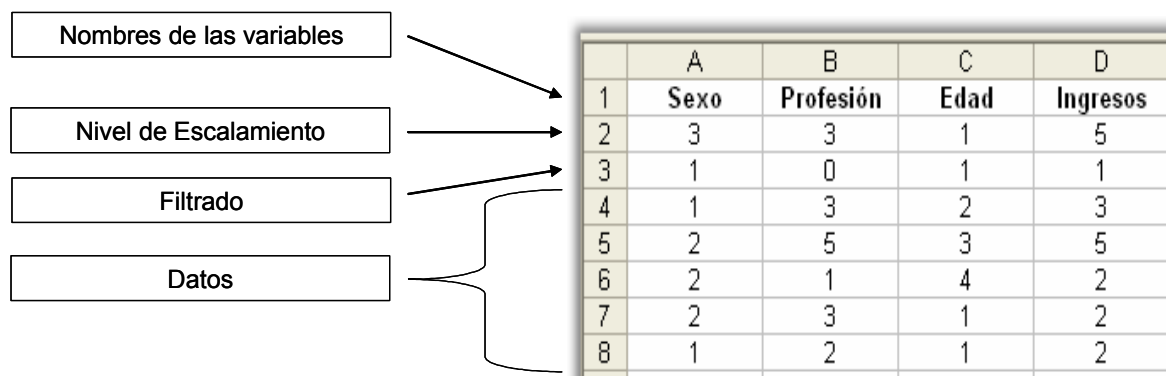


Figura 5.3. Estructura en Excel del archivo de entrada para la rutina CUANTIFICA.

Cuando se utiliza alguno de los dos niveles de escalamiento flotante, i. e., Numérico Flotante u Ordinal Flotante, la categoría flotante deberá identificarse con la etiqueta de mayor valor de la correspondiente variable. Considérese, por ejemplo, la variable *Ingresos*, escalada a nivel Ordinal Flotante, conformada por las siguientes categorías:

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

Para su identificación en la base de datos, se utilizarían respectivamente las etiquetas 5, 1, 2, 3 y 4. Alternativamente, podría utilizarse cualquiera de los siguientes conjuntos de etiquetas, sin que ello afecte el resultado: 4, 0, 1, 2, 3 ó 4000, 500, 600, 1500, 3000. Nótese que cualquiera de los conjuntos de etiquetas propuesto respeta la ordinalidad de la parte no flotante (categorías 2 a 5) y asigna la etiqueta de mayor valor a la categoría 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 Numérico Flotante, solo los dos primeros conjuntos de etiquetas propuestos serían equivalentes entre sí, 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 categoría *No responde* sea la última de su grupo, en este caso la hemos presentado intencionalmente encabezando el grupo de categorías, para enfatizar el hecho de que, sin importar su posición en la lista de referencia, la categoría *No responde* siempre se codificará con la etiqueta de mayor valor de su grupo de categorías.

DATOS FALTANTES

Cuando falta la información de algunas celdas, basta con dejarlas vacías o llenarlas con cualquier carácter o grupo de caracteres no numéricos. Es posible, incluso, combinar estas dos posibilidades: dejar celdas vacías y utilizar caracteres no numéricos en otras, los cuales no tienen que coincidir entre sí. Todos los valores no numéricos o ausentes que se encuentren de la cuarta fila en adelante serán manejados por el programa como datos perdidos.

En ningún caso debe utilizarse el cero (0) para señalar valores faltantes, pues este es un valor numérico que no se interpretaría como marcador de valor perdido, sino como la etiqueta de una categoría.

Aunque la importación de la base de datos es un proceso que normalmente se realiza sin problemas y de manera bastante rápida, en algunas ocasiones, la presencia de muchas celdas no numéricas (datos perdidos) puede afectar negativamente el desempeño de la función de importación. Para resolver esta situación, en caso de que se presente, basta con identificar todas las observaciones perdidas de la base de datos con el valor numérico -99. Este es el único valor numérico que siempre será interpretado por el programa como marcador de datos faltantes; cualquier otro valor numérico es manejado como la etiqueta de una categoría.

En adición al botón *¡ Leer Archivo de Excel ¡*, la interfaz gráfica permite elegir los resultados que se desean generar. Por defecto, todos aparecen marcados. En § 5.3.7 se describe en qué consiste cada uno de ellos.

El otro punto de interacción del usuario con la aplicación es a través del botón *Ayuda*, el cual le brinda acceso a una versión modificada de lo descrito en este apartado (§ 5.3.1), complementada con la descripción de los resultados, la cual presentamos más adelante en esta memoria (§ 5.3.7).

La comprensión de lo expuesto hasta aquí, en lo que tiene que ver con la rutina **CUANTIFICA**, es suficiente para su utilización por parte del usuario final, quien interactúa con la misma únicamente a través de la interfaz gráfica. A continuación

describimos el flujo general de la aplicación, así como los procesos específicos realizados por cada una de las demás 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 gráfica, la información es transmitida a la subrutina **Main**, la cual orquesta en adelante el proceso de cuantificación. 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 específicos 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 información para realizar el análisis, esto es, que la base de datos tenga como mínimo dos observaciones y dos variables (aunque una cuantificación para este caso extremo no resultaría demasiado interesante). De no ser así, el sistema genera el correspondiente mensaje de error y la aplicación se detiene.

A continuación, 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 así, se informa sobre dicho error y se detiene la aplicación.

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 aplicación. Asimismo, si el vector de filtrado se definió adecuadamente, el sistema informa sobre su estado indicando cuántas variables han sido filtradas o si todas han sido incluidas en el análisis.

Se verifica luego que todas las etiquetas de las categorías sean números enteros. De no ser así, se redondean al entero más próximo.

Se realiza a continuación la marcación de valores perdidos, informando por pantalla cuántos valores perdidos contiene la base de datos y lo que ello representa en términos porcentuales.

Puesto que ninguno de los pasos de la rutina **CUANTIFICA** está basado en operaciones que exijan matrices completas, no es necesario utilizar ninguna técnica de imputación de datos. Esta característica hace que el procedimiento sea transparente, por cuanto está basado únicamente en la información disponible, sin que se requiera forzar la estimación de categorías desconocidas mediante técnicas más o menos subjetivas y sin que la ausencia de valores específicos sea un impedimento para la utilización de los demás valores, usándose siempre el 100 % de la información susceptible de generar *Interdistancias*, esto es, todas las filas con al menos dos observaciones.

Lógicamente, las filas con menos de dos observaciones ni pueden usarse en el cálculo de las *Interdistancias* (cf. expresión [4.7]) ni contienen ninguna información multivariante. Se procede, por tanto, a retirarlas de la base de datos.

Puesto que el proceso de filtrado y eliminación de filas con menos de dos lecturas reducen el tamaño de la base de datos, la subrutina verifica nuevamente que la base de datos final contenga la información mínima para el análisis, esto es, dos filas y dos columnas. De no ser así, se informa que no se cuenta con información suficiente para realizar la cuantificación y la aplicación se detiene.

Aun cuando la reducción de la base de datos no sea tan drástica como para impedir su uso, cualquier reducción 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 cuantificación. 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. Así, por ejemplo, si la base de datos original aparece en una hoja llamada **Luz**, la base de datos reducida se copiará en una hoja llamada **Luz red**.

Finalmente, tras haber realizado todas las verificaciones y adecuaciones del caso, se genera un aviso informativo de la finalización del proceso de lectura y adecuación de la base de datos, indicando el número de filas y de columnas que serán utilizadas para la cuantificación.

5.3.3 OBTENCIÓN DE LA MATRIZ INICIAL DE CUANTIFICACIONES.

Una vez leída la base de datos y sus características (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 normalización por columnas de la matriz de datos, X .

La normalización 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 \sqrt{n} , siendo n , el número de observaciones (número de filas).

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

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

$$q_n = \frac{(q - \bar{q})}{\sqrt{(q - \bar{q})(q - \bar{q})}} \sqrt{n} \quad [5.1]$$

Donde:

- q_n : Columna normalizada de la matriz Q .
- q : Columna de la matriz Q antes de la normalización.
- \bar{q} : Media de la columna q .
- n : Número de filas de la matriz Q .

Si se trata de una columna con datos perdidos:

$$q_n = \frac{(q - \bar{q})}{\sqrt{\sum_{q_i \in \mathbb{R}} (q_i - \bar{q})^2}} \sqrt{n} \quad [5.2]$$

Donde:

q_n : Columna ‘normalizada’ de la matriz Q .

q : Columna de la matriz Q antes de la normalización.

\bar{q} : Media de la columna q , omitiendo los valores perdidos.

q_i : i -ésima observación de la columna q , con i evaluado sobre el conjunto de valores observados (valores no perdidos).

n : Número de filas de la matriz Q .

El entrecomillado en ‘normalizada’ indica que, puesto que la columna en cuestión tiene datos perdidos, no es posible obtener la norma mediante la expresión usual: $\|q\| = \sqrt{q'q}$,

sino que se obtiene una norma generalizada mediante la expresión: $\|q\| = \sqrt{\sum_{q_i \in \mathbb{R}} q_i^2}$.

Nótese que ambas expresiones coinciden cuando la columna está completa.

5.3.4 SUBROUTINAS DE INICIALIZACIÓN.

Antes de aplicar el ciclo general de minimización 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 más rápida hacia las *Interdistancias* mínimas.

SUBROUTINA “Numini”

La cuantificación óptima de variables escaladas a nivel Numérico 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 lógica de las cuantificaciones óptimas expuesta en el capítulo anterior, en particular, con base en la definición 4.1. Obsérvese, por ejemplo, que si un par de variables numéricas se correlacionan negativamente las categorías bajas de la primera tiende a asociarse con las altas categorías de la segunda y viceversa. Luego, las cuantificaciones de las categorías inferiores de la primera variable deberán estar cercanas a las cuantificaciones de las categorías superiores de la segunda variable y viceversa. Esto, desde luego, solo se logra utilizando constantes de signo contrario para las correspondientes transformaciones. Un razonamiento análogo se sigue para las variables correlacionadas positivamente.

Nótese que en cualquiera de los dos casos anteriores, bien sea que las variables originales se correlacionen positiva o negativamente, las variables transformadas se correlacionarán positivamente. Esto será así siempre que se consideren solo dos variables numéricas o que las correlaciones sean transitivas¹⁹.

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

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

$$\begin{aligned} A \text{ corr}(+) B \text{ y } B \text{ corr}(+) C &\Rightarrow A \text{ corr}(+) C \\ A \text{ corr}(+) B \text{ y } B \text{ corr}(-) C &\Rightarrow A \text{ corr}(-) C \\ A \text{ corr}(-) B \text{ y } B \text{ corr}(-) C &\Rightarrow A \text{ corr}(+) C \end{aligned}$$

datos con un elevado número de variables que interactúan de forma compleja, impidiendo que esto siempre sea así, tal y como se ilustra a continuación.

Supóngase, por ejemplo, que en un estudio sobre los hábitos de un colectivo estudiantil se incluyen, entre otras, las variables 1: ‘horas por semana dedicadas a realizar algún deporte’, 2: ‘número 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. Supóngase además que la matriz de correlaciones de estas tres variables, antes de ser transformadas es la que se muestra a continuación.

$$\begin{array}{c}
 D \quad C \quad L \\
 D \begin{bmatrix} 1,00 & -0,40 & 0,23 \\
 C \begin{bmatrix} -0,40 & 1,00 & 0,11 \\
 L \begin{bmatrix} 0,23 & 0,11 & 1,00 \end{bmatrix}
 \end{array}$$

Las variables deporte y lectura se correlacionan positivamente, por lo que podrían transformarse con sendas constantes positivas, manteniéndose la correlación 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 transformación para cerveza que se correlacione positivamente con las transformaciones de deporte y de lectura. Si se transforma mediante una constante positiva, tendrá correlación positiva con lectura, pero negativa con deporte; si se transforma mediante una constante negativa, tendrá correlación positiva con deporte, pero negativa con lectura.

Como se indicó anteriormente, en un sistema óptimamente 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 combinación de restricciones para las variables escaladas numéricamente (linealidad y normalización) hace posible obtener las cuantificaciones

cuasi definitivas desde el primer paso, donde se obtiene la matriz de cuantificaciones inicial mediante normalización, pudiendo estas sufrir cambios posteriores solamente en su dirección, es decir que bastará 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 demás variables, cambiando únicamente su signo. En general, si la suma de correlaciones de una variable con las demás es mayor al cambiar el signo que la suma de correlaciones existente antes de cambiarlo, dicho cambio estará justificado.

Tras la obtención de la matriz inicial de cuantificaciones, la subrutina **Numini** realiza un paso de direccionamiento de las variables Numéricas y Ordinales, para la maximización de la suma sus correlaciones, tratando a las variables Ordinales como si fueran Numéricas. Aquí se incluyen también las variables con nivel de escalamiento Numérico Flotante y Ordinal Flotante, ignorando durante el proceso las categorías flotantes. Este paso genera una estructura estable que permite obtener mejores estimaciones iniciales para las variables Nominales y, consecuentemente, una convergencia más rápida.

Puesto que el sistema de correlaciones usualmente es complejo y no transitivo, la subrutina **Numini** debe utilizar un mecanismo que permita encontrar rápidamente la máxima matriz de correlaciones. Para tal fin, se mide el efecto que tiene el cambio de signo de cada una de las variables, manteniendo las demás 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 evalúa nuevamente el cambio en las correlaciones totales producido por el cambio de dirección de cada una de las variables, y se sigue así 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 ningún cambio logra incrementar la suma de correlaciones, el proceso concluye, quedando las variables deporte y lectura con su orientación inicial y cambiándose la orientación 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 evolución de la correlación total.

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

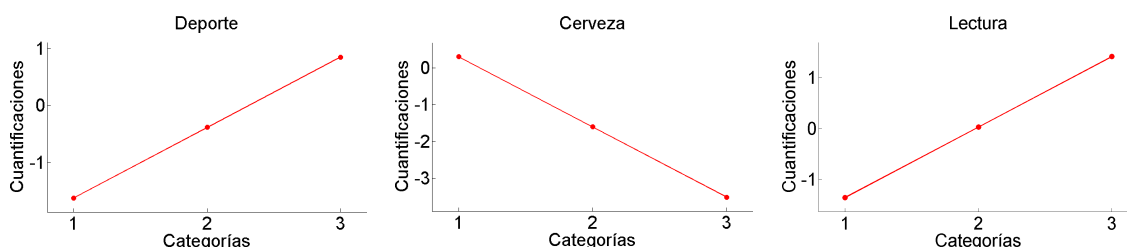


Figura 5.4. Gráficos de cuantificaciones para deportes, cerveza y lectura.

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

Aunque para fines de minimizar las *Interdistancias* se permitirá el libre cambio de dirección de las variables Numéricas y Ordinales (así como de la parte no flotante de las variables Numéricas Flotantes y Ordinales Flotantes), la dirección original de estas

variables se recuperará antes de presentar los resultados finales, mediante la multiplicación de las cuantificaciones por -1 , de manera que las cuantificaciones de estas variables se correlacionen positivamente con las etiquetas originales y, lo más importante, que las direcciones de las correlaciones iniciales entre estas variables se conserven, lo cual facilitará las interpretaciones de los resultados de otras técnicas basados en tales cuantificaciones.

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

SUBROUTINA “Floini”

Si se tienen variables con categorías flotantes, mediante esta subrutina se obtiene una cuantificación inicial de las mismas, con base en las variables Numéricas y Ordinales, tratando a estas últimas como Numéricas. Las categorías fijas de las variables con escalamiento Numérico Flotante y Ordinal Flotante también son usadas como base para la cuantificación inicial de las categorías flotantes.

Las categorías flotantes se cuantifican como si formaran parte de variables con escalamiento Nominal. Su cuantificación se obtiene, por tanto, como el promedio de las demás cuantificaciones, tal y como se detalla en el apartado correspondiente a la cuantificación de variables Nominales (subrutina **Nom**).

Si hay más de una variable con categorías flotantes, se evalúa inicialmente el cambio producido en la matriz de correlaciones al cuantificar las categorías flotantes, cuantificando inicialmente la que genere la mayor suma de correlaciones, y así sucesivamente hasta cuantificarlas todas.

Si bien es cierto que el concepto de correlación solo tiene sentido en variables Numéricas u Ordinales, su uso con las variables transformadas es totalmente válido puesto que estas tienen propiedades métricas.

INICIALIZACIÓN DE VARIABLES NOMINALES

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

SUBROUTINA “**Nomini**”

Si se tiene una mezcla de escalamiento Nominal con otros niveles de escalamiento, se realiza una categorización inicial de las variables Nominales con base en las demás variables, mediante un procedimiento análogo al usado para la cuantificación de las categorías flotantes (Subrutina **Floini**).

Partiendo de una estructura estable proporcionada por las variables Numéricas, 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 demás para la cuantificación de las siguientes variables Nominales, hasta cuantificarlas todas.

No está de más insistir en el hecho de que no obstante que el concepto de correlación solo tiene sentido en variables Numéricas u Ordinales, su uso con variables transformadas, aun de variables Nominales, es totalmente válido, pues todas las variables transformadas tienen propiedades métricas.

SUBROUTINA “Nominit”

Se utiliza cuando todas las variables están escaladas a nivel Nominal. En este caso, puesto que no existe una estructura de partida conformada por las variables Numéricas-Ordinales, se crea una estructura estable a partir de las dos variables que más se relacionen entre sí.

Si se considera aisladamente cualquier variable y se desea cuantificarla a partir de una única variable proveniente de un conjunto, lógicamente la mejor cuantificación será la obtenida con base en la variable con la cual esté más relacionada.

Aunque hemos insistido en la validez de utilizar el coeficiente de correlación como medida de asociación entre variables transformadas, debe notarse que en este primer paso, las variables no han recibido aún ninguna cuantificación, por lo que son nominales y no tendría ningún sentido usar el coeficiente de correlación. Por consiguiente, en esta parte se evalúan las asociaciones mediante el estadístico Ji cuadrado, el cual se calcula para todos los posibles pares de variables.

Puesto que el número de categorías de las variables puede diferir, los estadísticos Ji cuadrado no son directamente comparables entre sí. Lo más adecuado, desde el punto de vista teórico, sería utilizar los correspondientes p-valores, pero esta opción no resulta viable, puesto que, al nivel de precisión de MATLAB, estos se igualan a cero con mucha rapidez. Por tanto, se ha utilizado para fines de comparación un valor ajustado del estadístico Ji cuadrado, obtenido como la razón entre el estadístico y sus grados de libertad.

Una vez elegido el par de variables cuyo valor del Ji cuadrado ajustado sea máximo, se realiza un ciclo donde, omitiendo todas las demás 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 continuación se incluyen sucesivamente las demás variables, evaluando en cada paso las correlaciones totales y eligiendo la variable cuya inclusión dé lugar a la mayor correlación total, hasta incluirlas todas.

5.3.5 CÁLCULO DE LAS *INTERDISTANCIAS* INICIALES.

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

Es importante anotar que el cálculo de las *Interdistancias* no se ve afectado por los datos faltantes, siendo posible calcularlas en todos los casos, con la participación 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 demás, aplicando un paso de normalización tras cada recuantificación.

La cuantificación 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, así: **Num** (Numérico), **Ord** (Ordinal), **Nom** (Nominal), **Numflo** (Numérico Flotante) y **Ordflo** (Ordinal Flotante).

Un ciclo completo consta de la recuantificación de cada una de las variables con base en las restantes, eligiéndolas en un orden aleatorio, con el fin de evitar posibles sesgos. Tras la recuantificación de cada variable se aplica un paso de normalización y se calculan las *Interdistancias*, verificando que sean menores que las anteriores.

Si tras la ejecución de un ciclo se observa alguna disminución 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* mínimas y, en consecuencia, las cuantificaciones óptimas.

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

SUBROUTINA “Num” (ESCALAMIENTO NUMÉRICO)

Como se indicó anteriormente, la rigidez del escalamiento Numérico se convierte en una ventaja a la hora de generar cuantificaciones. Si todas las variables fueran Numéricas, tras la normalización de la matriz original de datos, bastaría con verificar los direccionamientos óptimos de cada variable que fueron generados por la subrutina **Numini**, cambiando el signo de las cuantificaciones donde fuera necesario, para tener las cuantificaciones óptimas.

En la práctica, cuando se tienen variables con diferentes niveles de escalamiento, las cuantificaciones eventualmente pueden cambiar con relación a las cuantificaciones iniciales generadas por la subrutina **Numini**, pero el único cambio posible es de dirección.

