UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

INTRODUCCION AL ANALISIS ESTADISTICO DE CONGLOMERADOS

T E S I S

QUE PARA OBTENER EL TITULO DE

A C T U A R I O

P R E S E N T A :

OSCAR PABLO HERRERA VILLALOBOS

FACULTAD DE CIENCIAS
UNAM

DIRECTOR DE TESIS: ACT. JAIME VAZQUEZ ALAMILLA

FACULTAD DE CIENCIAS
SECCION ESCOLAR

UNIVERSIDAD NACIONAL AVENOMA DE MEXICO

DRA. MARÍA DE LOURDES ESTEVA PERALTA
Jefa de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito:

Introducción al análisis estadístico de conglomerados

realizado por Oscar Pablo Herrera Villalobos

con número de cuenta 9613150-9 , quién cubrió los créditos de la carrera de Actuaria

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	Act. Jaime Vázquez Alamilla
Propietario	M. en C. José Antonio Flores Díaz
Propietario	M. en A. P. María del Pilar Alonso Reyes
Suplente	Dr. Luis Antonio Rincón Solís
Suplente	M. en C. María de Lourdes Guerrero Zarco

Consejo Departamental de Matemáticas

M. en C. José Antonio Flores Díaz

CONSEJO DEPARTAMENTAL DE MATEMÁTICAS

Introducción al análisis estadístico de conglomerados

Oscar Pablo Herrera Villalobos

Enero de 2004

Agradecimientos

A mis padres, hermanos
y amigos

Índice general

Prefacio

El análisis multivariado es un conjunto de técnicas que se concentran en el manejo de objetos que tienen más de una variable. Algunas de estas técnicas surgieron de la necesidad de una extensión del caso univariado, por ejemplo el Análisis Multivariado de la varianza (MANOVA) es una extensión del Análisis de la Varianza (ANOVA), y el resto de las técnicas no tienen analogías para el caso univariado, el análisis de conglomerados, los componentes principales y el escalamiento multidimensional son ejemplos de éstas. La gran aplicabilidad del análisis multivariado ha sido comprobada mediante un sinfin de textos en diversas disciplinas, no sólo en el área científica sino también en el área social, y le han dado un lugar muy importante en la estadística.

Cada técnica multivariada es distinta de las otras, lo que implica que se requiera hacer más de un análisis para resolver un problema. Así, el análisis de conglomerados no es la excepción. A lo largo del presente trabajo se hace uso de dos técnicas más, a saber, componentes principales y escalamiento multidimensional, aunque podrían ser más dependiendo del problema.

A diferencia de las demás técnicas multivariadas, una parte importante del análisis de conglomerados (métodos jerárquicos) puede ser entendida sin ser experto en matemáticas o estadística, pero esto no quiere decir que sea la técnica menos importante del análisis multivariado, sino al contrario, permite desarrollar un mayor número de aplicaciones en diversas áreas.

El objetivo principal de este trabajo es el dar un panorama general del análisis de conglomerados, pero diseñado para idear un sinfin de ejemplos que pueden ir desde una clasificación de seres vivientes en biología, hasta un estudio de preferencias de diversos consumidores en mercadotecnia.

El primer capítulo brinda una pequeña introducción de las características y las técnicas del ya mencionado análisis multivariado. Este capítulo es opcional por lo que esta marcado con un asterísco, "*", y su lectura no afecta el objetivo principal del presente trabajo.

El capítulo dos , Algebra de Matrices, reune el material suficiente para entender las técnicas del escalamiento multidimensional y el análisis de componentes principales presentadas en el capítulo tres.

En capítulo tres se presentan distintas formas para visualizar datos multivariados, ya sea por medio de alguna gráfica, o por medio de las dos técnicas multivariadas alternativas.

El cuarto capítulo profundiza en el tema del análisis de conglomerados, así como en los métodos más comunes en la paquetería estadística, los métodos jerárquicos.

El quinto capítulo proporciona la teoría de métodos de optimización y las mezclas de densidades (modelos probabilísticos).

El último capítulo brinda cuatro aplicaciones cuyo objetivo es mostrar que el análisis de conglomerados no solo es una herramienta para clasificar sino también es una técnica útil en otras situaciones, por ejemplo en la reducción de datos.

Además de los seis capítulos anteriores, se incluyen tres apéndices en los que se pueden encontrar la teoría de un árbol y los datos de dos aplicaciones, respectivamente, por lo que se recomienda leer primero el apéndice A.

Oscar Pablo Herrera Villalobos

Capítulo 1

Introducción

1.1. Analisis multivariado

El análisis multivariado puede ser definido como la aplicación de métodos que, junto con un número razonable de medidas o variables, son trabajadas como un mismo objeto con una o varias muestras al mismo tiempo.

El análisis multivariado se refiere a todos los métodos estadísticos que analizan múltiples medidas (o variables) al mismo tiempo de cada individuo u objeto en investigación.

Las técnicas multivariadas difieren del análisis univariado y del análisis bivariado en que éstas ponen atención en la media y la varianza de una sola variable, o la relación entre dos variables como su covarianza o su correlacion, esto refleja la necesidad de una extensión a tres o más variables. Cualquier análisis simultáneo de más de dos variables puede ser considerado como análisis multivariado.

Las medidas están relacionadas a características o atributos de los objetos que van a ser estudiados y más en general, podemos llamar variables a éstas.

Las técnicas analíticas del análsis multivarido son ampliamente aplicadas en la industria, gobierno y en centros de investigación universitarios. Más aún, pocos campos de estudio de investigación han fallado por incorporar técnicas multivaridas a su "caja de herramientas". Su uso se ha incrementado, numerosos artículos y libros han sido publicados para fines teóricos y matemáticos de estas herramientas, y textos introductorios han aparecido en casi todas las ramas, sin embargo pocos libros han sido escritos para investigadores que no son especialistas en matemáticas o en estadística.

1.2. Conceptos Básicos del Análsis Multivariado

En esta sección se dan algunas definiciones que serán usadas a lo largo de este trabajo.

1.2.1. Variable

Una variable es una combinación lineal con pesos empíricos determinados. Las variables son especificadas por el investigador y los pesos están determinados por el objeto específico de la técnica multivariada. Así, la variable de n variables ponderadas $(X_1, \ldots, X_n)$ puede ser vista como:

$$Variable = w_1X_1 + w_2X_2 + w_3X_3 + \ldots + w_nX_n,$$

donde X_n es la variable observada y w_n son los pesos determinados por la técnica multivariada.

El resultado es la representación de una sola variable como combinación lineal de un conjunto de variables.

1.2.2. Escalas de medidas

El análisis de datos involucra una partición, identificación y la asignación de una medida de variación a un conjunto de variables, ya sea entre ellas mismas, entre una variable dependiente o entre una o más variables independientes. La medición es importante para la representación acertada del concepto de interés y es esencial en la selección de algún método de análisis multivariado apropiado.

Hay dos tipos de datos: no-métricos (cualitativos) y los métricos (cuantitativos). Los datos no-métricos tienen atributos, características o propiedades categóricas que describen o identifican al sujeto. Los datos no-métricos describen diferencias en tipo o clase indicando la presencia o ausencia de una característica o propiedad. En cuanto a las variables métricas, éstas reflejan una cantidad relativa o una distancia. En este tipo de variables es posible hacer afirmaciones de cantidad o magnitud, tal como el nivel de satisfacción.

Escalas de medidas no-métricas

Las medidas no-métricas pueden ser hechas con escalas nominales u ordinales.

Nominal. La escala nominal o también conocida como escala categorica está descrita en términos de clases, esto es, los números asignados sólo llevan a colocar un objeto en una y sola una clase mutuamente excluyente y que no implica tener un orden. Los números o símbolos asignados a los objetos no tienen un significado cuantitativo más allá de indicar la presencia o ausencia del atributo o característica bajo investigación.

Ordinal. La escala ordinal clasifica a los datos, esto significa que se puede decir que objeto posee más cantidad o atributo que algún otro objeto. Las escalas ordinales poseen un nivel mayor de precisión de medición. Las variables pueden ser ordenadas o clasificadas con escalas ordinarias en relación con la cantidad del atributo poseído. Cada subclase puede ser comparada en términos de una relación "mayor que" o "menor que", por ejemplo, el producto A "satisface" más las necesidades del consumidor X que las que "satisface" el producto B.

Escalas de medidas métricas

Entre las medidas métricas se encuentran las escalas de intervalo y las de razón, las cuales proporcionan un nivel mayor de precisión, permitiendo un mayor número de operaciones matemáticas

Intervalo. La escala de intervalo indica que tanto más atributo o característica posee un objeto que otro.

Razón. Aqui se puede definir un origen que significa el cero del atributo en cuestion. Las medidas de razón dan la mayor precisión, porque éstas poseen las ventajas de las escalas menores y además tienen un cero absoluto y con este tipo de escala se pueden realizar cualquier tipo de operaciones matemáticas.

El comprender los diferentes tipos de medidas es importante por dos razones. Primero, el investigador debe determinar el tipo de escala usada en cada variable. Segundo, la escala de medición es crítica para determinar que

tipo de técnica multivariada se puede aplicar a los datos, en base a las consideraciones hechas para las variables dependientes y para las independientes.

1.3. Medición de los errores y las medidas multivariadas

Medición de los errores. Es el grado en el cual los valores observados no son representativos de los valores reales.

La meta del investigador de reducir la medida del error puede seguir muchas direcciones. Al valorar el grado de error presente en cualquier medida, el analista debe dirigir la validación y la fiabilidad de la medida. Validación es el grado en que la medida representa exactamente lo que se supone debe representar y Fiabilidad es el grado en que el valor real de las medidas de las variables observadas tienen un "error libre"; esto es lo opuesto de la medición de los errores. Si la medida es usada varias veces, por ejemplo, más medidas fiables presentarán una mejor consistencia que pocas medidas fiables. El investigador siempre debe valorar las variables que está usando y comprobar si es válido usar una medida alternativa, y escoger a la variable con más fiabilidad.

Medidas multivariadas. También conocidas como medidas resumidas, donde varias variables son unidas para representar una variable compuesta. El objetivo es evitar representar una sola variable para representar un sólo objeto, y en lugar de esto usar varias variables como *indicadores*, todas representando diferentes facetas del concepto para obtener una mejor perspectiva. El uso de muchos indicadores dirigen al investigador a una mejor respuesta a las preguntas planteadas.

1.4. Técnicas multivariadas

Se pueden mencionar las siguientes técnicas:

1. Regresión multiple.
2. Análisis de discriminante.
3. Componentes principales y análisis de factores comunes.
4. Análisis multivariado de la varianza y la covarianza (MANOVA).
5. Correlación canónica.

6. Análisis por conglomerados.
7. Escalamiento multidimensional.
8. Análisis conjunto.

Y entre las técnicas emergentes se encuentran:

9. Análisis de correspondencia.
10. Modelos de probabilidad lineales tales como el logit y el probit.
11. Modelación de ecuaciones simultáneas y modelación de ecuaciones estructuradas.

1.4.1. Regresión múltiple.

La regresión múltiple es el método de análisis apropiado cuando el problema del investigador involucra a una variable dependiente supuestamente relacionada a los cambios de una o más variables independientes. El objetivo del análisis de la regresión múltiple es el predecir los cambios de la variable dependiente en relación a los cambios en la(s) variable(s) independiente(s). El investigador está interesado en la predicción de la cantidad o magnitud de la variable dependiente.

1.4.2. Análisis discriminante

El análisis discriminante es útil cuando la muestra total puede estar dividida en grupos basados en una variable dependiente caracterizada por muchas clases conocidas. Los principales objetivos del análisis discriminante son el comprender las diferencias entre los grupos y la predicción de la probabilidad de que una entidad (individuo u objeto) pertenezca a una clase o grupo particular basado en la métrica de muchas variables independientes.

1.4.3. Análisis de factores y componentes principales

El análisis de factores puede ser usado para analizar intercorrelaciones entre un gran número de variables y explicar estas variables en términos de sus dimensiones subyacentes comúnes (factores). El objetivo es encontrar un camino de condensación de la información contenida en un conjunto original de variables en un conjunto menor de variables (factores) con un mínimo de pérdida de información. Asimismo, el análisis de componentes principales busca las combinaciones lineales de las variables originales que posean la maxima varianza, para asi crear unas nuevas variables que posean casi toda

la información de las originales, el término "casi" se refiere, al igual que el análisis de factores, a un mínimo de pérdida de información.

1.4.4. Análisis multivariado de la varianza

El análisis múltiple de la varianza o MANOVA por sus siglas en inglés es una técnica estadística que puede ser usada para explorar la relación entre varias variables categóricas independientes (tratamientos) en una o más variables dependientes. Es una extensión del Análisis de la varianza o ANOVA. El análisis multivariado de la covarianza o MANCOVA puede ser usada en conjunto con la MANOVA para quitar (después del experimento) el efecto de cualquier variable independiente no-controlada en las variables dependientes. La MANOVA es útil cuando el investigador diseña un experimento (la manipulación del tratamiento de variables no-métricas) a las pruebas de hipótesis concernientes a la varianza en grupos de respuesta en dos o más variables dependientes.

1.4.5. Correlación canónica

El análisis de correlación canónica puede ser visto como una extensión lógica del análisis de regresión multiple. Su objetivo es corelacionar varias variables dependientes e independientes al mismo tiempo, así, la regresión multiple involucra una variable dependiente en contraste a la correlación canónica que involucra varias. Su principio esta basado en desarrollar una combinación lineal de cada conjunto de variables, una para las variables dependientes y otra para las independientes, para maximizar la correlación entre los dos conjuntos. El procedimento involucra el obtener un conjunto de pesos para las variables dependientes y otro conjunto para las independientes que proveen la máxima correlación simple entre ambos conjuntos.

1.4.6. Análisis por conglomerados

El análisis por conglomerados es una técnica analítica para desarrollar subgrupos de individuos u objetos. Específicamente, el objetivo es clasificar una muestra de entidades (individuos u objetos) en un número pequeño de grupos mutuamente exclusivos basado en las similitudes entre las entidades. En el análisis por conglomerados los grupos no están predeterminados, en cambio, la técnica se usa para identificar los grupos.

El análisis por conglomerados usualmente involucra al menos dos pasos. El primero es la medición de alguna forma de similaridad o asociación entre las entidades para determinar cuantos grupos realmente existen en la muestra. El segundo paso es perfilar a las personas o variables para determinar su composición. Este paso puede lograrse aplicando el análisis discriminante a los grupos identificados por la técnica del conglomerado.

1.4.7. Escalamiento multidimensional

Su objetivo es transformar los juicios del consumidor de similitud o preferencia (por ejemplo, preferencias de tiendas o marcas) en distancias representadas en un espacio multidimensional. Si dos objetos A y B son juzgados por ser los más similares comparados con cualquier otro par de parejas de objetos, entonces la técnica del escalamiento multidimensional proporcionará la distancia más corta entre cualquier otro par de objetos. Los mapas resultantes muestran la posición relativa de todos los objetos, pero un análisis adicional es requerido para valorar qué atributos predicen la posición de cada objeto.

1.4.8. Análisis conjunto

El análisis conjunto es una técnica emergente que ha traído nuevas sofisticaciones a la evaluación de objetos, si éstos son nuevos productos, servicios o ideas. La aplicación más directa es en el desarrollo de un nuevo producto o servicio permitiendo la evaluación de productos complejos, manteniendo un marco de decisión más realista para su respuesta. El investigador de mercados es capaz de comprender la importancia de los atributos tan bien como los niveles de cada atributo mientras que los consumidores sólo evalúan unos pocos pérfiles del producto. Más aún, cuando las evaluaciones del consumidor están completas, los resultados del análisis conjunto pueden ser usados en simuladores para el diseño de productos, que muestran la aceptación del cliente para cualquier número de formulaciones del producto y ayudan al diseño de un producto óptimo.

1.4.9. Análisis de correspondencias

Es una técnica reciente que facilita la reducción dimensional de la valuación de objetos (productos, servicios, etc.) en un conjunto de atributos y la idea perceptual de los objetos relativos a esos atributos. El análisis de

correspondencia se distingue por su habilidad de acomodar las relaciones no-lineales y los datos no-métricos.

1.4.10. Modelos de probabilidad lineales

Los modelos de probabilidad lineales, usualmente referidos como el análisis logit, son una combinación de regresión multiple y análisis de discriminante. Esta técnica es similar al análisis de regresión multiple en el hecho en que una o más variables independientes son usadas para predecir a una variable dependiente. Lo que distingue a los modelos de probabilidad lineales de la regresión multiple es que la variable dependiente es no-métrica como en el análisis discriminante. La escala no-métrica en la variable dependiente requiere diferencias en el método de estimación y suposiciones acerca del tipo de distribución, y en casi todo lo demás es similar a la regresión multiple. Así una vez que la variable dependiente y la técnica empleada esten bien específicadas, los factores considerados en la regresión multiple funcionan muy bien. Los modelos de probabilidad lineales se diferencian del análisis discriminante principalmente en que estos acomodan todos los tipos de variables independientes (métricas y no-métricas) y no requieren la suposición de la distribución normal.

1.4.11. Modelación de ecuaciones estructuradas

La modelación de ecuaciones estructuradas es una técnica que permite tratar por separado las relaciones de cada conjunto de variables dependientes, esto es, la modelación de ecuaciones estructuradas proveen la técnica apropiada y la más eficiente para estimar series de ecuaciones de regresiones múltiples separadas al mismo tiempo. Está caracterizada por dos conceptos básicos: (1) El modelo estructural y (2) El modelo de la medición. El modelo estructural relaciona a las variables independientes con las variables dependientes y el modelo de la medición permite al investigador el uso de varias variables (indicadores) para una sola variable dependiente o independiente, también el investigador puede valorar la contribución de cada escala e incorporar las escalas de medición al concepto en la estimación de las relaciones entre la variable dependiente y las variables independientes.

1.5. Un acercamiento a la construcción del modelo multivariado

Los pasos generales involucrados en cualquier análisis multivariado son los siguientes:

1.5.1. Definición del problema a investigar, objetivos y la técnica multivariada que va a ser usada

El primer punto a tratar para cualquier análisis multivariado es el definir el problema a investigar y analizar los objetivos en términos conceptuales antes de específicar cualquier variable o medida.

Un modelo conceptual no necesita ser complejo ni detallado sino una simple relación de lo estudiado.

1.5.2. Desarrollo para el plan del análisis

Con el establecimiento del modelo conceptual, se necesita la técnica a aplicar. Para cada técnica el investigador debe desarrollar un plan de análisis específico que dirija al conjunto de cosas particulares a su propósito o diseño. El rango del problema de las consideraciones generales de minimizar o desear tamaños de muestras pequeños, permitir o requerir tipos de variables (métricas contra las no-métricas) y métodos de estimación, especificar el tipo de asociación de las medidas usadas en el escalamiento multidimensional, la estimación de la agregación o desagregación de resultados en conjunto, o el uso de una variable especial para representar efectos no lineales o interactivos en regresión. En cada ejemplo, estas cosas resuelven detalles específicos y finalmente concluyen la formulación del modelo y los requerimientos para la colección de los datos.

1.5.3. Evaluación de las suposiciones fundamentales de la técnicas multivariadas

Con los datos coleccionados, el primer análisis es evaluar las suposiciones fundamentales. Todas las técnicas multivariadas tienen suposiciones, estadísticas y conceptuales, que sustancialmente impactan su habilidad para representar las relaciones multivariadas. Para técnicas basadas en inferencia es-

tadística, las suposiciones de la normalidad, linealidad, independencia de los términos de los errores, y la igualdad de varianzas en una relación de dependencia deben ser conocidas. También cada técnica tiene una serie de suposiciones conceptuales tratadas como cosas en la formulación del modelo y los tipos de relaciones representadas. Antes de que se intente cualquier estimación del modelo, el investigador debe conocer las suposiciones estadísticas así como las suposiciones conceptuales.

1.5.4. Estimación del modelo multivariado y el cálculo del modelo ajustado

Cuando las hipótesis son satisfechas, prosigue el análisis de la estimación del modelo multivariado y la valoración del cálculo del modelo ajustado. En el proceso de estimación, el analista puede elegir entre opciones para conocer características específicas de los datos (como por ejemplo el uso de covarianzas en la MANOVA) o el maximizar el ajuste de los datos (rotación de factores o las funciones discriminantes). Después el modelo es estimado, el cálculo del modelo ajustado es evaluado según si determina niveles aceptables de un cierto criterio estadístico (por decir, nivel de significancia), si identifica las relaciones propuestas, y si logra la importancia práctica. Muchas veces el modelo será reespecificado en un intento de lograr mejores niveles de ajuste y/o explicación. Un modelo aceptable debe ser obtenido antes de prosegir.

Sin importar que nivel de ajuste sea encontrado, el analista debe determinar si los resultados son considerablemente afectados por una o un conjunto pequeño de observaciones indicando que los resultados pueden ser inestables. Estos esfuerzos ayudan a que los resultados sean "robustos" y estables aplicando razonablemente bien todas las observaciones en la muestra. Las malas observaciones pueden ser indentificadas como datos atípicos, observaciones influenciadas o cualquier otro tipo de resultado disparado.

1.5.5. Interpretación de la variable

Con un aceptable nivel de ajuste, el interpretar la variable revela la naturaleza de la relación multivariada. La interpretación de los efectos de variables individuales es hecha examinando los coeficientes estimados (pesos) de cada variable en la variación (pesos de la regresión, los factores de carga o utilidades conjuntas). Más aún, algunas técnicas también estiman las variaciones

múltiples que representan las dimensiones fundamentales de comparación o asociación (funciones discriminantes o componentes principales). La interpretación puede llevar a reespecificaciones adicionales de las variables y/o la formulación del modelo, en que el modelo es reestimado y así interpretado de nuevo. El objetivo es identificar evidencia empírica de las relaciones multivariadas en los datos de la muestra que pueden ser generalizados de la población total.

1.5.6. Validación del modelo

Antes de aceptar los resultados, el investigador debe someterlos a un conjunto de diagnósticos finales para evaluar el grado de generalizabilidad de los resultados por medio de los métodos de validación disponibles.

Capítulo 2

Álgebra de matrices

Definición 2.1 *Una matriz* **A** *es un arreglo rectangular de números. Si* **A** *tiene n filas y p columnas se dice que es una matriz* $n \times p$. *Por ejemplo, n observaciones en p variables aleatorias, es decir* **A** *es una matriz* $n \times p$, *se escribe como*

$$\mathbf{A} = \begin{pmatrix} A_{1\,1} & A_{1\,2} & \dots & A_{1\,p} \\ A_{2\,1} & A_{2\,2} & \dots & A_{2\,p} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n\,1} & A_{n\,2} & \dots & A_{n\,p} \end{pmatrix} = (A_{i\,j}) \tag{2.1}$$

donde $A_{i\,j}$ *es el elemento que esta en la fila* i *y en la columna* j *de la matriz* **A**., $i = 1, ..., n$; $j = 1, ..., p$. *Se escribirá a la matriz* **A** *como* $\mathbf{A}(n \times p)$ *para enfatizar el número de filas y el de columnas. Si* $n = p$, *se le llamará matriz cuadrada.*

Definición 2.2 *La traspuesta de la matriz* **A** *es formada intercambiando las filas y las columnas:*

$$\mathbf{A}^T = \begin{pmatrix} A_{1\,1} & A_{2\,1} & \dots & A_{n\,1} \\ A_{1\,2} & A_{2\,2} & \dots & A_{n\,2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1\,p} & A_{2\,p} & \dots & A_{n\,p} \end{pmatrix} \tag{2.2}$$

Definición 2.3 *Una matriz con una columna es llamada un vector columna,*

así

$$\underline{A} = \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_n \end{pmatrix} \tag{2.3}$$

es un vector columna de n componentes. Los vectores fila son escritos como de vectores columna traspuestos, esto es

$$\underline{A}^T = \begin{pmatrix} A_1 & A_2 & \dots & A_n \end{pmatrix} \tag{2.4}$$

Notación. *Se escribe a las columnas de la matriz* $\mathbf{A}$ *como* $\underline{A}_{(1)}, \underline{A}_{(2)}, ..., \underline{A}_{(p)}$ *y las filas como* $\underline{A}_1, \underline{A}_2, ..., \underline{A}_n$ *tal que*

$$\mathbf{A} = \begin{pmatrix} \underline{A}_{(1)} & \underline{A}_{(2)} & \dots & \underline{A}_{(p)} \end{pmatrix} = \begin{pmatrix} \underline{A}_1 \\ \underline{A}_2 \\ \vdots \\ \underline{A}_n \end{pmatrix} \tag{2.5}$$

donde

$$\underline{A}_{(j)} = \begin{pmatrix} A_{1j} \\ A_{2j} \\ \vdots \\ A_{nj} \end{pmatrix}, \qquad \underline{A}_i = \begin{pmatrix} A_{i1} & A_{i2} & \dots & A_{ip} \end{pmatrix} \tag{2.6}$$

Definición 2.4 *Una matriz escrita en términos de submatrices es llamada una matriz particionada.* ***Notación.*** *Sean* $\mathbf{A}_{11}$, $\mathbf{A}_{12}, \mathbf{A}_{21}$ *y* $\mathbf{A}_{22}$ *submatrices tales que* $\mathbf{A}_{11}(r \times s)$ *tiene como elementos* A_{ij}, $i = 1, ..., r$; $j = 1, ..., s$ *entonces*

$$\mathbf{A}(n \times p) = \begin{pmatrix} \mathbf{A}_{11}(r \times s) & \mathbf{A}_{12}(r \times (p-s)) \\ \mathbf{A}_{21}((n-r) \times s) & \mathbf{A}_{22}(n-r) \times (p-s)) \end{pmatrix} \tag{2.7}$$

Así, esta notación puede ser extendida para más particiones de $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, *etcétera..*

Definición 2.5 *Sea* $\mathbf{I}(n \times n)$ *una matriz cuadrada. Se dice que* $\mathbf{I}$ *es la matriz identidad si*

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & 1 & 0 \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix}$$

Definición 2.6 *El determinante de una matriz cuadrada* $\mathbf{A}$ *es definido como*

$$|\mathbf{A}| = \sum (-1)^{|\tau|} A_{1\tau(1)} \cdots A_{p\tau(p)} \tag{2.8}$$

donde la suma es tomada sobre todas las permutaciones τ *de* $(1, 2, ..., p)$ *y* $|\tau|$ *es igual a 1 o -1 dependiendo si* τ *puede ser escrito como el producto de un número par o impar de transposiciones.*

Definición 2.7 *Una matriz cuadrada es no-singular si* $|\mathbf{A}| \neq 0$, *de lo contrario es singular.*

Definición 2.8 *La inversa de la matriz* $\mathbf{A}$, *es una matriz* $\mathbf{A}^{-1}$ *que satisface* $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. *La inversa existe si y sólo si* $\mathbf{A}$ *es no-singular, esto es, si y sólo si* $|\mathbf{A}| \neq 0$.

Definición 2.9 *Una matriz* $\mathbf{A}$ *cuadrada es llamada simétrica si* $A_{ij} = A_{ji}$ *para toda* $1 \leq i, j \leq n$

Definición 2.10 *La traza de una matriz simétrica se define como*

$$tr\mathbf{A} = \sum A_{i,i}$$

Definición 2.11 *Sea* $\mathbf{A}$ *una matriz cuadrada. Se dice que* $\mathbf{A}$ *es una matriz diagonal si*

$$\mathbf{A} = \begin{pmatrix} A_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & A_p \end{pmatrix} \tag{2.9}$$

2.1. Espacios vectoriales

Definición 2.12 *Un conjunto* $\mathbf{X}$ *se llama espacio vectorial si para cualquiera dos elementos de* $\mathbf{X}$, *a los que se les llamará vectores, satisfacen:*

1. $\mathbf{X}$ es cerrado bajo la suma. Si $\underline{X}_1, \underline{X}_2 \in \mathbf{X}$ entonces

$$\underline{X}_1 + \underline{X}_2 \in \mathbf{X}.$$

2. La suma es asociativa. Si $\underline{X}_1, \underline{X}_2, \underline{X}_3 \in \mathbf{X}$ entonces

$$(\underline{X}_1 + \underline{X}_2) + \underline{X}_3 = \underline{X}_1 + (\underline{X}_2 + \underline{X}_3).$$

3. Existe un elemento neutro para la suma. Existe $\underline{0} \in \mathbf{X}$ tal que

$$\underline{0} + \underline{X} = \underline{X} + \underline{0} = \underline{X}$$

para todo $\underline{X} \in \mathbf{X}$.

4. Cada elemento tiene inverso. Para cada $\underline{X} \in \mathbf{X}$ existe -$\underline{X} \in \mathbf{X}$ tal que

$$-\underline{X} + \underline{X} = \underline{X} + (-\underline{X}) = \underline{0}.$$

5. La suma es conmutativa. Si $\underline{X}_1, \underline{X}_2 \in \mathbf{X}$ entonces

$$\underline{X}_1 + \underline{X}_2 = \underline{X}_1 + \underline{X}_2.$$

6. El producto es cerrado bajo escalares. $\lambda\underline{X} \in \mathbf{X}$ para cualesquiera $\lambda \in \mathbb{R}$, $\underline{X} \in \mathbf{X}$.

7. El producto de un escalar se distribuye sobre la suma de vectores.

$$\lambda(\underline{X}_1 + \underline{X}_2) = \lambda\underline{X}_1 + \lambda\underline{X}_2$$

para cualesquiera $\lambda \in \mathbb{R}$, $\underline{X}_1, \underline{X}_2 \in \mathbf{X}$.

8. El producto de un escalar distribuye la suma de vectores.

$$(\lambda + \mu)\underline{X} = \lambda\underline{X} + \mu\underline{X}$$

para cualesquiera $\lambda, \mu \in \mathbf{X}$ y $\underline{X} \in \mathbf{X}$.

9. El producto de escalares puede asociarse de cualquier forma.

$$(\lambda\mu)\underline{X} = \lambda(\mu\underline{X})$$

para cualesquiera $\lambda, \mu \in \mathbf{X}$ y $\underline{X} \in \mathbf{X}$.

10. El elemento real 1 funciona como neutro multiplicativo.

$$1\underline{X} = \underline{X}$$

para cualquier $\underline{X} \in \mathbf{X}$.

Ejemplos:

1. $\mathbb{R}^n$ con $n \in \mathbb{N}$
2. $\mathcal{F}_R = \{f \text{ función} \mid f : \mathbb{R} \to \mathbb{R}\}$
3. $C_R = \{f \text{ función} \mid f : \mathbb{R} \to \mathbb{R} \text{ y } f \text{ es continua}\}$

Lema 2.1 *En un espacio vectorial* $\boldsymbol{X}$ *son válidas las igualdades siguientes para cualquier* $\underline{X}$ *en* $\boldsymbol{X}$*:*

1. $0\underline{X} = \underline{0}$
2. $(-1)\underline{X} = -\underline{X}$

Demostración. 1. Obsérvese que

$$0\underline{X} + \underline{X} = 0\underline{X} + 1\underline{X} = (0+1)\underline{X} = 1\underline{X} = \underline{X}$$

por tanto si se suma $-\underline{X}$ de ambos lados se tiene $0\underline{X} = \underline{0}$.

2. Obsérvese que

$$(-1)\underline{X} + \underline{X} = (-1)\underline{X} + 1\underline{X} = (-1+1)\underline{X} = 0\underline{X} = \underline{0}$$

por lo que implica que $(-1)\underline{X}$ funciona como el inverso de $\underline{X}$, el cual es único. ■

Definición 2.13 *Si* $\mathbf{X}$ *es un espacio vectorial y* $\mathbf{Y}$ *es un subconjunto de* $\mathbf{X}$*, entonces se dice que* $\mathbf{Y}$ *es subespacio vectorial de* $\mathbf{X}$*, si y sólo si se cumplen las dos condiciones siguientes:*

1. $\mathbf{Y} \neq \phi$.
2. $\lambda\underline{X} + \mu\underline{Y} \in \mathbf{Y}$ para cualesquiera $\lambda, \mu \in \mathbb{R}$ y $\underline{X}, \underline{Y} \in \mathbf{Y}$

Ejemplos. $\{\underline{0}\}$ y $\mathbb{R}^n$ son subespacios vectoriales de $\mathbb{R}^n$.

Definición 2.14 *Los vectores $\underline{X}_1, ..., \underline{X}_k$ son llamados linealmente independientes si existen números $\lambda_1, ..., \lambda_k$, no todos cero, tales que*

$$\lambda_1 \underline{X}_1 + \cdots + \lambda_k \underline{X}_k = \underline{0},$$

de lo contrario son llamados linealmente dependientes.

Definición 2.15 *Sea* **W** *un subespacio de $\mathbb{R}^n$, entonces una base de* **W** *es un conjunto máximo de vectores linealmente independientes.*

Las siguientes propiedades se tienen para una base de **W**.

1. Cada base (finita) de **W** contiene el mismo número de elementos. Este número es llamado la dimensión de **W** y se denota como dim **W**. En particular dim $\mathbb{R}^n = n$.

2. Si $\underline{X}_1, ..., \underline{X}_k$ es una base de **W** entonces cada elemento $\underline{X}$ en **W** puede ser expresado como combinación lineal de $\underline{X}_1, ..., \underline{X}_k$; esto es, $\underline{X} = \lambda_1 \underline{X}_1 + \cdots + \lambda_k \underline{X}_k$ para algunos números $\lambda_1, ..., \lambda_k$.

Definición 2.16 *El producto escalar o producto punto entre dos vectores $\underline{X}, \underline{Y} \in \mathbb{R}^n$ esta definido por*

$$\underline{X} \cdot \underline{Y} = \underline{X}^T \underline{Y} = \sum_{i=1}^{n} X_i Y_i.$$

Los vectores se llaman ortogonales si $\underline{X} \cdot \underline{Y} = 0$.

Definición 2.17 *La norma Euclidiana de un vector $\underline{X} \in \mathbb{R}^n$ se define como*

$$\|\underline{X}\| = \left(\underline{X}^T \cdot \underline{X}\right)^{1/2} = \left(\sum_{i=1}^{n} X_i^2\right)^{1/2}$$

Definición 2.18 *Una base $\underline{X}_1, ..., \underline{X}_k$ de un subespacio* **W** *de $\mathbb{R}^n$ es llamada ortonormal si todos sus elementos tienen norma 1 y son ortogonales entre si, esto es, si*

$$\underline{X}_i^T \underline{X}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

2.2. Valores y vectores propios

Si $\mathbf{A}(p \times p)$ es una matriz cuadrada entonces

$$q(\lambda) = |\mathbf{A} - \lambda \mathbf{I}| \tag{2.10}$$

es el polinomio de grado p en la variable λ. Las p raíces de $q(\lambda)$, $\lambda_1, \lambda_2, ..., \lambda_p$, posiblemente números complejos, son llamados valores propios, valores característicos o eigenvalores de $\mathbf{A}$. Algunas de las λ's pueden ser iguales.

Para cada $i = 1, 2, ..., p$, $|\mathbf{A} - \lambda_i \mathbf{I}| = 0$, así que $\mathbf{A} - \lambda \mathbf{I}$ es singular. Así existe un vector $\underline{\gamma} \neq \underline{0}$ que satisface

$$\mathbf{A}\underline{\gamma} = \lambda_i \underline{\gamma} \tag{2.11}$$

Cualquier vector que satisfaga (2.11) es llamado vector propio (derecho), vector característico o eigenvector de $\mathbf{A}$ para el valor propio λ_i. Si λ_i es complejo entonces γ puede tener entradas complejas. Un vector propio con entradas reales se dice que está estandarizado si $\underline{\gamma}^T \underline{\gamma} = 1$.

Si $\underline{X}$ y $\underline{Y}$ son vectores propios para λ_i, y $\alpha \in \mathbb{R}^n$, entonces $\underline{X} + \underline{Y}$ y $\alpha \underline{X}$ también son vectores propios para λ_i. De esta manera, el conjunto de todos los vectores propios para λ_i forma un subespacio vectorial que es conocido como *espacio propio o eigenespacio.*

Como el coeficiente de λ^p en $q(\lambda)$ es $(-1)^p$, podemos escribir a $q(\lambda)$ en términos de sus raíces como

$$q(\lambda) = \prod_{i=1}^{p} (\lambda_i - \lambda). \tag{2.12}$$

Si se sustituye $\lambda = 0$ en (2.10) y en (2.12) se tiene que

$$|\mathbf{A}| = \prod_{i=1}^{p} \lambda_i. \tag{2.13}$$

Sea $\mathbf{C}$ una matriz no-singular $(p \times p)$, entonces

$$\begin{aligned} |\mathbf{A} - \lambda \mathbf{I}| &= |\mathbf{C}| \left|\mathbf{A} - \lambda \mathbf{C}\mathbf{C}^{-1}\right| |\mathbf{C}^{-1}| \\ &= \left|\mathbf{C}\mathbf{A}\mathbf{C}^{-1} - \lambda \mathbf{I}\right|. \end{aligned} \tag{2.14}$$

Así $\mathbf{A}$ y $\mathbf{C}\mathbf{A}\mathbf{C}^{-1}$ tienen los mismos valores propios. Además si $\underline{\gamma}$ es un vector propio de $\mathbf{A}$ para λ_i, entonces $\mathbf{C}\mathbf{A}\mathbf{C}^{-1}(\mathbf{C}\underline{\gamma}) = \lambda_i \mathbf{C}\underline{\gamma}$, de donde si

$$\underline{\nu} = \mathbf{C}\underline{\gamma},$$

entonces $\underline{\nu}$ es un vector propio de $\mathbf{CAC}^{-1}$ para λ_i.

Sea $\alpha \in R$. Entonces $|\mathbf{A}+\alpha\mathbf{I}-\lambda\mathbf{I}| = |\mathbf{A}-(\lambda-\alpha)\mathbf{I}|$, tal que $\mathbf{A}+\alpha\mathbf{I}$ tiene valor propio $\lambda_i+\alpha$. Además, si $\mathbf{A}\underline{\gamma} = \lambda_i\underline{\gamma}$, entonces $(\mathbf{A}+\alpha\mathbf{I})\underline{\gamma} = (\lambda_i+\alpha)\underline{\gamma}$, así que $\mathbf{A}$ y $\mathbf{A}+\alpha\mathbf{I}$ tienen los mismos vectores propios.

