Universidad Nacional Autónoma de México

Facultad de Ciencias

Identificación de grupos vía análisis de conglomerados basado en modelos de mezclas

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
ACTUARIO

PRESENTA:
SYLVIA CEJA MAYÉS LIMÓN

DIRECTOR DE TESIS:
DRA. RUTH SELENE FUENTES GARCÍA

2011

Índice general

Prólogo

En diferentes áreas de la ciencia ha surgido la necesidad de separar, juntar o resumir grandes bases de datos. Clasificar, entonces, resulta ser de gran importancia.

En un sentido amplio, la clasificación de información puede servir simplemente para tener una estructura organizada de los datos de modo que el estudio de estos resulte más rápido y sencillo. Pero más allá de sólo clasificar se pueden describir patrones de similaridad entre los individuos de la población estudiada y proporcionar un resumen adecuado de la información.

El análisis de conglomerados tiene como finalidad ayudar a descubrir dichos patrones, dado que en la mayoría de los casos no se conoce previamente la estructura de los datos.

La presente tesis procura, primeramente, dar una introducción al análisis de conglomerados de manera clásica y, segundo, vía de modelos de mezclas discutir, con dos alternativas diferentes, el problema de determinar estructura conglomerada dentro de datos sin conocimiento previo del número de grupos o de alguna otra información acerca de su composición.

Este trabajo está estructurado como sigue:

El Capítulo 1 busca que el lector se familiarice con el concepto "clasificar", se plantea un panorama general, se dan razones del por qué un esquema de clasificación ayuda a entender un gran conjunto de datos; así como antecedentes históricos. En el Capítulo 2 se definen algunos de los conceptos básicos de la teoría de la probabilidad y estadística. En el Capítulo 3 se describen dos tipos de métodos comunes para determinar conglomerados: métodos de partición (algoritmo k-medias) y jerárquicos (vecino más próximo, vecino más lejano, etc.). A lo largo del Capítulo 4 se aborda la teoría del análisis de conglomerados basado en modelos de mezclas. Para la primera alternativa descrita en el presente trabajo, el criterio de máxima verosimilitud es utilizado para grupos fusionados. Se estiman los parámetros de los

componentes de la mezcla y se clasifican las observaciones en los grupos por sus probabilidades de pertenencia a las distintas poblaciones, las técnicas de traslado son usualmente basadas en el algoritmo EM. Los modelos son comparados usando una aproximación del factor de Bayes basado en el *Criterio de Información Bayesiana (BIC)*, para elegir el mejor modelo. Por otro lado, la segunda alternativa basada en distancia servirá para determinar el número desconocido de componentes de la mezcla, generalizando un método de prueba bayesiano basado en la distancia Kullback-Leibler (KL) propuesto por Mengersen y Robert (1996). Una alternativa, Kullback-Leibler ponderada, es propuesta como criterio de prueba. Un procedimiento paso a paso es sugerido para seleccionar el número mínimo de componentes adecuado para los datos. En el Capítulo 5, se muestra una aplicación utilizando los dos diferentes enfoques. Finalmente, en el último Capítulo se presentan las conclusiones respectivas sobre los enfoques que se muestran a lo largo de este trabajo y de esa manera determinar si son o no de utilidad.

La presente tesis fue realizada utilizando dos software libres;

1. `R`, programa estadístico utilizado para llevar a cabo todas las rutinas necesarias para analizar la información.

2. LaTeX, sistema de composición de textos el cual se utilizó para la creación y formato de este trabajo.

Ambos pueden ser obtenidos en los sitios:

www.r-project.org y *www.latex-project.org*,

respectivamente.

Capítulo 1

Introducción

La idea de agrupar cosas similares dentro de categorías es claramente una idea primitiva desde los primeros humanos. Todas las personas clasificamos ya sea por necesidad o por tendencia, esto se hace para procesar la realidad. Una forma muy eficiente de procesar la realidad es simplificarla (mucho o poco), y una forma de simplificar es clasificar.

Categorizar o clasificar ayuda a ver características generales de los objetos de un cierto grupo y los elementos que tienen en común; objetos que difieren en insignificantes detalles, generalmente, reciben el mismo nombre, pueden ser tratados de igual manera, se espera que actúen de modo similar y si algunos objetos en un grupo tienen cierta propiedad, se esperará que otros objetos en ese grupo tengan la misma característica.

El hecho de clasificar es muy útil, pero no es perfecto. Entre los defectos más evidentes está el asignar un objeto a una clase que no corresponde o el de crear clases que no deberían estar. Pero eso será discutible, pues es subjetivo.

1.1. Razones para clasificar

En cierto nivel, un esquema de clasificación simplemente puede representar un método conveniente para la organización de un gran conjunto de datos de modo que pueda ser más fácil de entender y obtener información de manera más eficiente. Si los datos pueden ser resumidos válidamente por un pequeño número de grupos o clases de objetos, entonces las etiquetas para cada grupo proveen de una descripción concisa de los patrones de similitudes y diferencias en los datos.

La determinación de las clases o grupos puede hacerse básicamente mediante dos criterios:

- **Clasificación supervisada**. Se parte de un conjunto de grupos o clases conocido previamente. Esta clasificación también es conocida como *discriminación*.
- **Clasificación no supervisada**. No se establece ningún grupo o clase previo, aunque es necesario determinar el número de clases que se quiere establecer y dejar que las defina, por ejemplo, un procedimiento estadístico.

En realidad, ambos procedimientos son utilizados porque son complementarios. La clasificación supervisada utiliza para su análisis el conocimiento del evento, pero si este conocimiento no es perfecto pueden escaparse cosas que una clasificación no supervisada detectaría.

Este trabajo se concentrará en la clasificación no supervisada, la cual utiliza algoritmos matemáticos de clasificación automática. Los más comunes son los **algoritmos de conglomerados** los cuales buscan descubrir grupos dentro de los datos.

El análisis de conglomerados se puede utilizar, en la práctica, para agrupar especies naturales (Taxonomía), clasificar seres vivos con los mismos síntomas y características patológicas (Medicina), puede ser útil para agrupar un gran número de encuestados de acuerdo a sus preferencias por producto en particular (Investigación de Mercado), como técnicas de reconocimiento de patrones o formar grupos de pixeles en imágenes digitalizadas enviadas por un satélite desde otro planeta para identificar los diferentes terrenos que existan.

Se debe tener en mente que una clasificación general de un grupo de objetos aunque tiene carácter exploratorio es una teoría científica, y debe ser juzgada por su utilidad más que en términos de verdadero o falso.

1.2. Antecedentes históricos

En la ***taxonomía de los animales y las plantas***, los conglomerados datan desde Aristóteles siendo el modelo moderno esencialmente del botánico suizo Carlos Lineo (1753), quien sentó las bases de la taxonomía moderna

clasificando todos los organismos conocidos en dos grandes grupos: los reinos Plantae y Animalia, cada especie pertenece a grupos que incrementan en tamaño y decrecen en el número de características comunes.

Por ejemplo; el hombre pertenece a los primates, a los mamíferos, a los vertebrados, a los animales. Esta lógica de orden fue originalmente desarrollada para nombrar objetos, tuvo una significancia física en las teorías de la Evolución de Darwin, la cual establece que el hombre tiene ancestros en común en distintos niveles del *árbol de la vida*.

Las técnicas de taxonomía se pueden emplear en otras áreas. La organización del árbol es ahora usada como una estructura de conglomerado estándar.

En la ***Medicina*** se presenta un problema en la clasificación de las enfermedades. La Organización Mundial de la Salud produjo el Manual de la Clasificación Estadística Internacional de Enfermedades, Heridas y Causas de Muerte (1965), el cual provee una nomenclatura estándar para colectar estadísticas especialmente de las muertes siendo estas comparables a través de distintos países y épocas, pero dicha teoría general de clasificación de las enfermedades aún resulta desconocida para la taxonomía biológica. Como remarca Feinstein (1972), la nosología (parte de la medicina que estudia la clasificación, descripción y diferenciación de las enfermedades) difiere de dicha taxonomía biológica al estar orientada al tratamiento; ya que es importante separar enfermedades que requieren diferente tratamiento.

Otros ejemplos de agrupamiento se ilustran en Baron y Fraser (1968), en un experimento de métodos de conglomerados sobre 50 pacientes de cirrosis medidos en 330 características, mostrando que el *algoritmo simple o del vecino más cercano* se ajusta menos al diagnóstico previo que el *algoritmo del promedio*; Bouckart (1971) usó el *algoritmo del vecino más cercano* para elegir grupos de pacientes de 85 personas que presentaban "gota" y también seleccionó tres síndromes (conglomerado de síntomas).

Las afecciones del cuerpo no son tan alusivas como las enfermedades de la mente, por lo tanto en ***Psiquiatría*** hay un acuerdo en la existencia de paranoia, esquizofrenia y depresión; dichas categorías pueden ser vistas en la clasificación de Kant publicada en 1970. La dificultad de clasificar características de una enfermedad mental es subjetiva, sutil y de carácter variable dependiendo de los síntomas. Una de las contribuciones tempranas más conocidas de conglomeración corresponde a Zubin, que discute el método del descubrimiento de subgrupos en pacientes esquizofrénicos. El algoritmo es claramente orientado a un cálculo manual, no claramente descrito. Su grupo

esquizofrénico estaba en promedio más cerca a un grupo "normal" que a uno con la enfermedad. Lo cual ilustra la variabilidad del problema.

La conglomeración en el campo de la ***Antropología y Arqueología*** se puede ver reflejada en el descubrimiento de objetos como herramientas de piedras, objetos funerarios, piezas de cerámica, estatuas ceremoniales, o cráneos que pueden ser clasificados dentro de grupos de objetos similares, cada grupo producido por una misma civilización.

Pasa lo mismo en la ***Fitosociología***, la cual se encarga de la distribución espacial de las distintas especies de plantas y animales, sustenta la misma relación con la taxonomía que la epidemiología con la clasificación de las enfermedades. La información típica consiste en contar el número de especies en varios cuadrantes y el conglomerado detecta cuadrantes similares al ser del mismo tipo de habitat.

En ***Lingüística***, Dyen (1967) usó la proporción de unir palabras de una lista de 196 significados, como medida de distancia entre dos lenguajes, con el fin de reconstruir un árbol de lenguajes evolutivo.

En ***Astronomía***, Abell (1960) encontró miles de grupos de galaxias a través de la búsqueda de placas fotográficas sobre una gran fracción del cielo, de hecho es uno de los pioneros en aportar información a la astronomía extragaláctica. Faúndez-Abans (1996) aplicaron técnicas para conglomerados, debido a Ward (1963), a la información en la composición química de 192 nebulosas planetarias. Seis grupos fueron identificados los cuales fueron similares en varios aspectos a una clasificación usada previamente de tales objetos, pero también mostraron diferencias interesantes.

Otro ejemplo aplicado a la astronomía viene de Celeux y Govaert (1995) quienes aplicaron modelos de mezclas normales a información estelar, que consistía en una población de 2370 estrellas descritas por su velocidad hacia el centro galáctico y la rotación galáctica. Usando un modelo de tres conglomerados, encontraron uno de gran tamaño y pequeño volumen, y dos de pequeño tamaño y gran volumen.

Con el análisis de conglomerados, hoy en día, ya se puede responder los siguientes cuestionamientos ¿Cuáles son las empresas en las que sería más deseable invertir?, ¿es posible identificar grupos de clientes a los que les pueda interesar un nuevo producto que una empresa va a lanzar al mercado?, ¿se pueden clasificar las cavas de vino en función de las características químicas y ópticas del vino que producen?, ¿es posible clasificar las estrellas del cosmos en función de su luminosidad?

Capítulo 2

Información preliminar

El marco matemático para la teoría de decisión estadística es proporcionado por la teoría de la probabilidad, que a su vez tiene su fundamento en la teoría de la medida. Este capítulo sirve para definir y recordar algunos de los conceptos básicos de estas teorías, notación y afirmar sin pruebas, en ciertos casos, algunos de los principales resultados.

2.1. Probabilidad

Definición. 1 (Espacio muestral)
El espacio muestral, generalmente denotado por Ω*, es el conjunto de todos los posibles resultados de un experimento aleatorio (ejemplo, lanzar un dado).*

Definición. 2 (Función)
Una función con dominio A *y contradominio* B*, es un conjunto de pares ordenados* (a, b) *que satisface los siguientes puntos:*

- $a \in A$ *y* $b \in B$*.*
- *Cada* a *aparecerá como el primer elemento de algún par ordenado en el conjunto y cada* b *será el segundo elemento de algún par ordenado.*
- *No existen dos pares ordenados distintos que tengan el mismo primer elemento.*

Definición. 3 (Espacio de probabilidad)
Un espacio de probabilidad está dado por la tripleta $(\Omega, \mathcal{a}, P[\cdot])$ *donde* Ω *es un espacio muestral,* $\mathcal{a}$ *es un* σ*-álgebra de eventos y* $P[\cdot]$ *es una función de probabilidad con dominio* $\mathcal{a}$*.*

Definición. 4 (Función de probabilidad)
Sea Ω *un espacio muestral y* $\mathcal{a}$ *un conjunto de eventos que forman una* σ*-álgebra de eventos a considerar para algún experimento aleatorio.*
Una función de probabilidad $P[\cdot]$ *es una función con dominio* $\mathcal{a}$ *(álgebra de eventos) y contradominio el intervalo* $[0, 1]$*, la cual satisface los siguientes axiomas:*

- $P[A] \geq 0 \quad \forall A \in \mathcal{a}$*.*
- $P[\Omega] = 1$*.*
- *Si* $A_1, A_2, ...,$ *es una secuencia de eventos mutuamente excluyentes en* $\mathcal{a}$ *(esto es* $A_i \cap A_j = \varnothing \quad \forall i \neq j$ *con* $i, j \in \{1, 2, ...\}$*) y si*

$$\bigcup_{i=1}^{\infty} A_i \in \ a \quad \textit{entonces} \quad P[\bigcup_{i=1}^{\infty} A_i] = \sum_{i=1}^{\infty} P[A_i].$$

Definición. 5 (Probabilidad condicional)
Sea A *y* B *dos eventos en* $\mathcal{a}$ *del espacio de probabilidad dado por* $(\Omega, \mathcal{a}, P[\cdot])$*. La probabilidad condicional del evento* A *dado el evento* B*, denotada por* $P[A|B]$*, se define por*

$$P[A|B] = \frac{P[A \cap B]}{P[B]}, \quad si \quad P[B] \neq 0.$$

Teorema. 1 (Probabilidad total)
Para un espacio de probabilidad dado por $(\Omega, \mathcal{a}, P[\cdot])$*. Si* $B_1, B_2, ..., B_n$ *es un conjunto de eventos mutuamente disjuntos o incompatibles dos a dos en* $\mathcal{a}$*, que satisface*

$$\Omega = \bigcup_{i=1}^{n} B_i \quad y \quad P[B_i] > 0, \quad (para \ i = 1, 2, ..., n).$$

entonces para toda $A \in \mathcal{a}$*, se tiene que*

$$P[A] = \sum_{j=1}^{n} P[A|B_j]P[B_j].$$

Teorema. 2 (Teorema de Bayes)
Para un espacio de probabilidad dado por $(\Omega, \mathcal{a}, P[\cdot])$. *Si* $B_1, B_2, ..., B_n$ *es un conjunto de eventos mutuamente disjuntos en* $\mathcal{a}$ *satisfaciendo que* $\Omega = \bigcup_{j=1}^{n} B_j$ *y* $P[B_j] > 0$ *para* $j = 1, 2, ..., n$. *Entonces para toda* $A \in \mathcal{a}$ *para el cual* $P[A] > 0$, *se tiene que*

$$P[B_k|A] = \frac{P[A|B_k]P[B_k]}{\sum_{j=1}^{n} P[A|B_j]P[B_j]}.$$

Definición. 6 (Variable aleatoria)
Se dice que una variable aleatoria (v.a.), X, *es una función real definida en el espacio muestral asociado a un experimento aleatorio.*

$$X : \Omega \to \mathbb{R}$$

El **rango** [1] de una v.a., denotado por R_X, es el conjunto de los valores reales que esta puede tomar.

Definición. 7 (Función de distribución de una v.a.)
Sea X una v.a. real, a la función definida como $F_X(x) = P[X \leq x]$, *se le llama función de distribución acumulada o simplemente función de distribución de X; de manera que se cumplan las siguientes condiciones:*

I.
$$\lim_{x \to -\infty} F_X(x) = 0 \qquad y \qquad \lim_{x \to \infty} F_X(x) = 1.$$

II. $F_X(x)$ *es continua por la derecha.*

[1] Al subconjunto del codominio formado por todos los valores que puede tomar la función o rango.

El codominio de una función $f : X \to Y$ es el conjunto Y que participa en esa función, y se denota Cod_f. Sea Im_f la imagen de una función , entonces $Im_f \subseteq Cod_f$.

III. $F_X(x)$ *es monótona creciente* [2].

Definición. 8 *Se dice que una función de distribución F de la v.a. X es continua si existe una función no negativa $f : \mathbb{R} \to \mathbb{R}$ integrable, tal que:*

$$F(x) = \int_{-\infty}^{x} f(y)\mathrm{d}y \quad \forall\ x \in \mathbb{R}.$$

En este caso la v.a. X es absolutamente continua y f es llamada función de densidad de X.

Propiedades de la función de densidad

Si X es una v.a. continua y f su función de densidad, entonces:

1) $f(x) \geq 0 \quad \forall\ x \in \mathbb{R}$.

2) $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$.

Distribución Normal

Para el presente trabajo, es de especial interés, la distribución Normal de parámetros μ y σ^2, definida como $N(\mu, \sigma^2)$ y cuya función de densidad es

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big[-\frac{(x-\mu)^2}{2\sigma^2} \Big] \ \ \textit{donde} \ \ x \in \mathbb{R}, \ \ \mu \in \mathbb{R} \ \ y\ \sigma^2 > 0.$$

Algunas propiedades de la distribución Normal $N(\mu, \sigma^2)$:

1) Si X es $N(\mu, \sigma^2)$ entonces $Y = aX + b$ es $N(a\mu + b, a^2\sigma^2)$.

2) Si X_1 es $N(\mu_1, \sigma_1^2)$, X_2 es $N(\mu_2, \sigma_2^2)$ y son variables independientes, entonces $X_1 + X_2$ es $N(\mu_1 + \mu_1, \sigma_1^2 + \sigma_2^2)$.

[2] Sea $f(x)$ una función definida en $[a, b]$.

a) f es *creciente* en $[a, b]$ si y sólo si se cumple que $x_1 < x_2 \Rightarrow f(x_1) < f(x_2)\ \forall\ x_1, x_2 \in [a, b]$.

b) f es *decreciente* en $[a, b]$ si y sólo si se cumple que $x_1 < x_2 \Rightarrow f(x_1) > f(x_2)\ \forall\ x_1, x_2 \in [a, b]$.

c) f es *monótona* en $[a, b]$ si y sólo si f es creciente o decreciente en $[a, b]$.

Definición. 9 (Esperanza de una v.a.)
La esperanza (o valor esperado) de una v. a. X, generalmente denotada por E(X), es el número que formaliza la idea de valor medio de un fenómeno aleatorio.

$$\mu_x = E[X] = \begin{cases} \sum_x x f_X(x) & \textit{caso discreto} \\ \int_{-\infty}^{\infty} x f_X(x) \mathrm{d}x & \textit{caso continuo} \end{cases}$$

El valor esperado es un concepto fundamental en el estudio de las distribuciones de probabilidad, es considerado como el promedio ponderado de los resultados que se esperan en el futuro.

Definición. 10 (Varianza de una v.a.)
En el caso continuo, la varianza queda definida como:

$$Var[X] = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x) \mathrm{d}x = E[X^2] - \mu_x^2,$$

siempre que $E[X^2]$ exista; además μ_x^2 representa una medida de centralización y σ^2 una medida de dispersión de una variable aleatoria X respecto a su esperanza E[X].

La varianza de una v.a. proporciona información acerca de la variabilidad de las observaciones alrededor de la media. A la raíz cuadrada de la varianza se le conoce como ***desviación estándar*** σ.

Definición. 11 (Vector aleatorio)
Se llama vector aleatorio o variable aleatoria multidimensional a una función $\underline{X} : (X_1, X_2, \ldots, X_p) : \Omega \to \mathbb{R}^p$, donde cada X_i $(i = 1, \ldots, p)$ es una variable aleatoria.

Un vector aleatorio (variable aleatoria *p-dimensional*) es el resultado de observar p características en un elemento de la población. Por ejemplo, si se observa la edad y el peso de los estudiantes de la universidad se obtienen los valores de una variable aleatoria bidimensional.

Función de densidad y distribución conjunta
La ***función de distribución conjunta*** de $X_1, ..., X_p$, vector aleatorio, se define como:

$$F_{X_1,\ldots,X_p}(x_1,\ldots,x_p) = P(X_1 \leq x_1,\ldots,X_p \leq x_p) : \mathbb{R}^p \to [0,1]$$

Se dice que el vector aleatorio $X_1,\ldots,X_p$ es absolutamente continuo si existe una ***función de densidad conjunta*** tal que la función de distribución de $X_1,\ldots,X_p$ verifique que

$$F_{X_1,\ldots,X_p}(x_1,\ldots,x_p) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_p} f(x_1,\ldots,x_p)\,\mathrm{d}x_1\,\mathrm{d}x_2\cdots\,\mathrm{d}x_p$$

Por tanto, la función de densidad conjunta, $f_{X_1,..,X_p}(x_1,..,x_p) : \mathbb{R}^p \to [0,1]$, debe cumplir:

- $$f(x_1,\ldots,x_p) \geq 0$$
- $$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1,\ldots,x_p)\,\mathrm{d}x_1\,\mathrm{d}x_2\cdots\,\mathrm{d}x_p = 1$$

Funciones de densidad y distribución marginal

Cada variable componente de un vector aleatorio $(X_1,\ldots,X_p)$ es una variable aleatoria unidimensional que recibe el nombre de variable marginal. Por ejemplo, para variables bidimensionales continuas (X_1,X_2) con función de densidad conjunta $f(x_1,x_2)$, la ***función de densidad marginal*** para cada variable se define como:

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1,x_2)\,\mathrm{d}x_2 \qquad y \qquad f_{X_2}(x_2) = \int_{-\infty}^{\infty} f(x_1,x_2)\,\mathrm{d}x_1$$

Sea una función de distribución conjunta $F_{X_1,X_2}(x_1,x_2)$. Se define

$$F_{X_1}(x_1) = \lim_{x_2\to\infty} F_{X_1,X_2}(x_1,x_2) \qquad y \qquad F_{X_2}(x_2) = \lim_{x_1\to\infty} F_{X_1,X_2}(x_1,x_2)$$

a cada una de estas funciones se les llama ***función de distribución marginal***.