La subrutina **Num** calcula las *Interdistancias* obtenidas al cambiar la dirección 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 invertirá la dirección de la variable; en caso contrario, se mantendrá en su estado inicial.

SUBROUTINA “Nom” (ESCALAMIENTO NOMINAL)

El escalamiento Nominal se logra mediante la aplicación de transformaciones isomórficas, que son las menos restrictivas y las más generales del conjunto de transformaciones aplicadas. Todas las demás transformaciones pueden derivarse como casos particulares de estas.

Una transformación isomórfica consiste en asignar la misma cuantificación a todas las observaciones de una categoría, sin que tenga que satisfacerse ninguna relación particular entre las cuantificaciones de diferentes categorías. En tal sentido, la

cuantificación de todas las categorías es libre. Es por ello que no hace falta definir un nivel de escalamiento Nominal Flotante cuando se tiene una categoría especial, del tipo “no sabe/no responde”, en una variable escalada a nivel Nominal. Basta con elegir el nivel de escalamiento Nominal, incluyendo la categoría especial como una más de las categorías del conjunto, sin ser necesario siquiera que se le asigne la etiqueta de máximo valor.

Puesto que el escalamiento Nominal no impone ningún tipo de restricción para la cuantificación de una categoría, siendo esta independiente de la cuantificación de las demás categorías, el proceso de cuantificación de las diferentes categorías de una variable escalada a nivel Nominal se lleva a cabo, asimismo, de manera independiente, mediante la asignación de cuantificaciones óptimas a cada una de las categorías, esto es, mediante la asignación de cuantificaciones que minimicen las *Interdistancias*.

Aunque la expresión [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 categorías de una variable cualquiera, como se indica en la expresión [5.3].

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

Donde:

Interdistancias: Criterio escalado de suma de interdistancias cuadráticas.

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

K : Número de categorías de la variable.

n_k : Número de observaciones de la k -ésima categoría, donde

$$\sum_{k=1}^K n_k + n_{miss} = n, \text{ con } n_{miss}, \text{ el número de observaciones perdidas.}$$

S_{ik}^2 : ‘Varianza’ de la i -ésima observación, de la k -ésima categoría.

Puesto que la cuantificación de cada categoría es independiente de la de las demás, es posible minimizar la expresión [5.3] minimizando cada uno de sus K sumandos o

Interdistancias parciales, esto es, $\sum_{i=1}^{n_k} S_{ik}^2$, $k=1, 2, \dots, K$. Para tal efecto, resulta

conveniente recurrir a la expresión original de la suma de interdistancias cuadráticas, tal y como se presenta en el apartado 4.4 (cf. expresión [4.6]). La suma de interdistancias cuadráticas para una categoría específica se denota así:

$$\sum_{i=1}^{n_k} \sum_{j=1}^{m-1} \sum_{j'>j}^m (q_{ij} - q_{ij'})^2, \quad j \text{ t. q. } q_{ij} \in \mathbb{R} \quad [5.4]$$

Donde:

- n_k : Número de observaciones (filas) de la categoría evaluada.
- m : Número de variables.
- q_{ij} : Cuantificación de la i -ésima categoría de la j -ésima variable.
- $q_{ij'}$: Cuantificación de la i -ésima categoría de la j' -ésima variable.

La restricción $j \text{ t. q. } q_{ij} \in \mathbb{R}$ indica que esta suma solo se hace sobre los valores efectivamente observados, ignorando las observaciones perdidas.

Supongamos, ahora, sin pérdida de generalidad, que la variable que estamos cuantificando es la m -ésima. Podemos particionar la expresión [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 interés) y otro debido a las distancias entre la cuantificación de interés (c) y las cuantificaciones de las restantes $m-1$ variables, así.

$$\sum_{i=1}^{n_k} \sum_{j=1}^{m-1} \sum_{j'>j}^m (q_{ij} - q_{ij'})^2 = \sum_{i=1}^{n_k} \sum_{j=1}^{m-2} \sum_{j'>j}^{m-1} (q_{ij} - q_{ij'})^2 + \sum_{i=1}^{n_k} \sum_{j=1}^{m-1} (q_{ij} - c)^2, \quad j \text{ t. q. } q_{ij} \in \mathbb{R}$$

La derivada de esta expresión con respecto a c es:

$$\begin{aligned}
 \frac{\delta \sum_{i=1}^{n_k} \sum_{j=1}^{m-1} \sum_{j'>j}^m (q_{ij} - q_{ij'})^2}{\delta c} &= 0 + \sum_{i=1}^{n_k} \sum_{j=1}^{m-1} (-2q_{ij} + 2c) \\
 &= \sum_{i=1}^{n_k} \sum_{j=1}^{m-1} (2c) - 2 \sum_{i=1}^{n_k} \sum_{j=1}^{m-1} q_{ij} \\
 &= 2n_k(m-1)c - 2 \sum_{i=1}^{n_k} \sum_{j=1}^{m-1} q_{ij}
 \end{aligned}$$

Esta derivada se hace cero cuando:

$$c = \frac{\sum_{i=1}^{n_k} \sum_{j=1}^{m-1} q_{ij}}{n_k(m-1)} \equiv \bar{Q}_k, \text{ con } \bar{Q}_k, \text{ la media de la } k\text{-ésima categoría.} \quad [5.5]$$

Luego, la cuantificación de una categoría que minimiza la correspondiente componente de las *Interdistancias* es la que se obtiene promediando las cuantificaciones de las categorías de las demás variables que aparecen conjuntamente con la categoría en cuestión.

SUBROUTINA “Ord” (ESCALAMIENTO ORDINAL)

La cuantificación óptima de las variables escaladas a nivel Ordinal se realiza asignando cuantificaciones a cada una de sus categorías, a través del mismo procedimiento utilizado por la subrutina **Nom**, esto es, promediando las cuantificaciones de las demás variables en el correspondiente grupo de observaciones. Seguidamente, se aplica un paso de restricción, en el que se verifica la ordinalidad de la cuantificación y se aplican los correctivos necesarios. Esto respalda la afirmación de que todas las demás transformaciones pueden derivarse como casos particulares de la transformación isomórfica. En este caso, se trata de una transformación isomórfica con restricciones.

Puesto que en el apartado correspondiente a la subrutina **Nom** se ha detallado cómo obtener una cuantificación óptima irrestricta, mediante la aplicación de transformaciones isomórficas, no redundaremos en ello y nos concentraremos en la restricción por ordinalidad.

La ordinalidad puede ser ascendente o descendente. Para facilitar la aplicación de la restricción por ordinalidad, se hace que todas las ordinalidades sean ascendentes, modificando temporalmente las descendentes. Para tal efecto, se calcula la correlación entre las etiquetas originales de las categorías y las cuantificaciones. En caso de resultar negativa, se multiplican todas las cuantificaciones por menos uno antes de aplicar la restricción por ordinalidad y posteriormente vuelven a llevarse al estado inicial multiplicándolas nuevamente por menos uno.

Una vez que las cuantificaciones estén correlacionadas positivamente con las correspondientes etiquetas de sus categorías, se verifica que se satisfaga una relación de ordinalidad ascendente entre las mismas, esto es, que $q_1 \leq q_2 \leq \dots \leq q_k \leq \dots \leq q_K$, donde q_k es la cuantificación de la k -ésima categoría.

Si se encuentran dos cuantificaciones q_k y $q_{k'}$, con $k < k'$, tales que $q_k > q_{k'}$, será necesario igualar ambas cuantificaciones, así como todas las que se encuentren en medio de estas ($k < c < k'$). La verificación se realiza comparando inicialmente el par más externo y se va avanzando hacia los pares más internos, así:

$$q_1 < q_K, q_1 < q_{K-1}, \dots, q_1 < q_2, q_2 < q_K, q_2 < q_{K-1}, \dots, q_2 < q_3, \dots, q_{K-2} < q_K, q_{K-2} < q_{K-1}, q_{K-1} < q_K$$

Si en todas las comparaciones se satisfacen las relaciones de ordinalidad, se tendrán ya las cuantificaciones óptimas, sin necesidad de ninguna modificación más allá de regresar a ordinalidad descendente las cuantificaciones que hayan sido previamente modificadas²⁰.

²⁰ En los casos en que no hay empates en las cuantificaciones de categorías 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 dirección.

En caso de que al realizar alguna de las comparaciones no se satisfaga la relación de ordinalidad, será necesario igualar las cuantificaciones del par en cuestión así como las de todas las categorías intermedias, en caso de que existan.

Puesto que hay infinitas formas de realizar tal procedimiento²¹, la aplicación de esta restricción no es una cuestión trivial. De todas las posibles formas de aplicar dicha restricción, nos interesa la que produzca las menores *Interdistancias*.

Puesto que cualquiera que sea la restricción aplicada, las categorías sometidas a la misma quedarán con la misma cuantificación, podemos ver este proceso como un colapsamiento de categorías. Para la cuantificación óptima de esa nueva categoría, conformada por las categorías colapsadas, nos basamos en el resultado obtenido al buscar la cuantificación que minimizara para una categoría específica, la correspondiente componente de las *Interdistancias*, (cf. Subrutina **Nom**, expresión [5.5]). Así, pues, puede afirmarse que la cuantificación óptima de la categoría colapsada estará dada por la cuantificación promedio de las categorías de las demás variables en el grupo colapsado.

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

SUBROUTINA “Numflo” (ESCALAMIENTO NUMÉRICO FLOTANTE)

Esta subrutina genera cuantificaciones que minimizan las *Interdistancias* para una variable escalada a nivel Numérico Flotante. Es decir que las primeras $K-1$ categorías se escalan a nivel Numérico y la K -ésima categoría 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 igualándolas a cualquier otro punto ubicado entre ambas.

En este caso se obtiene la cuantificación óptima para la categoría flotante, con base en el promedio de las categorías de las demás variables en el correspondiente grupo, acorde con el resultado [5.5]. Para la parte Numérica de la variable, es decir, las primeras $K-1$ categorías, se evalúan sus dos posibles configuraciones, es decir, la que traiga del paso anterior y esa misma configuración multiplicada por menos uno. Tras el paso de normalización, se elige aquella configuración que dé lugar a las menores *Interdistancias*.

SUBROUTINA “Ordflo” (ESCALAMIENTO ORDINAL FLOTANTE)

Esta subrutina genera cuantificaciones óptimas para una variable escalada a nivel Ordinal Flotante, es decir, que las primeras $K-1$ categorías se escalan a nivel Ordinal y la K -ésima categoría 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 -ésima categoría 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 óptimas para cada una de las variables del conjunto multivariante, respetando las restricciones propias de cada nivel de escalamiento.

Tal y como se indicó en los párrafos finales de la descripción de la subrutina **Numini** (§ 5.3.4), antes de presentar los resultados, se realizan los cambios de dirección necesarios para que las cuantificaciones de las variables Numéricas y Ordinales (así como de las Numéricas Flotantes y Ordinales Flotantes, sin incluir la categoría flotante), se correlacionen positivamente con las etiquetas originales de las categorías. Esto se obtiene multiplicando por menos uno las cuantificaciones de las variables en cuestión.

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

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

SUBROUTINA “Res” (RESULTADOS ALFANUMÉRICOS)

Además 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 cuantificación óptima de cada una de sus categorías. Allí se indica también el nivel de escalamiento con base en el cual se ha obtenido la cuantificación óptima 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 información. En la Figura 5.5 se muestra el aspecto general de tales tablas para la misma sección 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 obtención de las cuantificaciones²², en la que se han

²² En caso de que no se haya generado una matriz de datos reducida (cf. § 5.3.2), la comparación con la base de datos original se realiza omitiendo las filas de escalamiento y de filtrado. En los casos en que sí se haya generado una matriz de datos reducida, las estructuras de ambas matrices coinciden celda a celda.

reemplazado las etiquetas iniciales de las categorías por sus correspondientes cuantificaciones óptimas. Todas las variables de esta matriz gozan de propiedades métricas, lo que hace posible analizarlas mediante procedimientos multivariantes lineales.

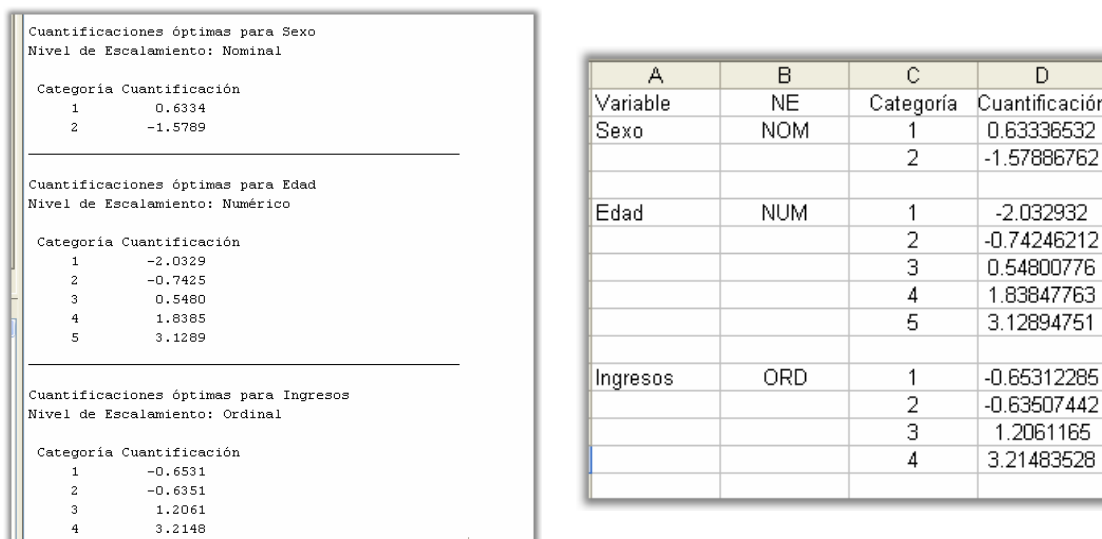


Figura 5.5. Aspecto general de la Tabla de Cuantificaciones, mostrada por pantalla (izquierda) y guardada en la hoja TC (derecha).

Puesto que el procedimiento de cuantificación óptima no utiliza ningún método de imputación para los valores faltantes, la matriz cuantificada tendrá 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 adición 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 información.

Debe verificarse que el archivo no contenga hojas llamadas TC ni MC, pues en tal caso serían sobrescritas con la información 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 análisis, 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 específico de salida de la información, dejando inalterado el resto de la hoja, lo que puede dar lugar a que se cometan errores con el posterior manejo de la información.

Supóngase, por ejemplo, que se inicialmente se realiza un análisis con 15 variables y que posteriormente, tras filtrar 5 de las variables, se realiza un nuevo proceso de cuantificación. La rutina de resultados solamente sobrescribirá en el rango correspondiente a las 10 primeras variables. Luego, la Tabla de Cuantificaciones de la nueva ejecución quedará sobrescrita en el rango de filas que corresponda a las 10 variables incluidas en la ejecución, pero esta información estará seguida por un grupo de filas remanentes correspondientes a las cuantificaciones de las últimas cinco variables de la ejecución anterior. Asimismo, la Matriz de Cuantificaciones seguirá teniendo 15 columnas: las 10 primeras correspondientes a las variables incluidas en el último análisis y las 5 siguientes correspondientes a las últimas 5 variables de la ejecución anterior.

Desde luego, tales precauciones no serán necesarias cuando no se modifique el vector de filtrado o cuando la modificación consista en agregar variables al análisis.

Finalmente, tras realizar modificaciones sobre el archivo de Excel, este deberá cerrarse para permitir su sobrescritura.

SUBROUTINA “Gcuant” (RESULTADOS GRÁFICOS)

Los resultados gráficos del proceso de cuantificación óptima 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 categorías en la abcisa y sus correspondientes cuantificaciones óptimas en la ordenada.

Para las variables escaladas a nivel Numérico, el gráfico será una línea recta con pendiente positiva; si la variable se escala a nivel Ordinal, se observarán relaciones monótonas no descendentes y si la variable se escala a nivel Nominal, el gráfico puede asumir cualquier forma. Para las variables con nivel de escalamiento flotante se satisfará la relación base entre las primeras $K-1$ categorías; la última categoría será flotante. Las Figura 5.6, 5.7, 5.8, 5.9 y 5.10 muestran gráficos de cuantificaciones típicos para variables escaladas a nivel Numérico, Ordinal, Nominal, Numérico Flotante y Ordinal Flotante, correspondientemente.

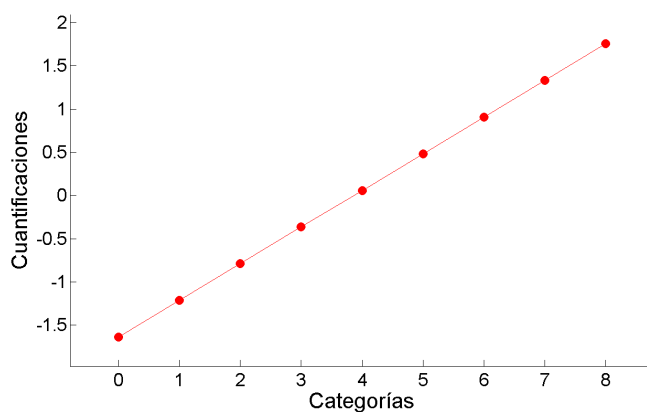


Figura 5.6. Línea recta, típica de los gráficos de cuantificaciones de variables escaladas a nivel Numérico.

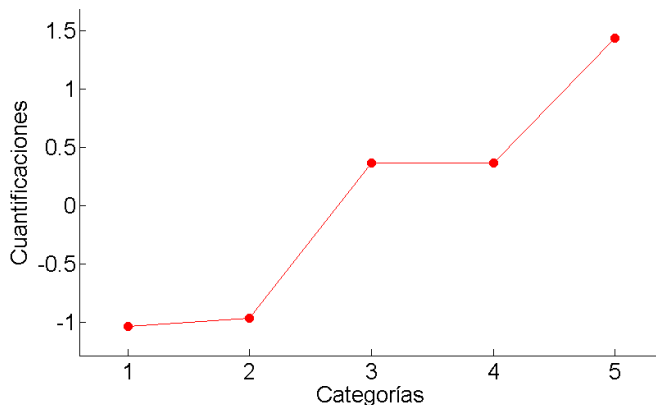


Figura 5.7. Función no decreciente, típica de los gráficos de cuantificaciones de variables escaladas a nivel Ordinal.

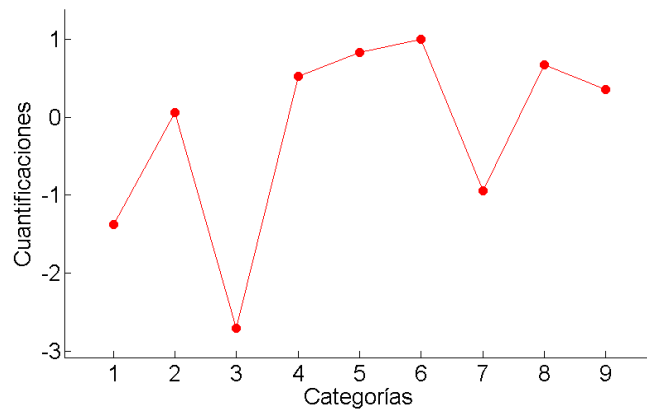


Figura 5.8. Gráfico sin ningún patrón definido, típico de las cuantificaciones de variables escaladas a nivel Nominal.

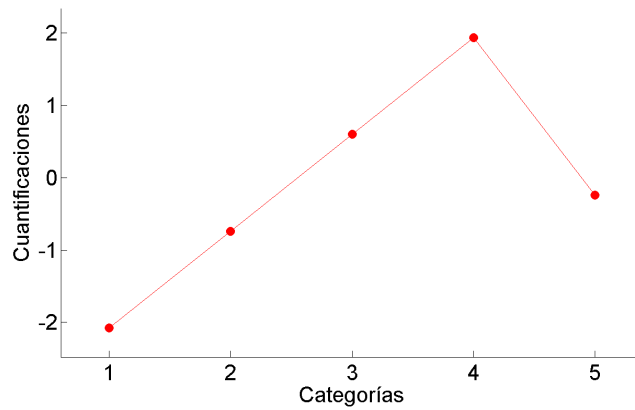


Figura 5.9. Línea recta para las primeras $K-1$ categorías, típica de los gráficos de cuantificaciones de variables escaladas a nivel Numérico Flotante.