Teorema 2.2 *Sea λ_1 cualquier valor propio de $\mathbf{A}$ matriz $(p\times p)$, con espacio propio $\mathbb{H}$ de dimensión r. Si k denota la multiplicidad de λ_i, en $q(\lambda)$, entonces $1 \le r \le k$.*

Demostración. Como λ_1 es un valor propio, existe al menos un vector propio no trivial, así $r \ge 1$.

Sea $\mathbf{e}_1, ..., \mathbf{e}_r$ una base ortonormal de $\mathbf{H}$ y extiéndase a todo $\mathbf{R}^p$, así existen $\mathbf{f}_1, ..., \mathbf{f}_{p-r}$ vectores tales que $\{\mathbf{e}_1, ..., \mathbf{e}_r, \mathbf{f}_1, ..., \mathbf{f}_{p-r}\}$ es una base ortonormal de $\mathbf{R}^p$. Sean $\mathbf{E} = (\mathbf{e}_1, ..., \mathbf{e}_r)$ y $\mathbf{F} = (\mathbf{f}_1, ..., \mathbf{f}_{p-r})$. Entonces $(\mathbf{E}, \mathbf{F})$ es una matriz ortonormal tal que $\mathbf{I}_p = (\mathbf{E},\mathbf{F})(\mathbf{E},\mathbf{F})^T = \mathbf{E}\mathbf{E}^T + \mathbf{F}\mathbf{F}^T$ y $|(\mathbf{E},\mathbf{F})| = 1$. También $\mathbf{E}^T\mathbf{A}\mathbf{E} = \lambda_i\mathbf{E}^T\mathbf{E} = \lambda_i\mathbf{I}_r$, $\mathbf{F}^T\mathbf{F} = \mathbf{I}_{p-r}$ y $\mathbf{F}^T\mathbf{A}\mathbf{E} = \lambda_i\mathbf{F}^T\mathbf{E} = \mathbf{0}$. Así,

$$\begin{aligned}
q(\lambda) &= |\mathbf{A}-\lambda\mathbf{I}| \\
&= \left|(\mathbf{E},\mathbf{F})^T\right| |\mathbf{A}-\lambda\mathbf{I}| \, |(\mathbf{E},\mathbf{F})| \\
&= \left|(\mathbf{E},\mathbf{F})^T \left[\mathbf{A}\mathbf{E}\mathbf{E}^T + \mathbf{A}\mathbf{F}\mathbf{F}^T - \lambda\mathbf{E}\mathbf{E} - \lambda\mathbf{F}\mathbf{F}^T\right](\mathbf{E},\mathbf{F})\right| \\
&= \begin{vmatrix} (\lambda_1-\lambda)\mathbf{I}_r & \mathbf{E}^T\mathbf{A}\mathbf{F} \\ 0 & \mathbf{F}^T\mathbf{A}\mathbf{F} - \lambda\mathbf{I}_{p-r} \end{vmatrix} \\
&= (\lambda_1-\lambda)^r q_1(\lambda)
\end{aligned}$$

por lo que se tiene que la multiplicidad de λ_1 como raíz de $q(\lambda)$ es al menos r. ■

Proposición 2.3 *Sean $\mathbf{A}$ y $\mathbf{B}$ dos matrices cuadradas y no singulares. Sea $c \in \mathbb{R}$ un número real. Se tienen las siguientes propiedades:*

1) $|c\mathbf{A}| = c^p\mathbf{A}$.

2) $|\mathbf{AB}| = |\mathbf{A}|\,|\mathbf{B}|$.

3) Si $\mathbf{A}(p\times p)$ y $\mathbf{B}(q\times q)$ son dos matrices cuadradas, entonces

$$\begin{vmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{0} & \mathbf{B} \end{vmatrix} = |\mathbf{A}|\,|\mathbf{B}|.$$

4) Si $\mathbf{A}$ se puede escribir como

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

entonces $|\mathbf{A}| = |\mathbf{A}_{11}|\left|\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right| = |\mathbf{A}_{22}|\left|\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right|$.

Demostración.

1) Se sigue de la Definición 2.6

2) Ver Strong (1982)

3) Para la prueba consultar Shilov (1977) páginas 20-23 o Leon (1994) página 95.

4) Sea $\mathbf{B} = \begin{pmatrix} \mathbf{I} & -\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$ una matriz, entonces

$$|\mathbf{BAB}^T| = \left| \begin{pmatrix} \mathbf{I} & -\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{12}^T & \mathbf{I} \end{pmatrix} \right|$$

$$\begin{aligned} |\mathbf{BAB}^T| &= \left| \begin{pmatrix} \mathbf{I} & -\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{12}^T & \mathbf{I} \end{pmatrix} \right| \\ &= \left| \begin{pmatrix} \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{0} \\ \mathbf{A}_{21} - \mathbf{A}_{22}^T & \mathbf{A}_{22} \end{pmatrix} \right| \\ &= |\mathbf{A}_{22}| \left| \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} \right| \end{aligned}$$

por la propiedad 3) se tiene $|\mathbf{B}| = |\mathbf{I}| = 1$ y $|\mathbf{B}^T| = |\mathbf{I}| = 1$. Por la propiedad 2)

$$|\mathbf{BAB}^T| = |\mathbf{B}|\,|\mathbf{A}|\,|\mathbf{B}^T| = |\mathbf{A}|$$

lo cual prueba la propiedad 4. ■

Observación 2.1 *Si* $\mathbf{A}$ *es simétrica entonces* $r = k$. *Sin embargo, Si* $\mathbf{A}$ *no es simétrica es posible que* $r < k$. *Por ejemplo*

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$

tiene valor propio de multiplicidad 2, pero el correspondiente eigenspace que es generado por el $(1,0)^T$ *tiene solamente dimensión 1.*

Observación 2.2 *Si* $r = 1$, *entonces el espacio propio para* λ_1 *es único.*

Ahora sea $\mathbf{A}(n \times p)$ y $\mathbf{B}(p \times n)$ dos matrices y supóngase que $n \geq p$. entonces de a.2.3j se tiene que

$$\begin{vmatrix} -\lambda\mathbf{I}_n & -\mathbf{A} \\ \mathbf{B} & \mathbf{I}_p \end{vmatrix} = (-\lambda)^{n-p} |\mathbf{BA} - \lambda\mathbf{I}_n| = |\mathbf{AB} - \lambda\mathbf{I}_n| \qquad (2.15)$$

Teorema 2.4 *Para* $\mathbf{A}(n \times p)$ *y* $\mathbf{B}(p \times n)$, *los valores propios distintos de cero de* $\mathbf{AB}$ *y* $\mathbf{BA}$ *son los mismos y tienen la misma multiplicidad. Si* $\underline{X}$ *es un vector no-trivial de* $\mathbf{AB}$ *para un valor propio* $\lambda(\neq 0)$, *entonces* $\underline{Y} = \mathbf{B}\underline{X}$ *es un vector propio no-trivial de* $\mathbf{BA}$.

Demostración. La primera parte se sigue de (2.15). Para la segunda parte, sea $\underline{Y} = \mathbf{B}\underline{X}$ y sustitúyase en la ecuación $\mathbf{B}(\mathbf{AB}\underline{X}) = \lambda\mathbf{B}\underline{X}$ obteniendo $\mathbf{BA}\underline{Y} = \lambda\underline{Y}$. El vector $\underline{Y}$ es distinto de cero si $\underline{X} \neq 0$. Así $\mathbf{BA}\underline{Y} = \mathbf{AB}\underline{X} = \lambda\underline{Y} \neq 0$. De donde se sigue que $\underline{Y} \neq 0$. ■

2.3. Matrices simétricas

Si $\mathbf{A}$ es simétrica es posible dar información más detallada acerca de sus valores y vectores propios.

Teorema 2.5 *Todos los valores propios de una matriz simétrica* $\mathbf{A}(p \times p)$ *son reales.*

Demostración. Sean

$$\underline{\gamma} = \underline{X} + i\underline{Y}, \qquad \lambda = a + ib, \qquad \underline{\gamma} \neq \underline{0} \tag{2.16}$$

De (2.11) e igualando las partes real e imaginaria, respectivamente, se obtiene

$$\mathbf{A}\underline{X} = a\underline{X} - b\underline{Y}, \qquad \mathbf{A}\underline{Y} = b\underline{X} + a\underline{Y}. \tag{2.17}$$

multiplicando por $\underline{X}$ y $\underline{Y}$ respectivamente, y restando, se obtiene que $b = 0$. Por tanto, si se sustituye este valor en (2.16) se tiene que λ es real. ■

Teorema 2.6 *(Teorema de la Descomposición Espectral o Teorema de la Descomposición de Jordan) Cualquier matriz simétrica* $\mathbf{A}(p \times p)$ *puede ser escrita como*

$$\begin{aligned} \mathbf{A} &= \mathbf{\Gamma\Lambda\Gamma}^T \\ &= \textstyle\sum \lambda_i \underline{\gamma}_{(i)}^T \underline{\gamma}_{(i)} \end{aligned} \tag{2.18}$$

donde $\mathbf{\Lambda}$ *es una matriz diagonal formada por los valores propios de* $\mathbf{A}$, *y* $\mathbf{\Gamma}$ *es una matriz ortogonal cuyas columnas son vectores propios estandarizados.*

Demostración. Supóngase que $\underline{\gamma}_{(1)}, ..., \underline{\gamma}_{(p)}$ son vectores ortonormales tales que $\mathbf{A}\underline{\gamma}_i = \lambda_i\underline{\gamma}_{(i)}$ para algunos números λ_i, en otras palabras, $\underline{\gamma}_{(i)}$ es el vector propio de $\mathbf{A}$ correspondiente al valor propio λ_i. Entonces

$$\underline{\gamma}_{(i)}^T \mathbf{A} \underline{\gamma}_{(i)} = \lambda_i \underline{\gamma}_{(i)}^T \underline{\gamma}_{(i)} = \begin{cases} \lambda_i, & si \quad i = j \\ 0 & si \quad i \neq j \end{cases},$$

o en forma matricial como

$$\mathbf{\Gamma}^T\mathbf{A}\mathbf{\Gamma} = \mathbf{\Lambda}. \tag{2.19}$$

Multiplicando por $\mathbf{\Gamma}$ y $\mathbf{\Gamma}^T$ por el izquierdo y por el lado derecho, respectivamente se obtiene (2.18). De (2.14), $\mathbf{A}$ y $\mathbf{\Gamma}$ tienen los mismos valores propios, así los elementos de $\mathbf{\Gamma}$ son exactamente los valores propios de $\mathbf{A}$ con las mismas multiplicidades.

Nótese que si $\lambda_i \neq \lambda_j$ son dos valores propios distintos con $\underline{X}$ y $\underline{Y}$ como sus respectivos vectores propios, entonces $\lambda_i \underline{X}^T\underline{Y} = \underline{X}^T\mathbf{A}\underline{Y} = \underline{Y}^T\mathbf{A}\underline{X} = \lambda_i\underline{Y}^T\underline{X}$ de aquí que $\underline{Y}^T\underline{X} = \underline{0}$. Así para una matriz simétrica, los vectores propios correspondientes a distintos valores propios son ortogonales uno del otro.

Supóngase que hay k valores propios distintos de $\mathbf{A}$ con los correspondientes espacios propios $\mathbb{H}_1, ..., \mathbb{H}_n$ de dimensiones $r_1, ..., r_k$. Sea

$$r = \sum_{j=1}^{k} r_j,$$

Como distintos espacios propios son ortogonales, existe un conjunto ortonormales de vectores $\underline{E}_1, ..., \underline{E}_r$ tales que los vectores etiquetados de la forma

$$\sum_{i=1}^{j-1} r_i + 1, ..., \sum_{i=1}^{j} r_i$$

forman una base de $\mathbb{H}_j$. Del Teorema (2.2) r_j es menor o igual a la multiplicidad del valor propio correspondiente. Se puede suponer que

$$\mathbf{A}\underline{E}_i = \lambda\underline{E}_i, \qquad i = 1, ..., r,$$

y $r \leq p$. (Si todos los valores propios son distintos, por el Teorema (2.2) se tiene que $r = p$).

Si $r = p$ entonces $\gamma_{(i)} = \underline{E}_i$ y la prueba termina. Se probará que $r < p$ conduce a una contradicción.

Sin pérdida de generalidad, supóngase que todos los valores propios son estrictamente positivos. (En caso contrario se puede reemplazar a la matriz $\mathbf{A}$ por $\mathbf{A}+\alpha\mathbf{I}$ para alguna α adecuada, porque $\mathbf{A}$ y $\mathbf{A}+\alpha\mathbf{I}$ tienen los mismos vectores propios). Sea

$$\mathbf{B} = \mathbf{A} - \sum_{i=1}^{r} \lambda_i \underline{E}_i \underline{E}_i^T,$$

entonces

$$tr\mathbf{B} = tr\mathbf{A} - \sum_{i=1}^{r} \lambda_i (\underline{E}_i^T \underline{E}_i) = \sum_{i=r+1}^{p} \lambda_i > 0,$$

con $r < p$. Asi $\mathbf{B}$ tiene al menos un valor propio estrictamente positivo, sea θ ese valor. Sea $\underline{X} \neq \underline{0}$ su correspondiente vector propio. Entonces para $1 \leq j \leq r$,

$$\theta \underline{E}_j^T \underline{X} = \underline{E}_j^T \mathbf{B} \underline{X} = \left\{ \lambda_j \underline{E}_j^T - \sum_{i=1}^{r} \lambda_i (\underline{E}_j^T \underline{E}_i) \underline{E}_i^T \right\} \underline{X} = \underline{0},$$

así $\underline{X}$ es ortogonal a $\underline{E}_j$, $j = 1, ..., r$. Además

$$\theta \underline{X} = \mathbf{B} \underline{X} = \left(\mathbf{A} - \sum_{i=1}^{r} \lambda_i (\underline{E}_i \underline{E}_i^T) \right) \underline{X} = \mathbf{A} \underline{X} - \sum_{i=1}^{r} \lambda_i (\underline{E}_i^T \underline{X}) \underline{E}_i = \mathbf{A} \underline{X}$$

por consiguiente $\underline{X}$ también es vector propio de $\mathbf{A}$. Así $\theta = \lambda_i$ para alguna i y $\underline{X}$ es una combinación lineal de algunas $\underline{E}_i$, que contradice la ortogonalidad entre $\underline{X}$ y $\underline{E}$. ■

Corolario 2.7 *Si* $\mathbf{A}$ *es una matriz cuadrada, entonces* $tr\mathbf{A} = \sum \lambda_i$.

Demostración. Por el teorema de la descomposición espectral $\mathbf{A}$ se puede escribir como

$$\mathbf{A} = \mathbf{\Gamma \Lambda \Gamma}^T$$

donde $\mathbf{\Lambda}$ es una matriz diagonal formada por los valores propios de $\mathbf{A}$, y $\mathbf{\Gamma}$ es una matriz ortogonal cuyas columnas son vectores propios estandarizados, entonces

$$tr\mathbf{A} = tr\left(\mathbf{\Gamma \Lambda \Gamma}^T\right)$$

pero

$$\begin{aligned} tr\left(\mathbf{\Lambda \Gamma \Lambda}^T\right) &= tr \sum \lambda_i \underline{\gamma}_{(i)} \underline{\gamma}_{(i)}^T \\ &= \sum tr \lambda_i \underline{\gamma}_{(i)} \underline{\gamma}_{(i)}^T \\ &= \sum \lambda_i tr \underline{\gamma}_{(i)} \underline{\gamma}_{(i)}^T \\ &= \sum \lambda_i \end{aligned}$$

el último paso se cumple debido a que $\underline{\gamma}_{(i)} \underline{\gamma}_{(i)}^T = 1$, por ser vectores propios normalizados. ■

Corolario 2.8 *Si* $\mathbf{A}$ *es una matriz simétrica no-singular, entonces para cualquier entero* n,

$$\Lambda^n = diag(\lambda_i^n) \text{ y } \mathbf{A}^n = \mathbf{\Gamma}\mathbf{\Lambda}^n\mathbf{\Gamma}^T. \tag{2.20}$$

Si todos los valores propios de $\mathbf{A}$ *son positivos entonces podemos definir las potencias racionales de la matriz* $\mathbf{A}$ *como*

$$\mathbf{A}^{r/s} = \mathbf{\Gamma}\mathbf{\Lambda}^{r/s}\mathbf{\Gamma}^T \tag{2.21}$$

con r *y* $s > 0$ *enteros. Si algunos de los valores propios son cero, entonces (2.20) y (2.21) se cumplen si los exponentes se restinguen a ser no-negativos.*

Demostración. Como

$$\begin{aligned} \mathbf{A}^2 &= (\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T)^2 \\ &= \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T \\ &= \mathbf{\Gamma}\mathbf{\Lambda}^2\mathbf{\Gamma}^T \end{aligned}$$

y

$$\mathbf{A}^{-1} = \mathbf{\Gamma}\mathbf{\Lambda}^{-1}\mathbf{\Gamma}^T, \qquad \mathbf{\Gamma}^{-1} = diag(\lambda_i^{-1}),$$

por lo que (2.20) puede ser probado mediante inducción. Para comprobar que las potencias racionales tienen sentido, nótese que

$$(\mathbf{A}^{r/s})^S = \underbrace{(\mathbf{\Gamma}\mathbf{\Lambda}^{r/s}\mathbf{\Gamma}^T)\cdots(\mathbf{\Gamma}\mathbf{\Lambda}^{r/s}\mathbf{\Gamma}^T)}_{S \text{ veces}} = \mathbf{\Gamma}\mathbf{\Lambda}^r\mathbf{\Gamma}^T = \mathbf{A}^r$$

■

2.4. Formas cuadráticas

Definición 2.19 *Una forma cuadrática en el vector* $\underline{X}$ *es una función de la forma*

$$Q(\underline{X}) = \underline{X}^T\mathbf{A}\underline{X} = \sum_{i=1}^{p}\sum_{j=1}^{p} A_{ij}X_iX_j, \tag{2.22}$$

donde $\mathbf{A}$ *es una matriz simétrica; esto es*

$$Q(\underline{X}) = A_{11}X_1^2 + ... + A_{pp}X_p^2 + 2A_{12}X_1X_2 + ... + 2A_{p-1}X_{p-1}X_p \tag{2.23}$$

Claramente $Q(\underline{0}) = 0$.

Definición 2.20 *$Q(\underline{X})$ es llamada una forma cuadrática positiva definida (p.d.) si $Q(\underline{X}) > 0$ para toda $\underline{X} \neq 0$.*

Definición 2.21 *$Q(\underline{X})$ es llamada una forma cuadrática semi-definida positiva (s.d.p.) si $Q(\underline{X}) \geq 0$ para toda $\underline{X} \neq 0$.*

Definición 2.22 *Una matriz simétrica $\mathbf{A}$ es llamada positiva definida (p.s.) si $Q(\underline{X})$ lo es, del mismo modo, $\mathbf{A}$ es semi-definida positiva (s.p.d.) si $Q(\underline{X})$ lo es. **Notación:** $\mathbf{A} > 0$ ó $\mathbf{A} \geq 0$ para $\mathbf{A}$ definida positiva o para $\mathbf{A}$ semi-definida positiva, respectivamente. Las formas definida negativa y semi-definida negativa se definen de una manera similar.*

Teorema 2.9 *Para cualquier matriz simétrica $\mathbf{A}$ existe una transformación ortogonal*

$$\underline{Y} = \mathbf{\Gamma}^T \underline{X} \tag{2.24}$$

tal que

$$\underline{X}^T \mathbf{A} \underline{X} = \sum \lambda_i Y_i^2 \tag{2.25}$$

Demostración. Considerése la descomposición espectral dada por

$$\mathbf{A} = \mathbf{\Gamma \Lambda \Gamma}^T, \tag{2.26}$$

de (2.24) se tiene que

$$\begin{aligned} \underline{X}^T \mathbf{A} \underline{X} &= \underline{Y}^T \mathbf{\Gamma}^T \mathbf{A} \mathbf{\Gamma} \underline{Y} \\ &= \underline{Y}^T \mathbf{\Gamma}^T \mathbf{\Gamma \Lambda \Gamma}^T \mathbf{\Gamma} \underline{Y} \\ &= \underline{Y}^T \mathbf{\Lambda} \underline{Y} \end{aligned}$$

de donde la última igualdad es una forma cuadrática, obteniendo (2.25). ■

Observación 2.3 *Es importante recalcar que $\mathbf{\Gamma}$ tiene como columnas a los vectores propios de $\mathbf{A}$ y que $\lambda_1, ..., \lambda_p$ son los valores propios de $\mathbf{A}$.*

Teorema 2.10 *Si $\mathbf{A} > 0$ entonces $\lambda_i > 0$ para $i = 1, ..., p$. Si además $\mathbf{A} \geq 0$ entonces $\lambda_i \geq 0$*

Demostración. Si $\mathbf{A} > 0$ se tiene por (2.9), para todo $\underline{X} \neq \underline{0}$,

$$0 < \underline{X}^T \mathbf{A} \underline{X} = \lambda_1 Y_1^2 + ... + \lambda_p Y_p^2$$

de (2.24) se cumple que si $\underline{X} \neq \underline{0}$ entonces $\underline{Y} \neq \underline{0}$. Eligiendo $Y_1 = 1$, $Y_2 = ... = Y_p = 0$ se obtiene que $\lambda_1 > 0$. De la misma forma se obtiene que $\lambda_i > 0$ para toda i. Claramente, mediante un razonamiento similar, se tiene que si $\mathbf{A} \geq 0$ entonces $\lambda_i \geq 0$. ■

Corolario 2.11 *(Descomposición simétrica)Cualquier matriz* $\mathbf{A} \geq 0$ *puede ser escrita como*

$$\mathbf{A} = \mathbf{B}^2,$$

donde es $\mathbf{B}$ *una matriz simétrica.*

Demostración. Sea $\mathbf{B} = \mathbf{\Gamma} \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{\Gamma}^T$ y sustituyendo en (2.26) se sigue el resultado. ■

Teorema 2.12 *Si* $\mathbf{A} \geq 0$ *es una matriz cuadrada* $(p \times p)$, *entonces para cualquier matriz* $\mathbf{C}$ $(p \times n)$, $\mathbf{C}^T \mathbf{A} \mathbf{C} \geq 0$. *Si* $\mathbf{A} > 0$ *y* $\mathbf{C}$ *es no-singular entonces* $\mathbf{C}^T \mathbf{A} \mathbf{C} > 0$.

Demostración. Si $\mathbf{A} \geq 0$ entonces para cualquier vector $\underline{X} \neq \underline{0}$ $(n \times 1)$ se tiene

$$\underline{X}^T \mathbf{C}^T \mathbf{A} \mathbf{C} \underline{X} = (\mathbf{C} \underline{X})^T \mathbf{A} (\mathbf{C} \underline{X}) \geq \underline{0}$$

de aqui que la matriz $\mathbf{C}^T \mathbf{A} \mathbf{C}$ sea semi-definida positiva $(\mathbf{C}^T \mathbf{A} \mathbf{C} \geq 0)$.

Si $\mathbf{A} > 0$ y $\mathbf{C}$ es no-singular entonces $\mathbf{C}^T \underline{X} \neq \underline{0}$, así $(\mathbf{C} \underline{\mathbf{X}})^T \mathbf{A} (\mathbf{C} \underline{\mathbf{X}}) > 0$, por tanto la matriz $\mathbf{C}^T \mathbf{A} \mathbf{C} > 0$. ■

Corolario 2.13 *Si* $\mathbf{A} \geq 0$ *y* $\mathbf{B} > 0$ *son matrices* $(p \times p)$, *entonces todos los valores propios, distintos de cero, de* $\mathbf{B}^{-1} \mathbf{A}$ *son positivos.*

Demostración. Como $\mathbf{B} > 0$, entonces $\mathbf{B}^{\frac{1}{2}}$ existe y por el Teorema (2.4) , $\mathbf{B}^{-1/2} \mathbf{A} \mathbf{B}^{-1/2}$, $\mathbf{B}^{-1} \mathbf{A}$ y $\mathbf{A} \mathbf{B}^{-1}$ tienen los mismos valores propios. Por el Teorema (2.12) $\mathbf{B} \mathbf{A} \mathbf{B} \geq 0$, así todos sus valores propios distintos de cero son positivos. ■

Capítulo 3

Análisis Exploratorio de Datos

3.1. Gráficas Multivariadas

Las gráficas de datos multivariados son importantes en todos los aspectos para su análisis. En términos generales, las gráficas pueden dar ideas acerca de la estructura de los datos, y en particular éstas pueden ser útiles para sugerir que los datos deben de ser agrupados. La utilidad de las gráficas se basan en el poder del sistema visual del hombre para detectar modelos. Sin embargo el tipo de gráficas que normalmente usamos es muy limitado respecto al número de dimensiones. Es fácil tener una forma gráfica en una, dos o tres dimensiones, pero en dimensiones mayores surgen limitantes debido al espacio en que vivimos, ya que éste es tridimensional y es difícil, y a veces hasta imposible, imaginar un espacio de más dimensiones. Así pues, se han desarrollado técnicas gráficas, asi como el uso de técnicas multivariadas para descubrir el número de conglomerados involucrados en los datos.

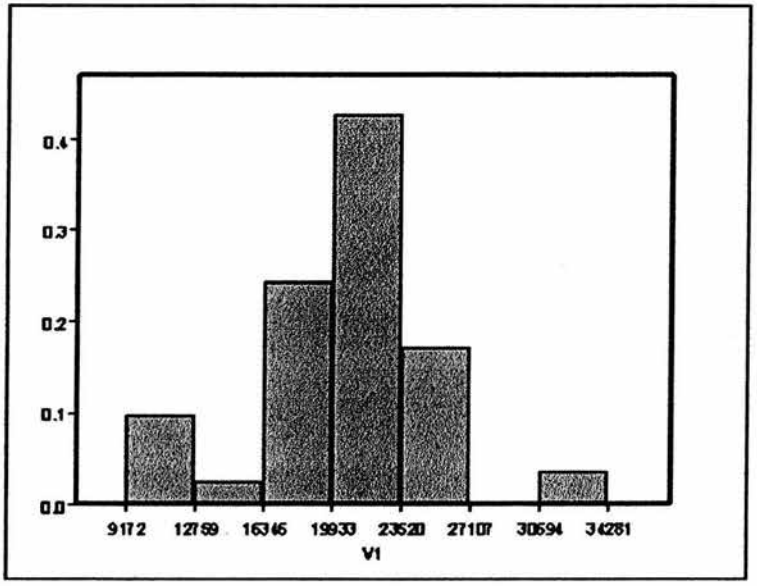
3.1.1. Detectando conglomerados en una o dos dimensiones

Generalmente se dice que una distribución unimodal corresponde a una población homogénea no agrupada por conglomerados y, por el otro lado, la existencia de distintas modas indica una población heterogénea agrupada por conglomerados, con cada moda correspondiente a cada conglomerado de las observaciones. Bajo esta suposición es posible visualizar gráficamente conglomerados de datos univariados y bivariados

Entre los tipos de gráficas que se pueden utilzar para este propósito, se pueden encontrar los siguientes tipos.

Histogramas y gráficas de dispersión

En una dimensión, un simple histograma puede ayudar a identificar modas en los datos, por ejemplo observése la figura

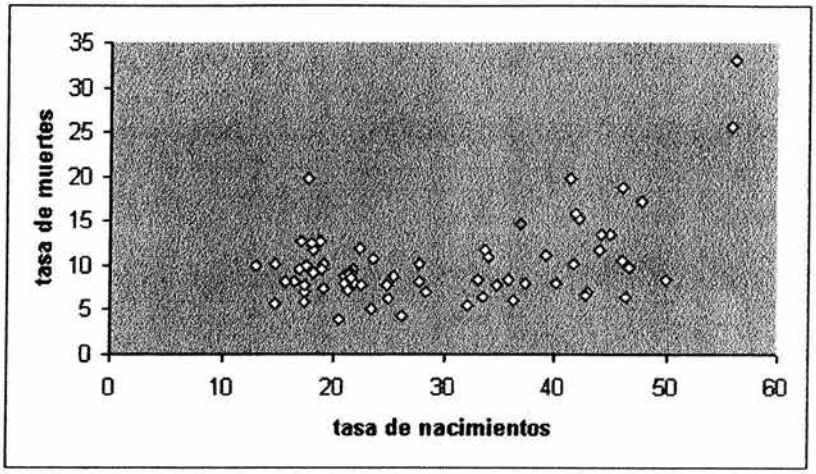


Histograma.

en el cual se puede observar un histograma de velocidades de 82 galaxias.

Para datos bivariados, una gráfica de dispersiones puede ser muy útil, por ejemplo la gráfica de dispersiones en el cual suguiere la existencia de al

menos dos conglomerados, como lo muestra la siguiente figura.



Grafica de dispersion.

Estimadores de los núcleos de una densidad

De la definición de una función de densidad, si X es una variable aleatoria con densidad f, entonces

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h).$$

Para una h dada, un estimador de $P(x-h<X<x+h)$ es la proporción de las estimaciones $X_1, ..., X_n$ que cayeron en el intervalo $(x-h, x+h)$, esto es,

$$\widehat{f}(x) = \frac{1}{2hn} [\text{no. de } X_1, ..., X_n \text{ que cayeron en el intervalo } (x-h, x+h)].$$

Si se introduce una función de ponderación W dada por

$$W(x) = \begin{cases} \frac{1}{2} & \text{si } |x| < 1 \\ 0 & \text{cualquier otro caso,} \end{cases}$$

entonces el estimador arriba mencionado para $\widehat{f}(x)$ se puede escribir como

$$\widehat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} W\left(\frac{x - X_i}{h}\right)$$

desafortunadamente este estimador no es una función continua, por consiguiente no es satisfactoriamente práctica para estimaciones de densidades. Sin embargo, da una idea para definir el estimador del núcleo el cual esta dado por:

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right),$$

donde K es conocido como la *función núcleo* y h es el ancho del intervalo. La función núcleo debe satisfacer la condición $\int_{-\infty}^{\infty} K(x)\,dx = 1$.

Las 3 funciones núcleo más usadas son la *rectangular, triangular y la Gausiana:*

- Rectangular,

$$K(x) = \begin{cases} \frac{1}{2} & \text{si } |x| < 1 \\ 0 & \text{cualquier otro caso} \end{cases}$$

- Triángular,

$$K(x) = \begin{cases} 1 - |x| & \text{si } |x| < 1 \\ 0 & \text{cualquier otro caso} \end{cases}$$

- Gausiana,

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

El estimador de núcleos de una densidad tiene una extensión a dos dimensiones, y de manera análoga a más de dos dimensiones. El estimador bivariado para datos de la forma $(X_1, Y_1), ..., (X_n, Y_n)$ está definido como

$$\widehat{f}(x, y) = \frac{1}{nh_x h_y}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_x}, \frac{y - Y_i}{h_y}\right).$$

En este estimador cada coordenada tiene su propio parámetro h_x y h_y. Una alternativa es utilizar el mismo párametro de estimación.

Para la estimación de una densidad bivariada comúnmente se usa la función núcleo de una normal bivariada estandarizada,

$$K(x, y) = \frac{1}{2\pi}\exp\left\{-\frac{1}{2}\left[x^2 + y^2\right]\right\}.$$

Otra posibilidad es el núcleo bivariado de Epanechnikov, dado por

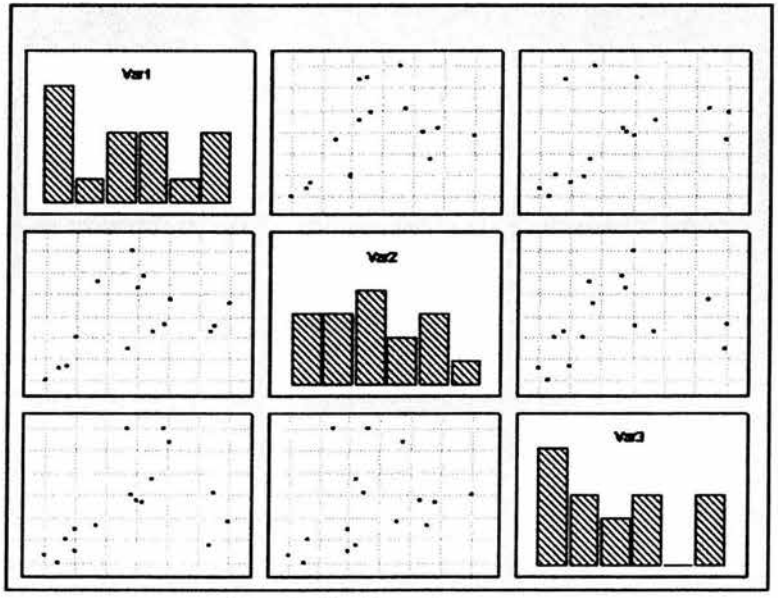
$$K(x, y) = \begin{cases} \frac{2}{\pi}(1 - x^2 - y^2) & \text{si } x^2 + y^2 < 1 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

3.1.2. Detectando conglomerados en más de dos dimensiones

Cuando cada objeto en los datos tiene más de 3 variables, los criterios anteriores son usados en forma indirecta, además surgen criterios nuevos (Diagramas Simbólicos) para la detección de conglomerados. Una posibilidad es crear gráficas de dispersión por cada par de variables y colocarlas en una matriz y así agregar estimaciones de densidades bivariadas a cada gráfica.

Matrices de dispersión para datos multivariados

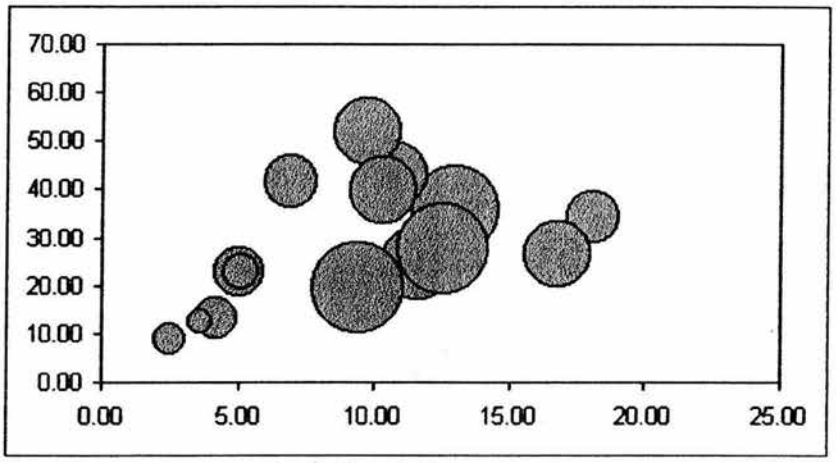
Una matriz de dispersión esta definida como una rejilla cuadrada y simétrica de gráficas de dispersión bivariadas. Esta rejilla tiene p filas y p columnas, cada una correponde a cada una de las p variables. Cada una de las entradas de la matriz muestra una gráfica de dispersión entre dos variables. Nótese que la matrices de dispersión son simétricas con respecto a su diagonal, puesto que la variable j es graficada contra la variable i en el espacio (i, j), y las mismas variables aparecen en el lugar (j, i) con los ejes intercambiados.



Matriz de dispersion.

Gráficas de Burbujas(Gower)

Este método consiste en acomodar radios de varios tamaños, a saber la medida de la tercera variable, para formar círculos cuyos centros son los puntos formados al graficar la primera y segunda variable en una gráfica de dispersión.



Burbujas de Gower.

3.1.3. Diagramas Simbólicos

Cuando sólo hay un número moderado de observaciones, cada una de éstas puede ser representada por algún tipo de símbolo o ícono, y la forma de este símbolo está representada por los valores de las variables de cada objeto.

Diagramas de Soles y Estrellas

Esta técnica consiste en graficar individualmente a los objetos. Dadas p variables, cada diagrama va a tener p rayos de tal forma que para cualesquiera dos rayos consecutivos, el espacio entre éstos sea el mismo. Así el primer paso consiste en dividir un círculo completo (o medio círculo) en p partes iguales, obteniendo el ángulo entre el j-ésimo y el $j+1$-rayo como θ. A partir de ésto se tiene que el ángulo entre un rayo dado y el j-ésimo rayo es

$$\text{círculo completo} \qquad \theta_j = 2\pi(j-1)/p \qquad j = 1, ..., p$$

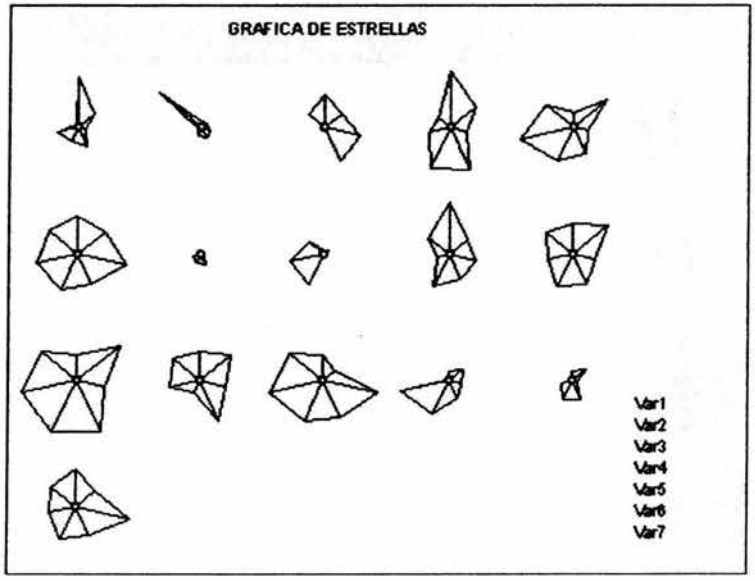
y

$$\text{medio círculo} \qquad \theta_j = \pi(j-1)/p \qquad j = 1, ..., p$$

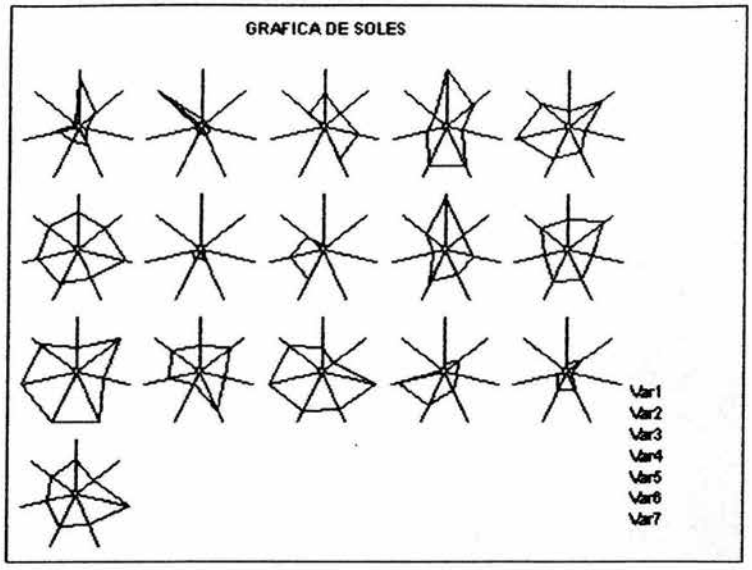
El segundo paso es localizar el centro en el plano cartesiano y estandarizar las variables al intervalo $[0, 1]$. Así el punto correspondiente a la j-ésima variable del i-ésimo objeto $(X_{i,j})$ tiene coordenadas polares

$$P_{i,j} = X_{i,j}\,(cos\theta_j, sen\theta_j)\,,$$

El Diagrama de Soles es muy parecido al Diagrama de Estrellas con la diferencia que el primero toma rayos de igual magnitud. Observesen las siguientes figuras:



Grafica de estrellas.



