UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

MODELOS DE MEZCLAS FINITAS APLICADOS A ESTIMACIÓN, CONGLOMERADOS Y ANÁLISIS DE DISCRIMINANTES

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIA

P R E S E N T A:

IRMA ROCÍO ZAVALA SIERRA

DIRECTORA DE TESIS:
DRA. RUTH SELENE FUENTES GARCÍA

2011

Hoja de Datos del Jurado

1. Datos del Alumno
Zavala
Sierra
Irma Rocío
55 58 42 52
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
303096997

2. Datos del tutor
Dra.
Fuentes
García
Ruth Selene

3. Datos del sinodal 1
Mat.
Margarita Elvira
Chávez
Cano

4. Datos del sinodal 2
M. en C.
María del Pilar
Alonso
Reyes

5. Datos del sinodal 3
Dr.
Ricardo
Ramírez
Aldana

6. Datos del sinodal 4
M. en C.
José
Antonio
Flores
Díaz

7. Datos del trabajo escrito
Modelos de mezclas finitas aplicados a estimación, conglomerados y análisis de discriminantes
132 p
2011

Índice general

Introducción

Los modelos de mezclas finitas son los modelos que consideran que la forma en que se distribuye una población es con base a una mezcla de distribuciones, es decir, una suma de distribuciones ponderadas. Se hacen supuestos sobre dichas distribuciones, como por ejemplo el de normalidad, pero en general puede ser cualquier familia de distribuciones. Lo importante es encontrar los parámetros de dicha distribución, para eso se ha utilizado el algoritmo de Esperanza-Maximización, el cual ayuda a encontrar los estimadores máximos verosímiles.

En el primer capítulo de este trabajo se describe más a fondo el modelo de mezclas finitas y se desarrolla en general el algoritmo de Esperanza-Maximización. Además se desarrolla dicho algoritmo para el caso univariado y multivariado de la distribución normal.

En ocasiones es necesario conocer la función de densidad de alguna población para poder tomar decisiones o simplemente porque se utiliza para encontrar más propiedades de la misma. Encontrar ésta no es fácil, se pueden hacer supuestos de la familia a la que pertenecen las observaciones, con base en la experiencia, pero se debe hacer algo para estimar el valor de los parámetros. En este caso se utiliza el método de momentos o el de máxima verosimilitud. Sin embargo, también se da el caso en que no se conoce ni si quiera la familia a la que pudiera pertenecer la población, para esto se puede suponer alguna familia y después decidir con base en una prueba de hipótesis si esto se acepta o se rechaza.

El segundo capítulo describe el tema en general de estimación de densidad y como se aplica el modelo de mezclas finitas en este problema, ejemplificando con algunos casos donde se supone distribución normal.

Un análisis de conglomerados, consiste en buscar dentro de una población ciertos grupos, es decir, el problema es poder decir si la población en general puede ser clasificada y la interpretación de esta clasificación, la cual debe ser coherente con el contexto del problema en general que se busca resolver. Para esto, hay diferentes métodos, los más usados son los métodos jerárquicos. Pero estos métodos no tienen una base estadística, no poseen algunas propiedades como la distribución o estimación de parámetros de cada grupo.

La descripción en general del tema de análisis de conglomerados se encuentra en el capítulo tres. En éste además se muestran ejemplos de la aplicación del modelo de mezclas finitas para encontrar grupos dentro de una población que se supone sigue una distribución

normal.

El análisis discriminante es aquél que, dado un número fijo de grupos, el problema es ahora encontrar una manera de clasificar dentro de algún grupo, una nueva observación. Si se conociera la manera que se distribuye cada grupo, se usaría una regla discriminante como la de Bayes o la de máxima probabilidad, para clasificar la nueva observación. Sin embargo, también existen métodos que no hacen supuestos sobre la distribución de la población, pero no siempre pueden servir.

Finalmente en el capítulo cuatro se habla del análisis discriminante, los modelos mas usados para discriminar y el modelo de mezclas finitas aplicado en este problema.

El modelo de mezclas finitas, puede ayudar a resolver los tres problemas antes mencionados. No se atacarán de la misma forma, puesto que la solución que se espera es distinta, pero si con el mismo modelo.

En el caso de estimación de densidad, se hace el supuesto de que la población tiene un comportamiento multimodal. Entonces la manera de encontrar la función de densidad es ajustando el modelo a la población, para lo cual se debe hacer un supuesto más, el de la familia de distribuciones a la que pertenece cada componente. Al final sólo interesa la función como tal, sin importar el número de componentes.

Para conglomerados, se asocia a cada componente un grupo. Lo que interesa es encontrar el número de componentes que mejor describa a la población y que además tengan una buena interpretación.

En el caso de análisis de discriminante, se conoce el número de grupos de la población en general. Se asocia a cada componente un grupo, por ende se está suponiendo una distribución para los mismos. Lo importante es encontrar el estimado de los parámetros, de tal manera que ayude a discriminar usando el discriminante de Bayes o el de máxima probabilidad.

Un mismo modelo aplicado a tres distintos problemas, en este trabajo se hace un análisis del cómo se ataca cada problema usando este modelo. En general, se puede usar con cualquier familia de distribuciones, pero este trabajo se basa en la distribución normal, ya que es la más usada.

Capítulo 1

Modelo de Mezclas Finitas

El pionero en usar este tipo de modelos fue Karl Pearson, el cuál usó unos datos sobre cangrejos obtenidos por Walter Frank Raphael Weldon, estos datos fueron ajustados por una mezcla de dos normales univariadas con varianzas diferentes, basándose en el método de momentos para estimar los parámetros. En 1924 Carl Vilhelm Ludwig Charlier y S. D. Wicksell extendieron el caso a una mezcla de componentes normales bivariadas y Doetsch (1928) se extendió a una mezcla de más de dos normales bivariadas. Tan y Chang (1972) y Fryer y Robertson (1972), entre otros, mostraron que el método de máxima verosimilitud era más eficiente que el método de momentos para este problema.

Butler (1986) escribe un poco sobre el uso de un algoritmo iterativo llamado de Esperanz-Maximización (EM) de Dempster, Laird y Rubin (1977), para estimar los parámetros de este modelo.

En el año de 1978 Schwarz G., toma en cuenta el Criterio de Información Bayesiana (BIC) para determinar las dimensiones del modelo y en 1988 Haughton también uso este método.

Bozdogan y Sclove en los años 1984 y 1987 respectivamente utilizaron el Criterio de Akaike (AIC), para determinar el número de componentes del modelo, en el caso particular de conglomerados.

A lo largo de estos últimos años se han desarrollado formas de atacar el problema, así como aplicaciones para el mismo considerando no solo distribuciones normales. Mas ejemplos de artículos en donde se aborda este tema, se encuentran en la referencia [17] pág. 35-37.

1.1. Idea General

Con base a la notación y resultados de la referencia [17], se explica a continuación la idea general del modelo de mezclas finitas. Sea $\mathbf{Y}_1, ..., \mathbf{Y}_n$ una muestra aleatoria de tamaño n, donde cada $\mathbf{Y}_j$, para $j = 1, ..., n$, es un vector de p dimensiones, al ser ésta una muestra aleatoria indica que son variables idénticamente distribuídas. Por lo tanto se puede denotar a la función de densidad para cada $\mathbf{Y}_j$ como $f(\mathbf{y})$. Se agrupan estas

variables aleatorias en una matriz de la siguiente forma $\mathbf{Y} = (\mathbf{Y}_1^{'}, ..., \mathbf{Y}_n^{'})^{'}$, cuando se usa el símbolo $(^{'})$ se está refiriendo al símbolo de vector o matriz transpuesta. Se denotan $\mathbf{y}_1, ..., \mathbf{y}_n$ como el conjunto de observaciones y de la misma manera que las variables, se pueden agrupar en una matriz a la que se llamará como la matriz de datos $\mathbf{y} = (\mathbf{y}_1^{'}, ..., \mathbf{y}_n^{'})^{'}$, donde $\mathbf{y}_j^{'}$ es la observación del vector aleatorio $\mathbf{Y}_j^{'}$.

Cuando se habla de un modelo de mezclas se supone que la función de densidad $f(\mathbf{x})$, con $\mathbf{x}$ un vector p-dimensional, puede ser escrita como:

$$f(\mathbf{x}) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{x}) \tag{1.1}$$

donde cada $f_i(\mathbf{x})$ es una función de densidad y cada π_i, llamadas proporciones o pesos de la mezcla, toman valores no negativos y menores que uno, $0 \leq \pi_i \leq 1$, de tal manera que $\sum_{i=1}^{g} \pi_i = 1$.

Se puede reescribir lo anterior considerando los parámetros de las funciones, entonces la función de densidad se expresa de la siguiente forma

$$f(\mathbf{x}, \Psi) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{x}, \theta_i) \tag{1.2}$$

donde Ψ corresponde a la matriz de parámetros desconocidos, $\Psi = (\boldsymbol{\pi}, \boldsymbol{\theta})$, donde

$$\boldsymbol{\pi} = (\pi_1, ..., \pi_g)^{'}$$

$$\boldsymbol{\theta} = (\theta_1, ..., \theta_g)^{'},$$

θ_i es el vector de parámetros para cada función de densidad $f_i(\mathbf{x})$.

Se suponen que las observaciones siguen una cierta distribución conocida, pero se desconocen los parámetros Ψ y g el número de componentes.

Uno de los métodos para estimar los parámetros de una densidad es el método de máxima verosimilitud (ML), la función de verosimilitud es entonces:

$$\begin{aligned} L(\Psi) &= \prod_{j=1}^{n} f(\mathbf{y}_j, \theta_i) \\ &= \prod_{j=1}^{n} \sum_{i=1}^{g} \pi_i f_i(\mathbf{y}_j, \theta_i) \end{aligned} \tag{1.3}$$

La log-verosimilitud, que en ocasiones es más fácil de manipular, tiene la siguiente expresión:

$$\begin{aligned} l(\Psi) &= Ln[L(\Psi)] \\ &= Ln[\prod_{j=1}^{n}\sum_{i=1}^{g}\pi_i f_i(\mathbf{y}_j,\theta_i)] \\ &= \sum_{j=1}^{n} Ln[\sum_{i=1}^{g}\pi_i f_i(\mathbf{y}_j,\theta_i)] \end{aligned} \tag{1.4}$$

Para obtener los estimadores de los parámetros usando la verosimilitud, es obteniendo las raíces de la derivada de la log-verosimilitud igualada a cero, ya que éstas son las que maximizan la verosimilitud.

$$\frac{\partial l(\Psi)}{\partial \Psi} = 0 \tag{1.5}$$

Obtener este resultado es algo complicado, para encontrar estimadores a los parámetros estimados se suele usar el algoritmo de esperanza maximizada (EM).

1.2. Algoritmo EM (Esperanza-Maximización)

Éste es un algoritmo iterativo propuesto en 1977 por Dempster, Laird y Rubin, usado frecuentemente para calcular los estimadores máximos verosímiles cuando los datos pueden ser vistos como incompletos, en esto radica su base. El método supone que existen datos de la muestra que no fueron observados, por lo tanto se debe maximizar la verosimilitud considerándolos. Son dos los pasos básicos de este algoritmo, primero se obtiene la esperanza y le sigue el paso de maximizar el valor esperado de la log-verosimilitud. Entre estos dos pasos se itera hasta llegar a una buena aproximación. Para comenzar a iterar el algoritmo necesita valores iniciales de los parámetros, los cuales se proponen.

Se llamará $\mathbf{z}_j$ al vector de datos no observados de los individuos j dentro de la muestra, de tal manera que

$$\mathbf{z}_j = (z_{1j}, ..., z_{gj})$$

donde

$$z_{ij} = \begin{cases} 1 & \text{si } \mathbf{y}_j \text{ sigue la densidad } f_i \\ 0 & \text{en otro caso} \end{cases}$$

Con base en lo anterior $P(z_{ij} = 1) = \pi_i$ y $P(z_{ij} = 0) = 1 - \pi_i$, por lo tanto al ser estas nuevas variables independientes e idénticamente distribuidas, siguen una distribución $Multinomial_g(1, \boldsymbol{\pi})$, es decir,

$$P(\mathbf{Z}_j = \mathbf{z}_j) = \prod_{i=1}^{g} \pi_i^{z_{ij}} \tag{1.6}$$

para todo $i = 1, ..., g$ y $j = 1, ..., n$.

Ahora la matriz de datos completos, de observados y no observados, se expresa de la siguiente forma

$$\mathbf{y}_c = (\mathbf{y}', \mathbf{z}')'$$

donde

$$\mathbf{z} = (\mathbf{z}_1^{'}, ..., \mathbf{z}_n^{'})'$$

Cuando se conocen los valores de $\mathbf{z}_j$, se puede calcular la función de densidad del individuo j, dicha densidad condicional queda expresada como

$$P(\mathbf{Y}_j = \mathbf{y}_j \mid \mathbf{Z}_j = \mathbf{z}_j) = \prod_{i=1}^{g} f_i(\mathbf{y}_j, \theta_i)^{z_{ij}} \tag{1.7}$$

Con esta información se calcula la densidad conjunta de $\mathbf{y}_j$ y $\mathbf{z}_j$, usando Bayes.

$$\begin{aligned} P(\mathbf{Y}_{cj} = \mathbf{y}_{cj}) &= P(\mathbf{Y}_j = \mathbf{y}_j, \mathbf{Z}_j = \mathbf{z}_j) \\ &= P(\mathbf{Y}_j = \mathbf{y}_j \mid \mathbf{Z}_j = \mathbf{z}_j) P(\mathbf{Z}_j = \mathbf{z}_j) \\ &= \prod_{i=1}^{g} f_i(\mathbf{y}_j, \theta_i)^{z_{ij}} \prod_{i=1}^{g} \pi_i^{z_{ij}} \\ &= \prod_{i=1}^{g} [f_i(\mathbf{y}_j, \theta_i)^{z_{ij}} \pi_i^{z_{ij}}] \\ &= \prod_{i=1}^{g} [f_i(\mathbf{y}_j, \theta_i) \pi_i]^{z_{ij}} \end{aligned} \tag{1.8}$$

Por lo tanto la función de densidad de los datos completos queda expresada de la siguiente forma

$$f_c(\mathbf{y}_{cj}) = \prod_{i=1}^{g} [f_i(\mathbf{y}_j, \theta_i) \pi_i]^{z_{ij}} \tag{1.9}$$

La verosimilitud de Ψ para los datos completos está dada por

$$\begin{aligned} L_c(\Psi) &= \prod_{j=1}^{n} f_c(\mathbf{y}_{cj}) \\ &= \prod_{j=1}^{n} \prod_{i=1}^{g} [f_i(\mathbf{y}_j, \theta_i) \pi_i]^{z_{ij}} \end{aligned} \tag{1.10}$$

En este caso es más fácil encontrar los máximos para la log-verosimilitud, la cual queda expresada como

$$\begin{aligned} l_c(\Psi) &= logL_c(\Psi) \\ &= log\prod_{j=1}^{n}\prod_{i=1}^{g}[f_i(\mathbf{y}_j,\theta_i)\pi_i]^{z_{ij}} \\ &= \sum_{j=1}^{n} log\prod_{i=1}^{g}[f_i(\mathbf{y}_j,\theta_i)\pi_i]^{z_{ij}} \\ &= \sum_{j=1}^{n}\sum_{i=1}^{g} log[f_i(\mathbf{y}_j,\theta_i)\pi_i]^{z_{ij}} \\ &= \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}log[f_i(\mathbf{y}_j,\theta_i)\pi_i] \\ &= \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}[logf_i(\mathbf{y}_j,\theta_i) + log\pi_i] \end{aligned}$$

Una vez establecida la log-verosimilitud de los datos completos

$$l_c(\Psi) = \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}[log\pi_i + logf_i(\mathbf{y}_j,\theta_i)] \tag{1.11}$$

se inicia con los pasos del algoritmo EM, mismos que se irán iterando conforme haya un acercamiento a los parámetros que maximicen dicha función.

1.2.1. Paso E

En este paso se calcula el valor de la esperanza de la log-verosimilitud de los datos completos dada la matriz de datos observados y suponiendo valores iniciales para Ψ. Para la primera iteración se proponen esos valores los que se denotarán como $\Psi^{(0)}$, esta esperanza queda expresa como

$$Q(\Psi;\Psi^{(0)}) = E_{\Psi^{(0)}}[l_c(\Psi) \mid \mathbf{y}] \tag{1.12}$$

En general cuando se está en la iteración $k+1$ los valores de los parámetros que se usan para calcular la esperanza se denotan como $\Psi^{(k)}$, entonces para esta iteración la esperanza está dada por

$$\begin{aligned} Q(\Psi;\Psi^{(k)}) &= E_{\Psi^{(k)}}[l_c(\Psi) \mid \mathbf{y}] \\ &= E_{\Psi^{(k)}}\{\sum_{j=1}^{n}\sum_{i=1}^{g} Z_{ij}[log\pi_i + logf_i(\mathbf{y}_j,\theta_i)] \mid \mathbf{y}\} \\ &= \sum_{j=1}^{n}\sum_{i=1}^{g} E_{\Psi^{(k)}}\{Z_{ij}[log\pi_i + logf_i(\mathbf{y}_j,\theta_i)] \mid \mathbf{y}\} \end{aligned} \tag{1.13}$$

En este paso se está suponiendo valores para los parámetros Ψ, entonces la expresión $log\pi_i + logf_i(\mathbf{y}_j,\theta_i)$ es un valor constante ya que se conoce el valor de todas las variables. Por lo tanto, al aplicar el operador esperanza no se ve afectado, es por eso que la función Q se reduce a

$$Q(\Psi;\Psi^{(k)}) = \sum_{j=1}^{n}\sum_{i=1}^{g}[log\pi_i + logf_i(\mathbf{y}_j,\theta_i)]E_{\Psi^{(k)}}(Z_{ij} \mid \mathbf{y}) \tag{1.14}$$

Por consiguiente, para conocer el valor de Q es necesario calcular $E_{\Psi^{(k)}}(Z_{ij} \mid \mathbf{y})$ ya que todos los otros términos son valores ya conocidos.

$$\begin{aligned} E_{\Psi^{(k)}}(Z_{ij} \mid \mathbf{y}) &= 0 * P_{\Psi^{(k)}}(Z_{ij}=0 \mid \mathbf{y}) + 1 * P_{\Psi^{(k)}}(Z_{ij}=1 \mid \mathbf{y}) \\ &= P_{\Psi^{(k)}}(Z_{ij}=1 \mid \mathbf{y}) \\ &= \frac{P_{\Psi^{(k)}}(Z_{ij}=1,\mathbf{y}_j)}{P_{\Psi^{(k)}}(\mathbf{y}_j)} \\ &= \frac{\pi_i^{(k)} f_i(\mathbf{y}_j;\theta_i^{(k)})}{f(\mathbf{y}_j;\Psi^{(k)})} \\ &= \frac{\pi_i^{(k)} f_i(\mathbf{y}_j;\theta_i^{(k)})}{\sum_{i=1}^{g}\pi_i^{(k)} f_i(\mathbf{y}_j;\theta^{(k)})} \end{aligned} \tag{1.15}$$

Esta expresión se denotará como

$$\begin{aligned} \tau_{ij}^{(k)} &= \tau_i(\mathbf{y}_j;\Psi^{(k)}) \\ &= \frac{\pi_i^{(k)} f_i(\mathbf{y}_j;\theta_i^{(k)})}{\sum_{i=1}^{g}\pi_i^{(k)} f_i(\mathbf{y}_j;\theta^{(k)})} \end{aligned} \tag{1.16}$$

Finalmente la esperanza de la log-verosimilitud dada la matriz de datos observados y bajo el supuesto de los parámetros $\Psi^{(k)}$ queda expresada de la siguiente forma

$$Q(\Psi;\Psi^{(k)}) = \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_i(\mathbf{y}_j;\Psi^{(k)})[log\pi_i + logf_i(\mathbf{y}_j,\theta_i)] \tag{1.17}$$

1.2.2. Paso M

En este paso se trata de maximizar $Q(\Psi, \Psi^k)$ con respecto a Ψ dados los parámetros $\Psi^{(k)}$. En este caso de modelos de mezclas, se calculan los estimadores $\boldsymbol{\pi}$ independientemente del resto de los parámetros.

Una manera de estimar π_i sería dividiendo el número total de variables $\mathbf{y}_j$, $j = 1, ..., n$ que siguen la función de densidad f_i, entre el número total de observaciones. La forma de calcular el número de variables que siguen dicha densidad es sumando sobre todas las j las variables z_{ij}. Debido que a estas se les asigna el valor de 1 si la variable $\mathbf{y}_j$ se distribuye conforme a la función de densidad f_i, de no ser así se le asigna el valor 0, de esta forma se estima π_i como

$$\hat{\pi}_i = \sum_{j=1}^{n} \frac{z_{ij}}{n} \qquad i = 1, ..., g \tag{1.18}$$

pero las variables z_{ij} no se conocen, por lo tanto se usará el valor esperado de las variables no observadas dada la observación y los parámetros de la iteración anterior o los valores inciales para el caso de la primera iteración. Este valor se denota por $\tau_i(\mathbf{y}_j; \Psi^{(k)})$, entonces la manera en la que se estima a π_i para la iteración $k + 1$ es

$$\begin{aligned} \pi_i^{(k+1)} &= \sum_{j=1}^{n} \frac{\tau_i(\mathbf{y}_j; \Psi^{(k)})}{n} \\ &= \sum_{j=1}^{n} \frac{\tau_{ij}^{(k)}}{n} \qquad i = 1, ..., g \end{aligned} \tag{1.19}$$

Para calcular el resto de los parámetros en la iteración $k + 1$ se maximiza la función Q con respecto a $\boldsymbol{\theta} = (\theta_1, ..., \theta_g)$, es decir, el resto de los parámetros no estimados aún y tomando en cuenta que $\boldsymbol{\pi}$ ya fue estimada en esa iteración por $\boldsymbol{\pi}^{(k+1)}$.

Para encontrar los máximos de la función Q, se deriva con respecto a $\boldsymbol{\theta}$ y se obtienen la raíz de dicha derivada, la cual se obtiene al resolver el sistema de ecuaciones.

$$\begin{aligned}
\frac{\partial Q(\Psi;\Psi^{(k)})}{\partial\theta} &= \frac{\partial\{\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_i(\mathbf{y}_j;\Psi^{(k)})[log\pi_i + logf_i(\mathbf{y}_j,\theta_i)]\}}{\partial\theta} \\
&= \sum_{j=1}^{n}\frac{\partial\{\sum_{i=1}^{g}\tau_i(\mathbf{y}_j;\Psi^{(k)})[log\pi_i^{(k+1)} + logf_i(\mathbf{y}_j,\theta_i)]\}}{\partial\theta} \\
&= \sum_{j=1}^{n}\sum_{i=1}^{g}\frac{\partial\{\tau_i(\mathbf{y}_j;\Psi^{(k)})[log\pi_i^{(k+1)} + logf_i(\mathbf{y}_j,\theta_i)]\}}{\partial\theta} \\
&= \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_i(\mathbf{y}_j;\Psi^{(k)})\frac{\partial[log\pi_i^{(k+1)} + logf_i(\mathbf{y}_j,\theta_i)]}{\partial\theta} \\
&= \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_i(\mathbf{y}_j;\Psi^{(k)})\{\frac{\partial[log\pi_i^{(k+1)}]}{\partial\theta} + \frac{\partial[logf_i(\mathbf{y}_j,\theta_i)]}{\partial\theta}\} \\
&= \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_i(\mathbf{y}_j;\Psi^{(k)})\frac{\partial\log f_i(\mathbf{y}_j,\theta_i)}{\partial\theta} \qquad (1.20)
\end{aligned}$$

Entonces el sistema de ecuaciones que se tiene que resolver es generado por la ecuación 1.21.

$$\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_i(\mathbf{y}_j;\Psi^{(k)})\frac{\partial\log f_i(\mathbf{y}_j,\theta_i)}{\partial\theta} = \mathbf{0} \qquad (1.21)$$

Estos pasos se repiten hasta que la diferencia $L(\Psi^{(k+1)}) - L(\Psi^{(k)})$ sea pequeña, tanto como se establezca. Es decir, se puede establecer una tolerancia para parar las iteraciones.

1.3. Criterios de Selección de Componentes

Cuando se usa el algoritmo E-M, se hace uso del número de componentes como si se supieran cuántos son en realidad. Pero por lo general se desconoce también este valor, es por eso que se han desarrollado ciertos criterios para seleccionar el número de componentes apropiado. Cabe señalar que estos criterios no cuentan con el rigor de una prueba de hipótesis, pero sirven para comparar los modelos con diferente número de componentes.

Los criterios descritos a continuación, tomados de la referencia [17], están basados en la información de Kullback-Leiber. Esta es una medida de similitud entre la verdadera distribución de los datos y la distribución que se le ajusta. Es decir,

$$I\{f(\mathbf{x});f(\mathbf{x},\hat{\Psi})\} = \int f(\mathbf{x})log[f(\mathbf{x})/f(\mathbf{x},\hat{\Psi})]d\mathbf{x} \qquad (1.22)$$

la idea es minimizar esta información, lo cual se logra al maximizar

$$logL(\hat{\Psi}) - b(F) \qquad (1.23)$$

donde $b(F) = E_F\{\frac{1}{n}\sum_{j=1}^{n} loff(\mathbf{x}, \hat{\Psi}) - \int logf(\mathbf{x}, \hat{\Psi})dF(x)\}$ y E_F es el valor esperado usando como distribución de la muestra aleatoria la función de distribución F. Se puede reescribir esto como

$$-2logL(\hat{\Psi}) + 2b(F) \tag{1.24}$$

y ahora la idea es minimizar este valor, entonces una vez calculado este valor para cada uno de los modelos, se elige el modelo donde este valor sea más pequeño.

1.3.1. Criterio de Información de Akaike (AIC)

En 1974 Hirotsugu Akaike [1] propuso una medida de bondad de ajuste del modelo estimado, usando los resultados de la información de Kullback-Leiber, demostró que $b(F)$ tiende al (d) número de parámetros del modelo, por lo tanto el AIC es

$$-2logL(\hat{\Psi}) + 2d \tag{1.25}$$

y se elige el modelo que minimice el valor absoluto de dicha expresión.

1.3.2. Criterio de Información Bayesiana (BIC)

En 1978 Gideon E. Schwarz [21] propuso este criterio de selección para saber si el modelo ajustado a una muestra es aceptable. De lo que se trata es de minimizar

$$-2logL(\hat{\Psi}) + dlogn \tag{1.26}$$

Donde el primer término de ambos criterios es el mismo, de hecho sólo se diferencian por el factor que multiplica a d. En ambos casos se penalizan los modelos con más componentes, es decir, entre mayor sea el número de componentes la segunda parte de cada criterio es mayor y por lo tanto incrementa el valor del criterio.

Note que en los casos donde $\log n > 2$ el criterio de Akaike favorece más un modelo con d número de parámetros que el criterio de Schwarz.

1.4. Mezclas Finitas de Normales

En particular se puede suponer que los datos tienen una distribución normal, es muy común aplicar este supuesto en la práctica. De hecho, muchos de los trabajos que se han realizado en modelos de mezclas, se desarrollan bajo este supuesto. Es por esto que es importante desarrollar un poco la mezcla de normales.

1.4.1. Caso univariado

Como notación se escribirá la función de densidad de cada componente de la mezcla como $\phi(x;\mu_i,\sigma_i)$, la cuál corresponde a la función de densidad de una normal con esperanza μ_i y varianza σ_i^2, es decir:

$$\phi(x;\mu_i,\sigma_i) = (2\pi\sigma_i^2)^{-\frac{1}{2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \qquad i = 1, ..., g \tag{1.27}$$

Entonces el modelo de mezcla se expresa como:

$$\begin{aligned} f(x;\Psi) &= \sum_{i=1}^{g} \pi_i \phi_i(x;\mu_i,\sigma_i) \\ &= \sum_{i=1}^{g} \pi_i (2\pi\sigma_i^2)^{-\frac{1}{2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \end{aligned} \tag{1.28}$$

Obteniéndose los parámetros Ψ usando el método de estimación máximo verosímil para lo cual se usará el algoritmo EM.