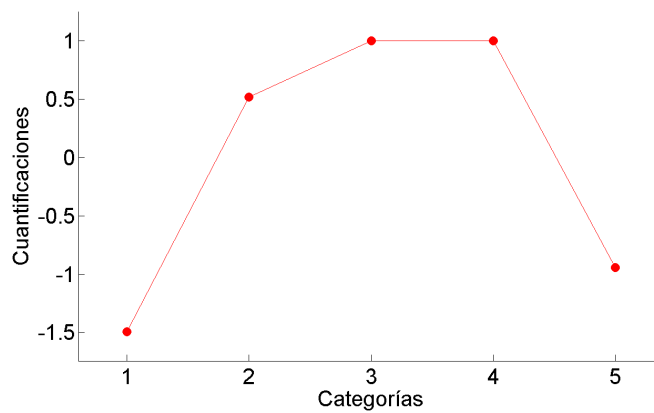


Figura 5.10. Función no decreciente para las primeras $K-1$ categorías, típica de los gráficos de cuantificaciones de variables escaladas a nivel Ordinal Flotante.

5.4 COMPARACIÓN ENTRE CUANTIFICA Y EL SISTEMA GIFÍ

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

Puesto que el producto único de la rutina **CUANTIFICA** es la cuantificación, compararemos nuestra propuesta con las alternativas del sistema Gifi cuando son utilizadas como primer paso para la obtención de cuantificaciones. Posteriormente, en § 6.10, contrastaremos la segunda faceta de las técnicas del sistema Gifi (generación de representaciones gráficas) con nuestra propuesta usada conjuntamente con las técnicas de representación Biplot (GABRIEL, 1971).

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

Aspecto	CUANTIFICA	PRINCALS	HOMALS	OVERALS
Manejo simétrico de todas las variables	Sí	Sí	Sí	No
Posibilidad de agrupación	No	No	No	Sí
Puede usar toda la información aun cuando hay datos faltantes	Sí	Sí	Sí	No
Flexibilidad en niveles de escalamiento	Sí	Sí	No	Sí
Escalamiento Numérico Flotante u Ordinal Flotante	Sí	No	No	No
Cuantificaciones únicas	Sí	No	No	No
Interdistancias mínimas	Sí	No	No	No

Tabla 5.2. Comparación entre **CUANTIFICA** y las principales técnicas del sistema Gifi.

Aunque en términos teóricos las técnicas del sistema Gifi están interrelacionadas entre sí, de modo que algunas pueden concebirse como casos particulares de otras más generales (cf. § 2.7), su implementación algorítmica conlleva algunas diferencias

prácticas. Es por esta razón que aunque PRINCALS y HOMALS pueden considerarse, desde el punto de vista teórico, casos particulares de OVERALS, los hemos incluido en la Tabla 5.2, pues así recogemos los tres módulos de escalamiento óptimo 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 característica de OVERALS, además de implicar un manejo diferente para las variables, exige observaciones completas por grupos, es decir, que para que una observación sea utilizada en el cálculo 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 información disponible, aun cuando haya datos faltantes.

En términos de escalamiento, HOMALS es la técnica más 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**, así como las demás técnicas 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 categorías de la variable con base en el esquema del escalamiento Numérico o del escalamiento Ordinal, pues el escalamiento Nominal implica que todas sus categorías 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 categoría de forma análoga a como lo hace **CUANTIFICA** con las categorías flotantes, es decir, que realiza una cuantificación libre para dicha categoría, respetando para las demás categorías de la variable el nivel de escalamiento definido (bien sea Numérico u Ordinal).

La diferencia entre el mecanismo utilizado por PRINCALS y la forma en que **CUANTIFICA** procesa la información cuando se define un nivel de escalamiento flotante, tiene su origen en la concepción de los niveles de escalamiento flotante. Durante el desarrollo de **CUANTIFICA** hemos considerado la posibilidad de que la variable posea categorías que no sigan el patrón general de las demás, tales como las que surgen cuando las encuestas incluyen un ítem con la etiqueta “*no sabe/no responde*”. Para contrastar esta categoría con las demás categorías de la misma variable así como con las categorías de las demás variables, deberá asignársele una cuantificación, la cual, desde luego, deberá ser libre o flotante, por no conocerse su rango en relación con las demás categorías 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 cuestión además contenga datos faltantes o perdidos, los cuales, en la mayoría de ocasiones no tiene sentido agrupar en una categoría para fines de cuantificación.

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 considerándolos como una categoría extra o no considerarlos en el análisis. PRINCALS no permite manejar parte de la información como una categoría extra (categoría flotante) y otra parte como datos faltantes. Esta diferencia de manejo en relación con **CUANTIFICA** proviene, como ya lo hemos indicado, de la concepción misma de ambos mecanismos. En las técnicas del sistema Gifi no se plantea la posibilidad de que exista una categoría extra que, sin ser información faltante, no siga el patrón de las demás categorías. El proceso señalado se concibe solamente como una forma de manejar la información faltante. Es por ello que en las técnicas del sistema Gifi este proceso no forma parte de los niveles de escalamiento, impidiendo, por tanto, el manejo simultáneo de una categoría extra o flotante y de datos faltantes por otras causas.

En resumen, aunque PRINCALS incluye un mecanismo para manejo de información perdida, que, en ausencia de información perdida, puede servir para realizar una cuantificación 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 técnicas del sistema Gifi no consideran los niveles de escalamiento flotante.

Esta opción de manejo de información faltante, implementada en PRINCALS, permite ejemplificar lo anotado anteriormente sobre técnicas del sistema Gifi que, siendo teóricamente iguales, pueden exhibir diferencias prácticas. En términos teóricos, la solución PRINCALS es equivalente a la solución OVERALS asignando una variable a cada grupo, esto es, un OVERALS sin agrupamiento. En la práctica, sin embargo, si se utiliza el módulo OVERALS con una variable por grupo, para llegar a la solución PRINCALS, no es posible manejar los datos faltantes como categoría extra, tal y como puede hacerse cuando se usa el módulo PRINCALS.

Tal y como se ha indicado y ejemplificado anteriormente, las técnicas del sistema Gifi generan varios posibles conjuntos de cuantificaciones, en función de la dimensionalidad elegida (cf. Capítulo 3). La detección de esta situación y la búsqueda de la mejor solución constituyeron la motivación inicial de la presente investigación. Aunque en el Capítulo 4 presentamos un criterio que permite seleccionar el mejor conjunto de cuantificaciones, partiendo de varios posibles conjuntos, nuestra mayor aportación consiste en la implementación del sistema de cuantificación óptima **CUANTIFICA**, el cual no involucra el concepto de dimensionalidad, proporcionando siempre un único conjunto de cuantificaciones óptimas.

En el Capítulo 4 definimos y sustentamos que 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. Una forma objetiva de medir la satisfacción de tal condición es mediante las *Interdistancias*, criterio escalado de la suma sobre todas las observaciones de las distancias cuadráticas entre las categorías cuantificadas de las diferentes variables. Un sistema de cuantificación óptima deberá tener, por tanto, *Interdistancias* mínimas. La implementación algorítmica de **CUANTIFICA** está basada justamente en la satisfacción de tal condición.

Como síntesis de las anteriores comparaciones, la técnica 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 información disponible aun cuando la matriz tenga datos faltantes y permite usar los niveles de escalamiento Numérico, Ordinal y Nominal.

Nótese que los aspectos que marcan la diferencia entre **CUANTIFICA** y **PRINCALS** son precisamente aquellos que le confieren superioridad a **CUANTIFICA** frente al conjunto de técnicas del sistema Gifi y que resumimos a continuación:

- 1) **CUANTIFICA** brinda la posibilidad de realizar escalamientos flotantes con independencia de si hay o no hay datos faltantes.
- 2) **CUANTIFICA** genera cuantificaciones únicas, al no basar su solución en el concepto de dimensionalidad.
- 3) **CUANTIFICA** genera cuantificaciones que minimizan las *Interdistancias*, con lo cual, las categorías más frecuentemente asociadas reciben cuantificaciones más 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 sintomatología relacionada con la misma, procedente de la Unidad de Cardiología Nuclear y de Ergometría del Hospital Universitario de Salamanca. Aunque la base de datos está estructurada en una matriz de 6.965 pacientes x 53 variables, el 53,12 % de las lecturas de dicha matriz no están presentes, lo cual representa un desafío tanto para la cuantificación como para el posterior análisis, pues en general es más directo trabajar con matrices completas que con matrices que tienen celdas vacías.

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 señalar que la presente base de datos, a pesar del alto porcentaje de información faltante, sigue contando con gran cantidad de información. Nótese que las lecturas efectivamente registradas ascienden a 173.042 observaciones, cifra nada despreciable.

El sistema multivariante objeto de este estudio está conformado por variables con diferentes escalas de medición, 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 Capítulo 5. Las variables óptimamente cuantificadas tienen propiedades numéricas, por lo que seguidamente pueden ser analizadas mediante cualquier técnica lineal.

Puesto que analizamos una base de datos con tanta información, en particular en lo que al número de variables se refiere, nuestro interés se centra en explorar las principales relaciones, más que en contrastar hipótesis específicas. Para el efecto nos apoyamos en los **Métodos Biplot** (GABRIEL, 1971), los cuales permiten resumir y representar los principales patrones de asociación entre variables, individuos, así como entre variables e individuos, partiendo de una matriz numérica de individuos x variables.

Dado que la rutina CUANTIFICA no requiere utilizar ningún método de imputación en el evento de datos faltantes, la matriz de variables óptimamente escaladas carece, asimismo, del 53,12 % de la información. Puesto que las representaciones Biplot, en su forma clásica se basan en la descomposición en valores singulares de la matriz que se desea representar, la cual no puede obtenerse para una matriz incompleta, es necesario construir una representación 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 capítulo de la aplicación práctica, más que en los capítulos de desarrollo teórico, por concebirlos como un complemento a la temática central de este trabajo que es la cuantificación óptima.

Para esta base de datos, en particular, una vez obtenidas las cuantificaciones óptimas, hemos analizado los patrones de asociación más sobresalientes a través de representaciones GCMP-Biplot de las primeras cuatro dimensiones factoriales.

Cerramos este capítulo 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 cardiovascular²³ está conformado por el corazón, los vasos sanguíneos y la sangre. Es el encargado de transportar nutrientes, hormonas y gases hacia todas las células del cuerpo y de recoger los desechos metabólicos que allí se generan. También desempeña un importante papel en la estabilización del pH celular y la temperatura corporal (GANONG, 2005).

Existen diversos factores que pueden causar una disminución transitoria o permanente del riego sanguíneo a un tejido, con la consecuente disminución del aporte de oxígeno al mismo. Al sufrimiento celular resultante de esta situación se le denomina isquemia. La ausencia de oxígeno puede causar daños o disfunción 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 oxígeno se encuentran los del riñón, el cerebro y el corazón (GUYTON y HALL, 2001).

El corazón es el órgano que se encarga de bombear la sangre a través de los vasos sanguíneos, mediante contracciones rítmicas repetidas. Está conformado esencialmente por un tejido muscular denominado miocardio y, en menor proporción, por tejido conectivo y fibroso (tejido de sostén, válvulas). Las arterias coronarias que rodean el corazón son las encargadas de su irrigación sanguínea (GANONG, 2005).

En las arterias sanas, las paredes internas son suaves y de grosor uniforme, permitiendo que los glóbulos rojos y demás sustancias que viajan en la sangre fluyan libremente hacia el corazón. No obstante, con frecuencia se generan depósitos 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 ál., 1997).

²³ En anatomía, la denominación de *sistema* corresponde a un conjunto de órganos formados predominantemente por el mismo tipo de tejidos (como en el caso del sistema nervioso). Puesto que los órganos que contribuyen a la función cardiovascular están formados por diferentes tejidos (miocardio, endotelio, sangre), desde este punto de vista sería más adecuada la denominación *aparato* cardiovascular. No obstante, si se acude a la definición más general de sistema: “conjunto de cosas que relacionadas entre sí contribuyen a determinado objeto”, puede validarse el uso del término *sistema*.

Entre los principales factores de riesgo involucrados en el desarrollo de la arteriosclerosis, se cuentan la predisposición genética, la hipertensión arterial, la diabetes mellitus, la dislipidemia, el sobrepeso/obesidad, el tabaquismo, el sedentarismo y algunos trastornos metabólicos como el hipertiroidismo o los síndromes post menopausia en mujeres (SELWYN et ál., 1997).

El espesor de la placa de ateroma puede ser suficiente para disminuir o bloquear el flujo sanguíneo al miocardio, lo que puede producir un daño cardíaco, técnicamente llamado infarto de miocardio. No obstante, la mayoría de infartos del miocardio se producen cuando la cubierta calcificada de la placa se rompe, dejando expuesto el núcleo graso al torrente sanguíneo, lo que hace que la sangre se coagule y forme trombos, los cuales pueden bloquear el flujo sanguíneo al corazón (BRAUNWALD et ál., 1997).

Por otra parte, las paredes de las arterias coronarias tienen músculo, 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 también por exposición al frío, por fuerte estrés emocional o por el uso de sustancias psicoactivas (BRAUNWALD et ál., 1997).

Los principales síndromes de la cardiopatía isquémica son la angina de pecho (dolor torácico y sensación de opresión aguda y sofocante), el infarto de miocardio y la muerte súbita²⁴.

²⁴ El término *infarto de miocardio* se refiere a la muerte de tejido muscular cardíaco, sin que se ocurra necesariamente la muerte del paciente. La mayoría de paradas cardíacas que conducen a la *muerte súbita* ocurren cuando los impulsos eléctricos se vuelven rápidos (taquicardia ventricular) o caóticos (fibrilación ventricular) o ambos. Este ritmo cardíaco irregular (arritmia) hace que el corazón deje de latir súbitamente, produciéndose la muerte del paciente. Luego, aunque un infarto de miocardio puede causar una parada cardíaca y la muerte del paciente, los términos *muerte por infarto de miocardio* y *muerte súbita* no son sinónimos.

6.3 IMPACTO DE LA CARDIOPATÍA ISQUÉMICA

La Organización Mundial de la Salud señala en su informe de 2004 sobre la salud en el mundo (OMS, 2004) que la cardiopatía isquémica es la principal causa de muerte a nivel mundial, cobrando 7,2 millones de vidas cada año, 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 cánceres de tráquea, bronquios y pulmón. En proyecciones realizadas por la Organización Mundial de la Salud (MATHERS and LONCAR, 2005) se estima que para el año 2030 esta enfermedad seguirá siendo el primer factor de mortalidad a nivel mundial (13,1 %).

Según el Ministerio de Sanidad y Consumo²⁵, durante 2005 se produjeron en España un total de 39.313 defunciones (22.188 en hombres y 17.125 en mujeres) por cardiopatía isquémica. La tasa de mortalidad por 10.000 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, después de la enfermedad cerebrovascular.

OLIVA et ál. (2004) estiman que la pérdida de productividad laboral en España por mortalidad prematura a causa de cardiopatía isquémica asciende a 460,45 millones de euros anuales. Las pérdidas por incapacidad laboral temporal se estiman en 187,06 millones de euros anuales, mientras que la incapacidad permanente genera pérdidas que están entre 431 y 488,5 millones de euros anuales. Luego, las pérdidas totales derivadas de la cardiopatía isquémica oscilarían entre 1.078 y 1.136 millones de euros anuales²⁶. Nótese que en estas estimaciones no se incluyen los costes directos de la prevención 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, España y comunidades autónomas. 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 año 2003.

6.4 BASE DE DATOS

La Unidad de Cardiología Nuclear y de Ergometría del Hospital Universitario de Salamanca evalúa anualmente numerosos pacientes²⁷ remitidos por cardiopatía isquémica diagnosticada o por sospecha de la misma. El historial completo de cada paciente, incluyendo antecedentes personales y familiares, evaluación general, resultados de pruebas y diagnósticos, así como el seguimiento se encuentra sistematizado, contándose actualmente con registros que datan desde 1988 hasta la fecha.

Aunque se han realizado numerosos estudios enfocados en hipótesis específicas (DIEGO DOMÍNGUEZ et ál., 2005; DIEGO DOMÍNGUEZ et ál., 2006; MARTÍN-MOREIRAS et ál., 2005; RAMÍREZ CASTRO et ál., 2006; RUANO et ál., 2005a; RUANO et ál., 2005b), la magnitud de la base de datos, así como su complejidad, dada por la mezcla de variables que la conforman, hacen difícil realizar un proceso exploratorio general que permita usar simultáneamente gran parte de la información disponible y resumir los patrones de asociación más característicos entre las variables, entre individuos, así como entre individuos y variables.

Mediante el presente estudio se pretende hacer una aportación 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 están ausentes, lo cual deja un total de 173.042 observaciones efectivas.

Las bases de datos con información incompleta surgen en la práctica quizá con más frecuencia que las bases de datos completas. Muchas de las bases de datos “completas” usadas en los análisis no son más que versiones reducidas de bases de datos mayores en las cuales se han eliminado las filas con observaciones faltantes.

Aunque esta es una práctica rebatible, eventualmente podría ser viable. Sin embargo, en algunas situaciones, como la presente, esta práctica es totalmente inviable. Si elimináramos todas las filas con al menos una observación faltante, pretendiendo

²⁷ Cada año se realizan alrededor de 2.000 pruebas de esfuerzo y 1.000 SPECT.

mantener las 53 variables, nos quedaríamos con cero pacientes. Si se eliminan sucesivamente las variables con menos lecturas, obtendríamos 107 pacientes con lecturas completas en 48 variables. Si seguimos reduciendo el número de variables, con el fin de incluir un mayor número de pacientes, tendríamos que trabajar con solamente 39 variables para poder tener 1.054 pacientes con observaciones en todas las variables. Esto, desde luego, iría en detrimento del análisis multivariante que pretendemos realizar.

Otra alternativa para el manejo de bases de datos incompletas consiste en utilizar algún método de imputación que permita estimar los valores faltantes con base en la información observada. Si bien, esto facilitaría bastante todos los procesos subsiguientes, pues, en general, es más fácil trabajar con matrices completas que con matrices en las que falte información, los resultados obtenidos con base en las matrices imputadas se verían afectados por el método de imputación utilizado.

Antes de plantear algún eventual método de imputación, hay que considerar que las matrices incompletas pueden tener diversos orígenes. Puede tratarse de matrices en las que las celdas vacías son producto de información efectivamente generada, que por algún motivo no llegó 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 vacías. Utilizaremos el término **datos faltantes** para definir esta segunda situación.

En el primer caso, puede ser viable y hasta recomendable utilizar algún método de imputación que permita estimar de la mejor manera posible la información perdida. En el segundo caso, ello no tendría sentido y **cualquier intento por estimar la información que nunca ha existido no haría más que introducir ruido**.

Las matrices con información faltante —mas no perdida— son comunes en bases de datos provenientes de la práctica clínica, donde cada paciente puede ser sometido a diferentes tipos de pruebas. El escenario del presente análisis se corresponde con dicha situación. En este caso, la Unidad de Cardiología Nuclear y de Ergometría utiliza diferentes pruebas diagnósticas, v. gr., SPECT, ecocardiografía y coronariografía.

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 específico solo es evaluado mediante SPECT, no tiene ningún sentido tratar de estimar los resultados de la ecocardiografía y de la coronariografía, pues no son datos perdidos, sino datos que nunca han existido.

En § 6.7 tomaremos en consideración este hecho para elaborar nuestra propuesta de representaciones Biplot. Entretanto, vale la pena recordar que nuestra propuesta de cuantificación, implementada en la rutina **CUANTIFICA**, utiliza el 100 % de la información disponible, sin hacer uso de ningún método de imputación (cf. § 5.2). Esto implica que la matriz cuantificada tiene exactamente la misma estructura de datos faltantes que la matriz original.

Más allá de la indiscutible importancia que desde el punto de vista clínico tiene el estudio de la cardiopatía isquémica, la base de datos utilizada en el presente estudio resulta óptima desde el punto de vista técnico para ilustrar la funcionalidad de la rutina **CUANTIFICA** como herramientas de cuantificación 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 más complejos que los que podrían realizarse sobre las matrices completas correspondientes a bases de datos sin información faltante. Por otra parte, el considerable tamaño de la base de datos, tanto en número de pacientes como en número de variables, aunado a las diferentes escalas de medición de las variables que conforman el sistema multivariante ponen a prueba los diferentes componentes de la rutina **CUANTIFICA**.