Grafica de soles.

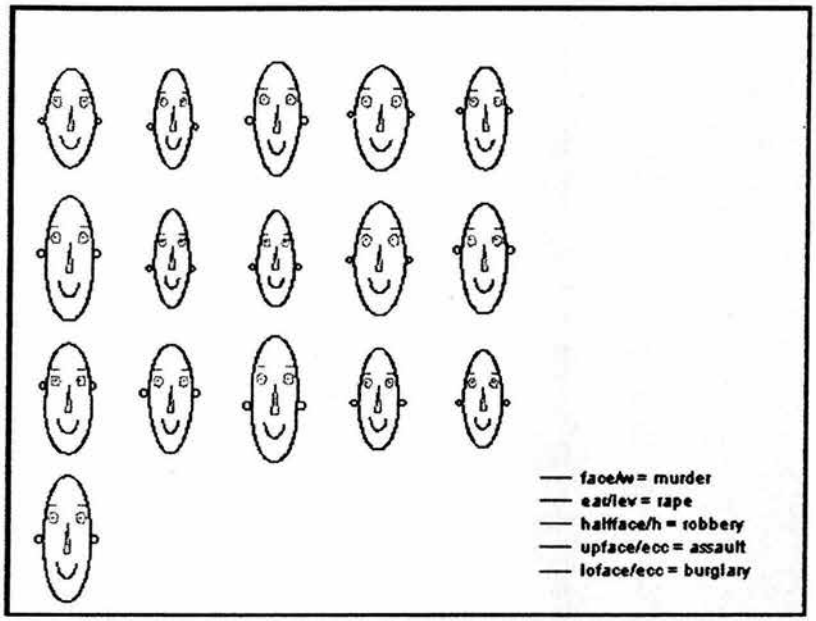
Caras de Chernoff

Ésta técnica fue desarrollada por Herman Chernoff (1971) y con la cual se pueden gráficar hasta 18 variables. Cada observación esta representada por un dibujo de una cara. La forma de la boca, la posición de los ojos, o de la nariz, etc., dependen de los valores de las variables. Las características que asemejan o distinguen a una persona de otra, son las que son afectadas por las

18 variables. Las características faciales que utiliza se listan a continuación:

NÚMERO	CARACTERÍSTICA FACIAL
1	Amplitud de la cara
2	Longuitud de la ceja
3	Altura de la cara
4	Separación de los ojos
5	Posición de las pupilas
6	Longuitud de la nariz
7	Ancho de la nariz
8	Diametro de la orejas
9	Nivel de las orejas
10	Longitud de la boca
11	Inclinación de los ojos
12	Curvatura de la boca
13	Nivel de la boca
14	Nivel de los ojos
15	Altura de las cejas

Originalmente, Chernoff propuso caras simétricas pero después Flurry duplicó la cantidad de variables dejando la simetría. Así, con esta técnica es posible graficar a lo más 36 variables.



Caras de Chernoff.

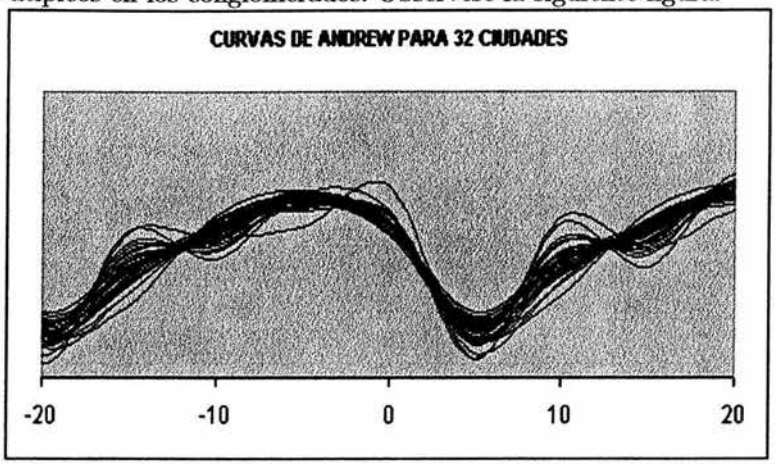
Curvas de Andrew

Andrew en 1972 propuso el siguiente método para graficar datos multivariados en dos dimensiones. Cada punto p-dimensional $\underline{X}^T = (X_1, ..., X_p)$, donde X_i, $i = 1, ..., p$ son los valores de las variables, está representada por la función

$$f_X(t) = \frac{1}{\sqrt{2}} X_1 + X_2 \sin(t) + X_3 \cos(t) + X_4 \sin(2t) + X_5 \cos(2t) + \cdots$$

sobre el rango $-\pi < t < \pi$. De esta manera, las n observaciones aparecerán como un conjunto de líneas dibujadas a través de la gráfica. Esta representación preserva la distancia Euclidiana, en el sentido de que dos observaciones con valores de las variables muy similares estarán representadas por curvas que permanezcan juntas para todos los valores de t. Por otro lado, dos observaciones cuyos valores de las variables que difieran notablemente, estarán representadas por curvas lejanas, al menos para algunos valores de t.

Por consiguiente las curvas de Andrew pueden permitir la identificación de datos atípicos en los conglomerados. Obsérvese la siguiente figura:



Muchas curvas de Andrew se pueden calcular con los mismos datos, simplemente al permutar las variables y recalculando las funciones $f_X(t)$. Así, es mejor asociar aquellas variables que se piense que tienen la más importante clasificación.

3.2. Técnicas Multivariadas

Como ayuda extra a las gráficas descritas en las secciones previas, también se puede hacer uso de algunas técnicas multivariadas para conseguir una idea acerca del número de conglomerados involucrados en la muestra. Entre estas técnicas se mencionan las siguientes:

3.2.1. Análisis de Componentes Principales

El Análisis de Componentes Principales es un método para transformar las variables en un conjunto nuevo de variables que son no-correlacionadas. Cada variable esta definida como una combinación lineal particular de las variables originales.

Definición 3.1 *Sea $\underline{X}$ un vector aleatorio p-dimensional y sea $\underline{L}$ un vector p-dimensional de constantes l_i's. Una combinación lineal estandarizada* (CLE)

de $\underline{X}$ *es una combinación lineal* $CLE = l_1X_1 + l_2X_2 + \cdots + l_pX_p$ *con la propiedad* $\sum_{i=1}^{p} l_i^2 = 1$.

Componentes Principales Poblacionales

Definición 3.2 *Sea* $\underline{X} \in \mathbb{R}^p$ *un vector aleatorio con media* $\underline{\mu}$ *y matriz de covarianzas* $\mathbf{\Sigma}$, *entonces la transformación de los componentes principales está dada por*

$$\underline{X} \to \underline{Y} = \mathbf{\Gamma}^T(\underline{X} - \underline{\mu})$$

donde $\mathbf{\Gamma}$ *es ortogonal,* $\mathbf{\Gamma}^T\mathbf{\Sigma}\mathbf{\Gamma} = \mathbf{\Lambda}$ *es diagonal y* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. *Las igualdades son estrictas si* $\mathbf{\Sigma}$ *es una matriz definida positiva. El i-ésimo componente principal de* $\underline{X}$ *es*

$$y_i = \gamma_{(i)}^T(\underline{X} - \underline{\mu})$$

donde $\gamma_{(i)}$ *es la i-ésima columna de* $\mathbf{\Gamma}$ *y puede ser llamado el i-ésimo vector de cargas de los componentres principales. La función* y_p *es llamada la última componente principal de* $\underline{X}$.

Teorema 3.1 *Sea* $\underline{X} \in \mathbb{R}^p$ *un vector aleatorio con media* $\underline{\mu}$ *y varianza* $\mathbf{\Sigma}$ *y* $\underline{Y}$ *la transformación de los componentes principales, entonces*

a) $\mathbf{E}(Y_i) = 0$;
b) $V(Y_i) = \lambda_i$;
c) $C(Y_i, Y_j) = 0$;
d) $V(Y_1) \geq V(Y_2) \geq \cdots \geq V(Y_p) \geq 0$;
e) $\sum_{i=1}^{p} V(Y_i) = tr\ \mathbf{\Sigma}$;
f) $\prod_{i=1}^{p} V(Y_i) = |\mathbf{\Sigma}|$

Demostración. a)

$$\begin{aligned}
\mathbf{E}(Y_i) &= \mathbf{E}\left(\boldsymbol{\gamma}_{(i)}^T(\underline{X}-\underline{\mu})\right) \\
&= \mathbf{E}\left(\boldsymbol{\gamma}_{(i)}^T\underline{X}-\boldsymbol{\gamma}_{(i)}^T\underline{\mu}\right) \\
&= \mathbf{E}\left(\boldsymbol{\gamma}_{(i)}^T\underline{X}\right)-\mathbf{E}\left(\boldsymbol{\gamma}_{(i)}^T\underline{\mu}\right) \\
&= \boldsymbol{\gamma}_{(i)}^T\mathbf{E}(\underline{X})-\boldsymbol{\gamma}_{(i)}^T\mathbf{E}(\underline{\mu}) \\
&= \boldsymbol{\gamma}_{(i)}^T\underline{\mu}-\boldsymbol{\gamma}_{(i)}^T\underline{\mu} \\
&= 0
\end{aligned}$$

b) Por definición $\underline{Y}=\boldsymbol{\Gamma}^T(\underline{X}-\underline{\mu})$ y por el terorema de la descomposición espectral se tiene que $\sum=\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T$, es decir $\boldsymbol{\Lambda}=\boldsymbol{\Gamma}^T\sum\boldsymbol{\Gamma}$ entonces

$$\begin{aligned}
V(\underline{Y}) &= \boldsymbol{\Gamma}^T V(\underline{X}-\underline{\mu})\boldsymbol{\Gamma} \\
&= \boldsymbol{\Gamma}^T V(\underline{X})\boldsymbol{\Gamma} \\
&= \boldsymbol{\Gamma}^T\sum\boldsymbol{\Gamma} \\
&= \boldsymbol{\Lambda}
\end{aligned}$$

por lo tanto $V(Y_i)=\lambda_i$.

c)

$$\begin{aligned}
C(\underline{Y},\underline{Y}) &= \mathbf{E}\left[\underline{Y}\underline{Y}^T\right] \\
&= \mathbf{E}\left[\boldsymbol{\Gamma}^T(\underline{X}-\underline{\mu})\left(\boldsymbol{\Gamma}^T(\underline{X}-\underline{\mu})\right)^T\right] \\
&= \mathbf{E}\left[\boldsymbol{\Gamma}^T(\underline{X}-\underline{\mu})(\underline{X}-\underline{\mu})^T\boldsymbol{\Gamma}\right] \\
&= \boldsymbol{\Gamma}^T\mathbf{E}\left[(\underline{X}-\underline{\mu})(\underline{X}-\underline{\mu})^T\right]\boldsymbol{\Gamma} \\
&= \boldsymbol{\Gamma}^T V(\underline{X})\boldsymbol{\Gamma} \\
&= \boldsymbol{\Gamma}^T\sum\boldsymbol{\Gamma} \\
&= \boldsymbol{\Lambda}
\end{aligned}$$

de esta manera si $i\neq j$ se tiene que $C(Y_i,Y_j)=0$.

d) De b) se sabe que $V(Y_i)=\lambda_i$, $1\leq i\leq p$ y como $\lambda_1\geq\ldots\geq\lambda_p$ son los valores propios de $\sum\geq 0$ entonces se tiene además que $\lambda_p\geq 0$. Por tanto, se concluye que $V(Y_1)\geq\ldots\geq V(Y_p)\geq 0$.

e) $\sum_{i=1}^{p}V(Y_i)=\sum_{i=1}^{p}\lambda_i=tr\ \boldsymbol{\Sigma}$ donde la primera igualdad se cumple por b) y la segunda por el corolario (2.7).

f) $\prod_{i=1}^{p}V(Y_i)=\prod_{i=1}^{p}\lambda_i=|\boldsymbol{\Sigma}|$ donde la primera igualdad se cumple por b) y la segunda por (2.13). ■

Teorema 3.2 *La varianza del primer componente principal Y_1 tiene una varianza más grande que la varianza de cualquier combinación lineal estandarizada (CLE) de $\underline{X}$, esto es $V(Y_1) \geq V(CLE)$ para toda CLE.*

Demostración. Sea $\underline{A}^T\underline{X}$ una combinación lineal estandarizada (CLE), donde $\underline{A}^T\underline{A} = 1$. Como $\gamma_{(1)}, \gamma_{(2)}, ..., \gamma_{(p)}$ son los vectores propios de $\mathbf{\Sigma}$, entonces $\{\gamma_{(1)}, \gamma_{(2)}, ..., \gamma_{(p)}\}$ constituye una base para $\mathbb{R}^p$, por lo que es posible escribir

$$\underline{A} = c_1\gamma_{(1)} + ... + c_p\gamma_{(p)} \tag{3.1}$$

Sea $\alpha = \underline{A}^T\underline{X}$, entonces

$$V(\alpha) = \underline{A}^T\mathbf{\Sigma}\underline{A} = \underline{A}^T\left(\sum_{i=1}^{p}\lambda_i\gamma_{(i)}\gamma_{(i)}^T\right)\underline{A} \tag{3.2}$$

(la primera igualdad por las propiedades de la varianza y la segunda por el Teorema (2.6)). Ahora $\gamma_{(i)}^T\gamma_{(j)} = \delta_{i,j}$ o mejor conocida como la delta de Kronecker. Al sustituir (3.1) en (3.2), se obtiene

$$V(\alpha) = \sum \lambda_i c_i^2 \tag{3.3}$$

y como $\underline{A}^T\underline{A} = 1$ y por (3.1) se tiene que

$$\sum c_i^2 = 1. \tag{3.4}$$

También se sabe que $V(Y_1) = \lambda_1$ es el más grande de los valores propios, el máximo de (3.3) sujeto a (3.4) es $V(Y_1) = \lambda_1$.

Así el máximo se obtiene cuando $c_1 = 1$ y $c_2 = ... = c_p = 0$. Por consiguiente, $V(\alpha)$ es maximizada cuando $\underline{A} = \gamma_{(1)}$. ■

Componentes Principales Muestrales

Sea $\mathbf{X} = (\underline{X}_1, ..., \underline{X}_n)^T$ una matriz muestral de datos, y considérese $\mathbf{S}$ la matriz de covarianzas muestrales de $\mathbf{X}$. Por el teorema de la descomposición espectral $\mathbf{S} = \mathbf{GLG}^T$ donde $\mathbf{L}$ es una matriz diagonal. Por analogía de la definición de los componentes principales poblacionales es posible definir a los componentes principales muestrales como

$$\underline{Y}_{(1)} = \left(\mathbf{X} - \mathbf{1}\underline{X}^T\right)\mathbf{g}_{(1)}$$

donde $\mathbf{g}_{(1)}$ es el vector estandarizado correspondiente al mayor valor propio de $\mathbf{S}$. Similarmente el i-ésimo componente principal esta definido como $\underline{Y}_{(i)} = \left(\mathbf{X} - \mathbf{1}\underline{X}^T\right)\mathbf{g}_{(i)}$, o al juntar todos los componentes, se obtiene

$$\mathbf{Y} = \left(\mathbf{X} - \mathbf{1}\underline{X}^T\right)\mathbf{G}$$

y su matriz de covarianzas, $\mathbf{S_Y}$, es

$$\begin{aligned}
\mathbf{S_Y} &= n^{-1}\mathbf{Y}^T\mathbf{H}\mathbf{Y} \\
&= n^{-1}\mathbf{G}^T\left(\mathbf{X} - \mathbf{1}\underline{X}^T\right)^T\mathbf{H}\left(\mathbf{X} - \mathbf{1}\underline{X}^T\right)\mathbf{G} \\
&= n^{-1}\mathbf{G}^T\mathbf{H}\mathbf{G} \\
&= \mathbf{G}^T\mathbf{S}\mathbf{G} \\
&= \mathbf{L}
\end{aligned}$$

donde

$$\mathbf{H} = \mathbf{I} - \left(n^{-1}\mathbf{1}\mathbf{1}^T\right);$$

en otras palabras, las columnas de $\mathbf{Y}$ son no-correlacionadas y la varianza de $\underline{Y}_{(j)}$ es l_j.

El r-ésimo elemento de $\underline{Y}_{(i)}$ es y_{ri} representa el valor del i-ésimo componente principal del r-ésimo individuo, esto es

$$y_{ri} = \mathbf{g}_{(i)}^T\left(\underline{X}_r - \overline{X}\right).$$

Observación 3.1 *La variación total es* $tr\,\mathbf{\Sigma} = \sum_{i=1}^{p}\lambda_i = \sum_{i=1}^{p}V\left(\underline{Y}_{(i)}\right)$ *por lo que se tiene que* $\left(\lambda_1, ..., \lambda_k\right)/\left(\lambda_1, ..., \lambda_p\right)$ *representa la proporción de la variación total "explicada" por los primeros k componentes principales.*

Observación 3.2 *Si la matriz de covarianzas de* $\underline{X}$ *tiene rango* $r < p$, *entonces la variación total es "explicada totalmente" por los primeros r componentes principales. Esto se debe al hecho que si* $\mathbf{\Sigma}$ *tiene rango r, entonces los últimos* $(p - r)$ *valores propios son idénticamente cero, así por lo anterior la variación total es explicada totamente.*

3.2.2. Escalamiento Multidimensional

El escalamiento multidimensional esta involucrado con el problema de construir una configuración de n puntos en el espacio Euclidiano usando información de una matriz de distancias entre n objetos. Las distancias entre

estos n puntos, esto es, los puntos de la configuración, pueden estar sujetos a un error. Por ejemplo considérese la siguiente tabla

	1	2	3	4	5	6	7	8	9	10	11	12
1												
2	244											
3	218	350										
4	284	77	369									
5	197	167	347	242								
6	312	444	94	463	441							
7	215	221	150	236	279	245						
8	469	583	251	598	598	169	380					
9	166	242	116	257	269	210	55	349				
10	212	53	298	72	170	392	168	531	190			
11	253	325	57	340	359	143	117	264	91	273		
12	270	168	284	164	277	378	143	514	173	111	256	

donde se aprecian las distancias de los caminos que comunican a 12 ciudades, y el objetivo es construir un mapa geográfico de las ciudades basado en esta información. Sin embargo las distancias no necesariamente son distancias Euclidianas, y pueden representar muchos tipos de disimilaridades entre objetos. También en algunos casos, se comienza con similaridades entre objetos en vez de disimilaridades o funciones distancia.

Solución Clásica

Definición 3.3 *Una matriz* $\mathbf{D}$ *es llamada Euclidiana si existe una configuración de puntos en algun espacio Euclidiano cuyas distancias entre los puntos que conforman la configuración estan dadas por* $\mathbf{D}$; *esto es, si para alguna* p, *existen* p *puntos, a saber,* $\underline{X}_1, ..., \underline{X}_p \in \mathbb{R}^p$ *tales que*

$$d_{r,s}^2 = (\underline{X}_r - \underline{X}_s)^t (\underline{X}_r - \underline{X}_s) . \tag{3.5}$$

El siguiente teorema establece cuando $\mathbf{D}$ es Euclidiana, y si lo es, como encontrar una configuración de puntos. Para este propósito sea $\mathbf{D}$ una matriz de distancias, y sea

$$\mathbf{A} = (A_{r,s}) , \qquad A_{r,s} = -\frac{1}{2} d_{r,s}^2 \tag{3.6}$$

y sea

$$\mathbf{B} = \mathbf{HAH} \tag{3.7}$$

donde $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T$ es una matriz $n \times n$ y es conocida como la matriz de centralidad.

Teorema 3.3 *Sea* $\mathbf{D}$ *una matriz de distancias y sea* $\mathbf{B}$ *definida como en (3.7), entonces* $\mathbf{D}$ *es Euclidiana si y sólo si* $\mathbf{B}$ *es una matriz definida positiva. En particular, se tiene que:*
a) Si $\mathbf{D}$ *es la matriz de distancias Euclidianas de los puntos de una configuración* $\mathbf{Z} = (\underline{Z}_1, ..., \underline{Z}_n)^T$, *entonces*

$$B_{r,s} = (\underline{Z}_r - \overline{Z})^T(\underline{Z}_s - \overline{Z}), \qquad r, s = 1, ..., n \tag{3.8}$$

La forma matricial de (3.8) es $\mathbf{B} = (\mathbf{HZ})(\mathbf{HZ})^T$, *así* $\mathbf{B} \geq 0$. *Nótese que* $\mathbf{B}$ *puede ser interpretada como la matriz centralizada de productos punto para la configuración* $\mathbf{Z}$.
b) Recíprocamente, si $\mathbf{B}$ *es una matriz definida positiva de rango* p *entonces una configuración correspondiente a* $\mathbf{B}$ *puede ser construida como sigue. Sean* $\lambda_1 \geq ... \geq \lambda_p$ *los valores propios positivos de* $\mathbf{B}$ *con sus correspondientes vectores propios* $\mathbf{X} = (\underline{X}_{(1)}, ..., \underline{X}_{(p)})$ *normalizados por*

$$\underline{X}_{(i)}^T \underline{X}_{(i)} = \lambda_i \qquad i = 1, ..., p. \tag{3.9}$$

entonces los puntos $\underline{P}_r \in \mathbb{R}^p$ *con coordenadas* $X_r = (X_{r,1}, ..., X_{r,p})^T$ *(o sea,* X_r *es la* r*-ésima fila de* $\mathbf{X}$*) tiene las distancias dadas por* $\mathbf{D}$. *Más aún, esta configuración tiene centro de gravedad* $\underline{X} = \underline{0}$, *y* $\mathbf{B}$ *representa la matriz centralizada de productos punto para esta configuración.*

Demostración. *a)* Por hipótesis

$$d_{r,s}^2 = -2A_{r,s} = (\underline{Z}_r - \underline{Z}_s)^T(\underline{Z}_r - \underline{Z}_s) \tag{3.10}$$

ahora

$$\begin{aligned} \mathbf{B} &= \mathbf{HAH} \\ &= (\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T)\mathbf{A}(\mathbf{I} - \mathbf{n}^{-1}\mathbf{1}\mathbf{1}^\mathbf{T}) \\ &= \mathbf{A} - n^{-1}\mathbf{AJ} - n^{-1}\mathbf{JA} + n^{-2}\mathbf{JAJ} \end{aligned}$$

donde $\mathbf{J} = \mathbf{1}\mathbf{1}^T$. Obsérvese que

$$\frac{1}{n}\mathbf{AJ} = \begin{bmatrix} \overline{A}_{1,\cdot} & \cdots & \overline{A}_{1,\cdot} \\ \vdots & & \vdots \\ \overline{A}_{n,\cdot} & \cdots & \overline{A}_{n,\cdot} \end{bmatrix}, \quad \frac{1}{n}\mathbf{JA} = \begin{bmatrix} \overline{A}_{\cdot,1} & \cdots & \overline{A}_{\cdot,1} \\ \vdots & & \vdots \\ \overline{A}_{\cdot,n} & \cdots & \overline{A}_{\cdot,n} \end{bmatrix},$$

$$\frac{1}{n^2}\mathbf{JAJ} = \begin{bmatrix} \overline{A}_{\cdot,\cdot} & \cdots & \overline{A}_{\cdot,\cdot} \\ \vdots & & \vdots \\ \overline{A}_{\cdot,\cdot} & \cdots & \overline{A}_{\cdot,\cdot} \end{bmatrix}$$

donde

$$\overline{A}_{r,\cdot} = \frac{1}{n}\sum_{s=1}^{n} A_{r,s}, \quad \overline{A}_{\cdot,s} = \frac{1}{n}\sum_{r=1}^{n} A_{r,s}, \quad \overline{A}_{\cdot,\cdot} = \frac{1}{n^2}\sum_{r,s=1}^{n} A_{r,s} \tag{3.11}$$

Así

$$B_{r,s} = A_{r,s} - \overline{A}_{r,\cdot} - \overline{A}_{\cdot,s} + \overline{A}_{\cdot,\cdot} \tag{3.12}$$

Sustituyendo (3.10) en (3.12), y simplificando se obtiene que

$$B_{r,s} = (\underline{Z}_r - \overline{Z})^T(\underline{Z}_s - \overline{Z})$$

así se termina la prueba de la parte *a)*.

b) Por hipótesis $\mathbf{B} \geq 0$ y considérese la configuración dada en el teorema. Sea $\mathbf{\Lambda} = diag(\lambda_1, ..., \lambda_p)$ y sea $\mathbf{\Gamma} = \mathbf{X\Lambda}^{-1/2}$, tal que las columnas de $\mathbf{\Gamma}$, $\boldsymbol{\gamma}_{(i)} = \lambda_i^{-1/2}\underline{X}_{(i)}$ los vectores estandarizados de $\mathbf{B}$, entonces por el Teorema (2.6) se tiene que

$$\mathbf{B} = \mathbf{\Gamma\Lambda\Gamma}^T = \mathbf{XX}^T;$$

esto es,

$$B_{r,s} = \underline{X}_r^T\underline{X}_s,$$

así $\mathbf{B}$ representa la matriz de productos punto para esta configuración.

Falta demostrar que $\mathbf{D}$ representa la matriz de distancias para esta configuración. Usando (3.12) para escribir a $\mathbf{B}$ en términos de $\mathbf{A}$ se obtiene

$$\begin{aligned} (\underline{X}_r - \underline{X}_s)^T(\underline{X}_r - \underline{X}_s) &= \underline{X}_r^T\underline{X}_r - 2\underline{X}_r\underline{X}_s + \underline{X}_s^T\underline{X}_s \\ &= B_{r,r} - 2B_{r,s} + B_{s,s} \\ &= A_{r,r} - 2A_{r,s} + A_{s,s} \\ &= -2A_{r,s} \\ &= d_{r,s}^2 \end{aligned} \tag{3.13}$$

porque $A_{r,r} = -\frac{1}{2}d_{r,r}^2 = 0$ y $-2A_{r,s} = d_{r,s}^2$.

Nótese que $\mathbf{B}\underline{1} = \mathbf{HAH}\underline{1} = \underline{0}$, así $\underline{1}$ es un vector propio de $\mathbf{B}$ correspondiente al valor propio $\underline{0}$. Así $\underline{1}$ es ortogonal a las columnas de $\mathbf{X}$, $\underline{X}_{(i)}^T\underline{1} = \underline{0}$, $i = 1, ..., p$. De aquí

$$n\overline{X} = \sum_{r=1}^{n}\underline{X}_i = \mathbf{X}^T\underline{1} = (\underline{X}_{(1)}^T\underline{1}, ..., \underline{X}_{(p)}^T\underline{1})^T = \underline{0}$$

así, el centro de gravedad de esta configuración es el origen. ■

Observación 3.3 *La matriz* **X** *puede visualizarse en términos de los vectores propios de* **B** *y los puntos correspondientes:*

$$\begin{array}{cccccc} & & \textit{Valores propios} & & & \\ & \lambda_1 & \lambda_2 & \cdots & \lambda_p & \textit{notacion vector} \\ \mathcal{P}_1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p} & \\ \mathcal{P}_2 & X_{2,1} & X_{2,2} & \cdots & X_{2,p} & \underline{X}_1^T \\ \vdots & \vdots & \vdots & & \vdots & \underline{X}_2^T \\ \mathcal{P}_n & X_{n,1} & X_{n,2} & \cdots & X_{n,p} & \vdots \\ & & & & & \underline{X}_n^T \\ & \underline{X}_{(1)} & \underline{X}_{(2)} & \cdots & \underline{X}_{(p)} & \\ & \textit{notacion vector} & & & & \end{array}$$

Centro de gravedad:

$$\underline{X}_1 = 0, \underline{X}_2 = 0, ..., \underline{X}_p = 0, \quad \overline{X} = \frac{1}{n}\sum \underline{X}_r = 0$$

En resumen, la r-ésima fila de* X *contiene las coordenadas del r-ésimo punto, mientras que la i-ésima columna de* X *contiene el vector propio correspondiente al valor propio λ_i.

Observación 3.4 *Geométricamente, si* **B** *es la matriz centrada de productos puntos para una configuración Z, entonces $B_{r,s}^{1/2}$ es igual a la distancia entre $\underline{Z}_r$ y $\underline{Z}_s$, y $B_{r,s}/(B_{r,r}B_{s,s})^{1/2}$ es igual al coseno del ángulo formado por los vectores $\underline{Z}_r$ y $\underline{Z}_s$ que parten de un punto común, a saber $\overline{Z}$.*

Observación 3.5 *Nótese que $\underline{1}$ es un vector propio de* **B** *no importanto si* **D** *es Euclidiana o no.*

Similaridades

En muchas situaciones se comienza con una matriz, **C**,de similaridades entre n objetos. Esta matriz $\mathbf{C} \in \mathcal{M}^{n\times n}$ satisface, por definición, $C_{r,s} = C_{s,r}$ y $C_{r,s} \leq C_{r,r}$ para toda r, s.

Definición 3.4 *La transformación estándar de una matriz de similaridad* **C** *a una matriz distancia* **D** *está definida por*

$$d_{r,s} = (C_{r,r} - 2C_{r,s} + C_{s,s})^{1/2}. \tag{3.14}$$

Teorema 3.4 *Si* $\mathbf{C} \geq 0$, *entonces la matriz de distancias* **D** *definida por (3.14) es Euclidiana, con matriz centralizada de productos punto* $\mathbf{B} = \mathbf{HCH}$.

Demostración. Como $\mathbf{C} \geq 0$,

$$d_{r,s}^2 = C_{r,r} - 2C_{r,s} + C_{s,s} = \underline{X}^T \mathbf{C} \underline{X} \geq 0,$$

donde $\underline{X}$ es un vector con $+1$ en la r-ésima entrada y -1 en la s-ésima entrada, para $r \neq s$. Así, la transformación estándar esta bien definida y $\mathbf{D}$ es una matriz distancia.

Sean $\mathbf{A}$ y $\mathbf{B}$ como en (3.6) y (3.7), respectivamente. Obsérvese que $\mathbf{HCH}$ también es una matriz definida positiva, por tanto basta demostrar que $\mathbf{B} = \mathbf{HCH}$ y así concluir que $\mathbf{D}$ es Euclidiana con matriz centralizada de productos punto dada por $\mathbf{HCH}$.

Ahora, usando la ecuación (3.12) y sustituyendo $A_{r,s} = -\frac{1}{2}d_{r,s}^2$ se obtiene

$$\begin{aligned} -2B_{r,s} &= d_{r,s}^2 - \tfrac{1}{n}\sum_{i=1}^{n} d_{r,i}^2 - \tfrac{1}{n}\sum_{j=1}^{n} d_{j,s}^2 + \tfrac{1}{n^2}\sum_{i,j=1}^{n} d_{i,j}^2 \\ &= C_{r,r} - 2C_{r,s} + C_{s,s} - \tfrac{1}{n}\sum_{i=1}^{n}(C_{r,r} - 2C_{r,i} + C_{i,i}) \\ &\quad -\tfrac{1}{n}\sum_{j=1}^{n}(C_{j,j} - 2C_{j,s} + C_{s,s}) + \tfrac{1}{n}\sum_{i,j=1}^{n}(C_{i,i} - 2C_{i,j} + C_{j,j}) \\ &= -2C_{r,s} + 2\overline{C}_{r,\cdot} + 2\overline{C}_{\cdot,s} - 2C_{\cdot,\cdot} \end{aligned}$$

así

$$B_{r,s} = C_{r,s} - \overline{C}_{r,\cdot} - \overline{C}_{\cdot,s} - C_{\cdot,\cdot}$$

o escrito en su forma matricial, $\mathbf{B} = \mathbf{HCH}$. ■

Supóngase que $\lambda_k > 0$ donde λ_k son los primeros k valores propios de la matriz $\mathbf{B}$. Dos posibles medidas para la proporción de una matriz de distancias $\mathbf{D}$ explicada por la solución clásica de k dimensiones de escalamiento multidimensional serían;

$$\alpha_{1,k} = \left(\sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{p} |\lambda_i| \right) \times 100\,\%, \tag{3.15}$$

y

$$\alpha_{2,k} = \left(\sum_{i=1}^{k} \lambda_i^2 \Big/ \sum_{i=1}^{p} \lambda_i^2 \right) \times 100\,\% \tag{3.16}$$

En la ecuación (3.15) se necesitan valores absolutos, puesto que los últimos valores propios pueden ser negativos.

Capítulo 4

Análisis de conglomerados

4.1. Introducción

El análisis de conglomerados es el nombre para un grupo de técnicas multivariadas cuyo propósito principal es el de agrupar objetos de acuerdo a las características que poseen. El análisis de conglomerados clasifica objetos tales que sean muy "similares" a los demás en el mismo conglomerado respecto a un criterio determinado. Los conglomerados resultantes deben presentar homogeneidad dentro de ellos y heterogeneidad fuera de ellos, en otras palabras, si la clasificación fue exitosa, al graficar los objetos en los conglomerados éstos deben estar "cercanos" y los de distintos conglomerados deben estar separados.

El análisis de conglomerados también es conocido como análisis Q, análisis de clasificación y taxonomía numérica. Esta variedad de nombres está ligada a su uso en diversas disciplinas como psicología, biología, sociología, economía, ingenería y negocios.

El análisis de conglomerados es una herramienta muy útil para el análisis de datos en muchas situaciones diferentes. De hecho, al menos se conocen tres campos de aplicación del análisis de conglomerados: 1) aplicaciones en la ciencia retro-deductiva y en la ciencia hipotético-deductiva, 2) aplicaciones en clasificaciones, y 3) aplicaciones en planeación e ingeniería. A continuación se detallan las tres categorías antes mencionadas:

4.1.1. Ciencias retro-deductivas e hipotético-deductivas

Creación de una pregunta

Aquí se recolecta una matriz de datos y éstos se exploran por medio de un análisis de conglomerados con la esperanza que este análisis revele patrones interesantes de similaridad para asi poder crear una pregunta relativa a la naturaleza de los patrones de similaridad.

Creación de una hipótesis

Si se tiene una buena pregunta en mente antes de realizar un análisis de conglomerados, esta pregunta sirve para recolectar algún tipo de datos correctamente y al realizar un análisis de conglomerados y estudiar los patrones de similaridad en el dendograma (ver apéndice A), se espera que los patrones sugieran una hipótesis para investigar. Cuando esto ocurra, se tendrá una tentativa respuesta a la pregunta.

Este uso del análisis de conglomerados es un ejemplo del método científico llamado *retroductivo*.

Probar una hipótesis

Si ya se tienen una pregunta y una hipótesis en mente, seguramente se obtuvieron sin la ayuda de un análisis de conglomerados, pero se quiere probarlas mediante la ayuda de dicha técnica y de esta forma, si la hipótesis es cierta, entonces el dendograma tendrá ciertas propiedades, es decir, si después de aplicar los patrones de similaridad predichos concuerdan en un grado aceptable de tolerancia con los actuales patrones, se dice que la hipótesis ha sido confirmada, en otro caso se rechaza.

Este uso del análisis de conglomerados es un ejemplo del método científico llamado *método hipotético-reductivo*.

4.1.2. Clasificaciones

Clasificaciones generales

Este tipo de clasificaciones sirve para propósitos generales científicos, como un catálogo básico de las entidades que estudian los científicos o bien

como la base para incorporar nuevas teorías. Se distinguen dos tipos de clasificaciones generales: 1) las biológicas (por ejemplo, clasificación de especies) y, 2) las no-biológicas.(por ejemplo, clasificación del tipo de suelos).

Clasificaciones específicas

Este tipo de clasificaciones difiere de las anteriores en que las clasificaciones específicas son utilizadas para una clasificación que ya se tiene en mente, por ejemplo un análisis de conglomerados para crear una clasificación de pacientes con enfermades del hígado por medio de algun estudio químico.

4.1.3. Planeación e ingenieria

Debido a que los cientificos están motivados de una manera distinta que los otros profesionistas, es necesario crear una categoría nueva, planeación e ingeniería. Los cientfficos estan motivados por la curiosidad de descubrir como funciona la naturaleza, en vez de aplicar su conocimiento para beneficio de la sociedad, y en cuanto a la planeación su objetivo primordial es el crear un mundo materialmente mejor.

Las decisiones o planes pueden ser usados como objetos de un análisis de conglomerados. Los atributos pueden describir rasgos de las alternativas o representar sus gastos esperados, de esta manera, al encontrar conglomerados de las alternativas que son similares entre si, el problema de desición se reduce a dos etapas: 1) Seleccionar el conglomerado que obtenga el mejor resultado para alcanzar el objetivo de planeación, y 2) Seleccionar la mejor alternativa dentro del mejor conglomerado.

Ejemplos de aplicaciones

- Un investigador que ha recolectado datos por medio de un cuestionario se puede enfrentar a un gran número de observaciones que no tienen sentido a menos que sean clasificadas en grupos. El análisis por conglomerados puede realizar objetivamente este procedimiento reduciendo la información de la población total o captar información específica en grupos más chicos. Por ejemplo, si se quieren entender las actitudes de grandes grupos de una población por medio de otros más pequeños. De esta forma el investigador tiene una descripción de los objetos más concisa y entendible, con una mínima pérdida de información.

