Universidad Nacional Autónoma de México

Facultad de Ciencias

Formación y evaluación del número de conlgomerados usando modelos basados en mezclas

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
ACTUARIO

PRESENTA:
FABIÁN SÁNCHEZ VALDOVINOS

DIRECTOR DE TESIS:
DRA. RUTH SELENE FUENTES GARCÍA

2010

Agradecimientos

A mis padres *Fabián* y *Angelita*. A mi hermano *Mauricio* y a toda mi familia por que sin su apoyo, cariño, motivación y comprensión, jamás hubiera podido cumplir mis metas y darles esta satisfacción.

A todos mis profesores por ser la punta de lanza de lo que lograré en el futuro. A mi alma matter la UNAM y la Facultad de Ciencias por darme sólo un poco de todo lo que nos ofrece a nosotros sus estudiantes.

A mi gran amigo Ulises, gracias a el tuve la oportunidad de conocer el maravilloso mundo de la facultad de ciencias, las matemáticas, la actuaría y de darle un rumbo diferente y grandioso a mi vida.

A mis amigos del CCH vallejo por confiar, creer en mi, apoyarme y acompañarme desde que los conozco.

A mis amigos de la facultad Hugo, Diego, César, Sylvia y a los demás ya que con su compañía y apoyo pude atravesar por la dura estancia en la facultad y terminar ésta tesis.

A la Dra. Ruth Fuentes García por compartir conmigo un poco de su conocimiento y brindarme la oportunidad de realizar este trabajo.

A mis sinodales Act. Jaime Vázquez Alamilla, M. en A.P. María del Pilar Alonso Reyes, Mat. Margarita Elvira Chávez Cano y M. en C. Ricardo Ramírez Aldana por su tiempo y dedicación en la revisión de mi trabajo, así como en sus imprescindibles comentarios para mejorarlo.

Gracias a todos.

Índice general

Prólogo

El propósito de este trabajo es ahondar en una de la técnicas del análisis multivariado denominado *Análisis de conglomerados* o también llamado *Clasificación no supervisada*, el cual tiene como principal propósito el agrupar objetos (personas, variables, productos, etc.) dado un cierto grupo de características que distingan a los elementos de la población que nos es de interés.

Es también objetivo del presente trabajo el introducir una técnica propuesta para la formación y evaluación del número de conglomerados obtenido mediante modelos basados en mezclas, y así darle otros elementos al investigador o persona que utilice la clasificación no supervisada para poder contestar preguntas del estilo *¿cuántos grupos deberán de ser?* o *¿será que sí son 4 grupos y no 5?* Esta técnica lleva implícita muchos conceptos que se irán desarrollando conforme al trabajo. Conceptos de probabilidad como lo son los el de función de distribución y función de densidad. Conocimientos de estadística tales como función de verosimilitud y modelos de mezclas o los conceptos de aproximación numérica mediante algoritmos de optimización de la verosimilitud del mejor modelo propuesto. En los capítulos siguientes se dará una explicación de los conceptos que se requerirán para la construcción de esta técnica que puede ser utilizada como un criterio de decisión al momento de formar conglomerados, dicha técnica propone un algoritmo para reducir el árbol generado por una jerarquía de un modelo basado en mezclas.

También se tiene el propósito de proporcionar al lector o estudiante de las carreras de actuaría, matemáticas o alguna carrera afín, una pequeña introducción o referencia bibliográfica para realizar consultas ya sea para realizar una aplicación de lo visto en él o como un pequeño libro de texto que le permita al lector poder aprender un poco sobre algunas técnicas del análisis multivariado que se presentan en el desarrollo del trabajo.

Este trabajo ha sido pensado para llevar una secuencia que permita poder comprender bien todos los elementos que lo constituyen. En el primer capítulo se tocarán algunos temas introductorios, definiciones y conceptos necesarios para poder comprender y saber de que trata el análisis de conglomerados,

sus ventajas y desventajas, sus limitaciones, controversias, etc. En el segundo capítulo se ilustrará la metodología del análisis de conglomerados así como cuestiones que se deben tomar en cuenta para llevar acabo un buen análisis de los datos de interés y se plantearán preguntas intentando dar soluciones. En el tercer capítulo se darán algunas de las herramientas más utilizadas dentro del análisis de conglomerados, modelos y algoritmos a los cuales se recurre para la formación de conglomerados dentro de un conjunto de datos. En el cuarto capítulo se darán la teoría y fundamentos en los que se basan la técnica y el algoritmo propuestos en este trabajo así como algunas aplicaciones y ejemplos en la práctica. En el quinto capítulo se llevará acabo la aplicación de los métodos comúnmente utilizados en el análisis de conglomerados a un conjunto de datos simulados y en este a su vez, se aplicará la herramienta propuesta en este trabajo para hacer comparaciones entre los métodos usuales y lo que esta herramienta nos puede proporcionar. Ya en sexto y último capítulo se llegará a las conclusiones respectivas de esta técnica y algoritmo aquí propuesto sobre las aplicaciones con un datos reales en donde se podrá concluir si es o no de gran utilidad esta técnica recientemente propuesta.

Este trabajo fue realizado con apoyo del software estadístico **R** y del programa LaTeX y toda la sintaxis utilizada en el desarrollo de este trabajo será expuesta en el apéndice para su consulta. Los software **R** y LaTeX son software libre y pueden ser obtenidos en los sitios de Internet:

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

Capítulo 1

Introducción al análisis de conglomerados

Para comenzar con el desarrollo de este trabajo, primero tenemos que definir todos los conceptos que se han de utilizar a lo largo del presente. Tal es el caso de los conceptos de grupo o conglomerado y similaridad.

Será de suma importancia plantearse preguntas como: ¿Qué es un grupo? ¿Qué es similaridad?¿Existe algún método para poder construir grupos de individuos u objetos de interés? Hoy en día ¿En qué contextos es importante agrupar datos u observaciones obtenidas mediante técnicas estadísticas?. Estas son las preguntas clave que serán discutidas y analizadas para poder proponer definiciones, soluciones y aplicaciones para así lograr resultados en base al óptimo uso y análisis de la información.

1.1. El análisis multivariado una herramienta de investigación

Muchas veces los académicos e investigadores se encuentran con situaciones cuya mejor forma de resolverlas es definiendo grupos de objetos homogéneos, tanto si son personas como productos, variables, comportamientos e incluso empresas. La pregunta que nos es de suma importancia en estos momentos es, ¿Cómo definiremos al análisis de conglomerados? Comenzamos respondiendo esta pregunta con la siguiente definición:

Definición 1.- El *Análisis de Conglomerados* es la herramienta del análisis multivariado que tiene por objeto el agrupar elementos de una cierta población o conjunto de datos en función de sus características, sus similaridades y disimilaridades.

1.1. El análisis multivariado una herramienta de investigación

Una vez definido lo que es el análisis de conglomerados, es importante centrarnos en lo que se fundamenta esta herramienta estadística, es decir en ¿Cómo podemos definir a un grupo o conglomerado de objetos? Una idea intuitiva es la que dice que un grupo es un conjunto de personas u objetos que son *parecidos* entre ellos o que tienen algo *similar* o en *común* que los caracteriza y/o los diferencia de otros individuos. Pues bien, podemos darle un poco mas de solidez a esta idea un poco vaga dando la siguiente definición de un grupo.

Definición 2.- Un o *conglomerado* es un conjunto de individuos u objetos de tal forma que cada objeto sea muy parecido a los que hay en el conglomerado y, además tienen que ser lo menos parecidos a los objetos de otros grupos con respecto a un criterio predeterminado.

Algunos ejemplos claros de la definición que hemos dado de *grupo* puede ser el conjunto de personas que prefieren cierto producto de características específicas, o también en las ciencias naturales la necesidad de la creación de una taxonomía biológica para la clasificación de varios grupos de animales o insectos. O en las ciencias sociales como el análisis de varios tipos de perfiles psiquiátricos.

Otra definición que será de mucha utilidad para comprender un poco más ¿en qué se fija este método para realizar la clasificación?, es la definición de similaridad entre los elementos de un grupo.

Definición 3.- La *similaridad* entre objetos es una medida de asociación, parecido o de correspondencia, entre objetos que van a ser agrupados.

La similaridad ente objetos puede medirse de varias formas, pero hay tres principalmente que dominan las aplicaciones del análisis de conglomerados. Estas formas son: medidas de correlación, de distancia y de asociación. Cada uno de los métodos representa una perspectiva particular de similitud, dependiendo tanto de sus objetivos como del tipo de datos. Tanto las medidas de distancia como las de correlación exigen datos que provengan de observaciones de variables continuas, mientras que las medidas de asociación son para datos que provengan de variables que no son continuas [1].

[1] En el capitulo 2 se darán la definición y características de lo que es una métrica y sus aplicaciones en el análisis de conglomerados.

1.2. ¿Cuántos grupos deben formarse?

Quizá el reto más importante para el investigador que utiliza el análisis de conglomerados es la determinación del número final de grupos a formar en cierto conjunto de datos de una población. Desafortunadamente no existe un procedimiento único o estándar, dado que no se utiliza un criterio estadístico interno para la inferencia, tal como una prueba de hipótesis. Durante estos últimos años se han desarrollado varios criterios y líneas a seguir para aproximarse a la solución del problema. La principal conclusión es que existen procedimientos que deben ser establecidos y calculados por el propio investigador, lo que muchas veces implica procedimientos complejos. Una clase de reglas o criterios para poder establecer el número final de conglomerados que es relativamente simple y muy utilizada, es proponer alguna medida de similaridad o distancia entre los conglomerados a cada paso sucesivo, donde la solución final se define cuando la medida de similaridad excede un valor dado o cuando los resultados sucesivos entre los pasos a seguir en algún algoritmo son muy influyentes.

También, el investigador o la persona que realiza el análisis de conglomerados debería complementar el juicio estrictamente empírico con cualquier conceptualización de las relaciones teóricas que pueda sugerir un número natural de conglomerados. Se puede empezar este proceso especificando algún criterio basándose en consideraciones prácticas, como el decir *el resultado esperado es lógico y fácil de comunicar si se tiene entre tres y seis conglomerados*, y a continuación resolver para este número de conglomerados y seleccionar la mejor alternativa después de evaluar todas ellas. En el análisis final, sin embargo, probablemente sea mejor calcular varias soluciones diferentes, es decir, probar con dos, tres o más soluciones diferentes y después decidir entre las soluciones alternativas utilizando criterios a priori, juicios prácticos, sentido común o fundamentos teóricos. Las soluciones se verán mejoradas mediante la restricción de la solución de acuerdo con los aspectos conceptuales del problema.

Cuando se identifica una solución aceptable, el investigador debería examinar la estructura fundamental representada en los conglomerados definidos. Es de particular interés un tamaño bastante grande en los conglomerados definidos o conglomerados que solo tengan uno o dos observaciones. Los investigadores deben examinar los tamaños de los conglomerados muy variables desde una perspectiva conceptual, comparando los resultados actuales con las expectativas formadas en los objetivos de investigación. Resultan mucho más problemáticos los conglomerados de un único miembro, que pueden ser atípicos no detectados en análisis anteriores. Si aparece un conglomerado con una única observación o uno de un tamaño muy pequeño en comparación con

los demás conglomerados, el investigador debe decidir si representa un componente estructural válido de la muestra o si debería ser eliminado por ser un dato atípico. Si se elimina cualquier observación, concretamente cuando se emplean modelos jerárquicos[2], el investigador debería repetir el análisis y empezar de nuevo el proceso de formación de conglomerados.

Y con esta pequeña introducción de lo que es el análisis de conglomerados, los problemas que aborda y las diferentes soluciones finales que puede darnos un análisis de este tipo sobre un conjunto de datos de interés, podemos ahora dar una pequeña reseña histórica de las aplicaciones del análisis de conglomerados y su amplio uso en muchas disciplinas que no solo involucran las ciencias exactas, sino las ciencias bioquímicas, de la salud y sociales.

1.3. Antecedentes históricos

Algunos de los antecedentes históricos sobre el análisis de conglomerados que pueden darnos muchas más ideas sobre el surgimiento y las extensas aplicaciones de esta rama de la estadística a diversos y variados problemas son:

- En la taxonomía de los animales y las plantas, los conglomerados datan desde Aristóteles siendo el modelo moderno esencialmente de Carlos Lineo (1753), cada Especie pertenece a grupos que incrementan en tamaño y decrecen en el número de características comunes.

- Las enfermedades del cuerpo no son tan elusivas como las enfermedades de la mente, por lo tanto en Psiquiatría hay un acuerdo en la existencia de la paranoia, la esquizofrenia y depresión dichas categorías pueden ser vistas en la clasificación de Kant publicado en 1970. La dificultad de clasificación para las características de una enfermedad mental es subjetiva, sutil y de carácter variable dependiendo de los síntomas.

- La conglomeración en el campo de la Antropología y la 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 cuál se encarga de la distribución espacial de las distintas especies de plantas y animales, sustenta

[2]En el capítulo 3 se dará toda la teoría sobre los diversos métodos de clasificación no supervisada.

la misma relación de la taxonomía que la epidemiología en cuanto a 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 hábitat.

- También en el campo de la Economía, Fisher (1969) considera una matriz de salidas-entradas en el cual las filas y las columnas tienen las mismas etiquetas para que el conglomerado de filas y columnas ocurran simultáneamente, Goronzy (1970) conglomera características tanto operativas como financieras, en el campo de la Investigación de Mercados, mientras que King (1966) lo hace manteniendo una reserva en el inventario de acuerdo al comportamiento del precio.

- En Lingüística, Dyen (1967) usa 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; Abell (1960) encuentra grupos de galaxias a través de la búsqueda de placas fotográficas en las más altas latitudes galácticas.

Y se podría dar una lista mucho más amplia sobre la historia y las aplicaciones del análisis de conglomerados, pero no es el objetivo de este trabajo. Ahora enlistaremos y haremos algunas comparaciones entre las ventajas y desventajas que están presentes en el análisis de conglomerados.

1.4. Inconvenientes del análisis de conglomerados

Junto con los beneficios y aplicaciones que tiene el análisis de conglomerados en una extensa rama del conocimiento humano, existen algunos inconvenientes. El análisis de conglomerados puede caracterizarse como descriptivo, ateórico y no inferencial. Este análisis no tiene bases estadísticas sobre las cuales deducir inferencias estadísticas para una población a partir de una muestra, y tiene una aplicación totalmente exploratoria de los datos.

Otro inconveniente de este análisis es que frecuentemente las soluciones no son únicas. En diversos paquetes estadísticos que son ampliamente utilizados en la actualidad como R, S-plus, SPSS, Minitab, etc. Tendremos soluciones diferentes con los mismos datos, las soluciones dependen de muchos elementos del procedimiento y se pueden obtener varias soluciones diferentes variando uno o más de estos elementos.

Además, el análisis de conglomerados siempre creará conglomerados a pesar de como estén estructurados los datos. Finalmente, la solución es to-

talmente dependiente de las variables utilizadas como base para la medida de *similaridad*. La incorporación u omisión de variables relevantes puede tener un impacto substancial sobre el resultado final del análisis. Por lo tanto, el investigador debe tener particular cuidado en llevar una evaluación del impacto de cada decisión implicada en el desarrollo o aplicación del análisis de conglomerados para no tener resultados erróneos o que no sean los esperados por el investigador. Es muy importante siempre tener presente este impacto para así obtener los resultados óptimos derivados de este análisis.

Como pudimos apreciar en los párrafos anteriores, el análisis multivariado tiene aplicaciones extensas que no solo abordan problemas de tipo estadístico o de un entorno puramente científico. También puede ser llevado a un entorno laboral donde se tienen situaciones que la vida laboral exige que sean estudiadas y respondidas. Pero a la vez, también tiene algunos inconvenientes que deben ser tomados muy en cuenta y siempre se debe estar consciente de que los resultados obtenidos mediante este análisis no siempre son los "*mejores*", y que siempre habrá algo de subjetividad implícita para la toma de decisiones en el análisis.

Con esta breve introducción daremos paso al siguiente capítulo donde se revisara de una forma rápida los conceptos básicos y fundamentales de la estadística inferencial clásica univariada y multivariada así como del álgebra lineal que serán de utilidad en la teoría y desarrollo del trabajo para después en capítulos posteriores dar una breve descripción de los diferentes algoritmos y métodos para realizar clasificación no supervisada .

1.5. Contribución del trabajo

La contribución principal de este trabajo es el de presentar herramientas de diagnóstico para la formación y evaluación del número de grupos basados en el grado de separación entre los componentes de una mezcla de distribuciones. Tales herramientas serán la matriz de clasificación errónea, medidas para evaluar el error que se puede cometer al clasificar elementos de la población en un grupo equivocado entre otras. También se propone un algoritmo de *podado* o *ajuste* de un árbol generado por el modelo jerárquico basado en mezclas. El algoritmo comienza con el árbol que corresponde al modelo basado en mezclas elegido mediante el criterio de información bayesiana (BIC)[3]. Entonces se combinan progresivamente componentes de la mezcla y se fusionan en un conglomerado que corresponde a una posible mejor clasificación de los datos. Es decir, cada conglomerado en la partición final obtenida me-

[3]En los capítulos posteriores se darán los conceptos necesarios para fundamentar bien el algoritmo.

diante un modelo de mezclas puede, por lo tanto, ser modelado por más de un componente de una mezcla de distribuciones. El procedimiento resultante puede ser considerado como un híbrido entre los métodos basados en mezclas y métodos no paramétricos para formar conglomerados.

Capítulo 2

Elementos de Inferencia Clásica

En este capítulo vamos a introducir, retomar y definir varios conceptos que serán de gran utilidad para el desarrollo teórico de este trabajo. Con ello se intenta dar las bases matemáticas para que se pueda formalizar y dar una explicación a un problema o fenómeno real en estudio. Empezaremos por retomar algunos conceptos de inferencia clásica y que son de gran importancia para sustentar los argumentos que se emplearán para pasos posteriores.

2.1. Conceptos previos

Comenzamos definiendo los conceptos de variable aleatoria y función de distribución que son muy importantes en probabilidad y son objetos de estudio y análisis en estadística.

Definición 4.- Sea $(\Omega, \psi, \Im)$ un espacio de probabilidad[1]. Se dice que una función $X : \Omega \longmapsto \mathbb{R}$ es una variable aleatoria real si $[X \leq x] \in \Im \ \forall \ x \in \mathbb{R}$.

Definición 5.- Sea X una variable aleatoria real, a la función $F : \mathbb{R} \longmapsto \mathbb{R}$, definida por $F(x) = P[X \leq x]$ se le llama función de distribución[2] de X

Como se demuestra en los cursos de probabilidad, las funciones de distribución cumplen con tres propiedades, estas son:

1. F Es una función monótona no decreciente y continua por la derecha.

2. $\lim_{x \mapsto \infty} F(x) = 1$.

[1]Se entiende por Ω un espacio muestral, ψ una medida de probabilidad y $\Im$ una σ-álgebra de Ω.

[2]También llamada función de distribución acumulada.

3. $\lim_{x \mapsto -\infty} F(x) = 0$.

La importancia de la función de distribución es que ésta tiene toda la información probabilística relativa a una variable aleatoria X, disponiendo de ella, se puede obtener la probabilidad de cualquier evento cuya ocurrencia o no ocurrencia dependa del valor que tome X. Dos variables aleatorias pueden ser distintas, vistas como funciones definidas sobre el espacio muestral Ω, pero pueden ser idénticas en cuanto a su distribución y entonces desde el punto de vista probabilístico, dan exactamente la misma información y pueden ser utilizadas indistintamente para el mismo propósito.

Para el caso multivariado, definimos una variable aleatoria como la función $X : \Omega \longmapsto \mathbb{R}^n$ con $n > 1$. En el caso de una familia de n variables aleatorias, el papel de central de las funciones de distribución cuando se trata de una sola variable aleatoria, no lo tiene la colección de las n funciones de distribución correspondientes, sino que se le llama la función de distribución conjunta, la cual definimos a continuación.

Definición 6.- Sean $X_1, X_2, \ldots, X_n$ n variables aleatorias. La función $F : \mathbb{R}^n \longmapsto [0, 1]$, definida por:

$$F(x_1, x_2, \ldots, x_n) = P[X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n]$$

es llamada la función de distribución conjunta de $X_1, X_2, \ldots, X_n$

Existen dos tipos de distribuciones de variables aleatorias, las discretas (toma valores en un conjunto de resultados a lo mas numerable) y las continuas (toman valores en conjuntos infinitos no numerables por ejemplo $\mathbb{R}$). Es en estas últimas es en las que nos enfocaremos en el desarrollo del trabajo ya que una distribución comúnmente utilizada es la distribución normal multivariada. Definimos una distribución continua como sigue.

Definición 7.- Se dice que una función de distribución F de la variable aleatoria X es continua si existe una función no negativa $f : \mathbb{R} \longmapsto \mathbb{R}$, integrable tal que:

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

En este caso se dice también que la variable aleatoria X es absolutamente continua y la función f es llamada una función de densidad de X.

Este otro concepto de función de densidad es fundamental en probabilidad y que es en el que se basan muchos de los resultados en estadística y daremos algunas de sus propiedades básicas.

Definición 8.- Sea X una variable aleatoria continua y f su función de densidad, entonces:

1. $f(x) \geq 0 \quad \forall\, x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

Para el caso de variables aleatorias multivariadas tendremos algo similar, definiremos una función de distribución conjunta continua como sigue:

Definición 9.- Se dice que la función de distribución conjunta F de las variables aleatorias $X_1, \ldots, X_n$ es absolutamente continua si existe una función $f : \mathbb{R}^n \longmapsto \mathbb{R}$ integrable tal que:

$$F(X_1, \ldots, X_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f(y_1, \ldots, y_n)\; dy_1 \ldots dy_n \quad \forall\, x \in \mathbb{R}$$

En este caso se dice también que las variables aleatorias $X_1, \ldots, X_n$ forman un vector aleatorio absolutamente continuo y la función f es llamada la función de densidad conjunta de $X_1, \ldots, X_n$.

También para el caso multivariado la función de densidad conjunta f cumple las mismas propiedades que las funciones de densidad de una sola variable aleatoria, es decir, es una función no negativa y su integral sobre todo el espacio muestral es 1.

2.2. Inferencia

De aquí en adelante se entenderá que la expresión $f(\mathbf{x}; \theta)$ representa una función de densidad arbitraria que depende de una variable aleatoria $\mathbf{x}$ y de un cierto vector de parámetros desconocidos θ. Este vector describe en su totalidad a la función de densidad de $\mathbf{x}$. La expresión $f(\mathbf{x}; \theta)$ puede ser cualquiera de las distribuciones conocidas como la distribución bernoulli, binomial, exponencial, normal, normal multivariada, gamma, etc. Denotaremos cuando una variable aleatoria $\mathbf{x}$ se *Distribuye* de cierta forma como $\mathbf{x} \sim exp(\lambda)$. Es decir, la variable aleatoria $\mathbf{x}$ tiene una distribución exponencial con parámetro λ. Ahora definiremos el concepto de parámetro que es de gran importancia en la estadística.

Definición 10.- Un *Parámetro* es una cantidad que se asume *fija* pero *desconocida*. Esta cantidad es la que *identifica* la ley o distribución usada para describir un fenómeno aleatorio.

Definición 11.- 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 12.- Si $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ es un conjunto de v.a's (*variables aleatorias*), se dice que es una *Muestra Aleatoria* si se satisfacen las condiciones:

1. $\mathbf{x}_i \sim f(\mathbf{x}; \theta) \quad \forall i = 1, 2, \ldots, n$
2. $f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} f(\mathbf{x}_i; \theta)$

Es decir, una *muestra aleatoria* es un conjunto de v.a's que son independientes (su función de densidad conjunta puede ser expresada como el producto de las funciones marginales de cada v.a.) y que cada una de ellas tiene la misma distribución con los mismos parámetros. También indicaremos a la muestra aleatoria como $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ y al vector de observaciones dadas de cada variable que conforman la muestra aleatoria lo expresamos por $x = (x_1, x_2, \ldots, x_n)$.

Definición 13.- Al conjunto de valores que puede tomar la muestra aleatoria se le llama *Espacio muestral* y se denota por χ.

Definición 14.- Sea $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ una muestra aleatoria de $f(\mathbf{x}_i; \theta)$. Una *Estadística* es cualquier función de la muestra aleatoria que no dependa de parámetros desconocidos.

Definición 15.- Sea $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ una muestra aleatoria de $f(\mathbf{x}_i; \theta)$. Un *Estimador* es una estadística $T(X)$ cuyos valores $t(x)$ sirven para aproximar los valores del(os) parámetro(s) desconocido(s) θ. Y denotaremos a el estimador del parámetro θ como $\widehat{\theta} = T(X)$.

Definición 16.- A los valores del estimador, es decir de $t(x)$ se les llama *Estimado* o *Estimada*.

Para que estos conceptos de estimador y estimado sean mas claros supongamos por ejemplo que se tiene una población normal $N(\mu, \sigma^2)$. Un posible estimador para μ es

$$\widehat{\mu} = \bar{X} = \sum_{i=1}^{n} \frac{\mathbf{x}_1 + \mathbf{x}_2 + \ldots + \mathbf{x}_n}{n}$$

Y un estimado sería un valor particular de $\bar{X}$ dada una muestra de tamaño n y lo denotamos como $\bar{x}$.

2.2.1. Estimación por Máxima Verosimilitud

Ahora se describirá uno de los métodos para encontrar estimadores de parámetros, este método es el método denominado de máxima verosimilitud [3]. Antes de enunciar el método de máxima verosimilitud, es preciso recordar el concepto fundamental que es la idea principal de este método, este concepto es el de la *Función de Verosimilitud* o simplemente llamada *Verosimilitud*

Definición 17.- Sea $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ una muestra aleatoria de $f(\mathbf{x}_i; \theta)$. Se define la *Función de Verosimilitud* como:

$$L(\theta) = f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} f(\mathbf{x}_i; \theta_i) \tag{2.1}$$

Nótese que ambas funciones, la función de densidad y la verosimilitud son iguales, pero la diferencia entre la función conjunta de densidad de la muestra X y la función de verosimilitud radica en que la función de densidad conjunta determina totalmente el comportamiento probabilístico de la muestra X y podemos realizar cálculos de probabilidades sobre ésta. Pero en el caso en que θ es desconocido, es el caso que nos interesa ya que la función de densidad conjunta se convierte en la función de verosimilitud y realizamos la estimación del parámetro θ considerando ahora a la verosimilitud como una función que depende ahora de θ. Es entonces que podemos definir al estimador máximo verosímil para el parámetro de interés θ como

Definición 18.- Sea $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ una muestra aleatoria de $f(\mathbf{x}_i; \theta)$ y $L(\theta)$ su respectiva función de verosimilitud. $\widehat{\theta} = T(X)$ será el estimador máximo verosímil si $\widehat{\theta}$ satisface que:

$$L(\widehat{\theta}) \ \geq \ L(\theta) \quad \forall \ \theta \ \in \Theta$$

El estimador máximo verosímil, es el valor de θ que maximiza la probabilidad de aparición de los valores observados de la muestra y se obtiene maximizando la función $L(\theta)$. Ahora enunciaremos el método de máxima verosimilitud.

Método de Máxima Verosimilitud

Sea $f(X; \theta_1, \theta_2, \ldots, \theta_k)$ una función de densidad con k parámetros que es diferenciable y que su máximo no se encuentra en un extremo de su dominio

[3] Este no es le único método de estimación de parámetros, otro método es el de momentos propuesto por Pearson.

de definición, su máximo lo alcanzará en el punto que satisface el sistema:

$$\frac{\partial L(\theta_1, \theta_2, \ldots, \theta_n)}{\partial \theta_i} = 0 \quad \text{para} \quad i = 1, 2, ..., k$$

Y el punto $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_k)$ corresponderá a un máximo si la matriz hessiana de segundas derivadas $\mathbf{H}$, evaluada en $\widehat{\theta}$ es definida negativa[4], entonces $\widehat{\theta}$ es el estimador máximo verosímil y lo denotaremos como $\widehat{\theta}_{MV}$.

Como por lo regular es difícil trabajar con la función de verosimilitud $L(\theta_1, \theta_2, ..., \theta_n)$ ya que por su complejidad es difícil de realizar operaciones como derivar o integrar. Para ello es conveniente utilizar la función $\log(L(\theta))$ ya que es una transformación que no afecta los resultados que se obtienen al momento de realizar la maximización de la función de verosimilitud, ya que ambas funciones alcanzan su máximo en el mismo punto crítico.

Definición 19.- A la función $l(\theta_1, \theta_2, ..., \theta_n)$ dada por

$$l(\theta_1, \theta_2, ..., \theta_n) = \log(L(\theta_1, \theta_2, ..., \theta_n)) \tag{2.2}$$

Se le define como la función *log-verosímil* o *log-verosimilitud*

El trabajar con la función de log-verosimilitud tiene tres ventajas principales:

1. Pasamos del producto de funciones de densidad a la suma de sus logaritmos y la expresión resultante suele ser más simple que la verosimilitud.

2. Al tomar logaritmos, las constantes multiplicativas de la función de densidad conjunta, que son irrelevantes para el máximo, se hacen aditivas y desaparecen al derivar.

3. Si multiplicamos la función $l(\theta)$ por -2 se obtiene un resultado asintótico importante en estadística que proporciona un método general para juzgar el ajuste de un modelo a los datos. Esta expresión es

   $$\mathbf{D} = -2l(\theta)$$

   A esta expresión se le llama *devianza* y mide la discrepancia entre los datos y el modelo, es decir, entre más grande sea $l(\theta)$, mayor será la concordancia entre el valor del parámetro y los datos y menor será la devianza.

Damos por terminada esta sección de inferencia univariada para dar paso a las definiciones y conceptos de la inferencia multivariada y seguir con el desarrollo teórico del trabajo.

[4]Véase J. Marsden Cálculo Vectorial.

2.3. Inferencia Multivariada

Ahora mencionaremos algunos de los conceptos principales de la inferencia multivariada como el de matriz de datos, vector de medias y la matriz de varianzas y covarianzas entre otros. Todos éstos son el sustento de la teoría que contiene el análisis multivariado y que sirven para dar paso a pruebas de hipótesis o la introducción de distribuciones multivariadas que son de gran utilidad para realizar la inferencia que se requiere para poder analizar un grupo de datos multivariados.

Debemos enfatizar que al igual que en la estadística univariada, se parte de una población en estudio que tiene características de interés y que queremos medir, para que de una pequeña muestra podamos hacer inferencia sobre toda la población. Entonces supondremos que cada una de estas características que posee la población es una variable.