Se comienza con el paso E, como el procedimiento de cada una de las iteraciones es el mismo, para efectos de notación se supondrá que se está en la iteración $(k+1)$. En este caso la función Q, es decir, la esperanza de la log-verosimilitud de los datos completos dadas las observaciones y_j, queda expresada como:

$$\begin{aligned} Q(\Psi;\Psi^{(k)}) &= \sum_{j=1}^{n}\sum_{i=1}^{g} \tau_{ij}^{(k)} [log\pi_i + log\phi_i(y_j,\mu_i,\sigma_i)] \\ &= \sum_{j=1}^{n}\sum_{i=1}^{g} \tau_{ij}^{(k)} \{log\pi_i + log[(2\pi\sigma_i^2)^{-\frac{1}{2}} e^{-\frac{(y_j-\mu_i)^2}{2\sigma_i^2}}]\} \end{aligned} \tag{1.29}$$

En el paso M, cuando maximizamos la función Q, se obtienen los valores de los parámetros para la iteración $(k+1)$, en el caso de π_i, como ya se especificó antes, se estima independientemente de los demás parámetros usando la siguiente expresión:

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \frac{\tau_{ij}^{(k)}}{n} \qquad i = 1, ..., g \tag{1.30}$$

donde

$$\begin{aligned}\tau_{ij}^{(k)} &= \frac{\pi_i^{(k)}\phi_i(y_j;\mu_i^{(k)},\sigma_i^{(k)})}{\sum_{i=1}^{g}\pi_i^{(k)}\phi_i(y_j;\mu_i^{(k)},\sigma_i^{(k)})} \\ &= \frac{\pi_i^{(k)}(2\pi\sigma_i^{(k)2})^{-\frac{1}{2}}e^{-\frac{(y_j-\mu_i^{(k)})^2}{2\sigma_i^{(k)2}}}}{\sum_{i=1}^{g}\pi_i^{(k)}(2\pi\sigma_i^{(k)2})^{-\frac{1}{2}}e^{-\frac{(y_j-\mu_i^{(k)})^2}{2\sigma_i^{(k)2}}}}\end{aligned} \tag{1.31}$$

Para calcular el resto de los parámetros se resuelve el siguiente sistema de ecuaciones:

$$\begin{aligned}\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}^{(k)}\frac{\partial\log\phi_i(y_j;\mu_i,\sigma_i)}{\partial\theta} &= \mathbf{0} \\ \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}^{(k)}\frac{\partial\log(2\pi\sigma_i^2)^{-\frac{1}{2}}e^{-\frac{(y_j-\mu_i)^2}{2\sigma_i^2}}}{\partial\theta} &= \mathbf{0}\end{aligned} \tag{1.32}$$

es decir, se deriva con respecto a cada uno de los parámetros correspondientes a cada componente, si tenemos g componentes y cada componente tiene 2 parámetros, entonces se obtienen $2g$ ecuaciones, las cuales son:

$$\left.\begin{aligned}\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}^{(k)}\frac{\partial\log\phi_i(y_j;\mu_i,\sigma_i)}{\partial\mu_h} &= 0 \\ \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}^{(k)}\frac{\partial\log\phi_i(y_j;\mu_i,\sigma_i)}{\partial\sigma_h^2} &= 0\end{aligned}\right\} h=1,...,g \tag{1.33}$$

Note que cuando se deriva la componente i, con respecto al parámetro de la componente h, esto es la derivada de una constante, por lo cual el resultado es cero, esto es:

$$\frac{\partial\log\phi_i(y_j;\mu_i,\sigma_i)}{\partial\theta_h} = \begin{cases}\frac{\partial\log\phi_i(y_j;\mu_i,\sigma_i)}{\partial\theta_h} & \text{si} \quad i=h \\ 0 & \text{si} \quad i\neq h\end{cases} \tag{1.34}$$

donde $\theta_h = (\mu_h, \sigma_h)'$

Cuando se suma sobre el número de componentes, los sumandos donde $i \neq h$ son cero, por lo tanto esta suma se reduce a:

$$\sum_{i=1}^{g}\tau_{ij}^{(k)}\frac{\partial\log\phi_i(y_j;\mu_i,\sigma_i)}{\partial\theta_h} = \tau_{hj}^{(k)}\frac{\partial\log\phi_h(y_j;\mu_h,\sigma_h)}{\partial\theta_h} \tag{1.35}$$

Resolviendo primero para μ_h, se calcula la derivada de la ecuación (1.34) la cual se expresa de la siguiente forma:

$$\begin{aligned}
\frac{\partial \log \phi_h(y_j;\mu_h,\sigma_h)}{\partial \mu_h} &= \frac{\partial \log[(2\pi\sigma_h^2)^{-\frac{1}{2}} e^{-\frac{(y_j-\mu_h)^2}{2\sigma_h^2}}]}{\partial \mu_h} \\
&= \frac{\partial\{[\log(2\pi\sigma_h^2)^{-\frac{1}{2}}] - \frac{(y_j-\mu_h)^2}{2\sigma_i^2}\}}{\partial \mu_h} \\
&= \frac{\partial[-\frac{(y_j-\mu_h)^2}{2\sigma_h^2}]}{\partial \mu_h} \\
&= -\frac{1}{2\sigma_h^2}\frac{\partial (y_j-\mu_h^{(k+1)})^2}{\partial \mu_h} \\
&= \frac{1}{2\sigma_h^2} 2(y_j-\mu_h^{(k+1)}) \\
&= \frac{1}{\sigma_h^2}(y_j-\mu_h^{(k+1)}) \qquad (1.36)
\end{aligned}$$

Usando las ecuaciones (1.35) y (1.36) para resolver la ecuación (1.33) para las μ_h, se tiene:

$$\begin{aligned}
\sum_{j=1}^{n}\sum_{i=1}^{g} \tau_{ij}^{(k)} \frac{\partial \log \phi_i(y_j;\mu_i,\sigma_i)}{\partial \mu_h} &= 0 \\
\sum_{j=1}^{n} \tau_{hj}^{(k)} \frac{1}{\sigma_h^2}(y_j-\mu_h^{(k+1)}) &= 0 \\
\frac{1}{\sigma_h^2}\sum_{j=1}^{n} \tau_{hj}^{(k)}(y_j-\mu_h^{(k+1)}) &= 0 \\
\sum_{j=1}^{n} \tau_{hj}^{(k)}(y_j-\mu_h^{(k+1)}) &= 0 \\
\sum_{j=1}^{n} \tau_{hj}^{(k)} y_j - \sum_{j=1}^{n} \tau_{hj}^{(k)} \mu_h^{(k+1)} &= 0
\end{aligned}$$

$\Rightarrow$

$$\sum_{j=1}^{n} \tau_{hj}^{(k)} y_j = \mu_h^{(k+1)} \sum_{j=1}^{n} \tau_{hj}^{(k)}$$

$\Rightarrow$

$$\mu_h^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)} y_j}{\sum_{j=1}^{n} \tau_{hj}^{(k)}} \qquad (1.37)$$

De manera análoga para encontrar los parámetros σ^2, la derivada de la ecuación (1.34) con respecto a σ_h^2

$$\begin{aligned}
\frac{\partial \log \phi_h(y_j;\mu_h^{(k)},\sigma_h^{(k)})}{\partial \sigma_h^2} &= \frac{\partial \log(2\pi\sigma_h^{(k)2})^{-\frac{1}{2}} e^{-\frac{(y_j-\mu_h^{(k)})^2}{2\sigma_h^{(k)2}}}}{\partial \sigma_h^2} \\
&= \frac{\partial\{[\log(2\pi\sigma_h^{(k)2})^{-\frac{1}{2}}]-\frac{(y_j-\mu_h^{(k)})^2}{2\sigma_i^{(k)2}}\}}{\partial \sigma_h^2} \\
&= -\frac{1}{2}\frac{\partial \log(2\pi\sigma_h^{(k)2})}{\partial \sigma_h^2} - \frac{\partial[\frac{(y_j-\mu_h^{(k)})^2}{2\sigma_i^{(k)2}}]}{\partial \sigma_h^2} \\
&= -\frac{1}{2}\frac{1}{(2\pi_h\sigma_h^{(k+1)2})}2\pi + (y_j-\mu_h^{(k)})^2\frac{2}{4\sigma_h^{(k+1)4}} \\
&= -\frac{1}{2\sigma_h^{(k+1)2}} + \frac{(y_j-\mu_h^{(k)})^2}{2\sigma_h^{(k+1)4}} \\
&= -\frac{1}{2}[\frac{1}{\sigma_h^{(k+1)2}} - \frac{(y_j-\mu_h^{(k)})^2}{\sigma_h^{(k+1)4}}] \qquad (1.38)
\end{aligned}$$

Usando las ecuaciones (1.35) y (1.38) para resolver la ecuación (1.33) para las σ_h^2, se tiene:

$$\begin{aligned}
\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}^{(k)}\frac{\partial \log \phi_i(y_j;\mu_i^{(k)},\sigma_i^{(k)})}{\partial \sigma_h^2} &= 0 \\
\sum_{j=1}^{n}\tau_{hj}^{(k)}\big[-\frac{1}{2}[\frac{1}{\sigma_h^{(k+1)2}} - \frac{(y_j-\mu_h^{(k)})^2}{\sigma_h^{(k+1)4}}]\big] &= 0 \\
\sum_{j=1}^{n}\tau_{hj}^{(k)}[\frac{1}{\sigma_h^{(k+1)2}} - \frac{(y_j-\mu_h^{(k)})^2}{\sigma_h^{(k+1)4}}] &= 0 \\
\sum_{j=1}^{n}\tau_{hj}^{(k)}\frac{1}{\sigma_h^{(k+1)2}} - \sum_{j=1}^{n}\tau_{hj}^{(k)}\frac{(y_j-\mu_h^{(k)})^2}{\sigma_h^{(k+1)4}} &= 0 \\
\sigma_h^{(k+1)2}\sum_{j=1}^{n}\tau_{hj}^{(k)} - \sum_{j=1}^{n}\tau_{hj}^{(k)}(y_j-\mu_h^{(k)})^2 &= 0
\end{aligned}$$

$\Rightarrow$

$$\sigma_h^{(k+1)2}\sum_{j=1}^{n}\tau_{hj}^{(k)} = \sum_{j=1}^{n}\tau_{hj}^{(k)}(y_j-\mu_h^{(k)})^2$$

$\Rightarrow$

$$\sigma_h^{(k+1)2} = \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)} (y_j - \mu_h^{(k)})^2}{\sum_{j=1}^{n} \tau_{hj}^{(k)}} \tag{1.39}$$

Entonces para el caso univariado, los estimadores máximos verosímiles para la mezcla de componentes normales, son los correspondientes a las ecuaciones (1.31) para los parámetros π_h, (1.37) para los parámetros μ_h y (1.39) para los parámetros σ_h^2, para toda $h = 1, ..., g$.

1.4.2. Caso multivariado

La notación que se usará para $f_i(\boldsymbol{x}, \boldsymbol{\theta}_i)$ en este caso será $\phi(\boldsymbol{x}; \boldsymbol{\mu}_i, \Sigma_i)$, es decir se supone que cada componente de la muestra tiene una distribución normal multivariada con esperanza $\boldsymbol{\mu}_i$ y matriz de covarianza Σ_i , es decir, $\boldsymbol{X}_i \sim N_p(\boldsymbol{\mu}_i, \Sigma_i)$ para toda $i = 1, ..., g$, por lo tanto la expresión de la función de densidad de la muestra será

$$\begin{aligned} f(\boldsymbol{x}) &= \sum_{i=1}^{g} \pi_i \phi_i(\boldsymbol{x}; \boldsymbol{\mu}_i, \Sigma_i) \\ &= \sum_{i=1}^{g} \pi_i \mid 2\pi\Sigma_i \mid^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)'\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)} \end{aligned} \tag{1.40}$$

Se obtendrán los parámetros Ψ haciendo uso del agoritmo EM, es decir, se calcularán los estimadores máximos verosímiles.

Se supondrá que estamos en la iteración $(k+1)$. La expresión para calcular la esperanza de la log-verosimilitud de los datos completos, con base en la ecuación (1.17):

$$\begin{aligned} Q(\Psi; \Psi^{(k)}) &= \sum_{j=1}^{n} \sum_{i=1}^{g} \tau_{ij}^{(k)} [log\pi_i + log\phi_i(\boldsymbol{y}_j, \boldsymbol{\mu}_i, \Sigma_i)] \\ &= \sum_{j=1}^{n} \sum_{i=1}^{g} \tau_{ij}^{(k)} [log\pi_i + \log((2\pi)^{-\frac{p}{2}} \mid \Sigma_i \mid^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{y}_j-\boldsymbol{\mu}_i)'\Sigma_i^{-1}(\boldsymbol{y}_j-\boldsymbol{\mu}_i)})] \end{aligned} \tag{1.41}$$

esto corresponde al paso E del algoritmo, para continuar con el algoritmo, se obtienen los estimadores en dicho paso.

Ahora para calcular el estimador de π_i, usamos el resultado de la ecuación (1.19):

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \frac{\tau_{ij}^{(k)}}{n} \qquad i = 1, ..., g \tag{1.42}$$

donde

$$\begin{aligned}\tau_{ij}^{(k)} &= \pi_i^{(k)}\phi_i(\boldsymbol{y}_j;\mu_i^{(k)},\Sigma_i^{(k)})/\sum_{i=1}^{g}\pi_i\phi_i(\boldsymbol{y}_j;\mu_i^{(k)},\Sigma_i^{(k)}) \\ &= \frac{\pi_i^{(k)}(2\pi)^{-\frac{p}{2}} \mid \Sigma_i^{(k)} \mid^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{y}_j-\boldsymbol{\mu}_i^{(k)})'\Sigma_i^{(k)-1}(\boldsymbol{y}_j-\boldsymbol{\mu}_i^{(k)})}}{\sum_{i=1}^{g}\pi_i^{(k)}(2\pi)^{-\frac{p}{2}} \mid \Sigma_i^{(k)} \mid^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{y}_j-\boldsymbol{\mu}_i^{(k)})'\Sigma_i^{(k)-1}(\boldsymbol{y}_j-\boldsymbol{\mu}_i^{(k)})}} \end{aligned} \tag{1.43}$$

para todo $i = 1, ..., g$ y $j = 1, ..., n$

El cálculo del resto de los parámetros se hace resolviendo el siguiente sistema de ecuaciones:

$$\begin{aligned}\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}^{(k)}\frac{\partial \log\phi_i(\boldsymbol{y}_j;\boldsymbol{\mu}_i,\Sigma_i)}{\partial\boldsymbol{\theta}} &= \mathbf{0} \\ \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}^{(k)}\frac{\partial\big[\log[(2\pi)^{-\frac{p}{2}} \mid \Sigma_i \mid^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{y}_j-\boldsymbol{\theta})'\Sigma_i^{-1}(\boldsymbol{y}_j-\boldsymbol{\mu}_i)}]\big]}{\partial\boldsymbol{\theta}} &= \mathbf{0}\end{aligned} \tag{1.44}$$

donde $\boldsymbol{\theta} = \{\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \Sigma_i) i = 1, ..., g\}$ .Es decir, derivamos con respecto a cada uno de los parámetros.

Para cada componente se tienen: p parámetros para la media y $p(p+1)/2$ para la matriz de covarianzas. Entonces para cada componente se tienen $p + p(p+1)/2$ y como hay g componentes, en total se tendrían $g(p + p(p+1)/2)$ ecuaciones.

En ocasiones, para reducir el número de parámetros se hacen suposiciones sobre Σ_k, con base en su descomposición espectral. Para esto se hará uso de algunos resultados tomados del libro *Matrix analysis for statistics* [22].

Teorema 1.1. *Sea $\boldsymbol{A}$ una matriz simétrica de $n \times n$ con valores propios $\lambda_1..., \lambda_n$ y se supone que $\boldsymbol{x}_1..., \boldsymbol{x}_n$ es el conjunto de vectores propios ortonormales asociados a los eigenvalores. Si $\Lambda = diag(\lambda_1..., \lambda_n)$ y $\boldsymbol{X} = (\boldsymbol{x}_1..., \boldsymbol{x}_n)$ entonces*

$$\boldsymbol{A} = \boldsymbol{X}\Lambda\boldsymbol{X}'$$

Como se sabe Σ es una matriz simétrica, se puede usar una descomposición espectral:

$$\Sigma_k = \mathbf{D}_k\Lambda_k\mathbf{D}_k' \tag{1.45}$$

donde Λ_k matriz diagonal con valores propios de Σ_k, $\mathbf{D}_k$ matriz ortonormal de vectores propios asociados a los valores propios.

También se puede expresar a la matriz como:

$$\Sigma_k = \lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k' \tag{1.46}$$

donde $\Lambda_k = \lambda_k\mathbf{A}_k$, λ_k es un valor constante que corresponde al primer valor propio de Σ_k, $\mathbf{A}_k$ es una matriz diagonal $\{a_{1k}, ..., a_{pk}\}$ y $1 = a_{1k} \geq ... \geq a_{pk} > 0$.

Identificador	Modelo
E	σ
V	σ_k
EII	λI
VII	$\lambda_k I$
EEI	λA
VEI	$\lambda_k A$
EVI	λA_k
VVI	$\lambda_k A_k$
EEE	$\lambda DAD'$
EEV	$\lambda D_k A D_k'$
VEV	$\lambda_k D_k A D_k'$
VVV	$\lambda_k D_k A_k D_k'$

Cuadro 1.1: Parametrización de la matriz de covarianzas Σ_k

Cada una de estas partes, determina el comportamiento de las curvas de nivel de la función de densidad, esto es:

$\mathbf{D}_k \longrightarrow$ Orientación (Inclinación)

$\mathbf{A}_k \longrightarrow$ Forma (Esférica/Elipsoide)

$\lambda_k \longrightarrow$ Volumen

El Cuadro 1.1, muestra los diferentes modelos sobre la Σ_k, se trata de ir de lo más sencillo a lo más complejo, fijando λ, D o A para todas las componentes. El modelo que se usa para los cálculos siguientes, es el más general, es decir el VVV .

Cuando se deriva para cada parámetro, lo que se hace es reducir esto a dos derivadas, la correspondiente al vector de medias y la correspondiente a la matriz de covarianzas. Por lo tanto, el sistema de ecuaciones será, como en el caso univariado:

$$\left.\begin{array}{l} \displaystyle\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}^{(k)}\frac{\partial \log \phi_i(\boldsymbol{y}_j;\boldsymbol{\mu}_i,\Sigma_i)}{\partial \boldsymbol{\mu}_h} = \mathbf{0} \\ \displaystyle\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}^{(k)}\frac{\partial \log \phi_i(\boldsymbol{y}_j;\boldsymbol{\mu}_i,\Sigma_i)}{\partial \Sigma_h^{-1}} = 0 \end{array}\right\} h = 1, ..., g \tag{1.47}$$

Al igual que en el caso univariado, cuando se deriva la componente i, con respecto al parámetro de la componente h, se deriva una constante, por lo cual el resultado es cero, esto es:

$$\frac{\partial \log \phi_i(\boldsymbol{y}_j;\boldsymbol{\mu}_i,\Sigma_i)}{\partial \theta_h} = \begin{cases} \frac{\partial \log \phi_i(\boldsymbol{y}_j;\boldsymbol{\mu}_i,\Sigma_i)}{\partial \theta_h} & \text{si} \quad i = h \\ 0 & \text{si} \quad i \neq h \end{cases} \tag{1.48}$$

donde $\theta_h = (\boldsymbol{\mu}_h, \Sigma_h)'$

Cuando se suma sobre el número de componentes, los sumandos donde $i \neq h$ son cero, por lo tanto esta suma la se reduce a:

$$\sum_{i=1}^{g} \tau_{ij}^{(k)} \frac{\partial \log \phi_i(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \Sigma_i)}{\partial \theta_h} = \tau_{hj}^{(k)} \frac{\partial \log \phi_h(\boldsymbol{y}_j; \boldsymbol{\mu}_h, \Sigma_h)}{\partial \theta_h} \tag{1.49}$$

Se desarrollará más la ecuación 1.44, usando el resultado de la ecuación (1.49), se tiene que:

$$\begin{aligned}
& \sum_{j=1}^{n} \sum_{i=1}^{g} \tau_{ij}^{(k)} \frac{\partial[\log(2\pi)^{-\frac{p}{2}} + \log \mid \Sigma_i \mid^{-\frac{1}{2}} - \frac{1}{2}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)'\Sigma_i^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)]}{\partial \theta_h} \\
= & \sum_{j=1}^{n} \tau_{hj}^{(k)} \frac{\partial[\log(2\pi)^{-\frac{p}{2}} + \log \mid \Sigma_h \mid^{-\frac{1}{2}} - \frac{1}{2}(\boldsymbol{y}_j - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_h)]}{\partial \theta_h} \\
= & \frac{\partial\{\sum_{j=1}^{n} \tau_{hj}^{(k)}[-\frac{p}{2}\log(2\pi) - \frac{1}{2}\log \mid \Sigma_h \mid - \frac{1}{2}(\boldsymbol{y}_j - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_h)]\}}{\partial \theta_h} \\
= & \frac{\partial\{-\frac{p}{2}\sum_{j=1}^{n} \tau_{hj}^{(k)} \log(2\pi) - \frac{1}{2}\sum_{j=1}^{n} \tau_{hj}^{(k)} \log \frac{1}{|\Sigma_h^{-1}|} - \frac{1}{2}\sum_{j=1}^{n} \tau_{hj}^{(k)}(\boldsymbol{y}_j - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_h)\}}{\partial \theta_h} \\
= & \frac{\partial\{\frac{1}{2}\sum_{j=1}^{n} \tau_{hj}^{(k)} \log \mid \Sigma_h^{-1} \mid - \frac{1}{2}\sum_{j=1}^{n} \tau_{hj}^{(k)}(\boldsymbol{y}_j - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_h)\}}{\partial \theta_h}
\end{aligned} \tag{1.50}$$

Proposición 1.1. *Es cierto que:*

$$\begin{aligned}
(\boldsymbol{y}_j - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_h) = & (\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h) \\
& + (\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h) \\
& + 2(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h)
\end{aligned}$$

donde $\hat{\boldsymbol{\mu}}_h = \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)} \boldsymbol{y}_j}{\sum_{j=1}^{n} \tau_{hj}^{(k)}}$

Demostración

$$(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h) + (\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h) + 2(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h)$$

$$\begin{aligned}
= & \boldsymbol{y}_j'\Sigma_h^{-1}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h) - \hat{\boldsymbol{\mu}}_h'\Sigma_h^{-1}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h) \\
& + \hat{\boldsymbol{\mu}}_h'\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h) - \boldsymbol{\mu}_h'\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h) \\
& + 2\hat{\boldsymbol{\mu}}_h'\Sigma_h^{-1}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h) - 2\boldsymbol{\mu}_h'\Sigma_h^{-1}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h) \\
= & \boldsymbol{y}_j'\Sigma_h^{-1}\boldsymbol{y}_j - \boldsymbol{y}_j'\Sigma_h^{-1}\hat{\boldsymbol{\mu}}_h - \hat{\boldsymbol{\mu}}_h'\Sigma_h^{-1}\boldsymbol{y}_j + \hat{\boldsymbol{\mu}}_h'\Sigma_h^{-1}\hat{\boldsymbol{\mu}}_h \\
& + \hat{\boldsymbol{\mu}}_h'\Sigma_h^{-1}\hat{\boldsymbol{\mu}}_h - \hat{\boldsymbol{\mu}}_h'\Sigma_h^{-1}\boldsymbol{\mu}_h - \boldsymbol{\mu}_h'\Sigma_h^{-1}\hat{\boldsymbol{\mu}}_h + \boldsymbol{\mu}_h'\Sigma_h^{-1}\boldsymbol{\mu}_h \\
& + 2\hat{\boldsymbol{\mu}}_h'\Sigma_h^{-1}\boldsymbol{y}_j - 2\hat{\boldsymbol{\mu}}_h'\Sigma_h^{-1}\hat{\boldsymbol{\mu}}_h - 2\boldsymbol{\mu}_h\Sigma_h^{-1}\boldsymbol{y}_j + 2\boldsymbol{\mu}_h\Sigma_h^{-1}\hat{\boldsymbol{\mu}}_h
\end{aligned}$$

Proposición 1.2. *Sea* $\boldsymbol{x}$ *y* $\boldsymbol{y}$ *vectores de* $p \times 1$ *y* $\boldsymbol{A}$ *una matriz de* $p \times p$*, entonces* $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{A}'\boldsymbol{x}$

Para eliminar términos equivalentes se hace uso de la proposición 1.2. En este caso es válido ya que Σ_h es una matriz simétrica, por lo tanto su inversa también lo es, es decir, $\Sigma_h^{-1} = (\Sigma_h^{-1})'$.

Una vez eliminados los términos equivalentes se tiene que:

$$\begin{aligned}
& \boldsymbol{y}_j'\Sigma_h^{-1}\boldsymbol{y}_j + \boldsymbol{\mu}_h'\Sigma_h^{-1}\boldsymbol{\mu}_h - 2\boldsymbol{\mu}_h\Sigma_h^{-1}\boldsymbol{y}_j \\
= \; & \boldsymbol{y}_j'\Sigma_h^{-1}\boldsymbol{y}_j - 2\boldsymbol{\mu}_h'\Sigma_h^{-1}\boldsymbol{y}_j + \boldsymbol{\mu}_h'\Sigma_h^{-1}\boldsymbol{\mu}_h \\
= \; & (\boldsymbol{y}_j - \boldsymbol{\mu}_h)'\Sigma_h^{-1}\boldsymbol{y}_j - \boldsymbol{\mu}_h'\Sigma_h^{-1}\boldsymbol{y}_j + \boldsymbol{\mu}_h'\Sigma_h^{-1}\boldsymbol{\mu}_h \\
= \; & (\boldsymbol{y}_j - \boldsymbol{\mu}_h)'\Sigma_h^{-1}\boldsymbol{y}_j - \boldsymbol{\mu}_h'\Sigma_h^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_h) \\
= \; & (\boldsymbol{y}_j - \boldsymbol{\mu}_h)'\Sigma_h^{-1}\boldsymbol{y}_j - \boldsymbol{\mu}_h(\boldsymbol{y}_j - \boldsymbol{\mu}_h)'\Sigma_h^{-1}\boldsymbol{\mu}_h \\
= \; & (\boldsymbol{y}_j - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_h)
\end{aligned}$$

♢

Regresando al problema de la ecuación (1.50). Cuando se suma el resultado de la proposición 1.1, sobre todas las observaciones se tiene que:

$$\sum_{j=1}^{n} \tau_{hj}^{(k)}[(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h) + (\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h) + 2(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h)]$$

$$\begin{aligned}
= \; & \sum_{j=1}^{n} \tau_{hj}^{(k)}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h) + \sum_{j=1}^{n} \tau_{hj}^{(k)}[(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)] \\
& +2(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)'\Sigma_h^{-1}\sum_{j=1}^{n} \tau_{hj}^{(k)}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h) \\
= \; & \sum_{j=1}^{n} \tau_{hj}^{(k)}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h) + \sum_{j=1}^{n} \tau_{hj}^{(k)}[(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)] \\
& +2(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)'\Sigma_h^{-1}\Big(\sum_{j=1}^{n} \tau_{hj}^{(k)}\boldsymbol{y}_j - \sum_{j=1}^{n} \tau_{hj}^{(k)}\frac{\sum_{j=1}^{n} \tau_{hj}^{(k)}\boldsymbol{y}_j}{\sum_{j=1}^{n} \tau_{hj}^{(k)}}\Big) \\
= \; & \sum_{j=1}^{n} \tau_{hj}^{(k)}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h) + \sum_{j=1}^{n} \tau_{hj}^{(k)}[(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)] \\
& +2(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)'\Sigma_h^{-1}\Big(\sum_{j=1}^{n} \tau_{hj}^{(k)}\boldsymbol{y}_j - \sum_{j=1}^{n} \tau_{hj}^{(k)}\boldsymbol{y}_j\Big) \\
= \; & \sum_{j=1}^{n} \tau_{hj}^{(k)}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_h) + \sum_{j=1}^{n} \tau_{hj}^{(k)}[(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)] \qquad (1.51)
\end{aligned}$$

La ecuación (1.51) es un escalar, en el momento de aplicarle la función traza, no se afecta. Entonces,

$$\begin{aligned}
&\sum_{j=1}^{n}\tau_{hj}^{(k)}(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}}_h)+\sum_{j=1}^{n}\tau_{hj}^{(k)}[(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)'\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)]\\
=\;& tr(\sum_{j=1}^{n}\tau_{hj}^{(k)}(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}}_h)'\Sigma_h^{-1}(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}}_h))+\sum_{j=1}^{n}\tau_{hj}^{(k)}tr\{[(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)'\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)]\}\\
=\;& tr(\sum_{j=1}^{n}\tau_{hj}^{(k)}(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}}_h)(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}}_h)'\Sigma_h^{-1})+\sum_{j=1}^{n}\tau_{hj}^{(k)}tr\{[(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)'\Sigma_h^{-1}]\}\\
=\;& tr(\sum_{j=1}^{n}\tau_{hj}^{(k)}(\Sigma_h^{-1}(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}}_h)(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}}_h)')+\sum_{j=1}^{n}\tau_{hj}^{(k)}tr\{[\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)']\}\\
=\;& tr(\Sigma_h^{-1}\sum_{j=1}^{n}\tau_{hj}^{(k)}(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}}_h)(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}}_h)')+\sum_{j=1}^{n}\tau_{hj}^{(k)}tr\{[\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)']\} \qquad (1.52)
\end{aligned}$$

Se llamará $\frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}}_h)(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}}_h)'}{\sum_{j=1}^{n}\tau_{hj}^{(k)}}=\boldsymbol{S}_h$, esto para efectos de espacio y además es un resultado que servirá en la demostración.

Retomando la ecuación (1.50) y con el resultado de las ecuaciones (1.51) y (1.52) se tiene:

$$\begin{aligned}
&\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}^{(k)}\frac{\partial\log\phi_i(\boldsymbol{y}_j;\boldsymbol{\mu}_i,\Sigma_i)}{\partial\theta_h}\\
=\;& \frac{\partial\{\frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}}{2}\log\mid\Sigma_h^{-1}\mid-\frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}}{2}tr(\Sigma_h^{-1}\boldsymbol{S}_h)\}}{\partial\theta_h}\\
-\;& \frac{\partial\{\frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}}{2}[tr[\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)']]\}}{\partial\theta_h} \qquad (1.53)
\end{aligned}$$

Antes de seguir con el desarrollo se enunciarán algunos resultados que serán útiles, éstos se pueden ver en el libro *Multivariate Analysis* [16].

Proposición 1.3. *Sea $\boldsymbol{x}$ vectores de $p\times 1$ y $\boldsymbol{A}$ una matriz de $p\times p$, entonces*

$$\frac{\partial\boldsymbol{x}'\boldsymbol{x}}{\partial\boldsymbol{x}}=2\boldsymbol{x}$$

y

$$\frac{\partial\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial\boldsymbol{x}}=(\boldsymbol{A}+\boldsymbol{A}')\boldsymbol{x}$$

Proposición 1.4. *Sea* $\boldsymbol{X}$ *una matriz de* $p \times p$*, entonces:*

$$\begin{aligned}\frac{\partial \mid \boldsymbol{X} \mid}{\partial x_{ij}} &= X_{ij} \quad \textit{si todos los elementos de} \quad \boldsymbol{X} \quad \textit{son distintos} \\ &= \left\{ \begin{array}{ll} X_{ij}, & i = j \\ 2X_{ij}, & i \neq j \end{array} \right\} \quad \textit{si} \quad \boldsymbol{X} \quad \textit{es simétrica}\end{aligned}$$

donde X_{ij} *es el* $(i,j)th$ *cofactor de* $\boldsymbol{X}$*.*

Proposición 1.5. *Sea* $\boldsymbol{X}$ *y* $\boldsymbol{Y}$*, matrices, entonces:*

$$\begin{aligned}\frac{\partial tr\boldsymbol{XY}}{\partial \boldsymbol{X}} &= \boldsymbol{Y}' \quad \textit{si todos los elementos de} \quad \boldsymbol{X}(n \times k) \quad \textit{son distintos} \\ &= \boldsymbol{Y} + \boldsymbol{Y}' - Diag(\boldsymbol{Y}) \quad \textit{si} \quad \boldsymbol{X}(n \times n) \quad \textit{es simétrica}\end{aligned}$$

Ahora se deriva con respecto a $\boldsymbol{\mu}_h$ y Σ_h^{-1} por separado, para obtener sus estimadores. Primero con respecto a $\boldsymbol{\mu}_h$:

$$\begin{aligned}&\frac{\partial\{\frac{\sum_{j=1}^n \tau_{hj}^{(k)}}{2} \log \mid \Sigma_h^{-1} \mid - \frac{\sum_{j=1}^n \tau_{hj}^{(k)}}{2} tr(\Sigma_h^{-1}\boldsymbol{S}) - \frac{\sum_{j=1}^n \tau_{hj}^{(k)}}{2}[(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)]\}}{\partial \boldsymbol{\mu}_h} \\ = \;& \frac{\partial\{-\frac{\sum_{j=1}^n \tau_{hj}^{(k)}}{2}[(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)]\}}{\partial \boldsymbol{\mu}_h} \\ = \;& -\frac{\sum_{j=1}^n \tau_{hj}^{(k)}}{2} \frac{\partial[(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)'\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)]}{\partial \boldsymbol{\mu}_h} \\ & \text{usando la Proposición 1.3 y la regla de la cadena} \\ = \;& -\frac{\sum_{j=1}^n \tau_{hj}^{(k)}}{2}(\Sigma_h^{-1} + \Sigma_h^{-1'})(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)\frac{\partial(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)}{\partial \boldsymbol{\mu}_h} \\ = \;& \frac{\sum_{j=1}^n \tau_{hj}^{(k)}}{2} 2\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h) \\ = \;& \Sigma_h^{-1}\sum_{j=1}^n \tau_{hj}^{(k)}(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h) \\ = \;& \Sigma_h^{-1}\{\sum_{j=1}^n \tau_{hj}^{(k)}\hat{\boldsymbol{\mu}}_h - \sum_{j=1}^n \tau_{hj}^{(k)}\boldsymbol{\mu}_h\} \qquad (1.54)\end{aligned}$$

igualando a cero y despejando el estimador:

$$\begin{aligned}\Sigma_h^{-1}\{\sum_{j=1}^n \tau_{hj}^{(k)}\hat{\boldsymbol{\mu}}_h - \sum_{j=1}^n \tau_{hj}^{(k)}\boldsymbol{\mu}_h^{(k+1)}\} &= \mathbf{0} \\ \sum_{j=1}^n \tau_{hj}^{(k)}\hat{\boldsymbol{\mu}}_h - \sum_{j=1}^n \tau_{hj}^{(k)}\boldsymbol{\mu}_h^{(k+1)} &= \mathbf{0}\end{aligned}$$

$\Rightarrow$

$$\sum_{j=1}^{n} \tau_{hj}^{(k)} \hat{\boldsymbol{\mu}}_h = \sum_{j=1}^{n} \tau_{hj}^{(k)} \boldsymbol{\mu}_h^{(k+1)}$$

$\Rightarrow$

$$\begin{aligned} \boldsymbol{\mu}_h^{(k+1)} &= \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)} \hat{\boldsymbol{\mu}}_h}{\sum_{j=1}^{n} \tau_{hj}^{(k)}} \\ &= \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)} \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)} \boldsymbol{y}_j}{\sum_{j=1}^{n} \tau_{hj}^{(k)}}}{\sum_{j=1}^{n} \tau_{hj}^{(k)}} \\ &= \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)} \boldsymbol{y}_j}{\sum_{j=1}^{n} \tau_{hj}^{(k)}} \end{aligned} \tag{1.55}$$

Este estimador es precisamente el que se había denotado por $\hat{\boldsymbol{\mu}}_h$, es decir, $\boldsymbol{\mu}_h^{(k+1)} = \hat{\boldsymbol{\mu}}_h$.