- Otra forma en la cual el análisis de conglomerados puede ser útil es cuando un investigador quiere desarrollar hipótesis concernientes a la naturaleza de los datos o examinar hipótesis previamente establecidas. Considerése el siguiente ejemplo, si un investigador desea saber las actitudes de la gente frente a los refrescos de dieta y los que no lo son, esta técnica puede ser utilizada para clasificar a los consumidores de refrescos por medio de sus actitudes y asi saber si prefieren los refrescos de dieta o no. De esta forma, los conglomerados resultantes mostrarán similitudes y diferencias entre los consumidores.

- En biología puede ser usada para clasificar todo tipo de seres vivientes, por ejemplo mamíferos, reptiles, etc.

- En psicología puede emplearse en clasificaciones basadas en personalidad y otras actitudes personales.

- En negocios puede ser utilizada para clasificar la estructura de los mercados o para evaluar diferencias y similitudes entre nuevos productos.

- En finanzas se puede usar un análisis de conglomerados para agrupar portafolios de inversión y, de esta forma, crear nuevos portafolios que tengan ciertas ventajas sobre otros.

4.2. Distancias y funciones de similaridad

Sea U un conjunto, no vacío, finito o infinito de elementos. Sea $\mathbb{R}$ el conjunto de los números reales y sean $\mathbb{R}^+$ y $\mathbb{R}^-$ el conjunto de los números reales positivos y negativos, respectivamente.

Definición 4.1 *Una función $d : U \times U \to \mathbb{R}$ (que asigna a cada par de elementos en U, un número real) se dice que es una distancia o disimilaridad, si para cada $x, y \in U$, se cumple lo siguiente:*
1) $d(x, y) \geq d_0$
2) $d(x, x) = d_0$
3) $d(x, y) = d(y, x)$
en donde d_0 es un número real arbitrario (incluso negativo).

Las relaciones entre las propiedades 1 y 2 significan que d se minimiza si los pares de elementos son iguales. La propiedad 3 expresa la propiedad escencial de simetría.

Definición 4.2 *Se dice que una distancia es métrica si, además de* 1, 2 *y* 3, *cumple las siguientes dos condiciones:*
4. Si $d(x,y) = d_0$ *entonces* $x = y$
5. $d(x,y) \leq d(x,y) + d(y,z), z \in U$
donde la condición 4 significa que el valor más pequeño de la función distancia siempre implica que los dos elementos son idénticos. La condición de la ecuación 5 corresponde a la desigualdad del triángulo de la geometría Euclidiana.

Si además d_0 es cero, se obtiene el concepto de métrica utilizado en análisis funcional.. Si d_0 es negativo, una métrica d' puede ser obtenida de cualquier función distancia que por definición es: $d'(x,y) = d(x,y) - d_0$.

Definición 4.3 *Sean* $A, B \subset U$ *dos subconjuntos no vacíos. Se define la distancia entre dos conjuntos como*
$d(A,B) = \inf_{y \in A} \{d(y,B)\} = \inf_{x \in B} \{d(A,x)\} = \inf_{y \in A, x \in B} \{d(y,x)\}$

Definición 4.4 *Una función de similaridad es una función* $s : U \times U \to R$ *con las siguientes propiedades:*
1. $s(x,y) \leq s_0$
2. $s(x,x) = s_0$
3.$s(x,y) = s(y,x)$
donde s_0 *es un número real. La distinción entre* d *y* s *radica en la primera propiedad en ambas definiciones.*

Definición 4.5 *Una función de similaridad es llamada métrica si, además de las condiciones 1, 2, y 3, satisface:*
4. Si $s(x,y) = s_0$ *entonces* $x = y$
5. $|s(x,y) + s(y,z)| . s(x,y) \geq s(x,y) . s(y,z)$
donde la condición 4 corresponde a la proposición de la máxima similaridad puede solamente ser alcanzada cuando son idénticos dos elementos, mientras la relación 5 corresponde es definida de manera análoga a la condición 5 de la definición 4.2.

Podría se observado que si d es una distancia entonces $\frac{1}{d}$ es una función de similaridad. Además, si d sólamente toma valores finitos, $\max d - d$, $\sqrt{\max d - d}$, y $\max d - d^2$ son funciones de similaridad métricas, y también $\exp(-d)$, pero esta última no requiere la condición de finitud de d.

De manera análoga, las aserciones con s reemplazadas para d también son válidas, por ejemplo, $\exp(-d)$ es reemplazada por -ln (s).

Si d es una función distancia métrica, entonces $\frac{Md}{1+d}$, donde M es una constante arbitraria positiva, también es una distancia métrica.

Observación 4.1 *Se dice que δ es un* coeficiente de semejanza *si δ es una función distancia o una función de similaridad. Así, el término semejanza involucra los dos conceptos, cuyo objetivo es usar un término general de una función que compara a dos objetos*

4.2.1. Perfiles de datos

Antes de discutir más profundamente a los coeficientes de semejanza, considerése a la matriz de datos dada por

$$
\begin{array}{lcccc|cc}
 & \multicolumn{4}{c|}{\textit{ATRIBUTOS EN EL TIEMPO}} & \overline{X}_j & S_j \\
1\ (\mathrm{B}) & 20 & 40 & 25 & 30 & 28.75 & 8.54 \\
2\ (\mathrm{B}{+}15) & 35 & 55 & 40 & 45 & 43.75 & 8.54 \\
3\ (\mathrm{B}{*}2) & 40 & 80 & 50 & 60 & 57.50 & 17.07 \\
4\ (\mathrm{TIE}) & 20 & 0 & 15 & 10 & 11.25 & 8.54 \\
 & & & & & & \\
\overline{X}_i & 28.75 & 43.75 & 32.5 & 36.25 & & \\
S_i & 10.31 & 33.51 & 15.55 & 21.36 & &
\end{array}
\qquad (4.1)
$$

donde los objetos representan cuatro tipos de acciones de cuatro compañías hipotéticas. Los atributos representan los precios de las acciones a través del tiempo. El objeto 1 es el caso base y esta denotado por B. El objeto 2 es 15 unidades mayor que B, denotado por B+15. El objeto 3 es dos veces mayor que B y, finalmente, el objeto 4 es la imagen de B bajo un espejo. La relación que existe entre el objeto 1 y 2 es conocida como "*traslación aditiva*". La relación entre 1 y 3 es conocida como "*traslación proporcional*". La de los objetos 1 y 4 se conoce como "*traslación de Imagen bajo un Espejo (TIE)*". Estos tipos de traslaciones hacen posible catalogar a los coeficientes de semejanza.

Rara vez los perfiles de datos son traslaciones aditivas o proporcionales. Sin embargo, en muchas aplicaciones puede haber una separación entre los perfiles, a la cual se le llama "*desplazamiento de tamaño*", y que en algunas veces se desea ignorar. Un dezplazamiento de tamaño ocurre cuando un perfil de un objeto es, atributo por atributo, mayor o menor que algún otro. Este

es un concepto muy útil si se desea medir la similaridad basada en la forma de los perfiles aparte de los desplazamientos entre ellos. Por ejemplo, si los objetos son huesos de fósiles y las atributos son las medidas de los huesos, probablemente se desea ignorar a los desplazamientos de los huesos, ya que los animales en consideración, probablemente, no eran de la misma edad.

4.2.2. Ejemplos

Distancias Euclidiana y de Mahalanobis

Sean $\underline{X}$ y $\underline{Y}$ dos vectores reales, a saber, $\underline{X} \doteq (X_1, \ldots, X_l)^T$ y $\underline{Y} = (Y_1, \ldots Y_l)^T$, que serán usados para representar dos objetos descritos como columnas en la matriz de datos.

La más conocida de todas las distancias es la que corresponde a la generalización a más de dos dimensiones de la distancia entre dos puntos en el plano, que es derivada de la norma $\mathbb{L}_2$, a saber

$$\|\underline{X}\|_2 = \sqrt{\sum_{k=1}^{l} X_k^2} = \sqrt{\underline{X}^T \underline{X}} \tag{4.2}$$

de donde se obtiene

$$d_2(\underline{X}, \underline{Y}) = \|\underline{X} - \underline{Y}\|_2 = \sqrt{(\underline{X} - \underline{Y})^T (\underline{X} - \underline{Y})} \tag{4.3}$$

La métrica Euclidiana tiene la propiedad de ser invariante, bajo todos sus valores, con respecto a cualquier mapeo ortogonal (rotaciones) de los vectores que son descritos por todas las matrices $Q \in M_{l \times l}$ tales que $Q^T Q = I$ (donde I es la matriz identidad). Así, se tiene que:

$$\|Q(\underline{X})\|_2 = \|\underline{X}\|_2 \tag{4.4}$$

y

$$d_2(Q(\underline{X}), Q(\underline{Y})) = d_2(\underline{X}, \underline{Y}). \tag{4.5}$$

La métrica Euclidiana puede ser generalizada de dos maneras. Una manera es la aplicación de la norma $\mathbb{L}^p$

$$\|\underline{X}\|_p = \left(\sum_{k=1}^{l} |X|_k^p \right)^{\frac{1}{p}} \tag{4.6}$$

con $(p \geq 1)$, de donde se obtiene, por analogía,

$$d_p(\underline{X}, \underline{Y}) = \|\underline{X} - \underline{Y}\|_p \tag{4.7}$$

La desigualdad $d_p(\underline{X}, \underline{Y}) \leq d_q(\underline{X}, \underline{Y})$ se cumple sí y sólo sí $p \geq q$.

Se tienen como casos especiales cuando $p = 1$ y $p = \infty$. Las normas

$$\|\underline{X}\|_1 = \sum_{k=1}^{l} |X_k| \text{ y } \|\underline{X}\|_\infty = \max_k |X_k| \tag{4.8}$$

y las distancias generadas por éstas son:

$$d_1(\underline{X}, \underline{Y}) = \|\underline{X} - \underline{Y}\|_1 \text{ y } d_\infty(\underline{X}, \underline{Y}) = \|\underline{X} - \underline{Y}\|_\infty . \tag{4.9}$$

El segundo tipo de generalización es obtenido definiendo

$$\|\underline{X}\|_{\mathbf{B}} = \sqrt{\underline{X}^T \mathbf{B} \underline{X}}, \tag{4.10}$$

en vez de (4.2). Donde $\mathbf{B}$ es una matriz definida positiva, esto es, una matriz simétrica tal que $\underline{X}^T \mathbf{B} \underline{X} \geq 0$ para toda $\underline{X}$ y $\underline{X}^T B \underline{X} = 0$ si y sólo sí $\underline{X} = \underline{0}$. La métrica correspondiente a (4.10) es entonces

$$d_{\mathbf{B}}(\underline{X}, \underline{Y}) = \sqrt{(\underline{X} - \underline{Y})^T \mathbf{B} (\underline{X} - \underline{Y})} \tag{4.11}$$

En casos particulares, cuando $\mathbf{B}$ es una matriz diagonal, los elementos de la diagonal son ponderadores positivos para los componentes de los vectores que corresponden a las variables en la matriz de datos. Especificando una $\mathbf{B}$ conveniente para la matriz de los datos, se obtiene la llamada distancia de *Mahalanobis*, la cual, como se verá posee la propiedad más general de invarianza, es decir, es invariante bajo toda transformación no singular C.

Sean $X_{\cdot k}$ la k-ésima columna y $X_{i \cdot}$ la i-ésima fila de la matriz de datos (X_{ik}). La matriz $\mathbf{S} = (s_{kj})$ definida por

$$s_{kj} = \frac{1}{m} \sum_{i=1}^{m} (X_{ik} - \overline{X}_{\cdot k})(X_{ij} - \overline{X}_{\cdot j}) \tag{4.12}$$

con $k, j = 1, \ldots, l$ es llamada la matriz de covarianzas de las variables. La matriz de covarianza T de los objetos esta definida como

$$t_{ij} = \frac{1}{l} \sum_{k=1}^{l} (X_{ik} - \overline{X}_{i \cdot})(X_{jk} - \overline{X}_{j \cdot}) \tag{4.13}$$

con $i, j = 1, \ldots, m$

Las matrices de correlación que se necesitan para las técnicas R y Q del análisis de factores estan dadas por

$$r_{kj} = \frac{s_{kj}}{\sqrt{s_{kk} s_{jj}}} \tag{4.14}$$

donde $k, j = 1, \ldots, l$ y

$$q_{ji} = \frac{t_{ij}}{\sqrt{t_{ii} t_{jj}}} \tag{4.15}$$

donde $i, j = 1, \ldots, m$ respectivamente.

Si se escribe

$$\widetilde{X} = \left(\widetilde{X}_{ik}\right) = \left(X_{ik} - X_{.k}\right), \tag{4.16}$$

entonces es posible escribir la matriz $\mathbf{S}$ como

$$\mathbf{S} = \frac{1}{m} \widetilde{X}^T \widetilde{X} \tag{4.17}$$

Cuando las columnas de $\widetilde{X}$ y, en consecuencia las de X, son linealmente independientes, como generalmente puede suponerse cuando $m \gg l$, la matriz S de covarianzas es definida positiva. Por consecuencia la matriz $\mathbf{S}$ es no-singular y así la matriz $\mathbf{S}^{-1}$existe y, además, también es definida positiva. Así la *distancia de Mahalanobis* puede ser definida como

$$d_{\mathbf{S}}\left(X_i^T, X_j^T\right) = \sqrt{\left(X_{i.} - X_{j.}\right) \mathbf{S}^{-1} \left(X_{i.} - X_{j.}\right)^T} \tag{4.18}$$

De esta manera la distancia es invariante bajo toda transformación $C \in M_{l \times l}$ no-singular aplicada a los vectores columna, esto es, si se escribe $y_i^T = C^T X_i^T$ tal que $Y_{i.} = X_{i.} C$ (con $i = 1, ..., m$), entonces

$$\frac{1}{m} \widetilde{Y}^T \widetilde{Y} = \frac{1}{m} \left(\widetilde{X} C\right)^T \widetilde{X} C$$

obteniendo,

$$\begin{aligned} d_{\mathbf{S}}\left(Y_{i.}^T, Y_{j.}^T\right) &= \sqrt{\left(X_{i.} - X_{j.}\right) C \left[\tfrac{1}{m} \left(\widetilde{X} C\right)^T \widetilde{X} C\right]^{-1} C^T \left(X_{i.} - X_{j.}\right)^T} \\ &= d_{\mathbf{S}}\left(X_{i.}^T, X_{j.}^T\right), \end{aligned}$$

donde la ecuación anterior representa la propiedad de invarianza. En particular, si C es una matriz diagonal con los elementos de la diagonal distintos de cero, las transformaciones de X por C significan que los valores de cada variable en X son multiplicados por una constante. En este caso se usa el término "*escalamiento*", y la distancia de Mahalanobis se dice que es una escala invariante, desde que proporciona la misma medida de distancia que cualquier otra unidad usada para medir las variables. Las otras medidas, en particular la Euclidiana, no poseen esta propiedad.

Observación 4.2 *A la distancia Euclidiana, de aquí en adelante, se le denotará como* $e(\underline{X}_i, \underline{X}_j)$.

Distancia Euclidiana Promedio o Coeficiente Taxonómico Promedio, $d(\underline{X}_i, \underline{X}_j)$.

Este coeficiente esta fuertemente ligado a la distancia Euclidiana, $e(\underline{X}_i, \underline{X}_j)$. Como se verá en la sección de "*Valores Faltantes*", $d(\underline{X}_i, \underline{X}_j)$ tiene la ventaja de poder manejar valores faltantes, mientras $e(\underline{X}_i, \underline{X}_j)$ no. A continuación se da la definición.

Definición 4.6 *La distancia Euclidiana promedio esta definida por*

$$d(\underline{X}_i, \underline{X}_j) = \left[\sum_{k=1}^{p}(X_{i,k} - X_{j,k})^2/n\right]^{1/2}$$

cuyo rango es $0 \leq d(\underline{X}_i, \underline{X}_j) \leq \infty$ *y así,* $d(\underline{X}_i, \underline{X}_j) = e(\underline{X}_i, \underline{X}_j)/n^2$.

Esto quiere decir que, al realizar dos análisis de conglomerados con la misma matriz de datos y el mismo método, pero el primero efectuado con la distancia Euclidiana , $e(\underline{X}_i, \underline{X}_j)$, y el segundo con la distancia Euclidiana promedio, $d(\underline{X}_i, \underline{X}_j)$, los dendogramas obtenidos de ambos análisis tendrán la misma topografía, esto es, los elementos serán unidos en el mismo orden y los valores a los que fueron unidos dos objetos diferirán por la constante $(1/n)^2$, de esta manera, ambos dendogramas proporcionarán las mismas conclusiones.

Ejemplo 4.1 *Sea*

$$\mathbf{X} = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{pmatrix} 10 & 5 \\ 5 & 10 \end{pmatrix} \end{array}$$

una matriz de datos, entonces

$$\begin{aligned} d(\underline{X}_i, \underline{X}_j) &= \left[\left((20-35)^2 + (40-55)^2 + (25-40)^2 + (30-45)^2\right) / 4\right]^{1/2} \\ &= 15. \end{aligned}$$

Coeficiente de Formas, $z(\underline{X}_i, \underline{X}_j)$.

Este es un coeficiente de similaridad y esta definido en el rango

$$0 \leq z(\underline{X}_i, \underline{X}_j) \leq \infty$$

y es función de $d(\underline{X}_i, \underline{X}_j)$ (la Distancia Euclidiana Promedio).

Definición 4.7 *El coeficiente de formas se define como*

$$z(\underline{X}_i, \underline{X}_j) = \left\{[n/(n-1)]\left(d^2(\underline{X}_i, \underline{X}_j) - q^2(\underline{X}_i, \underline{X}_j)\right)\right\}^{1/2}$$

donde

$$q^2(\underline{X}_i, \underline{X}_j) = (1/n^2)\left(\sum_{i=1}^{n} X_{ij} - \sum_{i=1}^{n} X_{ik}\right)^2.$$

Ejemplo 4.2 *Considerese la matriz dada por (4.1), entonces*

$$\sum_{i=1}^{4} X_{1,i} = 20 + 40 + 25 + 45 = 115$$

$$\sum_{j=1}^{4} X_{2,j} = 35 + 55 + 40 + 45 = 175$$

$$d^2(\underline{X}_1, \underline{X}_2) = (15)^2 = 225$$

así,

$$q^2(\underline{X}_1, \underline{X}_2) = \left[1/(4)^2\right](115 - 175)^2 = 225$$

por tanto,

$$z(\underline{X}_1, \underline{X}_2) = \left\{[4/(4-1)](225 - 225)\right\}^{1/2} = 0$$

En conclusión, $z(\underline{X}_i, \underline{X}_j)$ ignora las traslaciones aditivas de los perfiles, así $z(\underline{X}_i, \underline{X}_j)$ es igual a cero cuando los objetos i y j son iguales y cuando los perfiles entre ambos difieren por una constante.

Coeficiente del Coseno, $c(\underline{X}_i, \underline{X}_j)$.

El coeficiente del coseno es una función de similaridad definida en el rango $-1 \leq c(\underline{X}_i, \underline{X}_j) \leq 1$. El valor de $c(\underline{X}_i, \underline{X}_j) = 1$ indica maxima similaridad, pero no necesariamente los perfiles de $\underline{X}_i$ y $\underline{X}_j$ son iguales. El valor de $c(\underline{X}_i, \underline{X}_j) = -1$ indica máxima disimilaridad.

Definición 4.8 *El coeficiente del coseno esta definido por*

$$c(\underline{X}_i, \underline{X}_j) = \frac{\sum_{k=1}^{p} X_{i,k} X_{j,k}}{\left(\sum_{k=1}^{p} X_{i,j}^2\right)^{1/2} \left(\sum_{k=1}^{p} X_{j,k}^2\right)^{1/2}}.$$

Cuando los objetos i y j son vistos como puntos en el espacio Euclidiano y cada punto esta conectado con el origen, entonces $c(\underline{X}_i, \underline{X}_j)$ es el coseno del ángulo entre los dos vectores.

Ejemplo 4.3 *De la matriz (4.1) se tiene que*

$$\sum_{k=1}^{4} X_{1,k} X_{2,k} = (20)(35) + (40)(55) + (25)(40) + (30)(45) = 5,250.$$

$$\sum_{k=1}^{4} X_{1,k}^2 = (20)^2 + (40)^2 + (25)^2 + (30)^2 = 3,525.$$

$$\sum_{k=1}^{4} X_{2,k}^2 = (35)^2 + (55)^2 + (40)^2 + (45)^2 = 7,875.$$

por tanto

$$c(\underline{X}_1, \underline{X}_2) = 5,250 / \left[(3,525)^{1/2} (7,875)^{1/2}\right] = 0,996.$$

Claramente $c(\underline{X}_i, \underline{X}_j)$ *mide la similaridad evitando casi todos los desplazamientos de tamaño entre los perfiles de datos. Si dos perfiles tienen esencialmente las mismas formas, sin considerar los desplazamientos de tamaño, el coeficiente* $c(\underline{X}_i, \underline{X}_j)$ *proporcionará un alto grado de similaridad.*

Coeficiente de Correlación, $r(\underline{X}_i, \underline{X}_j)$

Este coeficiente esta definido en el rango $-1 \leq r(\underline{X}_i, \underline{X}_j) \leq 1$. El valor de $r(\underline{X}_i, \underline{X}_j) = 1$ indica maxima similaridad, pero no necesariamente los perfiles son iguales, y $r(\underline{X}_i, \underline{X}_j) = -1$ indica maxima disimilaridad.

Definición 4.9 *El coeficiente de correlación de Pearson está definido como*

$$r(\underline{X}_i, \underline{X}_j) = \frac{\sum_{k=1}^{p} X_{i,k}X_{j,k} - (1/n)\left(\sum_{k=1}^{p} X_{i,k}\right)\left(\sum_{k=1}^{p} X_{j,k}\right)}{\left\{\left[\sum_{k=1}^{p} X_{i,k}^2 - (1/n)\left(\sum_{k=1}^{p} X_{i,k}\right)^2\right]\left[\sum_{k=1}^{p} X_{j,k}^2 - (1/n)\left(\sum_{k=1}^{p} X_{j,k}\right)^2\right]\right\}^{1/2}}.$$

Ejemplo 4.4 *Considérense los datos de la matriz (4.1) y los calculos realizados en $z(\underline{X}_i, \underline{X}_j)$ y $c(\underline{X}_i, \underline{X}_j)$ se tiene que*

$$r(\underline{X}_i, \underline{X}_j) = \frac{5,250 - (1/4)(115)(175)}{\left\{\left[3,525 - (1/4)(115)^2\right]\left[7,875 - (1/4)(175)^2\right]\right\}^{1/2}} = 1.$$

El coeficiente $r(\underline{X}_i, \underline{X}_j)$ considera que los objetos 1 y 2 son 100% similares. Este coeficiente es totalmente insensible bajo traslaciones aditivas.

Observación 4.3 *El coeficiente $r(\underline{X}_i, \underline{X}_j)$ siempre juzga a cualquier perfil de datos y su traslación de imagen bajo un espejo como los más disimilares, esto es, considerando de nuevo el ejemplo anterior, $r(\underline{X}_{1i}, \underline{X}_4) = -1$.*

Coeficiente Métrico de Canberra, $a(\underline{X}_i, \underline{X}_j)$

Este es un coeficiente de disimilaridad y está definido en el rango $0 \leq a(\underline{X}_i, \underline{X}_j) \leq 1$, donde $a(\underline{X}_i, \underline{X}_j) = 0$ indica máxima similaridad, que ocurre cuando los perfiles de los objetos son idénticos.

Definición 4.10 *El coeficiente métrico de Canberra se define como*

$$a(\underline{X}_i, \underline{X}_j) = (1/n)\sum_{k=1}^{p} \frac{|X_{i,k} - X_{j,k}|}{(X_{i,k} + X_{j,k})}.$$

Ejemplo 4.5 *Tomése los datos de la matriz (4.1), por tanto*

$$\begin{aligned} a(\underline{X}_1, \underline{X}_2) &= (1/4)\left[\tfrac{|20-35|}{(20+35)} + \tfrac{|40-55|}{(40+55)} + \tfrac{|25-40|}{(25+40)} + \tfrac{|30-45|}{(30+45)}\right] \\ &= (1/4)(0{,}273 + 0{,}158 + 0{,}231 + 0{,}20) \\ &= 0{,}21 \end{aligned}$$

cada término de la suma esta entre 0 y 1, igualando la contribución de cada atributo sobre toda la similaridad. El término $(1/n)$ promedia las n proporciones.

Coeficiente de Bray-Curtis, $b(\underline{X}_i, \underline{X}_j)$

Este coeficiente es un coeficiente de disimilaridad definido en el rango $0 \leq b(\underline{X}_1, \underline{X}_2) \leq 1$, donde $b(\underline{X}_1, \underline{X}_2) = 0$ indica máxima similaridad, que ocurre cuando los perfiles son idénticos.

Definición 4.11 *El coeficiente de Bray-Curtis está definido por*

$$b(\underline{X}_i, \underline{X}_j) = \frac{\sum_{k=1}^{p} |X_{i,j} - X_{j,k}|}{\sum_{k=1}^{p} (X_{i,k} - X_{j,k})}.$$

Ejemplo 4.6 *Sea X la matriz dada por (4.1) entonces*

$$\begin{aligned} b(\underline{X}_1, \underline{X}_2) &= \frac{|20-35|+|40-55|+|25-40|+|30-45|}{(20+35)+(40+55)+(25+40)+(30+45)} \\ &= 60/290 \\ &= 0{,}21. \end{aligned}$$

Observación 4.4 *Para cualquier atributo k tal que $X_{i,k} = X_{j,k} = 0$, el denominador del k-ésimo término de $a(\underline{X}_i, \underline{X}_j)$ será cero. En tales casos $a(\underline{X}_i, \underline{X}_j)$ estará indefinido. El coeficiente $b(\underline{X}_i, \underline{X}_j)$ en este caso es mejor, pero no trabaja con pares de objetos i y j que tengan ceros en los p atributos, pues en tal caso el coeficiente $b(\underline{X}_i, \underline{X}_j)$ se indefine.*

Observación 4.5 *Una fórmula equivalente para $b(\underline{X}_i, \underline{X}_j)$ es*

$$b(\underline{X}_i, \underline{X}_j) = 1 - \frac{2\sum_{K=1}^{p} \text{mín}\,(X_{i,k}, X_{j,k})}{\sum_{k=1}^{p} (X_{i,k} + X_{j,k})}$$

con ésta fórmula equivalente algunos investigadores definen $b(\underline{X}_i, \underline{X}_j)$ como un coeficiente de similaridad en vez de un coeficiente de disimilaridad definiéndolo como

$$b'(\underline{X}_i, \underline{X}_j) = \frac{2\sum_{K=1}^{p} \text{mín}\,(X_{i,k}, X_{j,k})}{\sum_{k=1}^{p} (X_{i,k} + X_{j,k})}.$$

Cuando es definido de esta manera su rango sigue siendo $0 \leq b'(\underline{X}_i, \underline{X}_j) \leq 1$, pero ahora 1 es el punto máximo de similaridad.

Observación 4.6 *$a(\underline{X}_i, \underline{X}_j)$ usa una suma de términos normalizados igualmente ponderados a la contribución de cada atributo sobre toda la semejanza, mientras que $b(\underline{X}_i, \underline{X}_j)$ los pondera desigualmente.*

Más distancias y funciones de similaridad

1. Distancia Euclidiana:

$$\left\{\sum_{k=1}^{p} w_k \left(X_{rk} - X_{sk}\right)\right\}^{1/2}$$

   a) Se llama distancia Euclidiana no-estandarizada si $w_k = 1$.

   b) Se llama distancia Euclidiana estandarizada por su desviación estándar si $w_k = 1/s_k^2$ (distancia de Karl Pearson).

   c) Se llama distancia Euclidiana estandarizada por rango si $w_k = 1/R_k^2$ donde $R = \max_{i,j} |X_{ik} - X_{jk}|$

2. Distancia de Mahalanobis:

$$\left\{(X_r - X_s)' \sum^{-1} (X_r - X_s)\right\}$$

donde $\sum$ es cualquier matriz definida positiva.

3. Métrica por bloques (city-block distance) (métrica Manhattan):

$$\sum_{k=1}^{p} w_k \left|X_{rk} - X_{sk}\right|.$$

4. Métrica de Minkowski:

$$\left\{\sum_{k=1}^{p} w_k \left(X_{rk} - X_{sk}\right)^{\lambda}\right\}^{1/\lambda}.$$

5. Métrica de Canberra:

$$\sum_{k=1}^{p} \frac{|X_{rk} - X_{sk}|}{(X_{rk} - X_{sk})}.$$

6. Distancia de Bhattacharyya (proporciones):

$$\left\{\sum_{k=1}^{p}\left(x_i^{1/2}-y_i^{1/2}\right)^2\right\}^{1/2}.$$

7. Distancias entre grupos:

a) Coeficiente de disimilaridad de Karl Pearson:

$$\left\{\frac{1}{p}\sum_{k=1}^{p}\frac{\left(\overline{X}_{rk}-\overline{X}_{sk}\right)^2}{\left(s_{rk}^2/n_r\right)+\left(s_{sk}^2/n_s\right)}\right\}^{1/2}.$$

donde n_j es el tamaño de la j-ésima muestra, $j=r,s$; $\overline{X}_{jk}$, s_{jk}^2 son la media y la varianza de la k-ésima variable y para la j-ésima muestra, respectivamente.

b) Distancia de Mahalanobis;

$$\left\{(X_r-X_s)\widehat{\sum}^{-1}(X_r-X_s)\right\}.$$

Distancias y funciones de similaridad entre conglomerados

Sean m objetos, etiquetados desde 1 hasta m, y sea C un conglomerado etiquetado por una p, definido como un conjunto no vacío de índices tal que $C_p \subset \{1,...,m\}$. contiene m_p elementos x_i con $i \in C_p$. Dada la matriz de datos, la media (del conglomerado) se define como:

$$\overline{X}_p = \frac{1}{m_p}\sum_{i\in C_p} X_i^T$$

y puede ser usada para representar a los objetos en el conglomerado C_p

Algunas de las distancias entre conglomerados más usuales son:

$$d_M(C_p,C_q) = \left\|\overline{X}_{p.}^T - \overline{X}_{q.}^T\right\|_2, \tag{4.19}$$

$$d_M{}'(C_p,C_q) = \sqrt{\frac{m_p m_q}{m_p+m_q}}\left\|\overline{X}_{p.}^T - \overline{X}_{q.}^T\right\|_2, \tag{4.20}$$

$$d_{\mathbf{S}}(C_p,C_q) = \sqrt{\left(\overline{X}_{p.}^T - \overline{X}_{q.}^T\right)\mathbf{S}^{-1}\left(\overline{X}_{p.}^T - \overline{X}_{q.}^T\right)^T}, \tag{4.21}$$

en donde **S** esta definida como en (4.17). d_M es la distancia Euclidiana de los centroides de los conglomerados, $d_M{}'$ es la misma pero multiplicada por una función del número de elementos y d_S es la generalización de la distancia de Mahalanobis.

Algunos ejemplos de funciones distancia no-métricas son

$$d_Z\left(C_p, C_q\right) = \min_{\substack{i \in C_p \\ j \in C_q}} d_Z\left(X_{i.}^T, X_{j.}^T\right) \tag{4.22}$$

y algunas veces es posible definir de manera análoga funciones de similaridad no-métricas como

$$s_Z\left(C_p, C_q\right) = \max_{\substack{i \in C_p \\ j \in C_q}} s_Z\left(X_{i.}^T, X_{j.}^T\right) \tag{4.23}$$

donde Z representa cualquier función distancia o de similaridad.

Otras posibilidades son

$$s_Z\left(C_p, C_q\right) = \sum_{i \in C_p} \sum_{j \in C_q} s_Z\left(X_{i.}^T, X_{j.}^T\right) \tag{4.24}$$

o con ponderadores,

$$s_Z(C_p, C_q) = \frac{1}{m_p m_q} \sum_{i \in C_p} \sum_{j \in C_q} s_Z(X_{i.}^T, X_{j.}^T) \tag{4.25}$$

4.2.3. Sensibilidad a los desplazamientos de tamaño de los coeficientes de semejanza entre los pérfiles de datos

La sensibilidad a los desplazamientos de tamaño de los coeficientes de semejanza es de suma importancia para la elección de la correcta elección del coeficiente de semejanza.

Después de haber establecido el propósito de la investigación, continúa la elección del coeficiente de semejanza respecto al criterio de que si es necesario considerar a los dezplazamientos de tamaño en la similaridad o simplemente ignorarlos lo más que se pueda. En situaciones donde los desplazamientos de tamaño puedan afectar a la similaridad, coeficientes como $d(\underline{X}_i, \underline{X}_j)$, $a(\underline{X}_i, \underline{X}_j)$ y $b(\underline{X}_i, \underline{X}_j)$ son buenas alternativas. Por el otro lado, en situaciones donde los desplazamientos de tamaño no son necesarios, los

coeficientes $c(\underline{X}_i, \underline{X}_j)$ y $r(\underline{X}_i, \underline{X}_j)$ son buenas opciones. $z(\underline{X}_i, \underline{X}_j)$ podría ser buena opción si se considera que los desplazamientos de tamaño pueden ser aproximados por traslaciones aditivas.

A continuación se presenta un cuadro en donde se observan siete coeficientes de semejanza y sus "propiedades" respecto a las traslaciones aditivas y proporcionales:

		IGNORA	
COEFICIENTE	*RANGO*	*TRASLACIONES ADITIVAS*	*TRASLACIONES PROPORCIONALES*
Disimilaridad			
$e(\underline{X}_i, \underline{X}_j)$	$[0, \infty]$	NO	NO
$d(\underline{X}_i, \underline{X}_j)$	$[0, \infty]$	NO	NO
$a(\underline{X}_i, \underline{X}_j)$	$[0, 1]$	NO	NO
$b(\underline{X}_i, \underline{X}_j)$	$[0, 1]$	NO	NO
$z(\underline{X}_i, \underline{X}_j)$	$[0, \infty]$	SI	NO
Similaridad			
$c(\underline{X}_i, \underline{X}_j)$	$[-1, 1]$	NO	SI
$r(\underline{X}_i, \underline{X}_j)$	$[-1, 1]$	SI	SI

4.3. Manejo de datos nominales, ordinales y mixtos

4.3.1. Datos nominales y ordinales

Sean $\underline{X}$, $\underline{Y}$, $\underline{Z}$ vectores con elementos X_k, Y_k, Z_k $(k = 1, ..., p)$ tales que cada elemento es 0 ó 1, los cuales denotarán las filas de la matriz de datos. Ahora se definirá lo siguiente: sea

$$\alpha = \sum_{k=1}^{p} \text{mín}\,(X_k, Y_k) \tag{4.26}$$

así α, denota el número de veces que aparece 1 en ambos vectores, $\underline{X}$ y $\underline{Y}$,

$$\beta = \sum_{k=1}^{p} X_k - \alpha \tag{4.27}$$

que denota el número de 0's en X y el número de 1's en Y,

$$\gamma = \sum_{k=1}^{p} Y_k - \alpha \tag{4.28}$$

denota el número de 1's y 0's en X y Y respectivamente, y por último sea

$$\delta = p - (\alpha + \beta + \gamma) \tag{4.29}$$

el número de 0's que aparecen en X y Y al mismo tiempo. Así pues, se pueden nombrar algunas funciones de similaridad con las definiciones arriba mencionadas

			Rango
1	$\frac{\alpha}{\alpha+\beta+\gamma}$	Jaccard	$(0,1)$
2	$\frac{\alpha+\delta}{\alpha+\beta+\gamma+\delta}$	Simple	$(0,1)$
3	$\frac{2\alpha}{2\alpha+\beta+\gamma}$	Sorenson	$(-1,1)$
4	$\frac{\alpha+\delta}{\alpha+2(\beta+\gamma)+\delta}$	Rogers y Tanimoto	$(-1,1)$
5	$\frac{2(\alpha+\delta)}{2(\alpha+\beta)+\gamma+\delta}$	Sokal y Sneath	$(0,1)$
6	$\frac{\alpha}{\alpha+\beta+\gamma+\delta}$	Russell y Rao	$(0,1)$
7	$\frac{\alpha+(\alpha\delta)^{1/2}}{\alpha+\beta+\gamma+(\alpha\delta)^{1/2}}$	Baroni-Urbani y Buser	$(0,1)$
8	$\left(\frac{\beta+\gamma}{\alpha+\beta+\gamma+\delta}\right)^{1/2}$	Distancia Binaria de Sokal	$(0,1)$
9	$\frac{\alpha}{[(\alpha+\beta)(\alpha+\gamma)]^{1/2}}$	Ochiai	$(-1,1)$

y otras menos conocidas

10	$\frac{\alpha}{\alpha+2(\beta+\gamma)}$
11	$\frac{1}{2}\left(\frac{\alpha}{\alpha+\beta}+\frac{\alpha}{\alpha+\gamma}\right)$
12	$\frac{\alpha+\delta}{\alpha+\delta+2(\beta+\delta)}$
13	$\frac{1}{4}\left(\frac{\alpha}{\alpha+\beta}+\frac{\alpha}{\alpha+\gamma}+\frac{\delta}{\gamma+\delta}+\frac{\delta}{\beta+\delta}\right)$
14	$\frac{\alpha\delta}{\sqrt{(\alpha+\beta)(\alpha+\gamma)(\beta+\delta)(\gamma+\delta)}}$

Considérense tres ejemplos. Sean

$$\underline{X} = (1,1,0,1,1,0,1)^T$$

$$\underline{Y} = (1,0,1,0,0,0,1)^T$$
$$\underline{Z} = (1,0,1,1,1,0,0)^T$$

entonces se obtiene

	α/p	$\alpha/(\alpha+\beta+\gamma)$	$(\alpha+\delta)/p$
$s(x,y)$	2/7	1/3	3/7
$s(y,z)$	2/7	2/5	4/7
$s(x,z)$	3/7	1/2	4/7

En vez de variables binarias, ahora se considerarán variables con más de dos estados, así

$$1 \leq X_k \leq j_k \qquad (k = 1, ..., p) \tag{4.30}$$

donde x_k y j_k son enteros no-negativos, para las componentes $\underline{X}_1, ..., \underline{X}_p$ de la matrix de datos, en otras palabras, se considerará una matriz nominal.