Definición. 12 (Independencia)
Dadas $X_1, \ldots, X_p$ variables aleatorias con funciones de densidad de probabilidad $f_{X_1}(x_1), \ldots, f_{X_1}(x_p)$, respectivamente, se dice que son independientes si

$$f_{X_1,\ldots,X_p}(x_1, \ldots, x_p) = f_{X_1}(x_1) f_{X_1}(x_2) \ldots f_{X_p}(x_p), \quad \textit{donde}$$

$f_{X_1,\ldots,X_p}(x_1, \ldots, x_p)$ designa a la función de densidad conjunta de $X_1, \ldots, X_p$.

Funciones de densidad y distribución condicionada
Sean (X_1, X_2), vector aleatorio bidimensional con variables continuas, la ***función de densidad condicional*** de X_1 se define como

$$f_{X_1|X_2}(x_1, |x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} \quad \textit{cuando} \quad f_{X_2}(x_2) \neq 0.$$

La ***función de distribución condicional*** de X_1 se define como

$$F_{X_1|X_2}(x_1, |x_2) = \int_{-\infty}^{x_1} f_{X_1,X_2}(u, x_2) \mathrm{d}u$$

Definición. 13 (Covarianza)
Sean X_i, X_j variables aleatorias, la covarianza entre estas dos variables queda definida como

$$c_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)] = E[X_i \cdot X_j] - \mu_i \cdot \mu_j,$$

siempre que μ_i, μ_j y $E[X_i \cdot X_j]$ existan; siendo μ_i y μ_j las esperanzas de X_i y X_j, además

$$E[X_i \cdot X_j] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i x_j f(x_1, \ldots, x_p) \, \mathrm{d}x_1 \, \mathrm{d}x_2 \cdots \mathrm{d}x_p$$

Definición. 14 (Coeficiente de correlación)
El coeficiente de correlación queda definido como

$$\rho(X_i, X_j) = \frac{c_{ij}}{\sigma_i \sigma_j}.$$

El valor del índice de correlación varía en el intervalo $[-1, 1]$. Cuando $\rho = 1$, una variable es exactamente combinación de la otra, y si $\rho = 0$ no existe correlación lineal.

2.2. Inferencia estadística

La *inferencia estadística* comprende procedimientos que permiten deducir sobre población a partir de la información que proporciona una muestra tomada de ella.

En el campo de la estadística se denominan parámetros a todas aquellas medidas que expresan alguna característica general de una población, tales como la media de los valores que toma una variable en todos los individuos de la población, la varianza de estos valores, la proporción de individuos que poseen determinada característica, etc. Para todos estos ejemplos de parámetros el valor suele ser desconocido porque para su cálculo sería necesario observar a la totalidad de los individuos que componen la población, algo imposible en la mayoría de las situaciones; cuando más se podrá observar a una muestra (más o menos grande) de individuos de esta población.

Con la información recogida en los datos de una muestra se puede hacer una aproximación al conocimiento de la población, en particular, al valor de sus parámetros.

De forma general se distinguen dos grandes categorías de métodos de inferencia; métodos para estimación de parámetros y para contraste de hipótesis.

Algunos conceptos que se utilizan en este contexto son los siguientes:

Definición. 15 *Al conjunto de valores donde el o los parámetros de una distribución toma valores se le llama* espacio parametral *y se denota con la letra griega mayúscula* Θ.

Definición. 16 *Colección de* $X_1, \ldots, X_n$ *variables aleatorias independientes e idénticamente distribuidas, es decir,*

1)

$$X_1, ..., X_n \text{ es una muestra aleatoria si } X_i \sim f(x_i;\theta) \quad \forall i$$

2)

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\theta)$$

Definición. 17 *Al conjunto de valores que puede tomar la muestra aleatoria se le llama* espacio muestral.

2.2.1. Estimación de parámetros

El valor de un parámetro se estima a partir de alguna medida (estimador) calculada a partir de los datos de una muestra, que pueda proporcionar un valor aproximado (estimación) del parámetro.

Existen distintos métodos de estimación entre ellos está el método de los momentos, máxima verosimilitud o mínimos cuadrados. En este trabajo se trabajará con la estimación por máxima verosimilitud.

2.2.2. Estimación por máxima verosimilitud

El método de *máxima verosimilitud*, debido a Fisher, escoge como estimador de los parámetros a aquel valor que hace máxima la probabilidad de que el modelo a estimar genere la muestra observada. Sea $X = (x_1, x_2, ..., x_n)$ una muestra aleatoria de una población con función de probabilidad $f(x; \theta)$. Se define la *función de verosimilitud*, denotada como $l(\theta)$, como la densidad conjunta de la muestra aleatoria, es decir,

$$l(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

En general, para una muestra aleatoria de $f(x; \theta_1, ...\theta_k)$ se tiene que el estimador máximo verosímil $(\hat{\theta}_1, ..., \hat{\theta}_k)$ satisface las siguientes condiciones (es solución al sistema):

1. $$\frac{\mathrm{d}\, l(\theta_1, ..., \theta_k)}{\mathrm{d}\, \theta_1} = 0 \ , \cdots , \ \frac{\mathrm{d}\, l(\theta_1, ..., \theta_k)}{\mathrm{d}\, \theta_k} = 0$$

   donde $l(\theta_1, ..., \theta_k)$ es la función de verosimilitud correspondiente.

2. El estimador de máxima verosimilitud, es el valor de θ que hace máxima la probabilidad de aparición de los valores muestrales observados y se obtiene calculando el valor máximo de la función $l(\theta)$.

3. En la práctica, suele ser más sencillo obtener el máximo del logaritmo de la función de verosimilitud:

   $$L(\theta) = ln\, l(\theta)$$

   función que llamaremos *logverosimilitud*.

2.3. Introducción conceptual al análisis multivariado

El análisis multivariante, en esencia, se dedica al estudio de variables de modo simultáneo. Es decir, se toma un objeto y no sólo se puede medir una característica suya (ej. la estatura de una persona), sino que se consideran varios aspectos y se trata de determinar la relación entre estas medidas. Se suele resolver mediante álgebra matricial, cuyos cálculos constituyen la mecánica del análisis .

2.3.1. Conceptos elementales de álgebra matricial

Definición. 18 (Matriz transpuesta)
Sea A una matriz de $m \times n$, la matriz transpuesta *de A es aquella que se obtiene al cambiar en A las filas por columnas o las columnas por filas. Se representa por A^t y su dimensión es $m \times n$.*

Definición. 19 (Matriz cuadrada)
Una matriz de $n \times m$ es una matriz cuadrada si el número de filas es igual al número de columnas, es decir, $n = m$. Se dice entonces que la matriz es de orden n.

Definición. 20 (Matriz diagonal)
Una matriz diagonal *es aquella matriz $A_{n\times n}$, la cual todas sus entradas son nulas salvo la diagonal principal. Es decir, $A = (a_{i,j})$ será diagonal si $a_{i,j} = 0$ para $i \neq j$.*

Definición. 21 (Matriz identidad)
Sea $A_{n\times n}$ una matriz con unos en la diagonal principal y ceros en cualquier otra posición, denotada con $I_{n\times n}$, se conoce como matriz identidad *(o unidad). Para cualquier matriz A,*

$$A * I = I * A = A.$$

Definición. 22 (Traza de una matriz)
La traza de una matriz *es la suma de los elementos de la diagonal principal de la matriz. Si* A *es una matriz con elementos a_{ij} se verifica:*

$$tr(A) = \sum_{i=1}^{n} a_{ii}.$$

Definición. 23 (Matriz invertible)
Se dice que una matriz $A_{n\times n}$ es invertible, o matriz no singular*, si existe una matriz $B_{n\times n}$, la cual llamaremos matriz inversa de A, que cumple:*

$$A * B = I_{n\times n} \quad y \quad B * A = I_{n\times n}.$$

Una matriz invertible sólo tiene una inversa, es decir, la inversa es única y se denota por A^{-1}, donde,

$$A * A^{-1} = I_{n\times n} \quad y \quad A^{-1} * A = I_{n\times n}.$$

Definición. 24 (Matriz ortogonal)
Sea $A_{n\times n}$ una matriz. Se dice que la matriz es ortogonal *si es invertible y si $A^{-1} = A^t$. Nótese que con esta igualdad equivale a cualquiera de las siguientes relaciones:*

$$A^t * A = I_{n\times n}$$

$$A * A^t = I_{n\times n}$$

Definición. 25 (Matriz simétrica)
Una matriz de $m \times n$ es simétrica*, si es una matriz cuadrada $(m = n)$ y $a_{ij} = a_{ji}\ \forall\ i \neq j$ con $i, j = 1, 2, \ldots, n$.*

Nótese que la simetría es respecto a la diagonal principal, además se puede verificar:

$$A = A^t.$$

Definición. 26 (Matriz singular)
Una matriz $A_{n\times n}$ es singular *si su determinante es nulo. En tal caso se dice que dicha matriz no tiene inversa.*
Son equivalentes las siguientes observaciones:

- *$A_{n\times n}$ es no invertible.*
- *El determinante de $A_{n\times n}$ es nulo $(det|A| = 0)$, esto es, $A_{n\times n}$ es singular.*
- *$Ax = 0$ tiene soluciones infinitas.*

Definición. 27 (Matriz triangular)
Una matriz triangular *es aquella matriz cuadrada que sólo tiene registros ceros arriba (matriz triangular inferior) o abajo (matriz triangular superior) de la diagonal principal.*

El determinante de una matriz triangular es igual al producto de sus elementos diagonales.

Definición. 28 (Matriz definida positiva)
Una matriz A es definida positiva cuando $x^tAx > 0$. Siendo A una matriz cuadrada simétrica, x un vector columna distinto de cero y x^t el vector transpuesto de x.

2.3.2. Matriz de datos

El punto de partida del análisis multivariante es una ***matriz de datos*** que contiene n objetos y p variables. Estas variables serán presentadas como $X_1, \ldots, X_j, \ldots, X_p$.

	$Variables$					
$Objetos$	X_1	X_2		X_j		X_p
1	x_{11}	x_{12}	$\cdots$	x_{1j}	$\cdots$	x_{1p}
2	x_{21}	x_{22}	$\cdots$	x_{2j}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$
i	x_{i1}	x_{i2}	$\cdots$	x_{ij}	$\ldots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{nj}	$\ldots$	x_{np}

Algunas medidas de tendencia central y dispersión para el conjunto de variables, y medidas de tendencia lineal entre pares de variables son:

2.3.3. Vector de medias

$$E[\underline{X}] = (E(X_1), \ldots, E(X_P))^t = \bar{\mu},$$

donde cada $E(X_i)$ es la media aritmética de la i-ésima variable con $i = 1, \ldots, p$.

2.3.4. Matriz de varianzas y covarianzas

La matriz de varianzas y covarianzas queda definida como:

$$Var[\underline{X}] = E[(\underline{X} - E(\underline{X}))(\underline{X} - E(\underline{X}))^t] = E[(\underline{X} - \bar{\mu})(\underline{X} - \bar{\mu})^t]$$

$$(\underline{X} - \bar{\mu})(\underline{X} - \bar{\mu})^t = \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_n - \mu_n \end{pmatrix} (x_1 - \mu_1, x_2 - \mu_2, \ldots, x_n - \mu_n)$$

$$= \begin{pmatrix} (x_1 - \mu_1)^2 & (x_1 - \mu_1)(x_2 - \mu_2) & \cdots & (x_1 - \mu_1)(x_n - \mu_n) \\ (x_2 - \mu_2)(x_1 - \mu_1) & (x_2 - \mu_2)^2 & \cdots & (x_2 - \mu_2)(x_n - \mu_n) \\ \vdots & \ddots & & \vdots \\ (x_n - \mu_n)(x_1 - \mu_1) & (x_n - \mu_n)(x_2 - \mu_2) & \cdots & (x_n - \mu_n)^2 \end{pmatrix}$$

$$= \begin{pmatrix} Var(x_1) & Cov(x_1, x_2) & \cdots & Cov(x_1, x_n) \\ Cov(x_2, x_1) & Var(x_2) & \cdots & Cov(x_2, x_n) \\ \vdots & \ddots & & \vdots \\ Cov(x_n, x_1) & Cov(x_n, x_2) & \cdots & Var(x_n) \end{pmatrix} = \Sigma$$

Σ es una matriz simétrica que contiene en la diagonal las varianzas muestrales y fuera de ella las covarianzas muestrales entre cada uno de las variables.

2.3.5. Matriz de correlaciones

La matriz de correlaciones queda definida como sigue:

$$\boldsymbol{P} = \begin{pmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2n} \\ \vdots & \ddots & & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & \rho_{nn} \end{pmatrix}, \quad donde \quad \rho_{ii} = 1.$$

P es una matriz cuadrada y simétrica que contiene en la diagonal unos y fuera de ella todos los coeficientes de correlación lineal entre cada una de los pares de variables.

2.3.6. Eigenvalores y eigenvectores de una matriz

Sea $\Sigma_{n\times n}$ una matriz cuadrada, se dice que λ es un valor propio (*eigenvalor*) de Σ si existe un vector no nulo $\bar{x} \in \mathbb{R}^n$, tal que, $\Sigma\bar{x} = \lambda\bar{x}$, equivalente a

$$(\Sigma - \lambda I)\bar{x} = 0,$$

el vector $\bar{x}$ es llamado vector propio (*eigenvector*) de Σ. Los vectores propios son por lo general son normalizados.

El polinomio en λ de grado n, $P(\lambda) = det|\Sigma - \lambda I|$ es llamado *polinomio característico* de Σ y los n eigenvalores de Σ son las n raíces de $P(\lambda)$.

Algunas propiedades

- Si λ es un eigenvalor de Σ y $\bar{x}$ es su eigenvector correspondiente entonces λ y $B\bar{x}$ son eigenvalor y eigenvector, respectivamente, de BAB^{-1}.
- Si Σ es triangular entonces sus eigenvalores son los elementos de la diagonal.
- Si Σ es una matriz simétrica entonces todos sus eigenvalores son reales.
- Traza(Σ) es igual a la suma de los eigenvalores de Σ.
- Si Σ es simétrica entonces existe una matriz ortogonal U, tal que, $U^t\Sigma U = D$ donde D es una matriz diagonal cuyas entradas son los eigenvalores de Σ y las columnas de U son los eigenvectores correspondientes. De igual manera, $\Sigma = UDU^t$.

2.3.7. Descomposición espectral

Esta descomposición liga a los eigenvalores y eigenvectores de una matriz a su estructura. El procedimiento es el siguiente:

Sea Σ una matriz cuadrada con una dimensión de $k \times k$ y simétrica. Σ puede escribirse de la forma siguiente:

$$\Sigma = \lambda_k D_k A_k D_k^t,$$

donde λ es una constante, $A_k = diag(a_1, ..., a_p)$ y $D_k = (\underline{d}_1, ..., \underline{d}_k)$ es una matriz ortogonal que consiste de los eigenvectores de la matriz Σ.

Capítulo 3

Clasificación no supervisada

Como se dijo en la introducción, este trabajo se concentrará en la clasificación no supervisada en especial en el **análisis de conglomerados**. Este análisis es una técnica multivariante que tiene como propósito agrupar observaciones de forma que los datos sean muy homogéneos dentro de los grupos (mínima varianza) y que estos grupos sean lo más heterogéneos posible entre ellos (máxima varianza); pero ¿para qué agrupar o formar grupos? Se forman grupos para que de este modo se obtenga una clasificación de los datos multivariantes con la que se pueda comprender mejor los mismos y la población de la que proceden. El número de grupos y otra información sobre su composición puede ser desconocida.

La idea preliminar de este análisis es esencialmente sobre el descubrimiento de grupos de datos, teniendo entonces como el objetivo construcción de reglas para la clasificación de nuevos individuos dentro del mismo u otro grupo conocido. Los métodos de conglomerados no deberían ser confundidos con métodos de *discriminación* y *asignación* donde los grupos son conocidos previamente.

En general este análisis es una técnica notablemente exploratoria puesto que la mayoría de las veces, no utiliza algún modelo estadístico para llevar a cabo el proceso de clasificación. Se le podría calificar como una técnica adecuada para extraer información de un conjunto de datos sin imponer restricciones previas como los modelos estadísticos, al menos de forma explícita, y por ello puede llegar a ser muy útil como una herramienta de elaboración de hipótesis acerca del problema considerado sin imponer patrones o teorías previamente establecidas. Sin embargo, conviene estar siempre alerta ante el

peligro de obtener no una clasificación de los datos sino una disección de los mismos, así el conocimiento que el analista tenga acerca de esta clasificación decidirá cuál de los grupos, realmente, será importante y cuál no.

El análisis de conglomerados estudia diferentes tipos de problema:

a) *Partición de los datos*. Se quiere agrupar elementos en grupos homogéneos en función de las similitudes entre ellos, se tiene que:

   - Cada elemento pertenecerá a uno y sólo uno de los grupos.
   - Todo elemento quedará clasificado.

b) *Construcción de jerarquías*. En una clasificación jerárquica los datos se ordenan en niveles de modo que los niveles superiores contienen a los inferiores. Estos métodos no definen grupos, sino la estructura de asociación de cadena que pueda existir entre los elementos. Sin embargo la jerarquía construida permite obtener también una partición de los datos en grupos.

c) *Clasificación de variables*. Las variables pueden clasificarse en grupos o estructurarse en una jerarquía.

Los métodos de clasificación van desde aquellos que son muy heurísticos hacia procedimientos más formales basados en modelos estadísticos.

Hay básicamente tres tipos de métodos para conglomerados:

- Métodos de partición.
- Métodos jerárquicos.
- Métodos basados en modelos.

3.1. Métodos de partición

Los ***métodos de partición*** utilizan la matriz de datos.

El algoritmo más conocido dentro de los métodos de partición, es el *K-medias*, es muy utilizado en aplicaciones científicas e industriales.

El nombre le viene porque representa cada uno de los grupos por la media (o media ponderada) de sus puntos, es decir, por su centroide.

Se debe determinar los K grupos iniciales con los que comenzará el algoritmo, existen diferentes maneras de obtenerlos, una de ellas es seleccionar los K casos más distantes entre sí. El algoritmo comienza a leer la información asignando cada caso al centro más próximo y actualizando el valor de los centros a medida que se van analizando e incorporando más casos. Una vez que todos los casos han sido asignadas a uno de los K grupos, se inicia un proceso iterativo para calcular los centroides finales de esos K grupos.

El análisis de conglomerados de *K- medias* es especialmente útil cuando se dispone de un gran número de observaciones. Existe la posibilidad de utilizar la técnica de manera exploratoria clasificando los casos e iterando para encontrar la ubicación de los *centroides*, o sólo como una técnica de clasificación, catalogando las observaciones a partir de centroides conocidos determinados por el analista. Cuando se utiliza como una técnica exploratoria, es habitual que se desconozca el número idóneo de grupos, por lo que es conveniente repetir el análisis con distinto número de grupos y comparar las soluciones obtenidas; en estas ocasiones también puede utilizarse algún método de conglomerados *jerárquico* con una submuestra de casos.

Se considera una muestra de n elementos o casos y p variables, el objetivo es dividir esta muestra en K. El algoritmo de las *K- medias* requiere de las siguientes etapas:

- Seleccionar K puntos como centros de los grupos iniciales.
- Calcular las distancias euclidianas de cada elemento con respecto a los K centros y asignar cada elemento al grupo cuyo centro esté más próximo.
- Definir un criterio de optimalidad y comprobar si reasignando alguno de los elementos mejora el criterio.
- Si no es posible mejorar el criterio de optimalidad, fin del proceso.

3.1.1. Implementación del algoritmo *K-medias*

El i-ésimo caso de la j-ésima variable será denotado como *a(i,j)* ($i = 1, ..., n, j = 1, ..., p$). Las variables se escalan de manera que la distancia euclidiana es apropiada . La partición *p(n,G)* está compuesta de los grupos *1,2,...,G*. Cada uno de los n casos se encuentran en sólo uno de los G grupos. La media de la j-ésima variable sobre los casos en el k-ésimo grupo es denotada por *b(k,j)*. El número de casos en k será *n(k)*. La distancia entre el i-ésimo caso y el k-ésimo grupo es

$$d(k, j) = \sqrt{\sum_{j=1}^{p}[a(i, j) - b(k, j)]^2}. \tag{3.1}$$

El error en la partición es

$$e[p(n, G)] = \sum_{i=1}^{n} d[i, k(i)]^2, \tag{3.2}$$

donde *k(i)* es el grupo que contiene el i-ésimo caso.

El procedimiento general es la búsqueda de una partición con e pequeño moviendo los casos desde un grupo a otro. La búsqueda termina cuando ningún movimiento reduce e.

Esto es análogo al criterio de optimalidad, el cual minimiza la *suma de cuadrados dentro de los grupos (SCDG)* para las variables

$$SCDG = \sum_{k=1}^{G}\sum_{j=1}^{p}\sum_{i=1}^{n_k}(x_{ijk} - \bar{x}_{jk})^2, \tag{3.3}$$

donde x_{ijk} es el valor de la variable j en el elemento i del grupo k y $\bar{x}_{jk}$ es la media de ese grupo.