Ahora derivando con respecto a Σ_h^{-1}

$$\begin{aligned} & \frac{\partial \{ \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)}}{2} \log \mid \Sigma_h^{-1} \mid - \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)}}{2} tr(\Sigma_h^{-1} \boldsymbol{S}) - \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)}}{2} tr[\Sigma_h^{-1} (\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)'] \}}{\partial \Sigma_h^{-1}} \\ = & \frac{\partial \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)}}{2} \log \mid \Sigma_h^{-1} \mid}{\partial \Sigma_h^{-1}} - \frac{\partial \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)}}{2} tr(\Sigma_h^{-1} \boldsymbol{S})}{\partial \Sigma_h^{-1}} - \frac{\partial \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)}}{2} tr[\Sigma_h^{-1} (\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)']}{\partial \Sigma_h^{-1}} \\ = & \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)}}{2} \underbrace{\frac{\partial \log \mid \Sigma_h^{-1} \mid}{\partial \Sigma_h^{-1}}}_{(1)} - \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)}}{2} \underbrace{\frac{\partial tr(\Sigma_h^{-1} \boldsymbol{S})}{\partial \Sigma_h^{-1}}}_{(2)} \\ & - \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)}}{2} \underbrace{\frac{\partial tr[\Sigma_h^{-1} (\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)']}{\partial \Sigma_h^{-1}}}_{(3)} \end{aligned} \tag{1.56}$$

Se desarrollará por separado la derivada de cada uno de los sumandos:
Usando la Propisición 1.4

$$(1) = \frac{1}{\mid \Sigma_h^{-1} \mid} \frac{\partial \mid \Sigma_h^{-1} \mid}{\partial \Sigma_h^{-1}}$$

cuando se deriva con respecto a cada elemento de Σ_h^{-1} se tiene que:

$$\frac{\partial \mid \Sigma_h^{-1} \mid}{\partial \Sigma^{-1}_{h(i,j)}} = \begin{cases} V_{ii} & i = j \\ 2V_{ij} & i \neq j \end{cases} \tag{1.57}$$

donde V_{ij} es el $(i,j)th$ cofactor de Σ_h^{-1} $\quad i,j = 1,...,p$

$$\Rightarrow$$

$$(1) = \begin{cases} \dfrac{V_{ii}}{\mid \Sigma_h^{-1} \mid} & i = j \\ \dfrac{2V_{ij}}{\mid \Sigma_h^{-1} \mid} & i \neq j \end{cases} \tag{1.58}$$

como Σ_h^{-1} es simetrica, la matriz con elementos $\frac{V_{ij}}{|\Sigma_h^{-1}|}$ es igual a $(\Sigma_h^{-1})^{-1} = \Sigma_h$

$$\Rightarrow$$

$$(1) = 2\Sigma_h - Diag\Sigma_h \tag{1.59}$$

Por como está construida $\boldsymbol{S}_h$, es una matriz simética. Ahora desarrollando la segunda derivada y usando la Proposición 1.5:

$$\begin{aligned} (2) &= (\boldsymbol{S}_h + \boldsymbol{S}_h') - Diag\boldsymbol{S}_h \\ &= 2\boldsymbol{S}_h - Diag\boldsymbol{S}_h \end{aligned} \tag{1.60}$$

Finalmente se desarrolla para la última derivada usando de igual manera la Proposición 1.5:

$$\begin{aligned} (3) &= (\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)' + ((\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)')' - Diag(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)' \\ &= (\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)' + ((\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)')'(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)' - Diag(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)' \\ &= 2(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)' - Diag(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}_h)' \end{aligned} \tag{1.61}$$

Retomando la ecuación (1.56) y usando los resultados de las ecuaciones (1.59), (1.60) y (1.61) se tiene:

$$
\begin{aligned}
& \frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}}{2}\frac{\partial \log \mid \Sigma_h^{-1} \mid}{\partial \Sigma_h^{-1}} - \frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}}{2}\frac{\partial tr(\Sigma_h^{-1}\boldsymbol{S})}{\partial \Sigma_h^{-1}} \\
& -\frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}}{2}\frac{\partial tr[\Sigma_h^{-1}(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)']}{\partial \Sigma_h^{-1}} \\
= \quad & \frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}}{2}(2\Sigma_h - Diag\Sigma_h) \\
& -\frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}}{2}(2\boldsymbol{S}_h - Diag\boldsymbol{S}_h) \\
& -\frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}}{2}(2(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)' - Diag(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)') \\
= \quad & \frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}}{2}[2(\Sigma_h - \boldsymbol{S}_h - (\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)') \\
& -Diag(\Sigma_h - \boldsymbol{S}_h - (\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h)')] \qquad (1.62)
\end{aligned}
$$

Igualando a cero y se tiene:

$$
\begin{aligned}
\frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}}{2}[2(\Sigma_h^{(k+1)} - \boldsymbol{S}_h - (\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h^{(k+1)})(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h^{(k+1)})') & \\
-Diag(\Sigma_h^{(k+1)} - \boldsymbol{S}_h - (\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h^{(k+1)})(\hat{\boldsymbol{\mu}}_h-\boldsymbol{\mu}_h^{(k+1)})')] &= \mathbf{0} \\
\frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}}{2}[2(\Sigma_h^{(k+1)} - \boldsymbol{S}_h) - Diag(\Sigma_h^{(k+1)} - \boldsymbol{S}_h)] &= \mathbf{0} \\
2(\Sigma_h^{(k+1)} - \boldsymbol{S}_h) - Diag(\Sigma_h^{(k+1)} - \boldsymbol{S}_h) &= \mathbf{0}
\end{aligned}
$$

$\Rightarrow$

$$
\begin{aligned}
\Sigma_h^{(k+1)} - \boldsymbol{S}_h &= \mathbf{0} \\
\Sigma_h^{(k+1)} &= \boldsymbol{S}_h \\
&= \frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}}_h)(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}}_h)'}{\sum_{j=1}^{n}\tau_{hj}^{(k)}} \\
&= \frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}(\boldsymbol{y}_j-\boldsymbol{\mu}_h^{(k+1)})(\boldsymbol{y}_j-\boldsymbol{\mu}_h^{(k+1)})'}{\sum_{j=1}^{n}\tau_{hj}^{(k)}} \qquad (1.63)
\end{aligned}
$$

Entonces para el caso multivariado, los estimadores máximos verosímiles para la mezcla de componentes normales, son los correspondientes a las ecuaciones (1.43) para los parámetros π_h, (1.55) para los parámetros $\boldsymbol{\mu}_h$ y (1.63) para los parámetros Σ_h, para todo $h = 1, ..., g$.

1.5. Circunstancias en las Mezclas

Una vez que se conoce este modelo estadístico de mezclas finitas, una pregunta importante que surge es ¿cómo saber cuándo usar este modelo?, en el caso univariado y bivariado se podría decir que una manera de identificar este modelo, es graficando un histograma de las observaciones, si en éste se notan varias modas o jorobas es aceptable usar este tipo de modelos, en donde cada joroba correspondería a una componente del mismo.

Algunos ejemplos se tienen en la Figura 1.1, en esta se muestran algunas funciones de densidad donde son notorias las componentes, si se supone que los datos se pueden modelar por medio de mezclas, el histograma se podría aproximar por alguna densidad parecida a éstas.

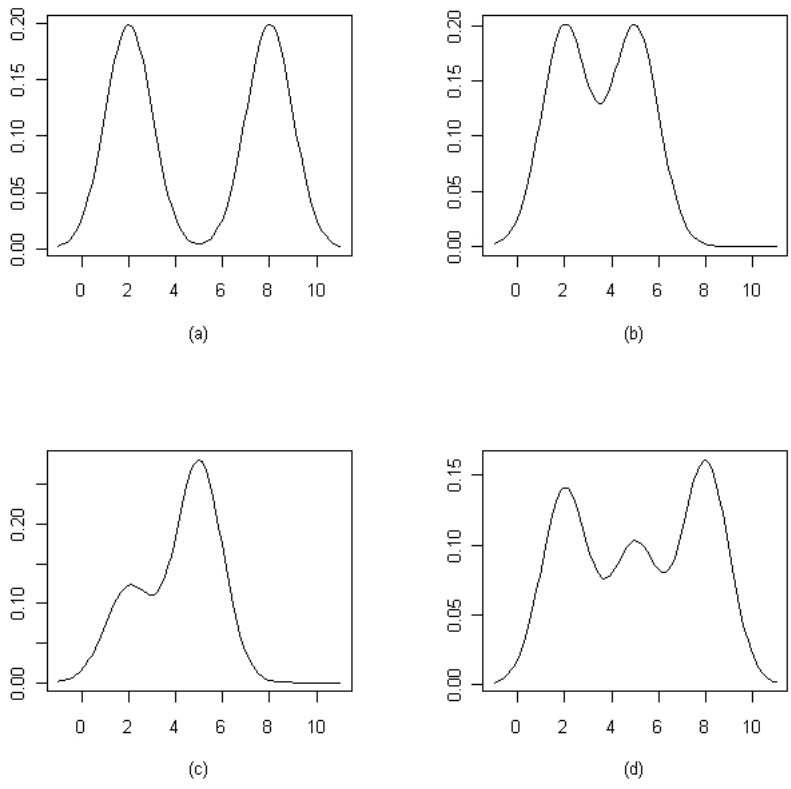

Figura 1.1: Plot de mezclas de densidades normales a) $0.5N(2,1) + 0.5N(8,1)$; b) $0.5N(2,1)+0.5N(5,1)$; c) $0.3N(2,1)+0.7N(5,1)$; d) $0.35N(2,1)+0.25N(5,1)+0.4N(8,1)$

Otra manera de saber si es factible usar un modelo de mezclas a través de la observación del histograma, esto es si dicha gráfica es asimétrica como la que se en la Figura 1.2. En ésta se muestran dos ejemplos de densidades asimétricas pertenecientes a una mezcla finita de densidades. Se puede pensar que la razón de esta asimetría es la cercanía entre las medias de cada densidad, que en estos casos así es, sin embargo esto no es argumento suficiente para afirmarlo.

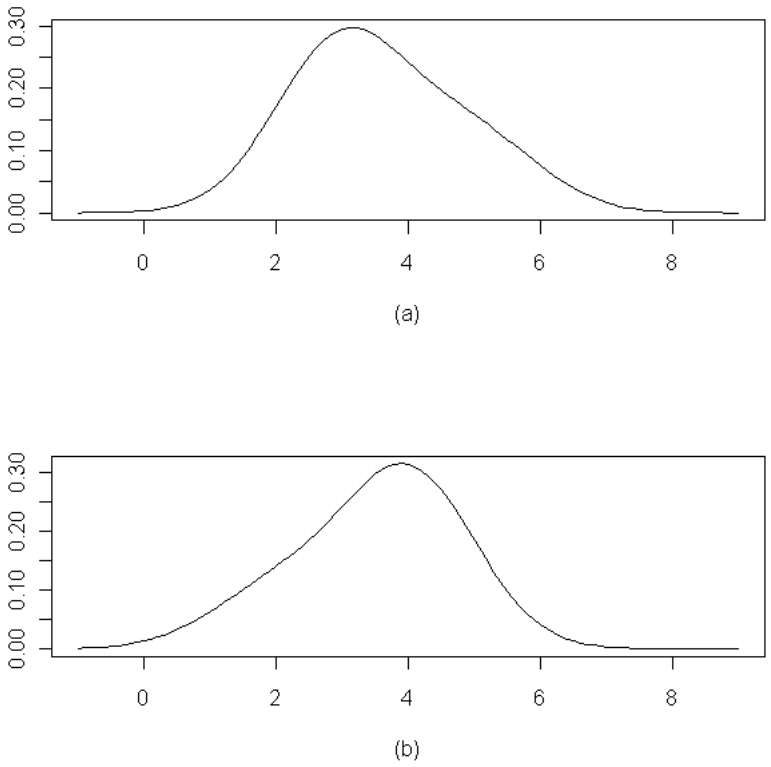

Figura 1.2: Plot de mezclas de densidades normales: a) $0.7N(3,1) + 0.3N(5,1)$; b) $0.75N(4,1) + 0.25N(2,1)$

Cuando existe asimetría en la muestra, además de poder suponer un modelo de mezclas, se puede argumentar una distribución log-normal o alguna otra que no sea simétrica. Esto lo ejemplificó Titterington (1985), él muestra gráficamente como una densidad log-normal con parámetros $\mu = log(10)$ y $\sigma^2 = 0.04$, es muy semejante a una mezcla de densidades normales univariadas expresada como $f(y) = 0.9\phi(y; 9.5, 2.5) + 0.1\phi(y; 13.5, 2.5)$. Ambas gráficas se muestran en la Figura 1.3, en donde además se muestra el histograma de una simulación de 1000 observaciones log-normales con los parámetros antes mencionados.

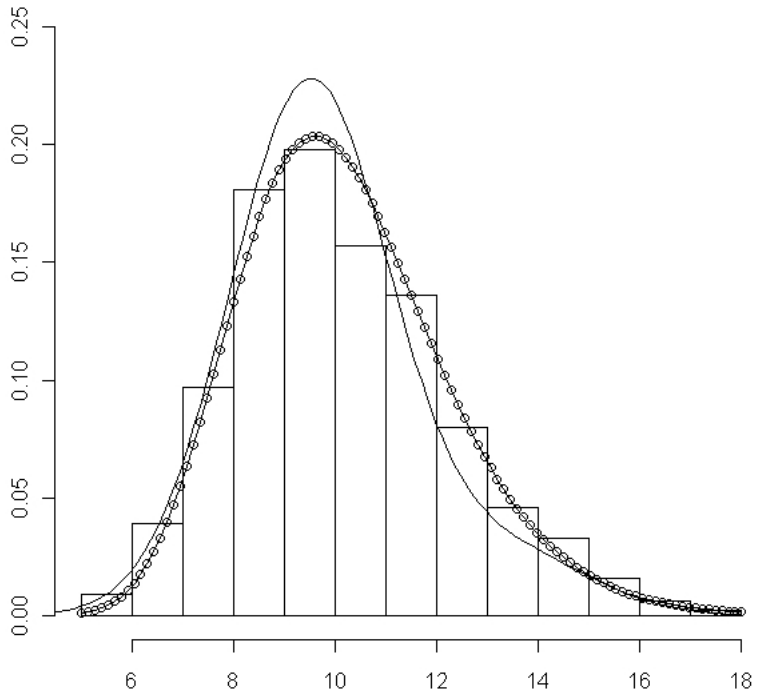

Figura 1.3: Histograma de 1000 observaciones lognormales con $\mu = \log(10)$ y $\sigma^2 = 0.04$ La línea sólida corresponde a la mezcla de normales y la línea con círculos corresponde a la log-normal

En este caso se ve que un modelo de mezclas puede ajustar bien una base de datos, sin embargo esto no implica que sea el mejor modelo. Es factible que otro modelo sea mejor en este caso, sobre todo porque no existe como tal una multimodalidad. Se podría pensar que una vez que se detectan jorobas en el histograma, esto implica que estos datos surgen de una modelo de mezclas, pero esto tampoco implica dicha afirmación.

Un ejemplo de esto lo hizo Day (1969), donde genera una muestra aleatoria $x_i, i = 1, ..., 50$ de normales $p = 10$ variadas y grafica el histograma de una nueva base de datos univariados de la forma $\mathbf{a}'\mathbf{x}_1, ..., \mathbf{a}'\mathbf{x}_{50}$, donde $\mathbf{a} = \hat{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)$, estos parámetros son los que se obtendrían si se supone que la muestra de normales se puede ver como una mezcla de dos normales multivariadas. En el histograma se aprecia multimodalidad, sin embargo esta muestra no surge de un modelo de mezclas.

Una simulación parecida se encuentra a continuación.

Ejemplo 1.1. Primero se generó una muestra aleatoria de 100 normales multivariadas de dimensión p=20 con $\boldsymbol{\mu} = \mathbf{0}$ y $\Sigma_0 = I$, $\mathbf{x}_j \sim N(\mathbf{0}, I)$, para toda $j = 1, ..., 100$. Haciendo uso del algoritmo EM, se encontraron los parámetros que tendría estas observaciones suponiendo un modelo de mezclas de dos normales de dimensión p=20 homocedásticas. Después se calcula un vector $\mathbf{a} = \hat{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)$, donde $\hat{\boldsymbol{\mu}}_g$, $g = 1, 2$ corresponden a los vectores de medias de cada uno de los componentes y $\hat{\Sigma}$ es la matriz de covarianza para ambas componentes. Una vez que se obtienen estos resultados se calculó una nueva muestra con base en la anterior, esto es $\mathbf{a}'\mathbf{x}_1, ..., \mathbf{a}'\mathbf{x}_{100}$. Finalmente la gráfica del el histograma de esta aparece en la Figura 1.4.

Por la forma en que se construyó esta muestra se sabe que no es una mezcla de funciones, sin embargo al graficar el histograma de estas observaciones se nota que existe multimodalidad, lo que haría pensar que es factible ajustar un modelo de mezclas. Casos como estos se podrán presentar en la base de datos, es por eso bueno hacerse de algún criterio que ayude a elegir el mejor modelo que ajusta los datos.

Conclusiones

El modelo de mezclas finitas es aquel cuya densidad es una suma ponderada de densidades. Estas densidades pertenecen a la misma familia, ya que se está suponiendo que pertenecen a una misma población, quizás con características diferentes. Po esta razón es bueno usarlo cuando se observan jorobas en el histograma de las observaciones.

Una vez que se ha decidido usar este modelo para ajustar alguna base de datos, se deben encontrar los parámetros del mismo. Anteriormente se usó el método de momentos para encontrarlos, pero las investigaciones demostraron que el método de máxima verosimilitud es mejor. Para encontrar estos estimadores por medio de este método, usamos el algoritmo de Esperanza-Maximización (EM). Si nosotros suponemos que los datos siguen una distribución normal, el cálculo para encontrar los estimadores es relativamente fácil.

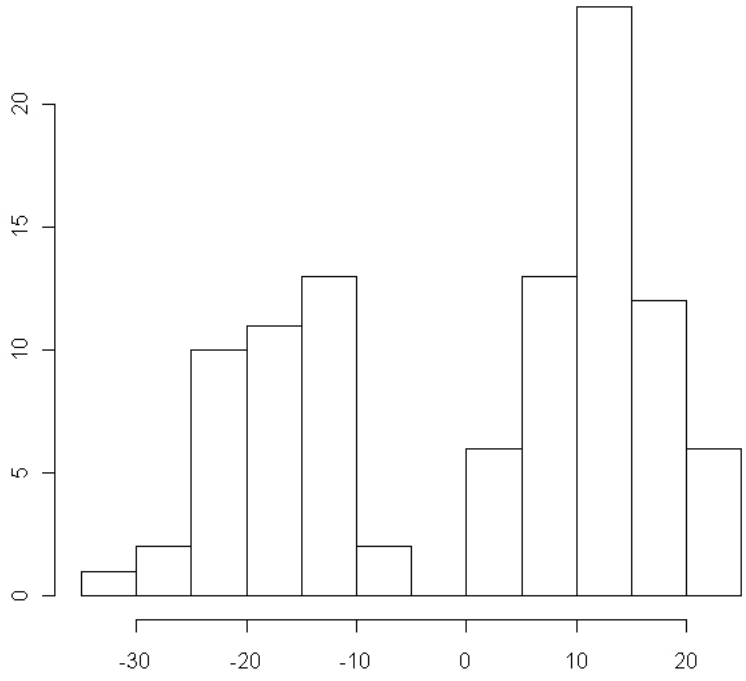

Figura 1.4: Histograma de las observaciones $\mathbf{a}'\mathbf{x}_1, ..., \mathbf{a}'\mathbf{x}_{100}$

Capítulo 2

Estimación de Densidad

Antes de hablar de una función de densidad se verán algunas nociones básicas de probabilidad.

Sea Ω un espacio muestral, es decir, el conjunto donde se agrupan todos los posibles resultados que puede tomar un experimento aleatorio.

Sea $\mathcal{F}$ la σ- álgebra de eventos, es una colección no vacía de subconjuntos de Ω, la cual cumple con lo siguiente:

1. $\Omega \epsilon \mathcal{F}$
2. $A \epsilon F \Rightarrow A^c \epsilon \mathcal{F}$
3. Si $A_1, A_2, ... \epsilon \mathcal{F} \Rightarrow \cup_{n=1}^{\infty} A_n \epsilon \mathcal{F}$

P es una medida de probabilidad si cumple lo siguiente

1. $P : \mathcal{F} \rightarrow [0, 1]$
2. $P(\Omega) = 1$
3. $P(A) \geq 0 \quad \forall \quad A \epsilon \mathcal{F}$
4. Si $A_1, A_2, ... \epsilon \mathcal{F} A_i \cap A_j = \emptyset \quad \forall \quad i \neq j$
   $\Rightarrow P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Una variable aleatoria $(v.a)$ es una función $X : \Omega \rightarrow \mathbb{R}^p$ tal que $\{w \epsilon \Omega : X(w) \leq x\} \epsilon \mathcal{F}$ $\forall \quad x \epsilon \mathbb{R}^p$

Entonces se llama función de distribución de X $v.a$ en $(\Omega, \mathcal{F}, P)$ a

$$\begin{aligned} F_X(x) &= P(\{w \epsilon \Omega : X(w) \leq x\}) \\ &= P(X \leq x) \end{aligned}$$

la cuál debe cumplir con lo siguiente:

1. F_X es no decreciente, si $a \leq b \Rightarrow F_X(a) \leq F_X(b)$

2. Continua por la derecha, $\underset{h \to 0^+}{lím} F_X(x+h) = F_X(x)$

3. $\underset{x \to \infty}{lím} F_X(x) = 1$ y $\underset{x \to -\infty}{lím} F_X(x) = 0$

Finalmente la función de densidad es $f_X(x) = \frac{d}{dx} F_X(x)$, esto en el caso continuo pero en el caso discreto es la probabilidad de que la variable aleatoria tome cierto valor, es decir, $f_x(x) = P(X = x)$,. Las propiedades que debe cumplir esta función son:

1. $f_X(x) \geq 0 \ \forall x \epsilon X(\Omega)$

2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$ en el caso continuo

3. $\sum_{x \epsilon X(\Omega)} f_X(x) = 1$ en el caso discreto

Es importante conocer el comportamiento de la función de densidad ya que con ella podemos describir el comportamiento de la población, podemos predecir la ocurrencia de ciertos eventos, como el valor esperado.

Es por eso que existen diferentes técnicas para poder identificar la manera en que se distribuyen los datos. Comenzando quizás con un análisis descriptivo que ayuda a orientar, de una manera visual, sobre su distribución y después alguna prueba de hipótesis que sirva para saber si la suposición es correcta.

En ocasiones se tiene conocimiento de la manera en que se distribuyen, ya sea de manera normal multivariada o como una exponencial, pero no se sabe nada de los páramentros con los que trabaja la función de densidad, en otras ocasiones no se conoce la forma en que se distribuye la muestra, aunque se podría tener conocimiento de la esperanza de la muestra o de la matriz de covarianzas, y con estas estimar los parámetros con base en algún método como por ejemplo el de máxima verosimilitud.

Cuando no se sabe la distribución a la que pertenece se pueden hacer suposiciones con base en un análisis descriptivo o por investigaciones anteriores, para tener una mayor certeza de esta afirmación se puede hacer uso de pruebas de hipótesis como las que se mencionarán a lo largo de este capítulo.

2.1. Análisis Descriptivo

Un análisis descriptivo es la primera etapa para desarrollar un análisis de los datos a través de tablas, gráfico y medidas de resumen. Como ya se mencionó antes, este análisis sólo ayudará a tener una idea del comportamiento de los datos, así como una mejor visualización de los mismos. Algunos procedimientos se mencionan a continuación.

2.1.1. Distribución Empírica

La distribución empírica es la descripción del comportamiento de la muestra con base en las observaciones registradas. Cuando existe una muestra representativa de la población la descripción de estas observaciones acerca mucho al comportamiento general de la población, es por eso que es importante conocerla. Antes de hablar de la distribución empírica, es importante hablar de lo siguiente.

Frecuencias

Dada una muestra $x_1, ..., x_n$ cada observación se asocia a un grupo k, esta asignación se hace con base al valor de la observación. Podemos definir k intervalos en el campo de los números reales, si la observación esta dentro del intervalo k, entonces la asignamos al mismo. Al número de observaciones dentro de cada intervalo se le llama frecuencia, éstas se denotan como g_i con $i = 1, ..., k$, teniendo como característica que $\sum_{i=1}^{k} g_i = n$.

Frecuencias Relativas

Es el cociente entre la frecuencia g_i y el número de total de observaciones n. A estas las denotaremos como $g'_i = \frac{g_i}{n}$, una característica es que $0 \leq g'_i \leq 1$, en este caso $\sum_{i=1}^{k} g'_i = 1$. A la asociación entre el intervalo k y su frecuencia relativa g'_k se le conoce como función de densidad empírica.

Frecuencias Relativas Acumuladas

La frecuencia relativa acumulada asociada a k es $G_k = \sum_{i=1}^{k} g'_i$, es la suma de las frecuencias relativas acumuladas hasta el intervalo k. Análogo a lo anterior a la asociación entre el intervalo k y su frecuencia relativa acumulada G_k se le puede llamar la función de distribución empírica.

Entonces la **Distribución Empírica**, es el comportamiento de la muestra con base en las frecuencia relativas, es decir, se puede hablar de que la población tiene la distribución obtenida de dichas frecuencias.

2.1.2. Medidas

Las medidas estadísticas pretenden de alguna forma resumir información de la muestra, para tener mejor conocimiento de la población en general y de su comportamiento.

Medidas de Tendencia Central

Mediana

Valor que ocupa la posición central del conjunto de datos ordenados de acuerdo a su magnitud en forma ascendente.

Media

También llamada promedio y se expresa como $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Moda

Valor cuya frecuencia relativa es mayor.

Medidas de Posición

Cuartiles

Con base en las observaciones ordenadas los cuartiles son aquellos valores que dividen a la muestra en cuatro partes iguales. El segundo Cuartil corresponde a la mediana, ya que divide a las observaciones en dos partes, el primer cuartil es aquel que divide la primera mitad de la muestra en dos partes y el tercer cuartil es el que divide la segunda mitad en dos partes.

Deciles

De nuevo con base en los datos ordenados, se trata de dividir en diez partes iguales. Usando la idea de los cuartiles para nombrar cada uno de los nueve deciles.

Percentiles

Estos son noventa y nueve, corresponde a cada uno de las observaciones que dividen a la muestra en cien partes iguales.

Medidas de Dispersión

Varianza muestral

Es el promedio de los cuadrados de la desviaciones de las observaciones con respecto a su media y se expresa como

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Desviación Estándar muestral

Es la raíz cuadrada positiva de la varianza y se expresa como $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$

Covarianza muestral

En este caso se supone que se tienen dos variables x y y, la covarianza es la media de los productos de las desviaciones correspondientes, es decir:

$$\hat{s} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Estas medidas son básicas cuando se está haciendo un análisis de los datos, sobre todo porque al momento de ajustar una función de densidad a los datos. Si se conoce como se distribuyen los datos, en muchas ocasiones se puede estimar sus parámetros con base en estas medidas, como ocurre en la mayoría de las funciones de densidad.

2.1.3. Gráficos

Una manera más visual de analizar los datos es a través de histogramas, diagrama de caja y brazos, caritas de Chernoff, diagrama de estrellas, etc. Estos ayudan a visualizar mejor las observaciones, ya que en ocasiones es más fácil hacer conjeturas con base en algo más visible. Aunque no siempre se puede visualizar los datos con algunos de estos gráficos. Esto es porque muchas gráficas se encuentran en a lo más tres dimensiones, pero a veces los observaciones son de más de dos variables y no es fácil graficar algo más allá de estas dimensiones. Por eso es bueno hacer uso de diagramas que ayuden a ver esos casos donde hay más de dos variables.

Histograma

Es una manera gráfica de representar las frecuencias relativas asociadas a cada k intervalo, que en este caso se llaman marcas de clase. En el eje de las abscisas están dichas marcas de clase y en el eje de las ordenas las frecuencias relativas. Ver la Figura 2.1

Cuando se trata de ajustar una función de densidad se procura que ésta se ajuste bien al histograma. Además de esta importancia, con este gráfico se puede observar el lugar donde se encuentra la moda de los datos, que correspondería a la joroba de la gráfica. Como ya se ha mencionado antes se puede detectar la multimodalidad de la muestra con ver esta gráfica. También se pude notar si la muestra tiene colas pesadas, si esta sesgada hacia alguna dirección, si es simétrica o no, con base en estas características se identifica cual distribución sería aceptable ajustarle.

En la actualidad existen muchos paquetes que realizan este gráfico fácilmente. La idea básica de la construcción de un histograma es la descrita a continuación:

Paso 1. Definir el rango de los datos. Es decir, diferencia entre el dato máximo y el dato mínimo.