Considérese la siguiente transformación del vector $\underline{X}$ de longitud p en el vector $\widetilde{X}$ de longitud $p' = \sum_{k=1}^{p} j_k$ dada por

$$\underline{X} \rightarrow \widetilde{X} = (0, ..., \overset{X_1}{\underset{---j_1---}{1}}, ..., 0; ...; 0, ..., \overset{X_p}{\underset{---j_p---}{1}}, ..., 0)^T$$

En $\widetilde{X}$ hay p vectores binarios de longitud j_k con un 1 en el lugar correspondiente a X_k y 0's en cualquier otro caso. Por ejemplo si se tiene el vector $\underline{X} = (2,3,1)^T$ entonces aplicando la transformación se tiene $\widetilde{X} = (0,1,0;0,0,1,0;1,0)^T$ para $p = 3$, $j_1 = 3$, $j_2 = 4$, $j_3 = 2$. De esta manera los vectores nominales son mapeados en vectores binarios, y así las funciones de similaridad (asi como las distancias generadas por éstas) explicadas anteriormente se pueden aplicar.

Si los valores j_k son muy diferentes, se puede usar la función de similaridad

$$s_H(X,Y) = \frac{\sum_{k=1}^{p} \ln j_k . r(X_k, Y_k)}{\sum_{k=1}^{p} \ln j_k} \tag{4.31}$$

donde $.r(X_k, Y_k) = 0$ si $X_{k \neq y_k}$ y 1 $X_k = Y_k$ y donde los pesos son elegidos de tal forma que son proporcionales a la información de la k-ésima variable.

Si la matriz de datos contiene solamente valores ordinales, entonces un rango $0 \leq X_k, Y_k \leq r_k$ análogo a la ecuación (4.30) es válido para las componentes de los dos vectores columnas $\underline{X}^T = (X_1, ..., X_p)$ y $Y^T = (Y_1, ..., Y_p)$. El cero a la izquierda puede ser obtenido, sin pérdida de generalidad, mediante una traslación. La distinción con la ecuación (4.30) es que, si los valores son ordinales, se puede establecer si $X_k \geq Y_k$ ó $X_k \leq Y_k$ y así tiene sentido preguntarse si mín (X_k, Y_k) tiene algún significado. Si se define

$$d_0(X,Y) = \frac{\sum_{k=1}^{p} X_k + \sum_{k=1}^{p} Y_k - 2\sum_{k=1}^{p} \text{mín}(X_k, Y_k)}{\sum_{k=1}^{p} X_k + \sum_{k=1}^{p} Y_k - \sum_{k=1}^{p} \text{mín}(X_k, Y_k)}, \tag{4.32}$$

entonces d_0 es una distancia tal que

$$0 \leq d_0(\underline{X}, \underline{Y}) \leq 1 \tag{4.33}$$

Supóngase que $1 \leq X_k \leq j_k$ (X_k, j_k *son enteros*, $k = 1, ..., p$), una tranformación del vector ordinal $\underline{X}$ en el vector binario $\widetilde{X}$ de longuitud $p' = \sum_{k=1}^{p} j_k$ sería

$$\underline{X} \to \widetilde{X} = (\underbrace{1, 1, ..., 1, 0}_{j_1}; ...; \underbrace{1, 1, ..., 1, 0, ..., 0}_{j_p})^T \tag{4.34}$$

Por ejemplo para $j_k = 5$ ($k = 1, ..., 4$) los vectores binarios

$$\widetilde{X} = (1,1,0,0,0;1,1,1,1,0;1,1,1,1,1;1,0,0,0,0)^T$$

$$\widetilde{Y} = (1,1,1,0,0;1,0,0,0,0;1,0,0,0,0;1,1,1,1,1)^T$$

son las imagenes de $\underline{X} = (2,4,5,1)^T$ y $\underline{Y} = (3,1,1,5)^T$ con una distancia de

$$d_0(\underline{X}, \underline{Y}) = \frac{12}{17}$$

El semi-orden es preservado por la transformación descrita en (4.34) porque si $\underline{X} \geq \underline{Y}$ se cumple elemento por elemento, entonces $\widetilde{X}$ es lexico-graficamente más grande que $\widetilde{Y}$.

El papel de los resultados 0-0

Algunos de los coeficientes de semejanza incluyen a δ, el número de los resultados 0-0, en sus numeradores mientras que otros no. A continuación se discute la significancia de esto.

Primero, con frecuencia los que no son investigadores usan coeficientes que incluyen a δ. Esta preferencia radica en que la mayoría de los coeficientes la incluyen así ellos siguen esta tendencia.

Segundo, los coeficientes de semejanza que excluyen a δ tienden a ser correlacionados con aquellos que la incluyen, de esta forma, los análisis de conglomeración que usan cualquier tipo de coeficiente, es decir, los que incluyen a δ y los que no la incluyen, frecuentemente proporcionan los mismos resultados. Por ésta razón el incluir a δ, muchas veces, no es crítica.

Tercero, la lógica frecuentemente hace claro que δ debería ser incluida para contribuir en la similaridad. Un ejemplo muy simple sería que dos personas expresaran sus gustos y de los que no lo son por medio de un *si* y un *no*, respectivamente. La mayoría de las personas diría que ambos resultados, α y δ, deberían contar en su similaridad

4.3.2. Datos mixtos

Se discutirá ahora el caso en que la matriz de datos contiene datos nominales, ordinales y métricos mezclados. La manera más simple de manejar datos cuantitativos y cualitativos, es simplemente ignorar esta diferencia, esto es, manejar los datos cualitativos como datos cuantitativos y, de esta manera, usar un coeficiente de semejanza para datos cuantitativos.

Una segunda estrategia es la partición de la matriz de datos en dos conjuntos de atributos, uno para datos cuantitativos y otro para datos cualitativos. Una vez hecha esta partición calcular la matriz de semejanza para cada conjunto con coeficientes de semejanza apropiados y analizarlos independientemente, y esperar que el resultado de los dendogramas sea, esencialmente, el mismo. Si los dos dendogramas "dicen lo mismo" entonces cualquiera puede ser usado.

La tercera estrategia consiste en permutar a las variables de tal forma que se formen bloques de los diferentes tipos de datos, así dos vectores fila pueden ser escritos como

$$X = \left(X^{(N)}, X^{(O)}, X^{(M)}\right)^T \text{ y } Y = \left(Y^{(N)}, Y^{(O)}, Y^{(M)}\right)^T \tag{4.35}$$

donde N significa variables nominales, O ordinales y M métricas. Cualquier tipo de variables puede ser omitido.

Una distancia que puede ser definida es d_A, que es además aditiva, es

$$d_A(\underline{X},\underline{Y}) = p_1 d^{(N)}\left(X^{(N)},Y^{(N)}\right) + p_2 d^{(O)}\left(X^{(O)},Y^{(O)}\right) + p_3 d^{(M)}\left(X^{(M)},Y^{(M)}\right) \tag{4.36}$$

donde p_1, p_2 y p_3 son pesos positivos adecuadamente normalizados y $d^{(N)}$, $d^{(O)}$ y $d^{(M)}$ son distancias para datos nominales, ordinales y métricas, respectivamente. Si además cada una de esas funciones distancia son métricas entonces d_A también lo es. Lo mismo pasa cuando se trata de funciones de similaridad.

Otra posibilidad es, con un poco de pérdida de información, convertir a los vectores $\underline{X}$ y $\underline{Y}$ dados por (4.35) en vectores binarios y usar las funciones distancia definidas anteriormente. Se ha visto también, como mapear una variable nominal en una binaria y una ordinal en una binaria (otra forma es mapear una variable ordinal en una binaria de la siguiente manera: se asigna 0 si todos los valores están debajo de la mediana y se asigna 1 si todos los valores son iguales o mayores a la mediana, pero hace falta notar que este mapeo es irreversible), así sólo hace falta tratar a las variables métricas. Éstas pueden ser tratadas de dos maneras. La primera es similar a la establecida para las variables ordinales, excepto que la media es usada es vez de la mediana. La segunda es mapearlas en variables ordinales en base a una apropiada elección de los valores límites.

Una última posibilidad es transformar todas las variables se transforman de tal forma que su rango de valores es el intervalo (0,1) y la métrica Euclidiana es usada, pero modificada por un factor

$$d_2\left(\widetilde{X},\widetilde{Y}\right) = \sqrt{\frac{\sum_{k=1}^{p}\left(\widetilde{X}_k - \widetilde{Y}_k\right)}{p}} \tag{4.37}$$

Para variables binarias no se aplica alguna transformación. Las variables nominales pueden ser transformadas en variables binarias. Variables ordinales, supóngase que toman valores desde $1,...,j$, puede ser transformada de tal forma que tome los valores $0,\ 1/(j-1),\ 2/(j-1),...,1$. Para las variables métricas la transformación

$$X'_{i,k} = \frac{X_{i,k} - a_k}{b_k - a_k} \tag{4.38}$$

donde $a_k = \min_i X_{ik}$ y $b_k = \max_i X_{ik}$.

4.4. Ponderación de las variables

Ponderar una variable significa darle mayor o menor importancia en relación a las otras variables cuando se usan para determinar la matriz de proximidad entre ellas. Más aún, el ponderar afecta el cálculo de similidades entre los objetos y de esta forma influye en el análisis de conglomerados

Los factores de ponderación, mejor conocidos como *pesos*, reflejan la importancia que el investigador asigna a las variables para la clasificación. Esta asignación puede ser, ya sea como el resultado del criterio del investigador o en algún aspecto basado en la matriz de datos.

Hay cuatro formas para ponderar los atributos. Primero, el investigador puede dejar fuera a algunos atributos, esto es dándoles como peso el valor de cero, de esta forma solo los atributos que posean un peso distinto a cero serán incluidos en el análsis.

Una segunda forma de ponderar a las variables requiere de un análisis-R o un análisis de los atributos antes de hacer un análsis-Q o análsis de objetos. Un análsis-R encontrará atributos altamente correlacionados. Los atributos correlacionados pueden ser vistos por el investigador como un conteo múltiple de cada atributo. Si es así, el investigador puede seleccionar un atributo de cada conjunto correlacionado como representate del conjunto, resultando una nueva matriz de datos teniendo un conjunto de atributos no correlacionados. Los atributos removidos de la matriz original recibirán un peso de cero y no contribuirán a la matrzi de semejanza entre los objetos.

Para matrices con datos cuantitativos se tiene una variación de este procedimiento que involucra el uso de componentes principales para el análisis-R. Da un nuevo conjunto de p atributos y sus valores: estos son los componentes principales. Los componentes principales tienen la deseable propiedad de que son no correlacionados dos a dos. Usando los componentes principales en lugar de los atributos originales son una forma de ponderación.

La tercera forma, es el uso de cualquier función estandarizadora (siguiente sección). Esto altera la contribución de un atributo sobre toda la semejanza de dos objetos.

La cuarta forma es introducir pesos para que los atributos contribuyan cantidades deseables a la semejanza entre dos objetos. Un peso de un atributo puede ser incrementado repitiendo su presencia en la matriz de datos. Para ilustrar esta forma de ponderación supóngase que la matriz de datos originales posee solamente dos atributos, y supóngase también que se desea que el primer atributo contribuya el 60 por ciento en toda la semejanza y

el segundo solo el 40 por ciento, relativo a lo que ellos contribuirían originalmente. Esto es, se repetiría dos veces el atributo 1 y una vez el atributo 2 y de esta manera obtener una nueva matriz de datos que contiene cinco atributos, donde el primer atributo aparece tres veces y el segundo solo dos. El primer atributo tendría el 60 por ciento, y el segundo el 40 por ciento.

Para realizar esta última forma, sin tener que repetir los atributos y así formar una nueva matriz de datos, usando la matriz original se necesita hacer una pequeña modificación a los coeficientes de semejanza, por ejemplo, si se usa la distancia Euclidiana promedio, $d(X_i, X_j)$, se reescribe como

$$d(X_i, X_j) = \left[\frac{\sum_{k=1}^{p} W_j \left(X_{i,k} - X_{j,k} \right)}{\sum_{k=1}^{p} W_k} \right]^{1/2}.$$

Los pesos W_j enteros, $1, 2, 3, \ldots$ Para el ejemplo arriba mencionado, $W_1 = 3$ y $W_2 = 2$ y $p = 2$. Esto proporcionará el 60 por ciento al atributo 1 y el 40 al segundo atributo.

Los métodos para ponderar variables pueden ser combinados para producir una variedad de esquemas. Por ejemplo, primero se puede estandarizar para remover pesos naturales y después ponderar la matriz estandarizada para imponer los pesos deseados.

Aún cuando los métodos para ponderar son fáciles de realizar, la decisión si se pondera a las variables o no y en que radios, es difícil. Para esto es necesario saber la meta de investigación y asi decidir si es necesario ponderar a las variables o no.

4.5. Estandarización

En muchas aplicaciones las variables que describen a los objetos que van a ser clasificados no serán medidos en las mismas unidades. Además éstas pueden ser de diferentes tipos, algunas categóricas, otras ordinales y tal vez algunas otras intervalares. Para estandarizar una matriz de datos primero se necesita elegir una *función estandarizadora* y aplicarla a la matriz de datos.

Sea

$$\mathbf{X} = \begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,j} & \cdots & X_{1,p} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,j} & \cdots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ X_{i,1} & X_{i,2} & \cdots & X_{i,j} & \cdots & X_{i,p} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \cdots & X_{n,j} & \cdots & X_{n,p} \end{pmatrix}$$

donde $X_{i,j}$ denota al j-ésimo atributo del i-ésimo objeto. El correspondiente valor estandarizado para $X_{i,j}$ es $Z_{i,j}$. Se pueden mencionar algunas funciones estandarizadoras. La función estandarizadora más usada en la práctica está dada por

$$Z_{i,j} = \frac{X_{i,j} - \overline{X}_j}{S_j} \tag{4.39}$$

donde

$$\overline{X}_j = \frac{\sum_{i=1}^{n} X_{i,j}}{n}$$

y

$$S_j = \left(\frac{\sum_{i=1}^{n} \left(X_{i,j} - \overline{X}_j \right)^2}{n-1} \right)^{1/2}.$$

Cuando todos los valores de la matriz de datos son no-negativos y no todos los atributos de un objeto son ceros, entonces se pueden estandarizar los valores en proporciones tales que $0 \leq Z_{i,j} \leq 1$. Hay tres formas de hacer esto. La primera es

$$Z_{i,j} = \frac{X_{i,j}}{CMAX_j} \tag{4.40}$$

donde $CMAX_j$ es el valor máximo en la j-ésima columna de la matriz de datos.

La segunda función estandarizadora es

$$Z_{i,j} = \frac{X_{i,j} - CMIN_j}{CMAX_j - CMIN_j} \tag{4.41}$$

donde $CMIN_j$ es el valor mínimo de la j-ésima columna de la matriz de datos, y a $CMAX_j - CMIN_j$ se le conoce como el rango de valores en la j-ésima columna. Por último, la tercer forma de estandarizar en proporciones es

$$Z_{i,j} = \frac{X_{i,j}}{\sum_{j=1}^{n} X_{i,j}}. \tag{4.42}$$

Las expresiones (4.39) a (4.42) son funciones columna porque los valores de la j-ésima columna dependen solamente en los valores de la j-ésima columna. Para estas ecuaciones se tienen las correspondientes funciones fila que siguen la misma lógica, los valores de la i-ésima fila dependen solamente en los valores de la i-ésima fila. Como función fila (4.39) se escribe como

$$Z_{i,j} = \frac{X_{i,j} - \overline{X}_i}{S_i} \tag{4.43}$$

donde

$$\overline{X}_i = \frac{\sum_{j=1}^{p} X_{i,j}}{p}$$

y

$$S_i = \left(\frac{\sum_{j=1}^{p} \left(X_{i,j} - \overline{X}_i \right)^2}{p-1} \right)^{1/2}$$

Como una función fila, la ecuación (4.40) es

$$Z_{i,j} = \frac{X_{i,j}}{FMAX_i} \tag{4.44}$$

aquí, $FMIN_i$ es el valor mínimo de la $i-$ésima fila de la matriz de datos.

La correspondiente función fila de (4.41) es

$$Z_{i,j} = \frac{X_{i,j} - FMIN_i}{FMAX_i - FMIN_i} \tag{4.45}$$

donde $FMIN_i$ es el mínimo de la i-ésima fila.

Finalmente, la ecuación (4.42) tiene su equivalente función fila dada por

$$Z_{i,j} = \frac{X_{i,j}}{\sum_{i=1}^{n} X_{i,j}}.$$

Observación 4.7 *La estandarización de variables es un caso particular de la ponderación de las mismas.*

Recuérdese que el análisis-Q es un análisis de conglomerados para los objetos y el análisis-R es una análisis de conglomerados para los atributos. Ambos análisis requieren el uso de funciones estandarizadoras fila o columna. En otras palabras, si se quieren comparar similaridades entre objetos,

se necesita una función estandarizadora columna, y de manera análoga, si se desea comparar similaridades entre atributos se necesita una función estandarizadora fila.

Algunas veces la meta de la investigación necesita un análisis-Q y un análisis-R aplicados a la misma matriz de datos. Cuando es necesario estandarizar, usualmente la matriz de datos debería ser estandarizada por separado para cada análisis, pero en cualquier caso, la matriz de datos no debería ser estandarizada por una función columna y una función fila en sucesión, la razón es porque al estandarizar por columnas primero y luego por filas, generalmente, es distinto que al estandarizar por filas y luego por columnas, así el orden si importa.

4.6. Transformaciones de datos y datos atípicos

El estandarizar es una forma de transformar los datos originales. Hay otros dos tipos de cambiar los datos: 1) transformando los datos, y 2) identificando y eliminando valores "ruidosos" de los datos llamados datos atípicos. Para transformar los datos se requiere la selección de una función "transformadora". Esta difiere de una función estandarizadora en que a pesar de que las dos cambian los valores de los atributos en la matriz de datos, una función estandarizadora tiene parámetros que estan determinados por los datos, mientras que una función transformadora no.

Ejemplos típicos de una función estandarizadora son $Z_{i,j} = \log X_{i,j}$ y $Z_{i,j} = \sqrt{X_{i,j}}$. Ahora un *outlier* es cualquier valor atípico en los datos con respecto a los otros valores en el conjunto de datos. Debido a que los datos atípicos son anticonvencionales, estos ejercen mucha influencia en la determinación de los parámetros de una función estandarizadora y en la semejanza entre los objetos.

4.7. Valores faltantes.

Los valores faltantes son comunes en todas las ramas del análisis multivariado. Estos pueden aparecer debido a una gran variedad de razones, por ejemplo, si se realizó una medición y luego se perdió o bien si la medida no

puede ser hecha por completo, así estos valores faltantes se pueden tratar de diferentes maneras.

El criterio más sencillo a este problema (pero no significa que sea el mejor) puede ser el utilizar solamente a los objetos que tengan todos los valores, esto puede, sin embargo, reducir considerablemente el número de individuos disponibles para su análisis. Para el manejo de datos faltantes también se puede utilizar el coeficiente de similaridad de Gower dado por

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

donde s_{ijk} es la similaridad entre el i-ésimo objeto y el j-ésimo como medida de la k-ésima variable y w_{ijk} es normalmente 1 ó 0 dependiendo si la comparación es válida, esto es, w_{ijk} es cero si el valor de la k-ésima variable falta, ya sea para uno o para ambos objetos, y w_{ijk} es uno si ambos valores se tienen. Además w_{ijk} puede usarse como cero si la k-ésima variable es binaria. Para variables binarias o categóricas con más de dos categorías, s_{ijk} puede ser 1 si ambos objetos tienen el mismo valor y 0 en cualquier otro caso. Para variables continuas se puede utilizar la medida de similaridad

$$s_{ij} = 1 - |x_{ij} - x_{jk}| / R_k$$

donde R_k es el rango de las observaciones de la k-ésima variable. Esta estrategia supone que la contribución a la similitud entre dos objetos que habría sido proporcionada por una variable con uno o más valores faltantes es igual a la media ponderada de las contribuciones proporcionadas por las variables para las cuales se han observado los valores.

Otra posibilidad para tratar de estimar estos valores es por medio de procedimientos iterativos. Un ejemplo de este tipo de procedimientos es: El primer paso se determina el proceso de agrupación de conglomerados pero solo para los objetos que tengan todos los valores. En el segundo paso los objetos con valores faltantes se asignan a cada uno de los conglomerados (por ejemplo usando las matrices de proximidades o distancias basado en las variables disponibles). En el tercer paso los valores faltantes son estimados por medio de un estadístico dentro del conglomerado (como por ejemplo la media para el caso de variables continuas, o la moda para el caso de las variables categóricas). En el cuarto paso se realiza un análisis de conglomerados con

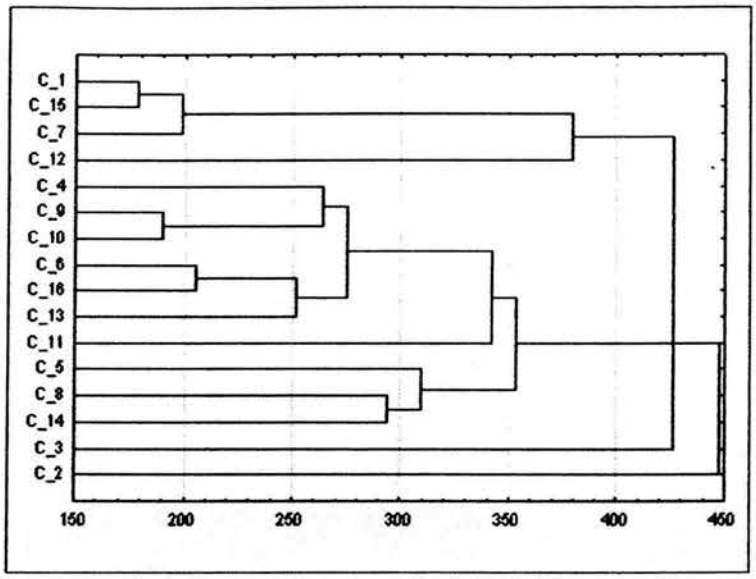
todas las variables, es decir, las variables que contienen todos sus datos y las variables que fueron estimadas anteriormente. El tercer y cuarto paso se repiten hasta que los valores estimados y los miembros del conglomerado no sufran más cambios.

Observación 4.8 *Si se desea estandarizar la matriz de datos, se deben ignorar los valores faltantes.*

4.8. Métodos jerárquicos

En la clasificación jerárquica o por jerarquías los datos no han sido particionados en un número particular de clases o conglomerados en cada paso. En vez de eso la clasificación consiste en una serie de particiones que pueden ir desde un solo conglomerado que contiene a todos los objetos hasta n conglomerados que contienen un sólo objeto. Los métodos jerárquicos pueden ser subclasificados en métodos *aglomerativos* y *divisivos*. El primer método **(aglomerativo)** consiste en una serie de pasos en la cual, en cada etapa, todos los objetos y/o todos los conglomerados son unidos en uno solo cada vez más grande, terminando en un gran conglomerado que contiene a todos los objetos. El otro procedimiento **(divisivo)** comienza con la separación de todo el grupo de objetos en partes cada vez más pequeñas, hasta que el último conglomerado, que contiene dos objetos, es roto en dos pedazos. En

ambos casos el resultado es un árbol de jerarquías.



Dendograma o arbol de jerarquias.

4.8.1. Métodos aglomerativos

De manera ilustrativa se hablará de los dos métodos más comunes, pero que sin embargo en la práctica no se usan mucho debido a que requieren de mucho tiempo máquina para muestras de tamaños considerablemente grandes. Estos métodos son los llamados "ligamiento simple" y "ligamiento completo". Además como ejemplo general se utilizará la matriz de distancias dada por

$$\mathbf{D}_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0{,}0 & & & & \\ 2{,}0 & 0{,}0 & & & \\ 6{,}0 & 5{,}0 & 0{,}0 & & \\ 10{,}0 & 9{,}0 & 4{,}0 & 0{,}0 & \\ 9{,}0 & 8{,}0 & 5{,}0 & 3{,}0 & 0{,}0 \end{pmatrix} \end{array}$$

Ligamiento simple o método de la vecindad más cercana

Este método probablemente sea uno de los métodos más simples para conglomerar. La distancia entre dos conglomerados es la menor distancia

entre los puntos de cada conglomerado, es decir

$$d(C_I, C_J) = \min_{x_i \in C_I, x_j \in C_J} \{d(X_i, X_j)\}$$

El procedimiento se explica a continuación:

a) Sean $C_1, ..., C_n$ los conglomerados iniciales conformados de un solo punto, por decir $X_i \in C_i$.

b) Sin pérdida de generalidad sea $d_{r_1,s_1} = \min_{r \neq s} \{d(X_r, X_s)\}$, en otras palabras, $C_{r_1} = \{X_{r_1}\}$ y $C_{s_1} = \{X_{s_1}\}$ son los conglomerados más cercanos. Entonces esos dos puntos son agrupados en un nuevo conglomerado, teniendo ahora $n-1$ conglomerados y donde $C_{r_1} + C_{s_1}$ es el nuevo conglomerado.

c) Sea d_{r_2,s_2} la segunda distancia más pequeña. Si ni r_1 ni s_1 son iguales a r_2 o s_2 , entonces dos nuevos conglomerados son formados, esto es, se tienen $n-2$ conglomerados $C_{r_1} + C_{s_1}$, $C_{r_2} + C_{s_2}$ más los antiguos conglomerados. Si $r_2 = r_1$ y $s_1 \neq s_2$ entonces los nuevos $n-2$ conglomerados son $C_{r_1} + C_{s_1} + C_{s_2}$ más los viejos conglomerados.

d) El proceso c) continua a través de las $\frac{1}{2}n(n-1)$ distancias. En el i-ésimo paso sea d_{r_i,s_i} la i-ésima menor distancia, entonces el conglomerado que contiene a r_i es unido al conglomerado que contiene a s_i. Nótese que si r_i y s_i están en el mismo conglomerado entonces no se forman nuevos grupos en este paso.

e) El proceso puede ser detenido antes que todos los conglomerados hayan sido unidos en un sólo conglomerado, parando cuando las distancias sean mayores a d_0, donde d_0 es un cualquier nivel umbral (threshold level).. Sean $C_1^*, ..., C_g^*$ los conglomerados resultantes. Esos conglomerados tienen la propiedad de que si d_0' $(> d_0)$ es un umbral mayor, entonces dos conglomerados C_j y C_k serán unidos en el umbral d_0' si al menos una distancia d_{rs} (o una unión simple) existe entre r y s con $X_r \in C_j$, $X_s \in C_k$ y $d_0 < d_{rs} \leq d_0'$.

A pesar de su criterio teórico, este método es frecuentemente considerado limitado debido a su pobre selectividad y su tendencia de producir árboles con largas cadenas.

Ejemplo 4.7 *Despúes de explicar este método considérese la matriz* $\mathbf{D}_1$. *La menor entrada de esta matriz es para los objetos 1 y 2, así pues estos dos elementos son unidos para formar un conglomerado de dos miembros. Despúes las distancias entre este nuevo conglomerado y los otros 3 objetos se calculan*

obteniendo una nueva matriz de distancias, digamos $\mathbf{D}_2$, *dada por*

$$\mathbf{D}_2 = \begin{array}{c} \\ (1,2) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (1,2) & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0{,}0 & & & \\ 5{,}0 & 0{,}0 & & \\ 9{,}0 & 4{,}0 & 0{,}0 & \\ 8{,}0 & 5{,}0 & 3{,}0 & 0{,}0 \end{pmatrix} \end{array}$$

Ahora la menor entrada es para los objetos 4 y 5, los cuales son unidos para formar otro conglomerado. Después se encuentra una 3a. matriz de distancias, $\mathbf{D}_3$*:*

$$\mathbf{D}_3 = \begin{array}{c} \\ (1,2) \\ 3 \\ (4,5) \end{array} \begin{array}{c} \begin{array}{ccc} (1,2) & 3 & (4,5) \end{array} \\ \begin{pmatrix} 0{,}0 & & \\ 5{,}0 & 0{,}0 & \\ 8{,}0 & 4{,}0 & 0{,}0 \end{pmatrix} \end{array}$$

La menor distancia es ahora $d_{(45)3}$ *y así el objeto 3 es unido al conglomerado que contiene a los objetos 4 y 5. Por último los grupos que contienen a los objetos 1,2 y 3,4,5 son unidos en un sólo conglomerado. Las particiones producidas a cada paso son:*

Paso	*Grupos*
P_5	[1], [2], [3], [4], [5]
P_4	[1, 2], [3], [4], [5]
P_3	[1, 2], [3], [4, 5]
P_2	[1, 2], [3, 4, 5]
P_1	[1, 2, 3, 4, 5]

Ligamiento completo o método de la vecindad más lejana

Este método evalúa la distancia entre dos conglomerados como la distancia más grande que puede ser encontrada entre cualquier par de puntos correspodientes a cada conglomerado, esto es:

$$d(C_I, C_J) = \max_{X_i \in C_i, X_j \in C_j} \{d(X_i, X_j)\}$$

El método del ligamiento completo realiza un mejor trabajo en cuanto a que no produce grupos prolongados. Además la ventaja del método del ligamiento completo es que el cálculo de las distancia pueden ser implementadas fácilmente en la computadora.

Ahora se explicará el procedimiento.

a) Se comienza con los conglomerados $C_1, ..., C_n$ que contienen a $X_1, ..., X_n$ respectivamente.

b) Se supone que $d_{1,2} = \min_{i \neq j} \{d(X_i, X_j)\}$ y sean $C_2^*, ..., C_n^*$ los grupos después de unir a la pareja 1,2 para formar C_2^*.

c) Se define una nueva matriz $((n-1)\,x\,(n-1))$ de distancias $D^* = \left(d_{ij}^*\right)$ con $d_{2j}^* = \max\{d_{1j}, d_{2j}\}$ para $j = 3, ..., n$ y $d_{ij}^* = d_{ij}$ para $i, j = 3, ..., n$. Y se encuentra $\min_{2 \leq i,j \leq n} d_{ij}^*$ y se prosigue como en b).

Se continúa hasta que todas las distancias son mayores que d_0, donde d_0 es cualquier umbral. Nótese que cuando se completa el proceso, se tiene que $\max_{i,j \in C} d_{i,j} \leq d_0$ para conglomerado C, así este método tiende a producir conglomerados compactos sin el efecto de encadenamiento.

Ejemplo 4.8 *Considérese la matriz* $\mathbf{D}_1$. *El primer paso es unir a los individuos 1 y el 2. Las distancias entre este grupo y los otros tres individuos se calculan de la siguiente forma:*

$$d_{(12),3} = \max[d_{1,3}, d_{2,3}] = d_{1,3} = 6{,}0$$

$$d_{(12),4} = \max[d_{1,4}, d_{2,4}] = d_{1,4} = 10{,}0$$

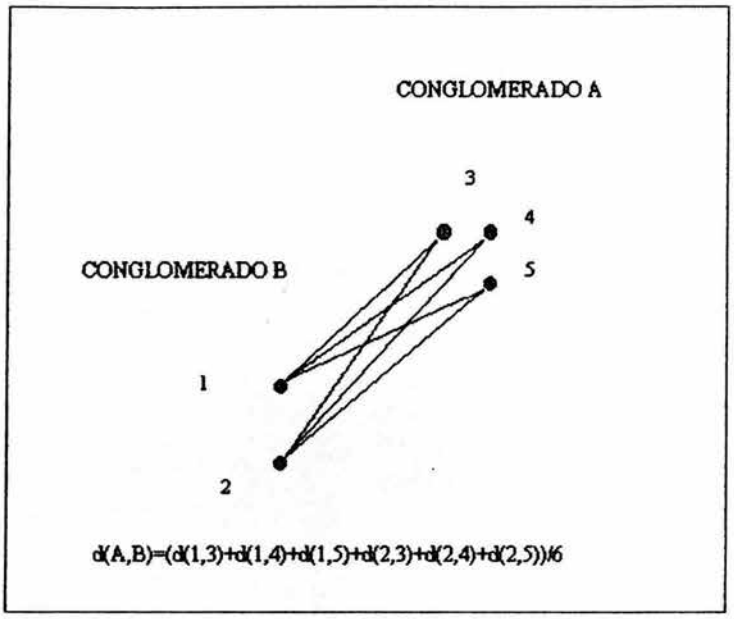
$$d_{(12),5} = \max[d_{1,5}, d_{2,5}] = d_{1,5} = 9{,}0$$

se elige la más chica y se une este último objeto al conglomerado, es decir el objeto 3 se une al conglomerado $C_{(1,2)}$ y se vuelven a calcular las distincias dadas por el máximo de las distancias repitiendóse el proceso.

Método del promedio grupal

Aqui la distancia entre los dos conglomerados es definida como el promedio de las distancias entre todos los posibles pares de individuos u objetos pero de distintos conglomerados. De manera ilustrativa considérese la sigu-

iente figura



Distancia del promedio grupal.

Ejemplo 4.9 *Aplicando este método a la matriz* $\mathbf{D}_1$*, el primer paso, al igual que en los dos métodos anteriores, es unir a los individuos 1 y 2. Las distancias serán calculadas de la siguiente forma:*

$$d_{(12),3} = \frac{1}{2}(d_{1,3} + d_{2,3}) = 5{,}5$$

$$d_{(12),4} = \frac{1}{2}(d_{1,4} + d_{2,4}) = 9{,}5$$

$$d_{(12),5} = \frac{1}{2}(d_{1,5} + d_{2,5}) = 8{,}5$$

colocando esta información en una nueva matriz $\mathbf{D}_2$ *dada por*

$$\mathbf{D}_2 = \begin{array}{c} \\ (1,2) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (1,2) & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0{,}0 & & & \\ 5{,}5 & 0{,}0 & & \\ 9{,}5 & 4{,}0 & 0{,}0 & \\ 8{,}5 & 5{,}0 & 3{,}0 & 0{,}0 \end{pmatrix} \end{array}$$

La menor entrada es $d_{4,5}$ y así los individuos 4 y 5 son unidos para formar un nuevo conglomerado. La distancia promedio entre los grupos ahora es

$$d_{(1,2)(4,5)} = \frac{1}{4}(d_{1,4} + d_{1,5} + d_{2,4} + d_{2,5})$$

y el procedimiento continua, es decir, se calcula una nueva matriz de distancias y se toma la menor de estas para formar un nuevo conglomerado.

Observación 4.9 *Los tres métodos arriba mencionados se basan únicamente en la matriz de distancias y no es necesario acceder a la matriz de datos original. Un ejemplo de un método que utiliza la matriz de datos original es el siguiente, el método del centroide.*

Métodos del centroide y de la mediana

En el método del centroide, una vez formados los grupos, los puntos de los conglomerados están representados por sus valores promedio de cada variable, esto es, su vector promedio, en otras palabras, los puntos del conglomerado están representados por un punto que está enmedio del conglomerado, como su centro de gravedad. La distancia entre dos conglomerados está dada por la distancia entre sus correspondientes medias.

El problema que se presenta al unir dos conglomerados de tamaños considerablemente distintos aparece en este método. El centroide calculado puede estar muy cerca de un conglomerado de tamaño muy grande, así la influencia de un conglomerado mucho menor puede perderse. Una forma de evitar esto es ponderar ambos grupos para dar al menor mayor influencia y así cambiar el centroide. Si el número de objetos en ambos conglomerados es igual, entonces se obtiene el método de la mediana.

Ejemplo 4.10 *Para ilustrar el método del centroide considérense los siguientes datos bivariados:*

individuo	*variable*1	*variable*2
1	1,0	1,0
2	1,0	2,0
3	6,0	3,0
4	8,0	2,0
5	8,0	0,0

Utilizando la distancia Euclidiana se obtiene una matriz de distancias dada por

$$\mathbf{D}_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0{,}00 & & & & \\ 1{,}00 & 0{,}00 & & & \\ 5{,}39 & 5{,}10 & 0{,}00 & & \\ 7{,}07 & 7{,}00 & 2{,}24 & 0{,}00 & \\ 7{,}07 & 7{,}28 & 3{,}61 & 2{,}00 & 0{,}00 \end{pmatrix} \end{array}$$

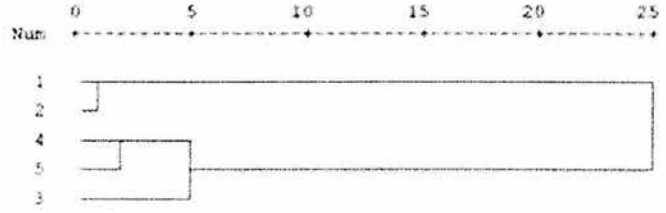
Se observa que $d_{1,2}$ es la menor de las entradas de la matriz $\mathbf{D}_1$ y los individuos 1 y 2 son unidos para formar un conglomerado. Se calcula el vector media, a saber, $(\ 1{,}0 \quad 1{,}5 \ )$, *y se calcula una nueva matriz de distancias*

$$\mathbf{D}_2 = \begin{array}{c} \\ (1,2) \\ 3 \\ 4 \\ 5 \end{array} \begin{pmatrix} 0{,}00 & & & \\ 5{,}22 & 0{,}00 & & \\ 7{,}02 & 2{,}24 & 0{,}00 & \\ 7{,}16 & 3{,}61 & 2{,}00 & 0{,}00 \end{pmatrix}$$

La menor de las entradas de $\mathbf{D}_2$ es $d_{4,5}$ y los individuos 4 y 5 son unidos para formar así un segundo conglomerado, ahora se calcula el vector media dado por $(\ 8{,}0 \quad 1{,}0 \ )$. *De nuevo se crea una matriz de distancias,*

$$\mathbf{D}_3 = \begin{array}{c} \\ (1,2) \\ 3 \\ (4,5) \end{array} \begin{pmatrix} 0{,}00 & & \\ 5{,}22 & 0{,}00 & \\ 7{,}02 & 2{,}83 & 0{,}00 \end{pmatrix}$$

En $\mathbf{D}_3$ la entrada más chica es $d_{(4,5),3}$ y así los individuos 3, 4 y 5 se unen para formar un nuevo conglomerado. El paso final consiste en unir a los 2 conglomerados en uno solo.