A continuación nombramos las 53 variables que conforman la base de datos:

BRI:	Bloqueo de rama izquierda del haz de his.
CI:	Cardiopatía isquémica en el diagnóstico preliminar.
EA:	Estado actual.
DM:	Diabetes mellitus.
HTA:	Hipertensión arterial sistémica.
Edad:	Edad.
IMC:	Índice de masa corporal.
Sexo:	Sexo.

SSS manual DA:	Score manual en estrés de defectos en territorio de la arteria descendente anterior.
SSS manual Cx:	Score manual en estrés de defectos en territorio de la arteria circunfleja.
SSS manual CD:	Score manual en estrés de defectos en territorio de la arteria coronaria derecha.
SSS manual Total:	Score manual total de defectos en estrés.
SSS auto DA:	Score automático en estrés de defectos en territorio de la arteria descendente anterior.
SSS auto Cx:	Score automático en estrés de defectos en territorio de la arteria circunfleja.
SSS auto CD:	Score automático en estrés de defectos en territorio de la arteria coronaria derecha.
SSS auto Total:	Score automático total de defectos en estrés.
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 automático diferencia de defectos en territorio de la arteria descendente anterior.
SDS auto Cx:	Score automático diferencia de defectos en territorio de la arteria circunfleja.
SDS auto CD:	Score automático diferencia de defectos en territorio de la arteria coronaria derecha.
SDS auto Total:	Score automático total de defectos por diferencia.
NSE:	Número de segmentos afectados en estrés.
NSR:	Número de segmentos afectados en reposo.
Clínica:	Resultado clínico de la prueba de estrés para angina de pecho.
Eléctrica:	Resultado eléctrico de la prueba de esfuerzo.
Necrosis:	Necrosis.
Isquemia:	Isquemia.
Di.VIE:	Dilatación de ventrículo izquierdo al estrés.
Di.VIR:	Dilatación de ventrículo izquierdo en reposo.
Di.VID:	Dilatación transitoria del ventrículo izquierdo, obtenida por diferencia.
VTDE:	Volumen telediastólico estandarizado de ventrículo izquierdo en estrés.
VTDR:	Volumen telediastólico estandarizado de ventrículo izquierdo en reposo.
VTSE:	Volumen telesistólico estandarizado de ventrículo izquierdo en estrés.
VTSR:	Volumen telesistólico estandarizado de ventrículo izquierdo en reposo.
FEE:	Fracción de eyección en estrés calculada por SPECT.
FER:	Fracción de eyección en reposo calculada por SPECT.
FE eco:	Fracción de eyección calculada por ecocardiografía.
FE hemo:	Fracción de eyección 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 cirugía de revascularización coronaria durante el seguimiento.
Infarto:	Infarto de Miocardio durante el seguimiento.
Vasos hemo:	Número de vasos afectados, evaluados por hemodinamia.

6.5 DESCRIPCIÓN DE ALGUNAS VARIABLES

Con el fin de facilitar la interpretación de los resultados que se presentan en § 6.8, a continuación se detalla la forma de obtención 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 células del cuerpo y de recoger los desechos metabólicos que allí se generan. Existen diversos factores que pueden causar una disminución transitoria o permanente del riego sanguíneo a un tejido, con la consecuente disminución del aporte de oxígeno al mismo. Al sufrimiento celular resultante de esta situación se le denomina **Isquemia**.

En general, la **Isquemia** miocárdica se produce a causa de un desequilibrio entre la demanda y el aporte de oxígeno al músculo cardíaco. Esto puede suceder ante dos eventos: en primer lugar, por un incremento de los requerimientos de oxígeno del miocardio (por ejemplo, al realizar ejercicio) que no vaya acompañado de un incremento del aporte de oxígeno por la sangre (en el caso de obstrucciones parciales de las arteriales coronarias²⁸) o, en segundo lugar, por una disminución 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 músculo cardíaco sean mínimos.

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 carácter irreversible.

Existen una serie de pruebas diagnósticas que permiten caracterizar la gravedad y las particularidades de la cardiopatía isquémica. El sistema multivariante que es objeto de este estudio incluye resultados de SPECT, de coronariografías y de ecocardiografías.

El **SPECT** (*Single Photon Emission Computed Tomography*) es una técnica que permite generar imágenes tridimensionales del corazón a partir de la detección de la

²⁸ A este evento se le denomina disminución de la reserva vasodilatadora coronaria.

radiación gamma emitida por un trazador radioactivo administrado previamente al paciente²⁹. La comparación de los grados de captación miocárdica bajo condiciones de estrés cardiaco y de reposo permite diagnosticar condiciones de normalidad, isquemia, y/o necrosis, entre otras.

El Gated SPECT es una técnica de punta que incorpora mejoras en relación con el SPECT estándar tanto en la forma en que la gammacámara realiza el barrido como en los algoritmos utilizados para la reconstrucción de las imágenes, lo cual permite obtener imágenes de diferentes etapas del ciclo cardiaco, incrementando la información clínica que de estas puede extraerse. En la Figura 6.1 se muestra algunas imágenes típicas de este tipo de pruebas.

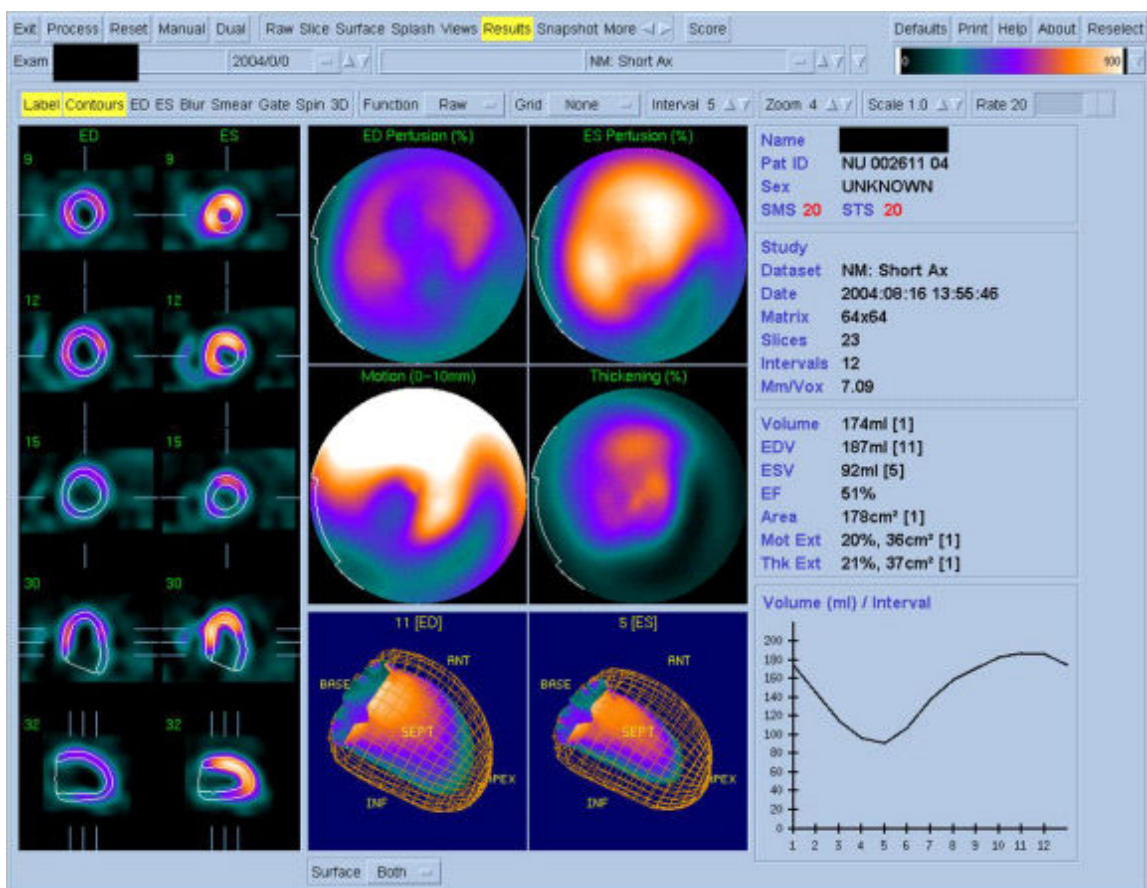


Figura 6.1. Imágenes típicas de un Gated SPECT.

²⁹ Suele utilizarse Tc-99m MIBI, Talio 201 y Tc-99m Tetrafosmín.

Como resultado del procesamiento del SPECT aparecen múltiples indicadores relacionados con la perfusión miocárdica. Estos indicadores pueden ser generados de forma automática por el sistema o pueden surgir del análisis de las imágenes 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 genérico de *Scores*, los cuales son sumas de puntajes de defectos en los vasos de una región determinada, con lo cual reflejan conjuntamente la intensidad y la extensión de los defectos. Su medición puede realizarse en estrés (*Summed Stress Score: SSS*), en reposo (*Summed Rest Score: SRS*) o pueden obtenerse como la diferencia entre el puntaje en estrés y el puntaje en reposo (*Summed Difference Score: SDS = SSS – SRS*). El estrés cardiaco puede alcanzarse mediante la realización de esfuerzo físico en una cinta sinfín, mediante la aplicación de fármacos como el dipiridamol y la dobutamina o combinando ambos métodos.

Las arterias coronarias, que rodean el miocardio formando una corona, son las encargadas de su irrigación. La **arteria descendente anterior (DA)**, que es la principal arteria coronaria, irriga la cara anterior y la punta del ventrículo izquierdo, así como el tabique interventricular; la **arteria coronaria derecha (CD)** es la responsable de irrigar el ventrículo derecho y la región inferior y posterior del ventrículo izquierdo; la **arteria circunfleja (Cx)** irriga la cara lateral del ventrículo izquierdo.

Esta diferenciación genera el concepto de territorios irrigados por cada una de las arterias coronarias. No obstante, es importante anotar que se trata de un concepto esquemático y relativamente amplio, dada la gran variabilidad en la delimitación de los territorios irrigados por cada rama coronaria entre pacientes y la existencia de circulación cruzada entre territorios.³⁰

³⁰ Como ejemplo de esta variabilidad, está el origen de la arteria descendente posterior, la cual irriga la cara diafragmática del ventrículo 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 (estrés o reposo) y la forma en que se genera la lectura (automática o analógica), surgen diferentes puntuaciones o scores que evalúan conjuntamente la extensión y la intensidad de los defectos en dicho territorio. Así, por ejemplo, la variable **SSS auto Cx** indica la extensión e intensidad de defectos en territorio de la arteria circunfleja, evaluado de manera automática bajo condiciones de estrés cardiaco. Los scores totales son la suma de los defectos en los tres territorios considerados.

Como resultado del SPECT, también se calculan los volúmenes telesistólico y telediastólico del ventrículo izquierdo, en condiciones de estrés o de reposo (**VTSE**, **VTDE**, **VTSR**, **VTDR**). Estos son los volúmenes que presenta el ventrículo izquierdo hacia el final de la sístole (contracción) y de la diástole (relajación), respectivamente.

La evaluación de los anteriores volúmenes da lugar al concepto de dilatación del ventrículo izquierdo en condiciones de estrés o de reposo (**Di.VIE**, **Di.VIR**). Asimismo, surge el concepto de dilatación transitoria del ventrículo izquierdo, obtenida por diferencia entre la dilatación en estrés y la dilatación en reposo (**Di.VID**).

Como resultado del SPECT también puede obtenerse el número de segmentos afectados en estrés o en reposo (**NSE**, **NSR**), como indicador general de afección cardiaca sin diferenciar territorios.

La **fracción de eyección** es el predictor más importante de morbimortalidad de origen cardiaco. Esta variable expresa en términos porcentuales la disminución del volumen del ventrículo izquierdo en sístole (contracción), con respecto a la diástole (relajación).³¹ Así, por ejemplo, una fracción de eyección del 50 % significa que el volumen del ventrículo izquierdo en máxima contracción (al final de la sístole) es la mitad de su volumen en máxima relajación (al final de la diástole). Un corazón con poca capacidad contráctil (fracción de eyección baja) tiene poca capacidad de bombeo. En estos casos se habla de insuficiencia cardiaca. La **fracción de eyección** es calculada de

³¹ $FE = \frac{Vol. VI Diástole - Vol. VI Sístole}{Vol. VI Diástole} \times 100$

forma automática por el SPECT, bien sea en condiciones de estrés o de reposo (**FEE**, **FER**).

Adicionalmente al SPECT, en ocasiones se requieren pruebas de tipo invasivo o hemodinámicas, conocidas con el nombre genérico de **coronariografías** o **cateterismos cardiacos**. Estas pruebas se realizan mediante la inserción de un catéter a través de la arteria femoral (región inguinal) o radial (brazo) hasta las arterias coronarias, en las que se inyecta un medio de contraste que facilita la visualización de imágenes en las que se pueden apreciar posibles estrechamientos del paso de sangre u oclusiones completas.

A partir de los cateterismos también puede evaluarse la fracción de eyección. En este caso, se denomina fracción de eyección calculada por hemodinamia (**FE hemo**). La fracción de eyección también puede obtenerse de forma incruenta usando ultrasonidos, mediante ecocardiografía (**FE eco**).

La coronariografía también permite determinar el grado de estenosis o estrechamiento de las arterias coronarias (**Estenosis DA**, **Estenosis Cx**, **Estenosis CD**) y el número de vasos afectados (**Vasos hemo**). Esta última variable contabiliza el número de vasos con algún 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 capítulo es de tipo mixto, puesto que está conformado por variables con diferentes escalas de medición, v. gr., Numérica, Ordinal y Nominal (cf. § 1.2).

Nuestra propuesta de análisis consiste en generar cuantificaciones óptimas con base en los desarrollos del Capítulo 5. Seguidamente, aprovechando las propiedades métricas de las cuantificaciones así obtenidas, construiremos una representación conjunta de los pacientes y las variables, utilizando un método lineal como es el de las representaciones Biplot (GABRIEL, 1971).

Para realizar el proceso de cuantificación, todas las variables, sin importar su nivel de escalamiento, deben someterse a un proceso de categorización o discretización (cf. § 5.3.1). A continuación se presenta la categorización utilizada para las variables que conforman el sistema multivariante objeto del presente estudio:

BRI:	No, Sí.
CI:	No, Sí.
EA:	Asintomático, Sintomático.
DM:	No, Sí.
HTA:	No, Sí.
Edad:	<30, 30-55, 56-80, >80.
IMC:	<25 (Normal), 25-30 (Sobrepeso), >30 (Obesidad).
Sexo:	Femenino, Varón.
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, elevación ST.
Necrosis:	No, Sí.
Isquemia:	No, Sí.
Di.VIE:	No, Sí.
Di.VIR:	No, Sí.
Di.VID:	No, Sí.
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, Sí.
CRC:	No, Sí.
Infarto:	No, Sí.
Vasos hemo:	0, 1, 2, 3.

En § 2.5 hemos establecido un paralelo entre la escala de medición de una variable y su nivel de escalamiento. Tal y como se indica al final de dicho apartado, la utilización de un nivel de escalamiento más flexible que aquel que correspondería naturalmente a la escala de medición de una variable puede resultar útil para evaluar posibles relaciones

no lineales de una variable con otras variables del sistema. Con base en tal consideración y teniendo en cuenta lo restrictivo que resulta el nivel de escalamiento Numérico, 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 demás variables se escalaron a nivel Ordinal.

El nivel de escalamiento utilizado para las variables binarias es irrelevante, obteniéndose siempre las mismas cuantificaciones o cuantificaciones con signo contrario, cualquiera que sea el nivel de escalamiento elegido. La única razón 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 categorías de ausencia (No, Asintomático) y positivas para las categorías de presencia (Sí, Sintomático), lo cual facilitará 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 aplicación CUANTIFICA.

Sexo	Edad	FE eco	Necrosis	Clínica
3	2	2	2	3
1	1	1	1	1
V	56-80	>50	No	Positiva
V	>80	41-50	No	Positiva
M	30-55	21-40	No	Negativa
M	30-55	21-40	Sí	Dudosa
M	56-80	>50	Sí	Positiva
V	>80	<20	Sí	Negativa
M	30-55	21-40	Sí	Positiva
V	<30	>50	No	Positiva
V	56-80	21-40	Sí	Negativa
M	>80	>50	No	Dudosa

Sexo	Edad	FE eco	Necrosis	Clínica
3	2	2	2	3
1	1	1	1	1
2	3	4	0	1
2	4	3	0	1
1	2	2	0	0
1	2	2	1	2
1	3	4	1	1
2	4	1	1	0
1	2	2	1	1
2	1	4	0	1
2	3	2	1	0
1	4	4	0	2

Figura 6.2. Preparación de la base de datos para su procesamiento mediante la aplicación 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 números entre 1 y 5, correspondientes a los niveles de escalamiento Numérico, Ordinal, Nominal, Numérico 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 cuantificación óptima (cf. § 5.3.1). Entre la

cuarta y la última fila se disponen las etiquetas de cada una de las categorías, las cuales deben estar conformadas por números enteros que respeten la ordinalidad de las categorías 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 óptimo a través de la aplicación **CUANTIFICA**, se genera la Tabla de Cuantificaciones, la Matriz de Cuantificaciones, así como un gráfico de cuantificaciones para cada una de las variables.

No obstante que en este capítulo, la cuantificación no es el fin último de nuestro análisis, sino que constituye un paso de adaptación del sistema multivariante para su posterior análisis con base en técnicas lineales, resaltaremos algunos aspectos relacionados con los resultados obtenidos al procesar la base de datos objeto de este estudio mediante la aplicación **CUANTIFICA**. Para una reseña general sobre los resultados que genera la aplicación, véase § 5.3.7.

Vale la pena resaltar que en solo 2 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 cuantificación 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 cuantificación para dos categorías adyacentes de una variable escalada a nivel Ordinal significa, en general, que no es posible diferenciar el comportamiento de las dos categorías en cuestión, en cuanto a sus relaciones con las categorías de las demás variables, bajo las restricciones impuestas por el escalamiento Ordinal. En particular, en este caso, podemos afirmar que, en la muestra analizada, no se observó 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 demás variables. Desde luego, lo mismo puede decirse en referencia a la variable **SDS manual Cx**.

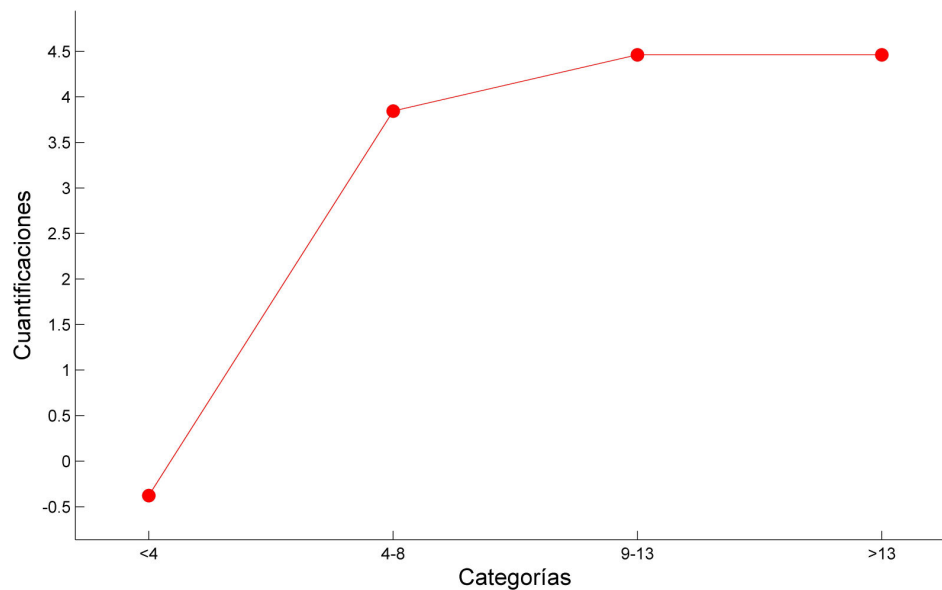


Figura 6.3. Gráfico de cuantificaciones para “Score manual en reposo de defectos en territorio de la arteria circunfleja”.