Paso 2. Determinar el número de clases o intervalos. En ocasiones este es igual a la raíz cuadrada del número de datos.

Paso 3. Establecer el rango de cada clase.

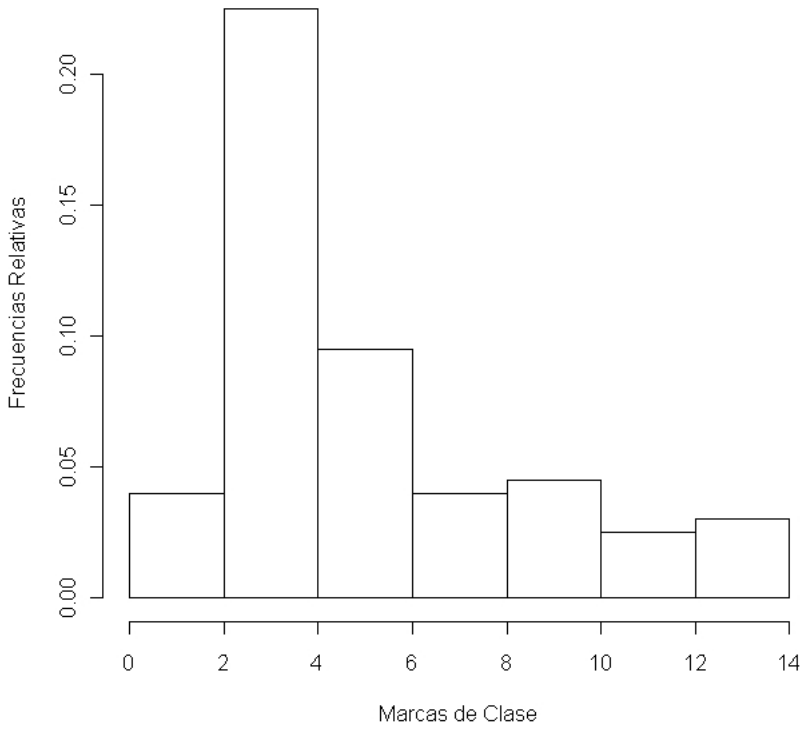

Figura 2.1: Histograma

Paso 4. Obtener la frecuencia de cada clase. Se cuentan los datos que caen en cada intervalo de clase.

Paso 5.Graficar la su frecuencia vs clase.

En el caso multivariado no se puede tener un histograma para la muestra a menos que estemos se considere de dos variables en ese caso se tendría un histograma en tres dimensiones. Lo que se puede hacer en estos casos es un histograma por cada una de las variables.

Diagrama de Caja y Brazos

Es una manera gráfica de ver las medidas de posición, así como que tan alejadas estas las observaciones ya que gráfica la mínima y máxima observación, además de mostrar aquellos datos que son atípicos. Ver Figura 2.2.

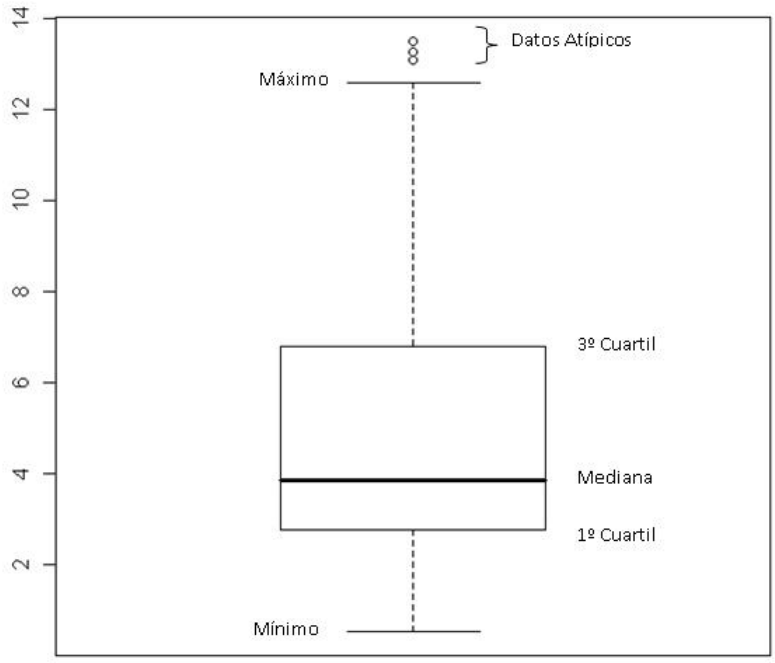

Figura 2.2: Diagrama de Caja y Brazos

Esta gráfica al igual que el histograma sirve para detectar características de la muestra, sobre todo de dispersión, ya que al graficar tanto los cuartiles como los extremos se nota que tanto se están alejando las observaciones de la mediana y hacia qué lado de la gráfica se están agrupando. Una gran desventaja de esto, es que cuando hay multimodalidad, no se puede ver en este gráfico.

Con este grafico podemos caer en la situación de no poder graficar en el caso de tener más de una variable, es por eso que en el caso multivariado se hace este diagrama por cada variable.

Caritas de Chernoff y Estrellas

El Diagrama de Caritas de Chernoff es especialmente para el caso multivariado, ya que se gráfica cada observación tomando rasgos humanos (caritas), cada rasgo del rostro, como el ancho de la cara, curvatura de la boca, localización de los ojos, longitud de la nariz, etc., representa una variable. Este gráfico fue introducido por Herman Chernoff en 1973. Ver Figura 2.3

Como este gráfico grafica cada observación, es muy bueno utilizarlo para aquellas bases de datos que no contienen muchos datos. Es fácil distinguir que variable es buena para identificar diferencias entre las observaciones, se podrían identificar datos atípicos. Una gran aplicación de este gráfico es para la identificación de grupos dentro de la muestra,

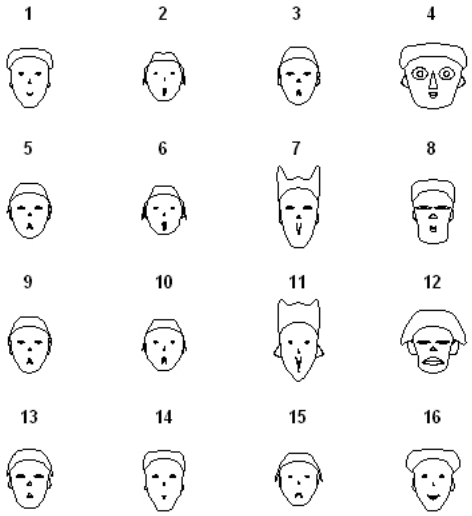

Figura 2.3: Caritas de Chernoff: 16 observaciones con 5 variables

ya que se pueden agrupar aquellas observaciones con caras similares y esto se daría una idea del comportamiento de la muestra, es decir, diría si es apropiado usar un modelo de mezclas finitas.

Un gráfico muy parecido es el Diagrama de Estrellas donde cada estrella del gráfico corresponde a una observación de la muestra. Cada rayo representa una variable y su longitud depende del valor de dicha variable para cada observación. Ver Figura 2.4

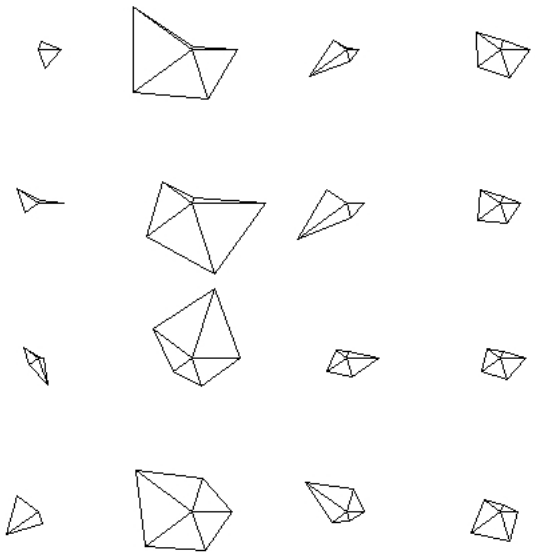

Figura 2.4: Diagrama de Estrellas: 16 observaciones con 5 variables

Una vez hecho un análisis descriptivo de los datos, se tiene una idea del comportamiento de los datos. Esto es, si tienen multimodalidad, un tanto la forma de la densidad en cuanto a la simetría o la pasadez en las colas. Quizás con base en investigaciones anteriores o conocimiento previo de la muestra, se sabe el modelo que siguen los datos, pero se desconocen los parámetros. Es para este caso que a continuación se mencionarán algunos de los métodos que se utilizan para conocer los parámetros de la función de densidad.

2.2. Estimación de Parámetros

También conocida como estimación puntual, ésta se utiliza cuando ya se conoce la familia de la distribución de los datos, como por ejemplo se supiera que se distribuyen Binomial, Poisson, Geométrica, Normalmultivariada, etc. Se llama θ al conjunto de parámetros que especifíaca la función de densidad y sea Θ el conjunto de posibles valores que puede tomar los parámetros θ.

2.2.1. Método de Momentos

Sea $f_X(x, \Theta)$ donde $x \quad \epsilon \quad \mathbb{R}$ y $\theta = (\theta_1, ..., \theta_k)$, una función de densidad, sea $\mu_r = E(X^r)$ el r-ésimo momento poblacional alrededor del cero, con base en un conjunto de

observaciones x_i para $i = 1, ..., n$ sea $M_r = \frac{\sum_{i=1}^{n} x_i^r}{n}$ el r-ésimo momento muestral alrededor del cero. Se igualan estos momentos

$$M_j = \mu_j \tag{2.1}$$

para cada $j = 1, ..., k$. En el caso de los momentos poblacionales se tienen por incógnitas a los k párametros de la función y para los momentos muestrales no se tienen incógnitas, por lo tanto se habla de k incógnitas y k ecuaciones. Se resuelve dicho sistema y se encuentran los estimadores de la función de densidad. A la solución $\hat{\theta}_M = (\hat{\theta}_1, ..., \hat{\theta}_k)$, se le conoce como el estimador por el método de momentos.

2.2.2. Método de Máxima Verosímilitud

Sea $X_1, ..., X_n \, \epsilon \, \mathbb{R}^p$ una muestra aleatoria. Se define como la función de verosímilitud a la densidad conjunta de dichas variables, como función de θ, como $L(\theta; X_1, ..., X_n)$ o simplemente $L(\theta)$, al ser estas variables independientes e identicamente distribuidas la función de verosímilitud queda expresada como:

$$L(\theta; \boldsymbol{X}) = \prod_{i=1}^{n} f_X(x_i, \theta) \tag{2.2}$$

Se llama estimador máximo verosímil ($\hat{\theta}_{MV}$) al valor que máximice a la función de verosimilitud, entonces $\hat{\theta}_{MV}$ es aquel que cumple la desigualdad $L(\hat{\theta}_{MV}) \geq L(\theta)$ para todo $\theta \quad \epsilon \quad \Theta$.

Para encontrar dicho máximo se hace uso de las derivadas parciales de la función de verosimilitud con respecto a cada uno de los k parámetros. En ocasiones no es fácil encontrar dichos máximos ya que es difícil despejar de las derivadas de la expresión de verosimilitud los parámetros. Es por eso que en general se obtiene el logaritmo de la verosimilitud, dado que el máximo de esta función es le mismo que el del logaritomo de la misma, entonces el objetivo ahora sería encontrar el máximo de $Ln(L(\theta)) = l(\theta)$.

Estos métodos ayudan a encontrar los estimadores de los parámetros, pero cuando se tiene un valor posible del parámetro, surge la duda sobre si dicho estimador ajusta bien a la muestra. Para esto es necesario hacer una prueba de hipótesis, donde nuestra hipótesis nula sería que el valor del parámetro de la función corresponde al valor que se propone como verdadero.

2.3. Pruebas de Hipótesis

Como su nombre lo dice una prueba de hipótesis sirve para comprobar si la aseveración, es decir, la hipótesis sobre la muestra es cierta o no. Por ejemplo, se puede decir que los datos son realmente aleatorios, pero quizás no lo sean, o quizás se puede suponer que la muestra tiene el mismo comportamiento de alguna otra, o que los datos se distribuyen

exponencial, etc. Para esto se hace una prueba de hipótesis con bases matemáticas y estadísticas más rigurosas que una simple suposición de algún análisis descriptivo.

En las pruebas de hipótesis, en el caso de Bondad de ajuste, se tiene una hipótesis nula (H_0), que por lo general es la que se quiere probar como verdadera, y una hipótesis alternativa (H_1), sería como la hipótesis complemento de la nula, en el caso de rechazar la hipótesis nula se puede decir se rechaza en favor de la alternativa.

2.3.1. Razón de Verosimilitudes

Se supone dada una muestra aleatoria $\boldsymbol{X} = X_1, ..., X_n$, de las cuales se conoce a la familia de funciones de densidad $f(X, \theta)$, se pueden hacer pruebas de hipótesis para cada uno de su parámetros, donde $H_0 : \theta \epsilon \Theta_0$, contra la hipótesis alternativa $H_1 : \theta \quad \epsilon \quad \Theta_1$, donde $\Theta_0 \subset \Theta$ y $\Theta_1 \subset \Theta$ y además estos dos conjuntos son disjuntos, usualmente $\Theta_1 = \Theta - \Theta_0$.

Se define la razón de verosimilitudes generalizada, es decir, en general para cualquier conjunto correspondiente a la hipótesis nula.

$$\lambda = \frac{\underset{\theta \epsilon \Theta_0}{máx} \; L(\theta; \boldsymbol{X})}{\underset{\theta \epsilon \Theta_1}{máx} \; L(\theta; \boldsymbol{X})} \tag{2.3}$$

En vista de que λ depende únicamente de la muestra aleatoria, se puede reescribir como una función de la misma, $\lambda = T(\boldsymbol{X})$.

El principio de razón de verosimilitudes dice que se rechaza $H_0 : \theta_i \epsilon \Theta_{i0}$ si y sólo si $\lambda \leq k$, donde k es alguna constante fija, la cual se especifica fijando el tamaño de la prueba. Esto es fijando el grado de significancia α, $P[T(\boldsymbol{X}) \leq k \mid H_0] = \alpha$, bajo el supuesto que se conoce $f(x)$ bajo H_0 se podría encontrar la forma en que se distruye $T(\boldsymbol{X})$ y así encontrar el valor de k.

Cuando no es fácil obtener la distribución de $T(\boldsymbol{X})$ se utiliza una aproximación con base a la distribución asintótica de λ, es decir, $-2Ln(\lambda) \sim \chi^2_{q-r}$, esto si Θ_1 pertenece a $\mathbb{R}^q$ y si Θ_0 pertenece a $\mathbb{R}^r$.

Cuando se desconoce totalmente la distribución de la muestra se puede suponer que tienen cierta distribución, esto con base en experiencia, investigaciones anteriores o quizás al hacer un análisis descriptivo. Como por ejemplo analizando el histograma de la distribución muestral y comparándolo con la gráfica de alguna distribución conocida. Pero estas suposiciones deben tener un respaldo estadístico, es por eso que se hace uso de algunas pruebas de Bondad de Ajuste donde la hipótesis nula es si la muestra se distribuye con alguna distribución conocida, cuando los parámetros no son especificados, se estiman. A continuación se mencionan algunas de estas pruebas.

2.3.2. Ji- Cuadrada de Pearson

Sea $x_1, ..., x_n$ una muestra aleatoria, se supone que dicha muestra tiene una función de distribución $F_0(x)$, entonces la hipótesis nula es $H_0 : F(x) = F_0(x)$ para todo x y la hipótesis alternativa es $H_1 : F(x) \neq F_0(x)$ para algún x.

Se supone que las observaciones están agrupadas en k categorías y las frecuencias de cada categoría observada son n_i para $i = 1, ..., k$, asimismo con base en la distribución en la hipótesis nula se obtiene la probabilidad de que una observación sea clasificada en cada una de las k categorías. Esta probabilidad multiplicada por n, da el valor esperado de frecuencias por cada celda bajo la hipótesis nula e_i. El criterio de Pearson está basado en la diferencia entre estos valores, salvo que se normalizan con las frecuencias esperadas. Entonces la estadística de prueba de Pearson es:

$$Q = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} \tag{2.4}$$

La distribución de esta variable se aproxima a una ji-cuadrada con $k - 1$ grados de libertad χ^2_{k-1} considerando lo anterior y además que $P(Q \leq k \mid H_0) = \alpha$, se rechaza H_0 si $Q > \chi^2_{k-1,1-\alpha}$.

En ocasiones la hipótesis nula no especifica los parámetros de la distribución, cuando ocurre esto se estiman los parámetros por máxima verosimilitud de la misma, lo que cambiaría en este caso son los grados de libertad de la χ^2, si los parámetros que no se especifican son s entonces Q tiene una distribución aproximada de una χ^2_{k-s-1}.

2.3.3. Kolmogorov-Smirnov

Sea $x_1, ..., x_n$ una muestra aleatoria, se define a $S_n(x)$ como la función de distribución empírica.

$$S_n(x) = \begin{cases} 0 & \text{si} \quad x < X_{(1)} \\ \frac{i}{n} & \text{si} \quad X_{(i-1)} \leq x < X_{(i)} \quad i = 1, ..., n \\ 1 & \text{si} \quad x \geq X_{(n)} \end{cases}$$

Se supone que dicha muestra tiene una función de distribución $F_0(x)$, de la cual se conocen sus parámetros, entonces la hipótesis nula es $H_0 : F(x) = F_0(x)$ para todo x y la hipótesis alternativa es $H_1 : F(x) \neq F_0(x)$ para alguna x.

Este criterio está basado en la diferencia entre la función de distribución empírica y la teórica bajo la hipótesis nula, para cada una de las observaciones. Entonces la estadística de prueba es:

$$D_n = \underset{\substack{x_i \\ i = 1, ..., n}}{máx} \{|S_n(x_i) - F(x_i)|, |S_n(x_{i-1}) - F(x_i)|\} \tag{2.5}$$

La región de rechazo está definida como:

$$C = \{D_n \mid D_n > D_{n,\alpha}\} \tag{2.6}$$

donde $D_{n,\alpha}$ = Cuantil de la distribución de D_n con un nivel de significancia α, es decir $\mathbb{P}(D_n > D_{n,\alpha}) = \alpha$.

La manera de calcular la función de distribución para D_n para cualquier $F(x)$, se puede deducir con base en el siguiente teorema:

Teorema 2.1. *Para* $D_n = \sup_x \{|S_n(x) - F(x)|\}$, *donde* $F(x)$ *es cualquier función de distribución, se tiene que:*

$$P(D_n < \frac{1}{2} + v) =$$

$$\begin{cases} 0 & \text{para} \quad v \leq 0 \\ \int_{1/2n-v}^{1/2n+v} \int_{3/2n-v}^{3/2n+v} \cdots \int_{(2n-1)/2n-v}^{(2n-1)/2n+v} f(u_1, u_2, ..., u_n) du_n \cdot\cdot du_1 & \text{para} \quad 0 < v < \frac{2n-1}{2n} \\ 1 & \text{para} \quad v \geq \frac{2n-1}{2n} \end{cases}$$

donde

$$f(u_1, u_2, ..., u_n) = \begin{cases} n! & \text{para} \quad 0 < u_1 < u_2 < \cdots < u_n < 1 \\ 0 & \text{en otro caso} \end{cases}$$

Usando este resultado se han tabulado los valores $D_{n,\alpha}$ para distintos valores de n con $\alpha = 0.01$ y $\alpha = 0.05$ como se enuncia en la referencia [12].

2.4. Mezclas Finitas para estimar densidades

Hasta el momento se ha explicado la manera de estimar densidades cuando la muestra es unimodal, pero en el caso de existir más de una moda, la manera de atacar el problema se modifica un poco, ya que se podría usar un modelo de mezclas finitas, usando las herramientas matemáticas explicadas en el Capítulo 1.

Al percatarse, a través de un análisis descriptivo, de la existencia de más de una moda, se puede ajustar un modelo de mezclas finitas. Por lo general si el problema es encontrar la función de densidad de la muestra, lo que realmente interesa es la expresión de dicha función para poder conocer la probabilidad de ocurrencia de ciertos eventos. Por ejemplo, en el caso de una aseguradora si la variable aleatoria a describir es el monto individual de las reclamaciones que llegan a dicha empresa le interesa conocer el monto esperado por reclamación, la probabilidad de que una reclamación exceda un valor, saber si tiene cola pesada por la derecha para efectos de deducible y a la izquierda para efectos de reaseguro, etc. Entonces al comenzar con un número determinado de componentes no interasa que sean muchos o pocos, sino a la variable como tal, así que se busca un modelo que mejor ajuste a los datos sin tener restricción en el número de componentes.

2.4.1. Ejemplos de Estimación de Densidad vía un Modelo de Mezclas Finitas

A continuación se muestran algunas simulaciones de mezclas, basadas en las densidades de la referencia [17]. A estas simulaciones les ajusta diferentes modelos de mezclas finitas como si no se conociera la verdadera distribución. La metodología para generar estas observaciones, está descrita en el Apéndice A. El ajuste de los modelos de mezclas finitas se realizó usando R versión 2.7.2 (2008-08-25) usando la función Mclust, la cual usa el algoritmo EM para encontrar los parámetros de la mezcla.

Ejemplo 2.1. La simulación corresponde a una muestra con 500 observaciones, de suma de normales univariadas de la siguiente forma:

$$\frac{1}{2}N\Big(-\frac{3}{2},(\frac{1}{2})^2\Big)+\frac{1}{2}N\Big(\frac{3}{2},(\frac{1}{2})^2\Big)$$

Antes de empezar con un modelo de mezclas finitas, se hace un análisis descriptivo de los datos simulados.

El valor de la media es de 0.028 y el de la varianza es de 2.443, estos resultados hacen pensar que se está hablando de una muestra cuya función de densidad, suponiendo que los datos se distribuyen normal, es centrada en el origen y al tener una varianza considerable se puede decir que es un poco gorda.

La figura 2.5 corresponde al diagrama de caja y brazos, en este se observa que el rango de los valores de las observaciones es amplio, lo que se esperaba en base a la varianza. El valor máximo y el mínimo no están muy alejados del cuerpo de la caja y los datos tienen una ligera concentración a la derecha.

Una vez que se obtienen estos resultados, ahora se analizará el histograma, el cual se muestra en la Figura 2.6. Lo que muestra el histograma es algo muy distinto a lo que se pudo deducir con base en el análisis anterior. En este es detectable a simple vista una bimodalidad en los datos, cosa que con una curva de una función normal no podría describir. Los valores de dichas modas estarían entre [-2,-1] y [1,2], al sumar estas modas el resultado estaría cercano a cero, lo cual explica porque el valor de la media muestral es cercano a este valor. El hecho de estar separadas ambas modas hace que el rango de las observaciones sea mayor, es por esto que la varianza anterior era grande.

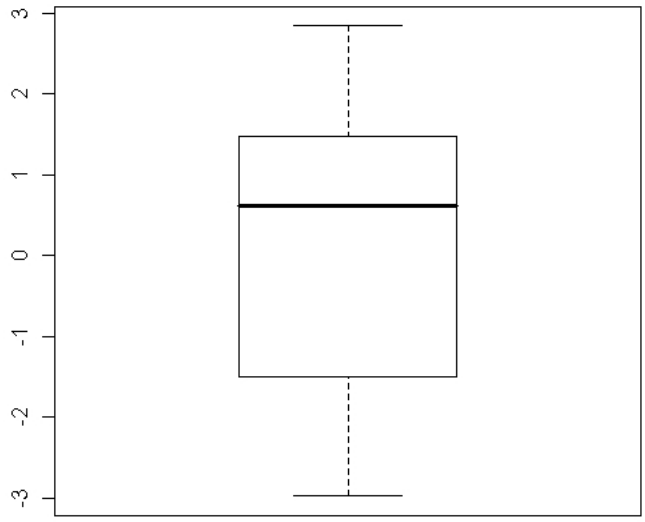

Figura 2.5: Diagrama de caja y brazos de la simulación del Ejemplo 2.1

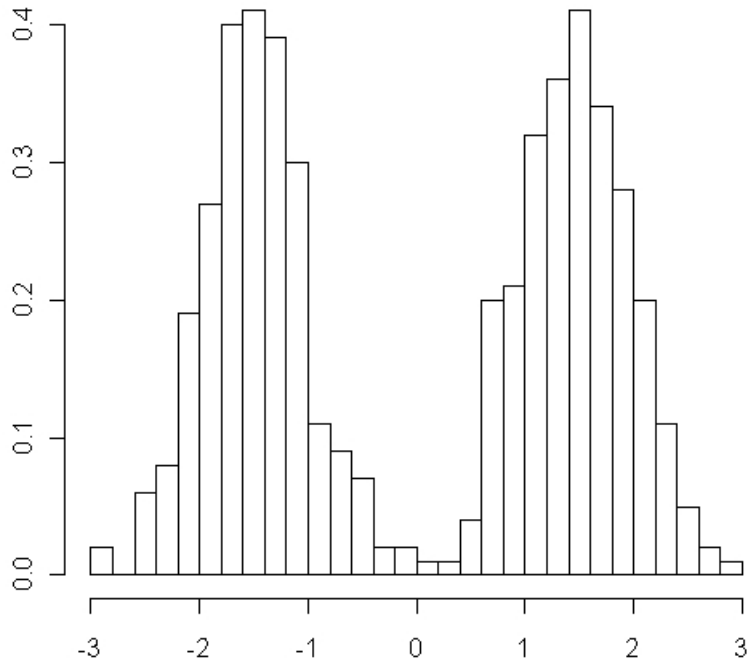

Figura 2.6: Histrograma de la simulacion del Ejemplo 2.1

En vista de los resultados no se puede usar una distribución conocida para ajustar los datos, una buena opción es usar una mezcla de funciones. En este caso lo primero es considerar dos componentes, pero no se debe descartar la idea de que pueden existir más. Por ejemplo podría ser que en el intervalo [1,2], donde se supone que hay una moda, realmente existan dos sólo que en el histograma no se alcanza a distinguir debido a que estén muy juntas. Por esta razón es prudente hacer un análisis para más de una componente.

En este caso, el análisis se hizo para uno, dos, tres y cuatro componentes. Los parámetros de estos ajustes y del BIC se encuentran en la Cuadro 2.1. En todos estos casos el modelo que se usó, es suponiendo la misma varianza, esta decisión se tomó ya que el histograma muestra dos jorobas muy parecidas en altitud como en longitud, lo que hace pensar que dichas componentes tiene una varianza igual.

Bimodal Separada				
Nº Componentes	Pesos	Media	Varianza	BIC
1	1	0.02792209	2.443798	-1877.143
2	0.4847575 0.5152425	-1.501169 1.466542	0.2391298 0.2391298	-1416.833
3	0.48394034 0.03732172 0.47873794	-1.503506 1.001352 1.500105	0.2309934 0.2309934 0.2309934	-1429.121
4	0.3049737 0.179643 0.2838219 0.2315614	-1.663501 -1.226975 1.223477 1.76374	0.1801206 0.1801206 0.1801206 0.1801206	-1440.868

Cuadro 2.1: Parámetros de los modelos de mezclas aplicados al Ejemplo 2.1.

Usando el Criterio de Información Bayesiana, el valor más pequeño en valor absoluto es con dos componentes, de ahí le sigue el de tres componentes, pero la diferencia es favorable al de dos. En cuanto a los otros casos se ve que el valor del BIC es muy grande, por lo tanto se descartan. En cuanto al valor de los parámetros cuando se ajustan con dos componentes, las dos medias estan dentro de los rangos que había considerando al ver el histograma, los pesos de las componentes son muy parecidos, de hecho es ligeramente mayor el de la componente con media en 1.466542, o sea al lado derecho, como se había observado en el diagrama de caja y brazos.

La Gráfica 2.7 muestra el histograma de las observaciones y la curva del modelo con dos componentes. En esta se puede ver que la curva se ajusta bien al histograma, por lo tanto la función de densidad de mezcla finita de normales describe bien las observaciones.

Una vez analizado el BIC y la curva en el histograma, se puede decir que la función de densidad, de una mezcla finita de normales, que mejor se ajusta a los datos es la de dos componentes expresada como:

$$0.4847575N(-1.501169, 0.2391298) + 0.5152425N(1.466542, 0.2391298)$$

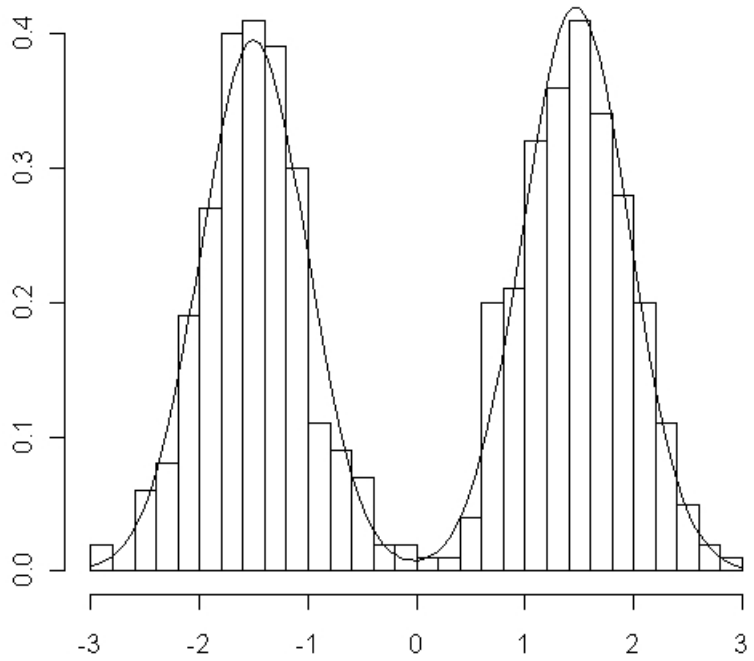

Figura 2.7: Histograma de la simulación del Ejemplo 2.1 y ajuste con dos componentes

Ejemplo 2.2. Este ejemplo corresponde a la simulación de una muestra con 500 observaciones de suma de normales univariadas de la siguiente forma:

$$\frac{9}{20}N\Big(-\frac{6}{5},(\frac{3}{5})^2\Big)+\frac{9}{20}N\Big(\frac{6}{5},(\frac{3}{5})^2\Big)+\frac{1}{10}N\Big(0,(\frac{1}{4})^2\Big)$$

De nuevo comenzamos el análisis suponiendo que no se sabe nada de los datos simulados.

El análisis descriptivo de la base de datos, arroja $\mu = -0.08262636$ y $\sigma^2 = 1.529288$, estos datos, en principio, describirían una gráfica centrada en el origen y ancha. Pero antes de especular vemos el histograma de las observaciones en la Figura 2.8.

Analizando el histograma, se ve que no existe una moda como tal, que la gráfica no es como nos gustaría ya que se aprecian varias modas, de entrada se podría decir que hay una en el intervalo $[-1.6,-1.2]$, otra en $[-1.2,-0.6]$, otra en $[-0.2,0.4]$ y otra en $[1,1.6]$, es decir, cuatro modas. Sin embargo, a simple vista no se puede asegurar nada, lo que si se puede decir es que esta muestra no se puede modelar con una distribución conocida.