La información de partida para la inferencia multivariada puede ser de varios tipos. La que utilizaremos es una tabla donde aparecerán los valores de p variables observadas en n individuos o elementos y con valores observados provenientes de variables continuas.

2.3.1. La Matriz de Datos

Supongamos que se ha observado p variables numéricas en un conjunto de n elementos o individuos. Cada una de estas p variables se denomina una variable *escalar* o *univariada* y al conjunto de las p variables forman una variable *vectorial* o *multivariada*. Entonces damos paso a definir a la matriz de datos

Definición 20.- Sea $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ una muestra aleatoria cuyas funciones de distribución $f(\mathbf{x}_i; \theta)$ son multivariadas. Entonces, los valores de las p variables escalares en cada uno de los n elementos pueden representarse en una matriz $\mathbf{X}$ de dimensiones $(n \times p)$ que llamaremos *Matriz de Datos*, y denotaremos por x_{ij} al elemento de esta matriz que representa el valor de la variable j sobre la observación o individuo i. Es decir

$$\mathbf{X} = x_{ij} \quad \text{donde} \;\; i = 1, \ldots, n \quad \text{y} \;\; j = 1, \ldots, p$$

La matriz de datos $\mathbf{X}$ puede ser representada de dos formas diferentes. por filas:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} \tag{2.3}$$

Donde cada variable $\mathbf{x}_i'$ es un vector fila de dimensión p x 1 que representa los valores de las p variables sobre el i-ésimo individuo. También podemos expresar de una manera alternativa esta matriz de datos $\mathbf{X}$ por columnas:

$$\mathbf{X} = [\mathbf{x}_{(1)} \ \dots \ \mathbf{x}_{(p)}]$$

Donde ahora cada variables $\mathbf{x}_{(j)}$ es un vector columna de dimensión n x 1 que representa la variables escalar x_j medida en los n elementos de la población. Llamaremos $\mathbf{x} = (x_1, \dots, x_p)'$ a la variable multivariada formada por las p variables escalares que toman valores particulares $\mathbf{x}_1, \dots, \mathbf{x}_n$, en los n valores observados.

2.4. Medidas de tendencia central y dispersión multivariada

A continuación daremos paso al análisis multivariado de los datos. Presentaremos como obtener medidas conjuntas de tendencia central y de dispersión para el conjunto de da variables y medidas de tendencia lineal entre pares de variables.

2.4.1. El vector de medias muestrales

La medida de tendencia central más utilizada para describir datos multivariados es el *vector de medias muestrales*, que es un vector de dimensión p cuyas entradas son las medias de cada una de las p variables. Puede calcularse de una manera similar al caso univariado donde se promedia cada una de las variables y se asigna a cada entrada correspondiente la media de cada variable.

Definición 21.- Sea $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ una muestra aleatoria, entonces se define el vector de medias $\mathbf{x}$ como un vector en $\mathbb{R}^p$ tal que

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \begin{bmatrix} \bar{\mathbf{x}}_1 \\ \bar{\mathbf{x}}_2 \\ \vdots \\ \bar{\mathbf{x}}_p \end{bmatrix} \tag{2.4}$$

Donde cada $\bar{\mathbf{x}}_i$ es la media aritmética de la i-ésima variable con $i = 1, \ldots, p$

Observemos que al igual que en el caso univariado, el vector de medias muestrales se encuentra en el centro de los datos en el sentido de hacer cero la suma de las desviaciones:

$$\sum_{1=i}^{n}(\mathbf{x}_i - \bar{\mathbf{x}}) = 0$$

Ya que escribiendo la suma de la forma $\sum_{i=1}^{n}\bar{\mathbf{x}} - n\bar{\mathbf{x}}$ y aplicando la definición del vector de medias, es evidente que esta suma se vuelve cero.

Las medidas de tendencia central univariadas basadas en el orden de las observaciones, no se pueden generalizar fácilmente al caso multivariado ya que por resultados de álgebra lineal, se sabe que el espacio vectorial $\mathbb{R}^n$ no tiene orden alguno, lo cual dificulta el hecho de intentar generalizar estas medidas. Por ejemplo, podemos calcular el vector de medianas, pero este no es un punto que necesariamente esta ubicado en el centro de los datos.

Antes de seguir con las definiciones y resultados del análisis multivariado debemos hacer énfasis en que la población tiene sus parámetros poblacionales que son de nuestro interés y objeto de estudio, los cuales se estimarán por algún método y obtener así un estimador para realizar inferencias sobre la población a partir de una muestra. Uno de estos parámetros poblacionales es el vector de valores esperados que consiste en un vector de dimensión px1 el cual tiene como entradas las esperanzas de cada una de las p variables de interés de la población. A este vector lo denotaremos como:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} E(\mathbf{x}_1) \\ E(\mathbf{x}_2) \\ \vdots \\ E(\mathbf{x}_p) \end{bmatrix} \tag{2.5}$$

Una vez que se ha aclarada la diferencia entre el vector de medias muestrales y el vector de valores esperados poblacional, podemos entonces seguir con otro definición importante para el análisis multivariado que es la matriz de varianzas y covarianzas, una matriz de suma importancia porque nos podrá decir o reflejar en cierto grado la relación entre las variables de la población que estamos midiendo.

2.4.2. La matriz de varianzas y covarianzas

En muchos casos, es de suma importancia saber si existe alguna relación entre algún par de variables como el peso y la estatura, o el ingreso y gasto de un hogar. Esto se puede medir mediante una relación que contemple a un par de variables, esta relación puede describirse a través de la covarianza entre las variables x_i y x_j que la definimos como sigue

Definición 22.- Sean $\mathbf{x}_i$ y $\mathbf{x}_j$ dos variables aleatorias, entonces definimos la covarianza $Cov(\mathbf{x}_i, \mathbf{x}_j)$ entre las variables $\mathbf{x}_i$ y $\mathbf{x}_j$ como

$$Cov(\mathbf{x}_i, \mathbf{x}_j) = E((\mathbf{x}_i - E(\mathbf{x}_i))(\mathbf{x}_j - E(\mathbf{x}_j)) \tag{2.6}$$

En la estadística univariada es de suma importancia encontrar la varianza de alguna población en estudio para saber que tan homogénea es, en el caso multivariado se cuenta con la matriz de *varianzas y covarianzas*[5] poblacional, que es una matriz de *nxp* que la definimos de la siguiente forma

Definición 23.- Sea $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ una muestra aleatoria, se define la matriz de *varianzas y covarianzas* poblacional a la matriz $\Sigma = Var(\mathbf{X})$ de *nxp* tal que

$$\Sigma = \begin{bmatrix} Var(\mathbf{x}_1) & Cov(\mathbf{x}_1, \mathbf{x}_2) & \ldots & Cov(\mathbf{x}_1, \mathbf{x}_p) \\ Cov(\mathbf{x}_2, \mathbf{x}_1) & Var(\mathbf{x}_2) & \ldots & Cov(\mathbf{x}_2, \mathbf{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\mathbf{x}_p, \mathbf{x}_1) & Cov(\mathbf{x}_p, \mathbf{x}_2) & \ldots & Var(\mathbf{x}_p) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2p} \\ \vdots & & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_{pp}^2 \end{bmatrix}$$

$$\Sigma = Var(\mathbf{X}) \tag{2.7}$$

Las entradas de esta matriz son las varianzas de cada una de las variables que tiene la población σ_i^2 y las covarianzas de cada par de variables σ_{ij} con $i, j = 1, 2, \ldots, p$

Como se ha comentado, para las variables univariadas la dispersión respecto a la media se mide comúnmente con la varianza, o por su raíz cuadrada que se define como la desviación estándar. La relación lineal existente entre dos variables se mide por la covarianza, la covarianza muestral se define como:

Definición 24.- Sea $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ una muestra aleatoria, se define la *covarianza* muestral como

$$\mathbf{S}_{jk} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \tag{2.8}$$

[5]También llamada simplemente matriz de covarianzas

Y ésta mide el grado de dependencia lineal entre ambas variables.

Para una variable multivariada definimos la matriz de varianzas y covarianzas muestral como:

Definición 25.- Sea $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ una muestra aleatoria, se define la matriz de *varianzas y covarianzas* muestral como la matriz $\mathbf{S}$ de n x p tal que

$$\mathbf{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' \tag{2.9}$$

Que es una matriz simétrica por que contiene en su diagonal las varianzas muestrales de las variables y fuera de ella las covarianzas muestrales entre las variables. En efecto, si multiplicamos los vectores:

$$\begin{bmatrix} x_{i1}-\bar{x}_1 \\ \vdots \\ x_{ip}-\bar{x}_p \end{bmatrix} \begin{bmatrix} x_{i1}-\bar{x}_1 & \ldots & x_{ip}-\bar{x}_p \end{bmatrix} = \begin{bmatrix} (x_{i1}-\bar{x}_1)^2 & \ldots & (x_{i1}-\bar{x}_1)(x_{ip}-\bar{x}_p) \\ \vdots & \ddots & \vdots \\ (x_{i1}-\bar{x}_p)(x_{ip}-\bar{x}_p) & \ldots & (x_{ip}-\bar{x}_p)^2 \end{bmatrix}$$

Se obtiene la matriz de cuadrados y productos cruzados de las p variables en el i-ésimo elemento. El sumar para todos los elementos y dividir entre n se obtienen las varianzas en la diagonal y las covarianzas en fuera de ella. La matriz de varianzas y covarianzas, que por simplicidad se le llamara matriz de covarianzas, es la matriz simétrica de dimensión nxp con forma:

$$\mathbf{S}^2 = \begin{bmatrix} s_1^2 & \ldots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \ldots & s_p^2 \end{bmatrix} \tag{2.10}$$

2.4.3. La matriz de correlación

La dependencia lineal entre dos variables se puede estudiar mediante el *coeficiente de correlación lineal*. Este coeficiente para las dos variables aleatorias $\mathbf{x}_i$ y $\mathbf{x}_j$ se define como:

Definición 26.- Sean $\mathbf{x}_i$ y $\mathbf{x}_j$ dos variables aleatorias, definimos el coeficiente de correlación entre $\mathbf{x}_i$ y $\mathbf{x}_j$ como:

$$\rho_{\mathbf{x}_i,\mathbf{x}_j} = \frac{Cov(\mathbf{x}_i\mathbf{x}_j)}{\sqrt{Var(\mathbf{x}_i)Var(\mathbf{x}_j)}}$$

Entonces podemos definir el coeficiente de correlación muestral entre las variables $\mathbf{x}_i$ y $\mathbf{x}_j$ como:

Definición 27.- Sean $\mathbf{x}_i$ y $\mathbf{x}_j$ dos variables aleatorias, se define el coeficiente de correlación lineal muestral de como

$$r_{ij} = \frac{s_{ij}}{s_i s_j} = \frac{\frac{1}{n}\sum_{i,j=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_j)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}}_i)^2 \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}}_j)^2}}$$

Y tiene las propiedades siguientes:

1. $0 \leq |r_{ij}| \leq 1$.
2. Si $x_{ij} = a + bx_{ik}$, entonces $|r_{jk}| = 1$.
3. r_{ij} es invariante ante transformaciones lineales de las variables.

La dependencia por pares entre las variables se mide por la matriz de correlación que se define como sigue:

Definición 28.- Definimos la *matriz de correlación* $\mathbf{R}$ a la matriz cuadrada y simétrica que tiene unos en la diagonal principal y fuera de ella todos los coeficientes de correlación lineal entre los pares de variables y la denotamos como:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{p2} & \dots & 1 \end{bmatrix} \tag{2.11}$$

Esta matriz es semidefinida positiva ya que puede contener entradas fuera de la diagonal negativas o iguales a cero.

2.5. Distancia y Variabilidad

2.5.1. ¿Qué importancia tiene la medición de la distancia en el análisis de conglomerados?

Como ya se había mencionado en capítulos anteriores, usaremos el concepto de similaridad, que asociamos con el de distancia de una forma intuitiva, ya que una distancia viene a la mente para saber si algo se encuentra cerca o lejos un punto geográfico, o una población respecto a otra y si esta distancia influye en ciertas características.

Se discutirá en esta sección las principales medidas de similaridad entre los datos y por tanto, las más comunes de utilizar en el análisis de conglomerados según cita Hair en su libro[5] son:

- Medidas de correlación
- Medidas de distancia
- Medidas de asociación [6]

Medidas de correlación

La medida de similaridad entre dos objetos que probablemente viene a la mente, en primera instancia, es la del coeficiente de correlación entre un par de objetos medido sobre varias variables ya que tiene un amplio uso en estadística y el análisis multivariado. En efecto, en lugar de hacer la correlación entre dos conjuntos de variables, transponemos la matriz de datos $\mathbf{X}$ de tal forma que las columnas representan observaciones o individuos y los renglones representan variables. Por tanto, el coeficiente de correlación entre las dos columnas de números es la correlación (o similaridad) entre los perfiles de los dos objetos. Una cifra cercana a 1 en valor absoluto indica una similitud o parecido muy alto y una cifra cercana cero indica una falta de similitud.

Sin embargo, las medidas de correlación se utilizan rara vez por que el interés de la mayoría de las aplicaciones del análisis de conglomerados esta en los elementos o individuos de la población.

Medidas de distancia

Aunque se mencionó que las medidas de correlación es una idea intuitiva de asociación y se utilizan en varias técnicas del análisis multivariado, no son las medidas de similaridad más utilizadas en el análisis de conglomerados. Las medidas de similitud que están asociadas al concepto de **distancia**, que son las que representan la simililaridad como la proximidad de las observaciones son las más utilizadas. Las medidas de distancia son en realidad medidas de diferencia o variación donde los valores elevados indican una menor similaridad o parecido entre las observaciones. Es así como la distancia se convierte en una medida de similaridad utilizando una relación inversa.

Medidas de asociación

Las medidas de asociación se utilizan para comparar objetos cuyas características se miden a través de variables categóricas, es decir, son variables nominales u ordinales. Un ejemplo que podemos dar sería el caso en que un

[6] Estas ultimas son empleadas en datos categóricos a diferencia de las dos primeras utilizadas en datos de origen continuo.

grupo de encuestados responden *sí* o *no* a cierto número de preguntas. Una medida de asociación podría evaluar el grado de acuerdo o de acercamiento entre cada par de encuestados. La forma más simple de medida de asociación sería el porcentaje de veces que existió un acuerdo (ambos dicen *sí* o ambos dicen *no*) para el mismo conjunto de preguntas. Se han desarrollado extensiones de este simple coeficiente de ajuste para acomodar variables nominales de varias categorías o incluso medidas ordinales.

2.5.2. El concepto de distancia

Así pues, una vez dada una breve explicación sobre los diferentes tipos de medidas que existen para establecer una similaridad entre un conjunto de datos, el procedimiento para estudiar la dispersión de las observaciones que utilizaremos es el concepto de **distancia** entre dos puntos. En el caso univariado, la distancia entre el valor de la variable $\mathbf{x}$ en un punto x_i, y la media de la variable $\bar{\mathbf{x}}$, se mide de una manera natural con la distancia euclidiana $\sqrt{(x_i - \bar{x})^2}$, o lo que es equivalente, por el valor absoluto de su diferencia $|x_i - \bar{\mathbf{x}}|$. La varianza es un promedio de estas distancias al cuadrado entre los puntos y su media. Cuando se dispone de una variable multivariada, cada dato es un punto en $\mathbb{R}^p$, y podemos pensar en construir medidas de variabilidad o dispersión promediando las distancias entre cada punto y el vector de medias. Esto requiere de generalizar el concepto de distancia para cualquier espacio de dimensión arbitraria.

Definición 29.- Definimos la función **distancia** $d_{(x_i,x_j)}$ entre dos puntos si dados dos puntos cualesquiera $\mathbf{x}_i$ y $\mathbf{x}_j \in \mathbb{R}^p$ se cumplen las propiedades

1. $d : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}^+$, es decir, dados dos puntos en el espacio de dimensión p su distancia con esta función es no negativa. En otras palabras, podemos decir que $d_{(x_i,x_j)} \geq 0 \quad \forall x_i, x_j \in \mathbb{R}$.

2. $d_{(x_i,x_j)} = d_{(x_j,x_i)}$, es decir, la distancia es una función simétrica.

3. $d_{(x_i,x_j)} = 0 \;\Leftrightarrow\; i = j \; i, j = 1, 2, \ldots, p$, es decir, la distancia entre un punto y el mismo es cero.

4. $d_{(x_i,x_j)} \leq d_{(x_i,x_k)} + d_{(x_k,x_j)}$, es decir, la distancia entre los puntos x_i y x_j cumple la desigualdad del triángulo.

Estas propiedades generalizan la noción intuitiva de distancia entre dos puntos para espacios de dimensiones mayores a 3 y que son difícil de ver. Además estas propiedades nos serán de utilidad para posteriormente definir

distancias entre puntos de una población y así poder medir su grado de dispersión o para poder formar conglomerados en base a las distancias entre individuos del mismo grupo.

2.5.3. Ejemplos de distancias

Ya definida la distancia entre dos puntos en un espacio de dimensión p, podemos dar algunos ejemplos de distancias comúnmente usadas en el estudio de la dispersión o variabilidad de los puntos de nuestra muestra.

Distancia de Minkowski

Una familia de medidas de distancia muy habituales en $\mathbb{R}$ es la familia de métricas o distancias de **Minkowski**, que se define como:

$$d^{(r)}_{x_i,x_j} = (\sum_{k=1}^{p}(x_{ik} - x_{jk})^r)^{\frac{1}{r}}$$

Donde las potencias más utilizadas son $r = 2$ que sería el caso particular de la distancia euclidiana

$$d_{x_i,x_j} = (\sum_{k=1}^{p}(x_{ik} - x_{jk})^2)^{\frac{1}{2}}$$

Y para $r = 1$ tenemos la distancia

$$d_{(x_i,x_j)} = \sum_{k=1}^{p}|x_{ik} - x_{jk}|$$

Distancia euclidiana

La distancia mas utilizada es la euclidiana dada por:

$$d_{(x_i,x_j)} = \sum_{k=i}^{p}\sqrt{(x_{ik} - x_{jk})^2}$$

Pero tiene el inconveniente de depender de las unidades de medida de las variables. Una forma de evitar este problema es dividir cada variable por un termino que elimine el efecto de la escala. Esto conduce a la familia de métricas euclídeas ponderadas que se definen como:

Distancias euclidianas ponderadas

Definimos las distancias euclidianas ponderadas de la siguiente forma:

$$d_{(x_i,x_j)} = [(\mathbf{x}_i - \mathbf{x}_j)'\mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)]^{\frac{1}{2}}$$

Donde $\mathbf{M}$ es una matriz diagonal que se utiliza para estandarizar las variables y hacer la medida invariante ante cambios de escala. Por ejemplo, si ponemos sobre la diagonal de $\mathbf{M}$ las desviaciones estándar de las variables, obtenemos:

$$d_{(x_i,x_j)} = (\sum_{k=1}^{p}(\frac{x_{ik} - x_{jk}}{S_k})^2)^{\frac{1}{2}} = (\sum_{k=1}^{p} S_k^{-2}(x_{ik} - x_{jk})^2)$$

Que puede verse como una distancia euclidiana donde cada coordenada se pondera inversamente proporcional a la varianza. En general la matriz $\mathbf{M}$ puede no ser diagonal, pero siempre debe ser no singular y definida positiva para que $d_{(x_i,x_j)} \geq 0$. En el caso particular en que tomemos M=$\mathbf{I}$ se obtiene de nuevo la distancia euclidiana. Si tomamos $\mathbf{M} = \mathbf{S^{-1}}$ se obtiene la distancia de Mahalanobis que estudiaremos a continuación.

Distancia de Mahalanobis

Definición 30.- Se define la **distancia de Mahalanobis** entre un punto y su vector de medias por:

$$d_M = [(\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})]^{\frac{1}{2}}$$

Es frecuente referirse al valor d_M^2 también como la distancia de Mahalanobis, en lugar de referirnos como el cuadrado de la distancia. Desarrollaremos esta distancia y comprobaremos que es una medida muy razonable de distancia entre variables correlacionadas y por tanto de gran utilidad. Consideremos el caso cuando $p = 2$. Entonces, si escribimos $s_{12} = rs_1s_2$, tenemos que

$$S^{-1} = \begin{bmatrix} s_1^{-2} & -rs_1^{-1}s_2^{-1} \\ -rs_1^{-1}s_2^{-1} & s_2^{-2} \end{bmatrix}$$

Y la distancia de Mahalanobis (al cuadrado) entre dos puntos (x_1, y_1), (x_2, y_2) puede escribirse como:

$$d_M^2 = \frac{1}{1-r^2}\left[\frac{(x_1 - x_2)^2}{s_1^2} + \frac{(y_1 - y_2)^2}{s_2^2} - 2r\frac{(x_1 - x_2)(y_1 - y_2)}{s_1 s_2}\right]$$

De donde r es el coeficiente de correlación. Si $r = 0$, esta distancia se reduce a la euclídea estandarizando las variables por sus desviaciones estándar.

Cuando $r \neq 0$ la distancia de Mahalanobis añade un término adicional que es positivo (y, por tanto "separa" los puntos) cuando las diferencias entre las variables tienen le mismo signo, cuando $r > 0$, o distinto cuando $r < 0$.

Ahora en base a estas definiciones y ejemplos de distancias, definiremos la matriz de distancias que es muy importante en el análisis de conglomerados.

2.5.4. Matriz de distancias

La matriz de distancias es importante en el análisis de conglomerados ya que en muchas de las ocasiones el criterio de similaridad entre las observaciones va a estar dada por alguna distancia de las que acabamos de enunciar. A esta matriz la definimos como:

Definición 31.- La matriz de distancias como la matriz de nxn cuyos elementos representan las distancias d entre los puntos x_i y x_j de un conjunto de datos, es decir, es la matriz cuyas entradas son $x_{ij} = d_{x_i,x_j}$ y tiene la forma

$$\begin{pmatrix} 0 & d_{x_1,x_2} & \dots & d_{x_1,x_n} \\ d_{x_2,x_1} & 0 & \dots & d_{x_2,x_n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{x_n,x_1} & d_{x_n,x_2} & \dots & 0 \end{pmatrix}$$

Esta matriz tiene las siguientes propiedades:

- Es una matriz simétrica
- Su diagonal contiene ceros ya que $d_{x_i,x_i} = 0$
- Es semi-definida positiva ya que sus entradas son mayores o iguales a cero

Por simplicidad, sólo se tomará la diagonal ya sea inferior o superior de la matriz ya que es simétrica y así es como esta matriz es empleada en los algoritmos jerárquicos de formación de conglomerados como son el encadenamiento simple, encadenamiento completo, etc, que mencionaremos en el siguiente capítulo de este trabajo.

Para terminar con este capítulo, podemos mencionar algunas diferencias que existen entre las medidas de correlación y de distancia que hay para formar conglomerados. Las diferencias entre las medidas de distancia y de correlación serían que las medidas de distancia se centran en los valores obtenidos de las observaciones y representan casos similares que están juntos. La

elección de una medida de correlación en lugar de una medida de distancia requiere una interpretación muy diferente de los resultados por el investigador. Los conglomerados basados en medidas de correlación pueden no tener valores similares y en lugar tener patrones similares. Los conglomerados basados en la distancia tienen valores más parecidos para el conjunto de variables, pero los patrones pueden ser bastante diferentes.

En el siguiente capítulo se abordarán algunos métodos clásicos utilizados en el análisis de conglomerados. Se dará una pequeña explicación y algunas comparaciones entre estos métodos, así como ventajas y desventajas de cada uno.

Capítulo 3

Métodos de clasificación no supervisada

Son varios los métodos que son empleados en el análisis de conglomerados para formar grupos de individuos u objetos. En este capítulo se mencionarán dichos métodos según son presentados por los autores Hair [5] y Peña [9] de una manera breve, pero que sirva para saber ¿cómo es que se dividen y clasifican los diferentes algoritmos y métodos de clasificación? y ¿en qué consiste cada uno de estos?.

Comenzaremos dando la clasificación de estos métodos para saber cuáles son las diferencias que marcan a cada grupo de métodos. Tenemos que dentro de los métodos de clasificación existen varios tipos:

- Métodos jerárquicos: Estos métodos no asumen ningún modelo estadístico para los datos.
- Métodos no jerárquicos o de partición: Estos métodos asumen un modelo definido para los datos.
- Modelos basados en mezclas: Estos métodos asumen un modelo de probabilidad, buscan expresar grupos a través de una combinación lineal convexa de distribuciones de probabilidad.[1]

Existen otros tipos de métodos de clasificación no supervisada en la literatura de aprendizaje automatizado que no se abordaran en este trabajo, ya que se estaría desviando mucho del objetivo principal, además de que se emplearan los métodos clásicos para poder realizar una comparación con el

[1] En la mayoría de las aplicaciones reales y por simplicidad se asume que este conjunto de distribuciones esta formado por distribuciones normales multivariadas.

método que se desarrollará para saber si existe alguna diferencia o si se pueden mostrar resultados similares al utilizar el método propuesto que con los algoritmos clásicos.

Ahora bien, daremos una breve descripción de algunos de estos algoritmos y métodos de clasificación para saber ¿cuál es su desarrollo y metodología para llegar a la formación de conglomerados y el número resultante de grupos que se pueden distinguir?

3.1. Métodos Jerárquicos

Los métodos jerárquicos buscan estructurar los elementos de un conjunto de forma jerárquica dada su similaridad. Por ejemplo, tenemos una encuesta de atributos de distintas profesiones y lo que queremos es darle un orden a cada una de estas mediante su similaridad. Una clasificación jerárquica implica que los datos se ordenan en niveles, de tal manera que los niveles superiores contengan a los inferiores. Este tipo de clasificación es muy frecuente en la Biología, al clasificar animales, plantas, insectos. etc. Estrictamente, estos métodos no definen grupos, sino la estructura de asociación en cadena que pueda existir entre los elementos.

Entre los métodos jerárquicos existen dos tipos de procedimientos que son los algoritmos de **Aglomeración** y los algoritmos **Divisivos**.

En los métodos de aglomeración, cada objeto u observación empieza dentro de su propio conglomerado. En etapas posteriores, los dos conglomerados más cercanos se combinan en un nuevo conglomerado agregado, reduciendo así el número de conglomerados paso a paso. En otros, un tercer individuo se une a los dos primeros en un conglomerado. Eventualmente todos los individuos se agrupan en un único conglomerado; por esta razón, los procedimientos de aglomeración son denominados aveces como métodos de **construcción**. Mientras que en el caso de los métodos divisivos, éstos parten del conjunto de elementos y lo van dividiendo sucesivamente hasta llegar a los elementos individuales. Los algoritmos de aglomeración requieren de menor tiempo de cálculo y son los más utilizados.

Una característica importante de los métodos jerárquicos es que los resultados obtenidos en un paso previo siempre deben encajar dentro de los resultados del siguiente paso, creando algo parecido a un árbol. Por ejemplo, una solución de seis conglomerados se obtiene uniendo dos de los conglomerados encontrados en el paso en que se tienen 7 conglomerados. Dado que los conglomerados se forman solo por unión de los conglomerados existentes, se puede rastrear hasta su origen de un conglomerado por observación.

A la representación gráfica del algoritmo se le denomina **Dendograma** o **Árbol** el cual describiremos a continuación.

El dendograma

El **Dendograma** o **Árbol Jerárquico**, es una representación gráfica del resultado del proceso de agrupamiento en forma de árbol. Los criterios para definir distancias que hemos presentado tienen la propiedad de que, si se consideran tres grupos **A**, **B** y **C**, se verifica que:

$$d(\mathbf{A}, \mathbf{C}) \leq \text{máx}(d(\mathbf{A}, \mathbf{B}), d(\mathbf{B}, \mathbf{C}))$$

Y una medida de distancia que tiene esta propiedad se denomina *ultramétrica*. Esta propiedad es más fuerte que la propiedad triangular, ya que una ultramétrica es siempre una distancia. En efecto si $d(\mathbf{A}, \mathbf{C})$ es menor o igual que el máximo de $d(\mathbf{A}, \mathbf{B})$,$d(\mathbf{B}, \mathbf{C})$ forzosamente será menor o igual que la suma $d(\mathbf{A}, \mathbf{B}) + d(\mathbf{B}, \mathbf{C})$. El dendograma es la representación de una ultramétrica, y se construye como sigue:

1. En la parte inferior del gráfico se tienen los n elementos iniciales.
2. Las uniones entre elementos se indican por tres líneas rectas. Dos dirigidas a los elementos que se unen, y que son perpendiculares al eje de los elementos, y una paralela a este eje, que se sitúa al nivel en que se unen.
3. El proceso se repite hasta que todos los elementos están conectados por lineas rectas.

Si cortamos el dendograma a un nivel de distancia dado, obtenemos una clasificación del número de grupos existentes a ese nivel y los elementos que los forman. El dendograma es útil cuando los puntos tienen claramente una estructura jerárquica pero puede ser confuso cuando se interpreta mecánicamente.

Cuando el proceso de obtención de conglomerados procede en dirección opuesta al método de aglomeración, se dice que es un método divisivo. En estos métodos, se empieza con un gran conglomerado que contiene todas las observaciones (individuos u objetos). En los pasos sucesivos, las observaciones que son más diferentes se dividen y se construyen conglomerados más pequeños. Este proceso continua hasta que cada observación es un conglomerado en si mismo.

Son cinco los algoritmos más utilizados para obtener conglomerados que se basan en los métodos jerárquicos que son:

1. Encadenamiento simple o del vecino más cercano
2. Encadenamiento completo o del vecino más lejano
3. Encadenamiento medio que es un promedio de los dos anteriores
4. Método de Ward
5. Método del centroide

3.1.1. Encadenamiento Simple

El procedimiento del encadenamiento simple se basa en la distancia mínima. Encuentra los dos objetos separados por la distancia más corta y las coloca en el primer conglomerado. A continuación se encuentra la distancia más corta, o bien un tercer objeto se une a los dos primeros para formar un conglomerado o se forma un nuevo conglomerado de dos miembros. La distancia entre dos conglomerados cualesquiera es la distancia más pequeña desde cualquier punto en un conglomerado a cualquier punto en el otro. Dos conglomerados se fusionan en cualquier nivel por la distancia más corta existente entre ellos. El proceso continua hasta que todos los objetos se encuentran en un conglomerado. A este procedimiento también es conocido como el método del *Vecino más Cercano*.