Si bien el hecho de obtener la misma cuantificación en categorías adyacentes de variables escaladas a nivel Ordinal no representa en sí un problema, vale la pena reflexionar al respecto cuando se presente esta situación, pues ello puede ser reflejo de diferentes circunstancias. Por una parte, podría estar indicando una sobrecategorización de la variable en cuestión, tratando de diferenciar niveles que, para fines prácticos, tienen el mismo comportamiento. Por otra parte podría estar indicando que la relación entre los niveles de la variable no es Ordinal, o al menos que no sigue el orden esperado.

En el caso de sobrecategorización de una variable no se requiere tomar ninguna medida adicional. Dependiendo de cuál 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 cuantificación en un solo nivel, pero ello no alteraría los resultados.

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

aplicarían cuantificaciones libres de restricciones que pudieran reflejar mejor las relaciones de las categorías de dicha variable con las categorías de las demás variables.

En el presente caso, en el que la ordinalidad de las categorías de las dos variables en cuestión es un hecho indiscutible, no es necesario realizar ninguna acción adicional, no requiriéndose ninguna modificación para utilizar estos resultados en la siguiente etapa.

Para satisfacer el objetivo final del presente capítulo, consistente en analizar las principales relaciones entre individuos, variables y variables e individuos a través de representaciones Biplot, la Matriz de Cuantificaciones constituye el resultado más importante de la aplicación 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 códigos 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 numéricos contenidos en la Matriz de Cuantificaciones tienen propiedades métricas, lo cual permite su uso como información de entrada de cualquier técnica lineal. En particular, en este capítulo, utilizaremos la Matriz de Cuantificaciones como información de entrada para la obtención 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 aplicación CUANTIFICA y que se utiliza como información de entrada para las representaciones Biplot.

En el panel (a) de la Figura 6.4 se presenta un sector hipotético de la base de datos original, dejando vacías las celdas cuando no se haya registrado la información de una variable para un paciente específico. El panel (b) corresponde a la base de datos categorizada, en la cual se reemplaza cada uno de los valores originales por la denominación de la categoría a la cual pertenece (cf. § 1.5). En el panel (c) se reemplaza cada una de las denominaciones de las diferentes categorías por una etiqueta numérica, 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 aplicación CUANTIFICA.

(a): BD Original

Sexo	Edad	FE.Eco	Necrosis	Clinica
V	57	53	No	Positiva
V	83	45	No	Positiva
M		34	No	
M	47		Si	Dudosa
M	65	67	Si	
V	89	15		Negativa
M	33	39	Si	Positiva
V	29		No	
V	60	33	Si	Negativa
M	81	55	No	Dudosa



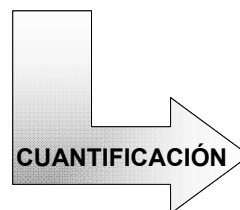
(b): BD Categorizada

Sexo	Edad	FE.Eco	Necrosis	Clinica
V	56-80	>50	No	Positiva
V	>80	41-50	No	Positiva
M		21-40	No	
M	30-55		Si	Dudosa
M	56-80	>50	Si	
V	>80	<20		Negativa
M	30-55	21-40	Si	Positiva
V	<30		No	
V	56-80	21-40	Si	Negativa
M	>80	>50	No	Dudosa



(c): BD Etiquetada

Sexo	Edad	FE.Eco	Necrosis	Clinica
2	3	4	0	1
2	4	3	0	1
1		2	0	
1	2		1	2
1	3	4	1	
2	4	1		0
1	2	2	1	1
2	1		0	
2	3	2	1	0
1	4	4	0	2



(d): M. Cuantificaciones

Sexo	Edad	FE.Eco	Necrosis	Clinica
0.73	0.36	1.85	-2.06	1.86
0.73	2.64	-1.11	-2.06	1.86
-1.37		-4.05	-2.06	
-1.37	-1.91		1.61	-0.33
-1.37	0.36	1.85	1.61	
0.73	2.64	-7.01		-0.59
-1.37	-1.91	-4.05	1.61	1.86
0.73	-4.19		-2.06	
0.73	0.36	-4.05	1.61	-0.59
-1.37	2.64	1.85	-2.06	-0.33

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 prácticos 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 aplicación CUANTIFICA (panel (c)), siendo suficiente con tomar nota de la denominación

correspondiente a cada categoría, para fines de la posterior interpretación y descripción de resultados.

Nótese que en los diferentes pasos se trabaja con la información disponible, manteniendo inalterada la información faltante, la cual sigue estando ausente en la Matriz de Cuantificaciones. Aunque hemos recomendado utilizar el código numérico -99 como marcador de las celdas faltantes en la base de datos que es leída por la aplicación CUANTIFICA (cf. § 5.3.1), en la Figura 6.4 hemos dejado tales celdas en blanco para facilitar la visualización 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 número de filas y número de columnas, sino también en lo que tiene que ver con información faltante. Es decir que, para el presente caso, la Matriz de Cuantificaciones carece del mismo 53,12 % de información 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 BILOT

6.7.1 BASE TEÓRICA.

La representación Biplot de una matriz X , propuesta por GABRIEL (1971), se basa en la obtención 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 -ésima fila y la j -ésima 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, así como en este caso, variables) de la matriz X en un subespacio de dimensión reducida; de ahí, la partícula 'Bi' que encabeza el nombre de la representación.

Para obtener la representación completa de una matriz de rango r se requeriría un espacio r -dimensional, lo cual no resulta práctico para matrices de rango mayor de dos. Es posible, sin embargo, obtener una representación aproximada de cualquier matriz por medio de la representación gráfica de su correspondiente matriz de aproximación a bajo rango.

Aunque la aproximación de rango dos es la más popular, por ser la que permite obtener la mejor representación bidimensional de los principales patrones de correlación y asociación del sistema multivariante, no es la única aproximación 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 prácticas 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 número de filas, m el número 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 factorización A y B con los marcadores vectoriales para las filas y las columnas de la matriz X . Las más populares se basan en la descomposición en valores singulares de X ($X = U\Sigma V'$), puesto que esta metodología hace posible obtener de manera simultánea la aproximación a bajo rango de X y los marcadores vectoriales basados en tal aproximación.

Para obtener la aproximación 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 (Σ_p , 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 $X'X$). $X_{(p)} = U_p \Sigma_p V_p'$ es la aproximación de rango p de X (HOUSEHOLDER and YOUNG, 1938).

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

$$A = U \quad ; \quad B = V\Sigma \quad \text{(GH-Biplot ó CMP-Biplot)}$$

$$A = U\Sigma \quad ; \quad B = V \quad \text{(JK-Biplot ó RMP-Biplot)}$$

$$A = U\Sigma^\alpha \quad ; \quad B = V\Sigma^{1-\alpha}, \text{ con } \alpha \in [0, 1] \quad \text{(Si } \alpha = 0,5, \text{ Biplot Simétrico)}$$

Como consecuencia de la elección de los marcadores vectoriales, la representación de las filas o de las columnas puede resultar favorecida en relación con la representación del otro conjunto.

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

de representación, quedando las columnas representadas en forma estándar. Es posible elegir cualquier configuración 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 RMP-Biplot; si $\alpha=0,5$ se tiene un Biplot Simétrico, en el que las filas y las columnas aparecen con igual calidad de representación.

En adición a estas opciones clásicas, existe una forma de elegir los marcadores tal que se maximice la calidad de representación tanto para filas como para columnas. Dicha solución, propuesta por GALINDO (1985, 1986), consiste en elegir $A=U\Sigma$ y $B=V\Sigma$. Nótese que esta solución es equivalente a tomar los marcadores columna, o H, del GH-Biplot, en el cual las columnas gozan de la máxima calidad de representación, y los marcadores fila, o J, del JK-Biplot, en el cual las filas tienen la máxima calidad de representación. Por esta razón, a la representación así obtenida se le denomina HJ-Biplot o RCMP-Biplot (*Row Column Metric Preserving Biplot*). Como contraparte a la máxima calidad de representación para filas y columnas brindada por esta representación, se sacrifica la propiedad de que el producto interno del i -ésimo marcador fila y el j -ésimo marcador columna reproduzca el elemento de la celda x_{ij} , el cual solo sería estimado en este caso. Nótese que, bajo esta selección de marcadores, $AB'=U\Sigma\Sigma V'=U\Sigma^2 V' \approx X$.

En el presente caso, donde no existen grupos naturales de pacientes, diferenciándose estos solamente por los valores de las variables, estamos particularmente interesados en representar la estructura de correlaciones de las variables con la máxima calidad posible. Debe tenerse en cuenta, sin embargo, el ingrediente adicional que aparece al intentar construir una representación Biplot para una matriz con datos faltantes, lo cual no puede hacerse mediante la aplicación directa de ninguna de las expresiones clásicas para la obtención de marcadores, pues todas ellas están basadas en la descomposición en valores singulares de la matriz, la cual no puede realizarse para matrices con datos faltantes.

Una opción viable en otras circunstancias consistiría en utilizar algún método para imputar la información faltante, tras lo cual podrían obtenerse los marcadores vectoriales mediante la descomposición en valores singulares. No obstante, tal y como

fue señalado anteriormente (cf. § 6.4), en el presente caso, la información faltante no es información perdida, sino información que nunca fue medida debido a que al paciente no se le realizó esa prueba particular. Luego, cualquier intento por estimar esa información que nunca ha existido no haría más que introducir ruido.

Nuestra propuesta para generar la representación Biplot de una matriz X con datos faltantes —mas no perdidos— consiste en obtener los marcadores columna de una hipotética 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 obtención de los marcadores fila de la matriz X .³²

Para el efecto, se parte del cálculo de las correlaciones por pares entre las variables de la matriz incompleta, X . La correlación de cada par de variables se calcula con base en toda la información disponible para ese par, sin importar si en esa fila hay información disponible o no para las demás variables. Considérese la información 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 información faltante.

³² En adelante, en este apartado, denotaremos por X a la matriz de cuantificaciones generada por la aplicación CUANTIFICA al realizar un proceso de cuantificación óptima sobre un sistema multivariante con datos faltantes. Si bien, esta matriz es un eslabón intermedio del proceso de análisis que se desarrolla en este capítulo, en este apartado nos referiremos a ella como la matriz de datos original.

Así, en el cálculo de r_{12} participarían 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 , así:

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mm} \end{pmatrix}$$

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

En consecuencia con lo indicado anteriormente acerca de nuestro interés por generar una representación Biplot en la que las variables tengan máxima calidad de representación, y teniendo en cuenta que la descomposición en valores singulares de la hipotética matriz X^* tiene la forma: $X^* = U\Lambda^{1/2}V'$, elegimos la matriz de marcadores columna como $B = V\Lambda^{1/2}$, elección correspondiente a los marcadores columna de un CMP-Biplot. Nótese que Λ es la matriz diagonal con los valores propios de la matriz $X^*{}'X^*$, mientras que $\Lambda^{1/2} \equiv \Sigma$ es la matriz diagonal con los valores singulares de X^* .

Puesto que la hipotética matriz X^* tiene una estructura de correlaciones igual a la de X , los marcadores columna obtenidos para X^* son igualmente válidos para X . Esta elección de marcadores bastaría para obtener una representación de las columnas de la matriz X , denominada H-Plot (GABRIEL, 1981, JOBSON, 1992). No obstante, es posible obtener también marcadores fila para generar una representación Biplot de X .

Para tal efecto, consideramos una hipotética descomposición en valores singulares de X , así $X = U\Lambda^{1/2}V'$. Obsérvese que la parte correspondiente a $\Lambda^{1/2}V'$ es conocida, siendo precisamente la que se obtuvo mediante la descomposición en valores y vectores propios de R . Si se conociera la matriz U , sus filas servirían de marcadores fila para las correspondientes filas de la matriz X , lo que permitiría, en conjunción con los marcadores columna ya definidos, elaborar una representación Biplot para la matriz X .

Si X fuera una matriz completa, podríamos obtener la matriz U así:

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

$$XB = UB'B$$

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

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

Aunque los datos faltantes en X impiden el uso directo de la anterior expresión para calcular la matriz U que contendría 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 ponderación cero para los valores faltantes.

La i -ésima fila de U , sin incluir aun la ponderación cero, se calcularía así:

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

Al trasponer esta expresión surge la familiar ecuación para la estimación del vector de parámetros β en regresión múltiple:

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

Luego, cada marcador fila es igual al vector de parámetros de una regresión múltiple de la correspondiente fila de la matriz original sobre los marcadores columna. Nótese, sin embargo, que a pesar de la analogía con el modelo de regresión, la anterior expresión no involucra un término de error. Por tanto, el producto entre la matriz de marcadores

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

Volviendo al hecho de que el vector x_i puede tener celdas vacías, y siguiendo la analogía con el modelo de regresión, u_i se calcula como el vector de parámetros de una regresión en la que se asigna ponderación 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 vacías.

La regresión ponderada tiene la siguiente forma:

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

Dado que la incorporación de ponderaciones nulas a la matriz $B'B$ puede generar problemas de singularidad en la matriz ponderada $B'W_iB$, impidiendo la obtención de su inversa, es necesario agregar una pequeña cantidad a su diagonal para corregir esta situación sin alterar significativamente los resultados. Este mecanismo es análogo al de la regresión *ridge* y es mencionado por GABRIEL et ál. (1998), quienes también lo utilizan para sortear problemas de singularidad. En este trabajo obtuvimos resultados satisfactorios sumando 1×10^{-9} a la diagonal de la matriz en cuestión.

La regresión ponderada para la obtención del i -ésimo marcador fila, incluyendo la matriz diagonal K que rompe la singularidad de $B'W_iB$, tiene la siguiente forma:

$$u_i = (B'W_iB + K)^{-1}B'W_ix_i$$

En resumen, la representación Biplot que proponemos utiliza los siguientes marcadores vectoriales:

$$B = V\Lambda^{1/2}; A = [a_1, a_2, \dots, a_i, \dots, a_n]', \quad a_i = (B'W_iB + K)^{-1}B'W_ix_i, \quad i = 1, \dots, n$$

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

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 .

$\Lambda^{1/2}$: Matriz diagonal de la raíz cuadrada de los valores propios de R .

a_i : Marcador fila correspondiente al i -ésimo individuo.

W_i : Matriz diagonal indicadora de observaciones faltantes para el i -ésimo individuo.

K : Matriz diagonal con una pequeña constante (1×10^{-9}).

x_i : i -ésima fila de la matriz original, en la que se remplazan los valores faltantes por un valor numérico arbitrario.

La representación Biplot obtenida con base en estos marcadores tiene las siguientes propiedades:

- 1) La longitud de los vectores que representan a las variables estiman la desviación estándar de las correspondientes variables.
- 2) Los cosenos de los ángulos entre los vectores que representan a las variables, aproximan la correlación entre las correspondientes variables representadas.
- 3) La proyección del i -ésimo marcador fila sobre el j -ésimo 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 interés en reproducir lo más fielmente posible la estructura de correlaciones observada, así como los datos observados. La utilización de los marcadores vectoriales propuestos satisface ambos objetivos.

El hecho de que las correlaciones originales solo sean estimadas en la representación Biplot en lugar de reproducidas no es producto de la elección de marcadores, sino de la

reducción de dimensionalidad. Nótese 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 máxima calidad posible para dicho subespacio de representación.

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 reducción de dimensionalidad. El producto interno entre los marcadores fila y columna sin reducción de dimensionalidad reproducen casi exactamente el dato observado. Tal ‘reproducción’ solo presenta una muy leve discrepancia con el dato original, a causa de la constante $K (1 \times 10^{-9})$ sumada a la matriz $B'W_i B$ 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 proyección de un marcador fila sobre un marcador columna, solamente debe usarse para estimar los valores observados en X , no teniendo ningún sentido utilizarlo como estimación de valores que nunca han existido.

Es notable la similitud entre nuestra propuesta de representación para una matriz incompleta y el CMP-Biplot de GABRIEL (1971). Los marcadores columna utilizados en nuestra propuesta de representación son equivalentes a los del CMP-Biplot de una matriz completa, las propiedades de la representación que proponemos son análogas a las del CMP-Biplot y la aplicación de nuestra propuesta a una matriz sin datos faltantes genera un CMP-Biplot. No obstante, no podemos decir, en sentido estricto, que nuestra representación sea un CMP-Biplot, al menos no si consideramos que este es un término ya acuñado para referirse a la representación Biplot de matrices completas.

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

6.7.2 RUTINA MARCADORES GCMP-BILOT.

Hemos elaborado una rutina computacional en MATLAB, que sistematiza el proceso de generación 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 gráfica como la que se muestra en la Figura 6.5.



Figura 6.5. Interfaz gráfica de la rutina **MARCADORES GCMP-BILOT**.

La información de entrada para esta aplicación 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 únicamente por la parte numérica. Puede utilizarse el código numérico –99 para identificar las celdas con información faltante o bien pueden dejarse vacías. La matriz de datos se leerá desde un archivo Excel.

Dado que esta rutina se ha elaborado como parte integral del proceso de análisis propuesto en este trabajo, la información de entrada se buscará 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 puede eliminar la marca de verificación de la casilla *Los datos están en la hoja "MC"*, con lo cual aparecerá una ventana adicional que permitirá 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 cálculo en el mismo archivo Excel que contiene la información de entrada, así: 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 representación GCMP-Biplot, en las salidas se incluyen también los valores propios de R , los cuales son utilizados en el cálculo de las calidades de representación.

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

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

6.7.3 RUTINA REPRESENTACIONES GCMP-BIPLLOT.

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-BIPLLOT** en MATLAB. Su instalación se realiza

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



Figura 6.6. Interfaz gráfica de la rutina **REPRESENTACIONES GCMP-BILOT**.

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

Los datos de entrada del programa están conformados por las tres hojas generadas por el módulo **MARCADORES GCMP-BILOT**, 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 información 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-BILOT**.

Aunque el programa también necesita leer la información de las etiquetas de las variables y de las filas, en caso de que esta última exista, el módulo ha sido diseñado para darle continuidad al proceso, evitando la intervención del usuario siempre que ello sea posible. En tal sentido, el módulo **REPRESENTACIONES GCMP-BILOT** lee las etiquetas de la primera hoja del archivo (la ubicada más hacia la izquierda), es decir, la que se utilizó como información de entrada del módulo **CUANTIFICA**.

Luego, aunque no se requiere ninguna modificación por parte del usuario, es necesario que la hoja con los datos utilizados en la cuantificación esté presente y que ocupe la primera posición. En caso de que el módulo **CUANTIFICA** haya generado una matriz de datos reducida (cf. § 5.3.2), esta deberá moverse a la primera posición.

Asimismo, el programa utiliza la información contenida en la matriz de datos para la conformación de grupos de individuos por niveles de una variable, cuando se activa la opción “*Pintar por niveles de variable...*” que aparece en la ventana de opciones de la representación (Figura 6.7).

No está de más insistir en el hecho de que los procesos han sido automatizados de tal forma que se logre minimizar la intervención del usuario. Luego, en caso de que el módulo **CUANTIFICA** no haya creado una matriz de datos reducida, no hace falta realizar ninguna modificación 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.

Así, pues, el archivo de entrada que será procesado por la rutina **REPRESENTACIONES GCMP-BILOT** deberá contar con 4 hojas obligatorias: 1) Datos utilizados para la Cuantificación, con los nombres de las variables en la primera fila. Esta información 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 podrá tener

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

Una vez leído el archivo Excel que contiene la información básica para construir la representación GCMP-Biplot, aparece una ventana como la de la Figura 6.7, en la que pueden precisarse las opciones particulares de la representación.



Figura 6.7. Interfaz gráfica para las opciones de la **REPRESENTACIÓN GCMP-BIPLLOT**.

DIMENSIONES

Dentro de las posibles dimensiones de representación (cualquier dimensión menor o igual que el rango de la matriz completa), se elige la de la abscisa y la de la ordenada.

PINTAR POR NIVELES DE UNA VARIABLE

Esta opción 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 información representada, lo que optimiza la visualización conjunta de los marcadores fila y los marcadores columna. Este artilugio no altera las propiedades de la representación GCMP-Biplot.

GRAFICAR ETIQUETAS DE FILAS

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

GRAFICAR ETIQUETAS DE COLUMNAS

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

CALCULAR CONTRIBUCIONES

Se calculan los siguientes indicadores:

- 1) Calidad de Representación Global.
- 2) Contribuciones Relativas del Elemento Fila al Factor.
- 3) Contribuciones Relativas del Elemento Columna al Factor.
- 4) Contribución Relativa del Factor al Elemento Fila (Calidad de Representación de los Individuos).
- 5) Contribución Relativa del Factor al Elemento Columna (Calidad de Representación 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 módulo 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.