En esta ocasión no se hará un análisis del diagrama de caja y brazos, ya que ya se ha detectado multimodalidad y este no ayuda en la descripción de un modelo de mezclas.

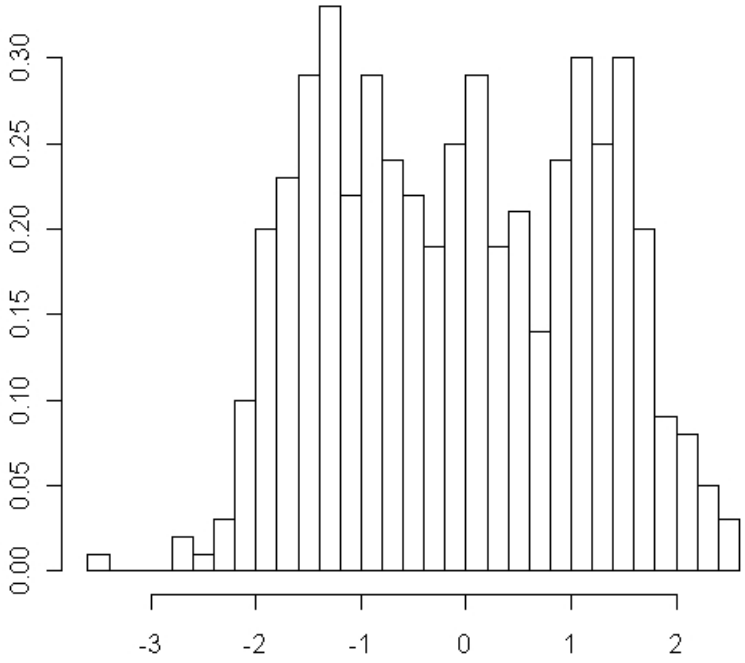

Figura 2.8: Histrograma de la simulación del Ejemplo 2.2

El número de componentes para ajustar los diferentes modelos, se elige con base en el histograma. Si estos modelos no funcionaran se probará con más. De entrada se usa de : 2, 3, 4 y hasta 5 componentes, esto es porque en el análisis del histograma se ven a lo más cuatro componentes y mínimo 2.

En cuanto a si se usará un modelo con varianza similar o diferente en los componentes, se prefiere ajustar ambos casos. Si lo más acertado es de cuatro componentes, los intervalos que parece deberían estar sus medias son de tamaños parecidos, esto hace pensar en modelos de igual varianza. Pero si se suponene tres componentes los intervalos de las modas ya no son tan iguales. Por lo tanto se prefiere ajustar ambos casos y discriminarlos con base en el criterio de información bayesiana.

En el Cuadro 2.2 se observa que el mejor modelo es el correspondiente a tres componentes con varianzas iguales. Sin embargo,es de interés también analizar el de dos componentes con misma varianza y el de tres componentes con diferente varianza. Se eligen

BIC				
Modelo \ Comp.	2	3	4	5
E	-1581.298	-1575.744	-1588.502	-1600.118
V	-1587.730	-1586.365	-1601.358	-1620.440

Cuadro 2.2: Criterio de Información Bayesiana para el Ejemplo 2.2.

Trimodal			
Nº Componentes	Pesos	Media	Varianza
3 (Var. iguales)	0.3757044 0.2882350 0.3360606	-1.3656340 -0.0479828 1.3220193	0.2443760
2	0.5263567 0.4736433	-1.0600080 1.0035310	0.4646390
3 (Var. distintas)	0.4215241 0.2466545 0.3318215	-1.2838379 0.0570647 1.3394763	0.2951920 0.1785097 0.2220076

Cuadro 2.3: Parámetros para los distintos modelos para el Ejemplo 2.2

estos modelos porque se considera que son los más cercanos a aquel con mejor BIC. No se busca otros modelos ya que se nota que existe una tendencia creciente (en valor absoluto) del BIC conforme se aumenta el número de componentes. Y si se disminuye el número de componentes, sería a una y este modelo ya quedó descartado desde el principio.

Los parámetros de estos ajustes en orden de mejor BIC, se encuentran en el Cuadro 2.3.

En vista de los resultados arrojados en el modelo con dos componentes, se ve una función de densidad un tanto simétrica al origen. El ajuste de este modelo al histograma se observa en la Figura 2.9. Se nota que a pesar de tener uno de los mejores BIC de los modelos ajustados, esta función de densidad no describe muy bien los datos, las observaciones cercanas al cero salen demasiado de la gráfica, aunque parece ajustarse bien en las otras dos jorobas, por lo que se prefiere analizar más a fondo el caso de los modelos con tres componentes.

Comparando los modelos con tres componentes, es claro que los valores de las medias son muy parecidos, los pesos son algo parecidos pero mantienen una misma tendencia. En cuanto a la varianza, si cambia, ya que la componente con media en el origen tiene una varianza mucho menor al resto, esto en el modelo de varianza diferente. Las otras dos varianzas están relativamente cerca a la del primer modelo. Se nota que las diferencias son mínimas en los números, se analizará ahora las curvas de densidad.

Para el primer modelo, la curva ajustada al histograma se muestra en la Figura 2.10.

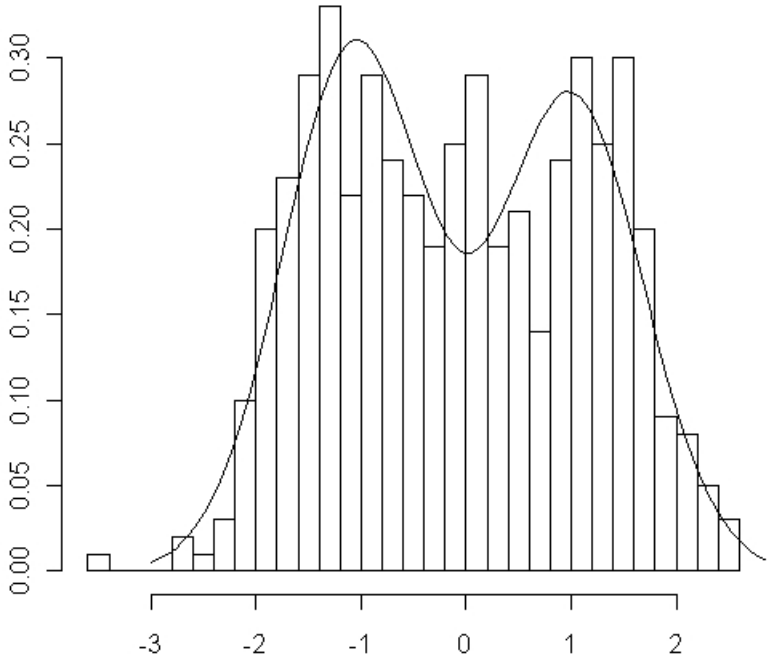

Figura 2.9: Curva del modelo con dos componentes e igual varianza del Ejemplo 2.2

Esta corresponde a la línea con puntos, en esta de nuevo hay muchos datos cercanos al origen que salen de la curva, pero no tanto como el caso anterior. La curva ajustada al histograma del tercer modelo analizado está en la Figura 2.10. Correspondiente a la línea sólida, una curva con las mismas características que la anterior, a simple vista no se encuentran diferencias como en el caso del análisis de los parámetros.

Analizando ambos ajustes. Se puede notar que en los intervalos $[-\infty, -1.8]$ y $[1.6, \infty]$ la diferencia es muy pequeña, casi nula. Sin embargo, para el intervalo $[-0.6, 0.6]$ la diferenica es notable, justo este intervalo es donde se encuentra la componente con media cero y la varianza para esta componente si tiene un gran cambio. Es aquí donde la curva del modelo con igual varianza es más delgada y está un poco hacia la derecha. Sin embargo, visualmente no se puede decidir si un modelo es mejor al otro, pues ambos son muy parecidos. Por esta razón el que se elige para modelar los datos es el de las varianzas iguales, con base en el BIC.

La función de densidad para las observaciones de este ejemplo es:

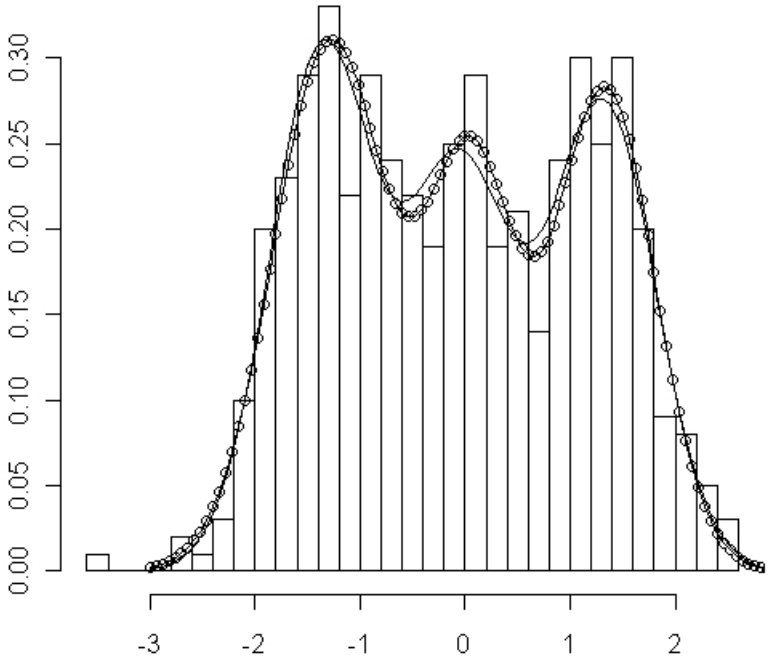

Figura 2.10: La curva con puntos representa el modelo con igual varianza y la curva seguida corresponde al modelo con varianza distinta.

$$\begin{aligned} &0.3757044N(-1.365634, 0.244376) \\ &+0.288235N(-0.0479828, 0.244376) \\ &+0.3360606N(1.3220193, 0.244376) \end{aligned}$$

Ejemplo 2.3. Este modelo no es exclusivo de muestras univariadas, también para el caso multivariado sirve. A continuación se presenta este modelo para el caso de una muestra bivariada de 500 observaciones con función de densidad dada por:

$$\frac{1}{2}N\left((1,1),\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)+\frac{1}{2}N\left((4,4),\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)$$

Se inicia con un análisis descriptivo de los datos simulados, se hace un histograma, diagrama de estrella, diagrama de rostro, etc.

En la Figura 2.11 se hace una clasificación de 100 de las observaciones. En este análisis se puede notar cuatro grupos, de los cuáles el dos y cuatro son grandes, los otros tienen menos datos. Aunque se ven distintos grupos aún no se puede asegurar multimodalidad. Se puede pensar en cuatro grupos y quizás dos, recuerde que este análisis es de solo el 20 % de la muestra total.

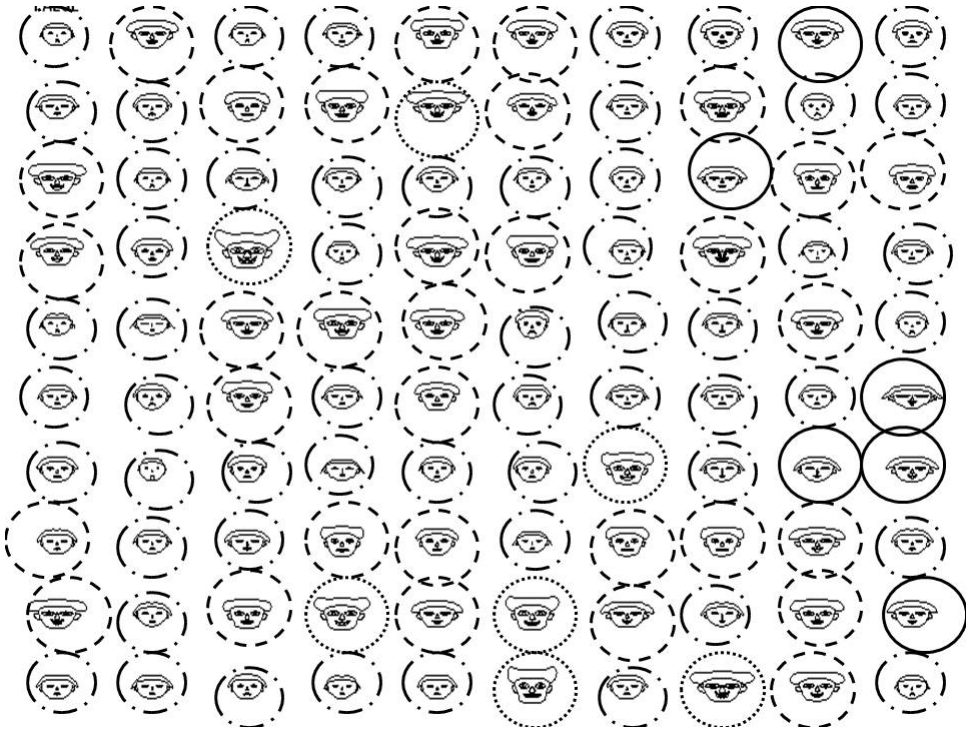

Figura 2.11: Diagrama de Rostros con 100 observaciones. 1- (Línea sólida); 2- $-\cdot-\cdot-\cdot$ (Línea punto); 3- $\cdots\cdots$ (Puntos); 4- $- - -$ (Línea punteada)

Un análisis parecido al anterior se podría hacer con un diagrama de estrellas, pero al tener la muestra solo dos variables, la figura no es clara.

La Figura 2.12 muestra la gráfica de las observaciones graficadas entrada 2 vs entrada 1. En esta se observa una separación de dos grupos bien definidos, pero aún así no se debe descartar la posible existencia de grupos muy pegados a los que se ve. Por ejemplo se puede detectar un grupo extra en medio de las dos masas que se forman.

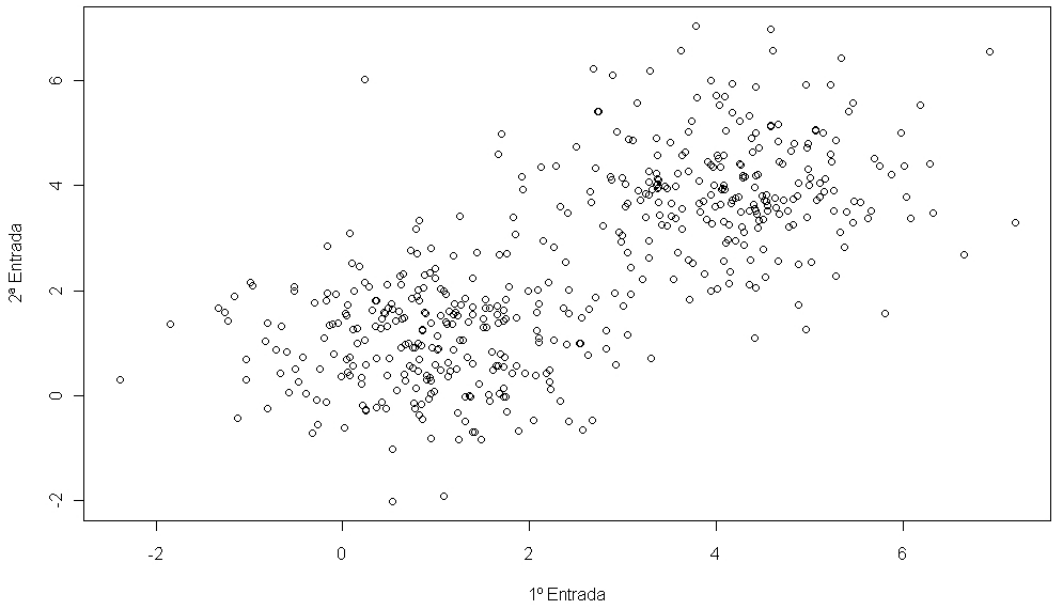

Figura 2.12: Gráfica de dispersión de las observaciones del ejemplo 2.3

Ahora se tiene una mejor idea de cómo se comportan los datos, la siguiente gráfica que se elaboró, es el histograma, el cual se muestra en la Figura 2.13. Esta figura da una mayor información de los datos, se ve claramente dos jorobas en la muestra. Una de ellas en el intervalo $[-2, 2] \times [-2, 2]$ y la otra en $[2, 6] \times [2, 6]$, pero no se debe descartar la posibilidad de que en el intervalo $[2, 4] \times [0, 4]$ exista algún otro grupo, con quizás un peso pequeño.

En conclusión se ajustarán modelos con 2, 3 y 4 componentes. Dos porque es el mínimo número de modas que se observan, tres por lo antes mencionado y cuatro por si hay grupos muy cercanos a las modas.

Los criterios de información bayesiana se grafican en la Figura 2.14, se ve que el mejor modelo es sin duda es de dos componentes, ahora la tarea es identificar con cual modelo en relación a la matriz de covarianzas se usará. Los BIC más pequeños están presentados en el Cuadro 2.4 junto con los parámetros de cada modelo.

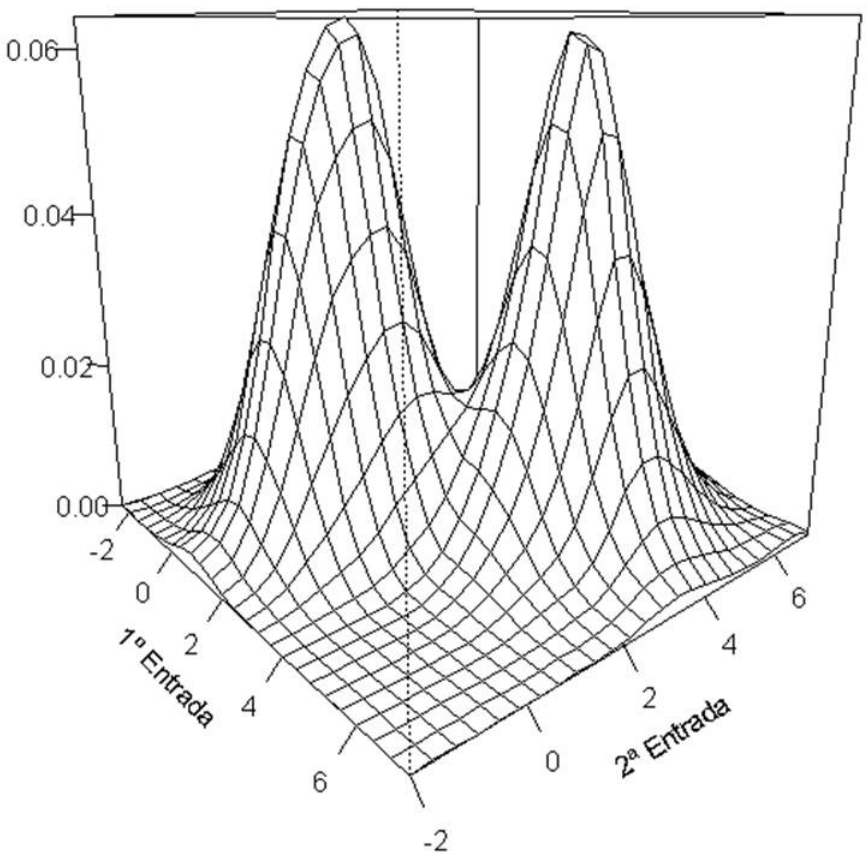

Figura 2.13: Histograma de las observaciones del Ejemplo 2.3

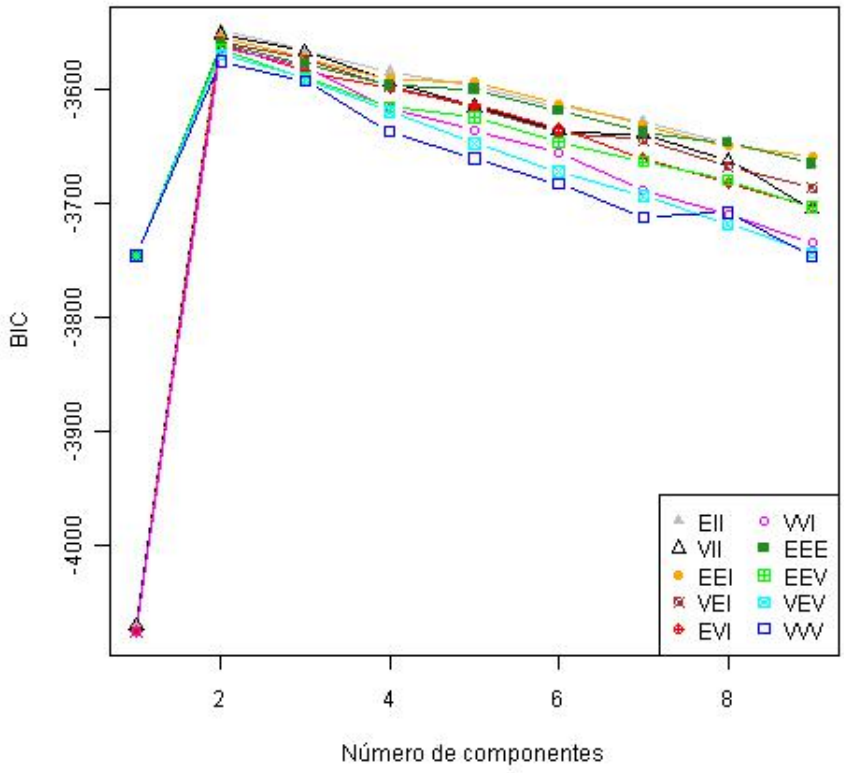

Figura 2.14: Resultados del Criterio de Información Bayesiana del Ejemplo 2.3

El modelo EII significa que ambas componentes tienen la misma matriz de covarianzas, las curvas de nivel son esféricas y el volumen de cada componente es el mismo. Esto es $\Sigma_k = \lambda I = 1.02259 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

El modelo VII dice que la matriz de covarianzas es diferente para las componentes, son esféricas y el volumen de cada componente es diferente. En este caso $\Sigma = \lambda_k I$, es decir , sólo difieren en la λ. $\lambda_1 = 0.937963$ y $\lambda_1 = 1.115871$, como este valor determina el volumen de cada componente, vemos que la primera tiene un mayor volumen, esto mismo se ve ligeramente en la Figura 2.13. Este modelo se ajusta mejor al análisis descriptivo.

El modelo EEI determina que ambas componentes tienen la misma matriz de covarianzas, el volumen y la forma de las curvas de nivel es el mismo.

Los parámetros se encuentran muy cercanos, las diferencias entre ellos son casi imperceptibles, aún más cercanos que en el caso del Ejemplo 2.2. Por ende lo mejor es elegir el modelo de dos componentes con EII, con base al BIC. Aunque el análisis descriptivo diría que se eligiera el VII, se ve que no hay mucha diferencia en los parámetros, por eso

Bivariada				
Modelo	Pesos	Vector de Media	Matriz de Covarianzas	BIC
EII	0.523101	$\begin{pmatrix} 0.9265902 \\ 1.0363458 \end{pmatrix}$	$\begin{pmatrix} 1.02259 & 0 \\ 0 & 1.02259 \end{pmatrix}$	-3548.936
	0.476899	$\begin{pmatrix} 4.120204 \\ 3.967556 \end{pmatrix}$	$\begin{pmatrix} 1.02259 & 0 \\ 0 & 1.02259 \end{pmatrix}$	
VII	0.5178435	$\begin{pmatrix} 0.910884 \\ 1.021847 \end{pmatrix}$	$\begin{pmatrix} 0.93796 & 0 \\ 0 & 0.93796 \end{pmatrix}$	-3552.473
	0.4821565	$\begin{pmatrix} 4.102249 \\ 3.951166 \end{pmatrix}$	$\begin{pmatrix} 1.11587 & 0 \\ 0 & 1.11587 \end{pmatrix}$	
EEI	0.5230801	$\begin{pmatrix} 0.9245523 \\ 1.0384619 \end{pmatrix}$	$\begin{pmatrix} 0.9957 & 0 \\ 0 & 1.049544 \end{pmatrix}$	-3554.898
	0.476919	$\begin{pmatrix} 4.122299 \\ 3.965106 \end{pmatrix}$	$\begin{pmatrix} 0.9957 & 0 \\ 0 & 1.049544 \end{pmatrix}$	

Cuadro 2.4: Parámetros de los diferentes modelos en base al BIC

se toma como criterio el BIC. La función de densidad que mejor ajusta nuestros datos es:

$$0.523101N\left(\begin{pmatrix} 0.9265902 \\ 1.0363458 \end{pmatrix}, \begin{pmatrix} 1.022590 & 0 \\ 0 & 1.022590 \end{pmatrix}\right)$$
$$+0.476899N\left(\begin{pmatrix} 4.120204 \\ 3.967556 \end{pmatrix}, \begin{pmatrix} 1.022590 & 0 \\ 0 & 1.022590 \end{pmatrix}\right)$$

cuya gráfica de las curvas de nivel ajustadas a los datos, se muestra en la Figura 2.15.

Como ya se había mencionado antes, este modelo no es único para el caso de dos variables, se puede extender a más. Sin embargo, el visualizar estos casos es más difícil, pero la idea general es la misma.

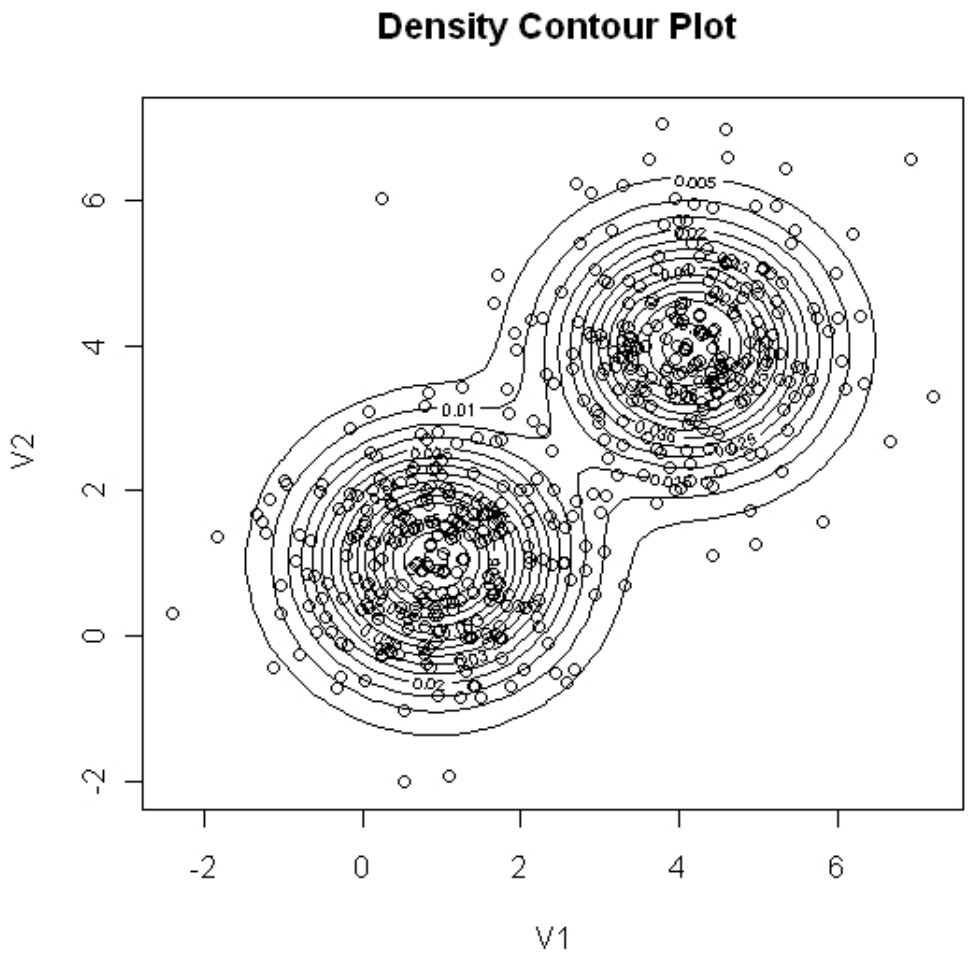

Figura 2.15: Ajuste de las curvas de nivel y los datos del Ejemplo 2.3

Conclusiones

El tema de estimación de densidad es muy extenso, existen diferentes métodos para encontrar dicha función, la mayoría de los cuales hace suposiciones sobre la muestra, como por ejemplo, la familia de distribuciones a la que pertenece. Una vez que se hacen supuestos sobre la familia a la que se le asocia, se pretende encontrar los parámetros de dicha función, para esto se utilizan diversos métodos, como lo son el de momentos y el de máxima verosimilitud.

Si se usa el modelo de mezclas finitas, se está haciendo una suposición sobre los datos. Cuando estamos hablando de estimación de densidad, no interesa mucho que tantas componentes tenemos, lo importante es encontrar esa función que describa mejor los datos.

Capítulo 3

Análisis de Conglomerados

Un análisis de conglomerados consiste básicamente en encontrar grupos dentro de un conjunto de n observaciones, X_1,X_2,X_3,...,X_n, con p variables observadas para cada objeto, es decir, $X_i = (x_{i1}, x_{i2}, ..., x_{ip})$, de tal manera que tengan características similares.

Como el objetivo es encontrar diferentes grupos, lo más lógico es escoger aquellas variables que ayuden a diferenciar mejor la muestra, no existe una regla para elegirlas. Lo que se puede hacer es con base en investigaciones pasadas o suposiciones, lo que se debe de considerar es que dichas variables caractericen lo mejor posible a los objetos y además que se relacione específicamente con el objetivo del análisis de conglomerados. Si se incluye alguna variable que es irrelevante a el objetivo se podría entorpecer los resultados.

Hay que considerar además, que pudieran existir observaciones que afecten mucho los resultados, datos que son muy diferentes a todos los demás. Posiblemente sean datos que no son representativos de la población en general o quizás estos datos si tengan una representación real en la población en general, pero no hubo un buen muestreo que arrojara más datos similares a este, es por eso que se debe asegurar que los datos sean una buena representación de la población en general.

Una vez establecidas las variables que se utilizarán y la base de datos en general, es importante determinar el método o algoritmo que se va elegir para encontrar los grupos. Alguno de los métodos utilizados son los jerárquicos, de partición, métodos basados en modelos, etc. Para la mayoría de estos procedimientos es necesario establecer una medida de disimilaridad o de distancia, entre las cuales destacan la distancia Euclideana, de Mahalanobis, o quizás el coeficiente de correlación entre las observaciones.

3.1. Disimilaridades y Similaridades

Las disimilaridades son tipos de medida, tratan de dar un grado a las diferencias o lejanías entre los grupos, entre más grande sea el valor de la medida mayor es la diferencia o la lejanía entre los dos elementos. Cuando se calcula la medida entre un mismo valor, el resultado de la medida es cero.