En base al procedimiento descrito en al párrafo anterior, podemos enunciar la forma matemática del mismo que da lugar a la formación de los conglomerados. Supongamos que tenemos un grupo A con n_a y un grupo B con n_b elementos, y que ambos se fusionan para crear un nuevo grupo AB con $n_a + n_b$ elementos. La distancia del nuevo grupo AB a otro grupo C con n_c elementos en base al método de encadenamiento simple se calcula como sigue:

Tenemos que la distancia entre los dos nuevos grupos es la menor de las distancias entre grupos antes de la fusión. Es decir:

$$d(C; AB) = \text{mín}(d_{CA}, d_{CB})$$

Una forma alternativa y simple de calcular esta distancia cuando se desea programar o utilizar en computadora este procedimiento es reexpresar al mínimo entre las dos distancias como:

$$\text{mín}(d_{CA}, d_{CB}) = 1/2(d_{CA} + d_{CB} - |d_{CA} - d_{CB}|)$$

En efecto, si $d_{CB} > d_{CA}$ el término en valor absoluto es $d_{CB} - d_{CA}$ y el resultado de la operación es d_{CA}, que resulta ser la menor de las distancias. Si $d_{CA} > d_{CB}$ el segundo término es $d_{CA} - d_{CB}$ y se tiene que d_{CB} es la menor de las distancias obtenidas.

La ventaja que se tiene con este criterio es que sólo depende del orden de las distancias, será invariante ante transformaciones monótonas y obtendremos la misma jerarquía aunque las distancias sean numéricamente distintas. Ahora, una desventaja es que los problemas se producen cuando están mal definidos los conglomerados. En estos casos este criterio tiende a producir grupos muy grandes, que pueden incluir elementos muy distantes en los extremos de los conglomerados, y los individuos que se encuentran en los extremos pueden ser muy diferentes a los individuos que se encuentran más cerca entre sí.

3.1.2. Encadenamiento Completo

El método de encadenamiento completo es parecido al de encadenamiento simple excepto en que el criterio de aglomeración se basa en la distancia máxima. Por esta razón, a veces se le conoce también como el método del *vecino más lejano*. La distancia máxima entre individuos de cada conglomerado representa la esfera más reducida (diámetro mínimo) que puede incluir a todos los objetos en ambos conglomerados. A este método se le denomina encadenamiento completo por que todos los objetos de un conglomerado se vinculan con el resto a alguna distancia máxima o por la mínima similitud. Podremos decir entonces que la similitud dentro del grupo es igual al diámetro del grupo.

El planteamiento matemático para este método es el siguiente. La distancia entre los dos nuevos grupos es la mayor de las distancias entre grupos antes de la fusión. Es decir, tendríamos que esta distancia está dada por:

$$d(C, AB) = \text{máx}(d_{CA}, d_{CB})$$

También podemos reexpresar al máximo de dos números para simplificar cálculos y asi poder emplearlo para ser utilizado en la computadora como sigue:

$$\text{máx}(d_{CA}, d_{CB}) = 1/2(d_{CA} + d_{CB} + |d_{CA} - d_{CB}|)$$

La ventaja de este criterio es que también es invariante ante transformaciones monótonas de las distancias al depender, como en el anterior, del orden de las distancias.

3.1.3. Encadenamiento Medio

El método de encadenamiento medio comienza igual que los métodos de encadenamiento simple y completo, pero el criterio de aglomeración es la distancia media de todos los individuos de un conglomerado con todos los individuos de otro. Tales técnicas no dependen de los valores extremos, como se hace en el encadenamiento simple o completo y la partición se basa en todos los miembros de los conglomerados en lugar de un par único de miembros extremos.

Podemos enunciar este método de la siguiente forma. La distancia entre dos nuevos grupos es la media ponderada de las distancias entre grupos antes de la fusión. Es decir:

$$d(C, AB) = \frac{n_a}{n_a + n_b} d_{CA} + \frac{n_b}{n_a + n_b} d_{CB}$$

Una ventaja es que el enfoque del encadenamiento medio tiende a combinar los conglomerados con variaciones reducidas dentro del conglomerado. Las desventajas son que este criterio no es invariante ante transformaciones monótonas. También tiende a estar sesgado hacia la producción de conglomerados con aproximadamente la misma varianza.

3.1.4. Método de Ward

El Método de Ward, ha sido propuesto por Ward y Whishart y consiste en que la distancia entre los conglomerados es la suma de los cuadrados entre dos conglomerados sumados para todas las variables. En cada paso del procedimiento de aglomeración, se minimiza la suma de cuadrados dentro del conglomerado para todas las particiones (el conjunto completo de conglomerados disjuntos) obtenida mediante la combinación de dos conglomerados en un paso previo.

La diferencia entre la formulación de los otros métodos anteriores es que ahora se parte de los elementos directamente, en lugar de utilizar la matriz de distancias, y se define la medida de distancia **W** de una agrupación de observaciones en grupos. Esta medida **W** es precisamente la suma de las distancias al cuadrado entre cada elemento y la media de su grupo:

$$\mathbf{W} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_{ig} - \overline{x}_g)(x_{ig} - \overline{x}_g)'$$

Donde $\overline{x}_g$ es la media del grupo g. El criterio comienza suponiendo que cada dato forma un grupo, $g = n$ y, por lo tanto, **W** es cero. A continuación,

se unen los elementos que produzcan el incremento mínimo de **W**. Obviamente, esto implica tomar los más próximos con la distancia euclidiana. En la siguiente etapa, tenemos $n-1$ grupos, $n-2$ de un elemento y uno de dos elementos. Decidimos de nuevo unir dos grupos para que **W** crezca lo menos posible, con lo que pasamos a $n-2$ grupos y así sucesivamente hasta tener un único grupo. Los valores de **W** van indicando el crecimiento del criterio al formar grupos y pueden utilizarse para decidir cuantos grupos naturales contienen nuestros datos.

Puede demostrarse que, en cada etapa, los grupos que deben unirse para minimizar **W** son aquellos tales que:

$$min \frac{n_a n_b}{n_a + n_b} (x_{ig} - \overline{x_g})'(x_{ig} - \overline{x}_g)$$

Este procedimiento tiende a combinar los conglomerados con un número reducido de observaciones. La desventaja es que también esta sesgado hacia la producción de conglomerados con aproximadamente el mismo número de observaciones.

3.1.5. Método del Centroide

En el método del centroide, la distancia entre los dos conglomerados es la distancia (normalmente euclídea o cuadrada) entre sus centroides. Además, se aplica generalmente con solo variables continuas. Los **Centroides** de los grupos son los valores medios de las observaciones de las variables en el valor teórico del conglomerado. En este método, cada vez que se agrupa a los individuos, se calcula un nuevo centroide. Los centroides de los grupos cambian a medida que se fusionan conglomerados. En otras palabras, existe un cambio en un centroide de cada grupo cada vez que un nuevo individuo o grupo de individuos se añade al conglomerado existente.

Podemos llevar este método a su forma matemática de la manera siguiente: sabemos que la distancia entre dos grupos se hace igual a la distancia euclídea entre sus centros, donde se toman como centros los vectores de medias de las observaciones que pertenecen al grupo. Cuando de unen dos grupos se pueden calcular las nuevas distancias entre ellos sin utilizar los elementos originales. Y puede demostrarse que el cuadrado de la distancia euclídea de un grupo C a la unión de los grupos A, con n_a elementos y B con n_b elementos es:

$$d^2(C, AB) = \frac{n_a}{n_a + n_b} d^2_{CA} + \frac{n_b}{n_a + n_b} d^2_{CB} - \frac{n_a n_b}{(n_a + n_b)^2} d^2_{AB}$$

Este método es más popular entre los biólogos pero la desventaja es que pueden producirse resultados desordenados y a menudo confusos. La confusión se produce a causa de los cambios, esto es, casos donde la distancia entre los centroides de un par puede ser menor que la distancia entre los centroides de otro par fusionado en una combinación anterior. La ventaja de este método es que se ve menos afectada por datos atípicos que los otros métodos jerárquicos.

Ejemplos

Ahora mostraremos un ejemplo de un conjunto de datos simulados en el software **R** provenientes de dos poblaciones normales cuyas distribuciones son $N_1(-5, 4)$ y $N_2(0, 16)$ para ver como es que se representan gráficamente la obtención de conglomerados mediante algunos de estos métodos mencionados[2]. La forma como se construyeron los datos es la siguiente: La primera coordenada x de cada punto corresponde a un valor de la normal $N_1(-5, 4)$ y la segunda coordenada y corresponde a un valor de la normal $N_2(0, 16)$, luego, los puntos (x, y) se graficaron y es así como se obtuvo el diagrama de dispersión.

Primeramente mostraremos el diagrama de dispersión para los datos provenientes de estas dos poblaciones normales propuestas. Para este ejemplo se generaron 70 observaciones de cada población cuyo diagrama es el siguiente:

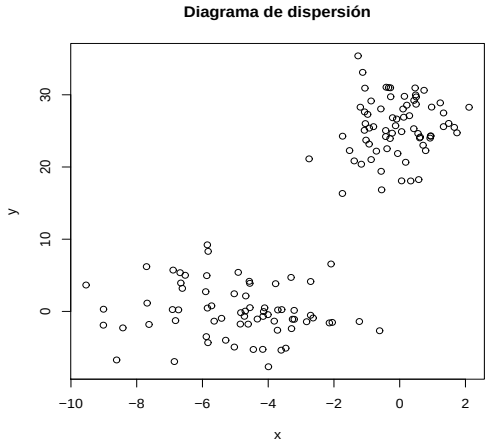

Figura 3.1: Datos simulados de dos poblaciones normales

[2]En el apéndice se proporcionara la sintaxis empleada en **R** para este ejemplo

Dado que éste es un ejemplo ideado para ilustrar de una forma clara y sencilla los resultados de la aplicación de estos métodos, es claro que en este diagrama se pueden distinguir dos grupos relativamente separados los cuales se intentara clasificar por los métodos de encadenamiento simple, encadenamiento completo y el encadenamiento medio. A continuación presentamos los resultados obtenidos de aplicar estos métodos en el conjunto de datos simulados.

Después de haber aplicado el método de encadenamiento simple, el dendograma que se obtuvo es el siguiente:

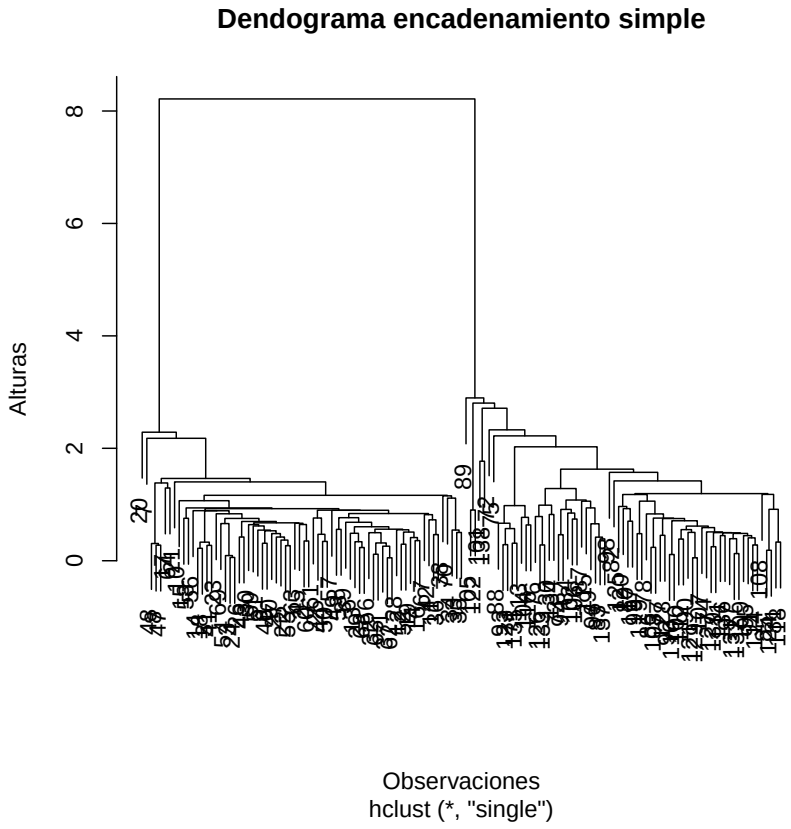

Figura 3.2: Dendograma de encadenamiento simple

Podemos observar en el dendograma que se tiene una clasificación del conjunto de datos en dos grupos ya que es muy notorio que existen dos ramificaciones de las observaciones lo cual lleva a determinar que el número de datos obtenidos por este método coincide con el número real de grupos.

También mostramos el dendograma obtenido mediante el método de encadenamiento completo el cual es:

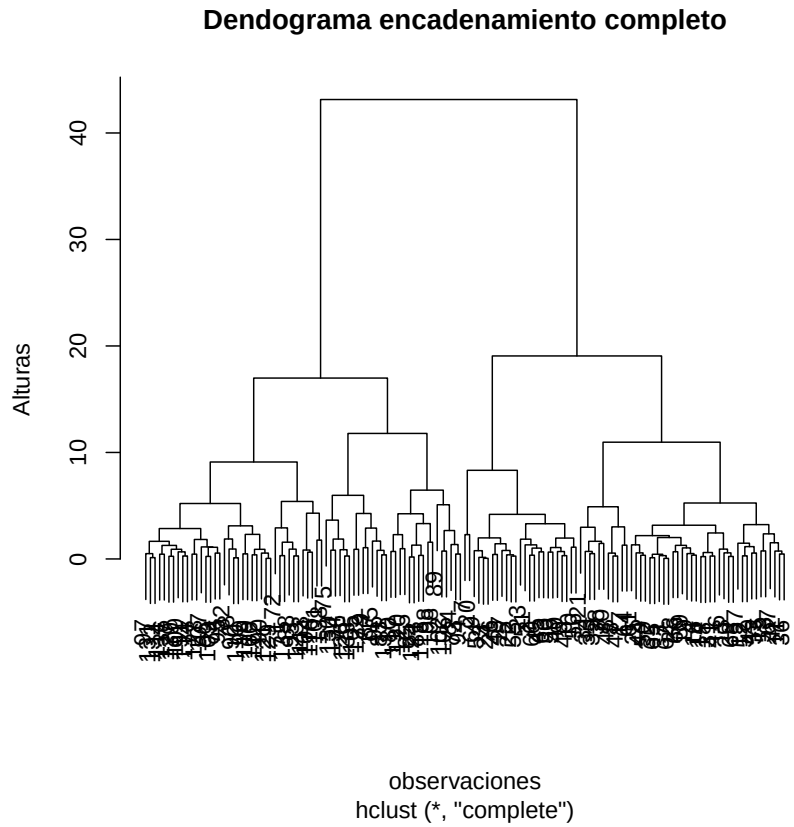

Figura 3.3: Dendograma resultado del encadenamiento completo

Aquí también es bastante notoria la formación de dos grupos mediante este método lo cual concuerda con los datos reales.

Por último mostraremos el dendograma correspondiente al método de encadenamiento medio que es un promedio de los dos anteriores.

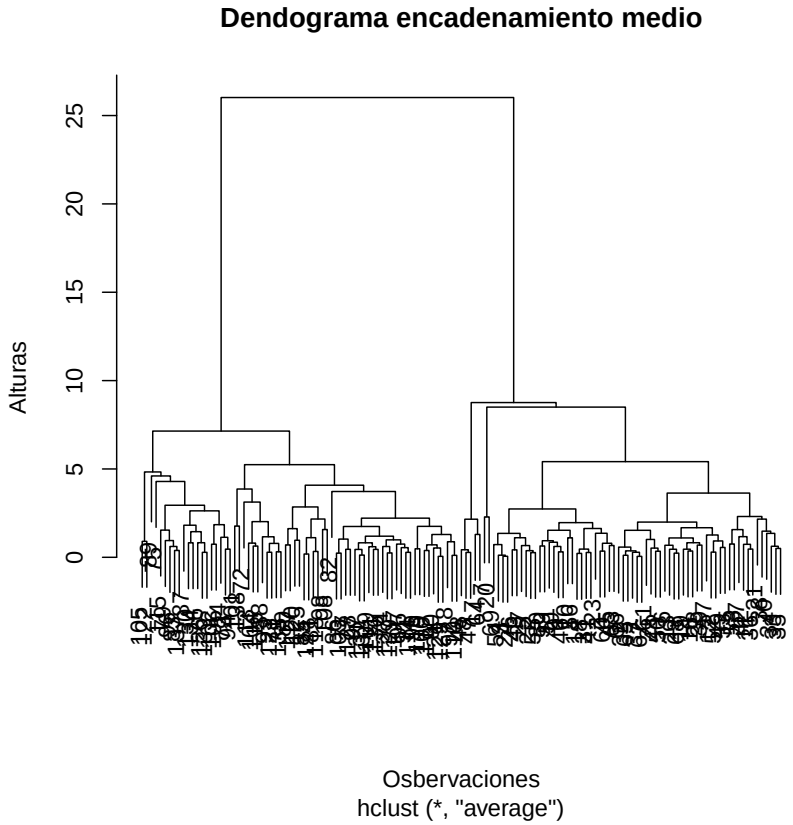

Figura 3.4: Dendograma resultado del encadenamiento promedio

Este dendograma viene a corroborar que efectivamente estos tres métodos con este conjunto de datos simulados en **R** concuerdan y dan una buena clasificación de los datos y los asocia de la siguiente manera a su correspondiente grupo como se muestra en la siguiente gráfica:

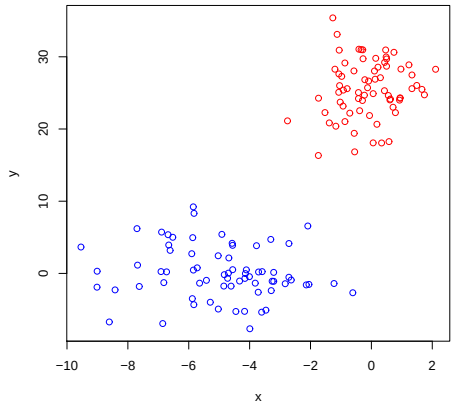

Figura 3.5: Clasificación de los datos

Es importante aclarar que esto no ocurre con frecuencia en la práctica, ya que la determinación del número de grupos reales es difícil y se pueden obtener en algunos métodos un cierto número de grupos y en otros otro número diferente.

Estos no son los únicos métodos propuestos para realizar clasificación no supervisada, a continuación mostraremos los métodos no jerárquicos para la formación de conglomerados.

3.2. Métodos No Jerárquicos

A diferencia con los métodos jerárquicos, los procedimientos no jerárquicos no implican los procesos de construcción de árboles. En vez de esto, se asignan los objetos a conglomerados una vez que el número de conglomerados a formar esta especificado. Por lo tanto, la solución de seis conglomerados no solo es una combinación de dos conglomerados desde una solución de siete conglomerados, sino que se basa sólo en la búsqueda de la mejor solución de seis conglomerados.

En un ejemplo simple, el proceso opera de la siguiente forma. El primer paso es seleccionar una **Semilla de conglomerado** como centro de conglomerado inicial, y todos los objetos (individuos) dentro de una distancia umbral previamente especificada se incluyen dentro del conglomerado resultante. Entonces se selecciona otra semilla de conglomerado y la asignación continua hasta que todos los objetos están asignados. Los objetos pueden entonces asignarse si están cercanos a otro conglomerado que no sea el original.

Existen diferentes aproximaciones para seleccionar las semillas de conglomerado y asignar objetos.

Ahora bien. En los métodos no jerárquicos o de partición disponemos de datos que sospechamos son heterogéneos y se desea dividirlos en un número de grupos prefijado, de tal forma que:

1. Cada elemento pertenezca a uno, y solo uno de los grupos.
2. Todo elemento quede clasificado.
3. Cada grupo sea internamente homogéneo.

Un ejemplo que podemos dar para la aplicación de este método es el siguiente. Supongamos que tenemos que analizar y clasificar una base de datos de compras en una empresa y se desea hacer una clasificación de clientes en función de sus características de consumidor.

3.2.1. Algoritmo de K-Medias

Los procedimientos de aglomeración no jerárquicos se denominan frecuentemente como aglomeración de **K-medias**, y normalmente utilizan una de las siguientes tres aproximaciones para asignar las observaciones individuales de uno de los conglomerados.

1. Umbral Secuencial
2. Umbral Paralelo
3. Optimización

El **Umbral Secuencial** empieza seleccionando una semilla de conglomerado e incluye todos los objetos que caen dentro de una distancia previamente especificada. Cuando los objetos dentro de la distancia están incluidos, se selecciona una segunda semilla de conglomerado y se incluyen todos los objetos dentro de la distancia previamente especificada. A continuación se selecciona una tercera semilla, y el proceso continua cono se ha descrito. Cuando un objeto se incluye en un conglomerado con una semilla, no se considera a efectos de anteriores semillas.

El **Umbral Paralelo** en contraste, selecciona varias semillas de conglomerado simultáneamente al principio y asigna objetos dentro de la distancia umbral hasta la semilla más cercana. A medida que el proceso avanza, se puede ajustar las distancias umbral para incluir más o menos objetos en los conglomerados. También en algunas variantes de este método, los objetos

permanecen fuera de los conglomerados si están fuera de la distancia previamente especificada desde cualquiera de las semillas de conglomerado.

La **Optimización** es parecida a los otros dos procedimientos no jerárquicos excepto en que permite la reubicación de los objetos. Si en el curso de la asignación de los objetos, un objeto se acerca más a otro conglomerado que no es el que tiene asignado en este momento, entonces un procedimiento de optimización cambia el objeto al conglomerado mas cercano.

Implementación del Algoritmo

Supongamos una muestra de n elementos con p variables. El objetivo es dividir esta muestra en un número de grupos prefijados, K. Requiere de las 4 etapas siguientes:

1. Seleccionar K puntos como centros de los grupos iniciales. Esto puede hacerse:

   a) asignando aleatoriamente los objetos a los grupos y tomando los centros de los grupos formados.

   b) tomando como centros los K puntos más alejados entre si.

   c) construyendo unos grupos iniciales con información *a priori* y calculando sus centros, o bien seleccionando los centros *a priori*.

2. Calcular las distancias euclidianas de cada elemento a los centros de los K grupos, y asignar cada elemento al grupo cuyo centro este más próximo. La asignación se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalculan las coordenadas del nuevo centro del grupo.

3. Definir un criterio de optimización y comprobar si resignando alguno de los elementos mejora el criterio.

4. Si no es posible mejorar el criterio de optimización, terminar el proceso.

El criterio de similaridad o de optimización que se utiliza en el algoritmo de k-medias, es minimizar la **suma de cuadrados dentro de los grupos** ($SCDG$) para todas las variables dada por:

$$SCDG = \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{i=1}^{n_k} (x_{ijk} - \overline{x}_{jk})^2$$

Donde x_{ijk} es el valor de la variable j en el elemento i del grupo k y $\overline{x_{jk}}$ es la media de esta variable en el grupo. Este criterio es equivalente a la suma

ponderada de las varianzas de las variables de los grupos, ya que se puede escribir de la siguiente forma:

$$\text{mín}\, SCDG = \text{mín} \sum_{k=1}^{K} \sum_{j=1}^{p} n_k s_{jk}^2$$

Donde n_k es el número de elementos del grupo g y s_{jk}^2 es la varianza de la variable j en dicho grupo.

Las varianzas de las variables en los grupos son claramente una medida de la heterogeneidad de la clasificación y al minimizarlas obtendremos grupos más similares. Un criterio alternativo de similaridad sería minimizar las distancias al cuadrado entre las observaciones y los centros de cada grupo. Si medimos las distancias con la norma euclidiana, este criterio se escribe como:

$$\text{mín} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ijk} - \overline{x}_{jk})'(x_{ijk} - \overline{x}_{jk}) = \text{mín} \sum_{k=1}^{K} \sum_{i=1}^{n_k} d^2(i,k)$$

Donde $d^2(i,k)$ es el cuadrado de la distancia euclidiana entre el elemento i del grupo g y su media de grupo. Se puede comprobar que ambos criterios son equivalentes. Ya que se tiene que un escalar es igual a su traza, podemos escribir este último criterio de la forma:

$$\text{mín} \sum_{k=1}^{K} \sum_{i=1}^{n_k} tr(d^2(i,k)) = \text{mín}\, tr[\sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \overline{x}_k)(x_{ik} - \overline{x}_k)']$$

Y nombrando a $\mathbf{W}$ la matriz de suma de cuadrados dentro de los grupos,

$$\mathbf{W} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \overline{x}_k)(x_{ik} - \overline{x}_k)'$$

tenemos que $\text{mín}\, tr(\mathbf{W}) = \text{mín}\, SCDG$. Y ambos criterios son equivalentes. Este criterio se denomina *criterio de la traza* y fue propuesto por Ward en el año de 1963.

La minimización de la suma de cuadrados dentro de cada grupo requiere calcular la suma

$$SCDG = \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{i=1}^{n_k} (x_{ijk} - \overline{x}_{jk})^2$$

Para todas las posibles particiones, lo cual resulta ser bastante complicado a menos que n tome valores pequeños. El algoritmo de k-medias busca la partición óptima con la restricción de que en cada iteración sólo se permite mover un elemento de un grupo a otro. el algoritmo funciona como sigue:

1. Partir de una asignación inicial.

2. Comprobar si moviendo algún elemento se reduce la $tr(\mathbf{W})$.

3. Si es posible reducir $tr(\mathbf{W})$ moviendo un elemento hacerlo, recalcular las medias de los dos grupos afectados por el cambio y volver a (2). Si no es posible reducir $tr(\mathbf{W})$, terminar.

En consecuencia, el resultado del algoritmo puede depender de lo asignación inicial y del orden de los elementos. Conviene siempre repetir el algoritmo con distintos valores iniciales y permutando los elementos de la muestra. El efecto del orden suele ser pequeño, pero conviene asegurarse en cada posible caso.

Número de Grupos

En la aplicación habitual del algoritmo de k-medias hay que fijar el número de grupos K. Resulta ser que no puede estimarse este número con un criterio de homogeneidad, ya que la forma de conseguir grupos muy homogéneos y a la vez minimizar la $SCDG$, es hacer tantos grupos como observaciones, con lo que siempre $SCDG$=0. Muchos métodos han sido propuestos para fijar un número de grupos. Uno de los métodos más utilizados es realizar una prueba de hipótesis basado en una estadística **F** de reducción de la variabilidad, comparando la $SCDG$ con K grupos con la de $(K+1)$ grupos, y calculando la reducción relativa de la variabilidad al aumentar un grupo adicional. La estadística de prueba es:

$$\mathbf{F} = \frac{SCDG(K) - SCDG(K+1)}{SCDG(K+1)/(n-K-1)}$$

Esta estadística compara la disminución de la variabilidad al aumentar un grupo con la varianza promedio. El valor obtenido se compara con el cuantíl de una distribución F de Fisher con p y $p(n-K-1)$ grados de libertad denotado por $F_{p,p(n-K-1)}$. Rechazando la hipótesis nula si la estadística **F** es mayor al cuantil $F_{p,p(n-K-1)}$ al nivel α. Este método no tiene una buena justificación ya que los datos no tienen porque seguir un comportamiento parecido a una distribución F para la aplicación de esta. Otra regla empírica comúnmente utilizada en varios paquetes estadísticos es introducir un grupo más si la estadística **F** es mayor a 10.

Una vez que los distintos métodos de clasificación no supervisada más comunes en la práctica han sido descritos y detallados, es momento de pasar a otro tipo de métodos. Estos métodos son de gran utilidad en la formación de

conglomerados dada su gran flexibilidad y sustento en la teoría de la probabilidad, son capaces de dar mas justificación y mas certidumbre al momento de preguntarse ¿cuantos grupos existen? Estos son los métodos *modelos de mezclas* y en la siguiente sección daremos una descripción de ellos.

3.3. Conglomerados basados en modelos de mezclas

La importancia de los modelos de mezclas en el análisis estadístico de datos es evidente en el rango tan amplio de artículos que han utilizado a estos modelos para aplicaciones estadísticas y en la literatura científica en general. Los modelos basados en mezclas han provisto una base matemática para la aproximación en la modelación estadística en una amplia variedad de fenómenos aleatorios. Dada su gran utilidad así como su flexibilidad, estos modelos han continuado recibiendo una creciente atención a través de los años, desde ambos puntos de vista tanto teórico, como práctico. En verdad, en décadas pasadas el extenso potencial de sus aplicaciones se ha ampliado considerablemente. Campos como la astronomía, biología, ingeniería, genética, mercadotecnia, medicina, psiquiatría y ciencias sociales entre muchos otros son algunos en los cuales los modelos basados en mezclas han sido aplicados con gran éxito.

Cabe señalar que en estas aplicaciones los modelos basados en mezclas respaldan muchas de las áreas de la estadística, incluyendo el análisis de conglomerados, análisis discriminante, análisis de supervivencia, sumados a estos el papel mas importante en al análisis de datos e inferencia es proveer de modelos descriptivos para distribuciones.

La utilidad de los modelos basados en mezclas en la modelación de la heterogeneidad en un contexto de análisis de conglomerados es obvio ya que como se verá mas adelante, a cada conglomerado del conjunto de datos se le puede asociar un componente de la mezcla. En otro ejemplo donde existe una estructura de un grupo, estos tienen un papel de mucho uso en la evaluación de los índices de error (sensibilidad y especificación) en el diagnostico y proceso de monitoreo en la ausencia de algún estándar. Pero como una distribución continua puede ser aproximada arbitrariamente bien por una mezcla de densidades normales con una varianza común (o una matriz de covarianzas en el caso multivariado), los modelos de mezclas proveen de un sistema semiparamétrico conveniente en el cual un modelo con una forma desconocida en su distribución puede ser aproximado no importando cual sea el objetivo aun cuando sea, por ejemplo, una estimación de una distribución

a priori en estadística bayesiana.

Ya que los modelos basados en mezclas son de mucha utilidad en estadística debido a que fundamentan varias ramas de esta misma, es necesario plantear la teoría que esta detras de ellos y mencionar algunas de sus propiedades y los conceptos que hacen de estos una herramienta muy versátil para la modelación de grupos. Daremos a continuación las definiciones y conceptos básicos que son empleados para la construcción de estos modelos.