Paso 1.
Asumir agrupaciones iniciales *1,2,...,G*. Calcular las medias de los grupos *b(k,j)* ($k = 1, ..., G, j = 1, ..., p$) y el error inicial

$$e[p(n, G)] = \sum_{i=1}^{n} d[i, k(i)]^2,$$

donde *d[i,k(i)]* denota la distancia euclidiana entre k y la media del grupo donde está contenida i.

Paso 2.
Para el primer caso, calcular para cada grupo k

$$\frac{n(k)d(1,k)^2}{n(k)+1} - \frac{n[k(1)]d[i,k(1)]^2}{n[k(1)]-1}. \tag{3.4}$$

Con la fórmula anterior, se obtiene el incremento en el error al transferir el primer caso del grupo *k(1)*, al cual pertenece actualmente, al grupo *k*. Si el mínimo de esta cantidad sobre todos los $k \neq k(1)$ es negativo, transferimos el primer caso desde *k(1)* al mínimo de los k grupos, y agregamos el incremento de error (el cual es negativo) *e[p(n,G)]*.

Paso 3.
Repetir el paso 2 para el i-ésimo caso ($2 \leq i \leq n$).

Paso 4.
El algoritmo se detiene, si ya no se mueven los casos desde un grupo a otro. De otra manera, regresar al paso 2.

Ejemplo del Algoritmo k-medias

La tabla *Nutrients in Meat, Fish and Fowl* (Hartigan [1]) contiene 27 alimentos en total, para una mejor manipulación de la información, sólo los primeros 8 serán considerados para este ejemplo según sus aportes en calorías, proteínas y calcio como un porcentaje de cantidades diarias recomendadas.

No. de Casos		*Energia* (*kcal*)	*Proteina* (*g*)	*Calcio* (*mg*)
BB	Beef, Braised	11	29	1
HR	Hamburguer	8	30	1
BR	Beef, Roast	13	21	1
BS	Beaf, Steak	12	27	1
BC	Beaf, Canned	6	31	2
CB	Chicken, Broiled	4	29	1
CC	Chicken, Canned	5	36	1
BH	Beef, Heart	5	37	2

Cuadro 3.1: Tabla de Aplicación del Algoritmo *k-medias* a información de alimentos.

Paso 1.
Se obtienen los conglomerados iniciales a través de la suma de los casos, denotado por **SUM(*i*)** con un valor mínimo MIN y un valor máximo MAX. Para obtener los k grupos iniciales, se establece el caso i dentro del *k-ésimo* grupo, donde k es parte integral de la suma de

$$k\frac{SUM(i) - MIN}{MAX - MIN} + 1, \tag{3.5}$$

donde las sumas de los casos son:

i	$Sum(i)$
1	$11 + 29 + 1 = \mathbf{41}$
2	$8 + 30 + 1 = \mathbf{39}$
3	$13 + 21 + 1 = \mathbf{35}$
4	$12 + 27 + 1 = \mathbf{40}$
5	$6 + 31 + 2 = \mathbf{39}$
6	$4 + 29 + 1 = \mathbf{34}$
7	$5 + 36 + 1 = \mathbf{42}$
8	$5 + 37 + 2 = \mathbf{44}$

Los grupos correspondientes a las sumas son 3, 2, 1, 2, 3, 1, 3 y 3, una vez aplicada la fórmula (3.5).
De este modo, la partición inicial queda (BR CB) (HR BS) (BB BC CC BH).
Los valores de *b(k,j)* ($1 \leq k \leq 3, 1 \leq j \leq 3$) serán presentados en la siguiente tabla.

Grupo	*No. de Casos*				e=154.9
		Energia	*Proteinas*	*Calcio*	
1.	BR, CB	8.5	25	1	
2.	HR, BS	10	28.5	1	
3.	BB, BC, CC, BH	6.75	33.25	1.5	

Cuadro 3.2: Medias de los grupos iniciales.

El error para la partición inicial es la suma de las distancias cuadradas de los casos y sus medias de grupo correspondiente:

$e[p(8,3)] =$

$(11-6.75)^2+(29-33.25)^2+(1-1.5)^2+(8-10)^2+(30-28.5)^2+(1-1)^2$

$+(13-8.5)^2+(21-25)^2+(1-1)^2+(12-10)^2+(27-28.5)^2+(1-1)^2$

$+(6-6.75)^2+(31-33.25)^2+(2-1.5)^2+(4-8.5)^2+(29-25)^2+(1-1)^2$

$+(5-6.75)^2+(36-33.25)^2+(1-1.5)^2+(5-6.75)^2+(37-33.25)^2+(2-1.5)^2$

$= 154.9$

Paso 2.

Para el primer caso (BB) las distancias a cada uno de los grupos son:

$$d(1,1)^2 = (11-8.5)^2+(29-25)^2+(1-1)^2 = 22.25$$
$$d(1,2)^2 = (11-10)^2+(29-28.5)^2+(1-1)^2 = 1.25$$
$$d(1,3)^2 = (11-6.75)^2+(29-33.25)^2+(1-1.5)^2 = 36.375$$

Aplicando la fórmula (3.4), el incremento de error en transferir el primer caso al grupo 1 es

$$\frac{2(22.25)}{3} - \frac{4(36.375)}{3} = -33.67$$

y al grupo 2

$$\frac{2(1.25)}{3} - \frac{4(36.375)}{3} = -47.7$$

Por lo tanto, el primer caso (BB) se traslada al grupo 2 cuya reducción de error es 47.7.
El nuevo valor de $e[p(8,3)]$ es por consiguiente $154.9 - 47.7 = 108.2$
Es necesario actualizar las medias de los grupos 2 y 3, ya que ahora el grupo 2 posee a BB. Las nuevas medias son

$d(2,1) =$	$11+8+12/3$	$= 10.33$	$d(3,1) =$	$6+5+5/3$	$= 5.33$
$d(2,2) =$	$29+30+27/3$	$= 28.67$	$d(3,2) =$	$31+36.5+37/3$	$= 34.67$
$d(2,3) =$	$1+1+1/3$	$= 1$	$d(3,3) =$	$2+1+2/3$	$= 1.67$

Grupo	*No. de Casos*			e=108.2
		Energia	*Proteinas*	*Calcio*
1.	BR, CB	8.5	25	1
2.	HR, BS, BB	10.33	28.67	1
3.	BC, CC, BH	5.33	34.67	1.67

Cuadro 3.3: Primer cambio

Paso 3.
Repetir el paso 2 para todos los demás casos (HR, BR, BS, BC, CB, CC, BH).
Nótese que para el caso 2 (HR), el grupo 2 está mucho más cerca que cualquier otro. Esto es

$$d(2,1)^2 = (8-8.5)^2 + (30-25)^2 + (1-1)^2 = 25.25$$
$$d(2,2)^2 = (8-10.33)^2 + (30-28.67)^2 + (1-1)^2 = 7.19$$
$$d(2,3)^2 = (8-5.33)^2 + (30-34.67)^2 + (1-1.67)^2 = 29.38$$

Aplicando la fórmula (3.4), el incremento de error en transferir el caso 2 al grupo 1 es

$$\frac{2(25.25)}{3} - \frac{3(7.19)}{2} = 6.048$$

y al grupo 3

$$\frac{3(29.38)}{4} - \frac{3(7.19)}{2} = 11.25$$

Por lo tanto, el caso 2 permanece en el grupo 2.

No se reportaron cambios hasta el caso 6 (CB), donde

$$d(6,1)^2 = (4-8.5)^2 + (29-25)^2 + (1-1)^2 = 36.25$$
$$d(6,2)^2 = (4-10.33)^2 + (29-28.67)^2 + (1-1)^2 = 40.18$$
$$d(6,3)^2 = (4-5.33)^2 + (29-34.67)^2 + (1-1.67)^2 = 34.37$$

El incremento de error en transferir el caso 6 al grupo 2 es

$$\frac{3(40.18)}{4} - \frac{2(36.25)}{1} = -42.5$$

y al grupo 3

$$\frac{3(34.37)}{4} - \frac{2(36.25)}{1} = -46.72$$

Por lo tanto, el primer 6 se traslada al grupo 3 cuya reducción de error es 46.7.

El nuevo valor de $e[p(8,3)]$ es por consiguiente $108.2 - 46.7 = 61.4$

Paso 4.
Ya que algunos cambios ocurrieron en el paso anterior, nuevamente se deberán actualizar las medias de los grupos.

Grupo	*No. de Casos*				e=61.4
		Energia	*Proteinas*	*Calcio*	
1.	BR	13	21	1	
2.	HR, BS, BB	10.33	28.67	1	
3.	BC, CC, BH, CB	5	33.25	1.5	

Cuadro 3.4: Segundo cambio

Como ningún otro cambio ocurre, el algoritmo se detiene quedando los últimos grupos (BR) (HR BS BB) (BC CC BH CB). Estos grupos son determinados por las variables, de la siguiente manera: El grupo 1 tiene alto contenido en energía y bajo en proteína, el grupo 2 es alto en energía y proteína, y el grupo 3 es bajo en energía y tiene alto contenido en proteína. Para este ejemplo en particular, el calcio no aportó información importante.

3.2. Métodos jerárquicos

Los ***métodos jerárquicos*** se plantean a partir de una matriz de distancias o similaridades entre los elementos de la muestra y construyen una jerarquía basada en estas distancias. Sean los puntos $X = (x_1, x_2, ..., x_n)$ y $Y = (y_1, y_2, ..., y_n) \in \mathbb{R}^n$;

Distancia Minkowsky

$$d(X,Y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}},$$

si $p = 1$ se refiere a la distancia de Manhattan y si $p = 2$, a la euclidiana.
Distancia Manhattan

$$d(X, Y) = \sum_{i=1}^{n} |x_i - y_i|$$

Distancia Euclidiana

$$d(X, Y) = (\sum_{i=1}^{n} |x_i - y_i|)^{\frac{1}{2}}$$

En general, si todas las variables son continuas, la distancia más utilizada es la euclidiana entre las variables estandarizadas de manera univariada.

Esta distancia no es invariante ante cambio de escalas por lo que es aconsejable estandarizar los datos si las unidades de medida de las variables no son comparables.

Para decidir si estandarizar las variables o no antes del análisis debemos considerar el objetivo del estudio. Primeramente, si no estandarizamos la distancia dependerá sobre todo de las variables con valores mayores y el resultado puede cambiar totalmente al modificar su escala de medida. En segundo, si estandarizamos, estamos dando *a priori* un peso semejante a las variables, con independencia de su variabilidad original, lo que puede no ser siempre adecuado.

Si en la muestra existen variables continuas y discretas el problema se complica, ya que puede ocurrir que las variables continuas tengan mayor importancia en el procedimiento de clasificación. Cuando esto no sea deseable la solución es trabajar con similaridades.

La estructura jerárquica es representada en forma de árbol y es llamada **Dendrograma**. Estos son fáciles de interpretar pero pueden conducir a falsas conclusiones ya que la estructura jerárquica del dendograma no representa fielmente las verdaderas distancias o similaridades entre los objetos distintos del conjunto de datos.

Si seguimos con la idea de agrupar datos homogéneos entre grupos y que estos sean lo más heterogéneos entre sí, como sabemos ¿qué tan similar es un individuo en un grupo a otro? Para contestar esto nos podemos ayudar de las medidas de *proximidad*, *similaridad* o *disimilaridad* que miden el grado de semejanza entre dos objetos de forma que, cuanto mayor (resp. menor) es su valor, mayor (resp. menor) es el grado de similaridad existente entre ellos

y con más (resp. menos) probabilidad los métodos de clasificación tenderán a ponerlos en el mismo grupo.

El coeficiente de similaridad entre dos elementos i y h, con base en la variable $j = 1, ..., p$ se define como una función s_{jih} no negativa y simétrica que satisface:

$$s_{jih} = 1,$$
$$0 \leq s_{jih} \leq 1,$$
$$s_{jih} = s_{jhi}.$$

Existen diversas medidas de semejanza y distancia dependiendo del tipo de variables y los diferentes tipos de datos como de intervalo, frecuencias ó datos binarios. Queda a consideración del analista con cuales medidas trabajará .

Los métodos jerárquicos pueden ser:

1) *El algoritmo de división* asume en un primer paso que todos los datos conforman un sólo conglomerado. Este grupo o conglomerado se va dividiendo sucesivamente en conglomerados más pequeños de acuerdo a algún criterio seleccionado previamente.

2) En el *algoritmo de aglomeración* cada observación o caso inicialmente es un conglomerado y en cada paso se asocian los conglomerados más similares hasta llegar a un sólo grupo.

Los Métodos aglomerativos más utilizados para determinar el número de grupos son:

- *Método Simple o vecino más cercano (Single Linkage Method).* Comienza seleccionando y uniendo los dos elementos de la matriz de distancias que se encuentren más próximos. La distancia de este nuevo conglomerado respecto a los restantes elementos de la matriz se calcula como la menor de las distancias entre cada elemento del conglomerado y el resto de los elementos de la matriz. En los pasos sucesivos, la distancia entre dos conglomerados es la mínima de las distancias entre un elemento de un grupo y un elemento del otro grupo. Así, la distancia d_{AB} entre los conglomerados A y B se calcula

$$d_{AB} = min(d_{ij}),$$

donde d_{ij} es la distancia entre los elementos $i \in A$ y $j \in B$.

- *Método Completo o vecino más lejano (Complete Linkage Method).* La distancia entre dos grupos se calcula como la distancia entre sus elementos más alejados. Es decir, la distancia entre dos conglomerados A y B se calcula

$$d_{AB} = max(d_{ij}).$$

Este método tiende a producir grupos alargados.

- *Método del Promedio ó enlace medio entre grupos(Average Method).* A diferencia de los dos métodos anteriores, este método aprovecha la información de todos los miembros de los dos grupos que se comparan. La distancia entre dos conglomerados se calcula como la distancia promedio existente entre los objetos de un grupo y los objetos del otro grupo

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{j \in B} d_{ij},$$

donde n_A es la cardinalidad de A y n_B es la cardinalidad de B.

- *Método de Ward ó de la suma de cuadrados.* Es un procedimiento en el cual en cada etapa, se unen los dos grupos para los cuales se tenga el menor incremento en el valor total de la suma de cuadrados de las diferencias (dentro de cada grupo) de cada individuo al centroide del grupo. Donde,

  - x_{ij}^k es el valor de la j-ésima variable sobre el i-ésimo individuo del k-ésimo grupo, suponiendo que dicho grupo tiene n_k individuos.
  - m^k es el centroide del grupo k con componentes m_j^k.
  - E^k es la suma de cuadrados de los errores del grupo k, o sea, la distancia euclidiana al cuadrado entre cada individuo del grupo k a su centroide.

$$E_k = \sum_{i=1}^{n_k} \sum_{j=1}^{n} (x_{ij}^k - m_j^k)^2.$$

- E será la suma de cuadrados de los errores para todos los grupos, o sea, si suponemos que hay G grupos

$$E = \sum_{k=1}^{G} E_k.$$

El proceso comienza con m grupos, cada uno de los cuales está compuesto por un solo individuo, por lo que cada individuo coincide con el centro del grupo y por lo tanto en este primer paso se tendrá $E_k = 0$ para cada grupo y con ello, $E = 0$. El objetivo del método de Ward es encontrar en cada etapa aquellos dos grupos cuya unión proporcione el menor incremento en la suma total de errores, E.

Los tres últimos métodos no son invariantes ante transformaciones monótonas, es decir no mantienen el orden de la información.

- *Método del Centroide (Centroid Method).* La distancia entre dos grupos es igual a la distancia entre sus vectores de medias. La matriz de distancias original sólo se utiliza en la primera fase. Sucesivamente se maneja la matriz de distancias actualizada en la fase anterior. En cada etapa, el algoritmo utiliza la información de los dos conglomerados (o elementos) fundidos en la etapa previa y el conglomerado (o elemento) que se intentará unir en esta fase. La distancia entre el grupo AB y C se deduce de la siguiente manera:

$$d_{(C;AB)} = \frac{n_A}{n_A + n_B} d_{CA} + \frac{n_B}{n_A + n_B} d_{CB} - \frac{n_A n_B}{(n_A + n_B)^2} d_{AB}.$$

3.2.1. Implementación de métodos jerárquicos

Para ejemplificar el uso de algunos métodos jerárquicos se trabajará con datos de la inflación de países de la Unión Europea de 2003 (Insituto Nacional de Estadística, España [5]) con el programa libre para cálculos estadísticos `R`.

Primeramente se deberá cargar la librería `cluster`.

```
> library(cluster)
```

Asignaremos a una matriz esta información y trabajeremos con la variables *Índice* y *Tasa* (incremento). El objetivo será separar los países inflacionistas de los no inflacionistas, por ello de antemano podremos intuir un número determinado de grupos en este caso, 2.

```
> matrizdatos <- matrix(c(109, 1, 108.9, 1.6, 111.3, 1.2, 117.3,
+3.5, 112.4, 1.4, 108.7, 1.8, 125.8, 3.8, 119.5, 3.7, 122.4,
+4.5, 114.3, 1.8, 113, 2.2, 119.1, 3.8, 114.1, 2.5, 108.7,
+1, 110.2, 1.2), nrow = 15, ncol = 2, byrow = TRUE,
+dimnames = list(c("Alemania", "Austria", "Belgica",
+"España", "Finlandia", "Francia", "Grecia", "Holanda",
+"Irlanda", "Italia", "Luxemburgo", "Portugal",
+"Dinamarca", "Reino Unido", "Suecia"), c("Indice", "Tasa")))
```

La **Figura 3.1** muestra la gráfica de dispersión de los datos para identificar visiblemente algún grupo.

Se puede decir que hay dos grupos, arriba a la derecha están los países con mayor tasa y abajo a la izquierda los de menor tasa de incremento, se plasmará esta idea en `R` y ver la forma de agrupamiento de los países.

Con la instrucción `dist` se calcula distintas distancias entre los puntos (las filas) de una base de datos. Esta instrucción también admite el argumento `method`, que permite especificar que tipo de distancia entre los puntos queremos calcular, siendo la distancia euclídea la opción por defecto (más detalles con `help(dist)`).

```
> dist.matrizdatos <- dist(matrizdatos)
```

```
> plot(matrizdatos)
```

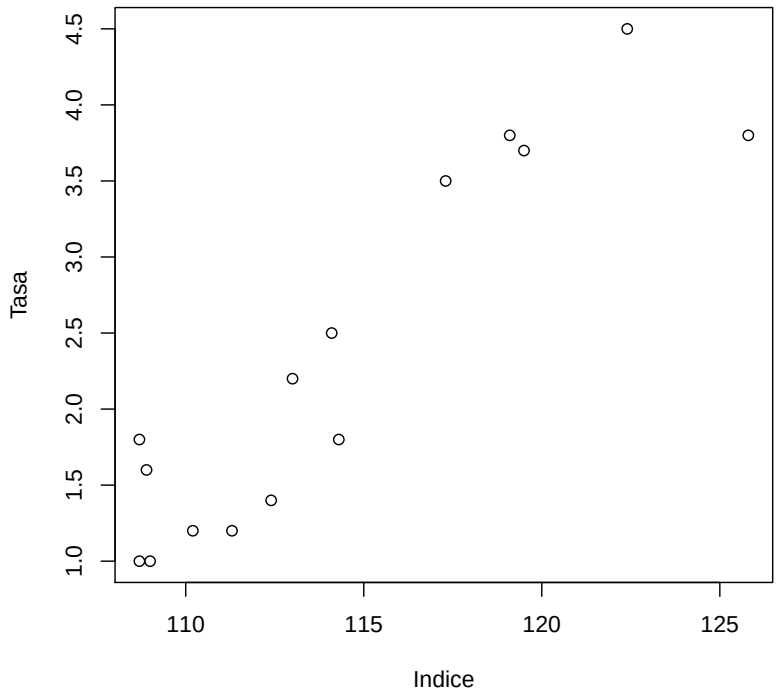

Figura 3.1: Gráfica de dispersión

Para realizar el procedimiento aglomerativo, se usa la instrucción `hclust` que parte de una matriz de distancias para llevar a cabo la jerarquía de clasificación de los datos; `hclust` admite como argumento `method` que puede tomar alguno de los siguientes valores: *single*, *complete*(éste es el método por defecto), *average*, *centroid*, *ward*, etc.

```
> mas.cercano = hclust(dist.matrizdatos, method = "single")
```

A partir de la **Figura 3.2** se observa que la información pudiera estar

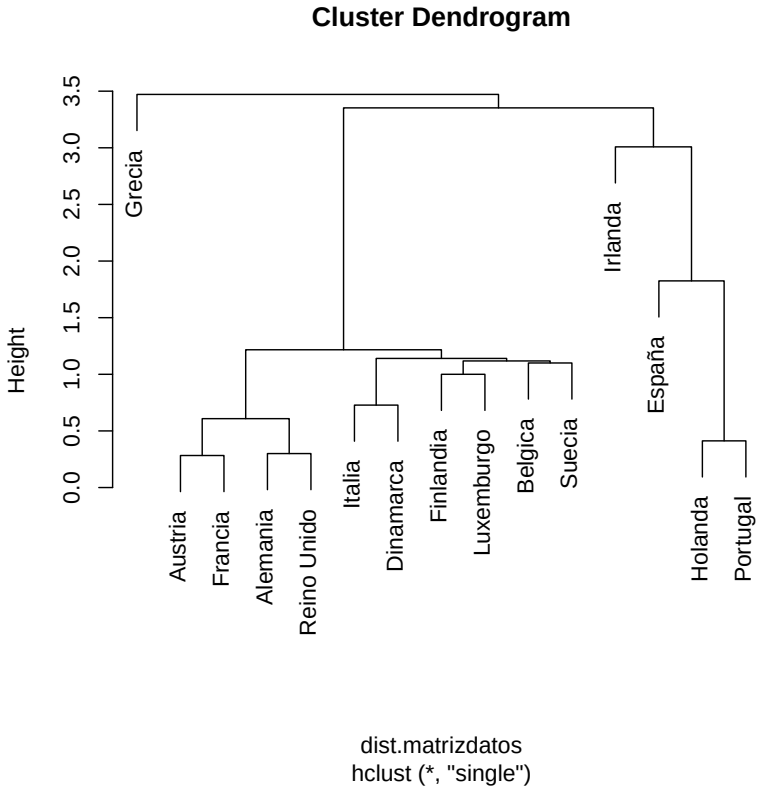

Figura 3.2: Vecino más cercano

distribuida en 3 grupos y no en 2 como se hubiése esperado. Se corta al jerárquico *mas.cercano* en 3 grupos (**Figura 3.3**).

```
> corte.mas.cercano = cutree(mas.cercano, k = 3)
> plot(matrizdatos, type = "n")
> points(matrizdatos[corte.mas.cercano == 1, 1],
+matrizdatos[corte.mas.cercano == 1, 2], col = "purple")
> points(matrizdatos[corte.mas.cercano == 2, 1],
+matrizdatos[corte.mas.cercano == 2, 2], col = "green")
> points(matrizdatos[corte.mas.cercano == 3, 1],
+matrizdatos[corte.mas.cercano == 3, 2], col = "darkred")
```