Una matriz de disimilaridades o de distancia es aquella que para cada elemento es una disimilaridad que cumple lo siguiente para todo i, j y k:

1. $d(i,j) \geq 0$
2. $d(i,i) = 0$
3. $d(i,j) = d(j,i)$
4. $d(i,j) \leq d(i,k) + d(k,j)$

Se tiene una matriz de datos $X_{n\times p} = (x_{ij})$, en donde la $i = 1, ..., n$ corresponde a la observación y la $j = 1, ..., p$ a la variable, por lo tanto la matriz de disimilaridades en una matriz de $n \times n$.

3.1.1. Distancias

Hay que tomar en cuenta al momento de utilizar algunas distancias los cambios de escala entre las variables. Si se utiliza alguna de las primeras cuatro distancias que se presentan a continuación, la escala con las que están cuantificadas la variables deben ser parecidas, es decir, no deben estar muy dispersas. Si ocurre esto se recomienda estandarizarlas de alguna manera, ya sea normalizarlas o quizás aplicar alguna función como logaritmo para suavizar alguna que este muy alejada. Otra opción es usar la distancia de Mahalanobis, la cual es menos sensible a cambios de escala.

Euclideana

Es la distancia ordinaria que hay entre dos puntos en un espacio euclideano

$$d(i,j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} = \|(X_i - X_j)\|$$

Manhattan o City Block

$$d(i,j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

Minkowski

$$d(i,j) = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^q\right)^{1/q}$$

Chebyschev

$$d(i,j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

Mahalanobis

Es una medida entre dos variables aleatorias considerando la correlación entre las observaciones

$$d(i,j) = \sqrt{(X_i - X_j)'\Sigma^{-1}(X_i - X_j)}$$

donde Σ es la matriz de covarianza de la muestra.

Las similaridades son medidas que evalúan el grado de parecido o proximidad entre dos elementos, entre más grande sea el valor de la medida mayor es el grado de parecido o proximidad, es por eso que cuando se calcula dicha medida entre un mismo valor, el resultado es el valor máximo de la misma.

3.1.2. Coeficiente

Congruencia

Este coeficiente corresponde al coseno del ángulo que hay entre los vectores formados por cada observación, toma valores entre [-1,1], para efectos de parecidos entre las observaciones se considera el valor absoluto, así si el resultado de la medida es muy cercano -1, esto significa que son muy parecidos.

$$s(i,j) = \frac{\sum_{k=1}^{p} x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{p} x_{ik}^2}\sqrt{\sum_{k=1}^{p} x_{jk}^2}} = \frac{X_i'X_j}{\|X_i\|\,\|X_j\|}$$

Correlación

Este corresponde al Coeficiente de Correlación de Pearson, el cual es un índice estadístico utilizado para medir la relación lineal entres dos variables. Una ventaja de este coeficiente es ser invariante a cambios de escalas.

$$s(i,j) = \frac{\sum_{k=1}^{p}(x_{ik} - \bar{X}_i)(x_{jk} - \bar{X_j})}{\sqrt{\sum_{k=1}^{p}(x_{ik} - \bar{X}_i)^2}\sqrt{\sum_{k=1}^{p}(x_{jk} - \bar{X_j})^2}} = \frac{\hat{Cov}(X_i, X_j)}{\sqrt{\hat{Var}(X_i)\hat{Var}(X_j)}}$$

donde $\bar{X}_m = \frac{\sum_{i=1}^{p} x_{mi}}{p}$

Existen diferentes métodos para agrupar las observaciones, dentro de estos se encuentran los llamados métodos jerárquicos. Estos métodos son secuenciales, paso a paso tratan de ir uniendo las observaciones o separándolas para formar los grupos. Esta manera de formalos involucra la construcción de una estructura de árbol o dendrograma. Existen dos tipos de procedimientos jerárquicos: los aglomerativos y los divisivos. Estos últimos parten de un solo grupo formado por todas las observaciones, en los pasos siguientes se va dividiendo el grupo hasta llegar a que cada observación es un grupo, la razón por la cual no son muy usados es debido a que las posibles maneras en dividir por primera vez el grupo son demasiadas $(2^{n-1}-1)$, muchas más que en el caso aglomerativo.

3.2. Métodos Jerárquicos Aglomerativos

Estos métodos parten de la idea que cada elemento es un grupo, en los subsecuentes pasos se van combinando dos grupos de estos hasta llegar a que todas las observaciones son parte de un mismo grupo, en la primera unión hay $\binom{n}{2} = \frac{n(n-1)}{2}$ posibles maneras de hacerlo. Estas agrupaciones se pueden ver gráficamente en un dendrograma la grafica de árbol y es con base en éste que por lo general se decide el número de grupos definitivo.

3.2.1. Vecino más cercano (Simple Linkage)

Como su nombre lo dice, este método toma en cuenta la distancia mínima que hay entre cada grupo, es decir, el primer grupo que se une esta formado por los dos elementos, vistos ahora como grupos, con la distancia más pequeña entre ellos. Con base en estos nuevos grupos se busca la distancia mínima entre ellos y los dos grupos que cumplan con esta característica se unen, hasta formar todos un mismo grupo. El criterio que se utiliza para medir la distancia entre los grupos es la distancia más pequeña entre sus miembros más próximos, es decir, suponga que se tienen dos grupos A y B de uno o más elementos cada uno, entonces la distancia entre ellos es:

$$d(A,B) = mín\,\{d(i,j) : i\epsilon A, j\epsilon B\}$$

3.2.2. Vecino más lejano(Complete Linkage)

Este método es muy parecido al anterior, de lo que se trata es de ir uniendo los grupos que se encuentren más cercanos, salvo que ahora la distancia que hay entre cada grupo está dada por la distancia máxima entre los elementos de los mismos, es decir, si se tienen dos grupos A y B de uno o más elementos cada uno, la distancia entre ellos es:

$$d(A,B) = máx\,\{d(i,j) : i\epsilon A, j\epsilon B\}$$

lo que representa la distancia máxima entre grupos es el mínimo diámetro en el que se pueden encerrar los elementos del grupo.

3.2.3. Promedio entre grupos (Average Linkage)

En este caso, de lo que se trata es de no irse a los extremos como en los casos anteriores, sino de encontrar un promedio para medir la distancia que hay entre los grupos que van surgiendo. La idea es la misma, unir aquellos grupos cuya distancia entre ellos sea pequeña. Se tienen dos conjuntos A y B, el promedio que hace es considerando el número de elementos de cada grupo así como la distancia que hay entre los pares que se forman tomando un elemento de A y otro de B, si n_A es el número de elementos de A y $n_B$es el número de elementos de B, entonces:

$$d(A,B) = \frac{1}{n_A n_B} \sum_{i \epsilon A} \sum_{j \epsilon B} d(i,j)$$

3.2.4. Método de Ward

Este método consiste en unir aquellos grupos cuya suma de cuadrados no se incremente de forma importante al momento de unirse, ya que si la suma de cuadrados es pequeña podría presentarse de una homogeneidad entre los elementos del mismo al no estar muy separados. Si se tiene un conjunto A cuyos elementos los denotaremos por $X_{Ai} = (x_{Ai1}, ..., x_{Aip}), i = 1, ..., n_A$ y un conjunto B cuyos elementos denotaremos por $X_{Bi} = (x_{Bi1}, ..., x_{Bip}) \quad i = 1, ..., n_B$, sea $(\bar{x}_{A1}, ..., \bar{x}_{Ap})'$ el vector de medias del grupo A y sea $(\bar{x}_{B1}, ..., \bar{x}_{Bp})'$ el vector de medias del grupo B, definimos la suma de cuadrados dentro un grupo como $SCD_A = \sum_{j=1}^{p} \sum_{i \epsilon A} (x_{Aji} - \bar{x}_{Aj})^2$, análogo para el grupo B y para un grupo $C = A \cup B$, entonces la disimilaridad que se usa para medir la distancia entre grupos es el incremento que existe cuando se unen los dos grupos, es decir,

$$d(A,B) = SCD_C - (SCD_A + SCD_B)$$

El procedimiento básicamente es el mismo que el de los métodos anteriores, lo que cambia es la manera de calcular la distancia entre grupos.

Para poder tomar una decisión sobre los grupos, en muchas ocasiones se hace uso de los dendrogramas que son una forma matemática y pictórica de representar dichos resultados, vea la Figura 3.1. Cada nodo del dendrograma representa un grupo y el largo de los tallos representa la distancia con la que los grupos son unidos.

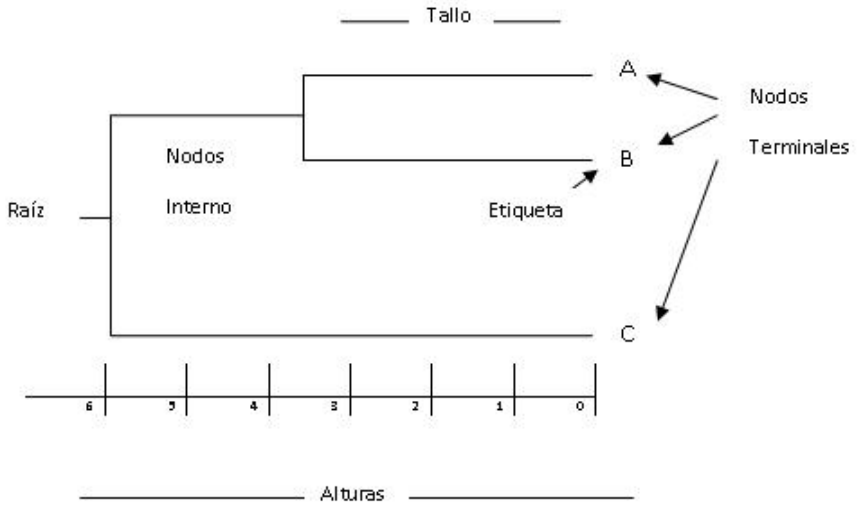

Figura 3.1: Dendrograma

3.2.5. Método de k-medias

Este es un método de partición e iterativo, el cual parte de suponer la existencia de k grupos, primero se escoge al azar los k grupos $A_1, ..., A_k$ y se calculan $\bar{X}_{A_i} = (\bar{x}_{A_i 1}, ..., \bar{x}_{A_i p})'$ para $i = 1, .., k$ que corresponden a los vectores de medias de cada grupo i. Después se calcula la distancia entre cada observación y cada media, es decir, por cada observación se calculan k distancias. En el momento que la distancia de alguna de las observaciones a la media de un grupo distinto al grupo al que pertenece dicha observación sea menor a la distancia que hay entre la observación y la media del grupo al que pertenece, se cambia dicha observación de grupo, al grupo donde la distancia era menor y se vuelven a hacer los cálculos de medias y a buscar algún otro cambio. Cuando ya no haya más cambios se detiene el algoritmo. Los grupos quedan conformados conforme a la última iteración.

Estos métodos arrojan grupos, de los cuales no se conocen sus propiedades estadísticas o de distribución, este es un defecto de estos métodos. Por esta razón en ocasiones se suele utilizar métodos que proporcione más de las propiedades de cada grupo como es el usar un modelo de mezclas finitas.

3.3. Mezclas Finitas para Conglomerados

Este es un método no jerárquico. Como ya se ha mencionado la idea principal de un modelo de mezclas es suponer que la muestra puede ser descrita a través de la suma

ponderada de normales (en general cualquier distribución). Aterrizando esta idea en el tema de conglomerados, se puede asociar a cada componente un grupo, es decir, el número de componentes dará el número de grupos con los cuales está conformada la muestra. Cada grupo compartiría las características de la componente, como media y varianza. De esta forma se tendría un mejor panorama de la distribución de cada grupo y con esto poder describirlo mejor.

En este caso lo más importante es encontrar el valor de g, se esperaría un número relativamente pequeño en comparación con el número de observaciones. El significado de un número grande es que se tendrían muchos grupos a describir, lo cual implica un mayor costo.

Por ejemplo, si se tiene una base de datos de un grupo de personas que asisten al psicólogo, a estas personas se les hace una encuesta y al hacer un análisis a estos datos, se encuentran diferentes opciones de grupos en base a las variables medidas. Una de ellas es: católicos practicantes, católicos no practicantes, protestantes practicantes, protestantes no practicantes, budistas practicantes, budistas no practicantes y ateos, son siete grupos con grandes diferencias, son muchos si la idea es contratar un psicólogo para cada grupo, tendría que contratar varios y esto sería costoso. Sin embargo, si se unen los grupos de tal manera que me queden como católicos, protestantes, budistas y ateos, estos son sólo cuatro, esta idea es más atractiva.

Por otro lado, la etiqueta de grupo, como católicos o no católicos, no me la da el modelo de mezclas, ni ningún modelo. Esta se elige con base en los elementos de cada componente en este caso. Las observaciones se asignan a cada componente con base al valor esperado de la variable $\mathbf{z}_j \quad j = 1, ..., n$ ya que ésta tiene el valor de uno en la entrada correspondiente a la componente a la que pertenece y cero en las demás.

3.3.1. Ejemplos de Conglomerados vía un Modelo de Mezclas Finitas

Ejemplo 3.1. Las variables de este ejemplo se obtuvieron de la página de internet del INEGI, todas estas variables son del año 2005. La base de datos consta de cuatro variables:

1. Homicidios - Porcentaje de muertes por homicidio con respecto al total de muertes por entidad federativa.

2. Derechohabientes - Porcentaje de población derechohabiente por entidad federativa.

3. Bajo Peso - Porcentaje de nacidos vivos con bajo peso al nacer por entidad federativa.

4. Alfabetas - Porcentaje de población de 15 y más años que es alfabeta por entidad federativa.

Las observaciones son 32, correspondientes a los 31 estados y el Distrito Federal. Lo que se espera es encontrar grupos en las entidades en base a estas variables.

Primero se hace un análisis descriptivo que indique de alguna manera cuántos grupos podrían existir.

En la Figura 3.2 se muestran las caras correspondientes a cada entidad federativa, a simple vista podría decirse, sin ser muy estrictos, que existen cinco grupos. Aunque podrían ser más grupos, ya que no se aprecia mucha similitud entre ellos, o por lo menos no en este gráfico.

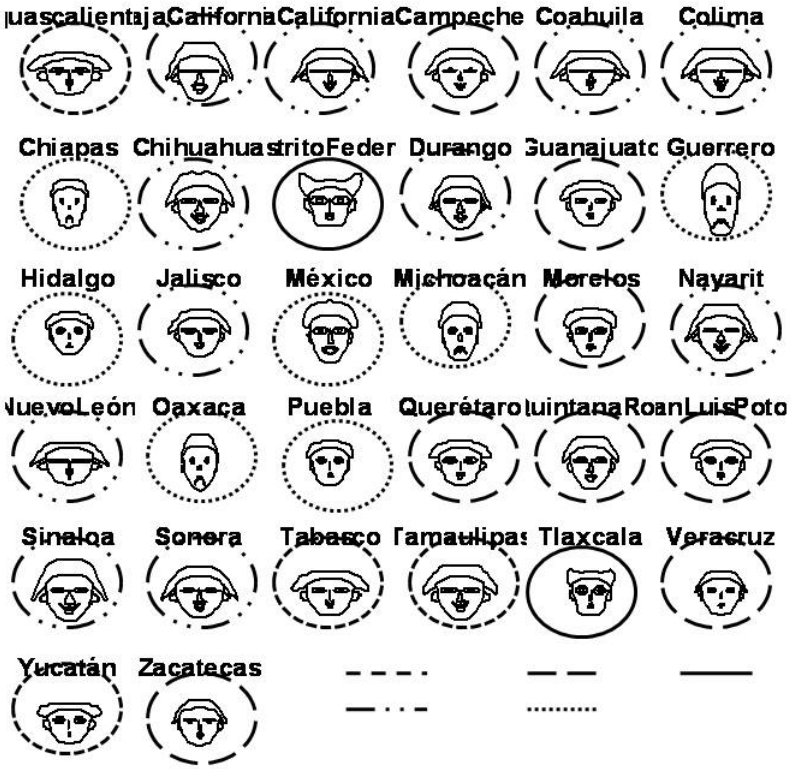

Figura 3.2: Diagrama de Caras para el Ejemplo 3.1

La Figura 3.3 muestra la gráfica de estrellas correspondientes a cada entidad, en este se puede ver cuatro grupos. En esta gráfica a diferencia de la de caras, se notan los grupos un poco más homogéneos. Sobre todo el formado por Chiapas, Guerrero y Oaxaca. Las estrellas de estos tres estados son muy delgadas y no se parecen a ninguna otra, sin embargo en el diagrama de caras se muestra parecido de estos tres estados con Michoacán.

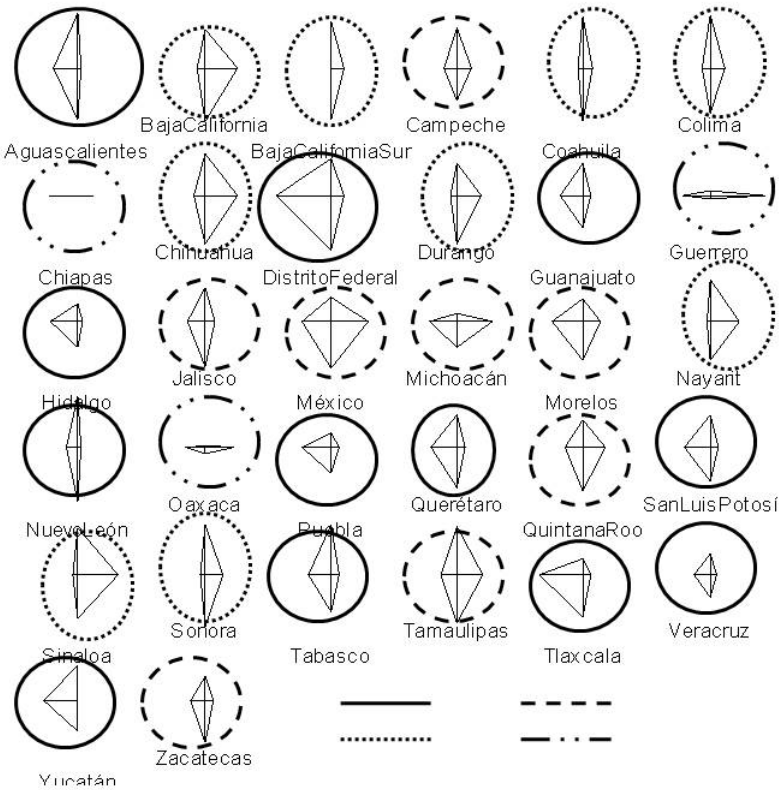

Figura 3.3: Diagrama de Estrellas para el Ejemplo 3.1

Para ver de una manera conjunta a los datos, debería existir una gráfica en cinco dimensiones, es por eso que se grafica en dos dimensiones cada par de variables, esto se muestra en la Figura 3.4. De estas gráficas no es fácil identificar grupos. Quizás la gráfica de Derechohabientes vs Alfabetas si se podría notar dos grupos, uno de ellos con de tres elementos. Pare ver esto un poco mejor las Figuras 3.5 y 3.6 se muestran las gráficas en 3D de algunas variables.

En la Figura 3.5 que grafica las variables de Bajo Peso, Derechohabientes y Alfabetas en 3D, se observan tres grupos, dos de ellos muy pequeños y un tercero que abarca la mayoría de las observaciones. Además de estos se nota una observación muy alejada del resto de las demás.

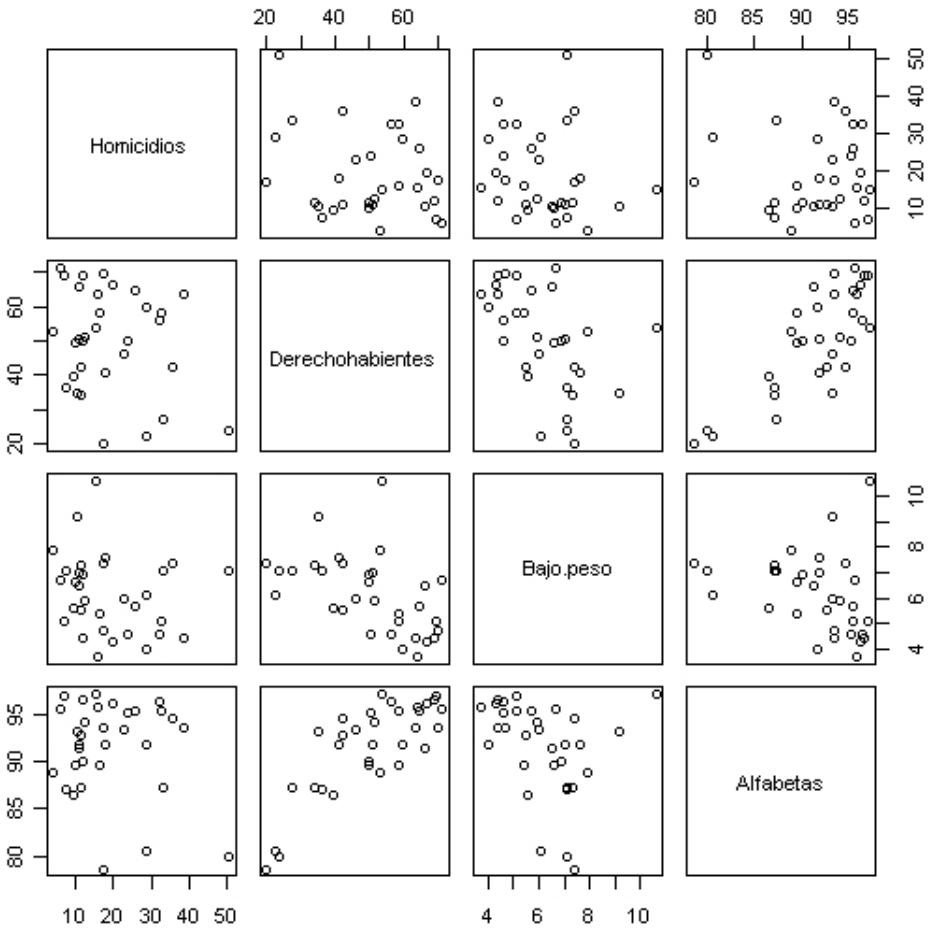

Figura 3.4: Gráfica de dispersión por cada par de variables del Ejemplo 3.1

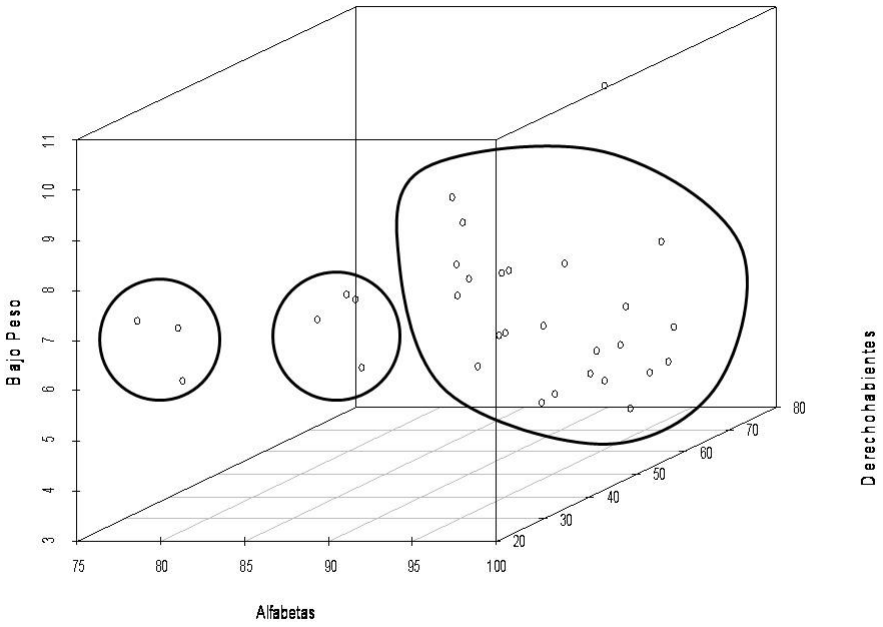

Figura 3.5: Gráfica de dispersión en 3D del Ejemplo 3.1

Sin embargo, en la Figura 3.6 cuyas variables graficadas son derechohabientes, Homicidios y Bajo Peso, se observan sólo dos grupos. El número de observaciones de cada grupo es muy parecido. De nuevo se puede ver aquel dato que se encuentra muy alejado del resto de las observaciones.

Una vez hecho este análisis, se aplica el modelo de mezclas finitas sobre toda la muestra para el caso de dos, tres, cuatro y cinco grupos. Los resultados de de estos análisis se muestran en el Cuadro 3.1. De este se deduce que el mejor modelo sería con tres grupos, sin embargo no se descarta la idea de dos.

N° de Componentes	BIC	Modelo
2	-813.9107	EEV
3	-811.5866	EVI
4	-814.8693	EEI
5	-826.5362	EEI

Cuadro 3.1: Criterio de Información Bayesiana para el Ejemplo 3.1.

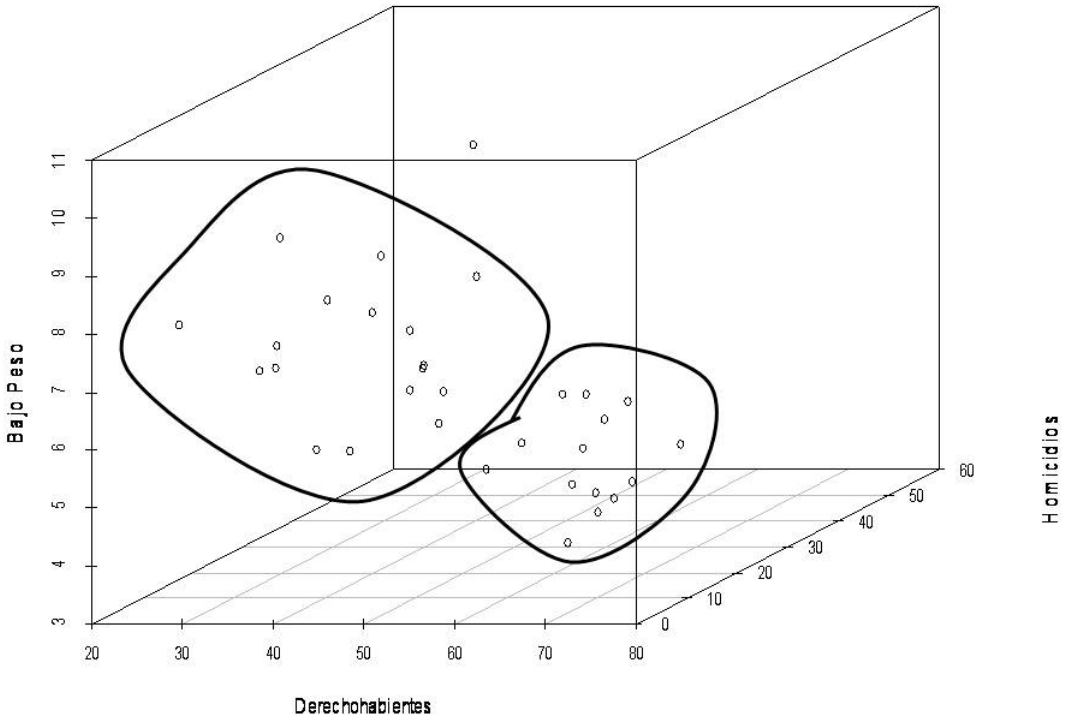

Figura 3.6: Gráfica de dispersión en 3D del Ejemplo 3.1

Como los datos son los estados de la República Mexicana, se usa un mapa para visualizar mejor los resultados.

Cuando hay dos grupos, estos se agrupan como en la Figura 3.7. En esta podemos ver que los grupos los se podrían etiquetar como estados del norte y estados de la costa sur en el océano Pacífico a grandes rasgos. Geográficamente se ven bien, aunque no son muchos los estados del segundo grupo, si ocupan una considerable región.

Figura 3.7: República Mexicana dividida en dos grupos.

En el Cuadro 3.2 están las medias de estos grupos. Con base en esto se puede decir que el grupo 1 tiene un mejor nivel educativo, de seguridad social y de seguridad pública que en el otro grupo. De alguna manera se diría que con base en estas variables se pueden agrupar los estados en dos donde uno corresponde a los estados mejor desarrollados y el otro con un desarrollo menor.

	Homicidios	Derechohabientes	Bajo Peso	Alfabetas
Grupo 1	16.53583	55.77083	5.625	92.65833
Grupo 2	26.19363	33.1875	7.8125	87.8875

Cuadro 3.2: Medias del modelo de mezclas finitas con dos grupos

Ahora para el caso de tres grupos, la distribución geográfica se muestra en la Figura 3.8. Se observa que los integrantes del grupo dos y tres, en azul y amarillo respectivamente,

se encuentran muy cercanos y en el caso del grupo 1, en rojo, aunque en el norte se ven juntos, también tiene integrantes en el sur y centro.

Figura 3.8: República Mexicana dividida en tres grupos. 1-Rojo; 2-Azul; 3-Amarillo

El Cuadro 3.3 están las medidas de cada grupo. De éstas se puede decir que una de las características de los grupos con respecto a los otros son:
Grupo 1 : Mejor educación, buena nutrición, buena seguridad social.
Grupo 2 : Con menor porcentaje de homicidios.
Grupo 3 : Menor educación, menor seguridad social y mayor cantidad de homicidios.

A grandes rasgos la etiqueta que usaría para describir a cada grupo es de nivel 1, 2 y 3 ; en donde el nivel uno significa que son estados muy atractivos y nivel 3 poco atractivos.

	Homicidios	Derechohabientes	Bajo Peso	Alfabetas
Grupo 1	22.699	60.75	5.05	94.97857
Grupo 2	11.309786	47.135714	7.078571	90.764286
Grupo 3	32.5715	23.4	6.925	81.625

Cuadro 3.3: Medias del modelo de mezclas finitas con tres grupos

En conclusión me quedaría con el modelo de tres grupos ya que a mi consideración agrupa mejor los estados.

Ejemplo 3.2. La base de datos que se usa para este ejemplo corresponde a las características de ciertos alimentos, estos datos se tomaron de la referencia [15]. De los 27 alimentos diferentes se consideran la cantidad de calorías, proteínas, grasa, calcio y hierro por cada uno de ellos. Es decir, tenemos una base de datos de 27 observaciones con cinco variables cada una.

Los datos se muestra en el Cuadro 3.4. Como se puede ver el rango de las variables es muy diferente, por esta razón lo más conveniente es estandarizar la base. Por lo tanto la base con la que se trabaja es la que se le resta el vector de medias y se multiplica por la inversa de la matriz de covarianzas a la un medio.