3.3.1. Definiciones Básicas

Definición 32.- Sea $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ una muestra aleatoria de tamaño n donde $\mathbf{x}_k$ representa un vector aleatorio de dimensión p con función de densidad $f(\mathbf{x}_k)$ en $\mathbb{R}^p$. De donde $\mathbf{x}_i$ contiene las p variables aleatorias correspondientes a la i-ésima observación de algún fenómeno. Y sea $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n')'$. De donde $(')$ denota el vector transpuesto. Usaremos $\mathbf{x}$ para denotar a toda la muestra representada en la matriz de datos. Y para denotar a una realización de alguno de los vectores aleatorios usaremos las letras minúsculas $x = (x_1', x_2', \ldots, x_n')'$ de donde x_i es el i-ésimo valor observado del vector aleatorio $\mathbf{x}_k$.

Definimos una mezcla de distribuciones como:

$$f(\mathbf{x}) = \sum_{k=1}^{g} \pi_k f_k(\mathbf{x}|\theta) \tag{3.1}$$

De donde g es el número de componentes que se supone es un número conocido y $f_k(\mathbf{x}|\theta)$ son funciones de densidad y π_k son constantes no negativas tales que cumplen con:

$$0 \leq \pi_k \leq 1 \qquad \text{y} \qquad \sum_{k=1}^{g} \pi_k = 1 \qquad (k = 1, \ldots, g)$$

A pesar de que aquí estamos suponiendo que una de las características del vector $\mathbf{x}$ es que sea un vector aleatorio continuo, podemos ver sin problema alguno a $f_k(\mathbf{x}|\theta)$[3] como una función de densidad discreta cuando se tienen variables de conteo.

Las constantes $\pi_1, \ldots, \pi_g$ son llamadas las *"proporciones"* o *"pesos"*. Como las funciones $f_1(\mathbf{x}), \ldots, f_g(\mathbf{x})$ son funciones de densidad y dado que una combinación lineal convexa de funciones de densidad da como resultado otra función de densidad, entonces la mezcla $f(\mathbf{x})$ es también una función de densidad. Las funciones $f_k(\mathbf{x})$ se les da el nombre de *densidades componente* de

[3]Podemos usar esta notación o también usaremos para mayor comodidad la expresión $f_k(\mathbf{x})$.

la mezcla. Nos referiremos a la función de densidad $f(\mathbf{x})$ como una distribución *mezcla* de g componentes y denotaremos por $F(\mathbf{x})$ a su correspondiente función de distribución llamada función de *distribución* de g componentes [4].

Durante toda la formulación de la teoría de los modelos de mezclas, el número de componentes g es considerado fijo. Pero en la gran mayoría de los casos y en las aplicaciones, el valor de g es desconocido y tiene que ser estimado por los datos disponibles, y junto con las proporciones de la mezcla y los parámetros en las distribuciones especificadas de las densidades componentes. En el contexto del análisis de conglomerados, se suele asociar un componente de la mezcla a un grupo presente en la población de interés.

3.3.2. Interpretación de un modelo de mezclas

Una manera muy fácil de generar un vector aleatorio $\mathbf{x}$ con una mezcla de distribuciones de g componentes $f(\mathbf{x}_k)$ es la siguiente. Sea $\mathbf{z}_j$ una variable aleatoria categórica, es decir que toma los valores $1, 2, \ldots, g$ con probabilidades $\pi_1, \pi_2, \ldots, \pi_g$ respectivamente, y suponiendo que las densidades condicionales de $\mathbf{x}_k$ dado $\mathbf{z}_j = i$ es $f_i(\mathbf{x}_k)$ con $k = 1, \ldots, g$. Entonces la función de densidad marginal de $\mathbf{x}_k$ esta dada por $f(\mathbf{x}_k)$. En este contexto, la variable $\mathbf{z}_j$ puede ser pensada como una etiqueta[5] del vector $\mathbf{x}_k$. Por simplicidad, es conveniente trabajar después con un vector de etiquetas de dimensión g en lugar de una única variable categórica z_j donde el i-ésimo elemento de $\mathbf{z}_j$, $\mathbf{z}_{ij} = (\mathbf{z}_j)_i$, es definido como cero o uno de acuerdo el componente de origen $\mathbf{x}_k$ en la mezcla es igual a k o no (i=1,. . . ,g). Entonces z_j tiene una distribución multinomial de g categorías con sus respectivas probabilidades $\pi_1, \ldots, \pi_g$ de la siguiente manera:

$$P(\mathbf{z}_j = z_j) = \frac{n!}{n_1! n_2! \cdots n_g!} \pi_1^{\mathbf{z}_{1j}}, \pi_2^{\mathbf{z}_{2j}}, \ldots, \pi_g^{\mathbf{z}_{gj}}$$

Y la denotaremos como

$$\mathbf{z}_j \sim \text{Multinomial}_g(1, \pi) \qquad \text{Donde } \pi = (\pi_1, \ldots, \pi_g)$$

En la interpretación que acabamos de dar de un modelo de mezclas, una situación muy obvia es cuando en el modelo de mezclas con g componentes es deseable una población consistente en g grupos $\mathbf{G}_1, \ldots, \mathbf{G}_g$ con proporciones $\pi_1, \ldots, \pi_g$. Si la función de densidad de $\mathbf{x}_k$ en el grupo $\mathbf{G}_i$ esta dada por

[4] Solo haremos referencia al uso de mezclas de un número finito de componentes y usualmente nos referiremos a estas simplemente como *mezclas*.

[5] Esto lo podemos pensar como la etiqueta o el numero de conglomerado al que pertenece

$f(\mathbf{x}_k)$ para $k = 1, \ldots, g$, entonces la función de densidad de $\mathbf{x}_j$ tiene el g-ésimo componente de la mezcla. En este caso los g componentes de la mezcla pueden ser físicamente identificados externamente con los g grupos existentes $G_1, \ldots, G_g$.

No es objetivo de este trabajo demostrar que existen muchos ejemplos en la práctica para estos modelos de mezclas donde la población es una mezcla de g distintos grupos que son conocidos de una manera *a priori*. Sin embargo, hay también ejemplos que envuelven el uso de modelos basados en mezclas donde los componentes no pueden ser identificados de una manera fácil o *a priori* como en le caso anterior. En algunos casos, los componentes son introducidos en un modelo de mezclas para permitir una mejor flexibilidad en la modelación de una población heterogénea que es aparentemente difícil de modelar por una simple distribución.

Comúnmente en el ajuste de modelos basados en mezclas se asume que los datos son generados por una mezcla de distribuciones de probabilidad subyacentes en que alguna de las distribuciones componentes es un grupo o un conglomerado. Dadas las observaciones $x = (x_1, x_2, \ldots, x_n)$ y sean $f_k(\mathbf{x}_i|\theta_k)$ la función de densidad de la i-ésima observación proveniente del k-ésimo componente, donde θ_k es el vector de parámetros de la distribución correspondiente y sea G el número de componentes en la mezcla. Este es el punto que se va a discutir en este trabajo ya que a veces la obtención de un modelo de mezclas, que por lo general serán distribuciones normales multivariadas, este modelo puede no ser tan práctico al final de todo. Esto debido a que se pueden tener más componentes de las que se esperan y así ser menos fácil de manejar o interpretar el modelo.

Este modelo es usualmente formulado en una de las siguientes dos formas.

1. Verosimilitud de la clasificación.

2. Verosimilitud de la mezcla.

El primer modelo trata de maximizar la expresión

$$L_c(\theta_1, \ldots, \theta_g; z_1, \ldots, z_n; \mathbf{x}) = \prod_{i=1}^{n} f_{\gamma_i}(\mathbf{x}_i; \theta_{\gamma_i}) \tag{3.2}$$

Donde z_i son valores discretos de una etiqueta de clasificación, $\gamma_i = k$ si x_i pertenece al k-ésimo grupo.

El segundo maximiza la expresión

$$L_c(\theta_1, \ldots, \theta_g; \pi_1, \ldots, \pi_n; \mathbf{x}) = \prod_{i=1}^{n} \sum_{k=1}^{g} \pi_k f_k(\mathbf{x}_i; \theta_k) \tag{3.3}$$

Donde π_k es la probabilidad de que una observación pertenezca al k-ésimo componente y además $\pi_k \geq 0$ y $\sum_{k=1}^{g} \pi_{=}1$

Estaremos principalmente interesados en el caso en que $f_k(x_i; \theta_k)$ sea una normal multivariada, este un modelo que ha sido utilizado con mucho éxito en un número de aplicaciones. En esta instancia, los parámetros θ_k consisten en un vector de medias μ_k y una matriz de varianzas y covarianzas Σ_k y su función de densidad está dada como

$$f_k(x_i; \mu_k, \Sigma_k) = \frac{\exp -\frac{1}{2}(x_i - \mu_k)' \Sigma_k^{-1}(x_i - \mu_k)}{(2\pi)^{\frac{p}{2}} |\Sigma_k|^{\frac{1}{2}}}$$

Los conglomerados tienen formas elipsoidales con centros en los vectores μ_k y la matriz Σ_k determina otras características geométricas de los conglomerados.

3.3.3. Estimación del modelo basado en mezclas

A través de los años, una gran variedad de aproximaciones han sido utilizadas para estimar las distribuciones de mezclas de funciones de densidad. Estas han incluido métodos gráficos, métodos de por momentos, distancias mínimas, máxima verosimilitud y aproximaciones bayesianas. Tal vez la razón principal para los investigadores y la literatura en la metodología de la estimación para modelos basados en mezclas, es de hecho las fórmulas explicitas para los parámetros a estimar no están disponibles de una forma fácil y práctica. Por ejemplo, el estimador máximo-verosímil para los pesos de la mezcla y las medias de las distribuciones componentes, así como para las varianzas y covarianzas no pueden ser escritas de una forma que puedan ser llevadas a una forma sencilla y práctica. Estos estimadores tienen que ser aproximados mediante métodos iterativos y computacionales. De hecho se mencionaran algunas técnicas para estimar estos parámetros mediante software y algoritmos, así como también se enunciara el algoritmo EM que es de gran utilidad para estos casos donde es difícil obtener un estimador máximo verosímil de una forma mas fácil o directa.

Cuando tenemos que realizar una estimación de uno o más parámetros, el método más común y conocido es el método de la máxima verosimilitud, que como ya se ha mencionado en secciones anteriores, consiste en maximizar la función de verosimilitud de cierta función de densidad $f(x; \theta_1, \ldots, \theta_g)$y encontrar el vector $(\widehat{\theta}_1, \ldots, \widehat{\theta}_g)$ que maximice la probabilidad de que la muestra obtenida sea seleccionada. En el caso de los modelos basados en mezclas, ya sabemos que su función de verosimilitud esta dada por la expresión (3.3). La cual puede ser una función bastante complicada de maximizar por méto-

dos normales de calculo vectorial, es entonces cuando se recurre a métodos iterativos.

Ya se ha expuesto la teoría acerca de estos modelos de mezclas. Dada la facilidad que se cuenta al manejar distribuciones normales ya que poseen propiedades probabilísticas que en estadística son muy útiles. Al suponer distribuciones normales tanto univariadas como multivariadas tiene mayor facilidad para realizar estimaciones, estimaciones por intervalos y pruebas de hipótesis, esto da lugar a suponer que las distribuciones componentes siguen una distribución normal. Formularemos a continuación esta teoría en el caso particular en el que $\mathbf{x} \sim N(\mu, \Sigma^2)$.

3.3.4. Estimación de modelos de mezclas normales

Un enfoque natural o intuitivo para realizar la subdivisión de la muestra en grupos o conglomerados es suponer que los datos se han generado como una mezcla de distribuciones normales multivariadas y estimar conjuntamente los parámetros de las distribuciones que forman la mezcla y los pesos de cada función de distribución componente o probabilidades de cada dato de pertenecer a determinado grupo. Vamos a presentar este enfoque.

Como ya se ha definido anteriormente, los datos provienen de una mezcla de distribuciones $f(\mathbf{x}) = \sum_{k=1}^{g} \pi_k f_k(\mathbf{x})$ que tiene como ya se ha visto su función de verosimilitud correspondiente

$$L_M(X;\theta) = \prod_{i=1}^{n}(\sum_{k=1}^{g} \pi_k f_k(\mathbf{x}_i)) \tag{3.4}$$

Y puede escribirse como la suma de g^n términos correspondientes a todas las posibles clasificaciones de las n observaciones entre los G grupos. Y su función log-verosimilitud es

$$l_M(X;\theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{g} \pi_k f_k(\mathbf{x}_i) \tag{3.5}$$

Ahora supongamos que cada función componente $f_k(\mathbf{x})$ es una normal p-variada con un vector de medias μ_k y matriz de varianzas y covarianzas Σ_k, de manera que el vector de parámetros desconocidos es

$$\theta = (\pi_1, \ldots, \pi_G, \mu_1, \ldots, \mu_G, \Sigma_1, \Sigma_2, \ldots, \Sigma_G) \tag{3.6}$$

Sustituyendo estas densidades por su expresión original en su log-verosimilitud tenemos lo siguiente:

$$l_M(X;\theta) = \sum_{i=1}^{n} \log(\sum_{k=1}^{G} \pi_k |\Sigma_k|^{-\frac{1}{2}} (2\pi)^{-\frac{p}{2}} \exp(-\frac{1}{2}(\mathbf{x}_i - \mu_k)' \Sigma_k^{-1} (\mathbf{x}_i - \mu_k))) \quad (3.7)$$

Si observamos que haciendo en esta función $\widehat{\mu}_k = \mathbf{x}_i$, la estimación de Σ_k es cero si $\pi_k \neq 0$, el cociente $\pi_k |\Sigma_k|^{-\frac{1}{2}}$ tiende a infinito y también lo hará la función log-verosímil. Por lo tanto, esta función tiene muchos máximos, ligados a soluciones donde cada densidad viene determinada exactamente por una observación. Para evitar estas singularidades supondremos que, como mínimo, hay p observaciones de cada distribución, y trataremos de encontrar un máximo local de esta función que proporcione un estimador consistente de los parámetros.

Un problema adicional es que las distribuciones normales no están identificadas, ya que el orden $1, \ldots, G$ es arbitrario. Para resolver este problema podemos suponer que las distribuciones $f_1, \ldots, f_G$ corresponden a $\pi_1 \geq \pi_2 \geq \ldots \geq \pi_G$ o definir el orden de las distribuciones por una medida del tamaño de la media o la matriz de covarianzas.

Para maximizar esta función con relación a las probabilidades π_k hay que tener en cuenta que $\sum_{k=1}^{G} \pi_k = 1$. Introduciendo esta restricción con un multiplicador de Lagrange en la log-verosimilitud, la función a maximizar es

$$l_M(X;\theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{G} \pi_k f_k(\mathbf{x}_i) - \lambda(\sum_{k=1}^{G} \pi_k - 1) \quad (3.8)$$

Derivando respecto a los pesos de cada grupo obtenemos lo siguiente:

$$\frac{\partial l_M(X;\theta)}{\partial \pi_k} = \sum_{i=1}^{n} \frac{f_k(\mathbf{x}_i)}{\sum_{k=1}^{G} \pi_k f_k(\mathbf{x}_i)} - \lambda = 0$$

y multiplicando por π_k, y como por hipótesis $\pi_k \neq 0$ ya que en otro caso el modelo k es redundante, podemos escribir

$$\lambda \pi_k = \sum_{i=1}^{n} \pi_{ik} \quad (3.9)$$

Donde

$$\pi_{ik} = \frac{\pi_k f_k(\mathbf{x}_i)}{\sum_{k=1}^{G} \pi_k f_k(\mathbf{x}_i)} \quad (3.10)$$

Los coeficientes π_{ik} representan la probabilidad de que, una vez observado un dato $\mathbf{x}_i$ haya sido generado por la distribución normal multivariada $f_k(\mathbf{x}_k)$. Estas probabilidades se denominan *a posteriori* y se calculan por el teorema de Bayes. Su interpretación es la siguiente. Antes de observar $\mathbf{x}_i$

la probabilidad de que cualquier observación, y en particular la $\mathbf{x}_i$, venga del grupo k es π_k. Sin embargo, después de observar $\mathbf{x}_i$, esta probabilidad se modifica en función de lo compatible que sea este valor con el modelo k. Esta compatibilidad se mide por $f_k(\mathbf{x}_i)$: si este valor es relativamente alto, aumentara la probabilidad de que venga del modelo k. Es obvio que para cada dato se cumple $\sum_{k=1}^{G} \pi_{ik} = 1$.

Para determinar el valor de λ, sumando en la expresión (3.9) que renombraremos como λ' para todos los grupos obtenemos que

$$\lambda = \sum_{i=1}^{n} \sum_{k=1}^{G} \pi_{ik} = n$$

Y susbtituyendo en λ' obtenemos las ecuaciones para estimar las probabilidades *a priori* son

$$\widehat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \pi_{ik} \tag{3.11}$$

Que proporcionan las probabilidades *a priori* como medio promedio de las probabilidades *a posteriori*.

Ahora hagamos los cálculos para realizar las estimaciones de los parámetros de las distribuciones normales multivariadas. Derivando la función log-verosimilitud respecto al vector de medias de la k-ésima distribución tenemos

$$\frac{\partial l_M(X;\theta)}{\partial \mu_k} = \sum_{i=1}^{n} \frac{\pi_k f_k(\mathbf{x}) \Sigma_k^{-1}(\mathbf{x}_i - \mu_k)}{\sum_{k=1}^{G} \pi_k f_k(\mathbf{x})} = 0 \qquad \text{Donde} \qquad k = 1, \ldots, G$$

Una vez resolviendo cada uno de estos sistemas para obtener el estimador máximo verosímil del k-ésimo vector de medias tenemos que es

$$\widehat{\mu}_k = \sum_{i=1}^{n} \frac{\pi_{ik}}{\sum_{i=1}^{n} \pi_{ik}} \mathbf{x}_i \tag{3.12}$$

Es decir, el vector de medias de cada distribución se estima como una media ponderada de todas las observaciones con pesos $\omega_{ik} = \frac{\pi_{ik}}{\sum_{i=1}^{n} \pi_{ik}}$, donde $\omega_{ik} \geq 0$ y $\sum_{i=1}^{n} \omega_{ik} = 1$. Los pesos ω_{ik} representan la probabilidad relativa de que la i-ésima observación pertenezca al k-ésimo grupo.

Análogamente obtendremos los estimadores para las matrices de covarianzas de cada distribución, derivamos respecto a Σ_k para obtener el sistema de ecuaciones y así el estimador máximo verosímil de Σ_k que es

$$\widehat{\Sigma}_k = \sum_{i=1}^{n} \frac{\pi_{ik}}{\sum_{i=1}^{n} \pi_{ik}} (\mathbf{x}_i - \widehat{\mu}_k)(\mathbf{x}_i - \widehat{\mu}_k)' \tag{3.13}$$

Este estimador tiene una interpretación parecida al anterior, es un promedio de las desviaciones de os datos respecto a sus medias, son pesos proporcionales a las probabilidades *a posteriori*.

Para resolver las ecuaciones para cada parámetro $\widehat{\pi}_k$, $\widehat{\mu}_k$ y $\widehat{\Sigma}_k$ y obtener estimadores necesitamos las probabilidades π_{ik}, y para calcular estas probabilidades con la expresión para π_{ik} necesitamos los parámetros del modelo.

En este trabajo, se hara énfasis en el uso del algoritmo EM para la estimación y ajuste de modelos de mezclas via máxima verosimilitud.

3.3.5. Algoritmo EM

El algoritmo esperanza-maximización o algoritmo EM se usa en estadística para encontrar estimadores de máxima verosimilitud de parámetros en modelos probabilísticos que dependen de variables no observables. El algoritmo EM alterna pasos de esperanza (paso E), donde se computa la esperanza de la verosimilitud mediante la inclusión de variables latentes como si fueran observables, y un paso de maximización (paso M), donde se computan estimadores de máxima verosimilitud de los parámetros mediante la maximización de la verosimilitud esperada del paso E. Los parámetros que se encuentran en el paso M se usan para comenzar el paso E siguiente, y así el proceso se repite.

El algoritmo EM viene expuesto por Arthur Dempster, Nan Laird y Donald Rubin de la Royal Statistical Society en una publicación de 1977. Los autores señalan que el método ya había sido propuesto muchas veces en situaciones especiales por otros autores, pero la publicación de 1977 generaliza el método y desarrolla la teoría detrás de él.

Para aplicar el algoritmo EM, ya se ha planteado al idea de introducir un conjunto de variables no observadas que llamamos etiquetas $(z_1, \ldots, z_n)$, que tienen como función indicar de que componente de la mezcla proviene cada observación. Con este objetivo, z_i será un vector aleatorio de dimensión Gx1[6] que tendrá una entrada igual a 1, que sera el correspondiente grupo del que proviene el dato x_i, y todas las entradas restantes serán igual a 0.

Un ejemplo sería que x_i proviene de la población número 1 si $z_{i1} = 1$ y $z_{i2} = z_{i3} = \ldots = z_{iG} = 0$, también ya se había mencionado la propiedad que $\sum_{g=1}^{G} z_{ig} = 1$ y $\sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} = n$. Con estas variables, la función de densidad

[6]En este caso asumiremos que G es el número de grupos

condicionada de $\mathbf{x}_i$ dada $\mathbf{z}_i$ es

$$f(\mathbf{x}_i|\mathbf{z}_i) = \prod_{g=1}^{G} f_g(\mathbf{x}_i)^{z_{ig}} \tag{3.14}$$

Podemos ver que, en $\mathbf{z}_i$ solo la entrada z_{ig} es distinto de cero y esa entrada definirá cual es la función de densidad de las observaciones. Análogamente, obtenemos la función de densidad de las variable $\mathbf{z}_i$ que es

$$f(\mathbf{z}_i) = \prod_{g=1}^{G} \pi_g^{z_{ig}} \tag{3.15}$$

por otro lado, la función de densidad conjunta es:

$$f(\mathbf{x}_i, \mathbf{z}_i) = f(\mathbf{x}_i|\mathbf{z}_i)f(\mathbf{z}_i) = \prod_{g=1}^{G} (\pi_g f_g(\mathbf{x}_i))^{z_{ig}} \tag{3.16}$$

y su función de log-verosimilitud es:

$$l(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_1, \ldots, \mathbf{z}_G|\theta) = \log(\prod_{i=1}^{n} f(\mathbf{x}_i, \mathbf{z}_i)) = \sum_{i=1}^{n} \log(f(\mathbf{x}_i, \mathbf{z}_i))$$

$$= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} log\pi_g + \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \log f_g(\mathbf{x}_i) \tag{3.17}$$

Si las variables z_{ig} que definen la población de la que proviene cada dato fueran conocidas, la estimación de los parámetros es la estudiada en el análisis discriminante que no es el objetivo de este trabajo abordar esa rama del análisis multivariado, pero para realizar esa estimación se tiene que la media de cada entrada se estima como el promedio de las observaciones generadas por el componente, que puede escribirse como

$$\widehat{\mu} = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\mathbf{x}_i \tag{3.18}$$

Y la matriz de covarianzas de cada grupo se calculará teniendo en cuenta sólo las observaciones de ese grupo mediante

$$\widehat{\Sigma}_g = \sum_{i=1}^{n} z_{ig}(\mathbf{x}_i - \overline{\mathbf{x}_g})(\mathbf{x}_i - \overline{\mathbf{x}}_g)' \tag{3.19}$$

Sin embargo, el problema es que ahora las variables de clasificación no son conocidas. La solución que proporciona el algoritmo EM es estimar las

variables z_{ig} mediante las probabilidades *a posteriori*, y después utilizar estas formulas.

El algoritmo EM comienza con una estimación inicial de los parámetros, a esta estimación la denotamos por $\widehat{\theta}^{(0)}$. En el paso **E** calcularemos el valor esperado de las observaciones ausentes en la verosimilitud completa condicionando a los parámetros iniciales y a los datos observados. Como la verosimilitud es lineal en z_{ig}, esto equivale a sustituir las variables ausentes por sus esperanzas. Las variables ausentes, z_{ig}, son variables binomiales con valores 0,1 y

$$E(z_{ig}) = p(z_{ig} = 1|\mathbf{X}, \widehat{\theta}^{(0)}) = p(z_{ig} = 1|\mathbf{x}_i, \widehat{\theta}^{(0)}) = \widehat{\pi}_{ig}^{(0)} \tag{3.20}$$

Donde $\widehat{\pi}_{ig}^{(0)}$ es la probabilidad de que la observación $\mathbf{x}_i$ venga del modelo j cuando ya se ha observado $\mathbf{x}_i$ y los parámetros de los modelos son los dados por $\widehat{\theta}^{(0)}$. Estas son las probabilidades *a posteriori* que se calculan por (3.10) utilizando como valores de los parametros especificados en $\widehat{\theta}^{(0)}$. Y al sustituir las variables ausentes por sus esperanzas se obtiene

$$l(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_1, \ldots, \mathbf{z}_G|\theta) = \sum_{i=1}^{n}\sum_{g=1}^{g} \widehat{\pi}_{ig}^{(0)} \log \pi_g + \sum_{i=1}^{n}\sum_{g=1}^{g} \widehat{\pi}_{ig}^{(0)} \log f_g(\mathbf{x}_g) \tag{3.21}$$

En el paso **M** se maximiza (3.21) respecto al vector de parámetros θ. Observamos que los parámetros π_g aparecen solo en el primer término y los de las normales solo en el segundo. Es así como podemos obtenerlos independientemente. Comenzando por los π_g estos parámetros están sujetos a que su suma debe ser igual a uno, por lo que la función a maximizar es

$$\sum_{i=1}^{n}\sum_{g=1}^{g} \widehat{\pi}_{ig}^{(0)} \log \pi_g - \lambda(\sum_{g=1}^{G} \pi_g - 1) \tag{3.22}$$

Que conduce a (3.11) con los valores π_{ig} ahora fijos a $\widehat{\pi}_{ig}^{(0)}$. Para obtener los estimadores de los parámetros de la distribución normal, derivando el segundo término se obtienen las ecuaciones (3.12) y (3.13), donde ahora las π_{ig} son iguales a $\widehat{\pi}_{ig}^{(0)}$. La solución de estas ecuaciones conducen a un nuevo vector de parámetros que llamaremos $\widehat{\theta}^{(1)}$, y el algoritmo se itera hasta obtener convergencia.

En resumen el algortimo es:

1. Parte de un valor $\widehat{\theta}^{(0)}$ y calcula $\widehat{\pi}_{ig}^{(0)}$ con (3.10).
2. Resuelve (3.11), (3.12) y (3.13) para obtener $\widehat{\theta}$.
3. Vuelve con este valor al paso 1 e itera 1 y 2 hasta convergencia.

Aplicación al análisis de conglomerados

Se han propuesto distintas aplicaciones de los modelos basados en mezclas de normales para resolver problemas de análisis de conglomerados. Autores como Banfield y Raftery (1993) y Dasgupta y Raftery (1988) han diseñado un método basado en mezclas de distribuciones normales y un algoritmo llamado MCLUST[7], que funciona bastante bien en la práctica cuando se tienen grupos bien separados y estos siguen significativamente una distribución normal. Las bases del procedimiento son comenzar el algoritmo EM con una estimación inicial obtenida mediante análisis jerárquico y reparametrizar las matrices de covarianzas para que puedan tener partes comunes y partes específicas. En resumen este procedimiento consiste en:

1. Seleccionar un valor M para el máximo número de grupos.

2. Estimar los parámetros de la mezcla con el algoritmo EM para $G = 1, \ldots, M$. Las condiciones iniciales del algoritmo se establecen con un método jerárquico de los ya mencionados anteriormente y la estimación se realiza para todas las posibles condiciones sobre las matrices de covarianzas. Si se consideran r condiciones se realizan un total de M_r estimaciones con el algoritmo EM.

3. Seleccionar finalmente el número de grupos y las condiciones de las matrices de covarianzas mas convenientes buscando la solución que maximiza el criterio de información bayesiana (BIC).

Solo por último nos resta mencionar ¿cómo es qué esta definido el BIC? y ¿por qué es utilizado como criterio para elegir un modelo final de mezclas? Y por tanto el número de grupos estimados por la mezcla.

3.3.6. Criterio BIC

El criterio para seleccionar el número de grupos es maximizar el **BIC**[8]. Si sustituimos la expresión de la verosimilitud en el máximo de la mezcla de normales en la expresión del **BIC** y eliminando constantes, este criterio en este caso equivale a:

$$\widehat{G} = \max_{g}((2 * L(G)) - r\ln(n)) \tag{3.23}$$

[7]Este algoritmo esta disponible en el Software **R**

[8]Por sus siglas en ingles **B**ayesian **I**nformation **C**riterion

Donde $L(G)$ es la log-verosimilitud del mejor modelo de mezcla con G componentes, r es el número de parámetros del modelo y n es el número de observaciones.

Podemos llevar acabo la obtención de varios modelos con G componentes por medio del algoritmo **EM** y entonces escoger el que maximice el **BIC**. Pero este procedimiento puede resultar largo y lento. Una sugerencia que propone Fraley y Raftery es usar una aproximación jerárquica: encontrar un modelo con $G-1$ componentes fusionando dos grupos del modelo con G componentes para el cual la fusión conduce a tener un decremento muy pequeño en la verosimilitud. Entre esta secuencia de pasos es como se elige el que maximice el **BIC**.

Importancia del criterio BIC

La importancia del criterio **BIC** radica en que este trata de seleccionar el modelo más adecuado, con máxima probabilidad *a posteriori*, y puede demostrarse que es un criterio consistente, de manera que la probabilidad de seleccionar el modelo correcto tiende a 1 cuando el tamaño de muestra $n \to \infty$.

Es así como damos paso al siguiente capítulo de este trabajo donde se da la teoría y procedimientos que utiliza la técnica que se está proponiendo en este trabajo. El objetivo del mismo es dar a conocer y mostrar en que casos es de utilidad el uso de esta técnica como una herramienta de decisión que amplíe un poco más el criterio al momento de llevar acabo la formación de conglomerados y por tanto obtener una aproximación del número real de grupos existentes en un conjunto de datos u observaciones.

Capítulo 4

Evaluación del grado de separación entre conglomerados

Como ya se revisó en la sección anterior, la razón por la cual se aborda con más detalle la formación de conglomerados mediante los modelos de mezclas es que los componentes de la mezcla en el modelo no necesariamente corresponden a los distintos grupos en el conjunto de datos. Si los grupos se comportan como una distribución normal multivariada entonces, los conglomerados resultantes verdaderamente tenderán a ser "*distintos*" en el sentido común de la palabra, es decir, serán contiguos, densos y con áreas relativamente vacías entre ellos y no se traslaparán elementos de grupos cercanos. Sin embargo, si los grupos no se comportan como tal, la correspondencia entre los componentes de un modelo de mezclas y los grupos en el conjunto de datos puede verse estropeada. Un grupo aislado con una distribución no elíptica, por ejemplo, puede ser modelado no solo por uno, sino por varios componentes de una mezcla, es decir, puede llegar a existir un traslape de los datos y los correspondientes conglomerados que se proponen en el modelo serán distintos a los reales. Es uno de los problemas más comunes que se puede llegar a encontrar el investigador al momento de modelar los grupos de una población, que haya áreas entre los conglomerados que tengan una gran cantidad de puntos en donde prácticamente la probabilidad de clasificación de un elemento en uno u otro conglomerado es igual. Ese es el propósito de este capítulo y trabajo el dar algunas herramientas para poder evaluar si existe o no este problema y como poder solucionarlo u obtener un modelo más adecuado que nos describa mejor los grupos de nuestra población, según lo proponen los autores *Tantrum J.*, *Murua A.* y *Wener S.*[2].