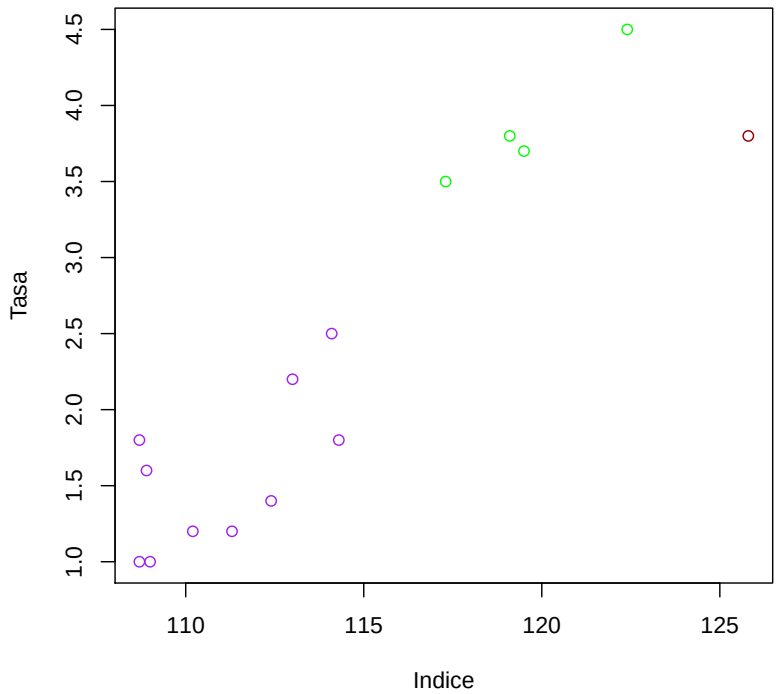

Figura 3.3: Más cercano

La información será analizada con el *método completo*.

```
> mas.lejano = hclust(dist.matrizdatos)
```

En la **Figura 3.4** se observa que el comportamiento de agrupación se asocia un poco más a lo que intuitivamente esperabamos. Se corta al jerárquico *mas.lejano* en dos grupos (**Figura 3.5**).

```
> corte.mas.lejano = cutree(mas.lejano, k = 2)
> plot(matrizdatos, type = "n")
> points(matrizdatos[corte.mas.lejano == 1, 1],
+matrizdatos[corte.mas.lejano == 1, 2], col = "purple")
> points(matrizdatos[corte.mas.lejano == 2, 1],
+matrizdatos[corte.mas.lejano == 2, 2], col = "green")
```

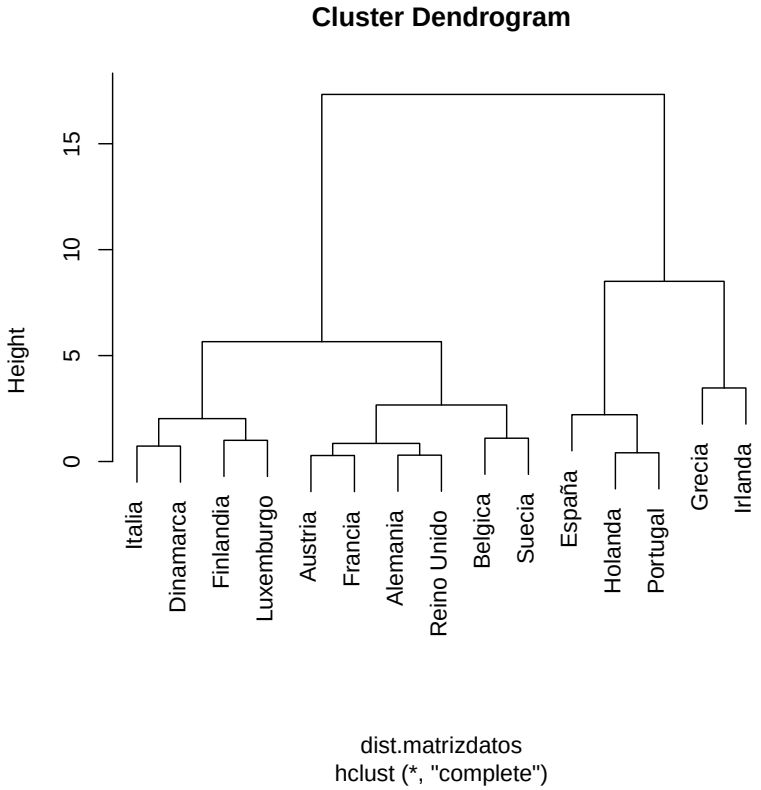

Figura 3.4: Vecino más lejano

Se realiza nuevamente un análisis jerárquico con el *método del promedio*, tomando de la misma manera los dos grupos.

```
> promedio = hclust(dist.matrizdatos, method = "average")
```

```
> corte.promedio = cutree(promedio, k = 2)
> plot(matrizdatos, type = "n")
> points(matrizdatos[corte.promedio == 1, 1],
+matrizdatos[corte.promedio == 1, 2], col = "purple")
> points(matrizdatos[corte.promedio == 2, 1],
```

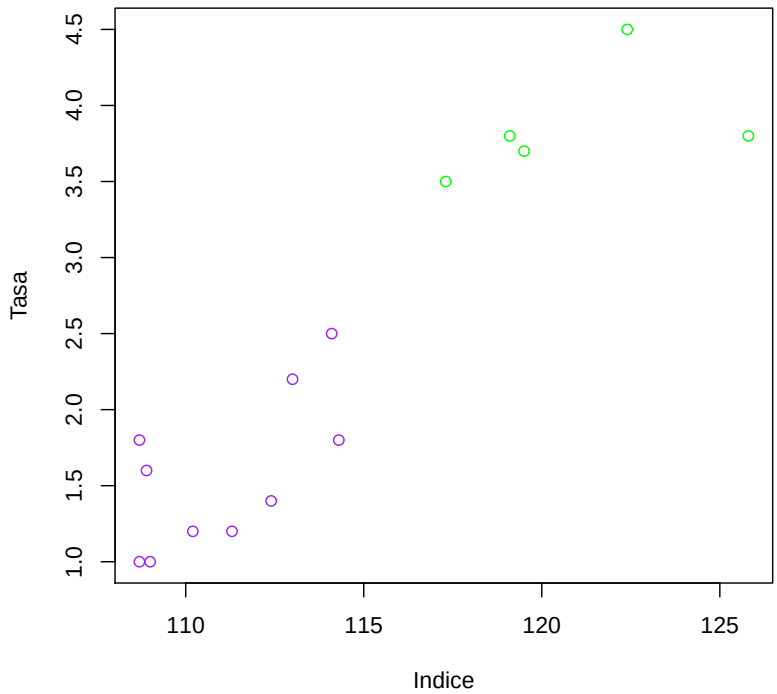

Figura 3.5: Más lejano

Al comparar las **Figuras 3.3**, **3.5**, y **3.6** se observan desacuerdos respecto al número de grupos a elegir. Aún antes del análisis habíamos insistido, intuitivamente, en que la información se clasificaría en dos grupos ya que el objetivo del ejemplo era separar los países inflacionistas de los que no, pero al aplicar el *método simple* crea gran confusión respecto a esto. Debemos tener muy presente que la jerarquía es basada en distancias y estas siempre fueron euclidianas y no se ha trabajado con métodos divisivos.

También se realizará el análisis con el algoritmo *k-medias*. Los centros iniciales fueron tomados aleatoriamente por el programa.

```
> kmedias <- kmeans(matrizdatos, 2)
```

Para ver a que grupo pertenece cada país unimos la *matrizdatos* con la variable $*cluster* que viene dentro de la instrucción `kmeans`:

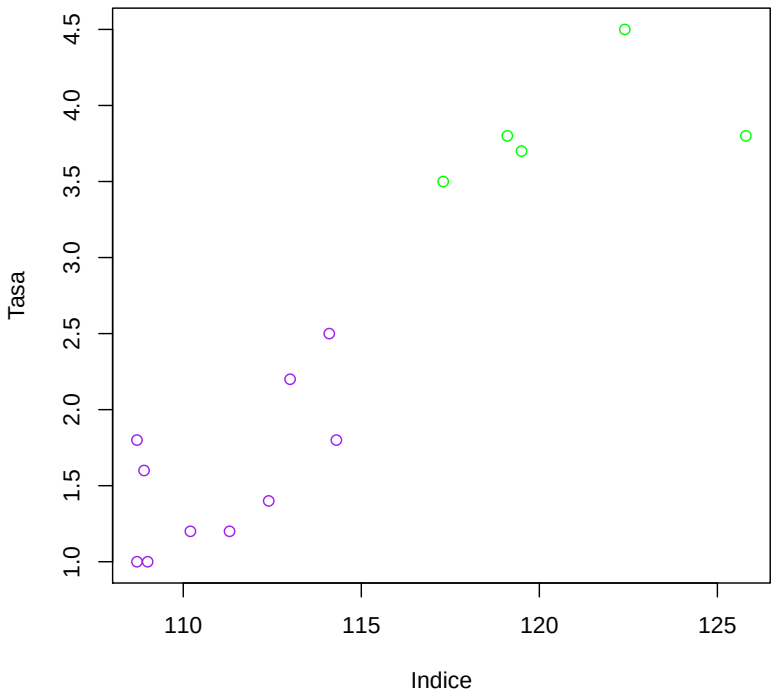

Figura 3.6: Promedio

```
> cbind(kmedias$cluster, matrizdatos)
```

		Indice	Tasa			Indice	Tasa
Alemania	2	109.0	1.0	Austria	2	108.9	1.6
Belgica	2	111.3	1.2	España	1	117.3	3.5
Finlandia	2	112.4	1.4	Francia	2	108.7	1.8
Grecia	1	125.8	3.8	Holanda	1	119.5	3.7
Irlanda	1	122.4	4.5	Italia	2	114.3	1.8
Luxemburgo	2	113.0	2.2	Portugal	1	119.1	3.8
Dinamarca	2	114.1	2.5	Reino Unido	2	108.7	1.0
Suecia	2	110.2	1.2				

Los menos inflacionistas los ha clasificado en el grupo 2 y los más inflacionistas (España, Grecia, Holanda, Irlanda y Portugal) en el grupo 1.

Veámoslo gráficamente.

```
> plot(matrizdatos, col = kmedias$cluster, pch = 15)
```

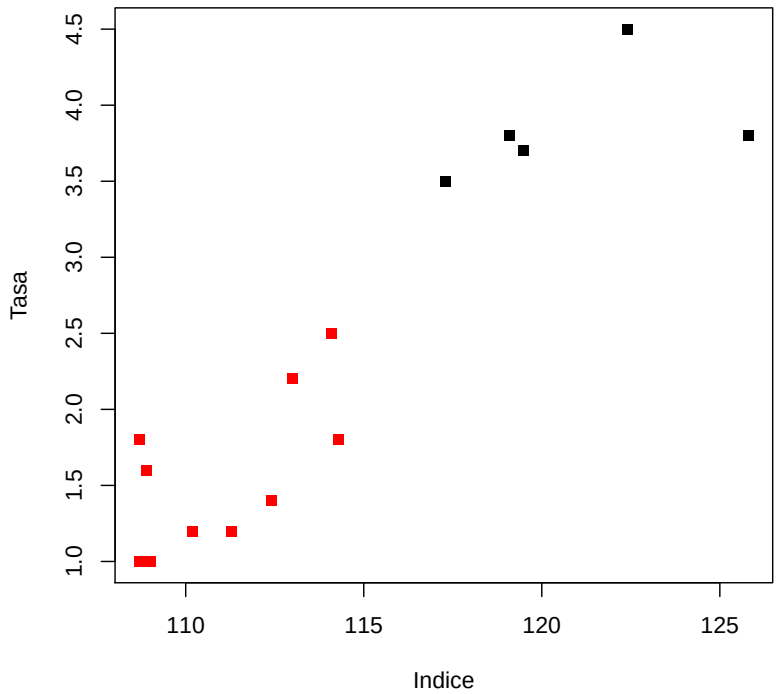

Figura 3.7: Agrupamiento

En la **Figura 3.7** se observa que el algortimo *k-medias* produce el mismo resultado que los métodos jerárquicos excepto por el *simple*, por esto se pudiera tener cierta confianza en que la información se clasifica en dos grupos. Sin embargo, aunque la información presentaba esta separación desde el principio, el resultado que arrojó *método simple* causó cierta incertidumbre. Por esta razón no es posible tomar un método como el correcto para conglomerar la información, al menos en un sentido estadístico, ya que al final los criterios serían sumamente relativos y quedaría en consideración del analista la clasificación final.

Capítulo 4

Métodos basados en modelos de mezclas

En el capítulo anterior se dió una introducción al análisis de conglomerados desde un punto de vista clásico, los problemas que estudia y los métodos más utilizados. Pero, ni los métodos jerárquicos ni los de partición se dirigen directamente al problema de determinar el número de grupos. Dos enfoques diferentes serán descritos en este capítulo.

Modelos de probabilidad han sido propuestos por algún tiempo como la base del análisis de conglomerados. En este tratamiento, la información es representada por un modelo de mezclas en el cual cada componente corresponde a un grupo diferente. Recientemente, métodos de este tipo han mostrado promesas en diferentes aplicaciones prácticas como en medicina, campos minados, detección de fallas sísmicas, etc.

Modelos de mezclas para conglomerados

Los modelos de mezclas finitas también son ampliamente usados para modelar distribuciones de una amplia variedad de fenómenos aleatorios y conjuntos de datos conglomerados.

Con la propuesta para conglomerados, se asume que la información es generada por una mezcla de distribuciones de probabilidad en el cual cada componente representa un grupo diferente. Especificando una forma paramétrica $f_j(x_i|\theta_j)$ para cada componente de la mezcla, podemos ajustar este modelo de mezclas paramétrico

$$f(x_i, \Psi) = \sum_{j=1}^{k} \pi_j f_j(x_i|\theta_j), \tag{4.1}$$

donde $\Psi = (\theta_1, ..., \theta_k; \pi_1, ..., \pi_k)$ y θ_j es el parámetro desconocido y correspondiente al j-ésimo componente de la mezcla, k el número de componentes en la muestra (corresponden a los k grupos) y las π_j son los pesos de la mezcla (cantidades no negativas) donde $\sum_{j=1}^{k} \pi_j = 1$.

Dadas las observaciones $\mathtt{X} = (x_1, x_2, \ldots, x_n)$, la propuesta de la verosimilitud de la mezcla es:

$$L_M(\theta_1, ..., \theta_k; \pi_1, ..., \pi_k|\mathtt{X}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j f_j(x_i|\theta_j). \tag{4.2}$$

Debido a que es más fácil maximizar $logL_M$ que L_M se trabajará con el logaritmo de la función de verosimilitud conjunta (4.2),

$$l(\theta_1, ..., \theta_k; \pi_1, ..., \pi_k|\mathtt{X}) = \sum_{i=1}^{n} \sum_{j=1}^{k} log(\pi_j f_j(x_i|\theta_j)). \tag{4.3}$$

Primeramente porque pasamos de un producto de densidades a la suma de sus logaritmos y la expresión resultante suele ser más simple que la verosimilitud y en segundo, al tomar logaritmos las constantes multiplicativas de la función de densidad, que son irrelevantes para el máximo, se hacen aditivas y desaparecen al derivar con lo que la derivada de $logL_M$ tiene la misma expresión y no depende de constantes arbitrarias. Además el logaritmo es una función monótona, por lo que ambas funciones tienen el mismo máximo.

Para el presente trabajo, se tiene especial interés en las mezclas de normales multivariadas, donde θ_j consiste en un vector de medias μ_j y una matriz de covarianzas Σ_j con $j = 1, ..., k$, por lo que la función de densidad para la i-ésima observación, x_i $(i = 1, ..., n)$, tiene la forma

$$f_j(x_i; \mu_j, \Sigma_j) = \frac{\exp[-\frac{1}{2}(x_i - \mu_j)' \Sigma_j^{-1}(x_i - \mu_j)]}{|\Sigma_j|^{1/2}(2\pi)^{p/2}}. \tag{4.4}$$

La ecuación (4.4) determina grupos elipsoidales, centrados en el vector de medias μ_j y Σ_j define sus otras características geométricas.

Jeffrey D. Banfield y Adrian E. Raftery[6] desarrollaron un esquema basado en modelos para grupos parametrizando la matriz de covarianzas en términos de su descomposición espectral como

$$\Sigma_j = \lambda_j D_j A_j D_j^t, \tag{4.5}$$

donde $\lambda_j = |\Sigma_j|^{1/p}$, D_j es la matriz ortogonal de eigenvectores de Σ_j y A_j es una matriz diagonal, tal que $|A_j| = 1$, con los eigenvalores normalizados de Σ_j en la diagonal en orden decreciente, siendo el escalar λ_j el mayor eigenvalor de la matriz.

λ_j especifica el tamaño o volumen del j-ésimo grupo, la orientación está determinada por D_j, mientras que A_j determina la forma.

Las características de las distribuciones (orientación, volumen y forma) son usualmente estimadas de los datos y se permiten variar entre los grupos u obligadas a ser las mismas para todos los grupos.

Este enfoque abarca varias propuestas basadas en mezclas de Normales multivariadas:

$\Sigma_j = \lambda I$, determina grupos esféricos con volúmenes iguales.

$\Sigma_j = \lambda D A D_j^t$, los grupos tienen la misma forma, volumen y orientación.

$\Sigma_j = \lambda_j D_j A_j D_j^t$ es la condición más general ya que todas las matrices de covarianzas son distintas, no se establece ninguna restricción sobre ellas.

$\Sigma_j = \lambda D_j A D_j$, modelo en el cual sólo la orientación de los grupos puede diferir.

4.1. Enfoque de estimación basado en maximización

Para calcular los estimadores de los parámetros contenidos en el vector $\Psi = (\pi_1, ...\pi_k, \mu_1, ..., \mu_k, \Sigma_1, .., \Sigma_k)$, la alternativa más usada es aplicar el algoritmo EM.

4.1.1. Algoritmo EM

El algoritmo EM fue introducido por Arthur Dempster, Nan Laird y Donald Rubin en 1977 y es un método general para encontrar el estimador de máxima verosimilitud de los parámetros de una distribución de probabilidad en presencia de información incompleta.

Supondremos un conjunto de datos $\mathtt{Y} = (\mathtt{X}, \mathtt{Z})$ donde $\mathtt{X}$ son los datos observados (también es llamado el conjunto de datos incompletos), $\mathtt{Z}$ los datos no observados (ocultos) y $\mathtt{Y}$ es la llamada información completa. Este algoritmo es utilizado en el marco donde una observación x_i de la muestra aleatoria observada $x_1, \ldots, x_n$ se piensa que es tomada de alguno de los componentes, mezclas o modas y el vector indicador $z_i = (z_{i1}, z_{i2}, ...z_{ik})$ denota su componente de origen de donde es tomado y no medido, donde

$$z_{ij} = \begin{cases} 1 & \text{si } x_i \text{ pertenece al grupo } j \\ 0 & \text{en otro caso} \end{cases}$$

Por ejemplo x_i vendrá de la población 1 si $z_{i1} = 1$ y $z_{i2} = ... = z_{ik} = 0$.
La función de densidad de x_i condicionada a z_i puede escribirse como

$$f(x_i|z_i) = \prod_{j=1}^{k} f_j(x_i)^{z_{ij}}.$$

Del mismo modo, la función de probabilidad de la variable z_i es

$$p(z_i) = \prod_{j=1}^{k} \pi_j^{z_{ij}}.$$

Por otro lado, la función de densidad conjunta de x_i y z_i es

$$f(x_i, z_i) = \prod_{j=1}^{k} (\pi_j f_j(x_i))^{z_{ij}}.$$

El logaritmo de la función de verosimilitud completa es

$$l(\Psi|\mathtt{X}, \mathtt{Z}) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij}[log\pi_j f_j(x_i|\theta_j)], \tag{4.6}$$

donde $\Psi = (\theta_1, ..., \theta_k; \pi_1, ..., \pi_k)$.

La fórmula anterior es equivalente a

$$l(\Psi|\mathtt{X}, \mathtt{Z}) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} log\pi_j + \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} log f_j(x_i|\theta_j).$$

Por otro lado, la cantidad

$$\hat{z}_{ij} = E[z_{ij}|\mathtt{X}, \hat{\Psi}_1, \hat{\Psi}_2, ..., \hat{\Psi}_k] = p(z_{ij} = 1|\mathtt{X}, \hat{\Psi}_1, \hat{\Psi}_2, ..., \hat{\Psi}_k)$$

$$= p(z_{ij} = 1|x_i, \hat{\Psi}_1, \hat{\Psi}_2, ..., \hat{\Psi}_k) = z^*_{ij},$$

para el modelo (4.6) es la esperanza condicional de z_{ij} dada la observación x_i y los valores de los parámetros. El valor z^*_{ij} de $\hat{z}_{ij}$ es la probabilidad condicional de que la observación i pertenezca al grupo j.