	Energía (cal)	Proteínas (g)	Grasas (g)	Calcio (mg)	Hierro (mg)
Carne de res a fuego lento	340	20	28	9	2.6
Hamburguesa	245	21	17	9	2.7
Carne de res asada	420	15	39	7	2
Carne de res filete	375	19	32	9	2.6
Carne de res enlatados	180	22	10	17	3.7
Pollo a la parrilla	115	20	3	8	1.4
Pollo en conserva	170	25	7	12	1.5
Carne corazón	160	26	5	14	5.9
Pierna de cordero asado	265	20	20	9	2.6
Lomo de cordero asado	300	18	25	9	2.3
Jamón ahumado	340	20	28	9	2.5
Cerdo asado	340	19	29	9	2.5
Carne de cerdo a fuego lento	355	19	30	9	2.4
Carne de la lengua	205	18	14	7	2.5
Carne de ternera chuleta	185	23	9	9	2.7
Pescado al horno	135	22	4	25	0.6
Almejas en bruto	70	11	1	82	6
Almejas enlatados	45	7	1	74	5.4
Carne de cangrejo en lata	90	14	2	38	0.8
Haddock fritos	135	16	5	15	0.5
Borrego a la parrilla	200	19	13	5	1
Borrego en lata	155	16	9	157	1.8
Perch fritos	195	16	11	14	1.3
Salmón enlatado	120	17	5	159	0.7
Sardinas en lata	180	22	9	367	2.5
Atún enlatado	170	25	7	7	1.2
Camarones enlatados	110	23	1	98	2.6

Cuadro 3.4: Nutrientes en Carne, Pescado y Pollo

En la Figura 3.9 se pueden notar tres grupos en la muestra, sin embargo, en la Figura 3.10 se notan al menos cinco grupos y uno de una observación. El orden de cada observación es el mismo en ambos casos.

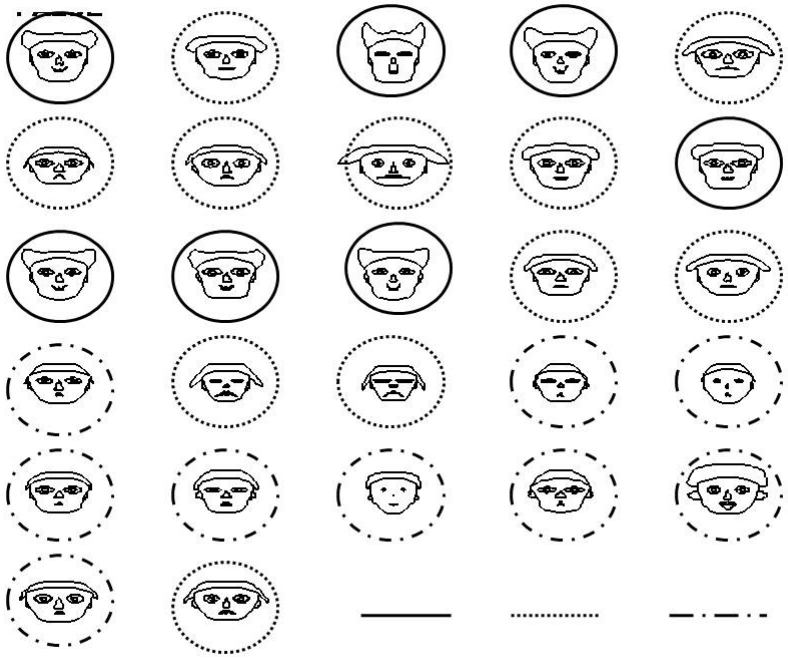

Figura 3.9: Gráfica de rostros del Ejemplo 3.2

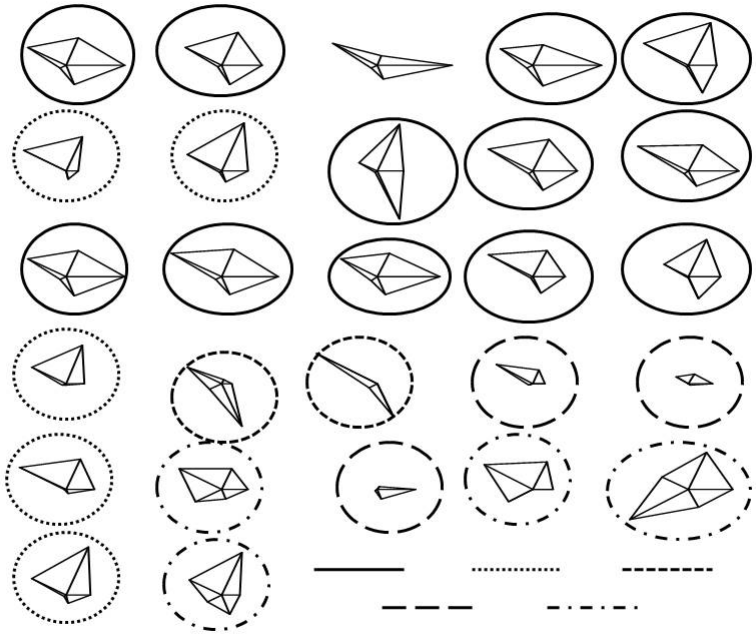

Figura 3.10: Gráfica de estrellas del Ejemplo 3.2

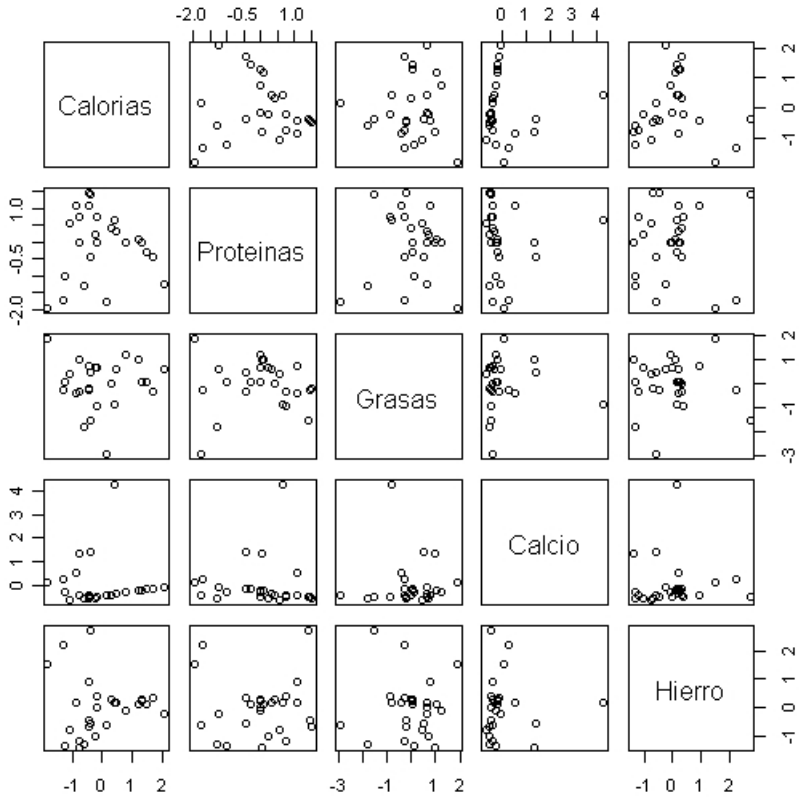

Figura 3.11: Grafica de dispersión por cada par de variables del Ejemplo 3.2

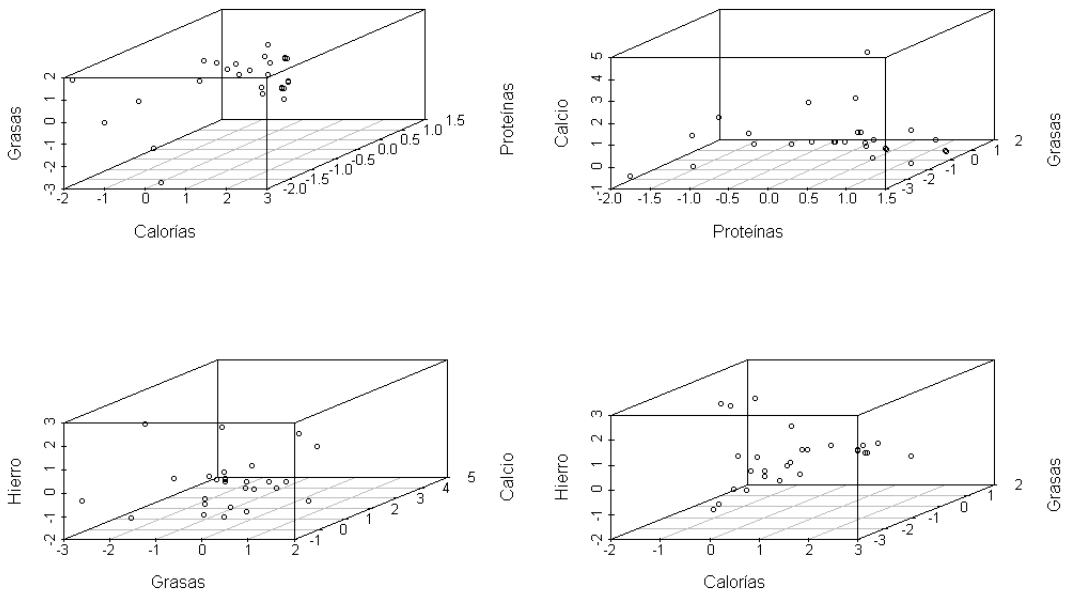

Figura 3.12: Gráfica de dispersión tres dimensiones de las variables del Ejemplo 3.2

Cuando se usa un modelo de mezclas finitas para encontrar los grupos se obtiene que el mejor resultado es de siete grupos. Los resultados del BIC se muestran en la Gráfico 3.13. Este valor no concuerda mucho con el análisis descriptivo anterior.

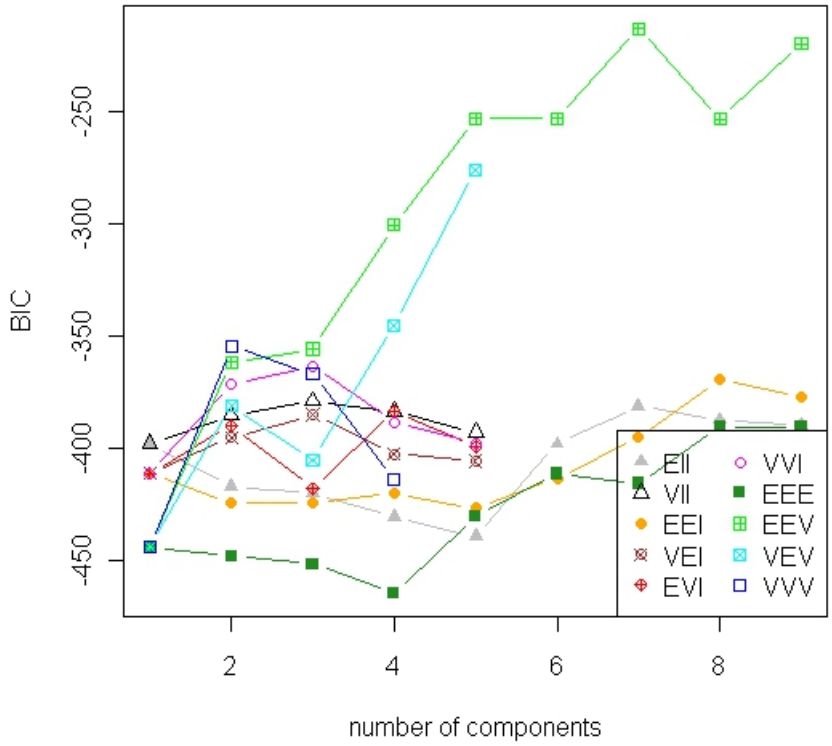

Figura 3.13: BIC del Ejemplo 3.2 usando un modelo de mezclas finitas.

Sin embargo, si se usa un modelo jerárquico, por ejemplo el método Ward, usando la distancia euclideana, se obtiene otro resultado. La Figura 3.14 corresponde al dendrograma de este modelo. Si se corta en una altura aproximada de 7, se notan tres grupos. Esto ejemplifica el hecho de que si los grupos no tienen una distribución normal, el Modelo de Mezclas Finitas detecta muchas componentes para describir los datos, sin embargo los métodos jerárquicos al no tener supuestos como éstos, pueden detectar menos grupos. Sin embargo no se conocen las propiedades estadísticas de estos.

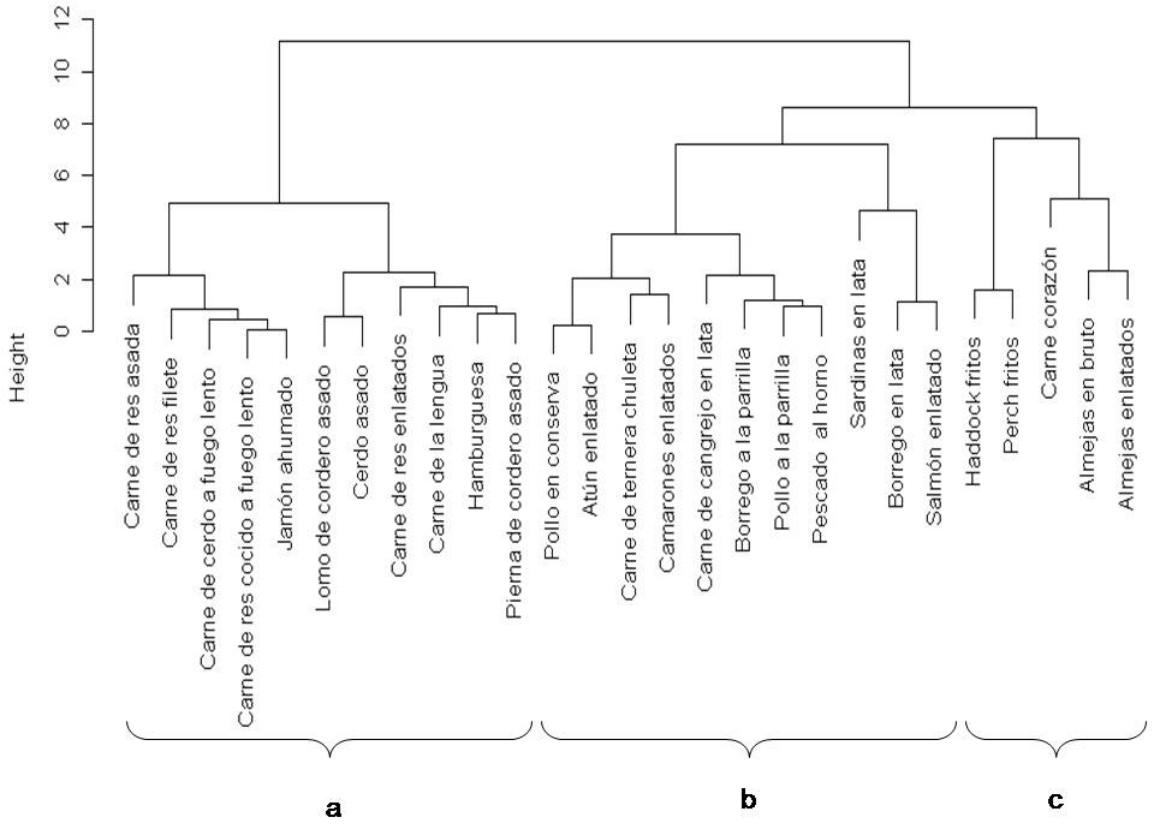

Figura 3.14: Dendograma del Ejemplo 3.2 usando el Método Ward y la Distancia Euclideana.

Las Gráficas 3.15, 3.16 y 3.17, muestran la gráfica de dispersión de cada grupo, esto para tratar de caracterizar cada uno de ellos.

El Grupo (a) tiene como principal característica que las grasas son muy altas a diferencia de los otros dos, la media es de 24.7272 gramos, mientras que las medias de los otros grupos son de 6.27 y 4.6 gramos. Además los valores de calorías también son grandes, pero no tan diferentes a los demás como en el caso de la grasa. Las comidas no tienen mucho calcio.

La principal característica del Grupo (b) es que el valor del calcio es superior a los otros dos grupos, los valores están entre 0 y 400 miligramos, mientras que para los otros grupos están entre (0,20) y (0,80). Tiene las menores cantidades de hierro.

El Grupo (c) tiene los alimentos con poca grasa, pocas calorías y pocas proteínas, pero hierro tiene más que en los otros grupos.

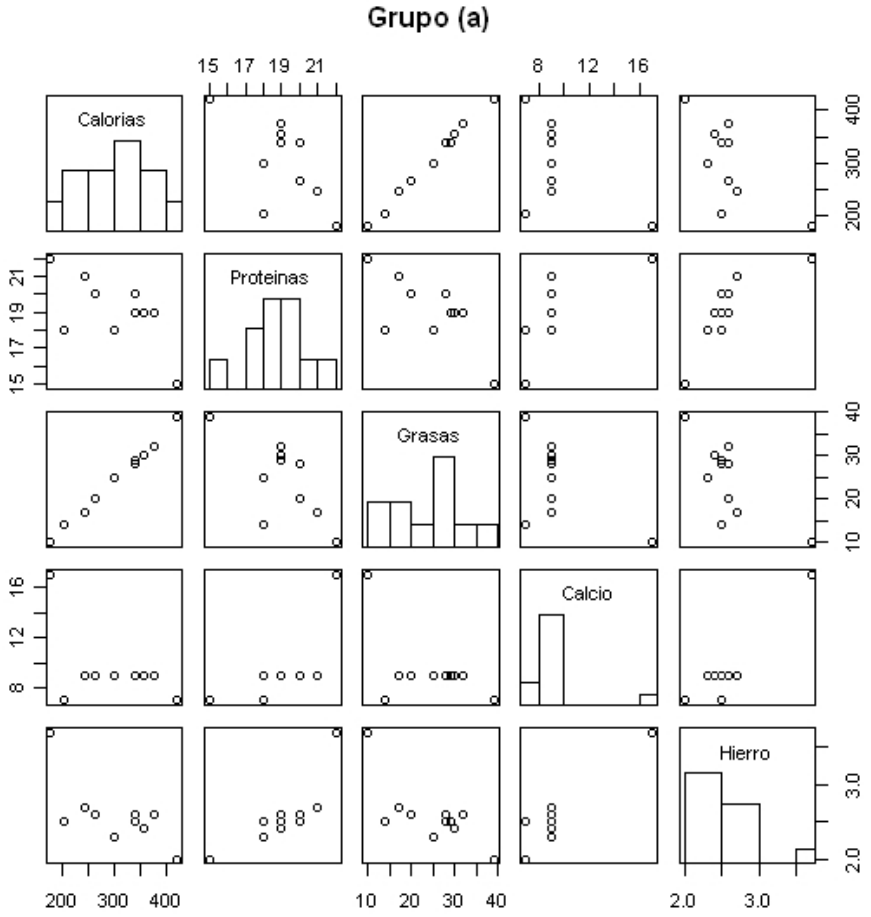

Figura 3.15: Gráfica de dispersión e histograma por variable del grupo (a) usando Método Ward

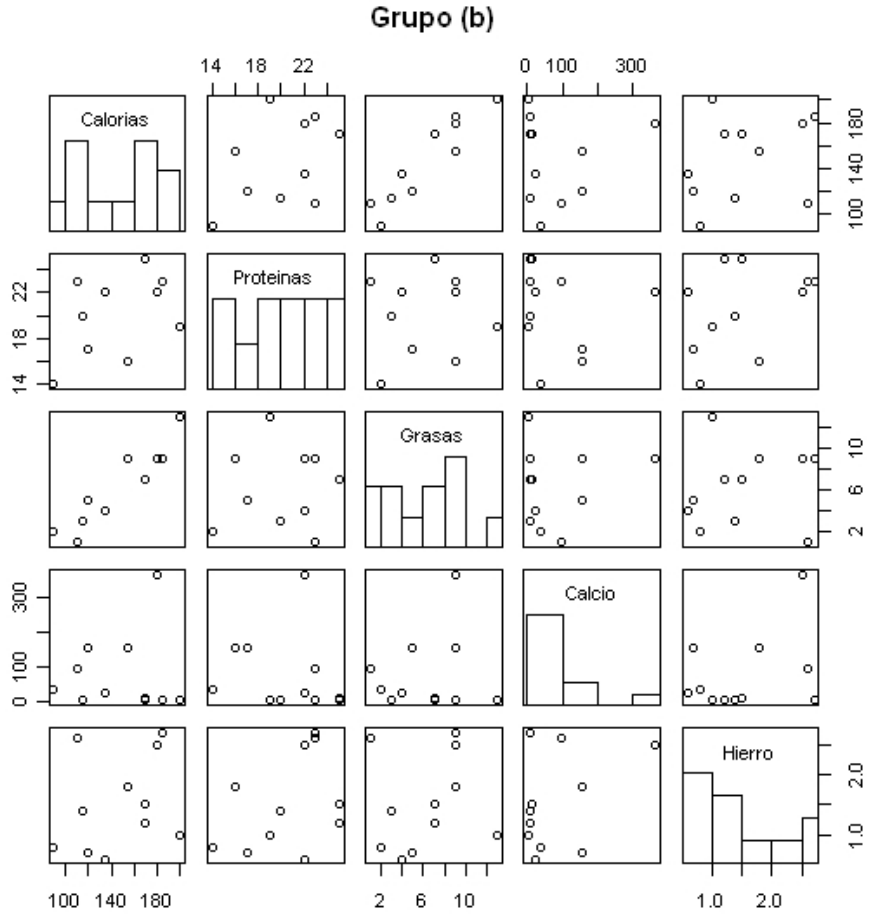

Figura 3.16: Gráfica de dispersión e histograma por variable del grupo (b) usando Método Ward

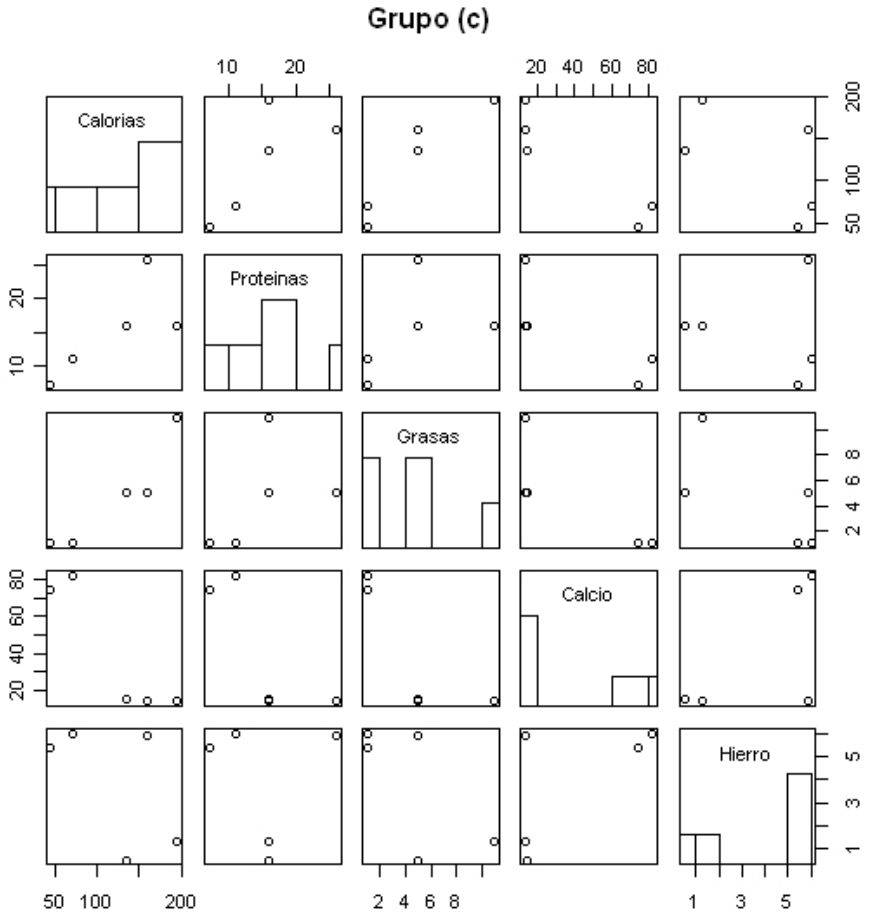

Figura 3.17: Gráfica de dispersión e histograma por variable del grupo (c) usando Método Ward

Se puede decir que este es un claro ejemplo en donde un modelo jerárquico muestra una mejor manera de describir los grupos detectados.

Ejemplo 3.3. Los datos que se analizarán a continuación corresponden a los puntajes alumnos de cuarto y sexto grado en 25 primarias de New Haven. Por cada grado se mide el grado de lectura y el grado de aritmética. Tomando esta base de datos de la Referencia [15]. Los cuales se muestran en el Cuadro 3.5.

Escuela	Cuarto Grado		Sexto Grado	
	Lectura	Aritmética	Lectura	Aritmética
Baldwin	2.7	3.2	4.5	4.8
Barnard	3.9	3.8	5.9	6.2
Beecher	4.8	4.1	6.8	5.5
Brennan	3.1	3.5	4.3	4.6
Clinton	3.4	3.7	5.1	5.6
Conte	3.1	3.4	4.1	4.7
Davis	4.6	4.4	6.6	6.1
Day	3.1	3.3	4	4.9
Dwight	3.8	3.7	4.7	4.9
Edgewood	5.2	4.9	8.2	6.9
Edwards	3.9	3.8	5.2	5.4
Hale	4.1	4	5.6	5.6
Hooker	5.7	5.1	7	6.3
Ivy	3	3.2	4.5	5
Kimberly	2.9	3.3	4.5	5.1
Lincoln Bassett	3.4	3.3	4.4	5
Lovell	4	4.2	5.2	5.4
Prince	3	3	4.6	5
Ross	4	4.1	5.9	5.8
Scranton	3	3.2	4.4	5.1
Sherman	3.6	3.6	5.3	5.4
Truman	3.1	3.2	4.6	5
West Hills	3.2	3.3	5.4	5.3
Winchester	3	3.4	4.2	4.7
Woodward	3.8	4	6.9	6.7

Cuadro 3.5: Puntajes de los escuelas primarias de New Haven

De nuevo, es bueno hacer un análisis descriptivo que dé una mejor idea del número de componentes. En este caso se tienen cuatro variables, no se puede trabajar con un histograma. Es por esto que se usarán aquellas gráficas que permitan una mayor visualización, como lo son la gráfica de rostros y de estrellas.

En la Figura 3.18 que corresponde a la gráfica de rostros, se observan tres grupos, uno de rostros grandes, otro de rostros pequeños y ojos medianos y una último de rostro chico y ojos pequeños. Sin embargo, para el grupo de rostro pequeño y ojos medianos sólo hay

cuatro escuelas, quizás sea conveniente agruparlo a alguno de los otros dos grupos.

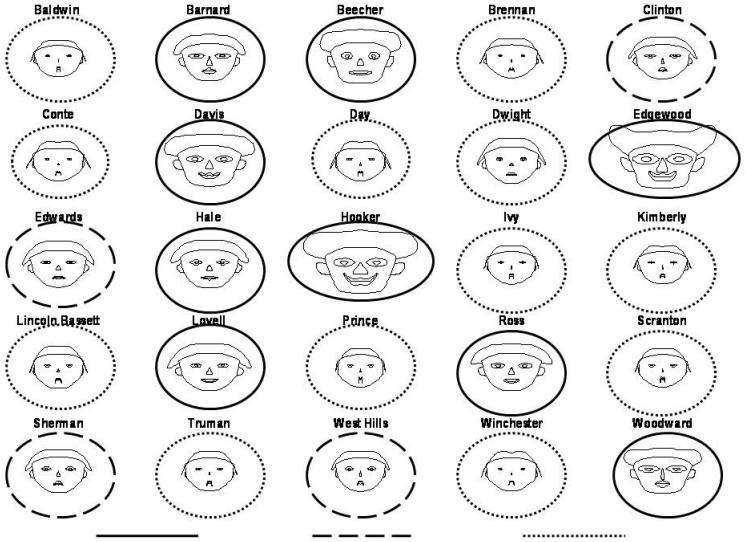

Figura 3.18: Diagrama de Caras del Ejemplo 3.3

A simple vista de Figura 3.19 no muestra la existencia de grupos, como es de esperarse en una gráfica de estas dimensiones. Sería más fácil de visualizar aglomeraciones si el número de variables correspondiera a la dimensión de la gráfica. Lo que si se observa es la existencia de una dependencia lineal en las variables. Esta dependencia se ve más en las gráficas de Lectura vs Aritmética de cualquiera de los grados.

Cuando se grafica en tres dimensiones, como se muestra en la Figura 3.20, se nota una misma tendencia lineal. En este gráfico es un poco más notoria la aglomeración de algunas observaciones, sobre todo en la parte baja de las gráficas. Además se ven dos datos en la parte superior que quizás formen parte de un grupo muy pequeño. No es tan despreciable esta idea ya que el número de observaciones es algo pequeño.

Una vez hecho este análisis descriptivo se desarrolla el análisis de modelo de mezclas finitas, el resultado del BIC de éste se ve en la Figura 3.21. Aquí podemos notar que se tiene el mejor modelo cuando existen dos grupos. El BIC para el caso de tres grupos se encuentra algo distante del de dos, no tendría mucho sentido suponer tres o más. Son tres modelos los que mejor describen los datos considerando dos componentes. El VVV, cada matriz de covarianzas es diferente, con BIC igual a 110.17. Le sigue el modelo EEE, con BIC igual a 109.0594 y finalmente el modelo VEV con BIC igual a 108.153. De estos valores, escogemos el modelo VEV.