Dendograma creado por el metodo del Centroide.

Coeficientes de similaridad y disimilaridad en los métodos jerárquicos

Esta sección se refiere a una diferencia importante que se debe tomar en cuenta antes de usar alguno de los métodos arriba mencionados. Esta diferencia radica en el uso de un coeficiente de similaridad o un coeficiente de disimilaridad en cualquier método. Nótese también que en las secciones anteriores se usarón coeficientes de disimilaridad (distancias). Al inicio de los métodos los dos individuos más similares fueron unidos a cada paso de conglomeración, esto es, los dos conglomerados con el menor valor del coeficiente de disimilaridad fueron unidos.

Cuando se utiliza un coeficiente de similaridad la situación cambia un poco, pues ahora la semejanza se invierte, en otras palabras, si a dos individuos, a decir los individuos 1 y 2, se les calcula el coeficiente de disimilaridad, $d(\underline{X}_1, \underline{X}_2)$, y si resulta que $d(\underline{X}_1, \underline{X}_2)$ es muy cercano a cero quiere decir que los individuos son muy semejantes, en cambio, si para los mismos dos individuos, 1 y 2, se les calcula el valor de un coeficiente de similaridad, $s(\underline{X}_1, \underline{X}_2)$, entonces $s(\underline{X}_1, \underline{X}_2)$ será muy cercano a 1 indicando ahora la semejanza en 1.

Para solucionar lo arriba mencionado, solo es necesario hacer una pequeña modificación al procedimiento, lo cual se expresa en la siguiente tabla y la cual puede ser generalizada para cualquier coeficiente de similaridad y cualquier método arriba mencionado.

COEFICIENTE DE DISIMILARIDAD

$$e_{(jk)(mn)} = \text{mín}\,(e_{jm}, e_{jn}, e_{km}, e_{kn})$$
$$e_{(jk)(mn)} = \left(\tfrac{1}{4}\right)(e_{jm} + e_{jn} + e_{km} + e_{kn})$$
$$e_{(jk)(mn)} = \text{mín}\,(e_{jm}, e_{jn}, e_{km}, e_{kn})$$

y

COEFICIENTE DE SIMILARIDAD

$$r_{(jk)(mn)} = \text{máx}\,(r_{jm}, r_{jn}, r_{km}, r_{kn})$$
$$r_{(jk)(mn)} = \left(\tfrac{1}{4}\right)(r_{jm} + r_{jn} + r_{km} + r_{kn})$$
$$r_{(jk)(mn)} = \text{máx}\,(r_{jm}, r_{jn}, r_{km}, r_{kn})$$

donde la primera fila representa a ligamiento simple, la segunda al método del promedio grupal y la última al ligamiento completo.

Observación 4.10 *La razón por la cual no se incluye el método de Ward, es porque este método prescinde del uso de un coeficiente de disimilaridad o similaridad como se verá a continuación:*

Método de Ward

El método de Ward esta basado en la minimización estadística entre conglomerados. A cada paso de este método se calcula el punto central de todas las combinaciones posibles entre dos conglomerados y se calcula la suma total del cuadrado de las distancias de este punto a todos los objetos, en otras palabras usa el criterio de la suma de los cuadrados de los errores(SCE). La asociación entre dos conglomerados se da cuando la suma de los cuadrados es mínima. Para ilustrar esto considérese a 10 individuos con valores (2 6 5 6 2 2 2 0 0 0). La pérdida de información que puede resultar al tratar los 10 valores como un grupo con una media de 2,5 es representada por SCE,

$$SCE = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

para este ejemplo

$$SCE_{\text{de un grupo}} = (2-2{,}5)^2 + (6-2{,}5)^2 + \cdots + (0-2{,}5)^2 = 50{,}5$$

De igual manera si los individuos son agrupados en cuatro conjuntos,

$$\{0,0,0\}, \{2,2,2,2\}, \{5\}, \{6,6\}$$

la SCE se puede calcular como la suma de

$$SCE_{\text{de 4 grupos}} = SCE_{\text{grupo1}} + SCE_{\text{grupo2}} + SCE_{\text{grupo3}} + SCE_{\text{grupo4}} = 0{,}0.$$

4.8.2. Métodos divisivos

Los métodos divisivos operan en la dirección contraria en comparación con los métodos aglomerativos, comenzando con un gran conglomerado, después partiendo este en más conglomerados y asi sucesivamente hasta tener conglomerados individuales, es decir, conglomerados con un sólo miembro. Éstos métodos exigen mucho tiempo-máquina si se consideran todas las $2^{k-1} - 1$ posibles particiones de k objetos a cada paso. Sin embargo, para datos cuyas p variables son binarias se dispone de métodos simples y computacionalmente eficientes. Éstos métodos divisivos dividen a los conglomerados conforme a los miembros que presentan ciertos características, ya sea la presencia o ausencia de cierto atributo.

Métodos divisivos monotéticos

La elección de la variable en este tipo de métodos en el cual se va hacer un corte depende de la optimización de un criterio que refleja la asociación con las otras variables, gereralmente conocido como *Análisis de Asociación*. Los conglomerados se cortan a cada paso de acuerdo a la posesión del atributo. Por ejemplo, para un par de variables V_i y V_j con valores 1 y 0, las frecuencias observadas pueden ser

$V_j \setminus V_i$	1	0
1	a	b
0	c	d

Ahora algunas medidas de asociación (sumadas sobre todas las variables) pueden ser:

1. $|ad - bc|$
2. $(ad - bc)^2$
3. $(ad - bc)^2 n / [(a+b)(a+c)(b+d)(c+d)]$
4. $\sqrt{(ad - bc)^2 n / [(a+b)(a+c)(b+d)(c+d)]}$
5. $(ad - bc)^2 / [(a+b)(a+c)(b+d)(c+d)]$

El corte a cada paso se realiza conforme a la presencia o ausencia del atributo cuya asociación con las otras (la suma sobre todas las variables) es un máximo.

Si hay valores faltantes de una variable, por decir V_1, se busca la variable que contiene a todos sus valores y que tiene la asociación más grande, por decir V_2, así el valor faltante de V_1 es reemplazado por el valor de V_2 de la misma observación.

Ejemplo Usando el criterio 3, sea

$$\mathcal{X}^2_{jk} = \frac{(ad - bc)^2 n}{(a+b)(a+c)(b+d)(c+d)}$$

y considérense los siguientes 5 individuos con 3 variables binarias

	Variables		
Individuos	1	2	3
1	0	1	1
2	1	1	0
3	1	1	1
4	1	1	0
5	0	0	1

Calculando las $\mathcal{X}^2_{jk}$'s de cada par de variables se tiene que $\mathcal{X}^2_{1,2} = 1{,}87$, $\mathcal{X}^2_{1,3} = 2{,}22$ y $\mathcal{X}^2_{2,3} = 0{,}83$, entonces

$$\mathcal{X}^2_{1,2} + \mathcal{X}^2_{1,3} = 4{,}09$$

$$\mathcal{X}^2_{1,2} + \mathcal{X}^2_{2,3} = 2{,}07$$

$$\mathcal{X}^2_{1,3} + \mathcal{X}^2_{2,3} = 3{,}05$$

Así al usar el criterio de máx$\sum \mathcal{X}^2$, se tiene que la primera división de los datos fue en 2 subconjuntos, aquellos que poseen la primera variable y aquellos que no la poseen, por consiguiente se tiene la división $[2, 3, 4]$ y $[1, 5]$.

Métodos Divisivos Politéticos

En este tipo de métodos se usan todas las variables al mismo tiempo y pueden trabajar con una matriz de proximidades. Un procedimiento que evita considerar todos los posibles cortes (un gran problema en los métodos divisivos) es el siguiente: se encuentra el objeto que esta más alejado de los otros miembros usándolo como el inicio de un nuevo conglomerado , entonces se considera cada uno de los otros miembros como candidato para ser parte de este nuevo conglomerado: cada objeto que esté cerca de este nuevo conglomerado será unido para formar parte de éste. Este proceso se repite, el siguiente conglomerado que será cortado es el que posea el diámetro más grande.

De manera ilustrativa considérese la siguiente matriz conformada por 7 individuos:

$$\mathbf{D} = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{pmatrix} 0 & & & & & & \\ 10 & 0 & & & & & \\ 7 & 7 & 0 & & & & \\ 30 & 23 & 21 & 0 & & & \\ 29 & 25 & 22 & 7 & 0 & & \\ 38 & 34 & 31 & 10 & 11 & 0 & \\ 42 & 36 & 36 & 13 & 17 & 9 & 0 \end{pmatrix}$$

El individuo usado para iniciar un nuevo conglomerado es aquel cuya distancia promedio a los otros individuos sea la más grande, en este caso es el individuo 1, así se tiene como grupos iniciales $A = [1]$ y $B = [2,3,4,5,6,7]$. Después se encuentra la distancia promedio de cada individuo del conglomerado principal, B, a los individuos del nuevo conglomerado y viceversa, la distancia promedio de los individuos del nuevo conglomerado al conglomerado principal y se calcula la diferencia entre éstas dos distancias promedio:

Individuo en B	Prom. distancia a A	Prom. distancia a B	$B - A$
2	10.0	25.0	15.0
3	7.0	23.4	16.4
4	30.0	14.8	-15.2
5	29.0	16.4	-12.6
6	38.0	19.0	-19.0
7	42.0	22.2	-19.8

La máxima diferencia es 16.4 para el individuo 3 que es agregado al conglomerado A. Ahora los conglomerados son $A = [1,3]$ y $B = [2,4,5,6,7]$. Repitiendo este proceso, obtenemos:

Individuo en B	Prom. distancia a A	Prom. distancia a B	$B - A$
2	8.5	29.5	12.0
4	25.5	13.2	-12.3
5	25.5	15.0	-10.5
6	34.5	16.0	-18.5
7	39.0	18.7	-20.3

Ahora el individuo 2 se une al grupo A, obteniendo los conglomerados $A = [1,3,2]$ y $B = [4,5,6,7]$ y el proceso continua:

Individuo en B	Prom. distancia a A	Prom. distancia a B	$B-A$
4	24.3	10.0	-10.5
5	25.3	11.7	-13.6
6	34.3	10.0	-24.3
7	38.0	13.0	-25.0

Ahora todas las diferencias son negativas, así el proceso podría continuar (si se desea) por cada conglomerado separadamente.

4.9. Eleccion del número de grupos

Muchas veces el investigador no está interesado en toda la jerarquía, sino que está interesado en una o dos particiones obtenidas de ésta; esto involucra la decisión del número de grupos presentes.

De esta forma se han sugerido muchos criterios que son informales, puesto que dependen de algun criterio gráfico, o bien, del criterio del investigador o del problema en cuestión. Muchos de éstos métodos se explicarán en la sección 5.1.5 del cápítulo 5. Pero en esta sección se tratarán los criterios que son idóneos para los métodos jerárquicos.

Mojena (1997) ha sugerido un procedimiento basado en los tamaños relativos de los niveles de fusión en el dendograma y que también es conocido como *la regla de la cola superior (upper tail rule)*. El propósito es seleccionar el número de grupos correspondiente al primer nivel en el dendograma que satisface

$$\alpha_{j+1} > \overline{\alpha} + kS_{\alpha}$$

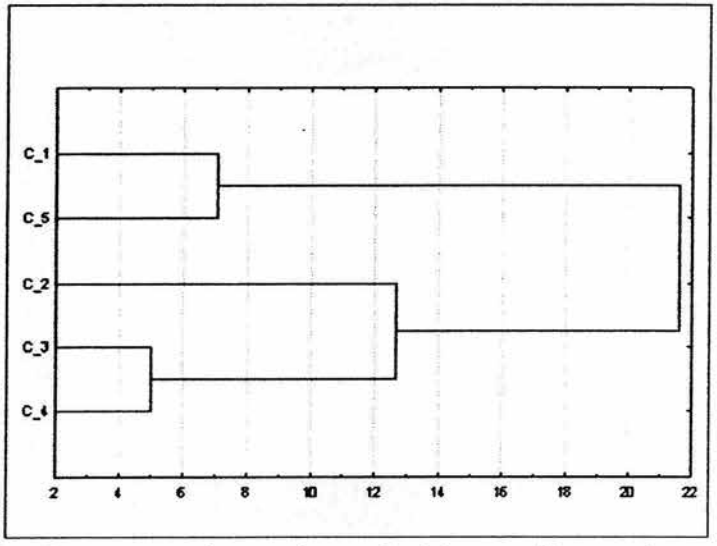
donde $\alpha_0, \alpha_1, ..., \alpha_{n-1}$ son los niveles de fusión correspondientes a $n, n-1, ..., 1$ conglomerados. Los términos $\overline{\alpha}$ y S_{α} son la media y la desviación estandar insesgada respectivamente, y k es una constante. Mojena sugirió que para valores de k en el rango (2,75, 3,50) proporcionan buenos resultados, aunque Milligan y Cooper (1985) sugirierón $k = 1{,}25$. Un criterio visual es identificar los cortes en la gráfica de los valores $(\alpha_{j+1} - \overline{\alpha})/S_{\alpha}$ contra el número de conglomerados j.

4.10. Coeficiente de Correlación Copeténica

Después de aplicar algún método de conglomeración mencionado anteriormente, es necesario saber qué tan bien un árbol o dendograma representa a la matriz de datos. Así el coeficiente de correlación $r_{X,Y}$, da una respuesta parcial. Este coeficiente compara el dendograma y la matriz de distancias. Para calcular este coeficiente es necesario convertir al dendograma en algún equivalente, esto es, una matriz que represente toda la información contenida en el dendograma, y viceversa, dada ésta matriz se puede construir el dendograma sin pérdida de información. Esta matriz es llamada la matriz copeténica (cophetenic matrix). Para ilustrar como la matriz copeténica es obtenida considérese el siguiente ejemplo número: Sea $\mathbf{X}_{(5\times 2)}$ una matriz de datos dada por

$$\mathbf{X} = \begin{pmatrix} 10 & 5 \\ 20 & 20 \\ 30 & 10 \\ 30 & 15 \\ 5 & 10 \end{pmatrix}$$

y su dendograma es



Dendograma creado por el metodo del promedio grupal.

la matriz de distancias es

$$\mathbf{D} = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & & & & & \\ 2 & 18.0 & & & & \\ 3 & 20.6 & 14.1 & & & \\ 4 & 22.4 & 11.2 & 5.0 & & \\ 5 & 7.07 & 18.0 & 25.0 & 25.5 & \end{array}$$

y la matriz copeténica es

$$\mathbb{C} = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & & & & & \\ 2 & 21.6 & & & & \\ 3 & 21.6 & 12.7 & & & \\ 4 & 21.6 & 12.7 & 5.0 & & \\ 5 & 7.07 & 21.6 & 21.6 & 21.6 & \end{array}$$

donde cada entrada de esta matriz representa la distancia en la cual fueron unidos dos elementos, por ejemplo, la entrada $C_{2,3} = 12{,}7$ es la distancia en la cual fueron unidos los elementos 2 y 3 en el dendograma anterior.

El siguiente paso es proporcionar por cada par de objetos $((X, Y)$ con $X \neq Y)$ los correspondientes valores de la matriz distancia $\mathbf{D}$ y los de la matriz copeténica $\mathbb{C}$, de la siguiente manera:

Pareja	X	Y
$(2,1)$	18.0	21.6
$(3,1)$	20.6	21.6
$(4,1)$	22.4	21.6
$(5,1)$	7.07	7.07
$(3,2)$	14.1	12.7
$(4,2)$	11.2	12.7
$(5,2)$	18.0	21.6
$(4,3)$	5.00	5.00
$(5,3)$	25.0	21.6
$(5,4)$	25.5	21.6

El siguiente paso es calcular el coeficiente de correlación de Pearson, $r_{X,Y}$, entre las listas X y Y. Para este uso particular éste coeficiente recibe el

nombre de Coeficiente de Correlación Copeténica. $r_{X,Y}$ esta dado por

$$r_{X,Y} = \frac{\sum xy - (1/n)(\sum x)(\sum y)}{\left\{\left[\sum x^2 - (1/n)\left(\sum x\right)^2\right]\left[\sum y^2 - (1/n)\left(\sum y\right)^2\right]\right\}^{1/2}} \tag{4.46}$$

Para el caso del ejemplo anterior la ecuación (4.46) resulta

$$r_{X,Y} = 0{,}93$$

Esto es un poco menor respecto a la concordancia perfecta ($r_{X,Y} = 1{,}0$); pero esta muy lejos de la discordancia ($r_{X,Y} = 0{,}0$); y esta fuera del rango de la concordancia negativa ($-1{,}0 \leq r_{X,Y} < 0{,}0$). En este ejemplo el valor de $r_{X,Y}$ representa bien a la estructura de similaridad entre objetos en la matriz de distancias.

El coeficiente $r_{X,Y}$ es un índice que indica que tanta distosion tuvo la información entrante después de aplicar el método empleado para el proceso de conglomeración para generar una salida.

Cabe destacar la hipótesis que se utiliza al usar este llamado coeficiente de correlación copeténica. Esta hipótesis esta basada en una dependencia lineal entre X y Y, la cual puede no cumplirse en la práctica. Así si se cumple ésta hipótesis, el coeficiente de correlación copeténica resulta ser una herramienta útil para saber que tan bueno es el dendograma comparado con la matriz de distancias.

4.11. Correlación de matrices

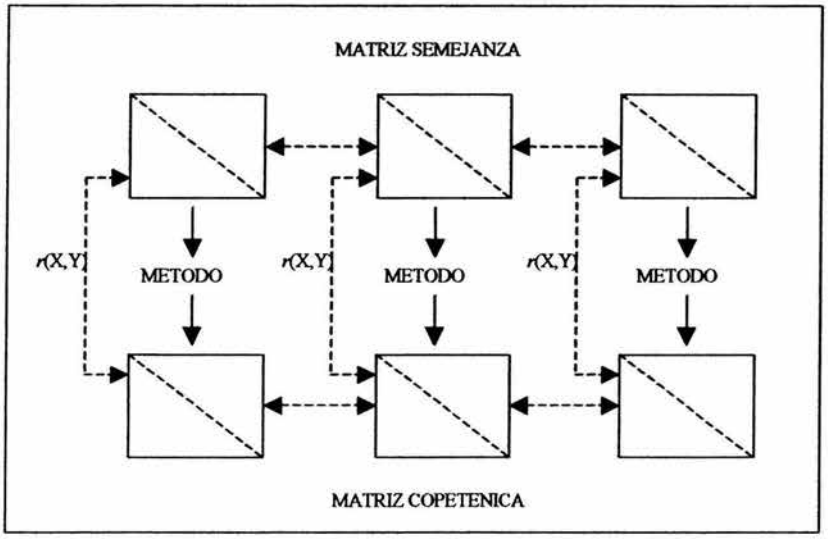
En la sección anterior se compararon dos matrices, la matriz original de datos y la matriz copeténica, ahora se hablará sobre la correlación de dos matrices en general. En esa sección el coeficiente de correlación de Pearson, $r_{X,Y}$, recibió el nombre de coeficiente de correlación copeténica para ese uso en particular. Exceptuando ese caso, a $r_{X,Y}$ se le conoce, simplemente, como coeficiente de correlación. Una vez hecha esta aclaración se iniciará con la correlación de dos matrices.

1. Supóngase que se han recolectado dos matrices provenientes del mismo conjunto de objetos. Para cada matriz de datos se calcula la matriz de semejanzas y se presenta a éstas dos matrices como listas, X y Y, entonces $r_{X,Y}$ representa un indice de concordancia respecto a los dos tipos de semejanzas.

2. Si se han recolectado dos matrices teniendo en comun un conjunto de objetos, pero cada una basada en distintos atributos, y si, además, se usa el mismo coeficiente de semejanza, el mismo método de conglomeración para cada matriz y cada matriz es representada como dos listas X y Y, entonces $r_{X,Y}$ es un indice de concordancia entre los dos diferentes tipos de atributos.

3. Dada una matriz de datos y si se calculan dos matrices de semejanza con distintos coeficientes y son representadas por listas X y Y, entonces $r_{X,Y}$ representa *un* indice de concordancia entre los dos coeficientes de semejanza.

4. Si se analiza una matriz de semejanza con dos métodos de conglomeración y las dos matrices copeténicas se representan por dos listas X y Y, entonces $r_{X,Y}$ representa un indice que indica qué tan bien los dos métodos de conglomeración dan el mismo resultado.



Posibles correlaciones entre matrices.

En conclusión, el coeficiente de correlación $r_{X,Y}$ proporciona un índice para indicar la correlación que existe entre dos matrices.

Capítulo 5

Otros métodos

5.1. Métodos de optimización

5.1.1. Introducción

Una clase de técnicas distintas a las del capítulo anterior radican en que éstas producen una partición de los individuos para un número particular de grupos, ya sea minimizando o maximizando algún criterio numérico. Estas técnicas de optimización difieren de los métodos jerárquicos en que no necesariamene forman clasificaciones por jerarquías en los datos.

5.1.2. Criterios para clasificar derivados de datos continuos

Los críterios más ampliamente usados de una matriz $\mathbf{X} \in \mathbf{M}_{n\times p}$ de datos continuos hacen uso de una descomposición de la matriz de dispersión total $\mathbf{T}$ dada por

$$\mathbf{T} = \sum_{m=1}^{g}\sum_{l=1}^{n_m} \left(X_{ml} - \overline{X}\right)\left(X_{ml} - \overline{X}\right)^T \tag{5.1}$$

donde X_{ml} es el vector p-dimensional del l-ésimo objeto en el grupo m y $\overline{X}$ es el vector media de la muestra.

Se definen la dispersión dentro de los grupos $\mathbf{W}$ y la dispersión entre

grupos **B** como

$$\mathbf{W} = \sum_{m=1}^{g} \sum_{l=1}^{n_m} \left(X_{ml} - \overline{X}_m\right) \left(X_{ml} - \overline{X}_m\right)^T \tag{5.2}$$

y

$$\mathbf{B} = \sum_{m=1}^{g} n_m \left(\overline{X}_m - \overline{X}\right) \left(\overline{X}_m - \overline{X}\right)^T \tag{5.3}$$

donde $\overline{X}_m$ es la media del m-ésimo conglomerado. De lo anterior se tiene la siguiente proposición.

Proposición 5.1 . $\mathbf{T} = \mathbf{W} + \mathbf{B}$

Demostración.

$$\begin{aligned}
\mathbf{T} &= \sum_{m=1}^{g} \sum_{l=1}^{n_m} \left(X_{ml} - \overline{X}\right) \left(X_{ml} - \overline{X}\right)^T \\
&= \sum_{m=1}^{g} \sum_{l=1}^{n_m} \left(\left(X_{ml} - \overline{X}_m\right) + \left(\overline{X}_m - \overline{X}\right)\right) \left(\left(X_{ml} - \overline{X}_m\right) + \left(\overline{X}_m - \overline{X}\right)\right)^T \\
&= \sum_{m=1}^{g} \sum_{l=1}^{n_m} \left(X_{ml} - \overline{X}_m\right) \left(X_{ml} - \overline{X}_m\right)^T + \sum_{m=1}^{g} \sum_{l=1}^{n_m} \left(X_{ml} - \overline{X}_m\right) \left(\overline{X}_m - \overline{X}\right)^T \\
&\quad + \sum_{m=1}^{g} \sum_{l=1}^{n_m} \left(\overline{X}_m - \overline{X}\right) \left(X_{ml} - \overline{X}_m\right)^T + \sum_{m=1}^{g} \sum_{l=1}^{n_m} \left(\overline{X}_m - \overline{X}\right) \left(\overline{X}_m - \overline{X}\right)^T \\
&= \mathbf{W} + \sum_{m=1}^{g} \sum_{l=1}^{n_m} \left(X_{ml}\overline{X}_m^T - X_{ml}\overline{X}_m^T - \overline{X}_m\overline{X}^T + \overline{X}_m\overline{X}^T\right) \\
&\quad + \sum_{m=1}^{g} \sum_{l=1}^{n_m} \left(\overline{X}_m X_{ml}^T - \overline{X}_m X_{ml}^T - X_{ml}X_{ml}^T + X_{ml}\overline{X}_m^T\right) \\
&\quad + \sum_{m=1}^{g} n_m \left(\overline{X}_m - \overline{X}\right) \left(\overline{X}_m - \overline{X}\right)^T \\
&= \mathbf{W} + \sum_{m=1}^{g} \left[\left(n_m\overline{X}_m\overline{X}_m^T - n_m\overline{X}_m\overline{X}_m^T\right) + \left(n_m\overline{X}_m\overline{X}^T - n_m\overline{X}_m\overline{X}^T\right)\right] \\
&\quad + \sum_{m=1}^{g} \left[\left(n_m\overline{X}_m\overline{X}_m^T - n_m\overline{X}_m\overline{X}_m^T\right) + \left(n_m\overline{X}\overline{X}_m^T - n_m\overline{X}\overline{X}_m^T\right)\right] + \mathbf{B} \\
&= \mathbf{W} + \sum_{m=1}^{g} 0 + \sum_{m=1}^{g} 0 + \mathbf{B} \\
&= \mathbf{W} + \mathbf{B}
\end{aligned}$$

■

Para $p = 1$ la proposición (5.1) representa una relación entre escalares, esto es, la suma total de los cuadrados para cada variable puede ser dividida entre la suma de los cuadrados dentro de los conglomerados y la suma de los cuadrados entre los conglomerados. En este caso un criterio natural podría ser elegir la partición correspondiente al mínimo valor de la suma de cuadrados dentro de los conglomerados.

Minimización de $tr(\mathbf{W})$

Una extensión obvia del criterio anterior sería la minimización de la suma de los cuadrados dentro de los conglomerados aplicada a todas las variables, esto es minimizar la $tr(\mathbf{W})$.

Sin embargo, a pesar de ser el criterio más usado, este método depende de la escala, ya que diferentes soluciones se pueden obtener de la matriz de datos originales y de la matriz estandarizada. Claramente esto es muy importante debido a la necesidad de la estandarización en muchas aplicaciones. Otro problema derivado del uso de este criterio es la imposición de estructuras esféricas en los conglomerados observados aún cuando los conglomerados "naturales" de los datos sean de otra forma, entiéndiéndose como conglomerados naturales aquellos conglomerados que uno esperaría que resultarán después de aplicar el proceso de agrupamiento, en otras palabras, los conglomerados que conservan una forma más idónea, no necesariamente esférica

Minimización del $\det(\mathbf{W})$

Un criterio derivado del Análisis Multivariado de la Varianza (MANOVA por sus siglas en inglés) derivado de una de las pruebas para diferencias en los vectores media de los conglomerados, basada en la razón de los determinantes de la dispersión total y la dispersión entre conglomerados, a saber $\det(\mathbf{T})/\det(\mathbf{W})$. Grandes valores de esta razón indican que los vectores media de cada conglomerado difieren, así Friedman y Rubin proponen minimizar el $\det(\mathbf{W})$.

Maximización de $tr(\mathbf{BW}^{-1})$

Un criterio más amplio sugerido por Friedman y Rubin es el maximizar la $tr(\mathbf{BW}^{-1})$. Esta función también es usada en el contexto del Análisis Multivariado de la Varianza. Grandes valores de $tr(\mathbf{BW}^{-1})$ indican que los vectores media de cada conglomerado difieren.

5.1.3. Criterios alternativos para clasificar conglomerados de distintos tamaños y formas

Como un intento de superar el problema de las formas similares que padece el criterio del $\det(\mathbf{W})$, se han sugerido otros criterios basados en la minimización de

$$\prod_{g=1}^{n_m} [\det(\mathbf{W}_m)]^{n_m} \tag{5.4}$$

donde $\mathbf{W}_m$ es la matriz de disperciones dentro del m-ésimo grupo

$$\mathbf{W}_m = \sum_{l=1}^{n_m} (X_{ml} - \overline{X}_m)(X_{ml} - \overline{X}_m)^T \tag{5.5}$$

y n_m es el número de individuos en el m-ésimo grupo. (Este método está restringido a soluciones de conglomerados donde cada conglomerado contiene al menos $p+1$ individuos, esto es, para evitar matrices de dispersiones singulares, lo que implicaría que sus determinantes serían iguales a cero). Un criterio alternativo es la minimización de

$$\sum_{m=1}^{g} (n-1)[\det(\mathbf{W}_m)]^{1/p}. \tag{5.6}$$

Un intento para solucionar criterios que proporcionen soluciones de conglomerados de distintos tamaños ha sido tratado por diversos autores. Entre los métodos propuestos se tiene la minimización de

$$\prod_{m=1}^{g} [\det(\mathbf{W}/n_m^2)]^{n_m} \tag{5.7}$$

(una pequeña modificación al criterio del determinante) y

$$\prod_{m=1}^{g} [\det(\mathbf{W}_m/n_m^2)]^{n_m}$$

(una modificación del criterio (5.4)).

Todos los criterios mencionados son esencialmente adecuados para datos donde todas sus variables son medidas en una escala continua. Cuando las variables no son continuas, se puede calcular una matriz de disimilaridades y usar un criterio de clasificación que opere en base con la matriz de disimilaridades.

5.1.4. Algoritmos de Optimización

Después de haber elegido un criterio de clasificación es necesario tomar consideraciones acerca de qué partición en g grupos optimiza el criterio. Puede haber, por supuesto, más de una partición que optimice el criterio. Sin embargo prácticamente calcular todas las posibles particiones, aún para las computadoras modernas, podría ser imposible para un tamaño de muestra considerablemente grande. Debido a este problema se han desarrollado diferentes técnicas cuyos pasos esenciales son:

- Encontrar una partición inicial de n objetos en g grupos.
- Calcular el cambio en el criterio de clasificación producido al mover cada objeto de su conglomerado original a cualquier otro conglomerado.
- Hacer el cambio que produzca la mejora óptima en el valor del criterio de clasificación.
- Repetir los dos pasos anteriores hasta que ningún cambio de los objetos produzca una mejora al criterio de clasificación.

Una partición inicial puede ser encontrada de diferentes formas. Por ejemplo, podría estar basada a un conocimiento anterior; o podría ser el resultado de un proceso de clasificación aplicado anteriomente. También se podría elegir una partición inicial aleatoriamente o cuando los objetos son representados como puntos en un espacio Euclidiano, g puntos se pueden elegir al azar para actuar como centros de los conglomerados. Sin embargo, los resultados del algoritmo de optimización pueden ser gravemente afectadas por la elección de la partición inicial, debido a que diferentes particiones iniciales pueden llevar a diferentes óptimos del criterio de clasificación. Por estas razones se recomienda correr cualquier algoritmo se optimización muchas veces usando diferentes particiones iniciales.

Algoritmos de k-medias El algorimo de k-medias es uno de los últimos algoritmos que consiste en la actualización constante de las particiones acomodando cada objeto al grupo cuya media este más cercana al objeto y recalculando las medias grupales. Tales algoritmos que involcran el cálculo de las medias (o centroides) de cada conglomerado son usualmente conocidos como algoritmos de k-medias.

Ahora considérese un ejemplo de estos algoritmos de k-medias. Para esto, tomemos los siguientes datos de siete individuos con dos variables.

Individuo	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Estos datos van a ser clasificados en dos conglomerados usando el criterio de la minimización de la $tr(\mathbf{W})$. Como primer paso para encontrar una partición inicial considérense a los dos individuos que están más alejados (usando la métrica Euclidiana), a saber los indivuduos 1 y 2, así se definen los conglomerados iniciales, dados por

	Individuo	Vector media
Grupo 1	1	(1,0,1,0)
Grupo 2	4	(5,0,7,0)

Los individuos restantes son examinados y asignados al grupo que esté más cercano, en términos de la distancia Euclidiana, a la media grupal. El vector media es recalculado cada vez que un nuevo miembro es agregado. Esto lleva a una serie de pasos:

	GRUPO 1		GRUPO 2	
	Individuo	Vector media	Individuo	Vector media
Paso 1	1	(1,0,1,0)	4	(5,0,7,0)
Paso 2	1,2	(1,3,1,5)	4	(5,0,7,0)
Paso 3	1,2,3	(1,8,2,3)	4	(5,0,7,0)
Paso 4	1,2,3	(1,8,2,3)	4,5	(4,3,6,0)
Paso 5	1,2,3	(1,8,2,3)	4,5,6	(4,3,5,7)
Paso 6	1,2,3	(1,8,2,3)	4,5,6,7	(4,1,5,4)

Esto da la partición inicial; los dos grupos tienen las siguientes dos características:

	Individuo	*Vector media*	$Tr(\mathbf{W}_j)$
Grupo 1	1,2,3	(1,8,2,3)	6.84
Grupo 2	4,5,6,7	(4,1,5,4)	5.38

Hasta este punto se tiene $Tr(\mathbf{W}) = 6{,}84 + 5{,}38 = 12{,}22$. La primera reagrupación del algoritmo de k-medias compara la distancia de cada individuo a su propia media y a la media del grupo opuesto. Se encuentra que:

Individuo	*Distancia al vector media del grupo 1*	*Distancia al vector media del grupo 2*
1	1.5	5.4
2	0.4	4.3
3	2.1	1.8
4	5.7	1.8
5	3.2	0.7
6	3.8	0.6
7	2.8	1.1

Solamente el individuo 3 está más cerca a la media del grupo opuesto (grupo 2) que a la de su propio grupo. Así, este individuo es reagrupado al grupo 2 resultando la nueva partición

	Individuo	*Vector media*	$tr(\mathbf{W}_j)$
Grupo 1	1,2	(1,3, 1,5)	0.63
Grupo 2	3,4,5,6,7	(3,9, 5,1)	7.9

Este movimiento causa una disminución en el criterio de la $tr(\mathbf{W}) = 0{,}63 + 7{,}9 = 8{,}53$. La reagrupación podría continuar con esta nueva partición, pero en este ejemplo cada individuo está cerca de su media grupal, así se elige la última partición como la solución final de clasificación.

5.1.5. Eligiendo el número de conglomerados

En muchas aplicaciones de los métodos de optimización, el investigador tiene que estimar el número de conglomerados en los datos. Una gran variedad de métodos han sido sugeridos que pueden ayudar al investigador en situaciones particulares. Muchos de éstos criterios son informales e involucran, esencialmente, el gráficar el valor del criterio de clasificación contra el número de grupos. Largos cambios en los niveles de la gráfica son usados para sugerir el número de grupos.

Otros criterios más formales se han sugerido que tratan de superar el problema de la subjetividad. Uno de estos criterios para datos continuos

para sugerir el número de conglomerados a elegir, corresponden a tomar el maximo valor de $C(g)$, donde

$$C(g) = \frac{tr(\mathbf{B})}{g-1} \Big/ \frac{tr(\mathbf{W})}{n-g}.$$

Como todas las técnicas para determinar el número de conglomerados, la evaluación de este criterio necesita conocimiento acerca del número de miembros del conglomerado para determinar las matrices **B** y **W**.

Otro procedimiento es seleccionar g de tal manera que $g^2 \det(W)$ sea mínimo. Él cual, también sugiere que los datos se consideren como un solo conglomerado si

$$\frac{g^2 \det(W)}{\det(T)} > 1$$

para todos los valores de g.

5.2. Mezclas de Densidades

Considérese a la familia de funciones de probabilidad de la forma

$$f(\underline{X}; \underline{P}, \underline{\theta}) = \sum_{j=1}^{g} p_j f_j(\underline{X}, \underline{\theta}_j) \tag{5.8}$$

donde $\underline{X}$ es una variable aleatoria de dimensión p, $\underline{P}^T = (p_1, ..., p_g)$ y $\underline{\theta} = (\underline{\theta}_1^T, ..., \underline{\theta}_g^T)$. p_j se conoce como la proporción de la mezcla y f_i, $j = 1, ..., g$, es la componente de la densidad, f_j esta parametrizada por θ_j. Las proporciones de la mezcla son no-negativas y satisfacen

$$\sum_{j=1}^{g} p_j = 1$$

y g es el número de componentes en la mezcla.

Éstas mezclas son modelos idóneos para el análisis de conglomerados si se supone que cada grupo de observaciones contiene conglomerados que provienen de distintas distribuciones de probabilidad. Los conglomerados pueden pertenecer a la misma familia, pero difieren en los parámetros de la distribución. Para entender mejor el concepto de mezclas de densidades

considérese el siguiente ejemplo: se quiere estudiar una muestra de individuos de alguna población midiendo las alturas de dichos individuos. Dicha muestra puede contener hombres y mujeres y la altura promedio (la de hombres y la de mujeres) es conocida y diferente. Ahora, la función de densidad de la altura tendrá la siguiente forma:

$$h(altura) = p(M)h_1(altura \mid M) + p(H)h_2(altura \mid H)$$

donde $p(M)$ y $p(H)$ son la probabilidad de que un miembro de la población sea mujer u hombre, respectivamente y h_1, h_2 son las funciones de densidad para mujer y hombre, respectivamente. Asi, la función distancia ha sido expresada como una combinación lineal de dos funciones condicionales.

Una vez que se hayan estimado los parametros de la mezcla, las observaciones pueden ser asociadas con conglomerados particulares por medio del máximo valor de

$$\mathbb{P}(C_j \mid \underline{X}_i) = \frac{\widehat{p}_j f_j(\underline{X}_i; \widehat{\underline{\theta}}_j)}{f(\underline{X}_j; \widehat{\underline{P}}, \widehat{\underline{\theta}})}, \qquad j = 1, 2, ..., g \tag{5.9}$$

5.2.1. Estimadores de Máxima Verosimilitud

Sean $\underline{X}_1, ..., \underline{X}_n$ una muestra de observaciones de una mezcla de densidades como en (5.8). La función log-máxima verosimilitud, l, es

$$l = \sum_{i=1}^{n} f(\underline{X}_i; \underline{P}, \underline{\theta}) \tag{5.10}$$

Los estimadores de los parámetros se obtienen como el las soluciones de las ecuaciones de máxima verosimilitud,

$$\frac{\partial l(\underline{\phi})}{\partial \underline{\phi}} = 0$$

donde $\underline{\phi}^T = (\underline{P}^T, \underline{\theta}^T)$.