El algoritmo EM itera entre el paso E (Cálculo de la Esperanza) en el cual los valores de $\hat{z}_{ij}$ son calculados de la información contenida en el parámetro actual y el paso M (Maximización) en el cual cada z_{ij} es reemplazado por su esperanza condicional actual $\hat{z}_{ij}$ en la logverosimilitud de la información completa (4.6).

4.1.2. Metodología del algoritmo EM

El algoritmo EM comienza con valores iniciales para el vector de parámetros, $\Psi^{(0)} = (\pi^{(0)}_1, ...\pi^{(0)}_k; \mu^{(0)}_1, ..., \mu^{(0)}_k; \Sigma^{(0)}_1, .., \Sigma^{(0)}_k)$ (iteración $m = 0$).

Calcularemos la esperanza condicional de z_{ij} con estos valores iniciales y la denotaremos $z^{*(0)}_{ij}$

$$\hat{z}^{(0)}_{ij} = E[z_{ij}|\mathtt{X}, \hat{\Psi}^{(0)}_1, \hat{\Psi}^{(0)}_2, ..., \hat{\Psi}^{(0)}_k] = z^{*(0)}_{ij},$$

donde $z^{*(0)}_{ij}$ es la probabilidad de que la observación i pertenezca al grupo j cuando ya se ha observado x_i.

Como la verosimilitud es lineal en z_{ij}, reemplazamos z_{ij} en (4.6) por sus esperanzas $z^{*(0)}_{ij}$,

$$l(\Psi|\mathtt{X},\mathtt{Z}) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}^{*(0)}[log\pi_j f_j(x_i|\theta_j)].$$

Paso M. Maximización

En este paso se maximiza la función anterior respecto a los parámetros de Ψ. Se calcula el estimador de máxima verosimilitud con los valores z_{ij} ahora fijos a $z_{ij}^{*(0)}$.

$$\hat{\pi}_j \leftarrow \frac{\sum_{i=1}^{n} z_{ij}^{*(0)}}{n} \; ; \qquad \hat{\mu}_j \leftarrow \frac{\sum_{i=1}^{n} z_{ij}^{*(0)} x_i}{\sum_{i=1}^{n} z_{ij}^{*(0)}}. \tag{4.7}$$

El cálculo $\hat{\Sigma}_j$ depende de su parametrización. En el artículo de Celeux y Govaert [8] se detalla el Paso M para $\hat{\Sigma}_j$ por la descomposición (4.5).

Al calcular cada parámetro de Ψ con los valores fijos $z_{ij}^{*(0)}$, generamos un nuevo vector $\Psi^{(1)} = (\pi_1^{(1)}, ...\pi_k^{(1)}; \mu_1^{(1)}, ..., \mu_k^{(1)}; \Sigma_1^{(1)}, .., \Sigma_k^{(1)})$.

Paso E. Cálculo de la Esperanza

Ahora calculemos $\hat{z}_{ij}^{(1)}$ con el parámetro estimado del Paso M, es decir, la iteración $m = 1$ con $\Psi^{(1)} = (\pi_1^{(1)}, ...\pi_k^{(1)}; \mu_1^{(1)}, ..., \mu_k^{(1)}; \Sigma_1^{(1)}, .., \Sigma_k^{(1)})$.

$$\hat{z}_{ij} \leftarrow \frac{\hat{\pi}_j f_j(x_i|\hat{\mu}_j, \hat{\Sigma}_j)}{\sum_{l=1}^{k} \hat{\pi}_l f_l(x_i|\hat{\mu}_l, \hat{\Sigma}_l)}. \tag{4.8}$$

donde f_j tiene la forma (4.4).

El proceso iterativo termina hasta obtener convergencia.

Aunque este método es la opción más utilizada y generalmente en modelos de mezclas da buenos resultados; además para cada i, da una medida de incertidumbre en la clasificación asociada $(1 - max_j\ z_{ij}^*)$, y en general su funcionamiento es satisfactorio, presenta algunas limitantes. Primero, cuando se ajustan los parámetros de un modelo de mezcla de normales sin ningún tipo

de restricción sobre las matrices de covarianza de los componentes, alguno de los π_j podría aproximarse a cero y por consiguiente, la matriz de covarianzas correspondiente a uno o más componentes se vuelve singular o casi singular. Segundo, la convergencia puede ser muy lenta. Sin embargo, esto no parece ser un problema para mezclas bien separadas, cuando la iteración comienza con valores razonables. Además, el algoritmo por si mismo, no permite la estimación del número de componentes del modelo.

Aparte del problema principal, determinar el número correcto de componentes, dentro del análisis de conglomerados se plantea otra cuestión básica: **seleccionar el mejor modelo**. La propuesta al problema de selección de modelo en conglomerados está basada en modelos de selección bayesiana, a través de los factores de Bayes y modelo de probabilidades posteriores.

Preliminares de la inferencia bayesiana

La inferencia bayesiana, en contraste a la clásica, le permite al investigador incorporar, para cualquier modelo probabilístico, la información previa que pueda poseer el fenómeno. Además, proporciona una manera satisfactoria de introducir explícitamente y da seguimiento a los supuestos sobre el conocimiento previo.

En el enfoque bayesiano de, la incertidumbre presente en un modelo dado, $p(\mathtt{X}|\theta)$, es representada a través de una distribución de probabilidad $p(\theta)$ sobre los posibles valores del parámetro desconocido θ (típicamente multidimensional) que define al modelo. El teorema de Bayes, permite entonces incorporar la información contenida en un conjunto de datos $\mathtt{X} = (x_1, x_2, \ldots, x_n)$, produciendo una descripción conjunta de la incertidumbre sobre los valores de los parámetros del modelo a través de la distribución final $(p|\mathtt{X})$ (Gutiérrez [9]).

En la práctica es común que la dimensión de θ sea muy grande. Por otro lado, excepto en aplicaciones muy sencillas tanto $p(\mathtt{X}|\theta)$ como $p(\theta)$ pueden llegar a tener formas muy complicadas. En la gran mayoría de los problemas las integrales requeridas no pueden resolverse analíticamente, por lo que es necesario contar con métodos numéricos eficientes que permitan calcular o aproximar integrales en varias dimensiones.

La aproximación de Laplace, integración numérica por cuadratura, el método de Monte Carlo, así como las técnicas de integración desarrolladas durante los últimos años conocidas con el nombre genérico de técnicas de

Monte Carlo vía cadenas de Markov, son algunos de los métodos clásicos para calcular integrales.

En términos generales, los métodos antes mencionados serán más eficientes y darán resultados más precisos en la medida en que la distribución final sea más parecida a la distribución normal. Es por esta razón que en la mayoría de los casos resulta conveniente trabajar en términos de reparametrización del modelo, de manera que cada uno de los nuevos parámetros tome valores en todo $\mathbb{R}$ y su distribución final sea aproximadamente normal.

Factores de Bayes

El uso de los factores de Bayes es una alternativa bayesiana a pruebas de hipótesis clásicas. Estos son el análogo bayesiano de las pruebas de razón de verosimilitud.

La comparación de modelos bayesianos es un método de selección de modelo basado en los factores de Bayes.

La idea básica es que si los datos $\mathtt{X}$ han surgido de alguna de las 2 hipótesis H_0 y H_1 de acuerdo con las densidades de probabilidad $P[\mathtt{X}|H_0]$ y $P[\mathtt{X}|H_1]$ y que existen las probabilidades *a priori* para cada una de las dos hipótesis. Estas probabilidades quedan automáticamente determinadas si establecemos una distribución *a priori*, $p(\theta)$, sobre θ, ya que entonces:

$$P[H_0] = P[\theta \in \Omega_0] = \int_{\Omega_0} p(\theta)\mathrm{d}_\theta,$$

$$P[H_1] = P[\theta \in \Omega - \Omega_0] = \int_{\Omega-\Omega_0} p(\theta)\mathrm{d}_\theta.$$

Las probabilidades *a posteriori* de las hipótesis las calcularemos mediante el teorema de Bayes

$$P[H_j|\mathtt{X}] = \frac{P[\mathtt{X}|H_j]P[H_j]}{P[\mathtt{X}|H_0]P[H_0] + P[\mathtt{X}|H_1]P[H_1]} \quad (j = 0, 1),$$

de modo que

$$\frac{P[H_0|\mathtt{X}]}{P[H_1|\mathtt{X}]} = \frac{P[\mathtt{X}|H_0]P[H_0]}{P[\mathtt{X}|H_1]P[H_1]}, \tag{4.9}$$

que puede expresarse como

$$momio\ a\ posteriori = Razón\ de\ Verosimilitudes \times momio\ a\ priori.$$

Esta expresión indica que la evidencia respecto a la hipótesis nula se obtiene multiplicando la evidencia proporcionada por los datos, con la evidencia *a priori*. A la razón entre las verosimilitudes se denomina **Factor de Bayes**, $\mathtt{B} = \frac{P[\mathtt{x}|H_0]}{P[\mathtt{x}|H_1]}$.

Ahora bien si varios modelos $M_1, M_2, \cdots, M_j$ son considerados con probabilidades *a priori* $p(M_j)$, $j = 1, 2, ..., k$, (a menudo tomadas iguales) entonces por el teorema de Bayes la probabilidad posterior del modelo M_j dada la información $\mathtt{X}$ es proporcional a la probabilidad de la información dado el modelo M_j, por la probabilidad *a priori* del modelo, a saber

$$p(M_j|\mathtt{X}) \propto p(\mathtt{X}|M_j)p(M_j),$$

donde hay parámetros desconocidos, entonces, por el teorema de la probabilidad total, $p(X|M_j)$ es obtenida integrando sobre los parámetros, es decir,

$$p(\mathtt{X}|M_j) = \int p(\mathtt{X}|\theta_j, M_j)p(\theta_j|M_j)\,\mathrm{d}\theta_j,$$

donde $p(\theta_j|M_j)$ es la distribución *a priori* de θ_j, el vector de parámetros para el modelo M_j. La cantidad $p(\mathtt{X}|M_j)$ es conocida como la *verosimilitud integrada* del modelo M_j.

Una aproximación natural bayesiana a la selección del modelo es elegir el modelo que es más probable *a posteriori*, y si el modelo de probabilidades *a priori*, $p(M_j)$, es el mismo, esto equivale a elegir al modelo con la más alta *verosimilitud integrada*. El factor de Bayes es utilizado para comparar modelos M_1 y M_2, es decir, $B_{12} = \frac{p(\mathtt{X}|M_1)}{p(\mathtt{X}|M_2)}$. En otras palabras, el factor de Bayes B_{12} representa la probabilidad posterior que la información fuera distribuida de acuerdo al modelo M_1 contra el modelo M_2 suponiendo que ningún modelo está favorecido *a priori*.

Interpretación

El factor de Bayes es un resumen de la evidencia provista por los datos a favor de una teoría científica, representada por un modelo estadístico, en oposición a otros. Jeffreys (1961) sugirió interpretar B o (B^{-1}) en unidades de la escala log_{10}. Fusionando dos de sus categorías ($1 < log_{10}(B_{12}^{-1}) < 1.5$ y $1.5 < log_{10}(B_{12}^{-1}) < 2$) para la simplificación e interpretando la evidencia contra modelo M_1, tenemos:

$0 < log_{10}(B_{12}^{-1}) < 0.5$	$No\ vale\ la\ pena\ mencionarlo$
$0.5 < log_{10}(B_{12}^{-1}) < 1$	$Hay\ evidencia\ Sustancial\ contra\ M_1$
$1 < log_{10}(B_{12}^{-1}) < 2$	$Hay\ evidencia\ Fuerte\ contra\ M_1$
$log_{10}(B_{12}^{-1}) > 2$	$Hay\ evidencia\ Decisiva\ contra\ M_1$

Cuadro 4.1: Interpretación del factor de Bayes según Jeffreys

4.1.3. Criterio de información bayesiana BIC

La dificultad principal en el uso de factores de Bayes es la evaluación de la integral que define la *verosimilitud integrada*. Para modelos no irregulares, y cuando el EM es usado para encontrar la mezcla de máxima verosimilitud, la *verosimilitud integrada* puede ser aproximada simplemente con el doble del logaritmo del factor de Bayes llamado **BIC**:

$$BIC_j = 2\log p(\mathtt{X}|M_j) \approx 2\log p(\mathtt{X}|\hat{\theta}_j, M_j) - v_j \log(n) \quad para\ j = 1, \ldots, k, \tag{4.10}$$

donde v_j es el número de parámetros independientes a ser estimados en el modelo M_j, n es el número de observaciones y $\hat{\theta}_j$ es el estimador de máxima verosimilitud para el vector de parámetros del modelo M_j, θ_j. El número de componentes no es considerado un parámetro independiente para el propósito de calcular el BIC.

El BIC puede ser utilizado para comparar modelos con diferentes parametrizaciones, diferente número de componentes o ambas; está cercanamente

relacionado con otros criterios de selección de modelos como el criterio de información Akaike [1], aunque el BIC tiende a seleccionar modelos más simples que los que seleccionaría AIC. El valor más grande del BIC es la evidencia más fuerte del modelo.

Es una ventaja el uso de factores de Bayes aproximados para comparar modelos de mezclas ya que da una manera sistemática de selección, no sólo para elegir la parametrización del modelo (y por lo tanto el método de conglomerado), sino también el número de grupos.

4.1.4. Estrategia para conglomerados vía BIC

En la práctica, los métodos jerárquicos aglomerativos basados en la clasificación de la verosimilitud con términos Gaussianos, con frecuencia, da buenas pero subóptimas particiones. El algoritmo EM puede refinar particiones si comenzaron suficientemente cercanas al valor óptimo. Una partición dada puede ser transformada en variables indicadoras, las cuales pueden ser utilizadas como probabilidades ó esperanzas condicionales en el paso M del algoritmo EM para la estimación del parámetro, inicializando una iteración de EM. Esto combinado con factores de Bayes como una aproximación por el BIC para selección de modelo, produce una estrategia comprensiva para conglomerar desarrollada por Chris Fraley y Adrian E. Raftery [7]:

1) Se determina un conjunto máximo de grupos (**M**) y un conjunto de parametrizaciones candidatas a considerar para el modelo Gaussiano.

2) Se realiza análisis jerárquico (aglomeración) para obtener clasificaciones correspondientes para máximo **M** grupos.

3) Se aplica el algoritmo EM a cada modelo y número de grupos, desde 2 hasta **M**, comenzando la clasificación desde el conglomerado jerárquico.

4) Se calcula el BIC para la mezcla de verosimilitud con parámetros óptimos desde el EM de 2 a **M** grupos y cada parametrización. Esto da una matriz de valores del BIC correspondientes a cada posible combinación de parametrización y número de grupos.

[1] AIC = -2($\log L(\hat{\theta})$ - número de parámetros)

5) Se grafican los valores del BIC para cada modelo. El primer máximo local decisivo indica evidencia fuerte para un modelo (parametrización + número de grupos).

Este criterio es considerado de suma importancia ya que trata de seleccionar el modelo más adecuado, con máxima probabilidad *a posteriori*, pudiendo demostrarse que es un criterio consistente, de manera que la probabilidad de seleccionar el modelo correcto tiende a 1 cuando el tamaño de la muestra tiende a infinito.

Existe relación entre el número de grupos a considerar y la complejidad de las matrices de covarianzas requeridas. Si se aceptan muchos grupos, se podrían obtener buenas soluciones imponiendo que las matrices de covarianzas sean iguales. De otro modo, con pocos grupos, se deberá dejar una amplia libertad a las matrices de covarianzas para obtener un buen ajuste del modelo de los datos. Por ello, las condiciones sobre las matrices de covarianzas se deciden simúltaneamente con el número de grupos, empleando el criterio BIC.

El marco probabilístico de conglomerados basado en modelos permite que los problemas, determinar el modelo más adecuado de conglomerados y el número de grupos, se reduzcan simultáneamente al problema de selección de modelo. Esto es importante porque hay una desventaja entre la elección del número de grupos y modelo de conglomerado. Por ejemplo, si un modelo simple es usado, más grupos pueden ser necesarios para proveer una buena representación de la información. Si un modelo más complejo es usado, menos grupos son suficientes para ajustar la información adecuadamente. Si no son Gaussianos los grupos, ¿qué pasa?, esto da paso a explorar el otro método.

4.2. Enfoque basado en distancia (KL)

En esta sección se trabajará con un enfoque basado en distancia para determinar el número desconocido de componentes de la mezcla, como se ha venido diciendo este es el problema principal que aborda la presente tesis. Se generalizará un método de prueba bayesiano basado en la distancia Kullback-Leibler (KL) propuesto por Mengersen y Robert (1996). Una alternativa, la distancia Kullback-Leibler ponderada, es planteada como criterio de prueba. Se enuncia un procedimiento paso a paso para seleccionar el número adecuado de componentes que representen a los datos.

Cabe mencionar que la distancia KL no es propiamente una métrica en el estricto sentido de la métrica de distancia. De cualquier forma, es generalmente reconocida por ser una cantidad conveniente para medir la "cercanía" entre las distribuciones, así que en este trabajo se usará el término "distancia" en un sentido informal. A pesar de que, en la literatura, existen muchas medidas de distancia posibles, la distancia KL es atractiva por su simplicidad y su tratamiento analítico para modelos de mezclas.

4.2.1. Introducción

Nuevamente consideremos una familia de funciones de probabibilidad de la forma

$$f^{(k)}(x) = \sum_{j=1}^{k} \pi_j f_j(x|\theta_j), \tag{4.11}$$

donde la densidad $f_j(x|\theta_j)$ es dada y parametrizada por θ_j y π_j son los pesos de la mezcla, $\sum_{j=1}^{k} \pi_j = 1$.

Estamos especialmente interesados en identificar la k más pequeña que explique adecuadamente la estructura de los datos.

Considerando un método reductivo paso a paso. Aquí una mezcla inicial de $k = k_0$ (>1) componentes, $f^{(k_0)}$, es ajustada a un grupo de datos donde se asume que el valor de k_0 es más que adecuado. Luego el valor de k es progresivamente reducido hasta que el ajuste ya no sea aceptable. En cualquiera de los pasos, asumiendo que $f^{(k)}$ es aceptable, el procedimiento es:

Paso 1: Ajustar una mezcla de $k-1$ componentes, $f^{(k-1)}$.

Paso 2: Comparar $f^{(k-1)}$ con $f^{(k)}$ usando algún contraste de hipótesis para ver si el ajuste reducido es todavía adecuado.

Estos pasos se repiten hasta que la reducción ya no de un ajuste aceptable.

En el enfoque bayesiano, el factor de Bayes es usado frecuentemente como el criterio de selección para determinar la adecuación del ajuste. Pero como se ha dicho anteriormente, su cálculo es usualmente difícil.

Una alternativa interesante es usar una medida de distancia para comparar dos distribuciones. Consideraremos el método de prueba bayesiano basado en el modelo de la distancia KL (1996).

Sujit K. Sahu y Russell C.H. Cheng [11] generalizaron un método basado en el de Mengersen y Robert, pero de una forma inusual. Esta propuesta contiene una modificación de cada uno de los Pasos 1 y 2 descritos previamente.

Primeramente, un reajuste completo en el paso 1 no es necesario. En lugar de eso consideremos si la *fusión* de dos componentes en $f^{(k)}$ empeora significativamente el ajuste. Esta fusión opera por la selección de dos componentes de $f^{(k)}$ y los reemplaza por un componente. Los parámetros de este componente serán recalculados, pero los otros $(k-2)$ componentes y sus ponderados se conservan fijos. Llamaremos a este ajuste reducido una *versión colapsada* de $f^{(k)}$ y denotada por $f^{*(k-1)}_{(ij)}$, con los subíndices ij indicando uno de los posibles pares de componentes, ${}^kC_2$ (el número de combinaciones de n objetos tomados de 2 en 2), que podrían fusionarse. Entre todas aquellas versiones colapsadas seleccionamos aquella cuya distancia desde $f^{(k)}$ es la más pequeña. A esta la llamamos la *mejor versión colapsada* denotada por $f^{*(k-1)}$. Su distancia desde $f^{(k)}$ está denotada por $d(f^{(k)}, f^{*(k-1)})$.

Una vez que la mejor versión colapsada es elegida de este modo, entonces tenemos que decidir si el ajuste es todavía el adecuado. Una prueba de hipótesis bayesiana, basada en el criterio de distancia, implica el cálculo de la probabilidad posterior que $d(f^{(k)}, f^{*(k-1)})$ sea menor que un valor preespecificado. Si la distancia KL es usada para $d(\cdot, \cdot)$ entonces la probabilidad posterior puede ser difícil de calcular.

La segunda modificación sugerida es usada como criterio de distancia alternativo, aún basado en la distancia KL, pero más fácil de calcular.

En el método global, se inicia con un valor de k razonablemente grande, y se reduce el número de componentes secuencialmente colapsando hasta que no se pueda colapsar más sin que haya una pérdida significante en el ajuste. Un nuevo ajuste del modelo con cada reducción en el valor de k es

así evitado. De cualquier modo, los parámetros estimados bajo el modelo con menor número de componentes en la mezcla son más precisos en general. Por consiguiente, es recomendado un reajuste del modelo de mezclas en el último paso para asegurar que la distribución posterior final haya sido identificada sin ambigüedades y que no sea posible otra reducción sin pérdida de calidad en el ajuste.