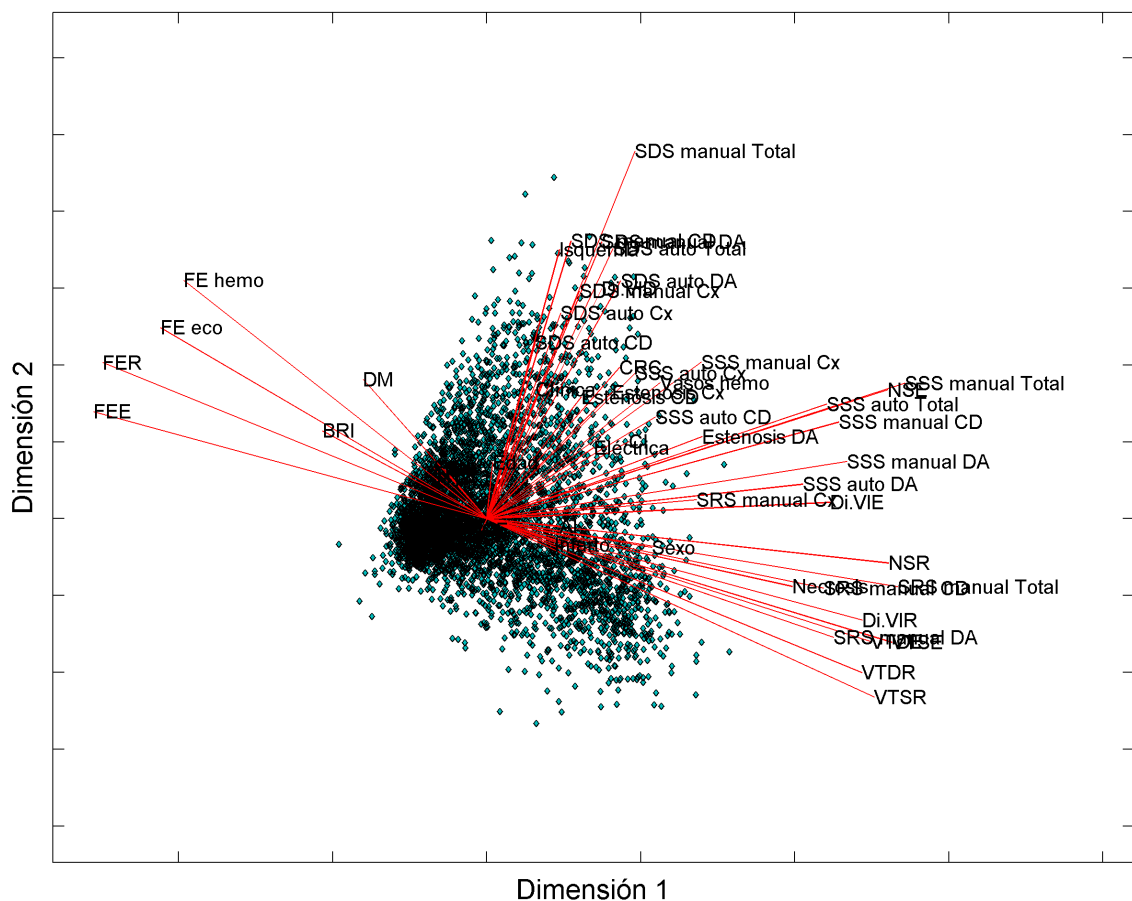


Figura 6.8. Representación GCMP-Biplot del primer plano factorial.

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

cuestión contiene 6.965 puntos y 53 líneas. 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 representación de 35,21 %.

Para el tipo de marcadores vectoriales elegidos (cf. § 6.7), la longitud al cuadrado de cada línea 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 línea puede interpretarse como calidad de representación de la variable representada en el correspondiente plano. Por tanto, las variables cuyas correspondientes líneas 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 visualización de las variables que gozan de una aceptable calidad de representación.

El coseno del ángulo formado por dos líneas proporciona una estimación de la correlación entre las dos variables representadas. Luego, un ángulo muy pequeño entre dos variables se interpreta como una alta correlación positiva entre tales variables. Análogamente, los ángulos obtusos se interpretan como correlación negativa, siendo esta de mayor magnitud, cuanto mayor sea el ángulo. Un ángulo llano o de 180 grados estima una correlación de -1 .

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

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

También es posible visualizar lo bien representada que está una variable en un plano, pintando de diferentes colores los puntos correspondientes a cada una de sus diferentes categorías. Si la variable no se encuentra bien representada, la separación entre los individuos que conforman las diferentes categorías no será muy clara. Por el contrario, una clara separación entre categorías indica una buena calidad de representación de la variable en dicho plano.

La Figura 6.9 es una versión modificada del Biplot de la Figura 6.8, en la que se han ocultado las variables con baja calidad de representación, se han eliminado los nombres de las variables y se han pintado los puntos que representan a los pacientes, acorde con la variable **Necrosis**: verde para ausencia de Necrosis; rojo para presencia de Necrosis.

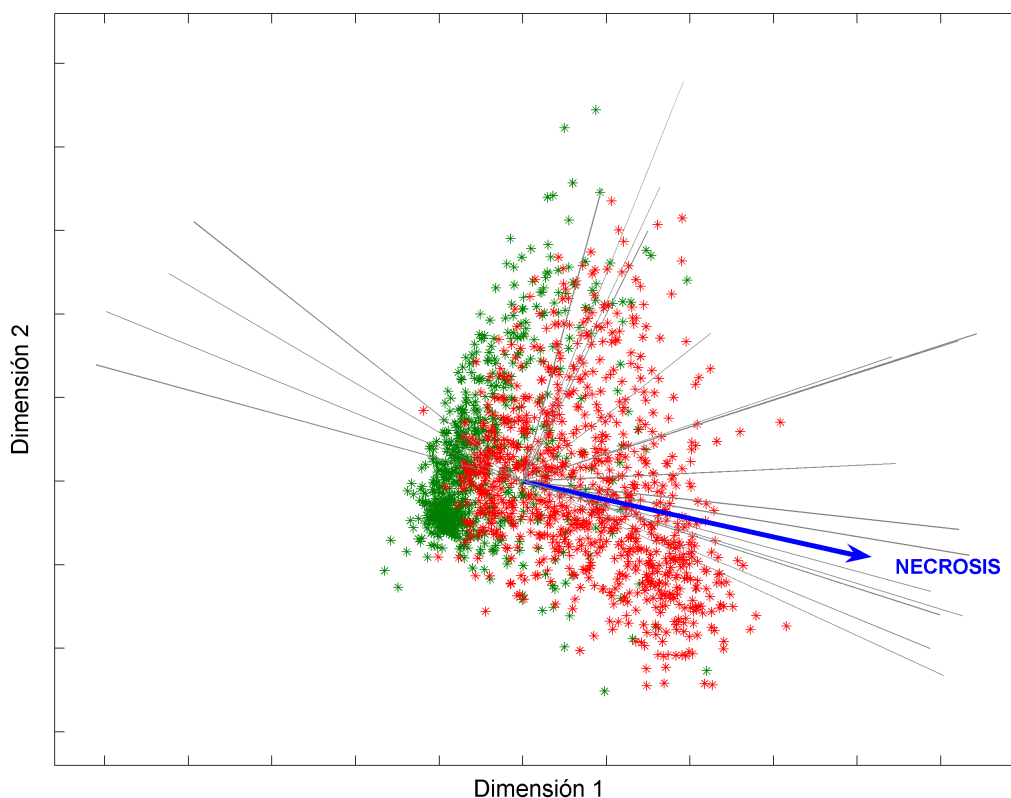


Figura 6.9. GCMP-Biplot del primer plano factorial con pacientes discriminados por Necrosis.

La marcación 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 número 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 conocía su lectura de la variable **Necrosis**.

En la Figura 6.10 se realiza una separación análoga 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.

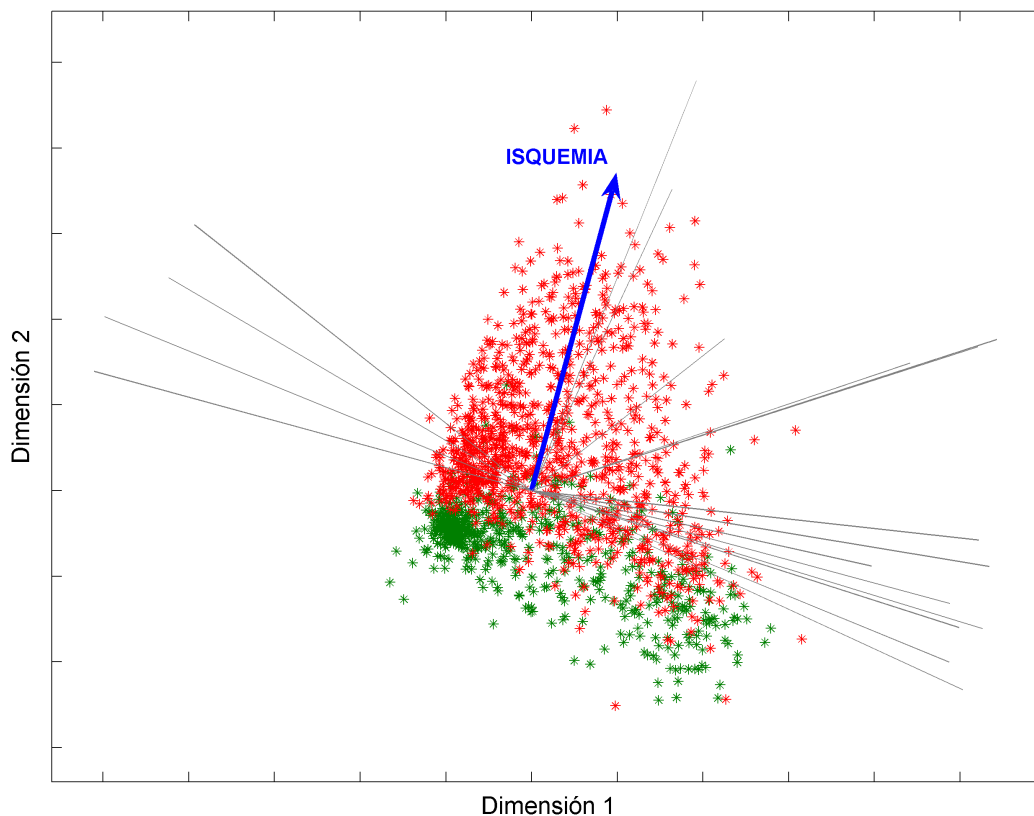


Figura 6.10. GCMP-Biplot del primer plano factorial con pacientes discriminados por Isquemia.

Con base en el análisis 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 afirmación más clínica que estadística,

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 conformación de los ejes, todas las demás variables con alta calidad de representación en el plano 1-2 son índices que de una u otra forma reflejan aspectos puntuales de una o ambas de las dos condiciones fundamentales que caracterizan la cardiopatía isquémica: **Necrosis** e **Isquemia**. Puede decirse en cierto modo que en la población de pacientes que conforma la base de datos analizada, donde la mayoría de variables son resultados de SPECT, la **Necrosis** y la **Isquemia** son los factores básicos, mientras que las otras variables son la expresión de tales factores.

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

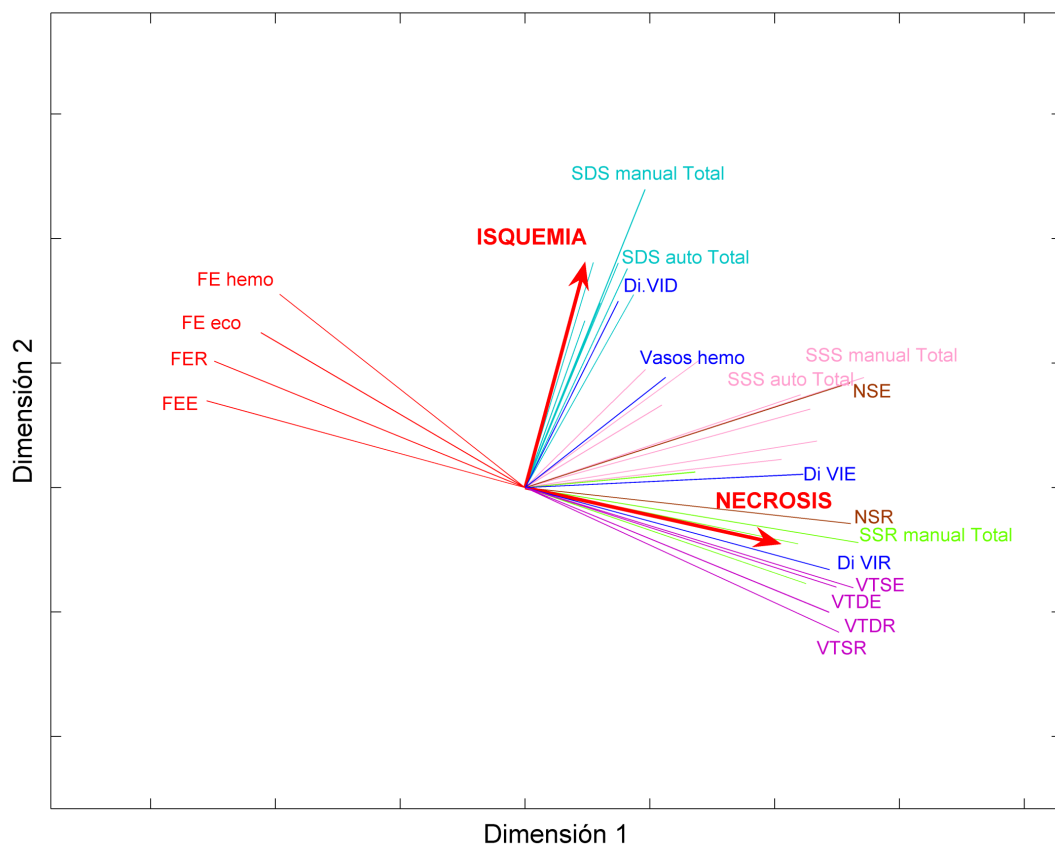


Figura 6.11. H-plot de las principales relaciones entre las variables en el plano 1-2.

En la Figura 6.11 se han utilizado colores para marcar grupos de variables, lo cual facilita su diferenciación. Así, se han resaltado con rojo las dos variables alrededor de las cuales gira la discusión, esto es, **Necrosis** e **Isquemia**, y además se ha agregado una cabeza de flecha en la dirección que apunta hacia sus correspondientes presencias. Se ha utilizado también el color rojo para marcar el grupo de vectores que representan las fracciones de eyección, medidas en diferentes condiciones y por diferentes métodos. Se han pintado con violeta los volúmenes telediastólico y telesistólico del ventrículo izquierdo. Se ha usado el azul encendido para señalar las diferentes medidas de dilatación del ventrículo izquierdo, así como el número 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 estrés, y de cian los scores diferencia. Se han señalado con marrón el número de segmentos afectados, tanto en estrés como en reposo.

En primer lugar, se destaca la conformación 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** están las fracciones de eyección, los volúmenes del ventrículo izquierdo así como su dilatación en estrés y en reposo, el número de segmentos afectados en reposo y los scores en reposo.

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

Aportando información sobre ambos factores están los scores en estrés, el número de segmentos afectados en estrés, así como el número de vasos afectados.

Se aprecia una estrecha relación entre **Necrosis** y los scores en reposo (**SRS**, pintados de verde). Con el objetivo de mantener el gráfico despejado, únicamente hemos etiquetado el score manual total de defectos en reposo (**SRS manual Total**), el cual como puede observarse está cerca de la bisectriz del ángulo formado por los dos marcadores más 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 habíamos indicado anteriormente, la **Necrosis** es una condición permanente, la cual no se asocia con un desequilibrio puntual entre la demanda y el aporte del oxígeno en el momento de la medición. Este es el motivo de que los scores en reposo (**SRS**) sean los más apropiados para evidenciar esta situación. Lo mismo puede decirse del número 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 dilatación del ventrículo izquierdo es una manifestación que se asocia frecuentemente con **Necrosis**. En particular, es llamativa la fuerte relación directa entre **Necrosis** y dilatación del ventrículo izquierdo en reposo (**Di.VIR**), así como la relación directa de los volúmenes telediastólicos y telesistólicos del ventrículo izquierdo (**VTDE**, **VTDR**, **VTSE**, **VTSR**) con las anteriores variables.

La dilatación del ventrículo izquierdo trae como consecuencia una disminución en la capacidad de bombeo del corazón. Esta capacidad de bombeo es medida por la fracción de eyección. Esto explica la fuerte correlación negativa observada entre las diferentes medidas de la fracción de eyección y los parámetros de dilatación del ventrículo izquierdo (**Di.VIR**, **VTSE**, **VTDE**, **VTDR** y **VTSR**).

Más allá de esta relación general, vale la pena resaltar que la fracción de eyección en reposo (**FER**) está particularmente asociada con los volúmenes telediastólico y telesistólico del ventrículo izquierdo en reposo (**VTDR** y **VTSR**), exhibiendo una correlación negativa perfecta con la primera de estas variables. Asimismo, la fracción de eyección en estrés (**FEE**) está particularmente asociada con los volúmenes telesistólico y telediastólico del ventrículo izquierdo en estrés (**VTSE** y **VTDE**).

En cuanto a la **Isquemia**, se destaca su estrecha relación con los scores diferencia, tanto manuales como automáticos (**SDS manual**, **SDS auto**), así como con la dilatación transitoria del ventrículo izquierdo, obtenida por diferencia (**Di.VID**).

Tal y como anotamos anteriormente, en estado de reposo se reflejan básicamente las afecciones permanentes producto de la **Necrosis**. En los resultados obtenidos por diferencia entre el estrés y el reposo (**SDS** y **Di.VID**) se está sustrayendo el efecto de la

Necrosis, aislando, por tanto, el efecto de la **Isquemia**. Esto explica la estrecha relación observada entre la **Isquemia**, los scores diferencia (SDS) y la dilatación del ventrículo izquierdo por diferencia (**Di.VID**).

Como es de esperarse, los scores totales, tanto manuales como automáticos (**SDS manual Total**, **SDS auto Total**), también se localizan en este caso cerca de la bisectriz de las dos líneas con mayor ángulo interno, que representan los scores diferencia en territorios específicos de cada una de las arterias coronarias.

Para analizar las variables que exhiben aproximadamente la misma correlación con **Necrosis** que con **Isquemia**, debe considerarse que el desempeño cardiaco en condiciones de estrés se ve afectado tanto por lesiones permanentes del miocardio (**Necrosis**), como por condiciones puntuales propias del desequilibrio entre la demanda y el aporte de oxígeno (**Isquemia**). Esto explica el hecho de que tanto el número de segmentos afectados en estrés (**NSE**) como los scores en estrés (**SSS manual** y **SSS auto**) aparezcan como indicadores compromiso entre **Necrosis** e **Isquemia**.

La dilatación del ventrículo izquierdo en estrés (**Di.VIE**), a pesar de estar lógicamente asociada con la dilatación permanente del ventrículo izquierdo (**Di.VIR**), incluye una componente adicional que es el estrés cardiaco, lo que la aproxima a las variables medidas en condiciones de estrés.

Puesto que el número de vasos afectados, medido por hemodinamia (**Vasos hemo**) solo contabiliza el número de vasos con algún tipo de estenosis o estrechamiento, sin calificar la severidad del mismo, aparece también 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 más probable es que no haya **Necrosis** ni **Isquemia**.

En la Figura 6.12 se muestra el plano 3-4 de la representación 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 correlación positiva entre todos los índices o resultados correspondientes al territorio de la arteria descendente anterior. Se observa también cómo dentro de cada territorio, las pruebas

manuales tienden a estar correlacionadas entre sí al igual que las pruebas automáticas lo están entre ellas.