4.1. Separación entre los componentes de una mezcla

En este capítulo se abordará el problema de la evaluación del grado de separación que existe entre los componentes de un modelo de mezclas ajustado a un grupo de datos, para así poder tomar una mejor decisión en cuanto al número final de grupos que existen. En las siguientes secciones de este capítulo, mostraremos tres métodos de evaluación de la separación entre componentes de una mezcla que son:

- Basados en la probabilidad a posteriori.
- Usando márgenes.
- Basados en probabilidades de clasificación errónea.

4.1.1. Probabilidades aposteriori

Estrictamente hablando, nosotros deberíamos de esperar componentes de una mezcla que modelan diferentes grupos en el conjunto de datos. Es decir, los componentes de la mezcla modelan grupos distintos dentro del conjunto de datos y se espera que no exista alguna superposición con algún otro grupo cercano. Pero cuando los grupos no presentan esta característica, se tiene una superposición entre dos o más grupos que están relativamente cercanos uno del otro.

Podemos llevar este problema a términos de probabilidad. Asumimos que ya hemos modelado la distribución de los datos observados mediante la función de densidad de una mezcla de distribuciones dada por (3.1). Podemos generar observaciones de esta función de densidad primeramente generando una variable componente de etiquetas $\mathbf{z}$ con $P(\mathbf{z} = k) = \pi_k$, y luego generando $\mathbf{x}$ de f_z.

Sabemos que por el teorema de Bayes, la probabilidad a posteriori de $P(\mathbf{z} = k|X)$ es

$$P(\mathbf{z} = k|X) = \frac{\pi_k f_k(X)}{\sum_{i=1}^{G} \pi_i f_i(X)} \tag{4.1}$$

Entonces el componente k de la mezcla está bien separado del resto de los componentes si $P(\mathbf{z} = k|X)$ sólo toma valores extremos, cercanos a 0 o a 1 donde 1 es para observaciones generadas por el componente k, y 0 en cualquier otro caso.

La evaluación exacta de las distribuciones de $P(\mathbf{z} = k|X)$ para los G componentes es generalmente imposible de realizar cuando el número de componentes crece. Para ver porque sucede esto, definimos la variable aleatoria $h(X) = P(\mathbf{z} = k|X)$. Y su función de distribución $F_h(u)$ está dada por.

$$F_h(u) = P(h(X) \leq u) = \int I(h(x) \leq u) f(x) dx \tag{4.2}$$

Donde $I(\cdot)$ denota la función indicadora. Excepto en casos triviales (cuando $G = 2, \Sigma_1 = \Sigma_2$) la región definida por la función indicadora que tiene una forma muy compleja descrita en términos secciones cónicas. Por lo tanto en general esta integral no puede ser evaluada con métodos directos y se recurre a la simulación por el método de simulación Monte Carlo.

4.1.2. Evaluación de la separación usando márgenes

Una forma alternativa de ver a las probabilidades a posteriori es considerando los márgenes. Sea $\widehat{\mathbf{z}}(X)$ la etiqueta estimada para X mediante la regla de Bayes, tenemos que:

$$\widehat{\mathbf{z}}(X) = \arg \max_k P(\mathbf{z} = k|X)$$

El margen de X obtenido del componente $\mathbf{z}$ del modelo es el siguiente:

$$margen(X, \mathbf{z}) = P(\widehat{\mathbf{z}}(X) = \mathbf{z}|\mathbf{z}) - \max_{k \neq Y} P(\widehat{\mathbf{z}}(X) = k|\mathbf{z})$$

Nótese que un margen negativo significa que $\mathbf{x}$ es asignado erróneamente a un componente, y un margen pequeño significa que X se sitúa en una región donde los componentes se sobreponen considerablemente.

4.1.3. Probabilidades de clasificación errónea

Cuando el número de conglomerados es moderado, podemos recurrir a la matriz de clasificación errónea como una herramienta para detectar componentes bien separados así como los sobrepuestos de un modelo de mezclas. Esta matriz la definimos de la siguiente manera:

Definición 33.- .- Sea $m_{gg'}$ la probabilidad que es asignada mediante la regla de Bayes a una observación que viene del componente g al componente g'.

Entonces la matriz de clasificación errónea es

$$M = \left(\begin{array}{cccc|c|c} m_{11} & m_{12} & \cdots & m_{1G} & 1-m_{11} & \pi_1 \\ m_{21} & m_{22} & \cdots & m_{2G} & 1-m_{22} & \pi_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ m_{G1} & m_{G2} & \cdots & m_{GG} & 1-m_{GG} & \pi_G \end{array}\right) \tag{4.3}$$

De donde m_{gg} es la probabilidad de clasificar la observación del grupo g correctamente en ese mismo grupo, $1-m_{gg}$ es la probabilidad de clasificación errónea, es decir, la probabilidad de que una observación del grupo g sea clasificada erróneamente en cualquier otro grupo g', y π_g es la probabilidad de pertenecer al grupo g.

Podemos extraer información importante directamente de la matriz de clasificación errónea en tres diferentes formas. En la primera podemos extraer la *probabilidad total de clasificación errónea* que definimos como:

$$P_{ce} = \sum_{g=1}^{G} \pi_g(1 - m_{gg}) \tag{4.4}$$

Esta probabilidad es una medida global de incertidumbre lo que hacemos es ponderar las probabilidades de pertenecer al grupo g-ésimo, es decir π_g por la probabilidad $(1 - m_{gg})$. Por ejemplo, si la probabilidad de clasificación errónea es cercana es diferente de 0, entonces se esta disminuyendo la probabilidad π_g, y dado que la suma de estas probabilidades es uno, cuando se ponderan por $1 - m_{gg}$ entonces esta suma ya no da 1 como resultado y esto nos esta indicando que algunos datos se están traslapando con otros provenientes de algún otro grupo. Entonces podemos concluir que mientras más bajas sea la probabilidad $(1 - m_{gg})$, mejor será la separación entre los componentes de la mezcla.

En otra forma, podemos observar las probabilidades de clasificación errónea del g-ésimo *componente* que denotamos por MC_g y que la definiremos como:

$$MC_g = \sum_{i \neq g}^{G-1} m_{gi} = (1 - m_{gg}) \tag{4.5}$$

Notemos que cuando esta probabilidad es relativamente alta, nos dice que se está teniendo una probabilidad considerable de llevar acabo una clasificación errónea de una observación que originalmente viene del grupo g a cualquier otro de los $G - 1$ grupos existentes, y en el caso contrario, si esta probabilidad es cercana a 0 quiere decir que la probabilidad de cometer un error de clasificación es prácticamente nula y los conglomerados asociados a los grupos corresponden casi perfectamente y no existe ningún traslape entre

los conglomerados del modelo. También será útil graficar estas probabilidades contra cada MC_g para así tener una herramienta gráfica más para tener otro criterio sobre si existe una clasificación errónea en el conjunto de datos.

La última forma en la que podemos extraer información de la matriz de clasificación errónea M es que los valores $m_{gg'}$ y $m_{g'g}$ indican que otro componente se traslapa con el componente g.

Estas herramientas propuestas para evaluar la separación entre los componentes de un modelo basado en mezclas son el primer paso para poder ahora llevar acabo la evaluación entre los conglomerados que se encuentran en el conjunto de datos que estamos estudiando. Describiremos a continuación que es lo que podemos hacer para evaluar la separación entre conglomerados.

4.2. Separación entre conglomerados

Un modelo basado en mezclas es sólo un estimador para la distribución real del conjunto de los datos. Por lo tanto el grado de separación entre los componentes de la mezcla (o la ausencia de esta) no siempre refleja exactamente la separación real entre los conglomerados.

No podemos calcular la matriz de probabilidades de clasificación errónea de los datos observados $(x_1, \ldots, x_n)$ ni los márgenes, porque estos requieren el conocimiento de las etiquetas verdaderas. Sin embargo, podemos calcular las probabilidades aposteriori $P(\mathbf{z} = k|x_i)$ y por tanto generar una gráfica llamada *Rootograma* que es una variante de un histograma, donde las alturas de las barras representan la raíz cuadrada de las frecuencias en cada categoría. Si el rootograma se carga hacia el extremo derecho, es decir, se tiene una barra de frecuencia muy alta en la categoría del número uno, quiere decir que la probabilidad de una mala clasificación en ese grupo es nula y la gran mayoría de los elementos pertenecientes a ese grupo están siendo correctamente clasificados en ese grupo por el modelo. Si sucede que las barras tienen una forma uniforme en las alturas de las frecuencias o las frecuencias son muy altas en 0, esto quiere decir que un número considerable de elementos que se han clasificado en ese grupo no pertenecen realmente a él, lo que nos indica que existe un traslape en los datos o que haya regiones entre conglomerados que no estén relativamente vacías y que para el modelo propuesto, clasificar a alguno de los elementos entre uno o otro de los conglomerados contiguos da prácticamente lo mismo. Esta gráfica hace más visibles los conteos bajos para tener una mejor perspectiva visual.

4.3. Conglomerados híbridos

Los métodos de obtención de conglomerados mediante modelos de mezclas, generan un jerarquía de modelos de mezclas. Es decir, el modelo con $m-1$ componentes en la mezcla, es obtenido fusionando dos conglomerados del modelos con m componentes para el cual, el cambio lleva al decremento mas pequeño en la log-verosimilitud. El resultado de este proceso de fusión, puede ser representado por un árbol binario T donde las hojas del árbol son las observaciones. A cada nodo N del árbol se le asignará una *generación* o nivel entre 1 y $n-1$ que indicará en que punto del proceso de fusión fue generado. El nodo interior corresponde a la i-ésima fusión y en la secuencia se le asigna la generación o nivel $n-i$; el nodo raíz tiene nivel 1. Cada nodo N es también asociado al conglomerado formado por las hojas descendentes.

La secuencia de fusión define una secuencia de arboles, T_n es obtenido de T removiendo la descendencia de todos los nodos con nivel mayor o igual a m. Por construcción, T_m tiene m hojas y corresponde al modelo de mezclas con m componentes mezcla. Entonces, si G es el número de componentes mezcla obtenido por el criterio **BIC**, definimos T_G el árbol correspondiente.

Si los distintos grupos en el conjunto de datos siguen una distribución normal, entonces esperaríamos *rigurosamente*[1] una correspondencia uno a uno entre los grupos y los componentes de la mezcla asociado al árbol T_G. También, los conglomerados asociados con las hojas de T_G serán similares a los del grupo. Si los grupos no tienen una distribución normal, cada grupo puede ser modelado por más de un componente mezcla, y en consecuencia puede ser el resultado de la unión de varios conglomerados.

La idea de los conglomerados híbridos es probar, para cada nodo de T_G el cual tiene hojas y aunque los correspondientes conglomerados estén bien separados, puede haber evidencia estadística suficiente para suponer que no lo están. Si no están bien separados, entonces los conglomerados probablemente correspondan al mismo grupo y entonces se debe fusionarlos. El nuevo conglomerado es entonces modelado por la suma de los componentes de la mezcla "modelando las hojas " que fueron "podadas". Este proceso de "*podado*" es repetido hasta que ningún conglomerado pueda ser fusionado.

4.4. Prueba de unimodalidad

Una herramienta más que tiene el algoritmo propuesto en la sección anterior, es la herramienta que nos ayuda a determinar si existe evidencia estadística suficiente para suponer que hay unimodalidad en los datos proyectados

[1] *Rigurosamente* por que G, después de todo sólo es un valor estimado.

en la dirección del discriminante de Fisher[2] provenientes de dos componentes de la mezcla, esto se traduce en que dos poblaciones bien separadas y que siguen una distribución normal multivariada, pueden ser modeladas por una mezcla de dos normales multivariadas y para este caso la función de densidad de la mezcla tendría dos modas. Pero si estas dos modas están significativamente juntas, entonces se podría suponer que en realidad la función de densidad no tiene dos, sino simplemente una moda. Para esto, necesitamos una forma de medir ¿cuánta evidencia de unimodalidad existe en un conjunto de datos?

Para probar si existe evidencia estadística suficiente para suponer que se tiene una distribución unimodal hay que realizar una prueba de hipótesis la cual mostramos a continuación:

Definición 34.- Sean $\mathbf{x}_1, \mathbf{x}_, \ldots, \mathbf{x}_n$ una muestra aleatoria (univariada) de $f(\mathbf{x})$, y sea $F_n(\mathbf{x})$ la distribución acumulada empírica de la muestra. Para probar la hipótesis nula $H_0 : f(\mathbf{x}) = H(\mathbf{x})$, de donde $H(\mathbf{x})$ es la mejor distribución unimodal, usaremos la estadística de prueba **DIP**[6] dada por la expresión:

$$D = sup_x|F_n(\mathbf{x}) - H(\mathbf{x})| \tag{4.6}$$

De donde $H(\mathbf{x})$ es la función de distribución acumulada mejor aproximada a $F_n(\mathbf{x})$

4.5. Implementación del algoritmo

Ahora daremos los pasos que emplea la formación de conglomerados híbridos y la evaluación de ¿qué tan bueno es el modelo? Esto es dado que, como ya se ha dicho antes, puede que un grupo no tenga una distribución normal y por lo tanto necesite de más de un componente para poder describir su forma, esto implica que también pueda existir un traslape de los datos de un grupo con otro. Esto se puede traducir en que puede ser llevada a cabo una fusión de dos conglomerados si es que se encuentran muy juntos, ya que la existencia de dos conglomerados muy juntos y con un traslape significativo quiere decir que existe una alta probabilidad de clasificación errónea en otro grupo que no corresponde al grupo real al cual pertenece la observación. La razón por la cual se le denominan formación de conglomerados híbridos es que se utilizan varias herramientas y técnicas que hacen posible la evaluación de la viabilidad del modelo obtenido mediante los modelos basados en mezclas de distribuciones. Esta técnica propone los siguientes pasos para la evaluación del grado de separación entre los componentes:

[2]Esto se debe a que es la mejor dirección que separa a los datos

1. Obtener el modelo de mezclas de distribuciones normales que maximice el BIC.
2. Obtener las probabilidades de posteriores mediante la regla de Bayes de cada observación.
3. Graficar los rootogramas de las probabilidades posteriores de las observaciones.
4. Calcular la matriz de clasificación errónea.
5. Obtener la estadística $\sum_{g=1}^{G} \pi_g(1 - m_{gg})$ para tener una medida de separación entre los componentes de la mezcla.
6. Graficar las probabilidades de clasificación errónea de cada componente MC_g.
7. Obtener la función de distribución empírica $F_n(\mathbf{x})$ para cada par de componentes candidatos a fusionarse.
8. Obtener la función unimodal $H(\mathbf{x})$ que mejor se aproxime a $F_n(\mathbf{x})$.
9. Llevar acabo el proceso de fusión de los nodos para obtener el árbol T_m.
10. Realizar una prueba de hipótesis de unimodalidad para la distribución $F_h(u) = P(h(X) \leq u)$ y rechazar H_0 con nivel de significancia α si $sup_x|F_n(\mathbf{x}) - H(\mathbf{x})| > q$ donde q es el cuantil que aparece en la tabla *qDiptab* del paquete *diptest*.
11. Repetir los pasos 7 a 11 hasta que no ya no pueda ser fusionado conglomerado alguno.
12. Generar el árbol binario T asociado al modelo.

4.5.1. Comentarios

El método de los conglomerados híbridos esta basado en que los grupos existentes pueden corresponder a una colección de mezclas de distribuciones. El propósito del método aquí descrito, es el de identificar esa colección. No es propósito de este el mejorar el ajuste del modelo de mezclas obtenido mediante el criterio **BIC**. El algoritmo de podado requiere que el nivel de significancia para las pruebas **DIP** sea especificado, mientras más grande sea la significancia, más hojas serán podadas del árbol jerárquico T_G. El nivel

de significancia no debería ser tomado tan a la ligera, ya que el total de procesos de poda no constituye un nivel α de significancia para la prueba de unimodalidad para el modelo multivariado.

Primero, existe un problema de multiplicidad. Si nosotros estamos llevando a cabo varias pruebas con nivel α, entonces la probabilidad errónea de rechazo de una o mas de las hipótesis nulas es más grande que α.

Segundo, estamos escogiendo la dirección de la proyección que maximiza la separación entre los conglomerados. Esto se convierte en un problema de dimensión del espacio donde se encuentran los dos conglomerados los cuales están siendo considerados. Por ejemplo, si nosotros tenemos un total de $n+1$ observaciones en un espacio de dimensión n, entonces siempre existirán direcciones para las cuales las observaciones en los dos conglomerados que estamos tomando en cuenta, se proyecten en exactamente en un mismo punto puntos, es decir un dato proyectado en esa dirección puede corresponder a uno o mas datos. Nosotros tratamos con este problema primero, proyectando las observaciones provenientes de dos conglomerados en sus k primeras componentes principales y luego encontrando la dirección del discriminante de Fisher en este subespacio de dimensión menor. Nosotros escogemos k como un tercio del total del número de observaciones en los dos conglomerados que se proyectarán.

Una vez expuesta la teoría que engloba este algoritmo que nos sirve para aproximar el número final de conglomerados en un conjunto de datos obtenidos mediante modelos de mezclas, es momento de llevar acabo los ejemplos correspondientes para ver que tan bueno resulta este en un conjunto de datos propuesto para su análisis. Llevaremos acabo una clasificación de las observaciones, formando los correspondientes conglomerados y obteniendo una aproximación al número real de grupos existentes en el conjunto de datos. Veremos que tan buena información nos dan las estadísticas propuestas en este capítulo, así como las gráficas y la prueba de hipótesis propuesta para saber si un conglomerado es modelado por mas de un componente de una mezcla o si esta bien modelado por un componente y a su vez esta bien separado de los $G-1$ componentes restantes.

Capítulo 5

Aplicaciones

Toda herramienta matemática, tiene como propósito el ser llevada a la aplicación y en este caso, la herramienta de diagnóstico propuesta en este trabajo tiene también como propósito ser aplicada a una base de datos obtenida mediante un experimento, simulación o recopilación de datos en un estudio estadístico. Ya hemos dado la teoría y fundamentos en los que esta basada esta técnica, así como en las situaciones en las que puede ser aplicada como herramienta de diagnóstico o de toma de decisión al momento de realizar un proceso de clasificación no supervisada, es decir, la formación y obtención de un determinado número de grupos en una población de interés.

5.1. Un primer análisis

Para este capítulo se tomará como información la matriz de datos $\mathbf{X}$ con datos provenientes de algunas simulaciones realizadas en el software $\mathbf{R}$[1]. Los datos de este nuestro ejemplo 1, fueron obtenidos de una simulación consistente en dos distribuciones normales $N(-15, 4)$ y dos distribuciones exponenciales $\exp(.2)$. Es decir, se generaron 70 observaciones análogamente al ejemplo del capítulo 3 y las coordenadas (x_1, y_1) y (x_2, y_2).

- $x_1 \sim N(-15, 4)$
- $y_1 \sim N(-15, 4)$
- $x_2 \sim \exp(.2)$
- $y_2 \sim \exp(.2)$

[1]La sintaxis utilizada para la obtención de los datos estará disponible en la sección de apéndices.

Esta simulación realizada dió como resultado los datos mostrados en la figura 5.1:

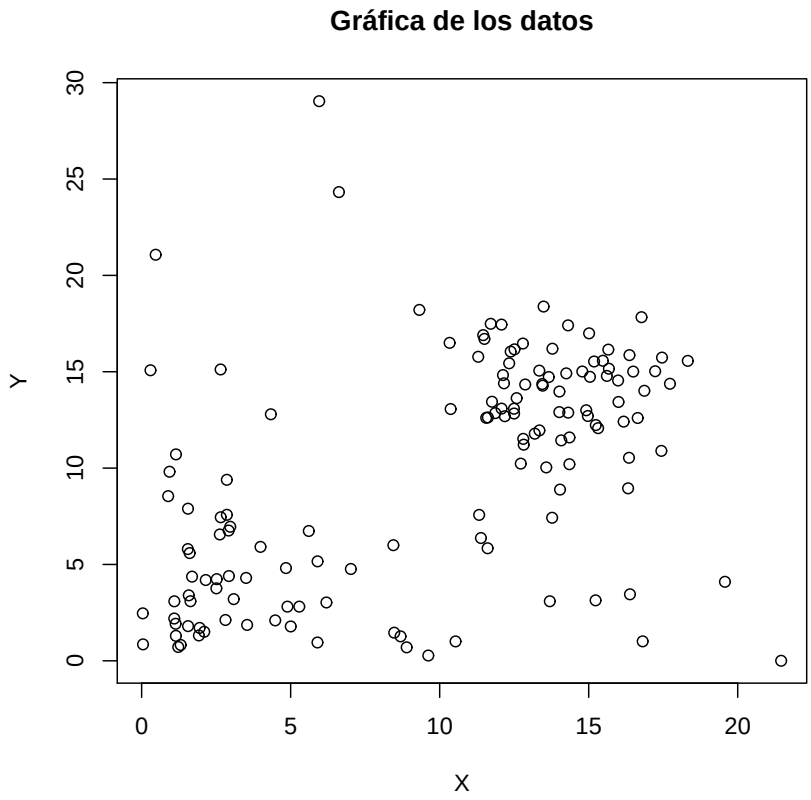

Figura 5.1: Datos simulados del ejemplo

Es muy probable que con esta gráfica sólo se puedan apreciar bien dos conglomerados ya que existen dos nubes de puntos que parecen estar muy cercanos entre ellos, existe una relativa área vacía entre ellos y también existen algunos puntos muy lejanos a los cuales seria difícil poder clasificar o pueden ser considerados como puntos discrepantes.

Podemos comenzar el análisis de estos datos aplicando los métodos clásicos de formación de conglomerados compararlos entre ellos y ver si son muy diferentes los resultados a los que se llegan con la aplicación de cada uno, que si recordamos, como en el capitulo 3, también se simularon datos a modo que todos los métodos coincidieran para concluir que en esa población existían 2 grupos.

Los métodos que se emplearán en esta primera formación de conglomerados consiste en:

- Método de encadenamiento sencillo
- Método de encadenamiento completo
- Método de encadenamiento medio
- Algoritmo de K-medias
- Modelos basados en mezclas de distribuciones normales

Este último método es el paso inicial para la aplicación de todos los pasos que se propusieron en el capítulo anterior para la evaluación y diagnóstico de la clasificación obtenida mediante el mejor modelo basado en mezclas escogido mediante el criterio de información bayesiana (BIC).

Mostraremos en la siguiente gráfica los 3 primeros dendogramas correspondientes a las clasificaciones obtenidas mediantes los primeros 4 algoritmos para nuestro primer ejemplo.

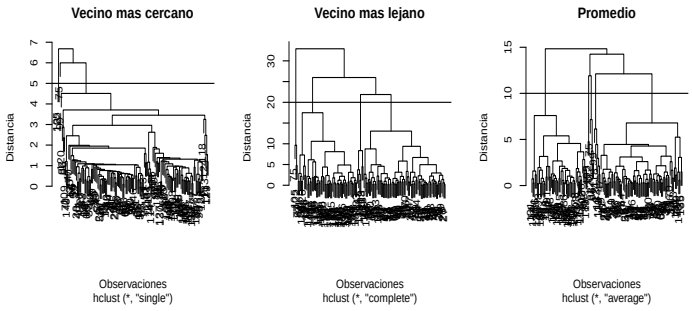

Figura 5.2: Clasificaciones obtenidas para el ejemplo

En el primer dendograma que corresponde al vecino más cercano, podemos ver que es muy difícil establecer un número final de conglomerados obtenidos mediante este método, ya que si cortamos a cierta altura el dendograma por ejemplo, a la altura 5, se tendrían 2 grupos, pero uno de ellos consta solo dos elementos y en el otro se concentraría la mayoría de los datos. Es decir, aquí se presenta una de las desventajas de aplicar este método. Puede formar grupos muy grandes y por lo tanto, en este ejemplo considera que casi todas las observaciones pertenecen a un solo grupo, esto se traduce que para fines prácticos este método no es muy bueno para la obtención del número final y la formación de conglomerados en este caso.

Si nosotros utilizáramos el método del vecino más lejano entonces obtendríamos el segundo dendograma. Aplicando el mismo criterio de el corte del dendograma pero ahora a la altura de 10, este nos sugiere que hay aproximadamente 4 grupos en los datos y si nos fijamos en el gráfico original de los datos, la clasificación asociada a este método con dos grupos concuerda un poco mas con la vista previa que nos ofrece la gráfica sobre los datos.

Ahora, si nosotros aplicamos el método del promedio de ambos, obtenemos el tercer dendograma y aplicando el mismo criterio, pero ahora realizando el corte a la altura de 10, este nos sugiere que hay aproximadamente 4 grupos en los datos y por lo tanto da otra clasificación diferente a los dos anteriores.

En conclusión, estos 3 primeros métodos dan 2 resultados distintos para un mismo grupo de datos, es decir, un método da un resultado con un número final de 2 conglomerados y los métodos restantes nos dan un resultado final de 4 conglomerados con una clasificación aunque no igual, si bastante parecida. En el siguiente gráfico de mosaico mostramos la clasificación correspondiente a cada método.

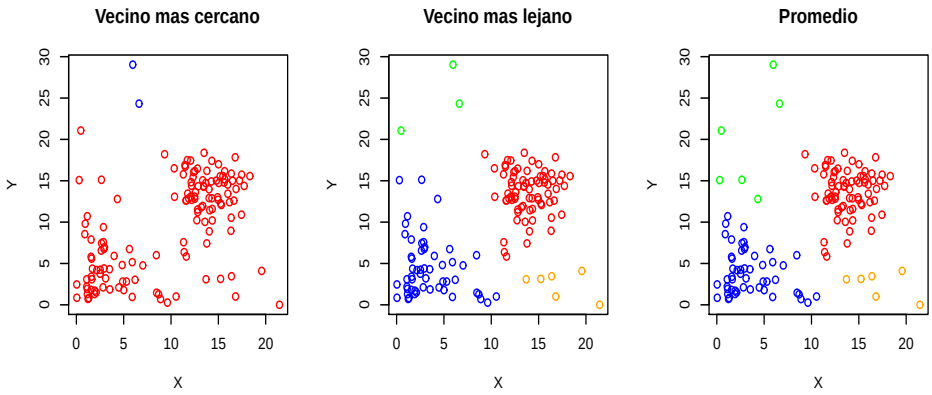

Figura 5.3: Clasificación para el ejemplo 1

En el caso que nosotros queramos aplicar el algoritmo de K-medias, suponemos $K = 4$ como sugieren dos de los 3 métodos anteriores y veamos cual es la clasificación obtenida mediante K-medias para los datos de nuestro ejemplo es la figura 5.4.

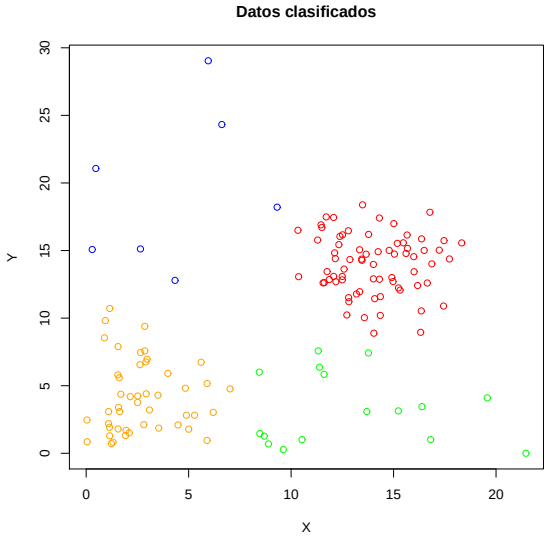

Figura 5.4: K-medias para el ejemplo

Estas clasificaciones obtenidas mediante encadenamiento completo, medio y K-medias son muy parecidas y al parecer tienen buenos resultados y coherencia con los datos ya que si es notorio que existen dos grupos identificables en la primera gráfica de los datos, pero existe una nube de puntos que al parecer no siguen una forma normal y ahí es donde se presentan los problemas de obtención de grupos mediante estos métodos del análisis de conglomerados.

Ahora como último método de análisis preliminar que aplicaremos al conjunto de los datos es el método de los modelos basados en mezclas. El modelo propuesto mediante el software **R** es el calculado mediante el paquete **mclust** que tiene que ser instalado desde cualquier espejo **CRAN** ya que en este paquete están la mayoría de las rutinas utilizadas para la modelación de conglomerados mediante mezclas de distribuciones.