4.2.2. Procedimiento de prueba

Este procedimiento consiste en los siguientes cinco pasos:

T1 Seleccionar un k_0 inicial y ajustar el modelo de mezclas (bayesiano) de k_0 componentes , $f^{(k_0)}$, a los datos. Se asume que k_0 es en definitiva lo suficientemente grande de tal manera que $f^{(k_0)}$ sea capaz de captar todas las variaciones presentadas en los datos.

Sea $k = k_0$.

T2 Calcular las distancias $d(f^{(k)}, f^{*(k-1)}_{(ij)})$ entre $f^{(k)}$ y cada una de sus ${}^kC_2$ versiones colapsadas $f^{*(k-1)}_{(ij)}$.

T3 Seleccionar la mejor versión colapsada, $f^{*(k-1)}_{(ij)}$, para la cual $d(f^{(k)}, f^{*(k-1)}_{(ij)})$ es minimizada. Esta versión colapsada se llamará $f^{*(k-1)}$.

T4 Evaluar la probabilidad posterior

$$P_{c_k} = P_r\{d(f^{(k)}, f^{*(k-1)}) \leq c_k | datos\}.$$

T5 Si $P_{c_k} > \alpha$ entonces reemplaza $f^{(k)}$ por $f^{*(k-1)}$ y repite todo desde T2 con k reducido por 1.

Por último, un paso de chequeo T6 puede ser implementado el cual reajusta la mezcla con el valor final de k.

El procedimiento de prueba evita el problema de *no identificabilidad* donde aparentemente diferentes combinaciones de valores de los parámetros en un modelo de k componentes realmente identifica la misma mezcla de $k-1$ componentes.

De cualquier modo varios aspectos del procedimiento de prueba necesitan discusión:

1) La elección de $d(\cdot, \cdot)$, la medida de distancia.

2) El método preciso para construir la distribución colapsada $f^{*(k-1)}$.

3) La elección de los parámetros de criterio: c_k y α.

4.2.3. Distancia Kullback-Leibler

Como fue sugerido en Mengersen y Robert [12], la distancia KL es una opción posible para la distancia d usada en el paso T2 del procedimiento de prueba. La distancia Kullback-Leibler entre dos densidades f y g, es definida como

$$d(f, g) = \int_{S(f)} f(x) log \frac{f(x)}{g(x)} \mathrm{d}x \tag{4.12}$$

donde $S(f)$ es el soporte de la densidad f, es decir, $S(f) = \{x : f(x) > 0\}$. La distancia $d(f, g)$ es siempre no negativa. Si el soporte de f y g son los mismos, simplemente se escribe

$$d(f, g) = \int f(x) log \frac{f(x)}{g(x)} \mathrm{d}x.$$

Para la notación, se usará $d(f^{(k)}, f^{*(k-1)})$ más que $d(f^{*(k-1)}, f^{(k)})$, poniendo el modelo más general como primer argumento. A pesar de que los soportes de $f^{(k)}$ y $f^{*(k-1)}$ son los mismos, $f^{(k)}$ es el modelo más general dado que el espacio de parámetros para $f^{*(k-1)}$ es un subconjunto del espacio de parámetros de $f^{(k)}$.

En general no es posible calcular la distancia KL, $d(f^{(k)}, f^{*(k-1)})$, como se define en (4.12) analíticamente. En Mengersen y Robert [12] utilizan una aproximación de Laplace en algunos casos especiales.

A pesar de que la evaluación de $d(f^{(k)}, f^{*(k-1)})$ por cuadratura es sencilla, de cualquier modo requiere de un esfuerzo adicional. A continuación, se sugiere una alternativa de distancia KL ponderada, $d^*(f, g)$, que puede ser usada en lugar de la distancia KL. En el presente contexto, se recomienda esta alternativa de distancia porque es mucho más fácil de implementar e igualmente satisfactoria (Sahu *et al.* [11]).

4.2.4. Distancia Kullback-Leibler ponderada

Sea π_j un conjunto de pesos fijos. Dos modelos de mezcla de k componentes, con estos mismos pesos, pueden ser comparados utilizando la siguiente medida de distancia. Sea $f^{(k)}$ definida como en (4.11), y $g^{(k)}$ una densidad de mezclas de k componentes con los mismos pesos de mezcla como en $f^{(k)}$

$$g^{(k)}(x) = \sum_{j=1}^{k} \pi_j g_j(x|\theta_j^*), \tag{4.13}$$

donde los componentes de g_j pueden ser diferentes de los f_j y puede depender posiblemente de diferentes parámetros θ_j^*. Entonces,

$$d^*(f^{(k)}, g^{(k)}) = \sum_{j=1}^{k} \pi_j d(f_j, g_j), \tag{4.14}$$

define una medida de distancia entre f_j y g_j. Para pesos fijos, esta medida claramente disfruta de las mismas propiedades que la distancia KL siendo simplemente la suma ponderada de distancias KL, $d(f_i, g_j)$, entre los componentes correspondientes. Nos referiremos a esto como la distancia KL ponderada.

En el problema de determinar k, el comportamiento de $d(f, g)$ y $d^*(f, g)$ es muy similar. La siguiente fórmula muestra como ellas están relacionadas:

$$\begin{aligned}
d(f^{(k)}, g^{(k)}) &= \int f^{(k)}(x) log \frac{f^{(k)}(x)}{g^{(k)}(x)}\, dx \\
&= \int \sum_{j=1}^{k} \pi_j f_j(x) log \left[\frac{f_j(x)}{g_j(x)} \frac{g_j(x)}{f_j(x)} \frac{f^{(k)}(x)}{g^{(k)}(x)} \right] dx \\
&= \int \sum_{j=1}^{k} \pi_j f_j(x) \left\{ log \left[\frac{f_j(x)}{g_j(x)} \right] + log \left[\frac{g_j(x)}{f_j(x)} \frac{f^{(k)}(x)}{g^{(k)}(x)} \right] \right\} dx \\
&= \sum_{j=1}^{k} \pi_j d(f_j, g_j) + \sum_{j=1}^{k} \pi_j \int f_j(x) log \left[\frac{g_j(x)}{f_j(x)} \frac{f^{(k)}(x)}{g^{(k)}(x)} \right] dx \\
&= d^*(f^{(k)}, g^{(k)}) + \sum_{j=1}^{k} \pi_j \int f_j(x) log \left[\frac{g_j(x)}{f_j(x)} \frac{f^{(k)}(x)}{g^{(k)}(x)} \right] dx.
\end{aligned} \tag{4.15}$$

El segundo término en (4.15) es no positivo. Esto sigue inmediatamente usando $ln\ z \leq z - 1$ (para $z > 0$). Tenemos que

$$\sum_{j=1}^{k} \pi_j \int f_j(x) log \left[\frac{g_j(x)}{f_j(x)} \frac{f^{(k)}(x)}{g^{(k)}(x)} \right] dx \leq \sum_{j=1}^{k} \pi_j \int f_j(x) \left[\frac{g_j(x)}{f_j(x)} \frac{f^{(k)}(x)}{g^{(k)}(x)} - 1 \right] dx$$
$$= \int \left[f^{(k)}(x) \right] dx - 1 = 0.$$

Entonces

$$d(f^{(k)}, g^{(k)}) \leq d^*(f^{(k)}, g^{(k)}).$$

Respecto a la elección de la medida de distancia, se propone que la distancia KL ponderada sea usada en el procedimiento para probar si se debe fusionar componentes. A menudo es fácil calcular $d^*(f_j, g_j)$. En consecuencia, $d^*(f, g)$ puede ser con frecuencia obtenida explícitamente cuando $d(f, g)$ no. En el contexto del método de colapso propuesto, la distancia $d^*(f^{(k)}, g^{(k)})$ toma una forma particularmente sencilla.

4.2.5. Colapso $f^{(k)}$

Sea $f_{(12)}^{*(k-1)}$ una versión colapsada de $f^{(k)}$ la cual asumimos es obtenida colapsando los dos primeros componentes de $f^{(k)}$. Para calcular $d^*(f^{(k)}, f_{(12)}^{*(k-1)})$, vemos a la versión colapsada como un caso especial de una mezcla de k componentes, es decir, $g^{(k)} = f_{(12)}^{*(k-1)}$. En el proceso de colapso, los parámetros de $f^{(k)}$ son considerados fijos. La versión colapsada es entonces definida de la siguiente manera

$$g^{(k)} = f_{(12)}^{*(k-1)} = \sum_{j=1}^{k} \pi_j f_j(x|\theta_j^*),$$

con la condición que $\theta_1^* = \theta_2^* = \theta^*$ y $\theta_j^* = \theta_j$ para $j = 3, \ldots, k$. Entonces, $g_j(x|\theta_j^*) = f_j(x|\theta_j)$ para $j = 3, \ldots, k$ y $d^*(f^{(k)}, f_{(12)}^{*(k-1)})$ se reduce a

$$\begin{aligned} d^*(f^{(k)}, f_{(12)}^{*(k-1)}) &= \sum_{j=1}^{2} \pi_j \, d(f_j(\cdot|\theta_j), f_j(\cdot|\theta^*)) \\ &= \sum_{j=1}^{2} \pi_j E_j \, ln \frac{f_j(X|\theta_j)}{f_j(X|\theta^*)} \end{aligned} \tag{4.16}$$

donde E_j denota la esperanza bajo la densidad $f(x|\theta_j)$ y

$$\theta^* = arg\left\{\min_{\theta}\sum_{j=1}^{2}\pi_j E_j\, ln\frac{f_j(X|\theta_j)}{f_j(X|\theta)}\right\}$$
$$= arg\left\{\min_{\theta}\sum_{j=1}^{2}\pi_j E_j\, ln\, f_j(X|\theta)\right\}.$$

La fórmula anterior define la versión colapsada obtenida de la fusión de los dos primeros componentes de $f^{(k)}$. La mejor versión colapsada, denotada por $f^{*(k-1)}$, es simplemente la que minimiza $d(f^{(k)}, f^{*(k-1)}_{(ij)})$ sobre todas las i,j con $i \neq j$, es decir, la mejor de todas las ${}^kC_2$ versiones colapsadas de $f^{(k)}$.

4.2.6. Distancia

En Mengersen *et al.* [12] se desarrollaron diferentes lemas para el cálculo del colapso.
Supongamos que $g_j(x|\theta^*_j, \Lambda_j) = N_p(\mathbf{x}; \theta^*_j, \Lambda_j)$ y

$$g^{(k)}(\mathbf{x}) = \sum_{j=1}^{k}\pi_j g_j(x|\theta^*_j, \Lambda_j).$$

Lema. 1 *La distancia (4.14) entre $f^{(k)}$ y $g^{(k)}$ está dada por,*

$$2d^*(f^{(k)}, g^{(k)}) = \sum_{j=1}^{k}\pi_j\left\{log|\Lambda_j\Sigma_j^{-1}| + tr(\Lambda_j^{-1}\Sigma_j) - p + (\theta_j - \theta^*_j)^t\Lambda_j^{-1}(\theta_j - \theta^*_j)\right\} \tag{4.17}$$

donde los dos primeros componentes de $g^{(k)}$ tienen los parámetros comunes θ^* y Λ, y todos los otros componentes tienen parámetros iguales a los parámetros correspondientes de $f^{(k)}$. Esto es $\theta^*_1 = \theta^*_2 = \theta^*$ y $\Lambda_1 = \Lambda_2 = \Lambda$ y $\theta^*_i = \theta_i$, $\Lambda i = \Sigma_i$, $i = 3, \ldots, k$. Consecuentemente, $g^{(k)}$ es una mezcla de $k-1$ componentes. Nótese que los pesos de la mezcla $\pi_1, \ldots, \pi_k$ no cambian.

El Lema 2.2 en Mengersen y Robert (1996) es un caso especial del Lema 1 anteriormente descrito y presentado en Sahu *et al.* [11].

4.2.7. Elección de c_k y α

Reducimos de k a *k-1* si $d(f^{(k)}, f^{*(k-1)})$ es pequeña. Una distancia grande enfatiza en que el colapso no puede ser hecho sin deteriorar significativamente el ajuste. La solución bayesiana es evaluar la probabilidad posterior

$$P_{c_k} = P_r\{d(f^{(k)}, f^{*(k-1)}) \leq c_k | datos\}. \tag{4.18}$$

Una regla de decisión simple es reducir de k a *k-1* si la probabilidad posterior anterior es grande (mayor que $\alpha = 0.5$).

Sahu *et al.* [11], para la selección de c_k, hacen diversos estudios de simulación y c_k tiene que ver con el contexto del problema planteado, es decir, el método es sensible a la elección de c_k; un valor grande de c_k seleccionará un menor número de componentes y un valor pequeño de c_k seleccionará un mayor número de componentes. Así que, basándonos en estas apreciaciones, se considerará que $c_k = 0.5$ es una elección razonable.

Seleccionar c_k es un problema específico y puede convertirse en algo sumamente complejo, y debería ser guiado por el tamaño requerido de la zona de indiferencia entre $f^{(k)}$ y su versión colapsada.

Tanto el cálculo de (4.18) como la elección de c_k se verán a detalle en el siguiente Capítulo a la par con la aplicación.

Capítulo 5

Aplicación

En este capítulo se llevará a cabo la aplicación de la teoría y fundamentos de los enfoques descritos en la presente tesis, desarrollados en los capítulos anteriores.

A partir de ahora, la manipulación de la información así como todos aquellos resultados serán efectuados con el programa libre para cálculos estadísticos `R` tal como se realizó en el capítulo 3 para ejemplificar la implementación de métodos jerárquicos y partición. La sintaxis de las rutinas realizadas para el análisis de la información están disponibles en el Apéndice B.

Se analizará el conjunto de datos Ruspini (1970 [13]), formado por 75 casos u observaciones en dos variables dadas por las coordenadas `x` y `y`. Cabe señalar que esta información ha sido ampliamente utilizada para ilustrar el desempeño de nuevas técnicas de agrupamiento.

Es bien sabido que existe una gran controversia sobre el número adecuado de grupos para estos datos, generalmente, se han agrupado en 4. Para el caso en particular se explorará la información partiendo de 5 grupos y así se procederá a aplicar cada uno de los enfoques descritos en el capítulo anterior.

La **Figura 5.1** muestra la dispersión de los datos. Es probable que con esta gráfica se aprecien cuatro conglomerados, sin mayor problema, pero aún así causa confusión la agrupación de algunas observaciones en la parte superior de la gráfica y eso es precisamente lo que se analizará.

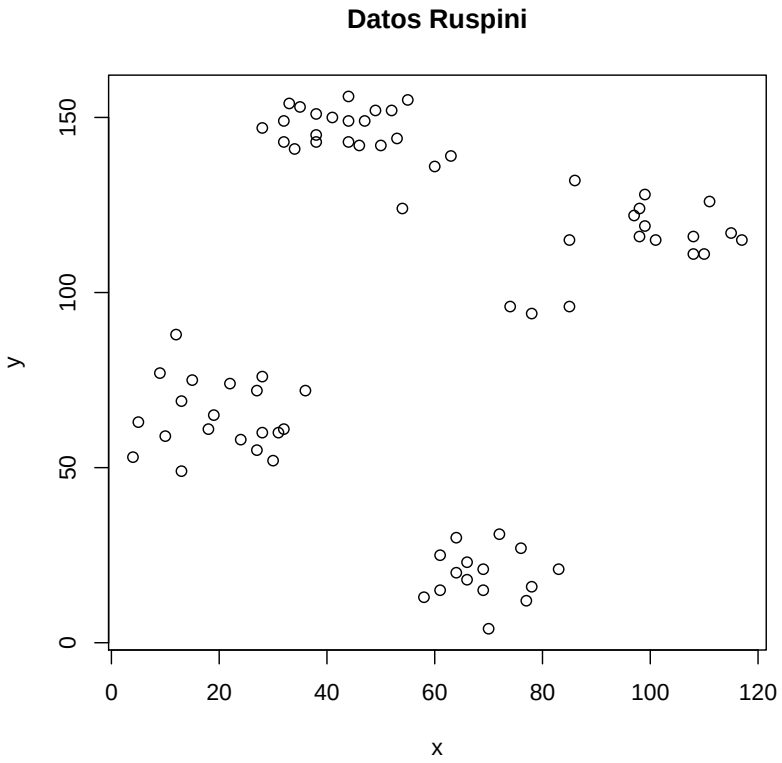

Figura 5.1: Dispersión de datos Ruspini

5.1. Implementación BIC

El cálculo del método propuesto mediante el programa `R` se realizará a través del paquete **`mclust`** [1], para realizar las rutinas referentes a la modelación de conglomerados basados en mezclas de distribuciones.

Los mejores valores obtenidos a través del BIC provisto en el paquete **`mclust`**, son los siguientes:

```
    VVV,4     VVV,5
-1380.783 -1389.082
```

[1] `mclust` está disponible como un paquete aportado por The Comprehensive `R` Archive Network (CRAN) en http://CRAN.R-project.org/

En el resultado anterior, se observa que el BIC se maximiza para el modelo VVV [2], es decir el modelo sin restricciones en la matriz de varianzas-covarianzas, con 4 grupos.

También se observa que en los mejores valores del BIC se incluye el ajuste con 4 y 5 grupos, y sus correspondientes valores de logverosimilitud tienen diferencia mínima.

Sin embargo se ha puesto énfasis en que la información se explorará a partir de 5 grupos. Por lo cual, el modelo de mezclas más adecuado con 5 componentes, es elipsoidal y consta de normales bivariadas con matrices de varianzas-covarianzas diferentes.

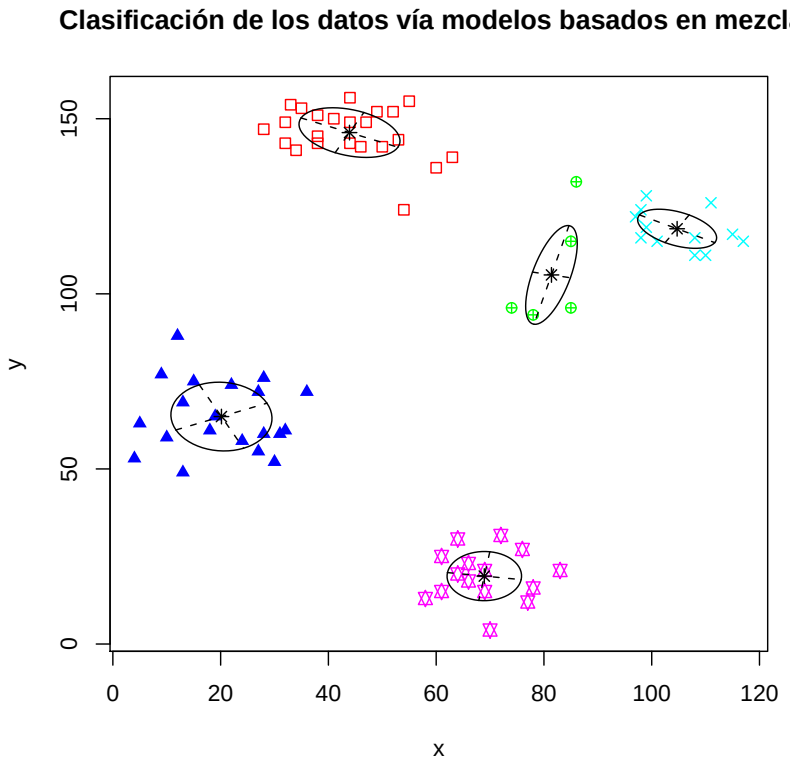

Figura 5.2: Asignación de los datos a cada grupo.

La gráfica anterior corresponde a la clasificación de los datos en 5 grupos.

[2]Todas los modelos disponibles pueden encontrar en el paquete **mclust**.

Donde los triángulos azules rellenos representan las observaciones clasificadas al grupo 1, los cuadros rojos sin relleno simbolizan el grupo 2, los círculos verdes el grupo 3, las cruces aguamarina el 4 y las estrellas fucsia el grupo 5.

La **Figura 5.3** muestra los mejores valores del BIC, y en -1380.783 se maximiza.

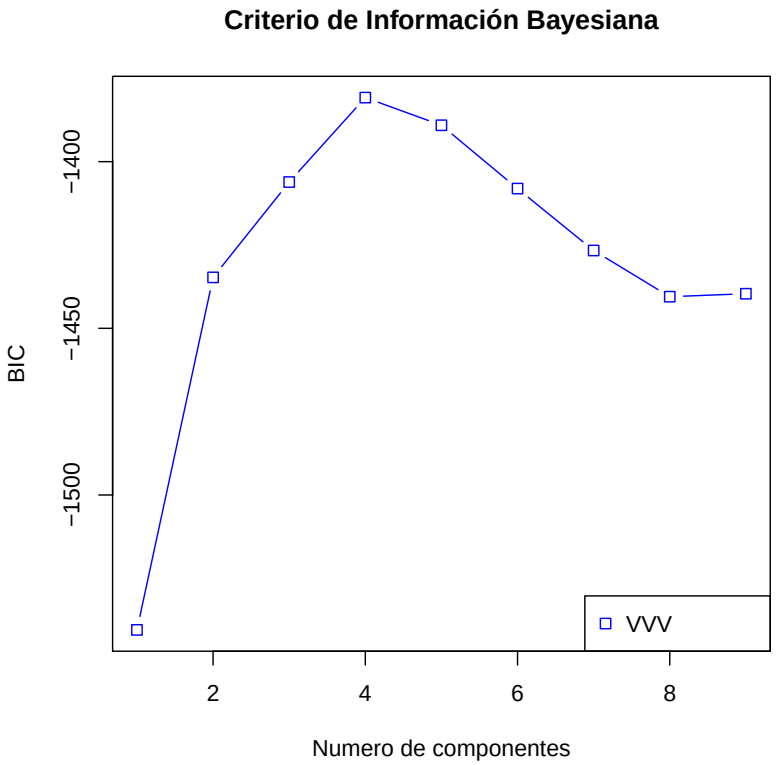

Figura 5.3: BIC

Anteriormente se elaboró el primer análisis del enfoque fundamentado en maximización. Posteriormente se trabajará con el planteamiento basado en distancia, pero antes se deplegarán los valores de los estimadores obtenidos por el programa, ya que son necesarios para el análisis de este segundo enfoque.