5.2.2. Mezclas de densidades normales y la estimación de sus parámetros

Sean

$$f(\underline{X}) = \sum_{i=1}^{g} p_i \alpha(\underline{X}; \underline{\mu}_i, \mathbf{\Sigma}_i) \tag{5.11}$$

donde

$$0 \leq p_i \leq 1, \qquad \sum_{i=1}^{g} p_i = 1$$

y

$$\alpha(\underline{X}; \underline{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{p/2}} |\boldsymbol{\Sigma}_i|^{-1/2} \exp\left(-\frac{1}{2}(\underline{X} - \underline{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\underline{X} - \underline{\mu}_i)\right)$$

entonces (5.11) representa un modelo cuando los datos considerados contienen g conglomerados y las variables de cada conglomerado tienen una densidad normal multivariada.

La estimación de los parámetros es por máxima verosimilitud. La función log-máxima verosimilitud de (5.11), está dada por

$$L = \sum_{i=1}^{n} \ln \left[\sum_{j=1}^{g} p_j \alpha(\underline{X}_i; , \underline{\mu}_i, \boldsymbol{\Sigma}_i) \right] \tag{5.12}$$

Al resolver las ecuaciones de máxima verosimilitud de (5.12) se obtienen los estimadores

$$\widehat{p}_i = \frac{1}{n} \sum_{j=1}^{n} \widehat{p}(i \mid \underline{X}_j), \qquad i = 1, 2, ..., g-1 \tag{5.13}$$

$$\widehat{\mu}_i = \frac{1}{n\widehat{p}_i} \sum_{j=1}^{n} \widehat{p}(i \mid \underline{X}_j) \underline{X}_j, \qquad i = 1, 2, ..., g \tag{5.14}$$

$$\boldsymbol{\Sigma}_i = \frac{1}{n\widehat{p}_i} \sum_{j=1}^{n} \widehat{p}(i \mid \underline{X}_j)(\underline{X}_i - \underline{\mu}_i)(\underline{X}_i - \underline{\mu}_i)^T, \qquad i = 1, ..., g. \tag{5.15}$$

Las ecuaciones (5.13), (5.14) y (5.15) no dan las estimaciones de los parámetros explícitamente, por lo que deben de ser resueltos por medio de algun procedimiento iterativo.

5.2.3. Mezclas para Datos Categóricos- Análisis Latente de Clases

El modelo anterior para componentes Gausianas no es adecuado para conjuntos de datos donde las variables son categóricas. Para proveer un modelo idóneo para éste tipo de conjuntos, es necesario usar otro tipo de densidades

en la mezcla en lugar de las Gausianas. Las densidades más frecuentemente usadas son las densidades multivariadas Bernoulli, las cuales suponen que dentro de cada grupo, las variables categóricas son independientes entre sí. Se supone que hay g conglomerados en los datos y que en el i-ésima conglomerado, el vector $\underline{\theta}_i$ proporciona la probabilidad de la siguiente forma:

$$\mathbb{P}(X_{i,j} = 1 \mid C_i) = \theta_{i,j}$$

donde $X_{i,j}$ es el valor de la j-ésima variable del i-ésimo objeto, y C_i es el i-ésimo conglomerado. Así, de la hipótesis de independencia se tiene que la probabilidad de que una observación este en el conglomerado i es

$$\mathbb{P}(\underline{X} \mid C_i) = \prod_{j=1}^{p} \theta_{i,j}^{X_{i,j}} (1 - \theta_{i,j})^{1-X_{i,j}}$$

Entonces si $p_1, ..., p_g$ son las proporciones de cada conglomerado en la población, la probabilidad de la observación $\underline{X}$, esta dada por la mezcla

$$\mathbb{P}(\underline{X}) = \sum_{i=1}^{g} p_i \prod_{j=1}^{p} \theta_{i,j}^{X_{i,j}} (1 - \theta_{i,j})^{1-X_{i,j}}.$$

Capítulo 6

Aplicaciones

6.1. Introducción

Este último capítulo esta diseñado para mostrar cuatro distintas aplicaciones del análisis de conglomerados en diversas ramas, como son ingeniería, demografía-muestreo, finanzas y botánica. Estas aplicaciones explican cómo elaborar un análisis de este tipo en la práctica, cada una con una pequeña conclusión.

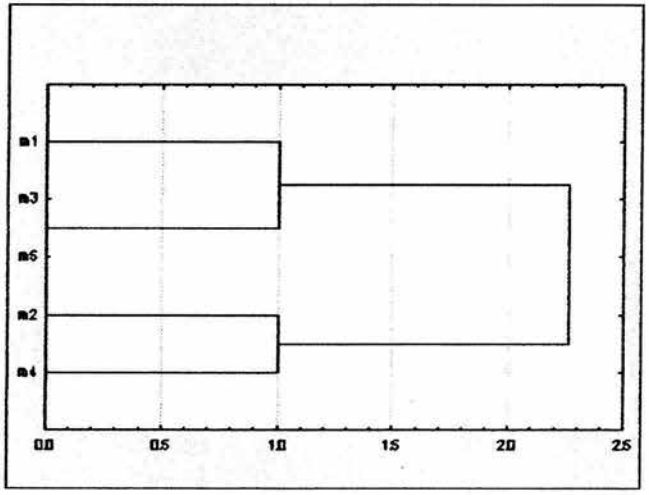
6.2. Ingeniería

Esta aplicación ilustra con un ejemplo un poco teórico el uso de los análisis Q y R, es decir, un ejemplo que requiere el uso de un análsis Q (o de objetos) y un análisis R (o de atributos) para resolver el problema. Para esto supóngase que una fábrica produce n componentes cada una requiere pasar de máquina en máquina con un total de p máquinas. Ahora, el problema es evitar rutinas innecesarias de los componentes de máquina en máquina, o sea, ¿Cómo se pueden acomodar las máquinas eficientemente de manera que se evite la pérdida de tiempo?. Para resolver este problema, identifíquese al 1 como un paso necesario que tiene que hacer un componente por una máquina y al 0 si no es necesario pasar por dicha máquina. Así la siguiente figura tiene

significado,

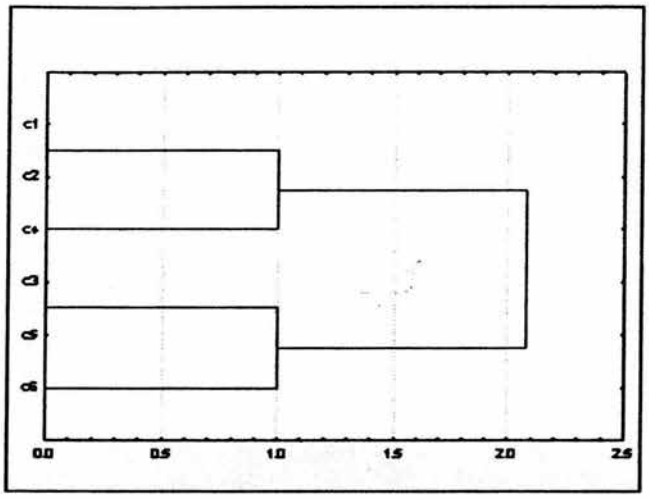
	$m1$	$m2$	$m3$	$m4$	$m5$
$c1$	1	0	1	0	1
$c2$	1	0	1	0	1
$c3$	0	1	0	1	0
$c4$	0	0	1	0	1
$c5$	0	1	0	1	0
$c6$	0	0	0	1	0

donde c_i representa el i-ésimo componente y m_j representa la j-ésima máquina. Por ejemplo, el componente c_1 sólo debe pasar por las máquinas m_1, m_3 y m_5. Se efectuarón, con la ayuda de la distancia euclidiana, dos análisis de conglomerados, un análisis Q para los componentes y un análisis R para las máquinas. Resultando los siguientes dendogramas



ANALISIS Q PARA CINCO MAQUINAS

y



ANALISIS $\mathcal{R}$ PARA SEIS COMPONENTES

y reordenando los datos se obtiene

	$m4$	$m2$	$m5$	$m3$	$m1$
$c3$	1	1	0	0	0
$c5$	1	1	0	0	0
$c6$	1	0	0	0	0
$c2$	0	0	1	1	1
$c1$	0	0	1	1	1
$c4$	0	0	1	1	1

Así las máquinas (m_4, m_2) y (m_5, m_3, m_1) deben estar juntas, así como los componentes (c_3, c_5, c_6) y (c_2, c_1, c_4).

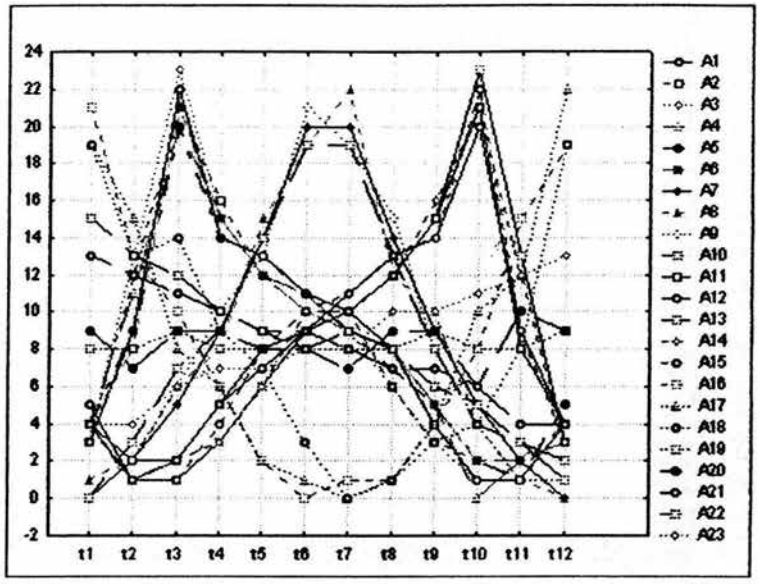
Debido a que en este ejemplo no se requiere visualizar el número de conglomerados con cualquier técnica mencionada en el capítulo 3, se procedió con la elección del coeficiente de la distancia Euclidiana y el método UPGMA (o método del promedio grupal). Debido a que al usar los coeficientes simple, cuadrado de la distancia Euclidiana, Manhatan con el método UPGMA propocionan los mismos resultados, no fueron incluidos. De igual forma, al utilizar métodos del ligamiento simple y completo con los coeficiente simple, Jaccard, Rusell y Rao, Rogers y Tanimoto y los métodos del centroide y de Ward con el cuadrado de la distancia Euclidiana.

6.3. Finanzas

Para ilustrar el gran potencial de esta técnica en cuando a la reducción de datos considérense 23 curvas hipotéticas que representan a 23 acciones a lo largo de 12 puntos en el tiempo (por ejemplo, días,meses, etc.). A pesar de ser un ejemplo hipotético sirve para representar una considerable reducción de datos y este principio puede ser usado para aplicaciones reales. Las características se han tomado de forma tal que $\sum_{j=1}^{12} X_{i,j} = 100$ para toda $i = 1, ..., 23$. Los datos de las curvas estan dadas por la siguiente matriz:

t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12
3	9	22	14	13	11	10	8	4	1	1	4
3	12	21	16	12	10	10	6	3	2	1	4
3	13	23	15	12	10	9	6	3	1	1	4
4	11	20	15	12	11	9	8	5	0	2	3
4	9	20	14	13	11	9	7	5	2	1	5
4	8	21	15	12	10	9	8	5	2	2	4
0	2	5	9	14	20	20	14	9	5	2	0
1	2	6	8	15	19	22	12	10	4	1	0
0	2	6	10	13	21	19	15	8	4	1	1
0	3	7	9	14	19	19	13	8	4	3	1
4	2	2	5	8	9	10	12	15	21	8	4
5	1	2	5	7	9	11	13	14	20	9	4
4	1	1	3	6	9	10	12	15	23	13	3
4	1	2	3	6	10	10	12	16	21	12	3
4	1	1	4	8	10	11	13	14	22	9	3
21	13	10	6	2	0	1	1	4	8	15	19
19	15	8	6	2	1	0	1	4	10	12	22
19	13	14	9	7	3	0	1	3	4	8	19
8	8	9	8	8	9	8	8	9	8	8	9
9	7	9	9	8	8	7	9	9	6	10	9
13	12	11	10	9	9	8	7	7	6	4	8
15	13	12	10	9	8	9	8	6	5	3	2
4	4	6	7	7	8	8	10	10	11	12	13

donde **ti** representa la acción en el tiempo i, $i = 1, ..., 12$, o bien, gráficamente

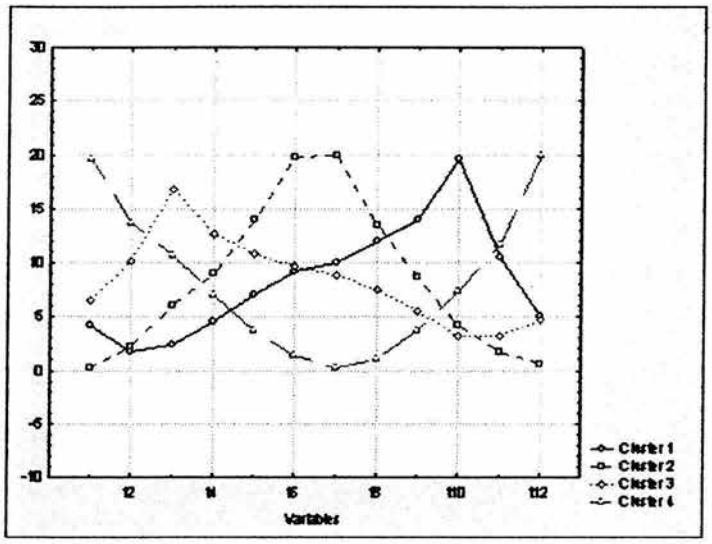


23 ACCIONES

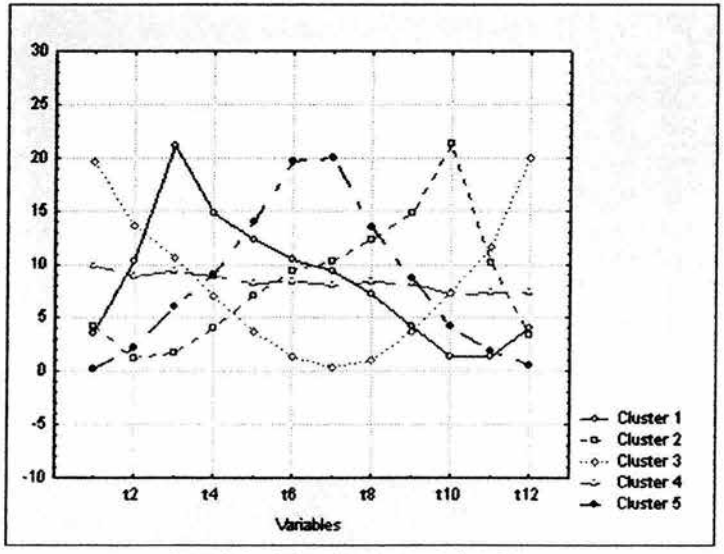
Frecuentemente cuando se analizan acciones, es imposible trabajar con todas, puesto que los valores cambian constantemente, de esta manera, el análisis de conglomerados es una alternativa para reducir los datos involucrados y asi concentrar esfuerzos para analizar unas pocas.

Tomando en consideración el algoritmo de k medias se puede reducir notablemente los datos paulatinamente y asi la reducción de los mismos, por ejemplo si se consideraran 4,5,6 y 7 conglomerados en el análisis, al graficar

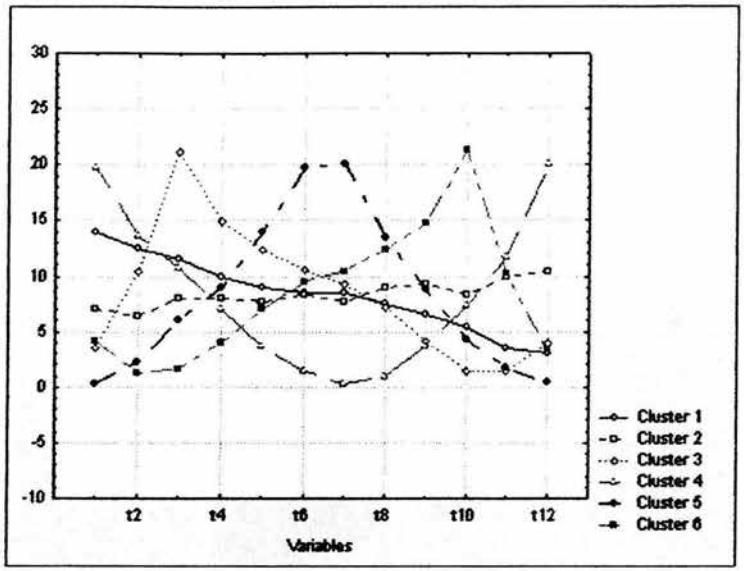
sus respectivas medias se obtienen



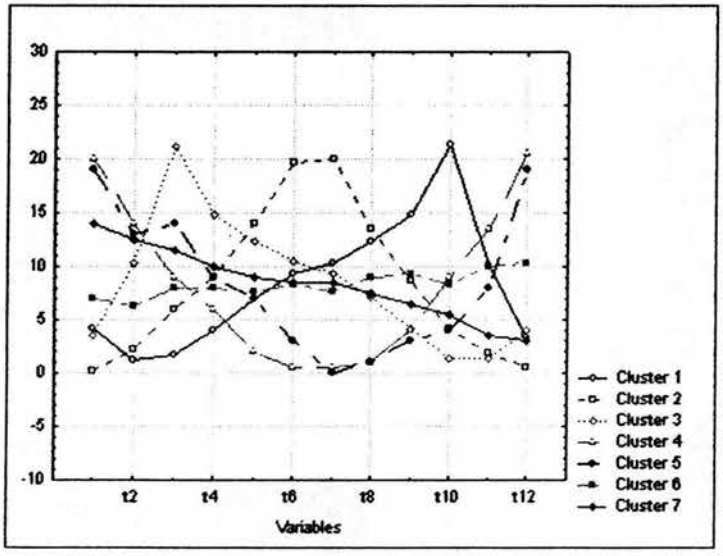
CUATRO CONGLOMERADOS



CINCO CONGLOMERADOS



SEIS CONGLOMERADOS

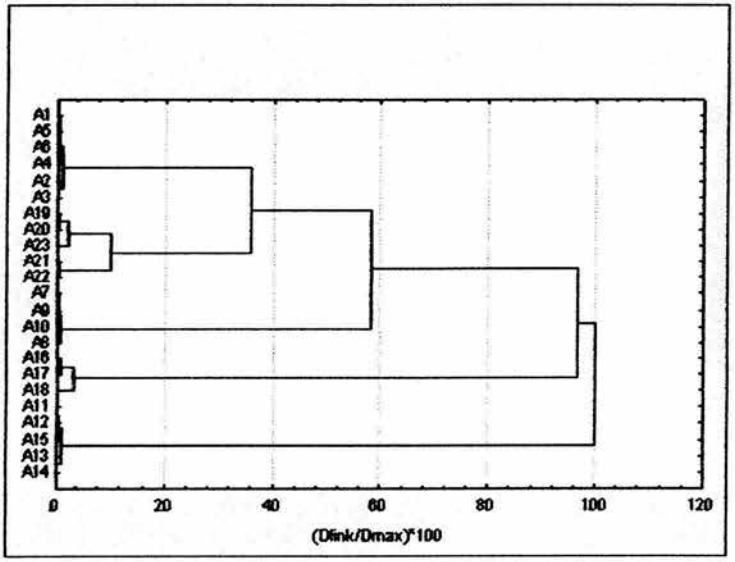


SIETE CONGLOMERADOS

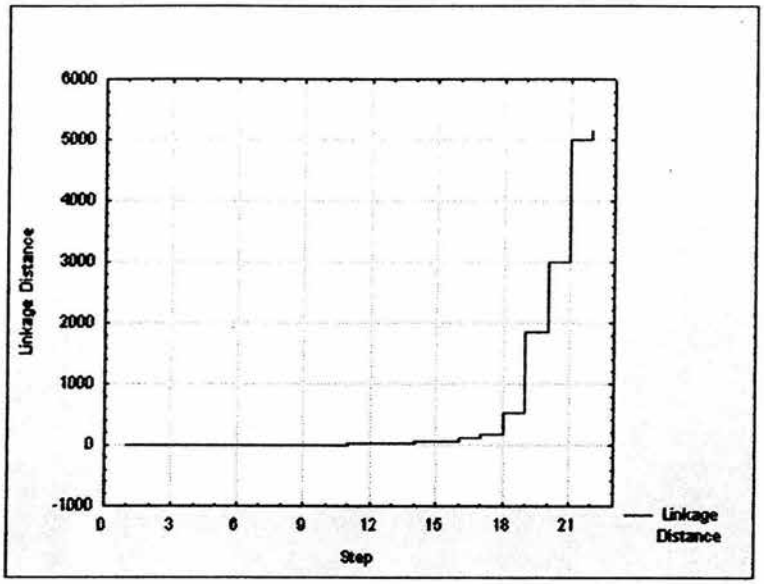
De esta manera se puede recopilar información de todos los datos, en este

caso curvas, y utilizarla para un análisis más profundo.

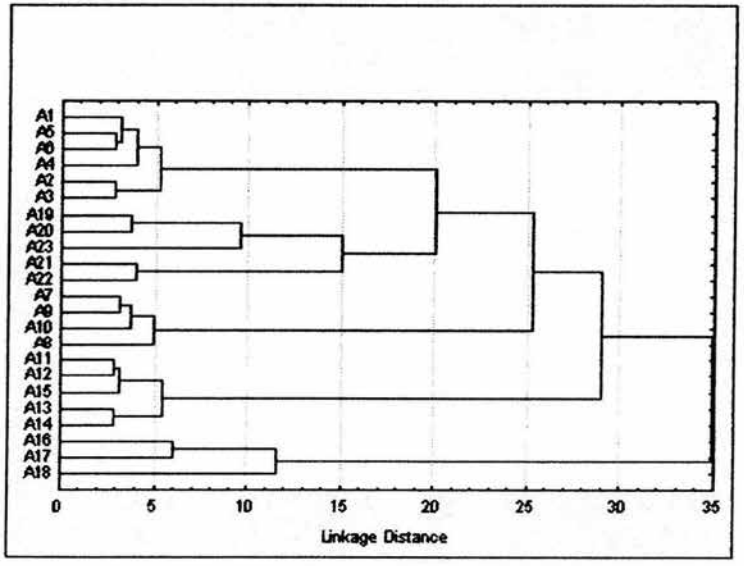
Para este ejemplo en particular, después de elaborar un análisis con los métodos de Ward y UPGMA, se decidio cortar el dendograma en 5 conglomerados debido a cambios bruscos en las distancias de ligamiento, los cuales se ven reflejados en los dendogramas y en las graficas de ligamiento



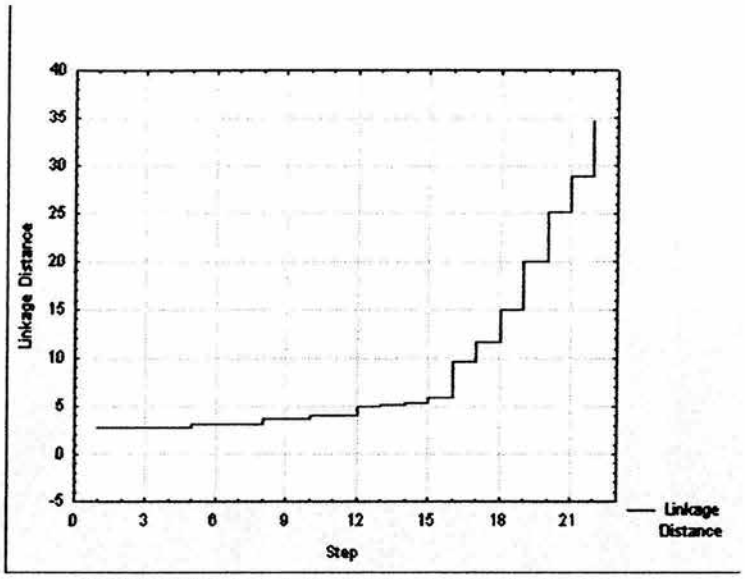
METODO DE WARD PARA 23 ACCIONES



LIGAMIENTO DEL METODO DE WARD



PROMEDIO GRUPAL DE LAS 23 ACCIONES



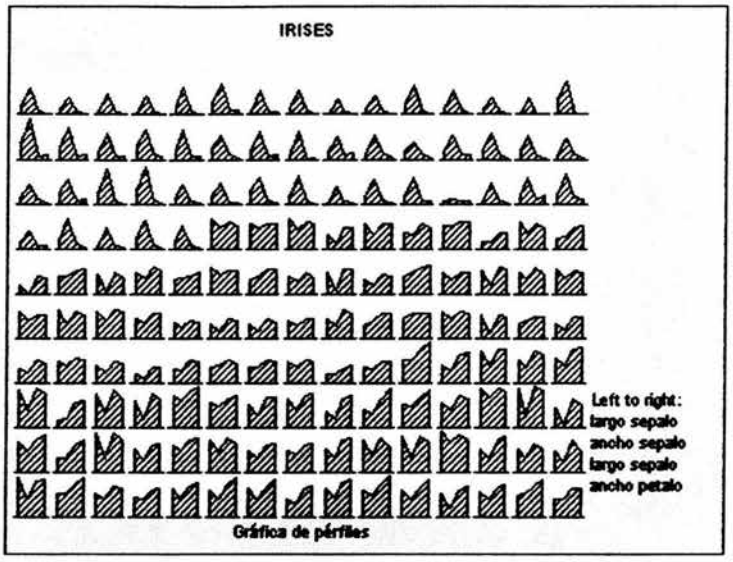
LIGAMIENTO DEL PROMEDIO GRUPAL

y después de ejecutar ambos análisis, se decidió optimizar los conglomerados con la ayuda del algoritmo de k medias.

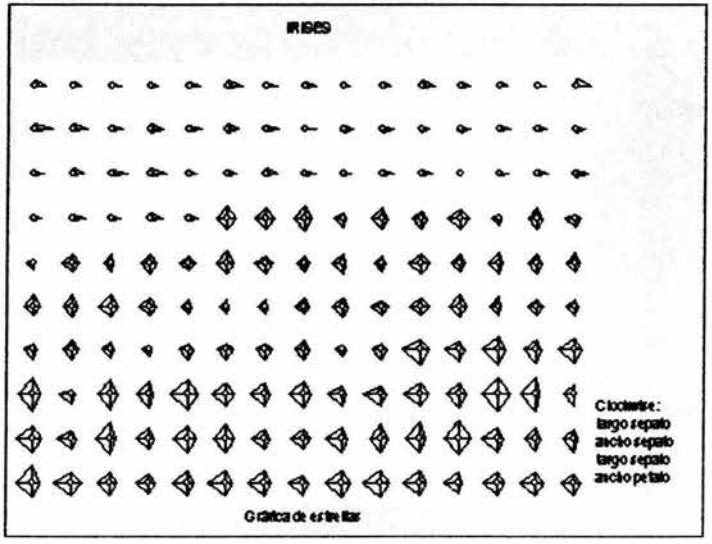
6.4. Botánica

Este es un ejemplo clásico, y fue propuesto por Fisher. Se trata de una clasificación de irises. Los datos de las 150 irises se encuentran en el apéndice B. Como primer paso se visualizarán los conglomerados. Las gráficas usadas para visualizar los conglomerados son gráficas de pérfiles (ver la subsección de Pérfiles de datos, cap 4), así como gráficas de estrellas y líneas (cápitulo

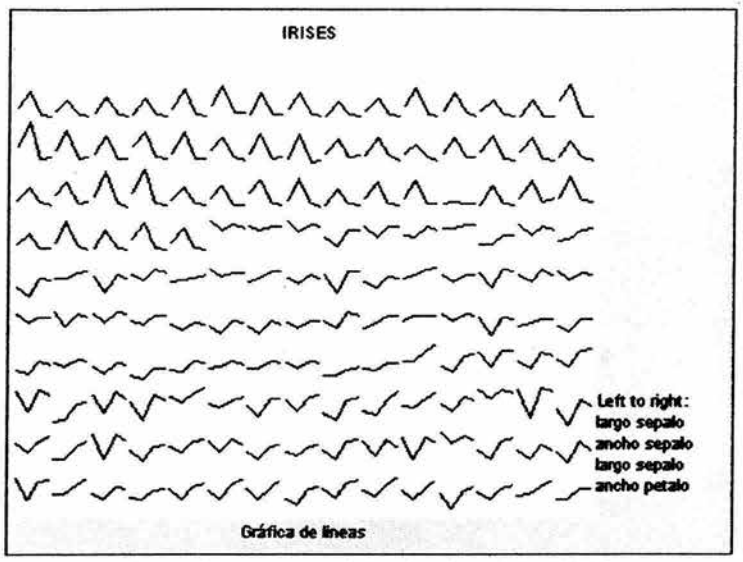
3)



GRAFICA DE PERFILES PARA 150 IRISES



GRAFICA DE ESTRELLAS PARA 150 IRISES



GRAFICA DE LINEAS PARA 150 IRISES

De esta manera se pueden observar al menos 3 patrones en los datos, esto es, 3 conglomerados.

El uso de estas gráficas es debido a lo que interesa es la forma de las irises, lo cual se verá reflejado en el uso de los coeficientes de semejanza. Por consiguiente, se han usado los coeficientes del coseno y el coeficiente de correlación de Pearson al método del ligamiento completo, asi como el uso del método de Ward y de un algoritmo de optimización. Los resultados son los siguientes:

- **COMPLETO-COSENO.** Concordancia del **84 %**.
- **COMPLETO-PEARSON.** Brindó el **85 %** de la clasificación original.
- **WARD.** Este método concordó con el **90 %** de la clasificación original.
- **ALGORITMO DE K-MEDIAS.** Este método proporcionó el **89 %** de los datos agrupados correctamente.

El uso de la estandarización no es necesaria, pues todos los datos de las columnas (largo sepalo, ancho sepalo, largo petalo y ancho petalo) afectan

a la similaridad total casi de igual forma, en otras palabras, los rangos de las 4 columnas son muy parecidos, lo que implica que la similaridad no está "dominada"por alguna columna.

Para mostrar que la estandarización es innecesaria se aplicaron las estandarizaciones por rango en 0 y 1 y la desviación estandar (DE) a 1 al método del ligamiento completo. Los resultados obtenidos fueron los siguientes:

- **COMPLETO-DES-COSENO. 75 %** de concordancia.
- **COMPLETO-RANGO-COSENO. 60 %** de concondancia respecto a la clasificación original.
- **COMPLETO-DE-PEARSON.** Brindó el **69 %.**
- **COMPLETO-RANGO-PEARSON.** Este método concordó con el **74 %.**

Excepciones

- **WARD-DE. 83 %.**
- **WARD-RANGO.** El **89 %.**

El apéndice B muestra la clasificación real de los irises (iris setosa, versicolor y virginica) y dos modelos ajustados a los datos, a saber, el método de ward y el algoritmo de k-medias.

Conclusión. Como se vio, el uso de los coeficientes de semejanza "especiales"para los pérfiles de los objetos brindaron muy buenos resultados, con muy buena compatibilidad en la clasificación ya existente. Al estandarizar los datos no se obtienen resultados contundentes respecto a la clasificación, pero existen algunas excepciones. Desafortunadamente, no se puede saber con exactitud que métodos brindan mejores resultados, o si es necesario estandarizar los datos, y si es necesario estandarizarlos cuál es la mejor función estandarizadora y que coeficientes de semejanza usar.

6.5. Demografía-muestreo

El objetivo de esta última aplicación es dar un procedimiento más general para la elaboración de un análisis de conglomerados, así como la detección de datos atípicos y la validación de los resultados, además de proporcionar datos valiosos, no sólo para los demógrafos sino para el público en general.

6.5.1. Clasificación de los estados de la república mexicana de acuerdo a ciertas características demográficas

El conjunto de datos utilizado corresponde a una publicación anual realizada por el Consejo Nacional de Población (CONAPO) titulado "La situación demográfica de México (2000)", en el cual se publican varios estudios demográficos, tales como fecundidad, mortalidad, población económicamente activa (PEA), migración, etc. Para este próposito se han considerado los siguientes datos

- *Esperanza* de vida al nacimiento por entidad federativa
- Tasa de *mortalidad* infantil por entidad federativa (Decesos menores de un año de edad por cada mil nacidos vivos).
- Tasa global de *fecundidad* y cobertura de métodos anticonceptivos por entidad federativa.
- Tasa Bruta de *inmigración* interestatal por entidad federativa (por mil).
- Tasa bruta de *emigración* interestatal por entidad federativa (por mil).
- Tasa media anual de *crecimiento* demográfico por entidad federativa.

De las siguientes estados

- Aguascalientes (AGS)
- Baja California (BC)
- Baja California Sur (BCS)
- Campeche (CAMP)

- Coahuila (COAH)
- Colima (COL)
- Chiapas (CHIS)
- Chihuahua (CHIH)
- Distrito Federal (DF)
- Durango (DGO)
- Guanajuato (GTO)
- Guerrero (GRO)
- Hidalgo (HID)
- Jalisco (JAL)
- Estado de México (MEX)
- Michoacán (MICH)
- Morelos (MOR)
- Nayarit (NAY)
- Nuevo León (NL)
- Oaxaca (OAX)
- Puebla (PUE)
- Queretaro (QRO)
- Quintana Roo (QR)
- San Luis Potosí (SLP)
- Sinaloa (SIN)
- Sonora (SON)
- Tabasco (TAB)

- Tamaulipas (TAMPS)
- Tlaxcala (TLAX)
- Veracruz (VER)
- Yucatán (YUC)
- Zacatecas (ZAC)

Los datos de estas ciudades se pueden consultar en el Apéndice C.

6.5.2. Objetivos del análisis

Se tienen 6 variables y 32 casos, por lo cual el investigador tiene 192 datos y lo que se desea es un una clasificación para así obtener datos más manejables, por ejemplo medias y varianzas, y de esta forma poder describir, en general, la situación del país.

Para dicho próposito se tendrá que descubrir el número de conglomerados involucrados en los datos, en este caso se utilizará el método de Mojena mencionado en la sección 4.9 del capítulo 4. Una vez obtenidos los conglomerados, se obtendrán estadísticos y serán tomados como representantes de cada clase o conglomerado para así poder brindar una descripción del país en base a conglomerados.

Observación 6.1 *Para el caso de muestreo, los conglomerados obtenidos pueden ser considerados como estratos, y de esta forma, realizar un muestreo estratificado si el objetivo lo requiere.*

6.5.3. Diseño del análisis

Como se ha mencionado, se requiere hacer una clasificación de las 32 ciudades del país conforme a seis características demográficas. Para dicho propósito se requieren los siguientes pasos:

1. *Estandarización de la matriz de datos.* Este paso es importante ya que al estandarizar se hará contribuir de forma más equitativa a todas las variables respecto a la semejanza total, de lo contrario, la columna de la esperanza de vida "dominaría" la semejanza total, puesto que las otras variables son tasas y la esperanza no. En este caso, se estandarizarán

los datos por su rango entre 0 y 1, esto es, por medio de (4.41). Los datos estandarizados se presentan en el apéndice C.

2. *Elección de la medida de semejanza y el método a usar.* El coeficiente de semejanza a utilizar es la distancia euclidiana al cuadrado, con distintos métodos, como son: 1) ligamiento simple, 2) ligamiento centroide, 3) método de Ward.

3. *Detección de datos atípicos.* El procedimiento a seguir en este caso es realizar un análisis completo para poder localizar estos datos atípicos. Otra opción para localizarlos es realizar un análisis que proporcione información respecto a éstos.

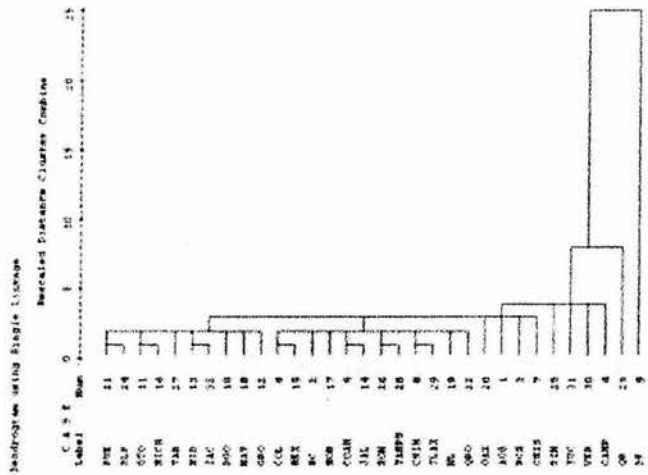
4. *Rearreglo de la matriz de datos.* El rearreglo en la matriz de datos brinda una manera visual más entendible de los mismos.

5. *Validación de los resultados.* Para la validación se tienen varias opciones; a) la primera podría ser, después de realizar un análisis total para obtener una clasificación 1 (C1), repetir el procedimiento pero a los componentes principales para obtenere una clasificación 2 (C2), si C1 es muy parecida a C2, y de hecho es lo que se espera, se obtendría cierta validez de nuestros resultados, y b) Realizar un muestreo aleatorio a los datos originales, para después emplear el mismo procedimiento y esperar que las dos clasificaciones sean muy parecidas. En este caso se empleará el primer método.

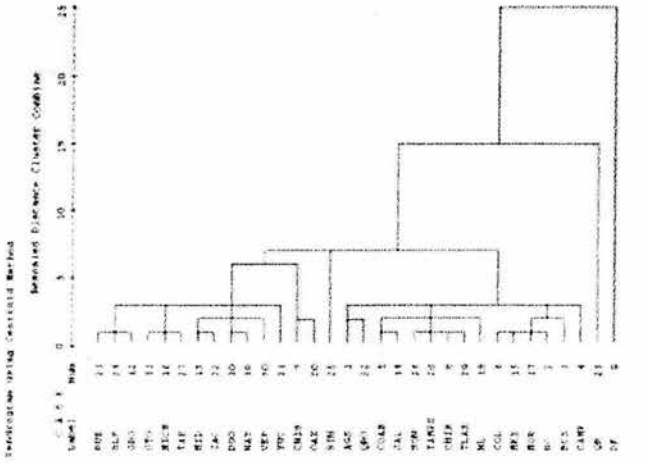
6. *Interpretación de los resultados.* Este paso es el último y se tratará de dar información en general.