El modelo de mezclas óptimo obtenido en **R** tiene 4 componentes que son normales bivariadas, con diferentes matrices de covarianzas muestrales y orientaciones cuyos vectores de medias muestrales son:

	X	Y
μ_1	13.965242	13.981728
μ_2	8.821011	3.326189
μ_3	2.005687	3.870258
μ_4	2.552594	15.144646

Cuadro 5.1: Vectores de medias del ejemplo 1

Y las matrices de varianzas y covarianzas de cada componente se muestran a continuación

$$\Sigma_1 = \begin{pmatrix} 3.8979942 & -0.4700239 \\ -0.4700239 & 5.2684712 \end{pmatrix}$$

$$\Sigma_2 = \begin{pmatrix} 27.019524 & -1.154484 \\ -1.154484 & 4.850281 \end{pmatrix}$$

$$\Sigma_3 = \begin{pmatrix} 0.7442849 & 1.119023 \\ 1.1190226 & 5.565788 \end{pmatrix}$$

$$\Sigma_4 = \begin{pmatrix} 4.929051 & 9.94526 \\ 9.945261 & 51.13972 \end{pmatrix}$$

Y por último tenemos las probabilidades de pertenecer a cada grupo, es decir, los pesos de la mezcla π_g que son:

$$(\pi_1, \pi_2, \pi_3, \pi_4) = (0.51908341,\ 0.21311127,\ 0.19591163,\ 0.07189368)$$

La clasificación que nos da este modelo de 4 componentes normales bivariadas que se muestra en la figura 5.5:

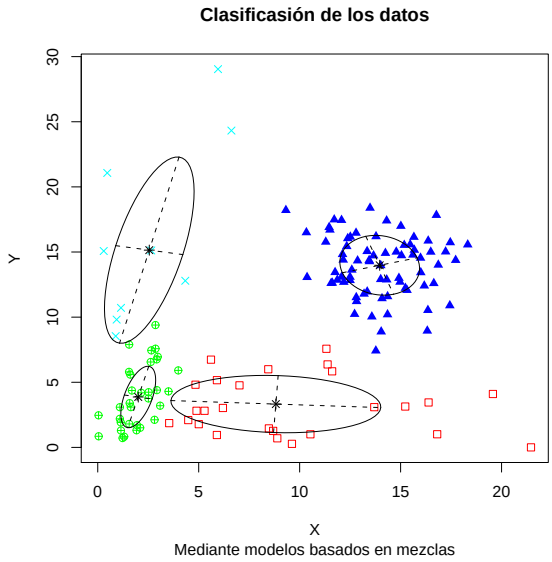

Figura 5.5: Modelos basados en mezclas para el ejemplo

La razón por la que el número de componentes G es igual a 4 se debe a que el **BIC** se maximiza para este modelo **VVV** en $G = 4$ como lo podemos apreciar en esta gráfica de los **BIC**'s correspondientes al número de componentes de la mezcla.

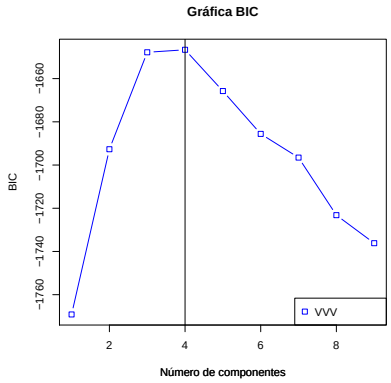

Figura 5.6: BIC para nuestro ejemplo

Podemos ver que cuando el número de componentes es 4, el **BIC** es igual a -1646.656 el cual es el valor máximo alcanzado para este modelo.

En resumen, podemos concluir que aplicando los métodos del vecino más lejano, promedio, K-medias y de mezclas, obtenemos que aproximadamente el número final de conglomerados es 4 y las clasificaciones obtenidas las pondremos en una sola gráfica para su comparación en la figura 5.7.

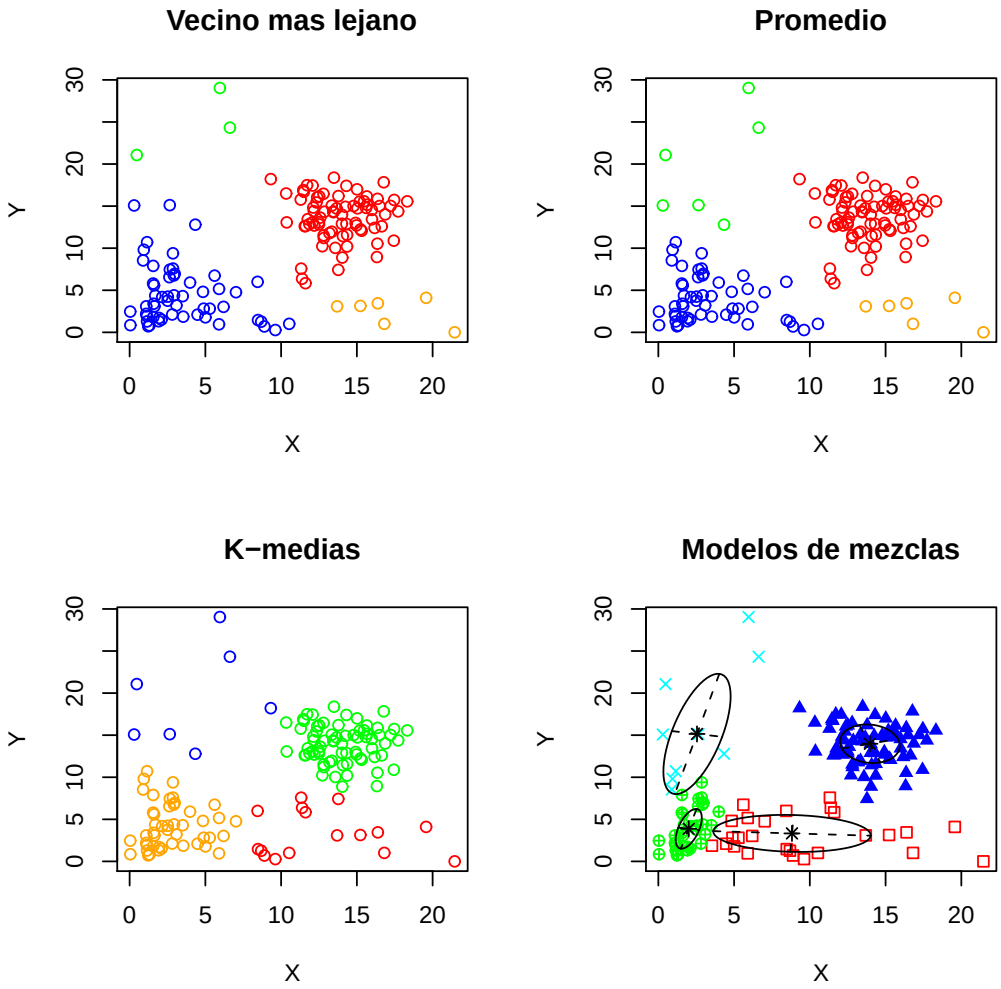

Figura 5.7: Clasificaciones para el ejemplo 1

Ahora ya que tuvimos un primer análisis de los datos con los métodos antes mencionados Es momento de poner en práctica el algoritmo antes propuesto para que podamos realizar la evaluación del número final de conglomerados del modelo de mezclas obtenido.

5.2. Implementación del algoritmo

El modelo óptimo calculado mediante **R** es en el que nos basaremos para suponer que existen 4 grupos en los datos, es decir, supondremos de aquí en adelante que $G = 4$. Una vez supuesto este número, entonces procederemos a implementar los pasos del algoritmo para saber si uno o más conglomerados no poseen una distribución normal y por tanto podrían ser modelado por dos o más componentes mezcla del modelo propuesto por **R**, es decir, si existe traslape entre uno o más componentes de la mezcla.

5.2.1. Matriz de probabilidades posteriores

Comenzaremos mencionando que la matriz de probabilidades posteriores asignadas a cada observación por la regla de Bayes no la mostraremos en esta sección ya que es una matriz de dimensiones $(140, 4)$, es decir, una matriz donde los renglones son las 140 observaciones de nuestro ejemplo y 4 columnas correspondientes a cada uno de los 4 componentes que conformar la mezcla, esta matriz se obtiene en **R** primero creando un objeto al cual se le asigna el resultado obtenido de aplicar la funcion Mclust incluida en le paquete mclust y mediante la sentencia.

```
>mezcla$z
```

5.2.2. Rootogramas

El siguiente paso es mostrar las gráficas de los rootogramas de probabilidades posteriores para cada componente de la mezcla, cada rootograma representa la raíz cuadrada de la frecuencia de cada intervalo y es una buena herramienta de diagnostico de separación de los componentes de la mezcla.

Debemos recordar que esta gráfica si está cargada hacia el 1, indica que nuestro grupo está bien separado del resto y clasificado, si está cargado hacia el cero esto quiere decir que nuestro grupo esta bien separado pero mal clasificado y si en el rootograma existen muchas frecuencias intermedias entre el 0 y el 1, esto nos quiere decir que no esta bien clasificado nuestro conglomerado y existe traslape con algún(os) componente(s) de la mezcla.

Empezaremos con el rootograma correspondiente al primer componente de la mezcla que se muestra en la figura 5.8

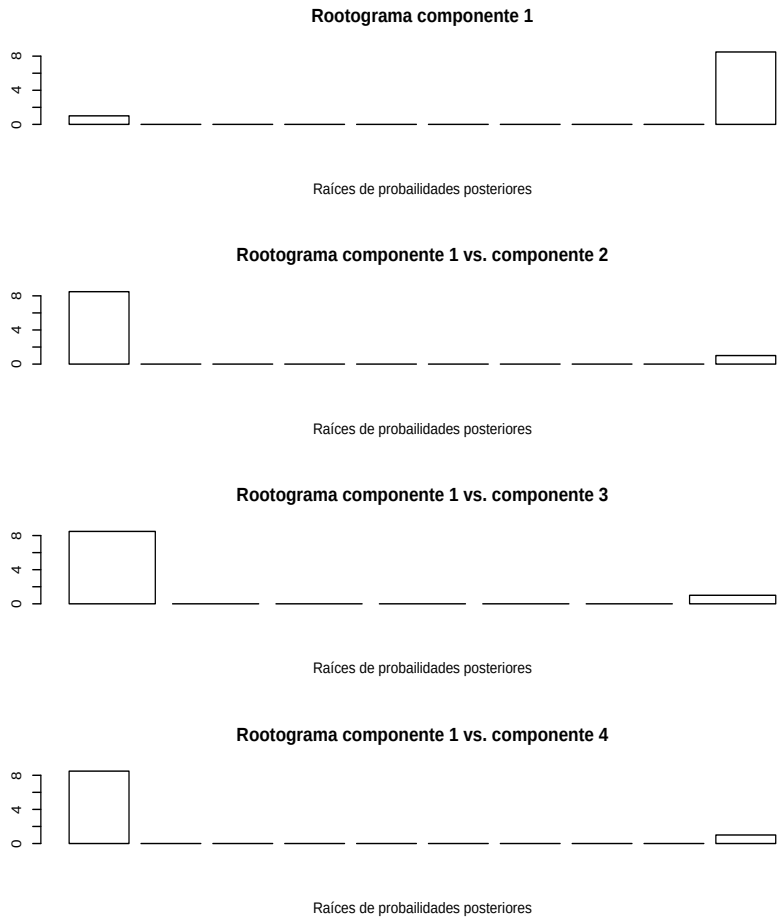

Figura 5.8: Rootograma del componente 1

Tenemos en este rootograma que en el primer renglón tenemos que las probabilidades posteriores de clasificar bien a una observación del grupo 1 en ese mismo grupo, son bastante altas y que éste está bien separado de los demás, es decir, es muy poco probable que una observación clasificada en el componente numero 1 de la mezcla, sea clasificado a cualquier otro de los 3 conglomerados restantes, lo cual indica que podemos suponer que este conglomerado sigue una distribución normal y sus elementos tienen una probabilidad posterior alta de ser bien clasificados.

Ahora mostraremos el rootograma correspondiente al componente número 2 de la mezcla en la figura 5.9.

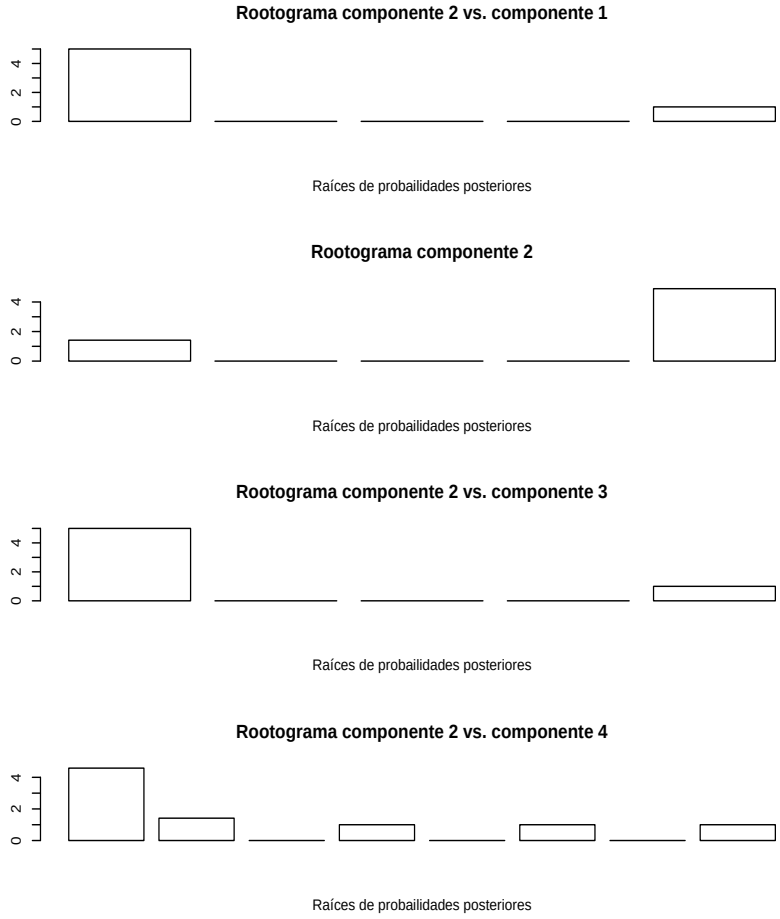

Figura 5.9: Rootograma del componente 2

Notemos que para el caso del rootograma del componente 2 de la mezcla, este nos indica que el componente esa bien separado del componente 1 de la mezcla y también está separado del componente 3, pero con respecto al componente 4, existen muchas frecuencias intermedias, esto nos esta indicando que puede existir un traslape entre estos dos componentes de la mezcla, lo cual los hace candidatos a ser fusionados en el proceso de podado del árbol binario que sera mostrado más adelante.

A continuación mostramos el rootograma correspondiente al componente 3 de la mezcla en la figura 5.10.

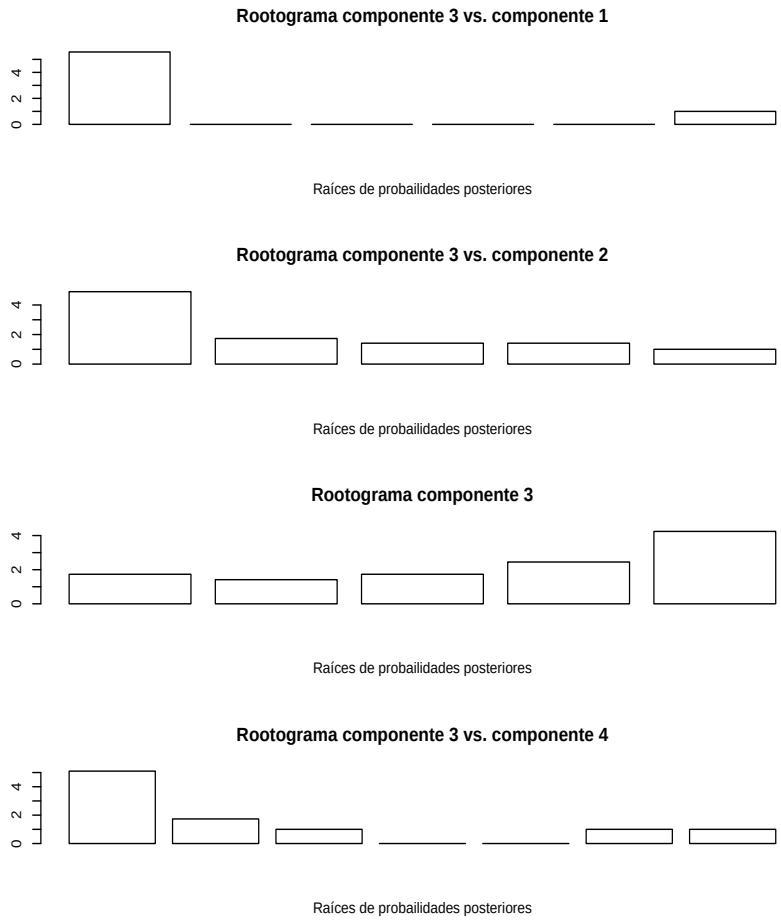

Figura 5.10: Rootograma del componente 3

Es notable para el caso del componente 3, que este componente si tiene problemas de clasificación ya que en el rootograma puede apreciarse que este componente esta bien separado del numero 1, pero existen demasiadas frecuencias intermedias en el rootograma del componente 3 y además existe traslape con el componente 2 y 4. Esto nos hace sospechar fuertemente que este componente es candidato también a ser fusionado con los componente 2 y 4 de la mezcla ya que el rootograma muestra evidencia de traslape entre los 3 componentes.

Por último mostramos el rootograma correspondiente al componente 4 número 4 de la mezcla en la figura 5.11.

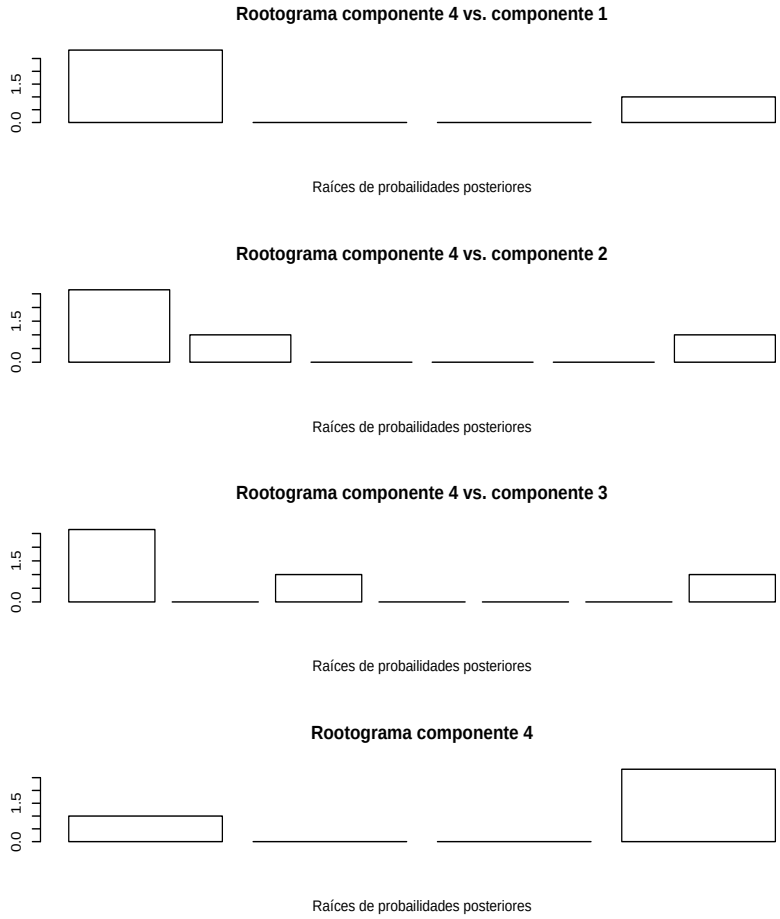

Figura 5.11: Rootograma del componente 4

Para este último caso, observamos que el componente 4 al parecer esta separado del resto, pero es posible que exista algún traslape con los componentes 2 y 3 y que existen frecuencias intermedias que así lo muestran.

Lo que podemos concluir después de haber dado el segundo paso del algoritmo de evaluación del número final de conglomerados, es que el componente 1 de la mezcla esta bien separado del resto, lo cual lo podemos ver en la gráfica de la clasificación obtenida mediante modelos de mezclas de la figura 5.5. En esta misma ya se había mencionado antes el posible traslape de los componentes restantes que serian el 2,3 y 4 de la mezcla, los cuales, con la obtención de los rootogramas se ha podido dar una prueba para poder supo-

ner que estos conglomerados no siguen una distribución normal multivariada y por tanto son candidatos a ser fusionados y así reducir el número final de conglomerados en el conjunto de datos.

5.2.3. Matriz de clasificación errónea

Como siguiente paso del algoritmo, tenemos el cálculo de la matriz de clasificación errónea, los elementos de esta matriz $m_{gg'}$ como ya se mencionó, $m_{gg'}$ es la probabilidad posterior de clasificar una observación del componente g al componente g'. Se tomó la observación del grupo g con la probabilidad posterior más pequeña porque es la que refleja que tan mal clasificada está en su respectivo componente, estas probabilidades $m_{gg'}$, son las correspondientes a la diagonal de la matriz, las demás entradas son los cruces de los componentes gg' y $g'g$.

La matriz de clasificación errónea obtenida con **R** es la siguiente:

	$C1$	$C2$	$C3$	$C4$	MC_g	Pi_g
$C1$	0.5111182	0.4888818	$8.629884e-51$	$1.697503e-12$	0.4888818	0.51908341
$C2$	$3.348326e-13$	0.7834997	0.2048378	0.01166257	0.2165003	0.21311127
$C3$	$5.891190e-09$	0.4639845	0.5056009	0.03041455	0.4943991	0.19591163
$C4$	$6.858079e-11$	0.1094927	0.06177568	0.8287316	0.1712684	0.07189368

La matriz de clasificación errónea nos da la siguiente información:

1. Para el componente 1 tenemos una probabilidad de clasificación errónea un poco alta con el componente 2, lo cual sugiere que es probable que haya traslape y para los componentes 3 y 4 esta probabilidad prácticamente es 0 indicando que esta bastante bien separado de estos dos.

2. Para el componente 2 hay una probabilidad de clasificación errónea considerable alta con los componentes 3 y 4, es decir, es probable que una observación del grupo 2 sea clasificado en alguno de estos dos componentes, pero es poco probable que una observación de este componente sea clasificado en el componente 1, lo cual nos dice que tal vez si estén bien separados estos componentes.

3. Para el componente 3 existen mayores probabilidades de clasificación erróneas con el componente 2 y 4, lo cual no dice que estos 3 componentes pueden ser fusionados más adelante, mientras que para el componente 1 la probabilidad es muy pequeña lo cual dice que pueden estar bien separados.

4. Por último para el componente 4 confirmamos que existe evidencia para suponer que hay un traslape considerable con los componentes 2 y 3 ya que estas probabilidades son algo altas mientras que la del componente 1 es casi 0.

La matriz de clasificación errónea ha dado otro punto para suponer que en el conjunto de datos si existe traslape entre los componentes de la mezcla, lo cual nos indica que uno o mas conglomerados no siguen una distribución normal multivariada y por lo tanto se requieren de mas de un componente mezcla para modelarlos y se esta teniendo un modelo con redundancia de parámetros.

5.2.4. Probabilidad total de clasificación errónea

Como siguiente paso del algoritmo, toca realizar el cálculo de la probabilidad total de clasificación errónea que se definió en (4.4). El resultado obtenido fue:

$$P_{ce} = \sum_{g=1}^{4} \pi_g(1 - m_{gg}) = 0.4090807$$

El resultado para nuestro ejemplo es que tenemos aproximadamente un 40 % de probabilidad de clasificar mal una obsevacion en cualquiera de los componentes.

5.2.5. Probabilidades de clasificación errónea de cada componente

Otro paso a seguir es el graficar las probabilidades de clasificación errónea para cada uno de los componentes de la mezcla, estas probabilidades las definimos en (4.5) y estas están dadas en la penúltima columna de la matriz de clasificación errónea y la gráfica sería la siguiente:

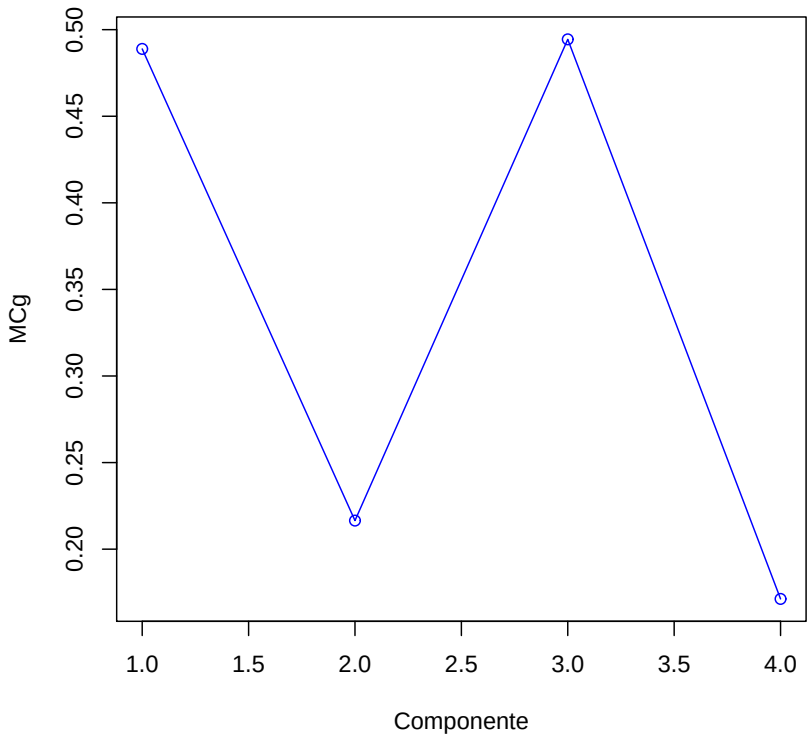

Figura 5.12: Probabilidades MCg por componente

5.2.6. Prueba de unimodalidad

Para realizar la prueba de unimodalidad se deben de realizar antes las proyecciones ortogonales de los datos originales en la dirección del discriminante lineal de Fisher[2]. Estas tienen que hacerse para cada par de componentes de la mezcla para después realizar la prueba de unimodalidad y rechazar o no rechazar la hipótesis nula de unimodalidad para cada caso. Para nuestro caso tendremos 7 proyecciones.

[2]Véase Peña D. *Análisis de datos multivariantes*[9]

La proyección para los componentes 1 y 2 se muestra en la figura 5.13:

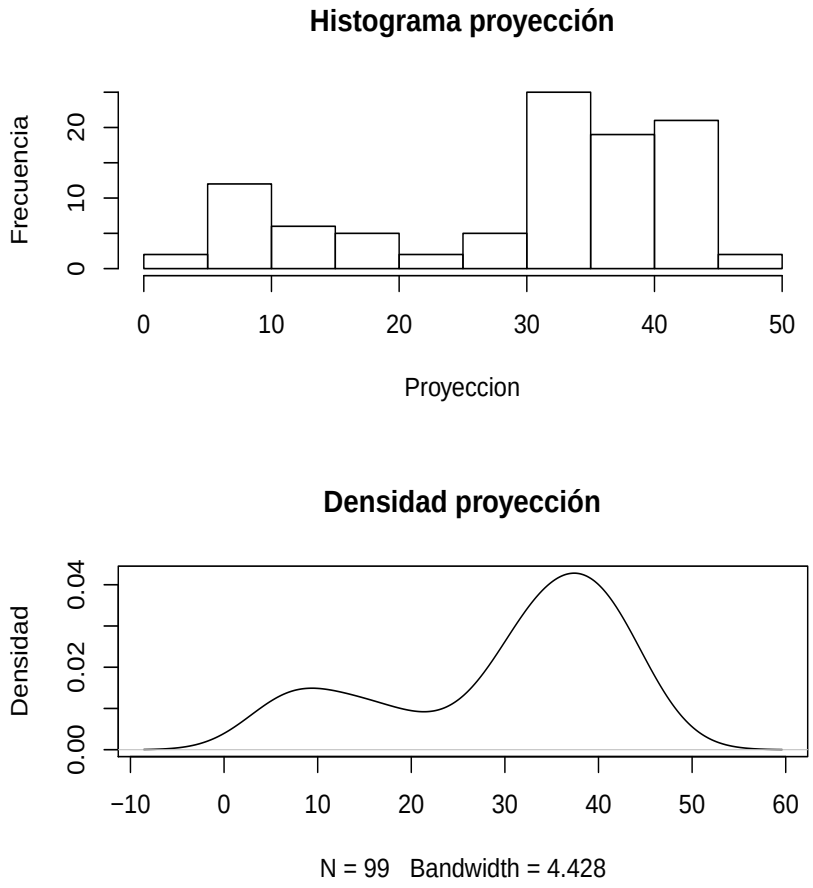

Figura 5.13: Histograma y densidad para componentes 1 y 2

En este par de gráficas podemos ver un comportamiento bimodal, lo cual nos hace pensar que estos componentes están bien separados y es muy probable que se rechace la hipótesis nula de unimodalidad.

Para los componentes 1 y 3 tenemos la siguiente proyección de la figura 5.14:

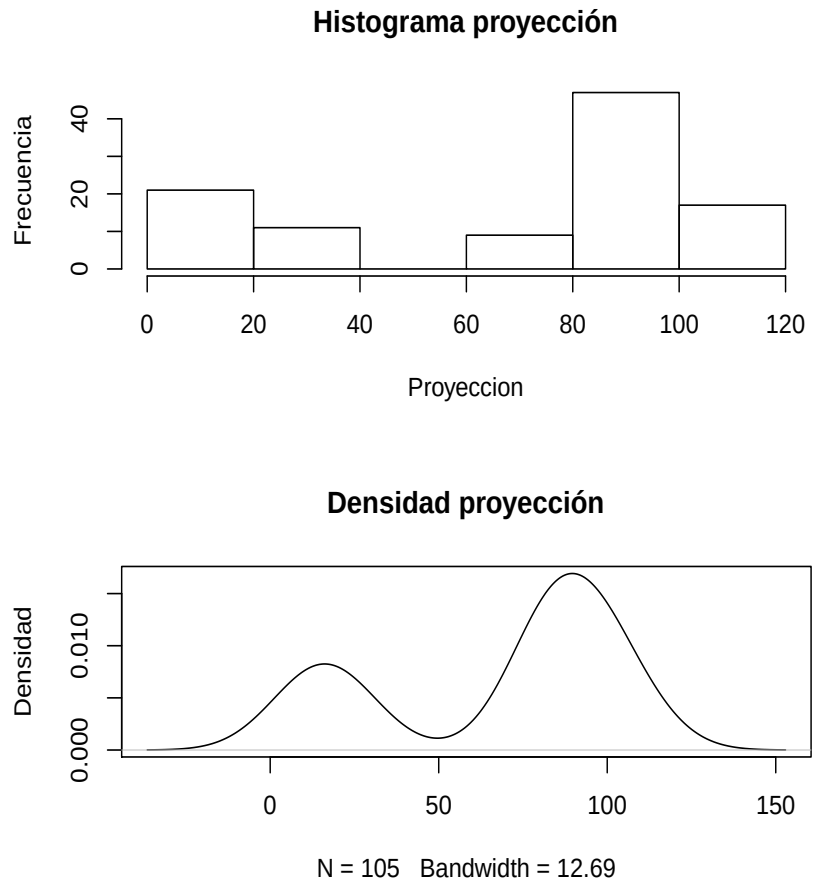

Figura 5.14: Histograma y densidad para componentes 1 y 3

Aquí también podemos observar que los componentes 1 y 3 al parecer si tienen una buena separación ya que las gráficas muestran un comportamiento bimodal de los datos proyectados para ambos componentes.