El Cuadro 5.1 muestra las probabilidades de pertenecer a cada grupo (pesos de la mezcla).

π_1	π_2	π_3	π_4	π_5
0.26666667	0.30667571	0.06354719	0.16311046	0.19999998

Cuadro 5.1: Vector de proporciones.

En el Cuadro 5.2 se observa los vectores de medias obtenidos en el ajuste.

	μ_1	μ_2	μ_3	μ_4	μ_5
x	20.15	43.914	81.3945	104.715	68.93
y	64.95	146.043	105.398	118.576	19.40

Cuadro 5.2: Vector de medias.

A continuación se muestran las matrices de varianzas-covarianzas de cada componente.

$$\Sigma_1 = \begin{pmatrix} 88.027527 & -4.792524 \\ -4.792524 & 96.447519 \end{pmatrix}$$

$$\Sigma_2 = \begin{pmatrix} 87.86817 & -22.27396 \\ -22.27396 & 50.30690 \end{pmatrix}$$

$$\Sigma_3 = \begin{pmatrix} 22.93867 & 42.91747 \\ 42.91747 & 198.99575 \end{pmatrix}$$

$$\Sigma_4 = \begin{pmatrix} 53.76398 & -18.81298 \\ -18.81298 & 31.18448 \end{pmatrix}$$

$$\Sigma_5 = \begin{pmatrix} 47.7955578 & 0.2266701 \\ 0.2266701 & 49.3066609 \end{pmatrix}$$

El procedimiento para obtener los parámetros anteriormente descritos se encuentra ubicado en el Apéndice B para su consulta.

5.2. Implementación enfoque basado en distancia

Para poner en funcionamiento este enfoque se trabajará en el supuesto de que existen 5 grupos en los datos, es decir, se procederá a implementar el procedimiento de prueba comenzando con $k_0 = 5$.

En el paso `T2` se espera el cálculo de las distancias $d(f^{(k)}, f^{*(k-1)}_{(ij)})$ entre $f^{(5)}$ y cada una de sus ${}^5C_2$ versiones colapsadas $f^{*(4)}_{(ij)}$. Dado que se comenzará el análisis con 5 grupos, las correspondientes combinaciones son: $(1,2),(1,3),(1,4),(1,5),(2,3),(2,4),(2,5),(3,4),(3,5),(4,5)$.

El resultado de las distancias se obtuvo aplicando el Lema desarrollado por los autores, descrito en el capítulo 4 del presente trabajo.

La distancia (4.14) entre $f^{(k)}$ y $g^{(k)}$ está dada por,

$$2d^*(f^{(k)}, g^{(k)}) =$$

$$\sum_{j=1}^{k} \pi_j \left\{ log|\Lambda_j \Sigma_j^{-1}| + tr(\Lambda_j^{-1}\Sigma_j) - p + (\theta_j - \theta_j^*)^t \Lambda_j^{-1} (\theta_j - \theta_j^*) \right\}.$$

Se comienza con $(1,2)$, suponiendo que es la combinación que nos proporcionó la mejor versión colapsada. Desarrollando, se tiene que,

$$\begin{aligned}
2d^*(f^{(5)}, g^{(5)}) = \\
&\pi_1 \left\{ log|\Lambda \Sigma_1^{-1}| + tr(\Lambda^{-1}\Sigma_1) - 2 + (\theta_1 - \theta^*)^t \Lambda^{-1} (\theta_1 - \theta^*) \right\} + \\
&\pi_2 \left\{ log|\Lambda \Sigma_2^{-1}| + tr(\Lambda^{-1}\Sigma_2) - 2 + (\theta_2 - \theta^*)^t \Lambda^{-1} (\theta_2 - \theta^*) \right\} + \\
&\pi_3 \left\{ \overbrace{log|\Lambda_3 \Sigma_3^{-1}|}^{0} + \overbrace{tr(\Lambda_3^{-1}\Sigma_3)}^{2} - 2 + \overbrace{(\theta_3 - \theta_3^*)^t \Lambda_3^{-1} (\theta_3 - \theta_3^*)}^{0} \right\} + \\
&\pi_4 \left\{ \overbrace{log|\Lambda_4 \Sigma_4^{-1}|}^{0} + \overbrace{tr(\Lambda_4^{-1}\Sigma_4)}^{2} - 2 + \overbrace{(\theta_4 - \theta_4^*)^t \Lambda_4^{-1} (\theta_4 - \theta_4^*)}^{0} \right\} + \\
&\pi_5 \left\{ \overbrace{log|\Lambda_5 \Sigma_5^{-1}|}^{0} + \overbrace{tr(\Lambda_5^{-1}\Sigma_5)}^{2} - 2 + \overbrace{(\theta_5 - \theta_5^*)^t \Lambda_5^{-1} (\theta_5 - \theta_5^*)}^{0} \right\} =
\end{aligned}$$

$$\pi_1 \left\{ log|\Lambda\Sigma_1^{-1}| + tr(\Lambda^{-1}\Sigma_1) - 2 + (\theta_1 - \theta^*)^t \Lambda^{-1}(\theta_1 - \theta^*) \right\} +$$
$$\pi_2 \left\{ log|\Lambda\Sigma_2^{-1}| + tr(\Lambda^{-1}\Sigma_2) - 2 + (\theta_2 - \theta^*)^t \Lambda^{-1}(\theta_2 - \theta^*) \right\}.$$

Donde los dos primeros componentes de $g(5)$ tienen parámetros comunes θ^* y Λ, y los demás componentes tienen parámetros iguales a los correspondientes en $f(5)$. Esto es $\theta_1^* = \theta_2^* = \theta^*$ y $\Lambda_1 = \Lambda_2 = \Lambda$ y $\theta_i^* = \theta_i$, $\Lambda_i = \Sigma_i$, $i = 3, 4, 5$. Nótese que los pesos de la mezcla $\pi_1, ..., \pi_5$ no cambian.

Para esta aplicación en particular, se considerará a los parámetros θ_1, θ_2, Σ_1 y Σ_2 como la media y varianza muestral, respectivamente, y que corresponden a los grupos 1 y 2 colapsados; ya que son con los que se está trabajando.

Y para los demás $\theta_i^* = \theta_i$, $\Lambda_i = \Sigma_i$ con $i = 3, 4, 5$, serán los obtenidos en el ajuste del BIC, en la sección anterior. Cabe señalar que se hizo lo mismo para las otras nueve combinaciones.

Las distancias obtenidas entre $f^{(5)}$ y cada una de las ${}^5C_2$ combinaciones son:

$(1,2)$	$(1,3)$	$(1,4)$	$(1,5)$	$(2,3)$
1.665471	2.113016	1.692010	1.394676	2.260789

$(2,4)$	$(2,5)$	$(3,4)$	$(3,5)$	$(4,5)$
1.481169	1.926839	1.201983	1.746112	1.608675

Se observa que con la combinación $(3,4)$ la $d(f^{(5)}, f_{(ij)}^{*(4)})$ es minimizada. Tal como se propone en `T3` del procedimiento de prueba.

En `T4` se pretende la evaluación de la probabilidad posterior

$$P_{c_k} = P_r\{d(f^{(k)}, f^{*(k-1)}) \leq c_k | datos\}.$$

Si bien se ha seguido la esencia de los autores en cuanto al desarrollo y estudio del enfoque basado en distancia, y donde para el cálculo de la probabilidad posterior ellos realizaron estudios de simulación; en el presente trabajo, el cálculo de dicha probabilidad estará orientado a la exploración de los datos *per se*. Es decir, nos dimos a la tarea de generar mil muestras

aleatorias con reemplazo obtenidas de los datos Ruspini originales, semejando al método **Bootstrap**, el cual precisamente consiste en generar un gran número de muestras de tamaño n efectuando un muestreo con reemplazamiento de esos valores y de esa manera calcular el valor de algún parámetro que se quiera estimar.

Sin embargo, no se reajustará cada muestra nuevamente, sino que los parámetros obtenidos en el ajuste de los datos originales son los que se utilizarán en cada una de las mil muestras generadas. Por lo que la exploración de los datos tiene cabida en que la obtención de la probabilidad de que la observación x_i de la N-ésima muestra, donde $N = 1000$, pertenezca a cierto grupo se obtendrá a través del reconocimiento de las $\hat{z}_{ij}$.

$$\hat{z}_{ij} \leftarrow \frac{\hat{\pi}_j f_j(x_i|\hat{\mu}_j, \hat{\Sigma}_j)}{\sum_{l=1}^{k} \hat{\pi}_l f_l(x_i|\hat{\mu}_l, \hat{\Sigma}_l)}.$$

De este modo, se hará un reclasificación de la información para cada muestra.

Posteriormente, se obtendrá la $d(f^{(5)}, f^{*(4)})$ de cada muestra. Haciendo remembranza, la combinación $(3, 4)$ fue la que minimizó $d(f^{(5)}, f^{*(4)}_{(ij)})$, entonces bajo esa circunstancia cada una de las mil muestras heredarán esta característica y nuevamente, se considerarán los parámetros θ_3, θ_4, Σ_3 y Σ_4 como la media y varianza muestral, respectivamente, y que corresponden a los grupos 3 y 4 colapsados.

Las distancias obtenidas serán depositadas en un vector d de dimensión 1000×1 para luego ser etiquetadas con uno si son menores o iguales a un determinado c_k o cero en caso contrario, y así proseguir con la regla de decisión donde se reduce de k a *k-1* si la probabilidad posterior es mayor a $\alpha = 0.5$.

Inicialmente, se probará con un $c_k = 0.5$ tal y como los propusieron los autores. La siguiente sentencia realizada en `R` asigna a `x`, justamente lo que se consideró en el párrafo anterior.

```
> x<-1*(d<=0.5)
```

Desplegando `x`.

```
> x
```

```
   [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [186] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [223] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [260] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [297] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [334] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [371] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [408] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [445] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [482] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [519] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [556] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [593] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [630] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [667] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [704] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [741] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [778] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [815] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [852] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [889] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [926] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [963] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[1000] 0
```

Dado el resultado anterior, P_{c_k} va ser igual a 0 ya que no existió algún caso favorable, del mismo modo P_{c_k} tampoco será mayor a un $\alpha > 0.5$ por lo que no se reemplazará $f^{(5)}$ por $f^{*(4)}$.

Este resultado indicaría que 5 es un número adecuado de grupos que representan a los datos. Sin embargo, se debe recordar que los autores hicieron estudios de simulación y el $c_k = 0.5$ utilizado para ejemplificar el enfoque estuvo sometido al contexto.

Se realiza nuevamente el mismo análisis con c_k distinto, uno que tenga que ver, en particular, con la presente aplicación. Se utilizará un c_k igual al promedio de las distancias obtenidas entre $f^{(5)}$ y cada una de las ${}^5C_2$ combinaciones, el cual es 1.7090.

Ahora se ejecuta la sentencia con $c_k = 1.7090$.

```
> x<-1*(d<=1.7090)
```

El resultado que se obtuvo fue $P_{c_k} = 0.664$, existiendo 664 casos favorables, además P_{c_k} fue mayor a un $\alpha > 0.5$ por lo tanto se reemplazará

$f^{(5)}$ por $f^{*(4)}$ y se deberá repetir el procedimiento de prueba desde `T2` con k reducido por 1.

Primeramente, se ajustará la mezcla con el valor final de k, basándonos en los valores obtenidos a través del BIC pero ajustados a 4 grupos.

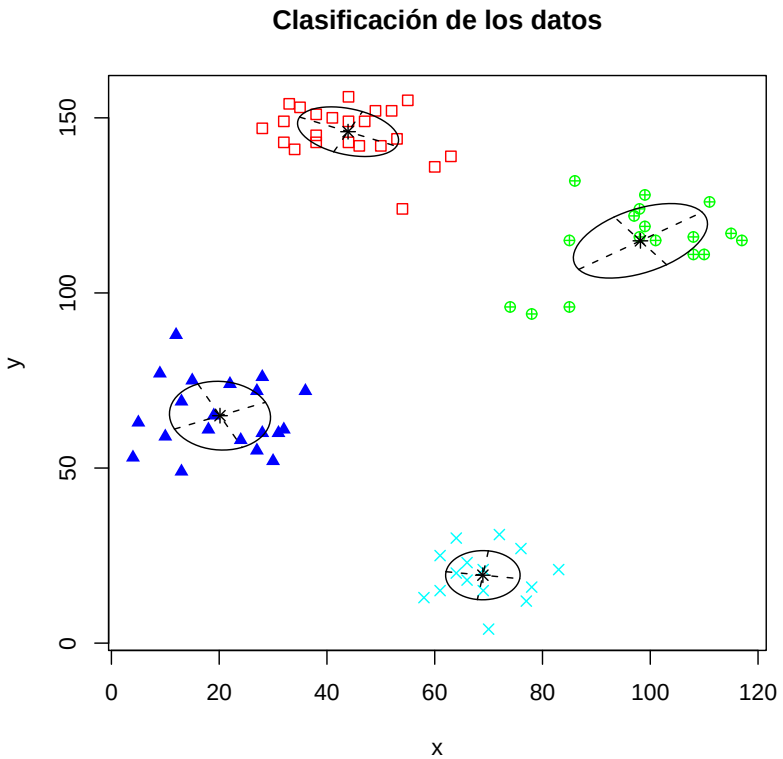

Figura 5.4: Asignación de los datos en 4 grupos

La gráfica anterior corresponde clasificación de los datos en 4 grupos. Donde los triángulos azules rellenos representan las observaciones clasificadas al grupo 1, los cuadros rojos sin relleno simbolizan el grupo 2, los círculos verdes el grupo 3 y las cruces aguamarina el 4.

Los nuevos parámetros generados son [3]:

[3]El procedimiento para obtener dichos parámetros está descrito en el Apéndice B para su consulta.

π_1	π_2	π_3	π_4
0.2666666	0.3066703	0.2266630	0.2000000

Cuadro 5.3: Vector de proporciones.

	μ_1	μ_2	μ_3	μ_4
x	20.15	43.91617	98.1731	68.93333
y	64.95	146.04395	114.8812	19.40000

Cuadro 5.4: Vector de medias.

$$\Sigma_1 = \begin{pmatrix} 88.027511 & -4.792537 \\ -4.792537 & 96.447524 \end{pmatrix}$$

$$\Sigma_2 = \begin{pmatrix} 87.97607 & -22.29453 \\ -22.29453 & 50.28116 \end{pmatrix}$$

$$\Sigma_3 = \begin{pmatrix} 155.05182 & 60.47337 \\ 60.47337 & 113.25518 \end{pmatrix}$$

$$\Sigma_4 = \begin{pmatrix} 47.7955558 & 0.2266676 \\ 0.2266676 & 49.3066659 \end{pmatrix}$$

Ahora se calculan las distancias $d(f^{(k)}, f^{*(k-1)}_{(ij)})$ entre $f^{(4)}$ y cada una de sus ${}^4C_2$ versiones colapsadas $f^{*(3)}_{(ij)}$ para 4 grupos. Las correspondientes combinaciones son: $(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)$.

Las distancias obtenidas entre $f^{(4)}$ y cada una de las ${}^4C_2$ combinaciones son:

$(1,2)$	$(1,3)$	$(1,4)$	$(2,3)$	$(2,4)$	$(3,4)$
1.665578	1.417203	1.394676	1.421433	1.926968	1.507903

Se observa que con la combinación $(1,4)$ la $d(f^{(4)}, f^{*(3)}_{(ij)})$ es minimizada.

Aplicando `T4`, nuevamente se probará con $c_k = 0.5$ ahora considerando cuatro grupos.

```
> x<-1*(d<=0.5)
```

Desplegando `x`.

```
> x
```

```
   [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [186] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [223] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [260] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [297] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [334] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [371] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [408] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [445] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [482] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [519] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [556] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [593] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [630] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [667] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [704] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [741] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [778] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [815] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [852] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [889] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [926] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [963] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[1000] 0
```

Dado el resultado anterior, P_{c_k} va ser igual a 0 ya que no existió algún caso favorable, del mismo modo P_{c_k} tampoco será mayor a un $\alpha > 0.5$ por lo que no se reemplazará $f^{(4)}$ por $f^{*(3)}$.
Este resultado es idéntico a lo realizado anteriormente para 5 grupos.
Ahora se ejecuta la sentencia con $c_k = 1.555627$. Este c_k es igual al promedio de las distancias obtenidas entre $f^{(4)}$ y cada una de las ${}^4C_2$ combinaciones con cuatro grupos.

```
> x<-1*(d<=1.555627)
```

Desplegando `x`.

```
> x
```

```
   [1] 1 1 0 1 1 1 1 1 0 1 1 1 1 1 0 0 0 1 1 1 0 0 1 1 1 1 0 1 1 1 0 1 1 1 1 1 1
  [38] 0 1 1 1 1 0 0 1 1 0 0 1 1 0 1 0 0 1 1 1 0 0 1 1 1 0 0 0 0 1 0 0 1 1 0 1 0
  [75] 0 0 1 1 0 0 1 0 1 1 1 1 1 0 1 0 0 1 1 1 1 1 0 1 0 1 1 0 1 1 1 0 1 0 1 1 1
 [112] 1 0 1 1 1 1 0 1 1 1 1 0 1 1 0 0 1 0 1 1 1 1 0 1 1 1 1 1 1 1 0 1 0 1 0 1 1
 [149] 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 0 0 0 1 0 0 1 1 1 1 1
 [186] 0 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 1 0 0 1 1 1 0 1 1 0
 [223] 0 1 1 1 0 0 1 0 1 1 0 1 0 1 1 0 0 1 0 1 1 1 0 1 1 1 1 1 1 0 1 1 1 0 1 1 1 1
 [260] 1 1 1 1 0 0 1 0 1 1 1 0 0 1 1 0 0 1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 0 1 1 0 1 1 1 1 1
 [297] 1 1 0 0 0 1 1 0 0 0 1 1 1 1 0 1 1 0 1 1 0 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 0 1 0
 [334] 0 1 1 1 1 1 1 0 1 0 1 1 1 0 1 1 1 0 0 1 1 1 1 1 1 1 1 0 1 0 0 1 1 0 0 1 1 1
 [371] 0 0 1 0 0 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 0 1 0 1 1 1 1 1 1 1 1 1
 [408] 1 1 1 1 0 1 0 1 0 0 1 1 1 0 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0
 [445] 1 0 0 1 1 1 0 0 0 1 1 0 1 1 1 1 1 0 1 0 0 1 1 1 0 0 1 1 1 1 0 1 1 0 0 1 1
 [482] 1 1 0 1 0 1 1 1 0 1 1 0 0 1 1 0 1 0 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 0 1 0 1 1
 [519] 0 1 1 1 1 0 1 0 1 0 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 1 0 1 0 1 0 1 1 0 1 0 1 1
 [556] 1 0 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 0 0 1 1 0 0 1 1 1 1 1 1 0 1 1 0 0 0 1 1 1 0
 [593] 1 1 1 1 0 1 1 0 1 1 1 0 1 0 0 1 1 0 0 0 1 0 1 0 1 0 0 1 1 1 1 1 1 1 0 0 1 1 0
 [630] 0 1 1 1 0 1 0 1 1 1 1 1 0 1 1 0 0 0 1 1 1 1 0 1 0 1 0 1 1 0 1 0 0 0 0 1 1 0 1
 [667] 0 1 1 0 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 0 0 1 1 1 1 1 1 0 1 0 1 1 1 0 1 1 1
 [704] 0 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 0 0
 [741] 1 1 1 0 0 1 1 1 1 1 1 0 1 0 1 1 1 1 0 1 1 1 0 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1
 [778] 1 0 1 1 1 0 1 1 1 1 1 1 1 0 1 1 0 1 0 0 1 1 0 1 1 1 1 1 0 1 0 1 1 1 1 1 1 0
 [815] 1 1 1 0 1 1 1 1 1 1 1 1 0 1 0 1 1 0 1 0 1 1 0 1 0 0 1 1 1 1 1 1 0 1 0 1 1 0 1
 [852] 0 1 0 1 1 0 0 0 1 0 0 1 1 1 1 0 0 0 1 1 1 0 1 1 1 1 1 1 0 0 0 1 0 1 1 1 1 0
 [889] 1 1 0 0 1 1 0 0 1 0 0 0 0 0 1 1 0 1 0 1 1 1 0 1 0 0 0 1 1 1 1 1 1 1 0 1 0 1
 [926] 1 1 0 1 1 0 1 1 1 0 1 0 0 1 1 1 0 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1 1
 [963] 1 1 1 1 1 0 0 0 0 1 1 1 1 0 0 1 0 1 1 1 1 1 0 1 1 1 0 0 1 1 0 1 1 0 1 1 1 1 0
[1000] 1
```

El resultado obtenido tras ejecutar la rutina fue $P_{c_k} = 0.689$, existiendo 689 casos favorables, además P_{c_k} fue mayor a un $\alpha > 0.5$ por lo que se sugiere reemplazar $f^{(4)}$ por $f^{*(3)}$ e implementar de nuevo el procedimiento de prueba.

Aunque, tanto intuitiva como gráficamente la sugerencia previa causa "ruido" esto aunado con el resultado arrojado por el BIC donde si recordamos los mejores valores eran -1380.783 y -1389.082 con $G = 4, 5$ respectivamente.

Por lo tanto, se podría decir que la metodología no arrojó los resultados deseados ya que de alguna manera se esperaba que el procedimiento se hubiera detenido al sugerir un modelo de 4 a 3 componentes.