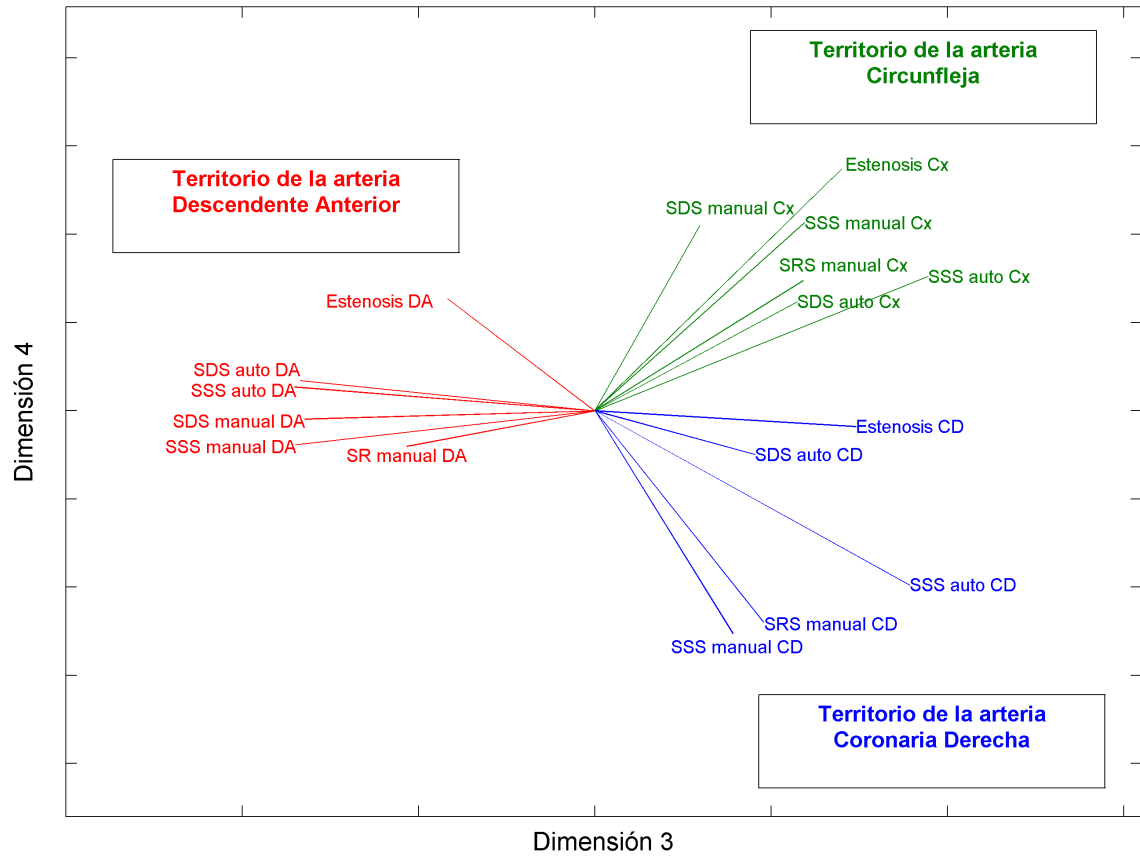


Figura 6.12. H-plot de las principales relaciones entre variables en el plano 3-4.

6.9 SÍNTESIS DE LA PROPUESTA DE ANÁLISIS

Los resultados ilustrados mediante la aplicación práctica de este capítulo, donde se han logrado separar los principales patrones de asociación de un sistema multivariante mixto de gran tamaño, con más de la mitad de datos faltantes, nos permiten cerrar este trabajo expresando nuestra satisfacción por el desempeño general de todas las técnicas aquí desarrolladas.

Es importante resaltar que el hecho de que las cuantificaciones generadas mediante la aplicación **CUANTIFICA** sean únicas y no estén basadas en una solución multidimensional —como sí lo están las generadas por las técnicas del sistema Gifi— no implica una reducción de la dimensionalidad del sistema cuantificado. La Matriz de Cuantificaciones no solo tiene propiedades métricas, sino que conserva todas las relaciones multidimensionales existentes en la base de datos original, las cuales pueden recuperarse y resumirse mediante alguna técnica lineal de reducción de la dimensionalidad, tal y como hemos ilustrado en este capítulo, mediante la obtención de representaciones GCMP-Biplot de la Matriz de Cuantificaciones.

Los módulos desarrollados como parte de este trabajo conforman una herramienta integral para el análisis de bases de datos multivariantes con información faltante. Es posible realizar un análisis análogo al presentado en este capítulo, con base en los siguientes pasos:

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

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

diferentes módulos (Figura 6.13). Para ingresar a la misma, se digita “*Modulos*” (Atención: “Modulos” sin tilde).

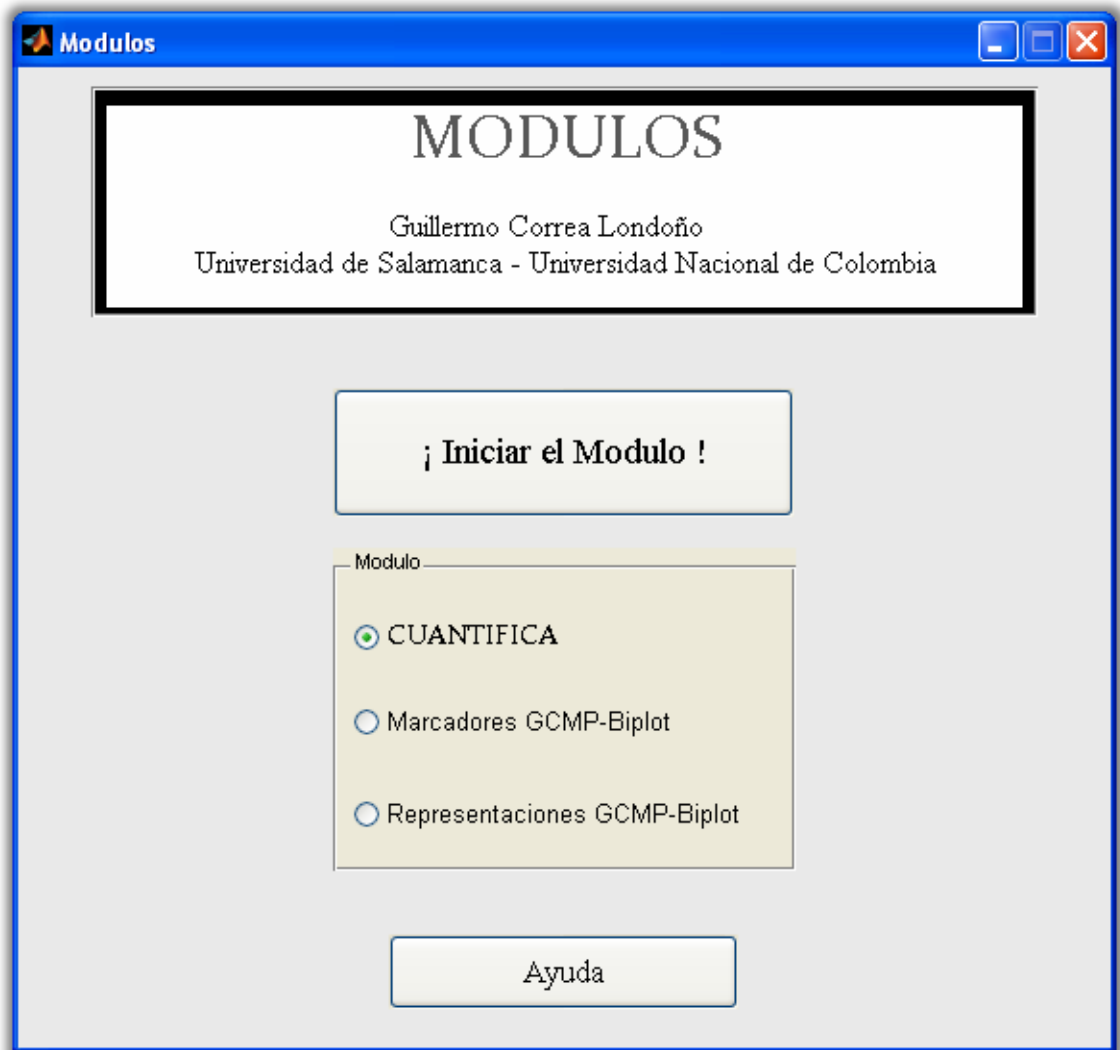


Figura 6.13. Interfaz de integración de módulos.

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

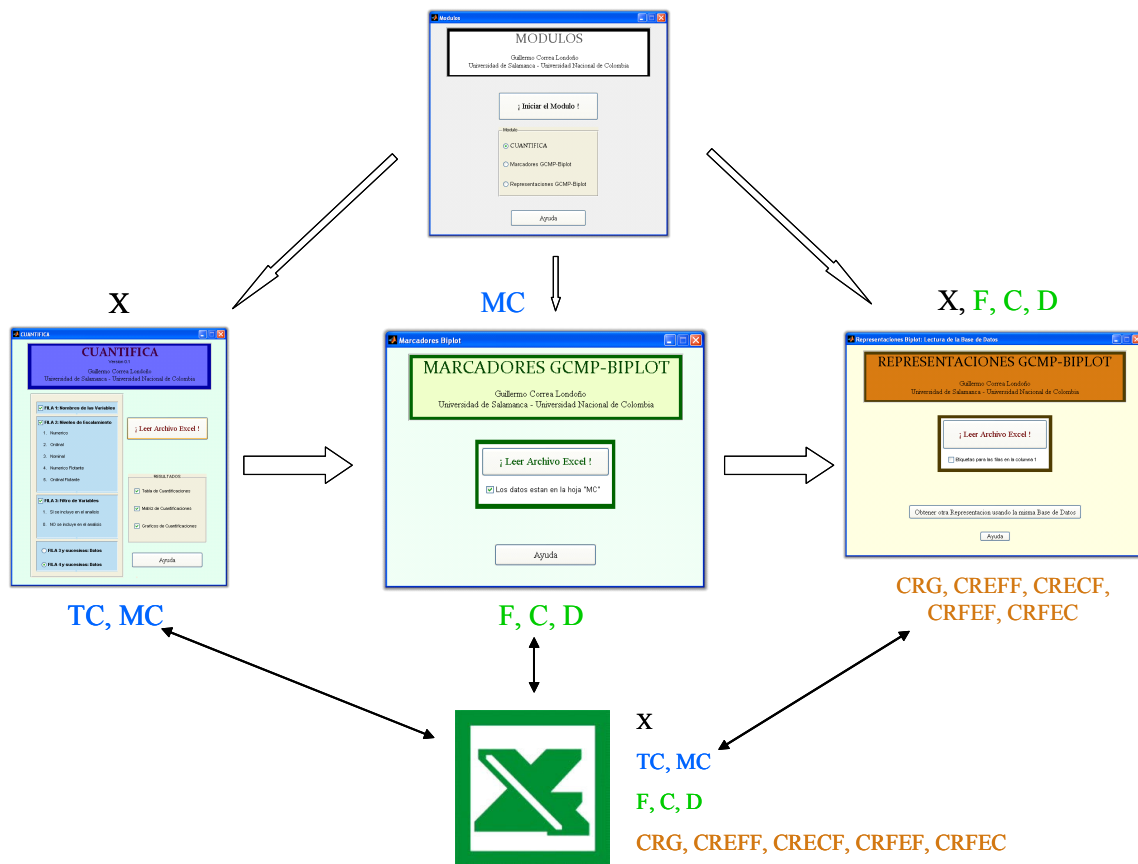


Figura 6.14. Resumen de los módulos de procesamiento de datos multivariantes mixtos.

En resumen, los pasos para el análisis de una base de datos multivariante mixta con información faltante, los módulos computacionales utilizados en cada uno de ellos, así como la información de entrada y salida, son:

- 1) Adecuar la base de datos para su cuantificación. (§ 5.3.1; § 6.6, Figura 6.4). En el esquema de la Figura 6.14, a los datos de entrada, usados para la cuantificación, los hemos denominado X.
- 2) Realizar la cuantificación óptima de la base de datos (X), mediante el módulo CUANTIFICA. Como resultados, se generan la Tabla de Cuantificaciones (TC) y la Matriz de Cuantificaciones (MC).

- 3) Usar la Matriz de Cuantificaciones (MC) como información de entrada del módulo **MARCADORES GCMP-BILOT** 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 información de entrada del módulo **REPRESENTACIONES GCMP-BILOT**, el cual, además de las representaciones, genera los índices de contribuciones y calidad de representación (CRG, CREFF, CRECF, CRFEF, CRFEC).

Los nombres de las salidas de un módulo 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 cuáles hojas se encuentra la información. El único caso en el que se requiere intervención 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* deberá moverse a la primera posición del archivo antes de ejecutar el módulo **REPRESENTACIONES GCMP-BILOT**.

Desde el punto de vista de la continuidad del proceso, habría sido posible omitir el módulo **MARCADORES GCMP-BILOT**, incorporándolo en el módulo **REPRESENTACIONES GCMP-BILOT**, haciendo que el cálculo de F, C y D, fuera un proceso interno de este último módulo. No obstante, hemos detectado que al trabajar con grandes bases de datos como la utilizada en este capítulo, el cálculo 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 más práctico calcular los marcadores GCMP-Biplot en una sola ocasión, dejándolos disponibles en el archivo de datos para la posterior obtención de representaciones GCMP-Biplot.

Finalmente, es importante resaltar que, si bien una de las fortalezas del sistema propuesto es el manejo de matrices con información faltante, todos los procesos funcionan igualmente bien cuando se cuenta con información completa. Al aplicar todo

el proceso descrito en este apartado sobre una base de datos completa, se obtiene una representación CMP-Biplot (cf. § 6.7.1).

Asimismo, cuando se parte de bases de datos completas, la Matriz de Cuantificaciones (MC) generada mediante la aplicación CUANTIFICA será, a su vez, completa, además de numérica. Por tanto, es posible utilizarla para obtener cualquier otra representación Biplot o como información de entrada de algún otro procedimiento lineal, acorde con los objetivos.

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

Antes de presentar las conclusiones de este trabajo, resulta obligatoria la comparación entre las técnicas de análisis aquí propuestas y las técnicas de análisis enmarcadas en el sistema Gifi de análisis multivariante no lineal. En § 5.4 hemos realizado una comparación general entre **CUANTIFICA** y los tres principales módulos de escalamiento óptimo del sistema Gifi de análisis multivariante no lineal, i. e., PRINCALS, HOMALS y OVERALS, cuando son usados como primer paso para la generación de cuantificaciones.

En este apartado, realizaremos una comparación particular entre las técnicas usadas para el análisis de la aplicación práctica de este capítulo (cuantificación óptima, usando **CUANTIFICA** + representaciones GCMP-Biplot) y las técnicas del sistema Gifi cuando son usadas para generar representaciones gráficas en subespacios de baja dimensionalidad. Para tal efecto, hemos elegido PRINCALS, pues tal y como se indica en § 5.4, es la técnica más similar a **CUANTIFICA**, difiriendo de ella solamente en los aspectos que le confieren superioridad a **CUANTIFICA** frente a las técnicas del sistema Gifi.

Puesto que en el análisis realizado en este capítulo, los principales patrones de asociación se manifiestan en el plano 1-2 de la representación GCMP-Biplot, hemos elegido inicialmente la solución PRINCALS con $p=2$ para fines de tal comparación (Figura 6.15).

Antes de analizar los resultados gráficos de PRINCALS, es pertinente aclarar que aunque la representación 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 teoría de GABRIEL (1971) (cf. § 6.7.1), es decir que no se trata de la representación gráfica conjunta de los marcadores fila y los marcadores columna de una matriz, obtenidos a partir de la descomposición factorial de su aproximación a bajo rango.

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

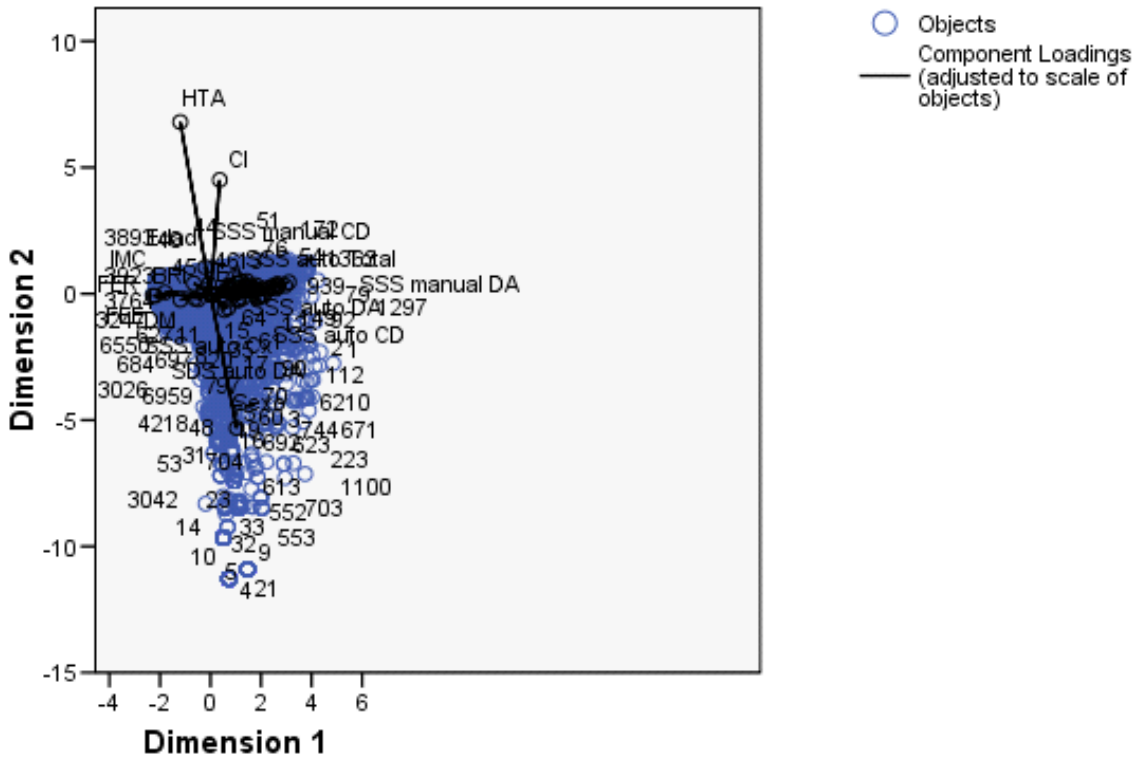


Figura 6.15. Plano biplot (puntuaciones + saturaciones) correspondiente a la solución PRINCALS, con $p=2$.

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

La Figura 6.16 permite contrastar el biplot (puntuaciones + saturaciones) resultante de la solución PRINCALS con $p=2$ y el GCMP-Biplot obtenido mediante las técnicas 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 análisis PRINCALS con $p=2$ no coincide con la generada mediante nuestra propuesta. Esto puede verse con mayor claridad en la Figura 6.17, versión modificada de la Figura 6.16, en la cual hemos omitido los puntos que representan a los pacientes y hemos resaltando las relaciones más sobresalientes entre variables.

La Figura 6.17 (a) permite apreciar que en la solución PRINCALS, casi todas las variables tienen su principal carga sobre la primera dimensión, mientras que las variables hipertensión arterial sistémica (**HTA**), cardiopatía isquémica en el diagnóstico inicial (**CI**) y **Sexo** son las únicas con altas cargas en la segunda dimensión.

Excepto por la correlación negativa entre las fracciones de eyección (**FEE**, **FER**, **FE hemo** y **FE eco**) y las demás variables, no se observa ninguno de los demás patrones detectados y analizados al usar las técnicas propuestas en este trabajo. En particular, no se observa la diferenciación entre **Necrosis** e **Isquemia**, ni sus relaciones con los correspondientes resultados del SPECT, principal patrón de la Figura 6.17 (b) y alrededor del cual giró nuestro análisis (cf. § 6.8).

Los principales patrones de asociación exhibidos en la Figura 6.17 (a) pueden resumirse de la siguiente forma. La primera dimensión está dominada por la correlación negativa entre las medidas de fracción de eyección (**FEE**, **FER**, **FE hemo** y **FE eco**) y todos los demás resultados de SPECT, ecocardiografía y coronariografía que de una u otra forma miden la gravedad de la cardiopatía isquémica. La segunda dimensión está dominada por las relaciones entre **Sexo**, hipertensión arterial sistémica (**HTA**) y cardiopatía isquémica en el diagnóstico inicial (**CI**), indicando que la prevalencia de hipertensión arterial es mayor en mujeres que en hombres y que la cardiopatía isquémica en el diagnóstico inicial es más frecuente en pacientes con hipertensión arterial sistémica que en pacientes con tensión arterial normal.