Esto nos dice que es muy probable que sea rechazada la hipótesis nula de unimodalidad para este par de componentes.

La proyección correspondiente para los componentes 1 y 4 es la de la figura 5.15:

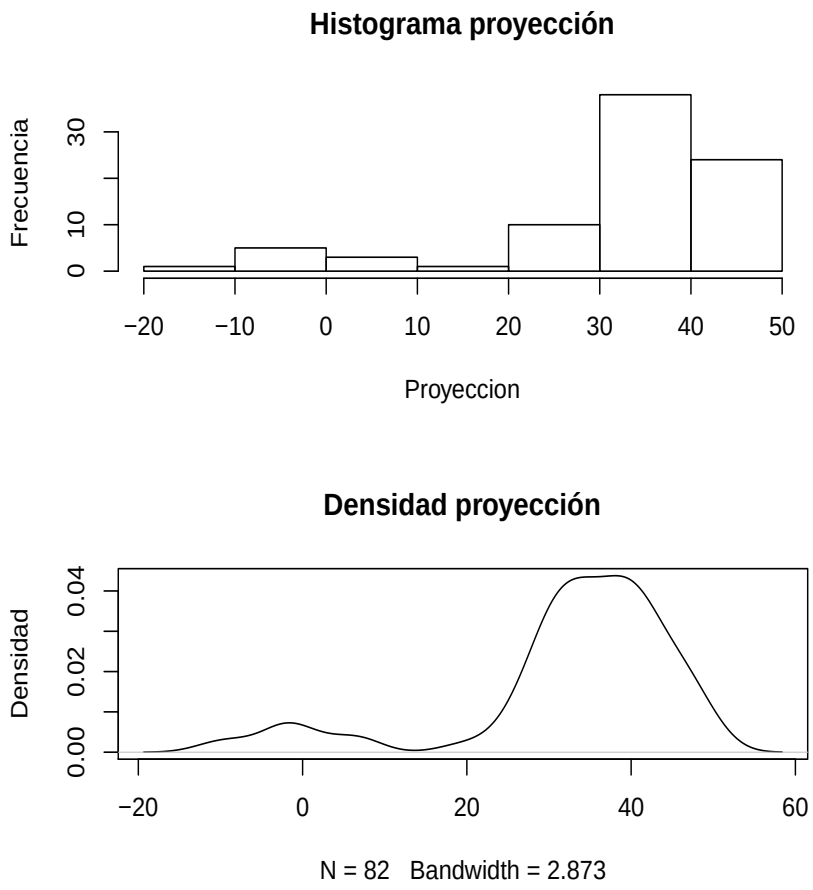

Figura 5.15: Histograma y densidad para componentes 1 y 4

Para este caso, tenemos un comportamiento bimodal, lo cual dice que estos componentes pueden estar suficientemente separados y por tanto se podría rechazar la hipótesis de unimodalidad.

En la figura 5.16 tenemos la siguiente proyección para los componentes 2 y 3:

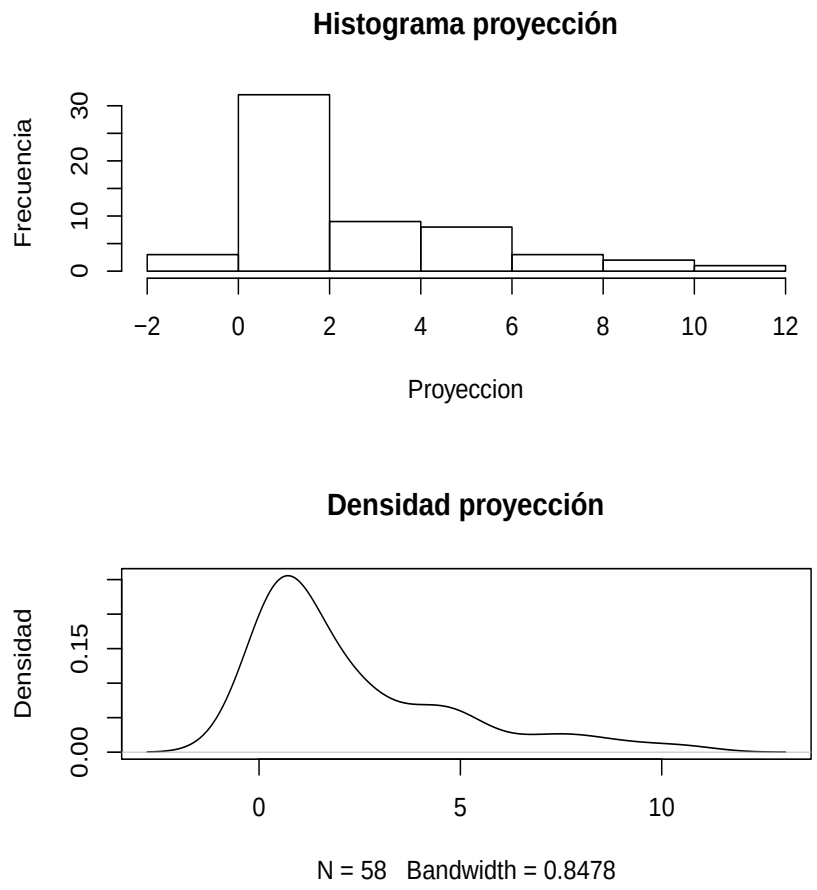

Figura 5.16: Histograma y densidad para componentes 2 y 3

En este caso como podemos ver en ambas gráficas, aquí se puede apreciar corroborar las sospechas de traslape entre componentes de la mezcla. Los datos proyectados de los componentes 2 y 3 muestran un comportamiento casi unimodal, lo cual se traduce en que estos componentes pueden ser fusionados si no se rechaza la hipótesis nula de unimodalidad en la prueba no paramétrica descrita anteriormente.

Los componentes 2 y 4 tienen asociada la siguiente proyección de los datos que se muestra en la figura 5.17:

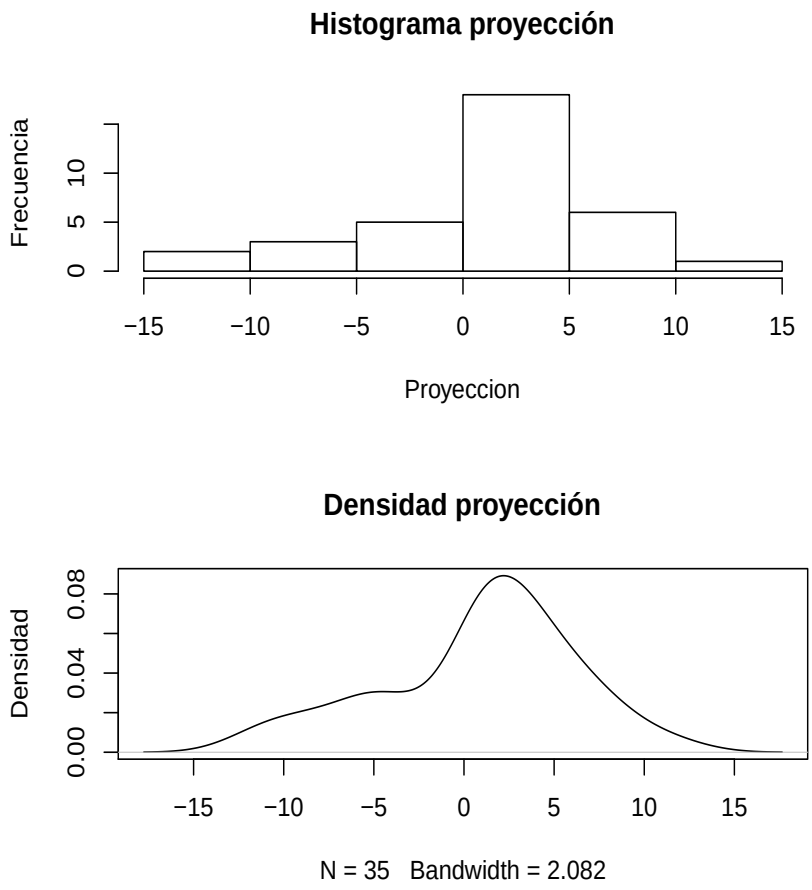

Figura 5.17: Histograma y densidad para componentes 2 y 4

Aquí también es notorio que los datos proyectados siguen un comportamiento que puede ser unimodal y como ya se había mencionado antes, algunas de las herramientas propuestas aquí revelan la evidencia para suponer que estos componentes están juntos por los resultados de los rootogramas y la matriz de clasificación errónea. Es muy probable que para estos componentes tampoco se rechace la hipótesis de unimodalidad al realizar esta prueba de hipótesis.

Por último, en la figura 5.18 se muestra para los componentes 3 y 4 tenemos las proyecciones siguientes:

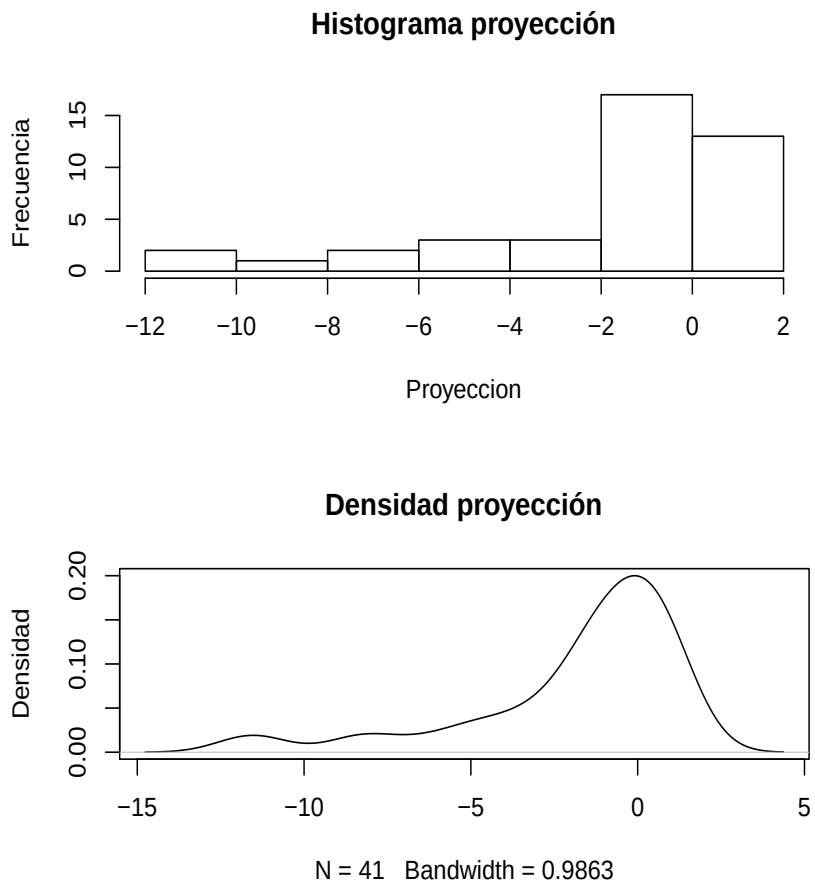

Figura 5.18: Histograma y densidad para componentes 3 y 4

En esta ultima proyección solo se viene a reiterar el traslape existente para este par de componentes ya que también en la gráfica se puede observar un comportamiento casi unimodal lo cual sugiere que estos componentes pueden ser fusionados al momento de obtener el árbol binario T asociado al modelo de mezclas.

En conclusión, las proyecciones también nos dieron más información para saber si los componentes de la mezcla están bien separados o no, estas proyecciones vienen a reafirmar las sospechas que se tenían de traslape entre los componentes 2,3 y 4 que ya se tenían al momento de obtener los rootogramas y matriz de probabilidad posterior.

Como siguiente paso se realizaran las pruebas de unimodalidad para cada par de componentes para así decidir si cada par es fusionado o no y así obtener un modelo con menos componentes y por tanto reducir la probabilidad de clasificación errónea y minimizar el error que se tiene al clasificar con el modelo de mezclas ya antes obtenido en **R**. La prueba de unimodalidad descrita en 4.6, se realizará en **R** con el paquete *diptest* que puede ser obtenido en la pagina de **R**. Para todos los pares de componentes a realizar la prueba, se utilizara un nivel de significancia $\alpha = 0.1$ y los cuantiles de la mejor distribución unimodal están dados en la tabla *qDiptab*.

Para los componentes 1 y 2 el valor de la estadística de prueba D es:

```
> dip(proyeccion)
$n
[1] 99

$dip
[1] 0.03911370
```

En este caso, el número de datos proyectados es $n = 99$ y el valor de la estadística es $D = 0.03911370$. El valor del cuantil en tablas de la estadística D es

```
 n     Pr   0.90
100   0.047152782
```

Es decir, con un nivel de $\alpha = 0.1$ tenemos que el cuantil 0.047152782 es mayor que D, es decir para los componentes 1 y 2 no se rechaza la hipótesis nula de unimodalidad, por lo tanto hay evidencia estadística suficiente para suponer que estos dos componentes no están bien separados.

Para los componentes 1 y 3 tenemos el siguiente valor de la estadística D

```
$n
[1] 105

$dip
[1] 0.09522506
```

Para los componentes 1 y 3 tenemos que el valor de la estadística es $D = 0.09522506$ y el cuantil en la tabla es el mismo que para el caso anterior ya que $n = 105$ y el cuantil es 0.047152782. Usando la misma regla de decisión que para el caso anterior, el cuantil es mayor que α, por tanto, se rechaza la hipótesis de unimodalidad para este caso al igual que el anterior.

Para el par de componentes 1 y 4 el valor de la estadística obtenido es:

```
$n
[1] 82

$dip
[1] 0.02903800
```

El valor del cuantil en tablas para el numero de datos proyectados $n = 58$ es:

```
 n    Pr   0.90
50   0.064913632
```

Como el valor del cuantil es mayor que el de la estadística, entonces para este caso también no se rechaza la hipótesis de unimodalidad.

Ahora para los componentes 2 y 3 tenemos que la estadística tiene el valor:

```
$n
[1] 58

$dip
[1] 0.02955439
```

El valor del cuantil en tablas es el mismo que para el caso anterior 0.064913632 y como es mayor que el valor de la estadística, entonces no se rechaza la hipótesis nula y se puede suponer que ambos componentes siguen una distribución unimodal lo cual concuerda con las gráficas de los histogramas y de las densidades estimadas.

Los componentes 2 y 4 dieron como valor de la estadística D el siguiente:

```
$n
[1] 35

$dip
[1] 0.04466146
```

El valor del cuantil en tablas es:

```
 n    Pr   0.90
30   0.081479138
```

Para este caso también se tiene que no se rechaza la hipótesis nula, ya que el valor del cuantil es mayor que el de la estadística. Esto también concuerda con lo observado en las gráficas anteriores.

Por último para los componentes 3 y 4 tenemos el valor de la estadística D:

```
$n
[1] 41

$dip
[1] 0.03810558
```

Y el valor del cuantil en tablas es el mismo que para 50 observaciones 0.064913632 y es mayor que el valor de la estadística D, por tanto aquí tampoco se rechaza la hipótesis nula de unimodalidad y se puede suponer que estos componentes también siguen un comportamiento unimodal, lo cual se traduce en un traslape de los componentes.

En conclusión, el componente 1 está bien separado del componente 3 y al parecer hay evidencia estadística suficiente para suponer que no esta bien separado de los componentes 2 y 4. Así como también hay evidencia que corrobora lo que ya se había dicho sobre los componentes 2, 3 y 4 que existe un traslape de los datos, lo cual se traduce en que estos componentes no siguen una distribución normal multivariada y se tendría que aplicar el algoritmo aquí propuesto para saber si hay o no fusión de uno o más conglomerados.

Para dibujar el árbol T, el primer paso es encontrar el par de componentes que estén más cercanos, este par lo obtenemos mediante el uso de la estadística D, pero tenemos dos valores de la estadística que tienen un número pequeño y parecido. Si nos fijamos en la estadística D de los componentes 1 y 4 es igual a $D = 0.029038$ y para los componentes 2 y 3 es $D = 0.02955439$, para ambos casos se tiene que no es rechazada la hipótesis de unimodalidad y por tanto están traslapados, pero en base a la información antes obtenida de los rootogramas y la matriz de clasificación errónea así como las probabilidades de clasificación errónea por componentes, optaremos por fusionar los componentes 2 y 3 ya que existe mayor evidencia de traslape entre estos dos conglomerados.

Entonces estos dos componentes serán los primeros en ser fusionados y después de ser fusionados se tiene que probar este nuevo conglomerado con los 2 restantes que serian el 1 y 4. Y después el conglomerado resultante tiene que ser probado contra el último conglomerado libre para saber si éstos son fusionados o no y así llegar al nodo raíz.

Ahora para el nuevo conglomerado que llamaremos $(2, 3)$, hay que realizar de nuevo las proyecciones contra los dos conglomerados restantes. La primera proyección mostrada en la figura 5.19 será contra el componente 1 y es la siguiente:

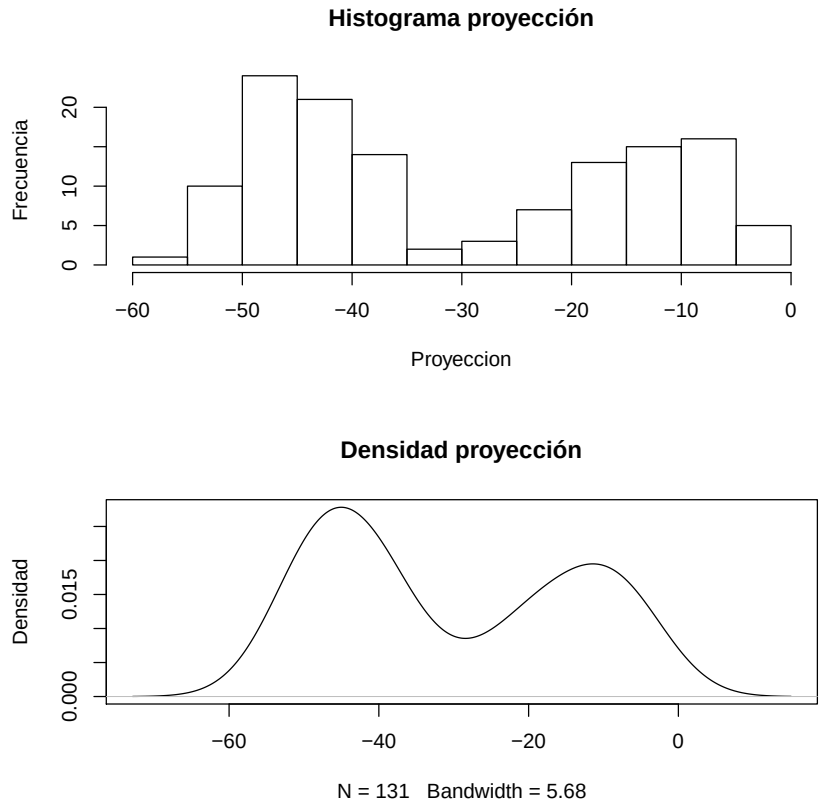

Figura 5.19: Histograma y densidad para componentes (2,3) y 1

En las gráficas del histograma y de la densidad de los datos proyectados de los componentes $(2, 3)$ y 1 se puede observar que no existe un comportamiento unimodal de los datos lo cual confirmamos con el resultado de la prueba de unimodalidad.

```
$n
[1] 131

$dip
[1] 0.06819015
```

El valor del cuantil en tablas es:

```
 n     Pr    0.90
100    0.047152783
```

En este caso, tenemos que el valor de la estadística $D = 0.06819015$ es mayor que el valor del cuantil en tablas 0.047152783, entonces se rechaza la hipótesis nula de unimodalidad para estos dos componentes y su puede suponer que ambos componentes están bien separados y no pueden ser fusionados.

Para el componentes $(2, 3)$ y el componentes 4 tenemos la siguiente proyección en la figura 5.20:

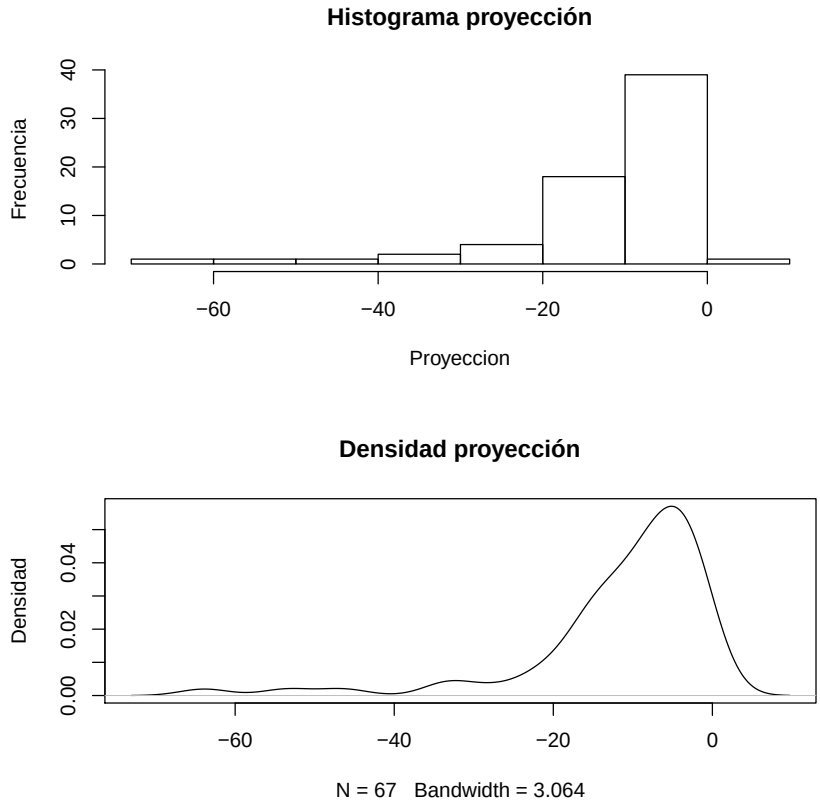

Figura 5.20: Histograma y densidad para componentes (2,3) y 4

Vemos que para este par de componentes, las gráficas tanto del histograma como de la densidad muestran un comportamiento unimodal lo cual se verifica aplicando la prueba de unimodalidad.

```
$n
[1] 67

$dip
[1] 0.03171908
```

Y el valor del cuantil en tablas es 0.064913632. Dado que el valor del cuantil en tablas es mayor que el de la estadística, entonces no se rechaza en este caso la hipótesis de unimodalidad, lo cual dice que estos componentes pueden ser fusionados.

En este paso del algoritmo tenemos que el nuevo conglomerado $(2, 3)$ debe ser fusionado con el conglomerado numero 4. A este nuevo conglomerado le llamaremos $(2, 3, 4)$ y por último resta verificar si puede ser fusionado o no con el último conglomerado restante numero 1.

La proyección de estos dos conglomerados $(2, 3, 4)$ y 1 se muestra en la figura 5.21:

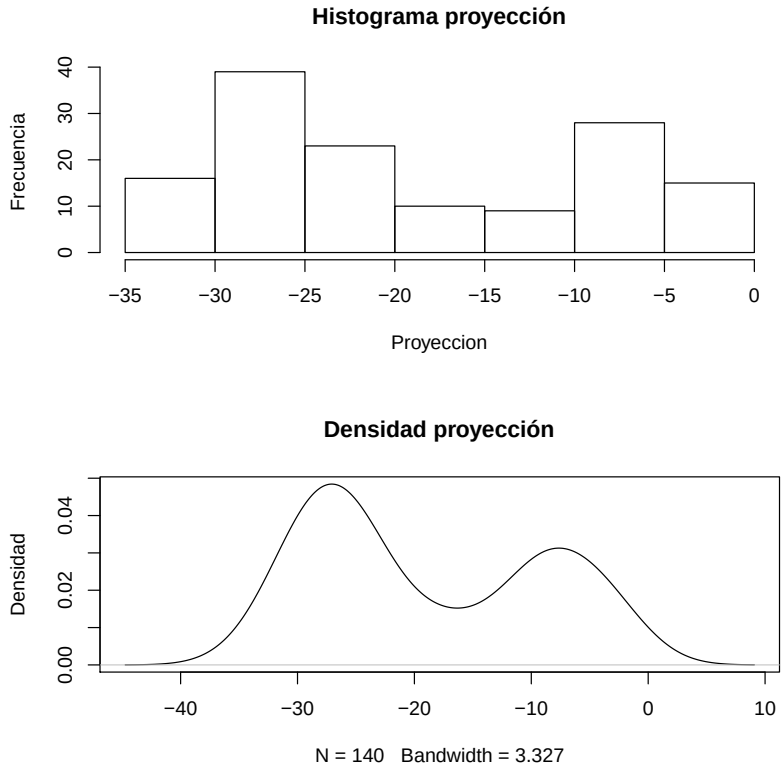

Figura 5.21: Histograma y densidad para componentes (2,3,4) y 1

Por último para este caso tenemos que las gráficas tanto del histograma como de la densidad muestran que los conglomerados $(2, 3, 4)$ y 1 siguen un comportamiento bimodal, lo cual verificaremos con la prueba de unimodadilad la cual arrojo como resultados:

```
$n
[1] 140

$dip
[1] 0.06336869
```

Y el valor del cuantil en tablas es:

```
 n    Pr   0.90
100   0.047152782
```

Como podemos ver, el valor de la estadística $D = 0.06336869$, que es mayor que el valor del cuantil en tablas 0.047152782, por lo tanto rechazamos la hipótesis nula de unimodalidad al nivel $\alpha = 0.1$ para los componentes $(2, 3, 4)$ y 1, lo cual se traduce en que estos conglomerados no pueden ser fusionados.

Sólo queda mencionar que la distribución del nuevo componente que es resultado de la fusión de los componentes 2,3 y 4 está dada por la suma de estos 3 componentes, es decir $f_{(2,3,4)}(\mathbf{x}) = f_2(\mathbf{x}) + f_3(\mathbf{x}) + f_4(\mathbf{x})$. Los parámetros de esta densidad son los correspondientes a la nueva variable aleatoria dada por la suma de las 3 densidades y la proporción $\pi_{(2,3,4)}$ está dada por $\pi_{(2,3,4)} = \pi_2 + \pi_3 + \pi_4$.

Mientras que los parámetros del primer componente vienen siendo los originales mostrados en la tabla de vectores media del modelo original de 4 componentes [5.1].

Podemos concluir entonces que los conglomerados 2,3 y 4 fueron fusionados mediante el uso de la prueba de hipótesis de unimodalidad. Mientras que el conglomerado numero 1 no fue fusionado con este criterio, así pues podemos pasar a dibujar el árbol binario T con la información que nos dio la prueba de hipótesis.

5.2.7. Árbol binario

Como ya se mencionó, mediante las pruebas de hipótesis realizadas en la sección, tenemos que los conglomerados fusionados son el 2,3 y 4 mientras que el 1 resultó no ser fusionado, con el árbol binario podemos ilustrar este proceso de fusión de conglomerados. En el primer árbol se muestran los nodos correspondientes a los 4 conglomerados del modelo y están relacionados de tal forma que los mas juntos son los que tienen estadísticas D menores, lo cual significa que son candidatos a ser fusionados, a continuación mostramos este primer árbol.

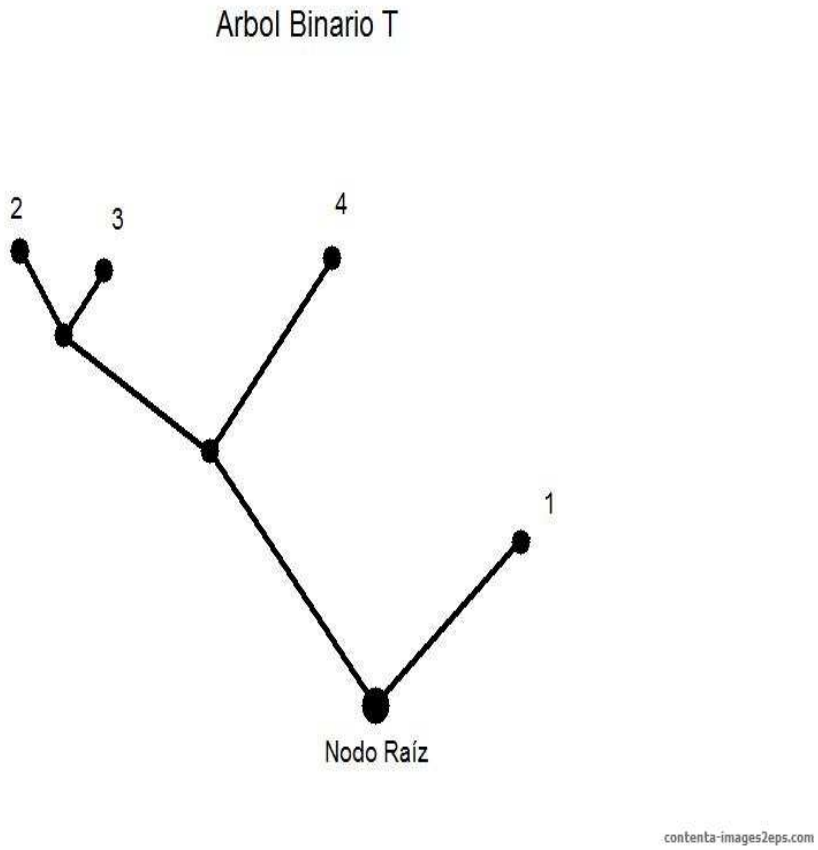

Figura 5.22: Árbol binario T

El árbol es la representación gráfica del proceso de fusión de los componentes de la mezcla. Como podemos ver, el árbol tiene 4 nodos que son los correspondientes a cada componente de la mezcla. Esta gráfica muestra el árbol binario antes del proceso donde se puede ver que el árbol tiene 3 niveles y las ramas son las observaciones correspondientes a cada conglomerado.