Pero antes de tomar alguna determinación se explorará nuevamente `T4`, para la evaluación de la probabilidad posterior, a manera de ver que tan sensible puede llegar a ser con un c_k generado por el promedio de las distancias más pequeñas con 4 grupos, es decir, con el promedio de:

$(1,3)$	$(1,4)$	$(2,3)$
1.417203	1.394676	1.421433

Se ejecuta la sentencia con $c_k = 1.411104$.

```
> x<-1*(d<=1.411104)
```

El resultado obtenido fue $P_{c_k} = 0.228$, por lo que P_{c_k} no fue mayor a un $\alpha > 0.5$ por lo tanto se sugiere que no se reemplace $f^{(4)}$ por $f^{*(3)}$ de modo que el procedimiento se detiene.

Con esta última exploración se llega a la conclusión de que 4 es un número adecuado de grupos que representan a los datos, con $c_k = 1.411104$ se obtuvo lo que intuitivamente se esperaba. Se consideró que este último valor fue más adecuado ya que se excluyeron las distancias que hacían más grande el promedio, estas estaban más lejos de minimizar $d(f^{(k)}, f^{*(k-1)}_{(ij)})$. Sin embargo estos resultados pueden causar ambigüedad, se entiende que esto es consecuencia debido a la elección de c_k, ya que como se ha venido señalando durante el desarrollo del enfoque basado en distancia, la preferencia por algún c_k se puede convertir en un problema específico o en algo sumamente complejo.

Capítulo 6

Conclusiones

Durante el desarrollo de la presente tesis, desde la introducción hasta la aplicación, se han facilitado las numerosas utilidades que tiene el análisis de conglomerados en diferentes áreas del conocimiento.

Fue pertinente iniciar con los métodos clásicos ya que históricamente fueron las primeras herramientas utilizadas para el estudio de conglomerados. Posteriormente, se proporcionó la teoría que envuelve a los métodos basados en mezclas y aunque se han propuesto distintas aplicaciones para estos modelos, autores como Banfield y Raftery diseñaron un algoritmo llamado **mclust** que funciona bastante bien en la práctica cuando se tienen grupos bien separados y estos siguen significativamente una distribución normal. Y es a partir de ahí donde se presentaron las dos diferentes propuestas dirigidas directamente al problema de determinar el número de grupos dentro de cualquier conjunto de datos.

El primer enfoque propuesto en esta tesis fue el basado en maximización, el razonamiento utilizado para seleccionar el número adecuado de grupos fue maximizar el criterio de información bayesiana (BIC) obteniendo así el mejor modelo.

Con base en este primer análisis se desprende una serie de conclusiones relevantes. Se analizó el conjunto de datos Ruspini formado por 75 observaciones incluídos en `R`. Los resultados obtenidos a través del BIC provisto en el paquete `mclust` fueron:

Gráficamente se puede apreciar que los resultados son adecuados, esto aunado a lo que la literatura se refiere respecto a estos datos.

Por otro lado, la segunda alternativa presentada fue basada en la distancia Kullback-Leibler (KL) propuesta por Mengersen y Robert (1996). Se

G	BIC
4	-1380.783
5	-1389.082

consideró un metódo reductivo paso a paso donde la mezcla inicial $k = k_0$ (>1) componentes es ajustada a un grupo de datos donde se asume que el valor de k_0 es más que adecuado, luego el valor de k es progresivamente reducido hasta que el ajuste ya no sea aceptable. Esta propuesta también sirvió para determinar el número desconocido de componentes de la mezcla.

Durante la aplicación de los enfoques a los datos Ruspini, se insistió que para la segunda propuesta se comenzaría el procedimiento de prueba con $k_0 = 5$. Siguiendo el paso `T2`, se obtuvo todas las $d(f^{(5)}, f^{*(4)}_{(ij)})$ entre $f^{(5)}$ y cada una de las diez versiones colapsadas donde (3,4) fue para la cual la distancia fue minimizada. Y de esa manera se seleccionó la mejor versión colapsada tal como se indica `T3`.

En `T4` se pretende la evaluación de la probabilidad posterior para lo cual se propuso la exploración de los datos, generándose mil muestras aleatorias con reemplazo obtenidas de los datos Ruspini originales. La exploración de los datos se dió a través del reconocimiento de las $\hat{z}_{ij}$, dándose así una reclasificación de la información en cada muestra para luego obtener la $d(f^{(5)}, f^{*(4)})$ de cada muestra. Estas distancias fueron depositadas en un vector d de dimensión 1000×1 y se etiquetaron con uno si fueron menores o iguales a un determinado c_k o cero en caso contrario, y así proseguir con la regla de decisión donde se reduce de k a *k-1* si la probabilidad posterior es mayor a $\alpha = 0.5$.

Primero se probó para un $c_k = 0.5$, tal y como lo propusieron los autores. El resultado que arrojó este ensayo fue $P_{c_k} = 0$ ya que no existió algún caso favorable, del mismo modo P_{c_k} tampoco será mayor a un $\alpha > 0.5$ por lo que no se reemplazará $f^{(5)}$ por $f^{*(4)}$.

Después se probó para $c_k = 1.7090$, es decir, el promedio de las distancias obtenidas entre $f^{(5)}$ y cada una de las ${}^5C_2$ combinaciones. El resultado obtenido fue $P_{c_k} = 0.664$, y como P_{c_k} fue mayor a un $\alpha > 0.5$ se sugirió reemplazar $f^{(5)}$ por $f^{*(4)}$.

Se aplicó nuevamente el procedimiento de prueba desde `T2`, ajustándo la mezcla con $k = 4$ y ahora la combinación de componentes $(1, 4)$ fue la que minimizó la $d(f^{(4)}, f^{*(3)}_{(ij)})$. Para la evaluación de la probabilidad posterior en

T4, se utilizó la misma metodología y nuevamente se probó para $c_k = 0.5$ donde el resultado logrado fue idéntico a lo realizado anteriormente para 5 grupos con ese mismo valor.

Por lo que, se trató para un c_k obtenido del promedio de las distancias conseguidas entre $f^{(4)}$ y cada una de las ${}^4C_2$ combinaciones con 4 grupos, es decir, $c_k = 1.555627$. Se logró $P_{c_k} = 0.689$, y como P_{c_k} fue mayor a un $\alpha > 0.5$ se sugiere reemplazar $f^{(4)}$ por $f^{*(3)}$.

El hecho de que en el ensayo anterior, el procedimiento haya sugerido nuevamente reemplazar $f^{(4)}$ por $f^{*(3)}$ indicaría que con 3 componentes la información tendría un ajuste más adecuado. Esto causa confusión ya que se esperaba que ambos enfoques concluyeran resultados similares, es decir, que 4 ó 5 componentes serían los adecuados para clasificar la información.

Sin embargo, se realizó un último ensayo con $c_k = 1.411104$, considerándolo más adecuado ya que se obtuvo del promedio de las distancias más pequeñas con 4 grupos, por lo que es menos pesado a diferencia de incluir todas las distancias. El resultado conseguido fue $P_{c_k} = 0.228$, donde P_{c_k} no fue mayor a un $\alpha > 0.5$ por lo tanto se sugiere que no se reemplace $f^{(4)}$ por $f^{*(3)}$ deteniéndose así el procedimiento.

Se concluye que aunque el enfoque basado en distancia es interesante y trabaja bien, para decidir si $f^{(k)}$ se reemplazará por $f^{*(k-1)}$, tiene mucha importancia la elección de c_k. En Sahu *et al.* [11], para la elección de c_k hicieron diversos estudios de simulación y c_k estuvo relacionado con el contexto del problema planteado. Por lo tanto, se puede ver que el método es sumamente sensible a la elección de c_k; un valor grande de c_k seleccionará un menor número de componentes y un valor pequeño de c_k seleccionará un mayor número de componentes. Asimismo se observó, a través de la aplicación del enfoque, que la elección c_k es un problema específico y puede convertirse en algo sumamente complejo y la investigación quedaría orientada al estudio de c_k y no al desarrollo de las propuestas presententadas en capítulos anteriores. También, a los resultados obtenidos en la aplicación del presente trabajo, se le dió peso la "intuición" y lo contribuido gráficamente, pero en dimensiones mayores ninguna gráfica aportará información acerca de la clasificación de los datos; por lo que el estudio se tendría que basar únicamente en los resultados obtenidos por el método.

Por estas razones, el BIC es actualmente un criterio sumamente recurrido y cuenta con fama mundial ya que arroja buenos resultados en diversas áreas como astronomía, biología, ingeniería, genética, mercadotecnia, medi-

cina, psiquiatría etc.

Para finalizar, téngase en cuenta que el propósito de la presente tesis fue ilustrar tan sólo dos enfoques orientados a solucionar el problema de determinar el número de grupos dentro de cualquier conjunto de datos, no significa que sean los únicos. Actualmente continúan apareciendo propuestas y efectuándose aplicaciones ó ensayos de diferentes procedimientos de clasificación, aumentando el volumen de trabajos publicados sobre este tema así como programas y rutinas de computación, lo que incrementa las posibilidades de obtener mejores resultados y así reducir la dificultad de encontrar soluciones satisfactorias, favorenciendo la investigación en esta área.

Apéndice A

Ejemplo. Eigenvalores y eigenvectores en R

Obtener los valores y vectores propios de la siguiente matriz simétrica.

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 1 \\ 3 & 1 & 5 \end{pmatrix}$$

Solución en R

```
> A = cbind(c(1, 2, 3), c(2, 4, 1), c(3, 1, 5))

> A

     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    2    4    1
[3,]    3    1    5

> eigen(A)

$values
[1]  7.5895980  3.3838454 -0.9734434

$vectors
           [,1]        [,2]       [,3]
[1,] -0.4799416 -0.01513700  0.8771698
[2,] -0.4732062 -0.83746472 -0.2733656
[3,] -0.7387367  0.54628172 -0.3947713
```

```
> B = eigen(A)

> attributes(B)

$names
[1] "values"  "vectors"

> D = diag(B$values)

> D

         [,1]     [,2]       [,3]
[1,] 7.589598 0.000000  0.0000000
[2,] 0.000000 3.383845  0.0000000
[3,] 0.000000 0.000000 -0.9734434

> U = B$vector

> U %*% D %*% t(U)

     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    2    4    1
[3,]    3    1    5
```

Apéndice B

Código en R

En este apéndice se encuentra la sintaxis realizada en el software R para la implementación de los dos enfoques estudiados (BIC y distancia), y que se ejemplifican en el Capítulo 5.

```
#####Cargar librería.
library(mclust)
library(cluster)
data(ruspini)


#####Gráfica de Dispersión de los datos Ruspini.
plot(ruspini, main="Datos Ruspini")

#####Conlgomerados usando modelos basados en mezclas. G=5
mezcla<-Mclust(ruspini, modelNames="VVV", G=5)
attach(mezcla)
attach(parameters)
mezcla

#####Obtención de parámetros con G=5, página 65.
mezcla$parameters$pro
mezcla$parameters$mean
mezcla$parameters$variance$sigma


#####Conlgomerados usando modelos basados en mezclas. G=4
mezcla<-Mclust(ruspini, modelNames="VVV", G=4)
attach(mezcla)
```

```
attach(parameters)
mezcla

#####Obtención de parámetros con G=4, página 71.
mezcla$parameters$pro
mezcla$parameters$mean
mezcla$parameters$variance$sigma


#####Gráfica de clasificación.
mclust2Dplot(ruspini, parameters=mezcla$parameters,z=mezcla$z)
title(main="Clasificasión de los datos",
+sub="Mediante modelos basados en mezclas")


#####BIC.
mezclaBIC<-mclustBIC(ruspini, modelNames="VVV")
mezclaBIC

#####Resumen de  como quedaron clasificadas las observaciones
##### y los 3 BIC's más altos.
mezclaSummary<-summary(mezclaBIC,data=ruspini)
mezclaSummary

#####Grafica BIC.
plot(mezclaBIC, xlab="Numero de componentes")

-----------------------------------------------------------------
-----------------------------------------------------------------
#####Obtención de las distancias correspondientes a cada combinación.
#####Lema 1. G=5

comb <- matrix(c(1,2,1,3,1,4,1,5,2,3,2,4,2,5,3,4,3,5,4,5),ncol=2,
+byrow=T)
d <- numeric( length(comb[,1]) )

for( i in 1:length(comb[,1]) ){
gp1 <- ruspini[which(mezcla$classification == comb[i,1]),]
gp2 <- ruspini[which(mezcla$classification == comb[i,2]),]
```

```
gp1 <- as.matrix(gp1)
gp2 <- as.matrix(gp2)

datos <- rbind(gp1,gp2)

mu <- c(mean(datos[,1]), mean(datos[,2]))
S1 <- cov(datos)

d[i] <-DISTANCIA(pro[comb[i,1]], pro[comb[i,2]],
+mean[,comb[i,1]], mean[,comb[i,2]],
+variance$sigma[ , ,comb[i,1]],
+variance$sigma[ , ,comb[i,2]], S1, mu, 2)
                          }
#####Despliegue de distancias para cada combinación.
d

#####Promedio de distancias.
mean(d)

#####Distancia mínima.
min(d)

-----------------------------------------------------------------
-----------------------------------------------------------------

#####Obtención de las distancias correspondientes a cada combinación.
#####Lema 1. G=4

comb <- matrix(c(1,2,1,3,1,4,2,3,2,4,3,4),ncol=2,byrow=T)
d <- numeric( length(comb[,1]) )

for( i in 1:length(comb[,1]) ){
gp1 <- ruspini[which(mezcla$classification == comb[i,1]),]
gp2 <- ruspini[which(mezcla$classification == comb[i,2]),]

gp1 <- as.matrix(gp1)
gp2 <- as.matrix(gp2)

datos <- rbind(gp1,gp2)
```

```
mu <- c(mean(datos[,1]), mean(datos[,2]))
S1 <- cov(datos)

d[i] <-DISTANCIA(pro[comb[i,1]], pro[comb[i,2]],
+mean[,comb[i,1]], mean[,comb[i,2]],
+variance$sigma[ , ,comb[i,1]],
+variance$sigma[ , ,comb[i,2]], S1, mu, 2)
                        }

#####Despliegue de distancias para cada combinación.
d

#####Promedio de distancias.
mean(d)

#####Distancia mínima.
min(d)
-----------------------------------------------------------------
-----------------------------------------------------------------
#####Procedimiento que para la evaluación posterior.

normalMV <- function(x,mu,S){

dens <- c(0)

dens <- (2*pi)^(-length(mu)/2) * det(S)^(-1/2) *
+exp( (-1/2) * t((x - mu)) %*% solve(S) %*% (x- mu) )

dens
                            }


DISTANCIA<-function(pro1,pro2,mu1,mu2,sigma1,sigma2,
+LAMBDA,MU,p){
TERMINO1<-(
pro1*(log(det(LAMBDA%*%solve(sigma1)))+
sum(diag(solve(LAMBDA)%*%sigma1))-p) +
(t(mu1-MU)%*%solve(LAMBDA)%*%(mu1-MU))
          )
```

```
TERMINO2<-(
pro2*(log(det(LAMBDA%*%solve(sigma2)))+
sum(diag(solve(LAMBDA)%*%sigma2))-p) +
(t(mu2-MU)%*%solve(LAMBDA)%*%(mu2-MU))
          )

(TERMINO1+TERMINO2)/2
             }

-----------------------------------------------------------------
-----------------------------------------------------------------
#####1000 muestras para G=5

ruspini <- as.matrix(ruspini)

n<-1000
d<-seq(0,0,length.out=n)
indexmuestra<-seq(0,0,length.out=n)
muestra<-matrix(seq(0,0,length.out=75*2),nrow=75)

for(N in 1:n){

indexmuestra<-sample(1:75,size=75,replace=TRUE)

for(i in 1:length(muestra[ ,2]))
  muestra[i, ]<-ruspini[indexmuestra[i], ]

M<-matrix(seq(0,0,length.out=length(muestra[ ,1])*length(pro)),
          nrow=length(muestra[ ,1]))

dmarg<-seq(0,0,length.out=length(pro))

for( i in 1:length(muestra[ ,1]) ){
 for( j in 1:length(pro) )
dmarg[j]<-pro[j]*normalMV(muestra[i, ],mean[ ,j],
+variance$sigma[ , ,j])

M[i, ]<-dmarg/sum(dmarg)
                                  }
```

```
prob<-matrix(seq(0,0,length.out=length(muestra[ ,1])*2),
+nrow=length(muestra[ ,1]))

 for( i in 1:length(muestra[ ,1]) ){
u<-runif(1)
acum<-M[i,1]
indice<-0
for( j in 1:length(pro) ){
if(acum>=u){
   indice<-j
   break
          }
else{
acum=acum+M[i,j+1]
         }
                                 }

prob[i,1]<-acum
prob[i,2]<-indice
                                    }

gp1 <- muestra[ which(prob[,2]==3), ]
gp2 <- muestra[ which(prob[,2]==4), ]

datos <- rbind(gp1,gp2)

mu <- c(mean(datos[,1]), mean(datos[,2]))
S1 <- as.matrix(var(datos))

d[N]<-DISTANCIA(pro[3], pro[4], mean[,3], mean[,4],
+variance$sigma[ , ,3], variance$sigma[ , ,4], S1, mu, 2)

}

#####Despliegue de vector de distancias.
d

-----------------------------------------------------------------
#####Criterio. Distancias menores o iguales que algún Ck=0.5
x<-1*(d<=0.5)
```

```
x
y<-sum(x)/1000
y
-----------------------------------------------------------------
#####Criterio. Distancias menores o iguales que Ck=1.7090
#####donde Ck es el promedio de las distancias.
x<-1*(d<=1.7090)
x
y<-sum(x)/1000
y

-----------------------------------------------------------------
-----------------------------------------------------------------
#####1000 muestras para G=4

ruspini <- as.matrix(ruspini)

n<-1000
d<-seq(0,0,length.out=n)
indexmuestra<-seq(0,0,length.out=n)
muestra<-matrix(seq(0,0,length.out=75*2),nrow=75)

for(N in 1:n){

indexmuestra<-sample(1:75,size=75,replace=TRUE)

for(i in 1:length(muestra[ ,2]))
  muestra[i, ]<-ruspini[indexmuestra[i], ]

M<-matrix(seq(0,0,length.out=length(muestra[ ,1])*length(pro)),
+nrow=length(muestra[ ,1]))

dmarg<-seq(0,0,length.out=length(pro))

for( i in 1:length(muestra[ ,1]) ){
for( j in 1:length(pro) )
dmarg[j]<-pro[j]*normalMV(muestra[i, ],mean[ ,j],variance$sigma[ , ,j])

M[i, ]<-dmarg/sum(dmarg)
                                    }
```

```
prob<-matrix(seq(0,0,length.out=length(muestra[ ,1])*2),
+nrow=length(muestra[ ,1]))

 for( i in 1:length(muestra[ ,1]) ){
u<-runif(1)
acum<-M[i,1]
indice<-0
for( j in 1:length(pro) ){
if(acum>=u){
   indice<-j
   break
          }
else{
acum=acum+M[i,j+1]
         }
                                 }

prob[i,1]<-acum
prob[i,2]<-indice
                                     }

gp1 <- muestra[ which(prob[,2]==1), ]
gp2 <- muestra[ which(prob[,2]==4), ]


datos <- rbind(gp1,gp2)

mu <- c(mean(datos[,1]), mean(datos[,2]))
S1 <- as.matrix(var(datos))

d[N]<-DISTANCIA(pro[1], pro[4], mean[,1], mean[,4],
+variance$sigma[ , ,1], variance$sigma[ , ,4], S1, mu, 2)

}

#####Despliegue de vector de distancias.
d
-----------------------------------------------------------------
#####Criterio. Distancias menores o iguales que algún Ck=0.5
```

```
x<-1*(d<=0.5)
x
y<-sum(x)/1000
y
-----------------------------------------------------------------
#####Criterio. Distancias menores o iguales que Ck=1.555627
#####donde Ck es el promedio de las distancias.
x<-1*(d<=1.555627)
x
y<-sum(x)/1000
y
-----------------------------------------------------------------
#####Criterio. Distancias menores o iguales que Ck=1.411104
#####donde Ck es el promedio de las distancias más pequeñas
x<-1*(d<=1.411104)
x
y<-sum(x)/1000
y
```

Bibliografía

[1] Hartigan, Jhon A. (1974) *Clustering Algorithms*. Departament of Statistics, Yale University, p.86.

[2] Sánchez, Miguel; Frutos, Gloria; L. Cuesta Pedro. (1996) *Estadística y matemáticas aplicadas*. Editorial Síntesis S.A., p.112.

[3] Kreyszig, Erwing. (1985) *Estadística Matemática, Principios y Métodos*. Editorial Limusa, p.219.

[4] Peña, Daniel. (2003) *Análisis de Datos Multivariantes*. Editorial McGrawHill.

[5] http://www.ine.es

[6] Banfield, J. D. y E. Raftery, Adrian. (1993) Model-based Gaussian and Non-Gaussian clustering. *Biometrics*, 803-821.

[7] Fraley, C y E. Raftery, Adrian. (1993) *How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis*. Departament of Statistics, University of Washington.

[8] Celeux, G. y Govaert, G. (1995) *Gaussian parsimonious clustering models*. Pattern Recognition, 781-793.

[9] Gutierrez, Eduardo. (1997) Métodos Computacionales en la Inferencia Bayesiana. *Monografías*. Instituto de Investigaciones en Matemáticas Aplicada y en Sistemas, UNAM, Vol. 6, No. 15, p.1.

[10] E. Kass, Robert y E. Raftery, Adrian. (1993) Bayes Factors and Model Uncertainty. *Technical Report* No. 254, p.7.

[11] Sahu, Sujit K. y Cheng, Russell C. H. (2002) *A Fast Distance Based Approach for Determining the Number of Components in Mixtures*. Faculty of Mathematical Studies, University of Southampton.

[12] Mengersen, K. y Robert, C. P. (1996) Testing for mixtures: A Bayesian Entropic Approach (with discussion). In *Bayesian Stadistics 5*, Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford University Press, 255-276.

[13] E. H. Ruspini (1970): Testing for mixtures: Numerical methods for fuzzy clustering. *Information Sciences*, 2, pp. 319-350.