Estos resultados, si bien son correctos, son relativamente evidentes y fáciles de predecir por un cardiólogo sin que se requiera para ello un análisis multivariante. A partir de esta

solución 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 capítulo (cf. § 6.8).

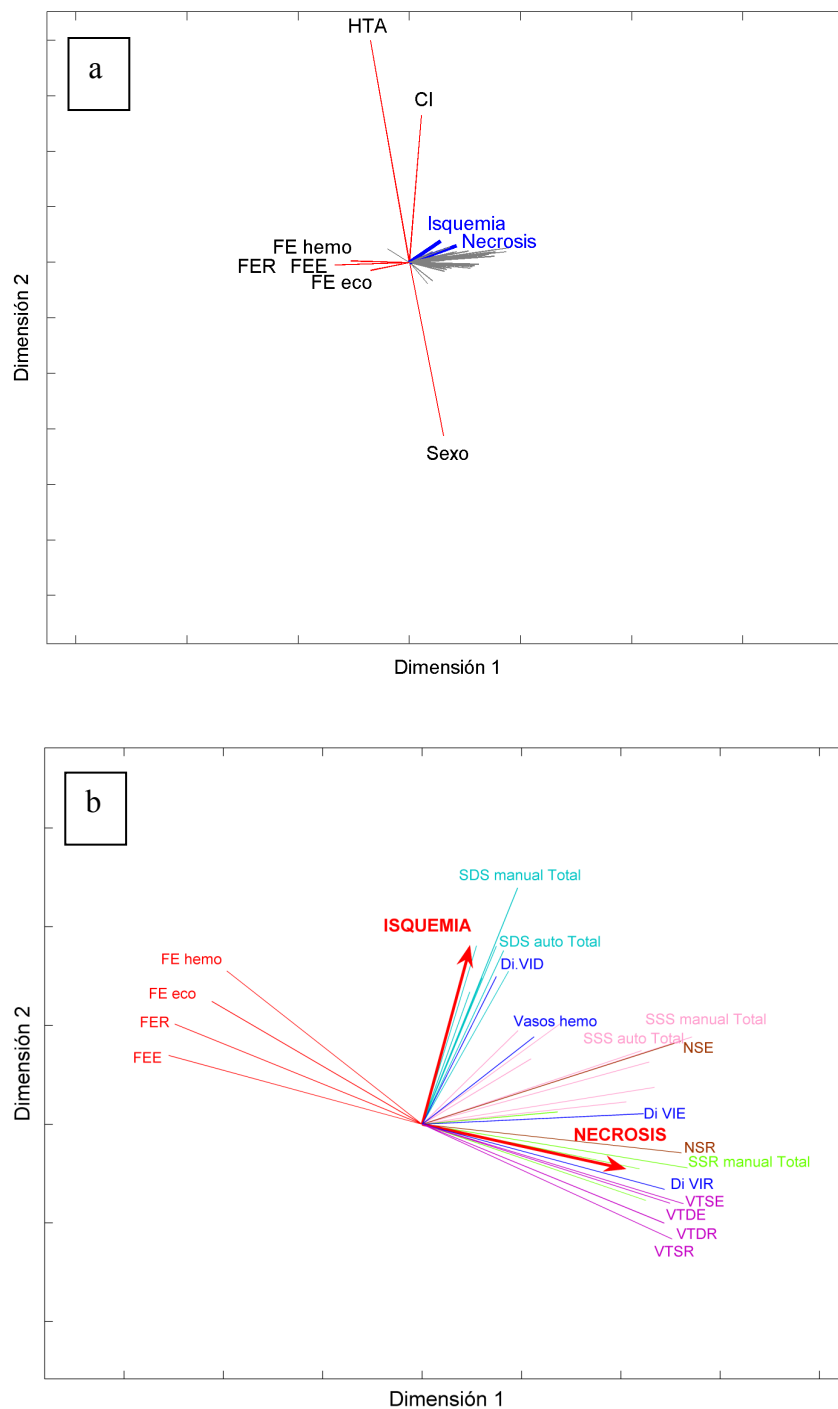


Figura 6.17. (a) Saturaciones correspondiente a la solución PRINCALS, con $p=2$ y (b) H-plot de la matriz de cuantificaciones generada mediante **CUANTIFICA**.

Tal y como hemos anotado a lo largo de toda la memoria, las soluciones del sistema Gifi cambian en función de la dimensionalidad elegida (cf. Capítulo 3). Resulta, por tanto, lógico evaluar otras soluciones PRINCALS de mayor dimensionalidad. Para el efecto, obtuvimos adicionalmente las soluciones PRINCALS desde $p=3$ hasta $p=10$. Se observó 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 más allá del 1-2 en búsqueda de los patrones que encontramos en nuestro análisis, pero estos no aparecieron.

La Figura 6.18 resulta suficiente para ilustrar esta búsqueda. Allí se presenta el plano 3-4 del biplot (puntuaciones + saturaciones) correspondiente a la solución 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 correlación positiva entre la fracción de eyección (**FE hemo**), el número de vasos afectados (**Vasos hemo**) y la estenosis de las arterias coronarias, que aparece reflejada en la parte inferior derecha de la representación.

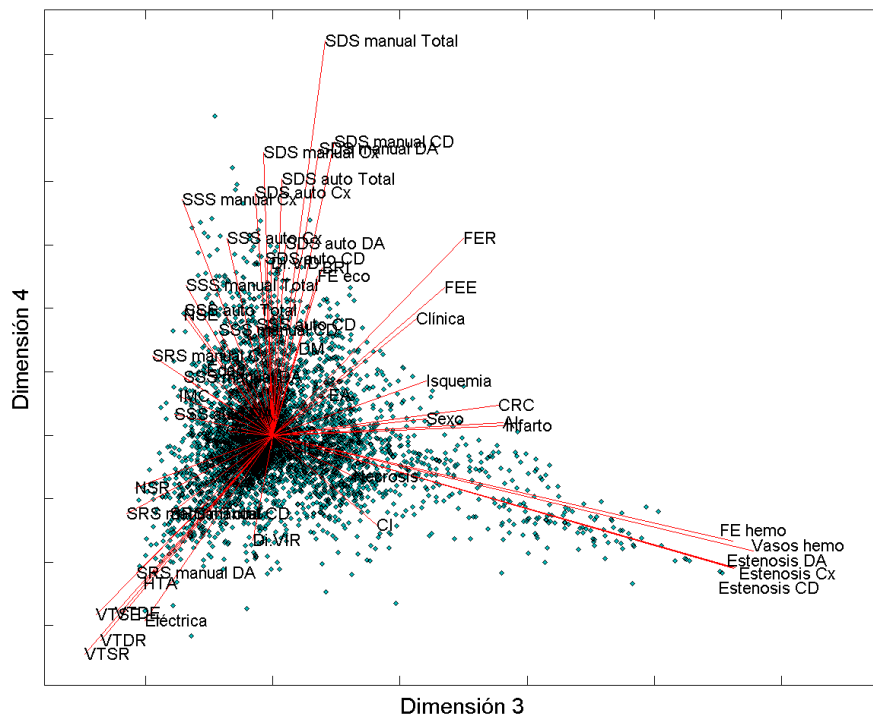


Figura 6.18. Plano 3-4 del biplot (puntuaciones + saturaciones) correspondiente a la solución PRINCALS, con $p=10$.

Finalmente, aunque se sabe que, por construcción, las cuantificaciones generadas mediante **CUANTIFICA** minimizan las *Interdistancias*, esto es, el criterio escalado de la suma sobre todas las observaciones de las distancias cuadráticas entre las categorías cuantificadas de las diferentes variables (cf. expresión [4.7] y § 5.2), cerraremos esta comparación mostrando que tal condición se satisface para este ejemplo específico. En la Figura 6.19 se muestran las *Interdistancias* calculadas para la solución **CUANTIFICA**, así como las correspondientes a las soluciones PRINCALS desde $p=1$ hasta $p=10$. Como era de esperarse, las *Interdistancias* para la solución **CUANTIFICA** son menores que cualquiera de las otras *Interdistancias*, lo que significa que la cuantificación de **CUANTIFICA** es la que asigna valores más similares a las categorías de las distintas variables que se asocian con mayor frecuencia.

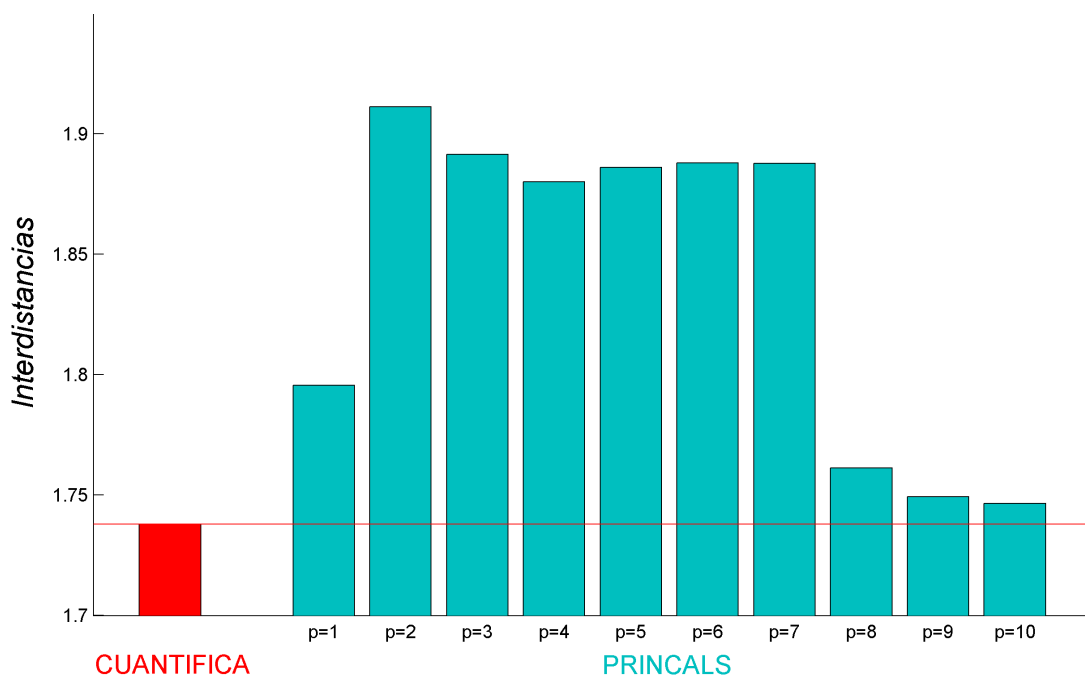


Figura 6.19. *Interdistancias* para **CUANTIFICA** y para soluciones PRINCALS desde $p=1$ hasta $p=10$.

En vista de los resultados de las anteriores comparaciones, y teniendo en cuenta la sustentación teórica presentada, así como la existencia de un soporte informático para nuestra propuesta, nos permitimos presentarla como alternativa a las técnicas del sistema Gifi para el análisis de sistemas multivariantes mixtos.

CONCLUSIONES

1. En este trabajo, hemos desarrollado un criterio de cuantificación óptima, el cual se basa en la asignación de valores más similares a las categorías de las distintas variables que se asocien con mayor frecuencia. Tales cuantificaciones óptimas se obtienen al hacer mínima la suma sobre todas las observaciones de las distancias cuadráticas entre las categorías cuantificadas de las diferentes variables.
2. Dado que las cuantificaciones generadas mediante el criterio de escalamiento óptimo propio de las técnicas del Sistema Gifi no son únicas, sino que, por el contrario, dependen de la dimensionalidad de la solución elegida, debe utilizarse algún criterio que permita elegir el mejor conjunto de cuantificaciones. Para tal efecto, hemos utilizado el criterio de cuantificación óptima desarrollado en este trabajo. La rutina computacional **INTERDIS** automatiza dicha elección.
3. Se ha elaborado una rutina computacional llamada **CUANTIFICA**, la cual implementa el criterio de cuantificación óptima desarrollado en este trabajo y automatiza su obtención. Mediante esta aplicación es posible asignar cuantificaciones óptimas a un sistema multivariante, tomando en consideración las restricciones propias del nivel de escalamiento de cada variable y sin que se requieran las cuantificaciones de ningún otro sistema como punto de partida. En adición a los niveles de escalamiento Numérico, Ordinal y Nominal, tradicionalmente usados en las técnicas del Sistema Gifi, en este trabajo incorporamos los niveles de escalamiento Numérico Flotante y Ordinal Flotante.
4. El hecho de que las cuantificaciones generadas mediante la aplicación **CUANTIFICA** sean únicas y no estén basadas en una solución multidimensional —como sí lo están las generadas por las técnicas del sistema Gifi— no implica una reducción de la dimensionalidad del sistema cuantificado. La matriz de cuantificaciones no solo tiene propiedades métricas, sino que conserva todas las relaciones multidimensionales existentes en el sistema original, las cuales pueden recuperarse y resumirse mediante alguna técnica lineal de reducción de dimensionalidad.

5. Se ha desarrollado una propuesta para generar la representación 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 automatización de dicho proceso.
6. La técnica de cuantificación desarrollada en este trabajo, implementada en la rutina computacional **CUANTIFICA**, en combinación con la propuesta de representaciones GCMP-Biplot, cuyos desarrollos teórico y algorítmico también constituyen aportaciones de este trabajo, conforman una herramienta integral que resulta efectiva para analizar grandes sistemas multivariantes mixtos con información faltante y, desde luego, también para sistemas multivariantes completos. En particular, como aplicación práctica, en este trabajo se analizó un sistema multivariante conformado por 6.965 pacientes con cardiopatía isquémica o síntomas de la misma y 53 variables, con 53,12 % de información faltante. Las técnicas propuestas permitieron detectar los principales patrones de asociación.
7. El primer plano principal de la representación GCMP-Biplot, para la matriz de cuantificaciones incompleta generada mediante la aplicación **CUANTIFICA**, permitió reconocer que las principales asociaciones en la población estudiada giran en torno a las variables Necrosis e Isquemia, como factores básicos de la cardiopatía isquémica. Se pudo establecer que la Necrosis, como defecto permanente, se asocia particularmente con las pruebas realizadas en reposo, así como con las diversas mediciones de dilatación, volumen y fracción de eyección del ventrículo izquierdo. La Isquemia, como defecto transitorio, se asocia especialmente con los *scores* diferencia. Los resultados de las pruebas realizadas bajo condiciones de estrés cardíaco, constituyen un índice general de los defectos de Necrosis e Isquemia.

-
8. Los patrones develados mediante el uso de las técnicas propuestas en este trabajo no fueron detectados al realizar un análisis análogo mediante las técnicas del sistema Gifi de análisis multivariante no lineal, en particular, al evaluar soluciones PRINCALS con dimensionalidades desde $p=2$ hasta $p=10$.

BIBLIOGRAFÍA

- BARTLETT, M. S. (1933). On the theory of statistical regression. *Proc. Roy. Soc. Edinburgh*, 53:260-283.
- BENZÉCRI, J. P. (1977). Sur l'analyse des tableaux binaires associés à une correspondance multiple. *Les cahiers de l'analyse des données*, 2:55-71.
- BENZÉCRI, J. P. et ál. (1973). L'analyse des données : L'analyse des correspondances. Paris: Dunod.
- BLASIUS, J. and GREENACRE, M. (2006). Correspondence analysis and related methods in practice. **In:** Michael Greenacre and Jörg 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 N° 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 Appliquée*. 38(3):35-51.

- CORREA, G. (2006). Caracterización multivariante no lineal de la cardiopatía isquémica. Memoria Interna, Departamento de Estadística, Universidad de Salamanca, España.
- CUADRAS, C. M. (1991). Métodos de análisis 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 ál.
- 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 Jörg 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 DOMÍNGUEZ, M.; RUANO, R.; RAMÍREZ, V. H.; SANTOS RODRÍGUEZ, I.; MARTÍN de ARRIBA, Á.; GARCÍA TALAVERA, J. R.; MARTÍN 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 Cardiología, Estocolomo (Suecia).

- DIEGO DOMÍNGUEZ, M.; RAMÍREZ, V. H.; RUANO, R.; SANTOS RODRÍGUEZ, I.; MARTÍN de ARRIBA, Á.; Del CAMPO, F; GARCÍA TALAVERA, P.; MARTÍN LUENGO, C. (2006). Valor predictivo del comportamiento de la fracción de eyección por Gated-SPECT para eventos y gravedad de la enfermedad coronaria. **En:** XII reunión de Cardiología Nuclear y Actualización del Diagnóstico por Imagen. VII Simposio Iberoamericano de Cardiología 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 representación simultánea de datos multidimensionales. Tesis doctoral. Universidad de Salamanca.
- GALINDO, M. P. (1986). Una alternativa de representación simultánea: HJ-biplot. *Questió*, 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-VILLARDÓN, J. L. (1998). Use of biplots to diagnose independence models in three-way contingency tables. **In:** Jörg Blasius and Michael Greenacre (Ed.), *Visualization of categorical data*. San Diego: Academic Press. p. 391-404.

- GANONG, W. F. (2005). Fisiología Médica. 20 ed. México, 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 Jörg 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 Fisiología Médica. 10ª ed. Madrid: McGraw Hill/Interamericana de España. 800 p.
- HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. (1999). Análisis Multivariante. 5 ed. Madrid: Prentice-Hall. 832 p.

- HAYASHI, C. (1950). On the quantification of qualitative data from the mathematico-statistical 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 symétrique 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 ál. p. 433-446.
- LECLERC, A. (1980). Quelques propriétés optimales en analyse de données en terme de corrélation entre variables. *Mathématique 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.
- MARTÍN-MOREIRAS, I.; CRUZ, M. D.; MARTÍN-HERRERO, F.; LEÓN, V.; SÁNCHEZ, M.; RAMÍREZ, V. H.; MARTÍN-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). Teoría Psicométrica. 3 ed. México: McGraw-Hill. 843 p.

- OLIVA MORENO, J.; LOBO ALEU, F.; LÓPEZ BASTIDA, J.; DUQUE GONZÁLEZ, B.; OSUNA GUERRERO, R. (2004). Costes no sanitarios ocasionados por las enfermedades isquémicas del corazón en España. *Cuadernos económicos de I.C.E.*, 67:263-298.
- OMS. (2004). Informe sobre la salud en el mundo 2004: cambiemos el rumbo de la historia. Ginebra: Organización 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.
- RAMÍREZ CASTRO, V. H.; DIEGO DOMÍNGUEZ, M.; RUIZ OLGADO, M; SÁNCHEZ FLORES, M; SANTOS RODRÍGUEZ, I; CASCÓN BUENO, M.; RODRÍGUEZ COLLADO, J. MARTÍN LUENGO, C. (2006). Valor y significado clínico de las imágenes de perfusión miocárdica en pacientes con angina de pecho y arterias coronarias sin lesiones angiográficas significativas. **En:** XI Congreso de la Sociedad Castellano-Leonesa de Cardiología (SOCALEC), Valladolid.
- ROY, S. N. (1957). Some aspects of multivariate analysis. New York: John Wiley & Sons. 214 p.
- RUANO, R.; DIEGO DOMÍNGUEZ, M.; MARTÍN de ARRIBA, Á; RAMÍREZ, V. H.; MARTÍN LUENGO, C.; GARCÍA-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 Asociación Europea de Medicina Nuclear. Estambul (Turquía), octubre de 2005.

- RUANO, R.; RAMÍREZ, V. H.; MARTÍN de ARRIBA, Á; DIEGO DOMÍNGUEZ, M.; TAMAYO, P.; LEÓN, V. GARCÍA-TALAVERA, J. (2005b). Seguimiento a largo plazo de pacientes con dolor torácico e isquemia demostrada mediante Gated-SPECT de perfusión miocárdica y coronarias angiográficamente normales. **En:** Congreso Español 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 données qualitatives. Thèse de Doctorat de 3ème 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. *American 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 Mathématique et Physique*. 10:113-121.

- VERHULST, P. F. (1845). Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux Mémoires de l'Académie Royale des Sciences, des lettres et des Beaux-Arts de Belgique*, 18:1-38.
- VERHULST, P. F. (1847). Deuxième mémoire sur la loi d'accroissement de la population. *Nouveaux Mémoires de l'Académie Royale des Sciences, des lettres et des Beaux-Arts de Belgique*, 20:1-32.
- VICENTE-VILLARDÓN, J. L.; GALINDO-VILLARDÓN, M. P. and BLÁZQUEZ-ZABALLOS, A. (2006). Logistic biplots. **In:** Michael Greenacre and Jörg 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.