Ahora mostramos el árbol binario después de la primera fusión de los componentes 2 y 3:

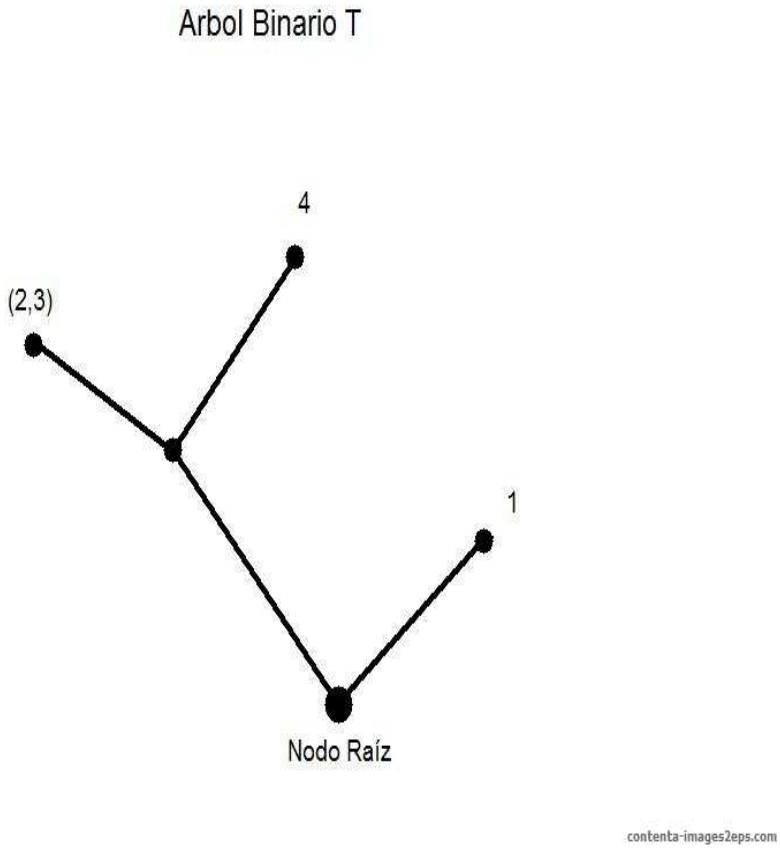

Figura 5.23: Árbol binario T

En este árbol se observa ya el primer corte y fusión de los componentes 2 y 3, ahora el nodo mas cercano es el 4 al que después se fusionará con el nuevo componente $(2, 3)$ para mostrar de forma gráfica este proceso.

Ahora mostramos el árbol binario final obtenido mediante la fusión de los componentes 2,3 y 4, y el componente restante 1 que este ya no pudo ser fusionado, dando como resultado un numero final de 2 conglomerados en el conjunto de datos.

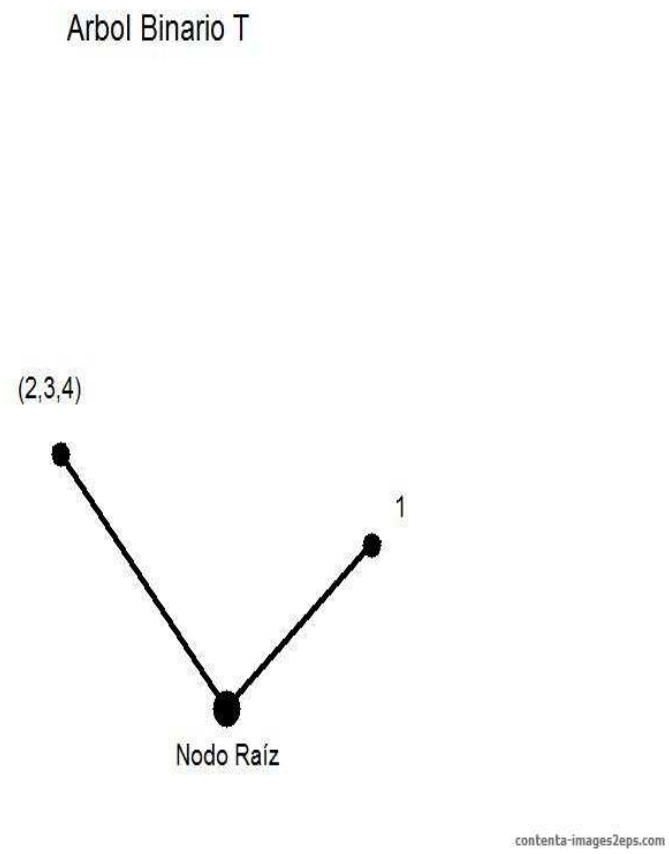

Figura 5.24: Árbol binario T

Aquí se muestra el ultimo paso de la fusión de componentes de la mezcla, se observa como los componentes 2,3 y 4 fueron fusionados, mientras que el componente 1 se considera un componente bien separado del resto. Esto concuerda con toda la información obtenida por las demás herramientas del algoritmo que dijo que existía un traslape entre los componentes 2,3 y 4, y ya fusionados se redujo el número de componentes de la mezcla a sólo 2 y es mas compacto el modelo de mezclas final.

Ahora pasamos a la última sección de este capitulo donde se evalúa el desempeño de este algoritmo para nuestro ejemplo.

5.2.8. Resultados finales del algoritmo

Como último paso queda ver ¿qué tan bueno fue el ajuste de este nuevo modelo? Podremos verificar esto mediante el uso de la matriz de clasificación errónea M y la probabilidad total de clasificación errónea P_{ce}. Para este caso el elemento m_{11} es el mismo descrito en la matriz M original, mientras que el componente $m_{1,(2,3,4)}$ esta dado por $m_{1(2,3,4)} = 1 - m_{11}$. Para el componente $m_{(2,3,4)1}$ que es la probabilidad posterior de que una observación del nuevo conglomerado $(2, 3, 4)$ sea clasificado en el grupo 1, tenemos que esta probabilidad la obtenemos sumando las probabilidades de clasificar una observación del conglomerado 2,3 o 4 en el conglomerado 1, es decir, $m_{(2,3,4)} = m_{21} + m_{31} + m_{41} = 5.960105623e - 09$. Y por tanto obtenemos la probabilidad posterior de clasificar un elemento del nuevo conglomerado $(2, 3, 4)$ en el mismo es $1-m(2, 3, 4)$. recordando que las probabilidades de clasificación errónea por componente MC_g se calculan como $MC_g = \sum_{i \neq g}^{G-1} m_{gi}$ entonces construimos nuestra matriz como sigue:

$$\begin{array}{c} \\ C1 \\ C2 \end{array} \begin{array}{c} \begin{array}{cc} C1 & C2 \end{array} \\ \left(\begin{array}{cc|c|c} 0.5111182 & 0.4888818 & 0.4888818 & 0.51908341 \\ 5.960105623e-09 & 0.9999990 & 5.960105623e-09 & 0.48091659 \end{array} \right) \end{array} \begin{array}{c} \\ MC_g \\ \\ \end{array} \begin{array}{c} \\ Pi_g \\ \\ \end{array}$$

Y la probabilidad total de clasificación errónea es:

$$P_{ce} = MC_1\pi_1 + MC_2\pi_2 = 0.2537737464$$

Podemos observar que la matriz de clasificación errónea se tiene una muy alta probabilidad de clasificar bien una observación de componente $(2, 3, 4)$ en el mismo y la misma probabilidad de clasificar una observación de componente 1 en si mismo. También podemos observar que tenemos una probabilidad total de clasificación errónea menor que la que se obtuvo con el modelo de mezclas de 4 componentes.

Tenemos que para el modelo de 4 componentes nos dio $P_{ce} = 0.4090807$ que es mayor a la probabilidad del modelo ya ajustado mediante el algoritmo que es $P_{ce} = MC_1\pi_1 + MC_2\pi_2 = 0.2537737464$, es decir, pasamos de un 40 % de probabilidad de clasificar mal las observaciones a un 25 % de probabilidad de clasificar mal las observaciones.

Por lo tanto podemos concluir que el algoritmo para este caso fue eficiente ya que se redujo el numero de componentes de la mezcla y por tanto también se redujo el numero de parámetros a estimar y se mejoraron las probabilidades de clasificación errónea lo cual indica que hay menos incertidumbre para clasificar haciendo al nuevo modelo mas confiable.

Conclusiones

Hasta este momento hemos visto las numerosas aplicaciones que tiene el análisis de conglomerados en diferentes ramas del conocimiento humano. También hemos visto los diferentes métodos clásicos para obtener conglomerados así como sus ventajas y desventajas. Se proporciono la teoría que existe detrás de los modelos basados en mezclas, así como su gran flexibilidad y empleo en diversos problemas. Pero estos modelos a pesar de ser tan flexibles y ampliamente utilizados poseen la desventaja de que si alguno o varios de los grupos no siguen una distribución normal multivariada, uno o más grupos pueden ser modelados por más de un componente de una mezcla de distribuciones normales.

Es en este punto donde presentamos el algoritmo propuesto en el artículo [2]. Como pudimos ver, algunos de los pasos del algoritmo como son los rootogramas y la matriz de clasificación errónea son muy útiles para detectar el traslape de uno o más componentes de la mezcla. Esto nos lleva a sospechar que dicho grupo no está siguiendo una distribución normal y por tanto se tiene un número más grande de componentes que de grupos en la población.

En este trabajo propusimos como ejemplo un conjunto de datos bivariados simulados en el paquete estadístico **R**, el cual consiste en 140 observaciones provenientes de las distribuciones mencionadas al principio del capítulo 5. Esta simulación fue propuesta ya que presentaba buenas propiedades. En primer lugar fue obtenida mediante distribuciones normales las cuales dieron origen al conjunto de datos que se puede apreciar aproximadamente encerradas en el rectángulo en $\mathbb{R}^2$ dado por $[10, 20] \times [10, 15]$ que por construcción si tiene una distribución normal, el resto de la nube de puntos que sigue una curva descendente se obtuvo mediante las distribuciones exponenciales dando esta forma a ese conjunto de datos. Obviamente por construcción este conjunto de datos no sigue una distribución normal y por tanto en **R** se obtuvo que para modelar esta nube de puntos eran necesarios 3 componentes normales bivariados, lo cual llevó a un total de 4 componentes en un primer modelo de mezclas de distribuciones normales bivariadas.

Como pudimos observar, la probabilidad total de clasificación errónea era

alta (0.4090807) dándonos sospechas de una clasificación mala en el conjunto de observaciones. También mediante la matriz de clasificación errónea pudimos obtener las probabilidades de este tipo de cada componente, teniendo probabilidades altas como en el caso del componente tres. Fue también importante el cálculo y uso de la matriz de clasificación errónea ya que los elementos $m_{gg'}$ que son las probabilidades posteriores de clasificar equivocadamente una observación del grupo g al grupo g' nos dieron información sobre que tan traslapados se encontraban cada par de componentes.

Por ejemplo los componentes 1 y 3 tenían una probabilidad prácticamente cero ya que $m_{13} = 8.629884e - 51$ lo cual indicaba que estos componentes estaban bastante bien separados y no podrían ser candidatos a una futura fusión. Sin embargo para los componentes 3 y 2 se tuvo una probabilidad $m_{32} = 0.4639845$ que es bastante alta, esta probabilidad nos decía que aproximadamente el 46 % de las observaciones del componente 3 podían ser mal clasificadas en el componente 2.

Los rootogramas también fueron de gran utilidad ya que éstos son una representación gráfica de lo que ocurre dentro del conjunto de datos. En los rootogramas correspondientes al componente 1 es claramente visible que no existe casi traslape alguno con cualquier otro componente de la mezcla, mientras que para el componente 3 es muy evidente que si existía traslape con los componentes 2 y 4. Esto era de esperarse ya que por construcción de los datos, esta terna de componentes no siguen una distribución normal.

También se propuso una herramienta estadística más fuerte para poder determinar si es significativo o no el traslape de los componentes y así determinar el número final de conglomerados. Esto fue mediante la prueba de hipótesis de unimodalidad que propone la estadística DIP para realizar la prueba. Aquí se entró en un debate ya que la menor de las estadísticas para los primeros pares de componentes fue la estadística $D = 0.029038$ que correspondía a los componentes 1 y 4. Pero dada la información previa que brindó la matriz de clasificación errónea, así como las probabilidades de clasificación errónea de cada componente y los rootogramas se opto por fusionar el par de componentes 2 y 3 que su estadística fue $D = 0.0295543$. En base a la información anterior fue como el proceso de fusión de los componentes se realizó ya que en ésta, se tenía mucha mayor probabilidad posterior de clasificar erróneamente elementos del componente 2 al 3 y viceversa que probabilidad posterior de clasificar erróneamente observaciones del grupo 1 al 4 y viceversa. Aquí se tuvo mucho cuidado ya que los resultados de la estadística D al parecer son muy sensibles, si nos hubiéramos fijado en sólo 3 decimales ya que en casos prácticos se realiza este truncamiento frecuentemente y prácticamente eran iguales. Se debían tener más herramientas y resultados para poder tener más criterio sobre cuáles conglomerados se deben

fusionar o cuales son más factibles, en este caso fueron mas factibles de ser fusionados los componentes 2 y 3.

Así por último realizamos el proceso de la obtención de árbol binario T, que es la parte visual del algoritmo, en este se pueden apreciar los pasos en los que los componentes 2,3 y 4 fueron fusionados hasta llegar a que no era factible fusionar el nuevo conglomerado formado por los componentes 2,3 y 4 con el restante componente 1, dando por resultado un número final de 2 conglomerados.

Se concluyó que este nuevo modelo conformado por el conglomerado correspondiente al componente 1 del modelo original y otro conglomerado correspondiente a una combinación de los componentes 2, 3 y 4 del modelo original. Se disminuyó el número de conglomerados de 4 a 2 y se mejoró la clasificación de los datos existentes y también se obtuvo una nueva matriz de clasificación errónea para este nuevo modelo y se redujo de un 40 % a un 25 % aproximadamente la probabilidad total de clasificar erroneamente las observaciones del conjunto de datos.

Por lo tanto se puede concluir que el empleo de este algoritmo junto con todas las herramientas propuestas tiene que ser de una manera muy cuidadosa sin dejar pasar por alto la información obtenida por los rootogramas u otras herramientas del algoritmo antes mencionadas, ya que no es suficiente y no puede ser óptimo el tomar la mínima estadística D entre los primeros pares de componentes a fusionar ya que se puede llegar a resultados diferentes. Esta información junto con le criterio del investigador pueden influir mucho en el resultado final del algoritmo ya que si se hubieran fusionado los componentes 1 y 4, el resultado final hubiera sido que los componentes fusionado hubieran sido el 1, 2 y 4, y se tendría un componente no fusionado que hubiera sido el numero 3.

Esto puede deberse a que el uso del discriminante lineal de Fisher como la mejor dirección que discrimina cada par de componentes de la mezcla para así obtener la proyección univariada de los datos y así tener su distribución y posterior uso de la prueba de hipótesis de unimodalidad, tal vez no sea la óptima, pudiera ser que en trabajos mas recientes se propongan otras direcciones diferentes a la del discriminante lineal que sean mucho mejores ya que si recordamos, este discriminante es de forma lineal y no es tan flexible como un discriminante no lineal.

Apéndices

Apéndice A

Discriminante lineal de Fisher

En el análisis discriminante se aborda el problema de recuperar, mediante las variables x_i ¿a cual de las pobalciones pertenece la observación? Se trata de encontrar funciones discriminantes o reglas de decisión $h = (x_1, x_2, \ldots, x_p)$ cuyos valores en los distintos grupos (o poblaciones) estén lo mas separados posible, es decir, buscamos funciones h sencillas que permitan asignar cada uno de los individuos a una población concreta Ω_g con $g = 1, 2, \ldots, G$, minimizando la tasa de error en dicha asignación.

La función h más sencilla y conocida es la función discriminante lineal de Fisher, donde h es una función lineal de $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ y se enuncia a continuación:

Sean μ_1 y μ_2 los vectores media de dos poblaciones Ω_1 y Ω_2 respectivamente. Sea Σ la matriz de varianzas y covarianzas común para ambas poblaciones. Sea $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ una observación a clasificar.

El *criterio geométrico*, consiste en asignar la observación $\mathbf{x}$ a la población más cercana utilizando la distancia de Mahalanobis:

$$d_M = [(\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})]^{\frac{1}{2}}$$

La regla de decisión es la siguiente:

- ω se asigna a Ω_1 si $d_M^2(\mathbf{x}, \mu_1) < d_M^2(\mathbf{x}, \mu_2)$
- ω se asigna a Ω_2 en caso contrario

A partir de la diferencia $d_M^2(\mathbf{x}, \mu_2) - d_M^2(\mathbf{x}, \mu_1)$, se construye la función discriminante lineal como sigue:

$$L(\mathbf{x}) = (\mathbf{x} - \frac{\mu_1 - \mu_2}{2}) \Sigma^{-1} (\mu_1 - \mu_2)'$$

Y se expresa la regla de decisión en función de ésta:

- ω se asigna a Ω_1 si $L(\mathbf{x}) > 0$
- ω se asigna a Ω_2 en caso contrario

Esta función es la función discriminante de Fisher.

Clasificación cuando los parámetros son estimados

En las aplicaciones prácticas μ_1, μ_2 y Σ son desconocidos y se deberán estimar a partir de muestras de tamaños n_1 y n_2 de las dos poblaciones Ω_1 y Ω_2.

Sean $\bar{\mathbf{x}}_1$,$\bar{\mathbf{x}}_2$ y $\mathbf{S}_1$, $\mathbf{S}_2$ los vectores media y las matrices de varianzas y covarianzas muestrales. La versión muestral del discriminante lineal de Fisher es:

$$\widehat{L}(\mathbf{x}) = (\mathbf{x} - \frac{\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2}{2})\mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}_2})'$$

De donde $\mathbf{S}_p = \frac{(n_1\mathbf{S}_1 + n_2\mathbf{S}_2)}{(n_1+n_2-2)}$ es la matriz de varianzas y covarianzas muestral ponderada.

Apendice B

Sintaxis en R

En este apéndice se da la sintaxis utilizada para la simulación de datos y obtención de resultados en el software **R**.

Sintaxis ejemplo capítulo 3

```
########## Ejemplo capitulo 3 ############

x1=rnorm(70,mean=-5,sd=2) #Se simulan 70 observaciones de una N(-5,4)
y1=rnorm(70,mean=0,sd=4) ##Se simulan 70 observaciones de una N(0,16)
z1=cbind(x,y) # Se combinan los objetos x y y en columnas
z2=cbind(x1,y1) # Se combinan los objetos x1 y y1 en columnas
datos=rbind(z1,z2) # Se combinan los objetos z1 y z2 en renglones
plot(datos) #Se grafican los datos

h1=hclust(dist(datos),method="single") #Conglomerados usando vecino mas cercano
plot(h1, main="Dendograma", xlab="Observaciones", ylab="Distancia")
clusters1=cutree(h1,k=2)
plot(simulacion[],type="n", main="Datos clasificados")
points(simulacion[clusters1==1,1],simulacion[clusters1==1,2],col="red")
points(simulacion[clusters1==2,1],simulacion[clusters1==2,2],col="blue")

h2=hclust(dist(datos),method="complete") #Conglomerados usando vecino mas lejano
plot(h2, main="Dendograma", xlab="Observaciones", ylab="Distancia")
clusters2=cutree(h2,k=4)
plot(simulacion[],type="n", main="Datos clasificados")
points(simulacion[clusters2==1,1],simulacion[clusters2==1,2],col="red")
points(simulacion[clusters2==2,1],simulacion[clusters2==2,2],col="blue")

h3=hclust(dist(datos),method="average") #Conglomerados usando promedio
plot(h3, main="Dendograma", xlab="Observaciones", ylab="Distancia")
abline(h=10)
clusters3=cutree(h3,k=4)
plot(simulacion[],type="n", main="Datos clasificados")
points(simulacion[clusters3==1,1],simulacion[clusters3==1,2],col="red")
points(simulacion[clusters3==2,1],simulacion[clusters3==2,2],col="blue")

kmedias=kmeans(datos, 2) #Se aplica el algoritmo de k-medias con k=2
kmedias
plot(simulacion[],type="n", main="Datos clasificados")
points(simulacion[kmedias$cluster==1,1],simulacion[kmedias$cluster==1,2],col="red")
points(simulacion[kmedias$cluster==2,1],simulacion[kmedias$cluster==2,2],col="blue")
```

Sintaxis ejemplo 1

Sintaxis utilizada en el ejemplo simulado para el capítulo 5 y sobre el cual se realizaron todos los cálculos y se obtuvieron los resultados presentados en este trabajo.

```
###################### Simulacion de los datos ############################
x1=rnorm(70,mean=-15,sd=2)#Se simulan 70 obs. de una N(-15,4)
x2=rnorm(70,mean=-15,sd=2)#Se simulan 70 obs. de una N(-15,4)
```

```
x3=rexp(70,rate=.2)
x4=rexp(70,rate=.2)#Se simulan 70 obs. de una exp(.25)

z1=cbind(x1+29,x2+29) #Se agrupan las coordenadas
z2=cbind(x3,x4) #Se agrupan las coordenadas
simulacion=rbind(z1,z2) #Se hacen las coordenadas

plot(simulacion, main="Gráfica de los datos") #Se grafican los datos simulados

#################### Ingreso de los datos y analisis ###################
simulacion<-read.table("simulacion1exp.2exp.2.txt", header = TRUE)

plot(simulacion, main="Datos simulados")
par(mfrow=c(2,2)) #Grafico en mosaico
hist(x1)#Histograma de x1 #histograma de la variable x1
hist(x2)#Histograma de x2 #histograma de la variable x2
hist(x3)#Histograma de x3 #histograma de la variable x3
hist(x4)#Histograma de x4 #histograma de la variable x4

################### Obtencion de conglomerados ##########################

par(mfrow=c(1,3)) #Grafico en mosaico
plot(h1, main="Vecino mas cercano", xlab="Observaciones", ylab="Distancia")
abline(h=5) #Linea horizontal a la altura 5
plot(h2, main="Vecino mas lejano", xlab="Observaciones", ylab="Distancia")
abline(h=20) #Linea horizontal a la altura 20
plot(h3, main="Promedio", xlab="Observaciones", ylab="Distancia")
abline(h=10) #Linea horizontal a la altura 10

par(mfrow=c(1,3)) #Grafico en mosaico
clusters1=cutree(h1,k=2)
plot(simulacion[],type="n", main="Vecino mas cercano")
points(simulacion[clusters1==1,1],simulacion[clusters1==1,2],col="red")
points(simulacion[clusters1==2,1],simulacion[clusters1==2,2],col="blue")
points(simulacion[clusters1==3,1],simulacion[clusters1==3,2],col="green")
points(simulacion[clusters1==4,1],simulacion[clusters1==4,2],col="orange")

clusters2=cutree(h2,k=4)
plot(simulacion[],type="n", main="Vecino mas lejano")
points(simulacion[clusters2==1,1],simulacion[clusters2==1,2],col="red")
points(simulacion[clusters2==2,1],simulacion[clusters2==2,2],col="blue")
points(simulacion[clusters2==3,1],simulacion[clusters2==3,2],col="green")
points(simulacion[clusters2==4,1],simulacion[clusters2==4,2],col="orange")

clusters3=cutree(h3,k=4)
plot(simulacion[],type="n", main="Promedio")
points(simulacion[clusters3==1,1],simulacion[clusters3==1,2],col="red")
points(simulacion[clusters3==2,1],simulacion[clusters3==2,2],col="blue")
points(simulacion[clusters3==3,1],simulacion[clusters3==3,2],col="green")
points(simulacion[clusters3==4,1],simulacion[clusters3==4,2],col="orange")

#################### Modelos basados en mezclas ###################
##### Descargar, instalar y cargar el paquete Mclust

mezcla<-Mclust(simulacion, modelNames="VVV") #Conglomerados con modelos basados en mezclas
mezcla
plot(simulacion[],type="n")
points(simulacion[mezcla$classification==1,1],simulacion[mezcla$classification==1,2],
+col="blue")
points(simulacion[mezcla$classification==2,1],simulacion[mezcla$classification==2,2],
+col="green")
points(simulacion[mezcla$classification==3,1],simulacion[mezcla$classification==3,2],
```

```
+col="red")
points(simulacion[mezcla$classification==4,1],simulacion[mezcla$classification==4,2],
+col="black")
title(main="clasificasion")

mezclaBIC<-mclustBIC(simulacion, modelNames="VVV") #BIC
mezclaBIC
plot(mezclaBIC, xlab="Numero de componentes") #Grafica BIC para cada particion de Sigma
title(main="Gráfica BIC", xlab="Número de componentes")
abline(v=4)
abline(v=3)

mclust2Dplot(simulacion, parameters=mezcla$parameters,z=mezcla$z) #Grafica de clasificacion
title(main="Clasificasión de los datos", sub="Mediante modelos basados en mezclas")

################ Mosaico de comparación mezclas #######

par(mfrow=c(2,2)) #Grafico en mosaico

clusters2=cutree(h2,k=4)
plot(simulacion[],type="n", main="Vecino mas lejano")
points(simulacion[clusters2==1,1],simulacion[clusters2==1,2],col="red")
points(simulacion[clusters2==2,1],simulacion[clusters2==2,2],col="blue")
points(simulacion[clusters2==3,1],simulacion[clusters2==3,2],col="green")
points(simulacion[clusters2==4,1],simulacion[clusters2==4,2],col="orange")

clusters3=cutree(h3,k=4)
plot(simulacion[],type="n", main="Promedio")
points(simulacion[clusters3==1,1],simulacion[clusters3==1,2],col="red")
points(simulacion[clusters3==2,1],simulacion[clusters3==2,2],col="blue")
points(simulacion[clusters3==3,1],simulacion[clusters3==3,2],col="green")
points(simulacion[clusters3==4,1],simulacion[clusters3==4,2],col="orange")

kmedias=kmeans(simulacion, 4) #Se aplica el algoritmo de k-medias con k=4
kmedias
plot(simulacion[],type="n", main="K-medias")
points(simulacion[kmedias$cluster==1,1],simulacion[kmedias$cluster==1,2],col="red")
points(simulacion[kmedias$cluster==2,1],simulacion[kmedias$cluster==2,2],col="blue")
points(simulacion[kmedias$cluster==3,1],simulacion[kmedias$cluster==3,2],col="green")
points(simulacion[kmedias$cluster==4,1],simulacion[kmedias$cluster==4,2],col="orange")

########## Proyecciones

direccion<-DIRECCION(simulacion[mezcla$classification==2,],+
simulacion[mezcla$classification==2,],mezcla$parameters$mean[,2],mezcla$parameters+
$mean[,2],mezcla$parameters$variance$sigma[,,2],mezcla$parameters$variance$sigma[,,2])
proyeccion<-PROYECCION(direccion,simulacion[mezcla$classification==2,],+
simulacion[mezcla$classification==2,])

par(mfrow=c(2,1))
hist(proyeccion, main="Histograma proyección", xlab="Proyeccion", ylab="Frecuencia")
plot(density(proyeccion), main="Densidad proyección", ylab="Densidad")

plot(density(pro1), main="Comparación", ylab="Densidad")
unimodal<-density(pro1, bw=30) #Mejor aproximacion unimodal
lines(unimodal, col="grey")
```

Funciónes programadas

Aqui se da la sintaxis de las funciones programadas para obtener algunos datos como lo es la matriz de clasificacion erronea, las probabilidades MC_g, la probabilidad total de clasificacion erronea P_{ce}.

```
####### Matriz de probabilidades posteriores  #######

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

dens <- 1:length(x[,1])
for( i in 1:length(x[,1]) ){
dens[i] <- (2*pi)^(-length(mu)/2) * det(S)^(-1/2) * exp( (-1/2) * t((x[i,] - mu)) %*%
+solve(S) %*% (x[i,]- mu) )
}
dens

}


simulacion <- as.matrix(simulacion)
attach(mezcla)

p1 <- MV( simulacion, parameters$mean[,1], parameters$variance$sigma[ , ,1] )
dMV <- p1

for( i in 2:length(parameters$pro) ){
p <- MV( simulacion, parameters$mean[,i], parameters$variance$sigma[ , ,i] )
dMV <- cbind(dMV,p)
}

wdMV <- parameters$pro * dMV[1,]

for(i in 2:length(dMV[,1]) ){
wdMV <- rbind(wdMV,parameters$pro * dMV[i,])
}

wdMV <- matrix(wdMV,ncol=4)

M <- matrix( rep(0,times=length(wdMV[,1])*length(wdMV[1,]) ), nrow=length(wdMV[,1]),
  +ncol = length(wdMV[1,]) )

for( i in 1:length(wdMV[,1]) ){
for( j in 1:length(wdMV[1,]) ){
M[i,j] <- wdMV[i,j] / sum(wdMV[i,])
}
}

####### Probabilidad total de clasificacion erronea #######
PTCE<-function(MCE,pesos){
   probce<-1:4
   for(i in 1:4)
    probce[i]<-(1-MCE[i,i])*mezcla$parameters$pro[i]
    sum(probce)
}

###### Probabilidades MC_g #############
MCg<-function(MCE){
   mcg<-1:4
    for(j in 1:4){
```

```
    mcg[j]<-(1-MCE[j,j])
    }
 mcg
}

######### Proyecciones

DIRECCION<-function(clas1,clas2,mu1,mu2,sigma1,sigma2){
    Sponderada<-((length(clas1)*sigma1)+(length(clas2)*sigma2))/(length(clas1)+
    +length(clas2))
    W= solve(Sponderada)%*%(mu1-mu2)
}

PROYECCION<-function(dir,clas1,clas2){
    pro1<-t(dir)%*%t(clas1)
    pro2<-t(dir)%*%t(clas2)
    pro<-rbind(t(pro1),t(pro2))
}
```

Estas funciones programadas se combinaron para obtener los diferentes resultados numéricos en este trabajo.

Bibliografía

[1] Aurea Grané *Análisis discriminante y clasificación*, Universidad Carlos III de Madrid, España, PDF.

[2] Fraley C. y Raftery A. (1998), How many clusters? wich clustering method? answer via model-based cluster analysis *The computer journal*, 41(8): 578-588.

[3] García Álvarez M. A. (2005) *Introducción a la teoría de la probabilidad Vol. 1*, México, Fondo de Cultura Económica

[4] García Álvarez M. A. (2005) *Introducción a la teoría de la probabilidad Vol. 2*, México, Fondo de Cultura Económica

[5] Hair J.F. Jr., Anderson E. R., Tatham R. L., Black W. C.(2007), *Análisis multivariante*, España, Prentice Hall.

[6] J.A. Hartigan y P.M. Hartigan. The dip test for unimodality *Annals of Statistics*, 13:70-84, 1985.

[7] Marsden J., Tromba A., *Cálculo Vectorial*, México, Person Education.

[8] McLachlan, G.J. and Peel, D. (2000) *Finite mixture models*, Wiley.

[9] Peña D. (2002), *Análisis de datos multivariantes*, España, Mc-Graw Hill

[10] Tantrum J., Murua A., Wener S. Assesment and pruning of hierichal model based clustering SIGKDD 2003 Washington, DC, USA.