6.5.4. Desarrollo

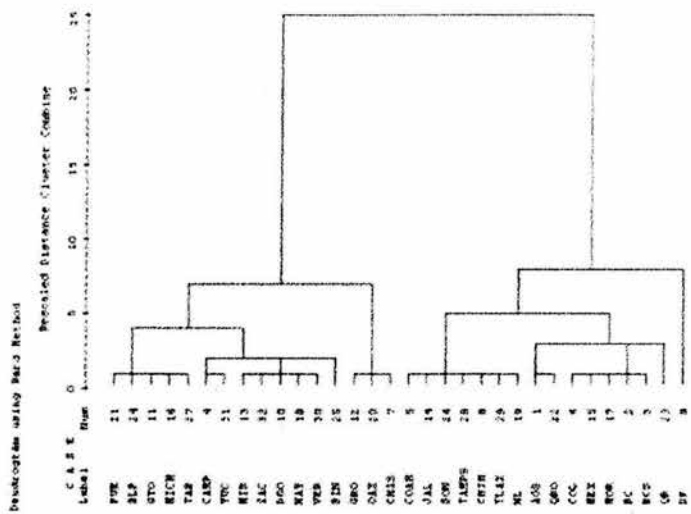
Al efectuar los análisis de los métodos simple, centroide y Ward, se obtuvieron los siguientes dendogramas:



LIGAMIENTO SIMPLE



METODO DEL CENTROIDE



METODO DE WARD

en el primer dendograma, método del ligamiento simple, se aprecia una mala clasificación llamada *encadenamiento*, la cual es comun en el método del ligamiento simple para un número "considerablemente"grande de datos, pero no es del todo mala, pues es una opción para detectar datos atípicos, de hecho, los candidatos a ser datos atípicos son los estados del DF y CHIS y OAX ya que son los que poseen una mayor distancia de ligamiento. Los mismos datos atípicos se observan en el metodo del centroide. Sin embargo, el último dendograma muestra solo el DF.

Ahora, los datos necesarios para calcular el número de conglomerados, utilizando el método de Mojena de la sección 4.9 del capítulo 4, son:

	$\overline{\alpha}$	S_α	*umbral*
simple	0.0925	0.1245	0.2480
centroide	0.2121	0.3402	0.6374
Ward	1.3796	2.1620	4.0821

por tanto se tienen las siguientes clasificaciones

simple

- Grupo 1: AGS, BC, BCS, CAMP, COAH, COL, CHIS, CHIH, DGO, GTO, GRO, HID, JAL, MEX, MICH, MOR, NAY, NL, OAX, PUE, QRO, QR, SLP, SIN, SON, TAB, TAMPS, TLAX, VER, YUC, ZAC.
- Grupo 2: DF.

centroide

- Grupo 1: AGS, BC, BCS, CAMP, COAH, COL, CHIH, DGO, GTO, GRO, HID, JAL, MEX, MICH, MOR, NAY, NL, PUE, QRO, QR, SLP, SIN, SON, TAB, TAMPS, TLAX, VER, YUC, ZAC.
- Grupo 2: CHIS, OAX
- Grupo 3: DF

Ward

- Grupo 1: PUE, SLP, GTO, MICH, TAB, CAMP, YUC, HID, ZAC, DGO, NAY, VER, SIN, GRO, OAX, CHIS.
- Grupo 2: COAH, JAL, SON, TAMPS, CHIH, TLAX, NL, AGS, QR, COL, MEX, MOR, BC, BCS, QRO.
- Grupo 3: DF.

El mejor resultado es proporcionado por el método de Ward, mientras que los métodos del ligamiento simple y del centroide dieron una clasificación muy mala. De los resultados anteriores se concluye que el DF es un dato atípico, y posiblemente CHIS y OAX. Por lo que la clasificación que se va a tomar es la proporcionada por el método de Ward. El rearreglo de la matriz de datos puede consultarse en el Apéndice C.

6.5.5. Validación de la clasificación

Como se mencionó en el diseño del análisis se recurrirá a la validación por medio de los componentes principales. De la matriz estandarizada se obtiene

la matriz de correlación

	esperanza	mortalidad	fecundidad	inmigración	emigración	crecimiento
esperanza	1.0000	-0.9996	-0.7519	0.5795	0.1432	0.2095
mortalidad	-0.9996	1.0000	0.7538	-0.5814	-0.1466	-0.2087
fecundidad	-0.7519	0.7538	1.0000	-0.5458	-0.2587	0.00734
inmigración	0.5795	-0.5814	-0.5458	1.0000	0.3380	0.5957
emigración	0.1432	-0.2587	-0.2587	0.3380	1.0000	-0.4681
crecimiento	0.2095	0.00734	0.00734	0.5957	-0.4681	1.0000

Ahora, el siguiente paso es encontrar los valores y vectores propios de esta matriz, los cuales están dados por

	eigenvalor	% varianza total	acumulado eigenvalor	% acumulado
1	3.2828	54.7140	3.2828	54.7140
2	1.5366	25.6099	4.8194	80.3238
3	0.8641	14.4023	5.6836	94.7261
4	0.3057	5.0949	5.9893	99.8210
5	0.0104	0.1732	5.9997	99.9942
6	0.0004	0.0058	6.0000	100.0000

y

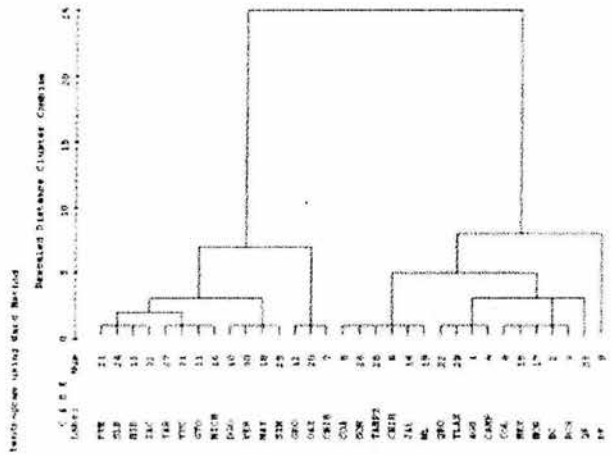
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
esperanza	0.5174	-0.0175	0.2967	0.3818	-0.0119	0.7057
mortalidad	-0.5180	0.0193	-0.2937	-0.3784	-0.0147	0.7082
fecundidad	-0.4633	0.1976	-0.2083	0.8032	-0.2404	-0.0065
inmigración	0.4424	0.1672	-0.5940	-0.1308	-0.6373	-0.0104
emigración	0.1471	-0.6347	-0.5826	0.2197	0.4333	0.0097
crecimiento	0.1786	0.7277	-0.3001	0.0229	0.5898	0.0108

respectivamente, por lo que se tomarán los tres primeros componentes principales, ya que proporcionan casi el 95 % de la variación total, para el análisis de la validación, en otras palabras, se tomará una nueva matriz de datos de tamaño 32×3 obtenida de la siguiente manera: se multiplicarán la matriz original y la matriz de factores para obtener una matriz 32×6 de la cual sólo se tomarán las primeras tres columnas. Ésta matriz puede consultarse en el apéndice C.

Una vez obtenida esta nueva matriz de datos, se le aplicará el mismo análisis hecho anteriormente.

Observación 6.2 *Nótese que los eigenvectores están estandarizados.*

El dendograma de la matriz de los tres primeros componentes principales se muestra a continuación:



METODO DE WARD PARA 3 COMPONENTES

Los datos que utiliza el criterio de Mojena de la sección 4.9 del capítulo 4 son:

$\overline{\alpha}$	S_α	*umbral*
1.0707	1.9711	3.5346

De esta manera, las clasificaciones usando el método de Ward y la matriz 32×3 de los tres primeros componentes principales son:

- Grupo 1: PUE, SLP, HID, ZAC, TAB, YUC, GTO, MICH, DGO, VER, NAY, SIN, GRO, OAX, CHIS.
- Grupo 2: COA, SON, TAMPS, CHIH, JAL, NL, QRO, TLAX, AGS, CAMP, COL, MEX, MOR, BC, BCS, QR.
- Grupo 3: DF.

Conclusión de la validación Por tanto, al comparar la clasificación de la matriz original con la clasificación mediante el uso de los tres componentes principales se tienen que solo difieren en el estado de Campeche (CAMP), lo que se puede atribuir a que los componentes principales representan casi el 95 % de la variabilidad, de esta manera se puede concluir que la clasificación que proporcionó el método de Ward es confiable, y, además también se concluye que el DF es un dato atípico y que la clasificación obtenida consta de dos conglomerados y un dato atípico.

6.5.6. Interpretación de la clasificación

Como último paso, sólo hace falta dar la información que brindó el análisis de conglomerados. Los estadísticos de la clasificación están dados por

CONGLOMERADO 1

	Media	Desviación estándar	Varianza
Esperanza	74.4	0.8729	0.7620
Mortalidad	26.9	2.2240	4.9463
Fecundidad	2.6	0.3015	0.0909
Inmigración	8.6	5.1531	26.5540
Emigración	9.3	2.5173	6.3367
Crecimiento	1.5	0.5114	0.2615

CONGLOMERADO 2

	Media	Desviación estándar	Varianza
Esperanza	75.9	0.6588	0.4340
Mortalidad	23.0	1.6315	2.6617
Fecundidad	2.3	0.2610	0.0681
Inmigración	11.5	3.9585	15.6695
Emigración	7.4	2.0070	4.0281
Crecimiento	1.8	0.4063	0.1651

Al realizar el análisis de conglomerados se encontró una estructura involucrada de dos conglomerados, de esta manera, se pueden clasificar a los estados de la república mexicana como sigue: 1) Los estados que conforman el conglomerado 1 presenta una menor esperanza de vida y una mortalidad alta, asimismo, un índice de emigración mayor que la tasa de inmigración. Su tasa de crecimiento es menor que la de los estados que conforman el conglomerado 2. 2) Para el segundo conglomerado se tiene una esperanza de vida alta junto con una mortalidad baja, pero presentan una fecundidad baja. El sentido de

la migración, a comparación con el primer conglomerado, se invierte aqui, es decir, la inmigración es más alta que la emigración. Los integrantes de este conglomerado presentan un crecimiento mayor. 3) Se podría decir que los resultados más confiables son los que poseen una desviación estándar baja, en otras palabras, los datos proporcionados por fecundidad, crecimiento y esperanza de vida son los más confiables en ambos conglomerados, en contraste, con la peor desviación estándar proporcionada por la inmigración. 4) Era de esperarse que el DF fuera un dato atípico, ya que el DF es la ciudad más grande y con un movimiento migratorio mayor que el de los otros estados, pero curiosamente, posee una esperanza de vida bastantemente alta, además las tasas de crecimiento y fecundidad son las peores.

Las conclusiones del párrafo anterior puede resumirse mediante la siguiente tabla:

	cluster1	cluster2	cluster3 (DF)
Esperanza	baja	media	alta
Mortalidad	alta	media	baja
Fecundidad	alta	media	baja
Inmigración	baja	media	alta
Emigración	media	baja	alta
Crecimiento	media	alta	baja

Conclusiones y recomendaciones

A la largo del presente trabajo se presentó material suficiente para elaborar un análisis de conglomerados en cualquier disciplina, sin embargo, como su título lo indica, este trabajo es una introducción a esta técnica multivariada y todavía falta mucho material que abarcaría el volumen II, sólo por mencionar algunos de los temas faltantes se tiene al análisis de busqueda de densidades, métodos de conglomeración traslapada (ideales para análisis de mercadotecnía), y los métodos difusos, mejor conocidos como fuzzy cluster, los cuales requieren del uso de la lógica difusa (fuzzy logic).

Una vez aclarados algunos de los temas no incluidos en este trabajo se mencionarán algunas conclusiones y/o recomendaciones:

1. Antes de realizar cualquier análisis se recomienda el uso de varias gráficas multivariadas mencionadas en el capítulo 3, así como las técnicas adicionales, componentes principales y/o escalamiento multidimensional.

2. Es de suma inportancia la elección del coeficiente de semejanza que va a ser usado, ya que este influye sustancialmente en el proceso de conglomeración.

3. No se puede decir qué método es el mejor, pero si se tiene conocimento de las características de cada uno de ellos, por ejemplo, el método del ligamiento simple , para muestras grandes, tiende a presentar el fenómeno encadenamiento, el del ligamiento completo suele formar grupos esféricos y de pocos individuos y el método de Ward también forma conglomerados esféricos casi del mismo tamaño. La elección del método prácticamente radica en el problema a investigar y en el criterio del investigador. Es indispensable realizar más de un análisis a la vez para encontrar un buen análisis.

4. La estandarización es un paso importante, y al igual que los métodos de conglomeración no se puede decir qué función estandarizadora es la mejor, pero a través de la práctica se sabe que la función estandarizadora dada por (4.41) brinda buenos resultados.

5. Se recomienda el uso de $k = 1{,}25$ en el método de Mojena presentado en la sección 4.9 del capítulo 4, ya que brindó buenos resultados.

6. La estabilidad de la solución puede verificarse por medio de 1) Análisis de conglomerados individuales, es decir, se toman muestras aleatorias de los datos originales y se aplican dos análisis por separado, similarmente, este procedimiento también se debe aplicar a las variables y deben compararse los resultados. Una clasificación buena no debe de afectarse considerablemente debido a un muestreo de datos, y 2) A través del uso de los componentes principales, el cual fue empleado para validar los resultados de la aplicación de demografía-muestreo en la sección 6.5 del capítulo 6.

Por último se recomienda la lectura de Romesburg (1990) y Everitt (2001) como textos introductorios y de fácil entendimiento al tema.

Conclusiones y recomendaciones

A la largo del presente trabajo se presentó material suficiente para elaborar un análisis de conglomerados en cualquier disciplina, sin embargo, como su título lo indica, este trabajo es una introducción a esta técnica multivariada y todavía falta mucho material que abarcaría el volumen II, sólo por mencionar algunos de los temas faltantes se tiene al análisis de busqueda de densidades, métodos de conglomeración traslapada (ideales para análisis de mercadotecnía), y los métodos difusos, mejor conocidos como fuzzy cluster, los cuales requieren del uso de la lógica difusa (fuzzy logic).

Una vez aclarados algunos de los temas no incluidos en este trabajo se mencionarán algunas conclusiones y/o recomendaciones:

1. Antes de realizar cualquier análisis se recomienda el uso de varias gráficas multivariadas mencionadas en el capítulo 3, así como las técnicas adicionales, componentes principales y/o escalamiento multidimensional.

2. Es de suma inportancia la elección del coeficiente de semejanza que va a ser usado, ya que este influye sustancialmente en el proceso de conglomeración.

3. No se puede decir qué método es el mejor, pero si se tiene conocimento de las características de cada uno de ellos, por ejemplo, el método del ligamiento simple , para muestras grandes, tiende a presentar el fenómeno encadenamiento, el del ligamiento completo suele formar grupos esféricos y de pocos individuos y el método de Ward también forma conglomerados esféricos casi del mismo tamaño. La elección del método prácticamente radica en el problema a investigar y en el criterio del investigador. Es indispensable realizar más de un análisis a la vez para encontrar un buen análisis.

4. La estandarización es un paso importante, y al igual que los métodos de conglomeración no se puede decir qué función estandarizadora es la mejor, pero a través de la práctica se sabe que la función estandarizadora dada por (4.41) brinda buenos resultados.

5. Se recomienda el uso de $k = 1{,}25$ en el método de Mojena presentado en la sección 4.9 del capítulo 4, ya que brindó buenos resultados.

6. La estabilidad de la solución puede verificarse por medio de 1) Análisis de conglomerados individuales, es decir, se toman muestras aleatorias de los datos originales y se aplican dos análisis por separado, similarmente, este procedimiento también se debe aplicar a las variables y deben compararse los resultados. Una clasificación buena no debe de afectarse considerablemente debido a un muestreo de datos, y 2) A través del uso de los componentes principales, el cual fue empleado para validar los resultados de la aplicación de demografía-muestreo en la sección 6.5 del capítulo 6.

Por último se recomienda la lectura de Romesburg (1990) y Everitt (2001) como textos introductorios y de fácil entendimiento al tema.

Apéndice A

Árboles

El inicio del árbol es llamado "raíz", las ramas son llamadas "nodos" o "vértices" y las conecciones entre los vértices son llamadas "bordes". El último vértice o vértice terminal que no tiene ramas es llamado "hoja" y representa a los objetos que han sido unidos en algún conglomerado. Cada vértice en el árbol, incluyendo a la raíz, representa un conglomerado específico de todos los objetos que puede ser alcanzado siguiendo la trayectoria del árbol.

Desde el punto teoríco de gráficas, un árbol es una gráfica conectada sin ciclos.Así, un árbol de jerarquías es una gráfica conectada que significa que sólo hay un camino entre cualquiera dos nodos o vértices en la gráfica, además, el árbol de jerarquías no contiene ciclos, es decir, viajando desde cualquier vértice en la dirección definida, fuera de la raíz, se termina siempre al final de un vértice (hoja) representando a un solo objeto. El vértice inicial, del cual todos los otros vértices estan conectados a este, es llamado raíz. En el caso que solo hay dos ramas, para cada vértice, es llamado "árbol binario".

De la definición anterior, se sigue, evidentemente, que el árbol ligado por N objetos, siempre tiene exactamente 2N-1 vértices.

En la mayoría de las aplicaciones, los árboles que son usados son binarios debido a que cualquier otro árbol puede ser sustituido, al menos teoricamente, por un árbol binario, además que los árboles binarios pueden ser implementados más fácilmente en la computadora.

Si los valores númericos, esto es, las longitudes de las ramas, son introducidos en cada vértice en el árbol de jerarquías, entonces el árbol es llamado "dendograma".

Las longitudes de las ramas pueden ser absolutas o relativas respecto a

una distancia máxima. La escala de similaridad está relacionada a la función de similaridad dada por

$$s(\overline{X}_i, \overline{X}_j) = 1 - d(\overline{X}_i, \overline{X}_j)/\,\text{máx}(d(\overline{X}_l, \overline{X}_m))$$

donde X_n es el vector en el espacio de medida.En teoria de gráficas se maneja el término "distancia"pero es un concepto distinto a la distancia entre objetos y/o conglomerados. En la teoria de gráficas la distancia esta asociada al número de bordes (la cual se denotará aquí como distancia-gráfica para evitar confusión alguna), así en el árbol de jerarquías se define la "longitud"del árbol como la distancia.

Apéndice B

Irises

1.

DATOS ORIGINALES															
Iris setosa					Iris versicolor					Iris virgínica					
Clave	Largo	Ancho	Largo	Ancho	Clave	Largo	Ancho	Largo	Ancho	Clave	Largo	Ancho	Largo	Ancho
1	5,1	3,5	1,4	0,2	51	7	3,2	4,7	1,4	101	6,3	3,3	6	2,5
2	4,9	3	1,4	0,2	52	6,4	3,2	4,5	1,5	102	5,8	2,7	5,1	1,9
3	4,7	3,2	1,3	0,2	53	6,9	3,1	4,9	1,5	103	7,1	3	5,9	2,1
4	4,6	3,1	1,5	0,2	54	5,5	2,3	4	1,3	104	6,3	2,9	5,6	1,8
5	5	3,6	1,4	0,2	55	6,5	2,8	4,6	1,5	105	6,5	3	5,8	2,2
6	5,4	3,9	1,7	0,4	56	5,7	2,8	4,5	1,3	106	7,6	3	6,6	2,1
7	4,6	3,4	1,4	0,3	57	6,3	3,3	4,7	1,6	107	4,9	2,5	4,5	1,7
8	5	3,4	1,5	0,2	58	4,9	2,4	3,3	1	108	7,3	2,9	6,3	1,8
9	4,4	2,9	1,4	0,2	59	6,6	2,9	4,6	1,3	109	6,7	2,5	5,8	1,8
10	4,9	3,1	1,5	0,1	60	5,2	2,7	3,9	1,4	110	7,2	3,6	6,1	2,5
11	5,4	3,7	1,5	0,2	61	5	2	3,5	1	111	6,5	3,2	5,1	2
12	4,8	3,4	1,6	0,2	62	5,9	3	4,2	1,5	112	6,4	2,7	5,3	1,9
13	4,8	3	1,4	0,1	63	6	2,2	4	1	113	6,8	3	5,5	2,1
14	4,3	3	1,1	0,1	64	6,1	2,9	4,7	1,4	114	5,7	2,5	5	2
15	5,8	4	1,2	0,2	65	5,6	2,9	3,6	1,3	115	5,8	2,8	5,1	2,4
16	5,7	4,4	1,5	0,4	66	6,7	3,1	4,4	1,4	116	6,4	3,2	5,3	2,3
17	5,4	3,9	1,3	0,4	67	5,6	3	4,5	1,5	117	6,5	3	5,5	1,8
18	5,1	3,5	1,4	0,3	68	5,8	2,7	4,1	1	118	7,7	3,8	6,7	2,2
19	5,7	3,8	1,7	0,3	69	6,2	2,2	4,5	1,5	119	7,7	2,6	6,9	2,3
20	5,1	3,8	1,5	0,3	70	5,6	2,5	3,9	1,1	120	6	2,2	5	1,5
21	5,4	3,4	1,7	0,2	71	5,9	3,2	4,8	1,8	121	6,9	3,2	5,7	2,3
22	5,1	3,7	1,5	0,4	72	6,1	2,8	4	1,3	122	5,6	2,8	4,9	2
23	4,6	3,6	1	0,2	73	6,3	2,5	4,9	1,5	123	7,7	2,8	6,7	2
24	5,1	3,3	1,7	0,5	74	6,1	2,8	4,7	1,2	124	6,3	2,7	4,9	1,8
25	4,8	3,4	1,9	0,2	75	6,4	2,9	4,3	1,3	125	6,7	3,3	5,7	2,1
26	5	3	1,6	0,2	76	6,6	3	4,4	1,4	126	7,2	3,2	6	1,8
27	5	3,4	1,6	0,4	77	6,8	2,8	4,8	1,4	127	6,2	2,8	4,8	1,8
28	5,2	3,5	1,5	0,2	78	6,7	3	5	1,7	128	6,1	3	4,9	1,8
29	5,2	3,4	1,4	0,2	79	6	2,9	4,5	1,5	129	6,4	2,8	5,6	2,1
30	4,7	3,2	1,6	0,2	80	5,7	2,6	3,5	1	130	7,2	3	5,8	1,6
31	4,8	3,1	1,6	0,2	81	5,5	2,4	3,8	1,1	131	7,4	2,8	6,1	1,9
32	5,4	3,4	1,5	0,4	82	5,5	2,4	3,7	1	132	7,9	3,8	6,4	2
33	5,2	4,1	1,5	0,1	83	5,8	2,7	3,9	1,2	133	6,4	2,8	5,6	2,2
34	5,5	4,2	1,4	0,2	84	6	2,7	5,1	1,6	134	6,3	2,8	5,1	1,5
35	4,9	3,1	1,5	0,2	85	5,4	3	4,5	1,5	135	6,1	2,6	5,6	1,4
36	5	3,2	1,2	0,2	86	6	3,4	4,5	1,6	136	7,7	3	6,1	2,3
37	5,5	3,5	1,3	0,2	87	6,7	3,1	4,7	1,5	137	6,3	3,4	5,6	2,4
38	4,9	3,6	1,4	0,1	88	6,3	2,3	4,4	1,3	138	6,4	3,1	5,5	1,8
39	4,4	3	1,3	0,2	89	5,6	3	4,1	1,3	139	6	3	4,8	1,8
40	5,1	3,4	1,5	0,2	90	5,5	2,5	4	1,3	140	6,9	3,1	5,4	2,1
41	5	3,5	1,3	0,3	91	5,5	2,6	4,4	1,2	141	6,7	3,1	5,6	2,4
42	4,5	2,3	1,3	0,3	92	6,1	3	4,6	1,4	142	6,9	3,1	5,1	2,3
43	4,4	3,2	1,3	0,2	93	5,8	2,6	4	1,2	143	5,8	2,7	5,1	1,9
44	5	3,5	1,6	0,6	94	5	2,3	3,3	1	144	6,8	3,2	5,9	2,3
45	5,1	3,8	1,9	0,4	95	5,6	2,7	4,2	1,3	145	6,7	3,3	5,7	2,5
46	4,8	3	1,4	0,3	96	5,7	3	4,2	1,2	146	6,7	3	5,2	2,3
47	5,1	3,8	1,6	0,2	97	5,7	2,9	4,2	1,3	147	6,3	2,5	5	1,9
48	4,6	3,2	1,4	0,2	98	6,2	2,9	4,3	1,3	148	6,5	3	5,2	2
49	5,3	3,7	1,5	0,2	99	5,1	2,5	3	1,1	149	6,2	3,4	5,4	2,3
50	5	3,3	1,4	0,2	100	5,7	2,8	4,1	1,3	150	5,9	3	5,1	1,8

2.

DATOS AL APLICAR EL METODO DE WARD					
Cluster Membership		Cluster Membership		Cluster Membership	
Case	3 Clusters	Case	3 Clusters	Case	3 Clusters
1	1	51	2	**107**	**2**
2	1	52	2	**114**	**2**
3	1	53	2	**115**	**2**
4	1	54	2	**120**	**2**
5	1	55	2	**122**	**2**
6	1	56	2	**124**	**2**
7	1	57	2	**127**	**2**
8	1	58	2	**128**	**2**
9	1	59	2	**134**	**2**
10	1	60	2	**135**	**2**
11	1	61	2	**139**	**2**
12	1	62	2	**143**	**2**
13	1	63	2	**147**	**2**
14	1	64	2	**150**	**2**
15	1	65	2	**78**	**3**
16	1	66	2	101	3
17	1	67	2	103	3
18	1	68	2	104	3
19	1	69	2	105	3
20	1	70	2	106	3
21	1	71	2	108	3
22	1	72	2	109	3
23	1	73	2	110	3
24	1	74	2	111	3
25	1	75	2	112	3
26	1	76	2	113	3
27	1	77	2	116	3
28	1	79	2	117	3
29	1	80	2	118	3
30	1	81	2	119	3
31	1	82	2	121	3
32	1	83	2	123	3
33	1	84	2	125	3
34	1	85	2	126	3
35	1	86	2	129	3
36	1	87	2	130	3
37	1	88	2	131	3
38	1	89	2	132	3
39	1	90	2	133	3
40	1	91	2	136	3
41	1	92	2	137	3
42	1	93	2	138	3
43	1	94	2	140	3
44	1	95	2	141	3
45	1	96	2	142	3
46	1	97	2	144	3
47	1	98	2	145	3
48	1	99	2	146	3
49	1	100	2	148	3
50	1	**102**	**2**	149	3

* Los valores en negritas no coinciden con la clasificación original

3.

DATOS AL APLICAR EL ALGORITMO DE *K*-MEDIAS					
Cluster Membership		Cluster Membership		Cluster Membership	
Case	*3 Clusters*	*Case*	*3 Clusters*	*Case*	*3 Clusters*
1	1	**114**	**1**	89	2
2	1	**115**	**1**	90	2
4	1	**120**	**1**	91	2
5	1	**122**	**1**	92	2
6	1	**124**	**1**	93	2
7	1	**127**	**1**	94	2
8	1	**128**	**1**	95	2
9	1	**134**	**1**	96	2
10	1	**139**	**1**	97	2
11	1	**143**	**1**	98	2
12	1	**147**	**1**	99	2
13	1	**150**	**1**	100	2
14	1	51	2	**3**	**3**
15	1	52	2	**28**	**3**
16	1	53	2	101	3
17	1	54	2	103	3
18	1	55	2	104	3
19	1	56	2	105	3
20	1	57	2	106	3
21	1	58	2	108	3
22	1	59	2	109	3
23	1	60	2	110	3
24	1	61	2	111	3
25	1	62	2	112	3
26	1	63	2	113	3
27	1	64	2	116	3
29	1	65	2	117	3
30	1	66	2	118	3
31	1	67	2	119	3
32	1	68	2	121	3
33	1	69	2	123	3
34	1	70	2	125	3
35	1	71	2	126	3
36	1	72	2	129	3
37	1	73	2	130	3
38	1	74	2	131	3
39	1	75	2	132	3
40	1	76	2	133	3
41	1	77	2	135	3
42	1	78	2	136	3
43	1	79	2	137	3
44	1	80	2	138	3
45	1	81	2	140	3
46	1	82	2	141	3
47	1	83	2	142	3
48	1	84	2	144	3
49	1	85	2	145	3
50	1	86	2	146	3
102	1	87	2	148	3
107	1	88	2	149	3

* Los valores en negritas no coinciden con la clasificación original

Apéndice C

Demografía-muestreo

1.

Datos de 2000						
	Esperanza	*Mortalidad*	*Fecundidad*	*Inmigracion*	*Emigracion*	*Crecimiento*
AGUASCALIENTES	76,4	21,9	2,61	12,6	4,8	2,44
BAJA CALIFORNIA	76,3	22	2,15	19,5	8,7	2,52
BAJA CALIFORNIA SUR	76,3	22,1	2,1	13,7	12	1,49
CAMPECHE	74,7	26	2,26	14,4	9,6	1,9
COAHUILA	76,2	22,3	2,39	7,7	8,9	1,3
COLIMA	76,4	21,9	2,11	17,2	8,2	2,11
CHIAPAS	72,4	31,9	2,94	2,4	5,8	1,6
CHIHUAHUA	75,8	23,4	2,2	8,7	4,7	1,67
DISTRITO FEDERAL	77,2	19,8	1,8	16,9	23,7	0,39
DURANGO	74,8	25,7	2,65	7,4	13,5	0,89
GUANAGUATO	75,1	25,1	2,75	5,2	5,2	1,62
GUERRERO	73,3	29,7	3,03	4,8	8,3	1,39
HIDALGO	74,2	27,4	2,6	9	10,1	1,38
JALISCO	76,3	22,1	2,51	6,1	6,9	1,35
ESTADO DE MEXICO	76,3	22	2,18	15,5	9	2,09
MICHOACAN	74,8	25,9	2,8	5,1	7	1,33
MORELOS	75,9	23,2	2,1	14,1	6,9	1,89
NAYARIT	75,2	24,8	2,43	8,6	11,5	1
NUEVO LEON	76,8	20,9	2,06	8	5,2	1,53
OAXACA	72,5	31,7	2,92	5,7	10,5	1,14
PUEBLA	74,1	27,6	2,98	6,6	8,7	1,61
QUERETARO	75,3	24,6	2,54	12	5,5	2,31
QUINTANA ROO	75,7	23,6	2,41	25,8	13,6	2,97
SAN LUIS POTOSI	74,2	27,3	2,94	7,2	8,9	1,46
SINALOA	75,4	24,3	2,12	6,6	13	0,52
SONORA	76,1	22,6	2,12	9	7,2	1,42
TABASCO	75,0	25,3	2,55	7,7	7,5	1,76
TAMAULIPAS	75,5	23,9	2,12	10,6	8,1	1,48
TLAXCALA	75,4	24,2	2,31	10,5	5,7	1,92
VERACRUZ	74,0	28	2,29	5,3	10,4	0,75
YUCATAN	74,3	27	2,21	8,1	5,2	1,5
ZACATECAS	74,4	26,8	2,68	7,9	10,6	1,12

2.

Rearreglo de la clasificación obtenida por el método de Ward						
	Esperanza	*Mortalidad*	*Fecundidad*	*Inmigracion*	*Emigracion*	*Crecimiento*
PUEBLA	74,1	27,6	2,96	6,8	8,7	1,61
SAN LUIS POTOSI	74,2	27,3	2,94	7,2	8,9	1,46
GUANAGUATO	75,1	25,1	2,75	5,2	5,2	1,62
MICHOACAN	74,8	25,9	2,8	5,1	7	1,33
TABASCO	75,0	25,3	2,55	7,7	7,5	1,76
CAMPECHE	74,7	26	2,26	14,4	9,6	1,9
YUCATAN	74,3	27	2,21	8,1	5,2	1,5
HIDALGO	74,2	27,4	2,6	9	10,1	1,38
ZACATECAS	74,4	26,8	2,68	7,9	10,6	1,12
DURANGO	74,8	25,7	2,65	7,4	13,5	0,89
NAYARIT	75,2	24,8	2,43	8,8	11,5	1
VERACRUZ	74,0	28	2,29	5,3	10,4	0,75
SINALOA	75,4	24,3	2,12	6,6	13	0,52
GUERRERO	73,3	29,7	3,03	4,8	8,3	1,39
OAXACA	72,5	31,7	2,92	5,7	10,5	1,14
CHIAPAS	72,4	31,9	2,94	2,4	5,8	1,6
COAHUILA	76,2	22,3	2,39	7,7	8,9	1,3
JALISCO	76,3	22,1	2,51	6,1	6,9	1,35
SONORA	76,1	22,6	2,12	9	7,2	1,42
TAMAULIPAS	75,5	23,9	2,12	10,6	8,1	1,48
CHIHUAHUA	75,8	23,4	2,2	8,7	4,7	1,67
TLAXCALA	75,4	24,2	2,31	10,5	5,7	1,92
NUEVO LEON	76,8	20,9	2,06	8	5,2	1,53
AGUASCALIENTES	76,4	21,9	2,61	12,6	4,8	2,44
QUINTANA ROO	75,7	23,6	2,41	25,8	13,6	2,97
COLIMA	76,4	21,9	2,11	17,2	8,2	2,11
ESTADO DE MEXICO	76,3	22	2,18	15,5	9	2,09
MORELOS	75,9	23,2	2,1	14,1	6,9	1,89
BAJA CALIFORNIA	76,3	22	2,15	19,5	8,7	2,52
BAJA CALIFORNIA SUR	76,3	22,1	2,1	13,7	12	1,49
DISTRITO FEDERAL	77,2	19,8	1,8	16,9	23,7	0,39

3.

MATRIZ COMPONENTES PRINCIPALES					
0,37170	0,76660	-0,44136	0,74376	0,02234	0,71085
0,69612	0,63483	-0,67613	0,43955	0,03168	0,70368
0,55523	0,18477	-0,50421	0,46522	0,03910	0,70976
0,17859	0,42335	-0,71675	0,29241	0,02697	0,70212
0,27605	0,23913	-0,29488	0,63637	0,03170	0,70553
0,65049	0,51251	-0,53934	0,42795	-0,00312	0,71179
-0,85515	0,50696	-0,66123	0,38941	0,06415	0,70786
0,26942	0,46363	-0,25379	0,39520	0,03005	0,71105
0,93864	-0,54858	-0,65396	0,52053	0,02653	0,70899
-0,11680	0,01999	-0,59370	0,63968	-0,00043	0,69809
-0,15176	0,50142	-0,35211	0,67039	0,01751	0,70635
-0,64766	0,38903	-0,68058	0,60222	-0,00934	0,71027
-0,19755	0,28007	-0,65696	0,46222	-0,00035	0,70924
0,20791	0,32720	-0,20804	0,71520	0,01764	0,70779
0,58173	0,47976	-0,53884	0,48115	0,04330	0,70366
-0,24521	0,36919	-0,41745	0,67296	-0,01501	0,70855
0,46080	0,47403	-0,45596	0,34133	0,00298	0,71422
0,06632	0,08959	-0,49689	0,52602	-0,01593	0,70420
0,51791	0,37231	-0,08886	0,46998	0,05644	0,71247
-0,76134	0,23990	-0,82117	0,42240	-0,01979	0,70996
-0,39655	0,43770	-0,66032	0,69429	0,00595	0,70547
0,14898	0,69934	-0,55415	0,53640	0,03804	0,70761
0,65322	0,68959	-1,15860	0,53701	0,02344	0,70933
-0,35914	0,38431	-0,63892	0,68425	-0,02674	0,70212
0,16284	-0,16294	-0,35421	0,38053	0,02918	0,70571
0,37386	0,29653	-0,25747	0,41687	0,03761	0,70886
-0,02099	0,45050	-0,47955	0,53945	0,07298	0,70506
0,29491	0,29908	-0,40130	0,33048	0,02818	0,69673
0,20971	0,53400	-0,42200	0,41391	0,03950	0,69925
-0,23928	0,01781	-0,47338	0,24368	0,02360	0,71575
-0,06939	0,40752	-0,41591	0,17745	0,01631	0,70098
-0,21538	0,19334	-0,60074	0,55876	-0,03383	0,70276

Bibliografía

[1] CONAPO. La Situación Demográfica de México (2000). *Proyecciones de la población de las entidades, los municipios y localidades.* pags. 19-20, 22, 24, 25 y 28.

[2] Hair, J. F., y otros (1995). *Multivariate Data Analysis, 4a edición.* Prentice Hall, New Jersey.

[3] Dillon, W. R. y Goldstein,.M. (1984). *Multivariate Analysis Methods with Applications.* John Wiley & Sons, Canada.

[4] Everitt, B. S. (2001). *Cluster Analysis, 4a. edición.*Edward Arnold, Londres.

[5] Everitt, B. S. y Dunn, G (2001). *Applied Multivariate Data Analysis, 2a. edición.* Edward Arnold, Londres.

[6] Flury, B. (1997). *A First Course in Multivariate Statistics.* Springer, New York .

[7] Jobson, J. D.. *Applied Multivariate Data Analysis: Categorical and Multivariate Methods, vol II.* Springer.

[8] Kaufman, L. y Rousseeuw, P. J. (1990). *Finding Groups in Data. An Introduction to Cluster Analysis.* John Wiley & Sons, Inc.

[9] Kiers, H. A.L. y otros (2000). *Studies in Classification, Data Analysis, and Knowledge Organization.* Springer.

[10] Krzanowski, W. J.y Marriott, F. H. C. (1995). *Multivariate Analysis, vol II.* Arnold, Londres.

[11] Leon, S. J. (1994) *Algebra with applications, 4a edición.* Prentice Hall

[12] Mardia K,. V. y otros (1979). *Multivariate Analysis.* Academic Press, Londres.

[13] Romesburg, H. C. y Krieger, R. E. (1990). *Cluster Analysis for Researchers.* Publishing Company.

[14] Shilov, G. E. (1977). *Linear algebra, 2a edición.* Dover publications. Canada.

[15] Späth, H. (1980). *Cluster Analysis Algorithms for Data Reduction and Classification of Objets.* Halsted Press: a division of Wiley & Sons.

[16] Strang, G. (1982). *Algebra lineal y sus aplicaciones, 1a edición.* Fondo Educativo Interamericano.

[17] Zupan, J. (1982). *Clustering of Large Data Sets.* Research Studies Press.