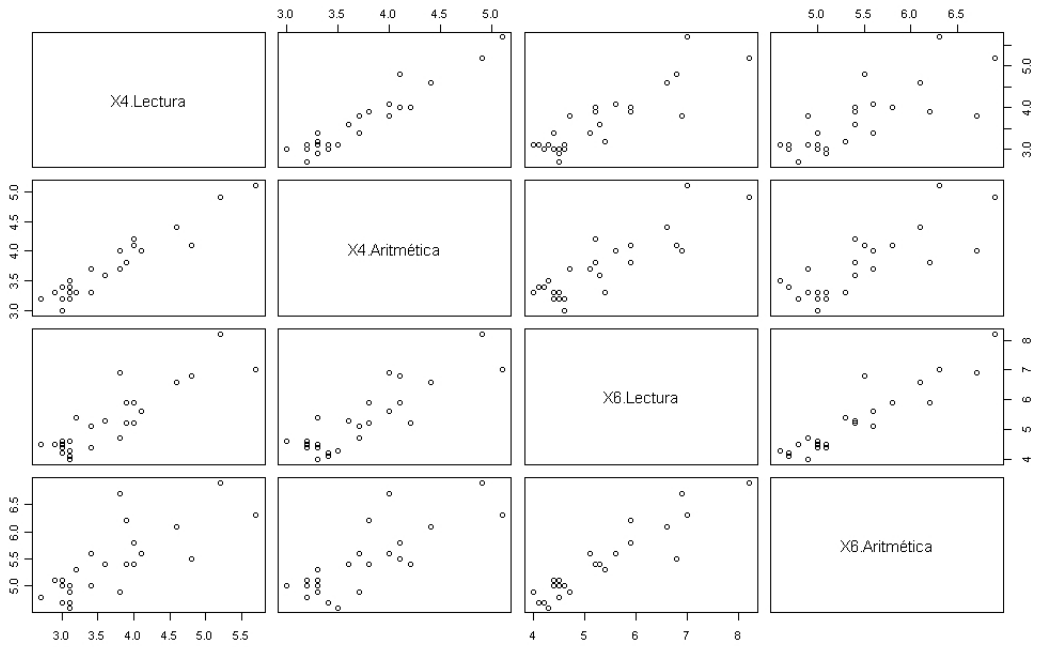

Figura 3.19: Grafica de dispersión por cada dos variables del Ejemplo 3.3

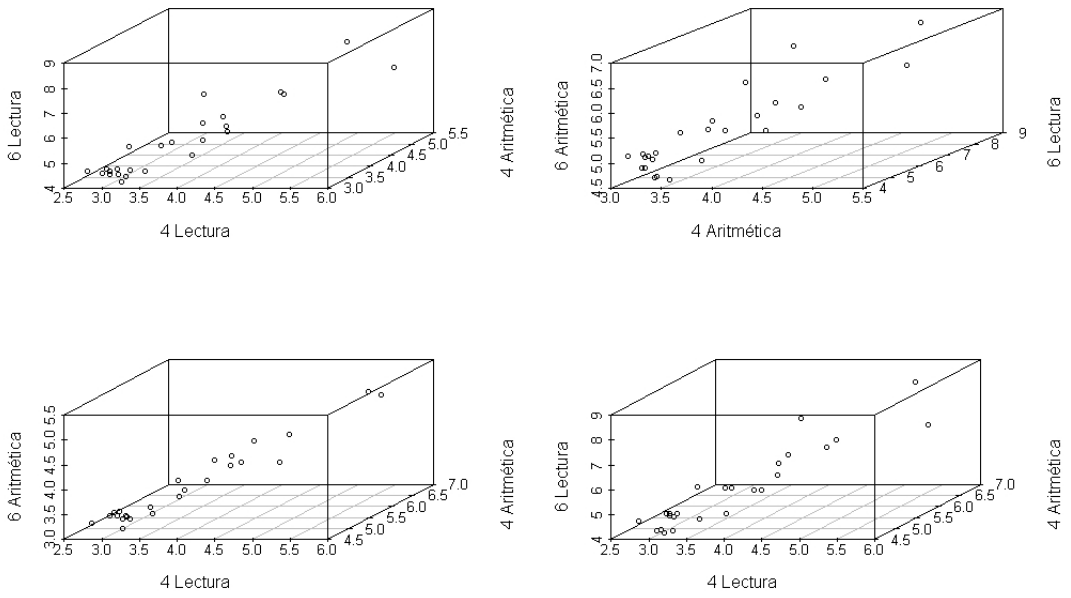

Figura 3.20: Gráfica de dispersión en 3D del Ejemplo 3.3

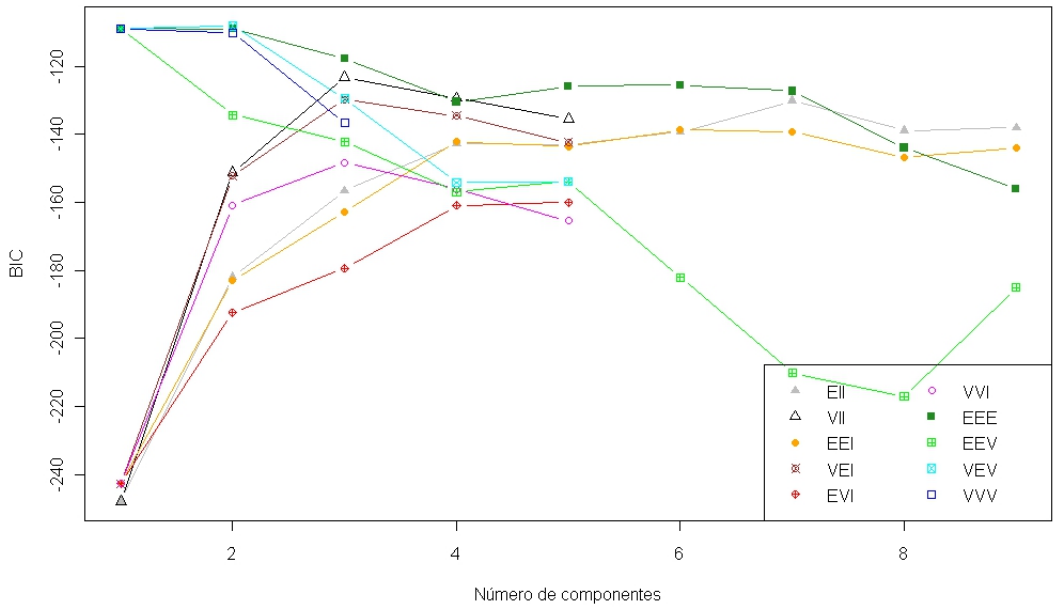

Figura 3.21: BIC del Ejemplo 3.3

Una vez que se elige el modelo con el número de componentes apropiado, se analiza cada uno de los grupos para verificar si esta partición es favorable para catalogar o etiquetar los grupos. Lo anterior recordando que la información de la pertenencia de cada observación dentro de algun grupo, la da el propio algoritmo de Esperanza-Maximización.

Las observaciones correspondientes a cada grupo se muestran en el Cuadro 3.6. El primer grupo está conformado por diez escuelas, mientras que el otro tiene quince. Esta diferencia no es tan grande.

La media y la varianza de cada grupo se muestran en el Cuadro 3.7. Cada entrada del vector de medias del Grupo 1, es menor a las del Grupo 2. Como las variables muestran el puntaje para cada escuela, entre mayor mejor es el nivel de la escuela, entonces esto indica que las escuelas del Grupo 2 tienen un mejor nivel educativo. En cuanto a la varianza, el Grupo 2 tiene los valores más altos en la matriz de covarianza, lo que indica que estos datos están más dispersos. Esto se ve en la Figura 3.22, en donde los triángulos que corresponden al Grupo 1 están más juntos que los otros. En esta gráfica se puede ver además que hay una separación definida, salvo por una o dos observaciones del Grupo 2, las cuales se ven mezcladas entre las del Grupo 1. Pero en la gráfica de 4 Aritmética vs 6 Aritmética la separación es más clara.

Ahora que ya se ha clasificado los datos, al observar esta gráfica es fácil ver estos dos grupos. A simple vista parecen bien agrupados y además es conveniente ver un grupo está por encima del otro. Esto ayuda al momento de querer nombrar o etiquetar los grupos.

Grupo 1				
Baldwin	Brennan	Conte	Day	Ivy
Lincoln Bassett	Prince	Scranton	Truman	Winchester
Grupo 2				
Barnard	Beecher	Clinton	Davis	Day
Dwight	Edgewood	Edwards	Hale	Hooker
Kimberly	Lovell	Ross	Sherman	West Hills

Cuadro 3.6: Agrupación de las escuelas segun el modelo de mezclas finitas con dos componentes.

Grupo	Vector de Media	Matriz de Covarianzas
1	$\begin{pmatrix} 3.050357 \\ 3.275072 \\ 4.357354 \\ 4.875214 \end{pmatrix}$	$\begin{pmatrix} 0.03060185 & 0.0164785 & -0.0229470 & -0.00571635 \\ 0.01647851 & 0.0468029 & -0.0630242 & -0.04757687 \\ -0.02294708 & -0.0630242 & 0.1020431 & 0.06380806 \\ -0.00571635 & -0.0475768 & 0.0638080 & 0.05769840 \end{pmatrix}$
2	$\begin{pmatrix} 4.035048 \\ 3.978953 \\ 5.850945 \\ 5.728443 \end{pmatrix}$	$\begin{pmatrix} 0.4006648 & 0.2572676 & 0.3990832 & 0.1560892 \\ 0.2572676 & 0.1959796 & 0.2714395 & 0.1396493 \\ 0.3990832 & 0.2714395 & 0.7165403 & 0.3492584 \\ 0.1560892 & 0.1396493 & 0.3492584 & 0.2523904 \end{pmatrix}$

Cuadro 3.7: Parámetros de cada grupo.

Con base en lo anterior, se pueden identificar dos grupos, escuelas de alto y de bajo nivel. Este análisis además muestra que son pocas las escuelas con bajo nivel. Es importante que solo hayan sido dos grupos si el interés de catalogarlos es para destinar fondos para becas por alto rendimiento, así solo habría dos tipos de subsidios.

En conclusión usando un modelo de mezclas finitas tenemos una buena clasificación de los datos, con dos grupos.

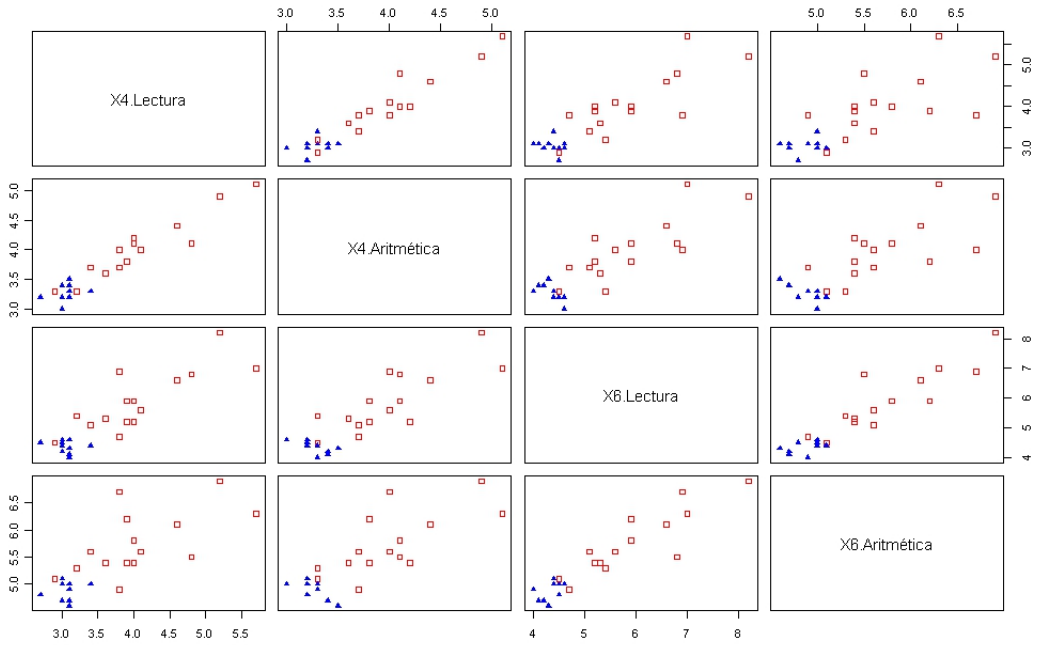

Figura 3.22: Gráfica de dispersión de la clasificacion del Ejemplo 3.3 usando un modelo de mezclas finitas con dos componentes

Conclusiones

Si se tiene un conjunto de observaciones las cuales se quieren agrupar, se pueden utilizar diferentes técnicas para encontrar dichos grupos. Es importante destacar, que cuando no hay alguna referencia sobre el número de componentes es adecuado hacer un análisis descriptivo de los datos.

Dentro de los métodos para encontrar el número de componentes, es usar un modelo de mezclas finitas, en donde a cada componente le corresponde un grupo. Entonces se hacen pruebas o se consideran ciertos criterios para determinar el número de grupos que mejor describa los datos. A veces un método jerárquico puede dar una mejor descripción, sobre todo si se está usando un supuesto fuerte como lo es la normalidad. Quizás los grupos no sigan esta distribución, entonces el modelo no servirá de mucho como se vió en el ejemplo 3.2.

Capítulo 4

Análisis Discriminante

Con frecuencia surge la necesidad de encontrar una manera de poder identificar a los individuos como parte de un grupo, es decir, una vez que se tiene bien definida la existencia de diferentes grupos dentro de una población y existe un grupo nuevo de individuos a se desea encontrar una manera de etiquetarlos dentro de algún grupo, como por ejemplo en el caso de un análisis de mercado, existe la pregunta si con base en las características del individuo compraría o no el producto, en medicina si el paciente responderá positivamente a algún medicamento, etc. Para estos casos puede hacer uso de un Análisis Discriminante que es una técnica multivariada que sirve para hallar esa regla o forma de discriminar a la población con base en las características medidas en cada individuo.

Suponga que se tiene una población de n individuos agrupados en q grupos, cada individuo con p variables, el tamaño de cada grupo es n_k con $k = 1, ..., q$, es decir, $\sum_{k=1}^{q} n_k = n$

En ocasiones se puede encontrar que las variables que se utilizan para hacer este tipo de análisis, no son útiles ya que los valores que toman las variables en la mayoría de los grupos o en todos son muy parecidos. Esto se puede ver en un análisis descriptivo en donde se compararían las medias de cada uno de los grupos, es decir, sea $(\bar{x}_{k1}, ..., \bar{x}_{kp})'$ el vector de medias del grupo k, y $(\bar{x}_{j1}, ..., \bar{x}_{jp})'$ el vector de medias del grupo j,. Si la distancia entre $\bar{x}_{j1}$ y $\bar{x}_{k1}$ es muy pequeña y ocurre lo mismo con la media de otros grupos, podría ser que dicha variable correspondiente a la entrada no aporta mucho para el análisis ya que no establece diferencias entre los grupos.

Dentro de los métodos más comunes para discriminar es el Discriminante Lineal de Fisher, sin embargo existen algunos otros que utilizan supuestos sobre la distribución de la muestra.

4.1. Discriminante Lineal de Fisher

Este método consiste en obtener funciones lineales a partir de q grupos en la muestra medida en p variables, se llamará $\prod_k$ al conjunto de elementos del grupo k, el número total de elementos de este grupo es n_k, de tal manera que maximicen la varianza entre

grupos y minimicen la varianza dentro de cada grupo, una vez calculadas estas funciones se buscan los criterios para discriminar.

Sea $X_{1\times p} = [x_1 ... x_p]$, el vector correspondiente a una observación, ahora se expresarán dichas funciones discriminantes, las cuales son lineales y deben ser no correlacionadas, en forma matricial

$$Y = X\boldsymbol{a} = \begin{bmatrix} x_1 & \cdots & x_p \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ & \ddots & \\ a_{p1} & & a_{pm} \end{bmatrix}$$

$\Rightarrow$

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p \\ &\vdots \\ y_m &= a_{1m}x_1 + a_{2m}x_2 + \cdots + a_{pm}x_p \end{aligned}$$

donde m corresponde al número de función es discriminantes que se usan $m = mín(q-1, p)$

El problema se trasforma en encontrar $\boldsymbol{a}$ que maximice la varianza entre los grupos y minimice la varianza entre los grupos, se puede calcular la matriz de covarianzas de la siguiente forma:

$$Cov(Y) = Cov(Xa) = a'Cov(X)a$$

Ahora, se puede calcular la covarianza total como la suma de la covarianza muestral entre los grupos y la covarianza dentro de cada grupo, con base en la matriz de datos X_{nxp}

$$Cov(x_i, x_j) = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) \text{ para } i, j = 1, ..., p$$

se considerará ahora las medias de las variables por cada uno de los q grupos, es decir, $\bar{x}_{kj} = \frac{1}{n_k}\sum_{i\epsilon\prod_k} x_{ij}$, corresponde a la media de la variable j del grupo k, entonces la media total de cada variable se puede ver de la siguiente forma:

$$\begin{aligned} \bar{x}_j &= \frac{1}{n}\sum_{i=1}^{n} x_{ij} \\ &= \frac{1}{n}\sum_{k=1}^{q}\sum_{i\epsilon\prod_k} x_{ij} \\ &= \frac{1}{n}\sum_{k=1}^{q} n_k\bar{x}_{kj} \\ &= \sum_{k=1}^{q}\frac{n_k}{n}\bar{x}_{kj} \end{aligned}$$

se puede reescribir la covarianza anterior usando lo siguiente

$$(x_{ki} - \bar{x}_i) = (x_{ki} - \bar{x}_{ki}) + (\bar{x}_{ki} - \bar{x}_i)$$

análogo para j

$$(x_{kj} - \bar{x}_j) = (x_{kj} - \bar{x}_{kj}) + (\bar{x}_{kj} - \bar{x}_j)$$

entonces

$$Cov(x_i, x_j) = \frac{1}{n}\sum_{k=1}^{q}\sum_{l\epsilon\prod_k}(x_{li} - \bar{x}_{ki})(x_{lj} - \bar{x}_{kj}) + \sum_{k=1}^{q}\frac{n_k}{n}(\bar{x}_{ki} - \bar{x}_i)(\bar{x}_{kj} - \bar{x}_j)$$

El primer sumando corresponde a la covarianza que hay dentro de los grupos $W(i,j)$ y el segundo a la covarianza entre los grupos $B(i,j)$, entonces para cada i y j, se puede ver la covarianza como $Cov(x_i, x_j) = W(i,j) + B(i,j)$, por lo tanto en forma matricial se escribe como $Cov(X) = W + B$, entonces las funciones lineales $Cov(Y) = a'(W+B)a = a'Wa + a'Ba$. Como el objetivo es maximizar la varianza entre los grupos y minimizar la varianza dentro de los grupos, esto equivale a maximizar la razón $a'Ba/a'Wa$, usando los multiplicadores de Lagrange se llega a que el vector que se usó para cada función discriminante corresponde al vector propio asociado a los valores propios. Tomando en cuenta que a la primer función discriminante le corresponde el valor propio más grande, al segundo el siguiente más grande y así sucesivamente. Esto es para que se cumpla que la primer función discriminante contiene la mayor explicación de la variabilidad y así sucesivamente, como estos valores propios son linealmente independientes se asegura que las funciones discriminantes no son correlacionadas.

Para medir la proporción que explicada por cada función se considera $\frac{\lambda_i}{\sum_{j=1}^{m}\lambda_j}$, en si $\sum_{j=1}^{m}\lambda_j$ es la variabilidad total que se explica con esas m funciones discriminantes.

Encontradas estas funciones, lo siguiente es hallar una regla o forma de discriminar las nuevas observaciones para poder decir a que grupo pertenecen con base en las mediciones de las p variables. Cuando se tiene una nueva observación x, se colocará en el grupo j si

$$|a'x - a'\bar{x}_j| < |a'x - a'\bar{x}_i|$$
$$\text{para toda } i \neq j$$

4.2. Regla Discriminante Basada en Distancia

Esta regla está basada en la distancia de Mahalanobis. Se calcula $D_i(x)$, que es la distancia entre la observación x y la media del grupo $\prod_i$ para toda $i = 1, ..., q$.

$$D_i = \sqrt{(x - \bar{x}_i)'\Sigma_i^{-1}(x - \bar{x}_i)} \tag{4.1}$$

Con esto se nota qué tan "cerca" está la observación x de la media del grupo en cuestión, que sería el punto centroide del grupo. Considerando la distancia que hay entre

la observación y cada uno de los grupos, se asigna esta observación al grupo $\prod_j$ si $D_j(x)$ es la menor para todos los casos. Es decir, se asigna a x en $\prod_j$si $D_j(x) < D_i(x)$, para toda $i \neq j$.

El cálculo de estas distancias es sencillo, pero no hay muchas herramientas estadísticas para saber que tan bien se está asignando o con qué probabilidad de error.

4.3. Regla Discriminante de Máxima Probabilidad

Cuando se conoce la distribución exacta de la muestra para cada uno de los grupos $\prod_1, ..., \prod_q$, se escribe la función de densidad de la población j como $f_j(x)$. La regla discriminante de probabilidad máxima para colocar o asignar a una nueva observación x en uno de los q grupos, es en aquel grupo cuya evaluación de la observación en la función de densidad del grupo sea la máxima. Es decir, x es asignada a $\prod_j$ si

$$f_j(x) = \underset{i}{máx}\ f_i(x) \qquad i = 1, ..., q \tag{4.2}$$

si varios grupos obtienen el valor máximo, se asigna la observación x en cualquiera de estos. En la Figura 4.1, se tienen dos grupos, cada uno con su respectiva función de densidad. Bajo esta regla de discriminación, asignaríamos la observación x al grupo uno, ya que al ser evaluado en ambas densidades, la mayor se obtiene en este grupo.

En ocasiones no se tiene la distribución exacta de cada uno de los grupos, se puede conocer sólo la distribución, pero no se tienen conocimientos sobre los parámetros que maneja cada distribución, si interesa encontrar una regla que discrimine a las nuevas observaciones, se puede usar aproximaciones a estos parámetros con base en la matriz de datos $X_{n\times p}$.

Para resolver este problema se hace uso de los métodos de la estadística paramétrica, para obtener una estimación puntual o quizás el intervalo en donde se encuentran los parámetros de la base de datos. Por ejemplo si se sabe que nuestra la se distribuye normal, se puede aproximar la media μ_i de cada grupo por $\bar{x}_i = \frac{1}{n_i}\sum_{j\epsilon\prod_i} x_j$ y la matriz de covarianzas Σ_i de cada grupo por $S_i = \frac{1}{n_i}\sum_{j\epsilon\prod_i}(x_j - \bar{x}_i)(x_j - \bar{x}_i)'$. Los cuales corresponden a los estimadores máximos verosímiles.

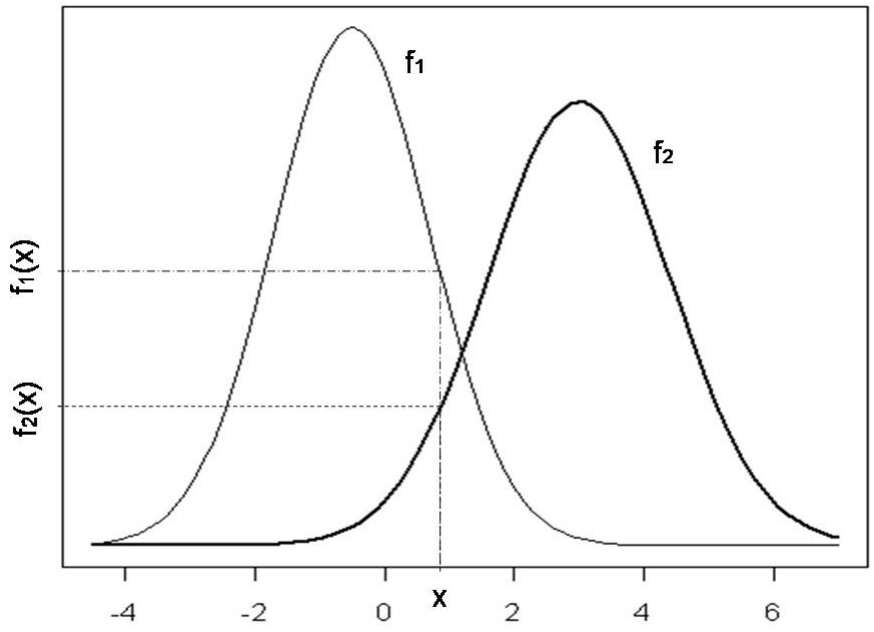

Figura 4.1: Funciones de densidad de dos grupos de una misma muestra

4.4. Regla Discriminante de Bayes

Con base en experiencia o investigaciones previas de la muestra se puede saber la probabilidad apriori de la población, esta información puede ser usada para calcular una nueva forma de discriminar.

Si para cada $\Pi_1, ..., \Pi_q$ se tienen las probabilidades apriori $\pi_1, ..., \pi_q$ de pertenecer a cada grupo. Además cada grupo sigue una distribución $f_j(\mathbf{x}, \boldsymbol{\theta}_j) \quad j = 1, .., q$. Siguiendo la teoría bayesiana, la probabilidad aposteriori de que la observación $\mathbf{x}$ pertenezca al grupo Π_j se expresa como:

$$\begin{aligned} \mathbb{P}(\mathbf{x} \in \Pi_j) &= \tau_j(\mathbf{x}, \boldsymbol{\theta}) \\ &= \frac{\pi_j f_j(\mathbf{x}, \boldsymbol{\theta}_j)}{\sum_{k=1}^{q} \pi_k f_k(\mathbf{x}, \boldsymbol{\theta}_k)} \end{aligned} \tag{4.3}$$

Siguiendo este esquema, se asigna la observación $\mathbf{x}$ al grupo Π_j, si

$$\tau_j(\mathbf{x}, \boldsymbol{\theta}) = \underset{i}{máx} \ \tau_i(\mathbf{x}, \boldsymbol{\theta}) \tag{4.4}$$

También se puede trabajar en términos del logaritmo de la razón de probabilidades. Para esto se tomará arbitrariamente el grupo q.

$$\begin{aligned}
\eta_{jq}(\mathbf{x},\boldsymbol{\theta}) &= log[\frac{\tau_j(\mathbf{x},\boldsymbol{\theta})}{\tau_q(\mathbf{x},\boldsymbol{\theta})}] \\
&= log[\frac{\pi_j f_j(\mathbf{x},\boldsymbol{\theta}_j)}{\sum_{k=1}^{q}\pi_k f_k(\mathbf{x},\boldsymbol{\theta}_k)} \bullet \frac{\sum_{k=1}^{q}\pi_k f_k(\mathbf{x},\boldsymbol{\theta}_k)}{\pi_q f_q(\mathbf{x},\boldsymbol{\theta}_q)}] \\
&= log[\frac{\pi_j f_j(\mathbf{x},\boldsymbol{\theta}_j)}{\pi_q f_q(\mathbf{x},\boldsymbol{\theta}_q)}] \\
&= log[\frac{\pi_j}{\pi_q}] + log[\frac{f_j(\mathbf{x},\boldsymbol{\theta}_j)}{f_q(\mathbf{x},\boldsymbol{\theta}_q)}]
\end{aligned} \tag{4.5}$$

En este sentido, se asigna la observación $\mathbf{x}$ al grupo q si $\eta_{jq}(\mathbf{x},\boldsymbol{\theta}) \leq 0 \quad \forall j = 1,..,q-1$. Esto significa que la razón de probabilidades es un número menor a uno, por lo tanto $\tau_q(\mathbf{x},\boldsymbol{\theta})$ es la mayor de dichas probabilidades.

4.5. Función de Riesgo

Se define $\phi_j(x)$ como la probabilidad de asignar la observación x en el grupo $\prod_j$ para $j = 1,...,q$, usando cualquiera de las reglas anteriores para discriminar, se puede definir la probabilidad de asignar una observación x al grupo $\prod_i$ cuando realmente pertenece al grupo $\prod_j$ como p_{ij}. Con base en la función ϕ_i y a la función de densidad del grupo j, dicha probabilidad queda expresada de $p_{ij} = \int \phi_i f_j(x)dx$. De donde se puede deducir que $1 - p_{ii}$ es la probabilidad de una mala clasificación de la observación x.

El asignar de forma incorrecta a un individuo en algún grupo, puede causar problemas con un grado o costo dependiendo al contexto del problema, como por ejemplo no es lo mismo si se está clasificando entre clientes que aceptan el producto y quienes no lo aceptan o pacientes que soportarán el tratamiento o no, puede traer mayores consecuencias el segundo caso ya que se puede causar la muerte del paciente. Visto desde un punto de vista de la teoría de decisión se puede expresar la función de pérdida la cual representa el costo o pérdida que se genera cuando se asigna una observación en el grupo $\prod_i$ cuando realmente pertenece al grupo $\prod_j$, se supone además que $c_{ij} > 0$ para todo $i \neq j$.

$$k(i,j) = \begin{cases} 0 & i = j \\ c_{ij} & i \neq j \end{cases}$$

Se puede entonces hablar de una función de riesgo definida por

$$\begin{aligned} R(d,i) &= E(K(d(x),j) \mid \prod_j) \\ &= \sum_{i=1}^{q} K(i,j) \int \phi_i(x) L_i(x) dx \end{aligned}$$

donde $d(x)$ es la función discriminante, es decir, dependiendo de que métodos estemos usando esta función es la que asigna la nueva observación a un grupo.

4.6. Mezclas Finitas para el Análisis Discriminante

Cuando se usa el modelo de Mezclas Finitas para encontrar una manera de discriminar los datos, se está suponiendo que cada clase tiene una distribución asociada a una componente. Es decir, se fija el valor de número de componentes g y el problema es encontrar tanto los parámetros de cada componente como los pesos para tener una mejor idea de las proporciones.

De igual manera que en los modelos anteriores, se escoge primero una muestra de los datos, a esta se le llamará muestra de entrenamiento. El resto de la muestra total, a la que se llamará muestra de prueba, servirá para verificar los resultados.

Una vez ya encontrada la distribución con base en este modelo, se tiene la aproximación de la distribución de cada grupo y además se puede ver a π_i como la probabilidad apriori de pertenecer al grupo Π_i. Esto es muy útil para el caso de que se quisiera hacer análisis de cada uno de los grupos.

Entonces, si se quiere encontrar una manera de discriminar nuevos datos, se usa la distribución de cada componente y se sigue la Regla Discriminante de Máxima Probabilidad o la Regla Discriminante de Bayes. Si se analiza un poco el algoritmo E-M, que sirve para obtener los parámetros de la mezcla finita de funciones. Se nota que al obtener el valor esperado de las z_{ij} (Ecuación 1.16), se utiliza de alguna forma la Regla Discriminante de Bayes.

Bajo el supuesto que $f(\mathbf{x}) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{x})$, se asigna una observación $\mathbf{y}$ al grupo Π_j usando la Regla Discriminante de Máxima Probabilidad si $f_j(y) = \max_i f_i(\mathbf{y})$, donde cada f_i con $i = 1, ..., g$, es la distribución de cada componente. Si se usara la Regla Discriminante de Bayes, se asigna esta observación $\mathbf{y}$ al grupo Π_j cuando $\pi_i f_i(\mathbf{y}; \theta_i) / \sum_{k=1}^{g} \pi_k f_k(\mathbf{y}; \theta_k)$ sea el valor maximo, para todos los grupos.

Si se hiciera un análisis general de la muestra sin suponer que se conoce el número de clases g, quizás el resultado sería un mejor modelo con más o menos componentes. La razón de esto pudiera ser que cuando la muestra realmente tenía un número h de clases, pero para quien hizo el análisis (quien elaboró la clasificación) este número de clases era grande o no se apegaba al objetivo del análisis, por lo tanto se decidió agrupar en g número

de grupos. Otra razón podría deberse a que este modelo supone cierta distribución en cada componente, como por ejemplo la normalidad, si uno de los grupos no cumple con este supuesto, este modelo puede partir al grupo de tal manera que por separado se vean como componentes normales. En general puede ser que no sean normales las observaciones o que su separación no sea fácilmente detectable.

Ejemplo 4.1. El siguiente ejemplo corresponde a la base de datos del artículo citado en la Referencia [20]. Se hacen tres pruebas a 145 adultos no obesos, estas pruebas ayudan al diagnóstico de pacientes con diabetes. Las pruebas son, intolerancia a la glucosa, esta prueba mide el nivel de glucosa en la sangre después de un ayuno; respuesta insulínica a la prueba de tolerancia oral de glucosa y finalmente una prueba de resistencia a la insulina, concentración de equilibrio de glucosa (steady-state plasma glucose, SSPG).

Los pacientes se clasifican en tres grupos, diabéticos u overt, quienes según sus niveles de glucosa padecen diabetes; pre-diabéticos o chemical, pacientes con amplias posibilidades de padecer diabetes ya que sus niveles de glucosa son muy altas pero no los suficientes para diagnosticarse diabéticos; y los normales que son los pacientes que no padecen diabetes ni pre-diabetes.

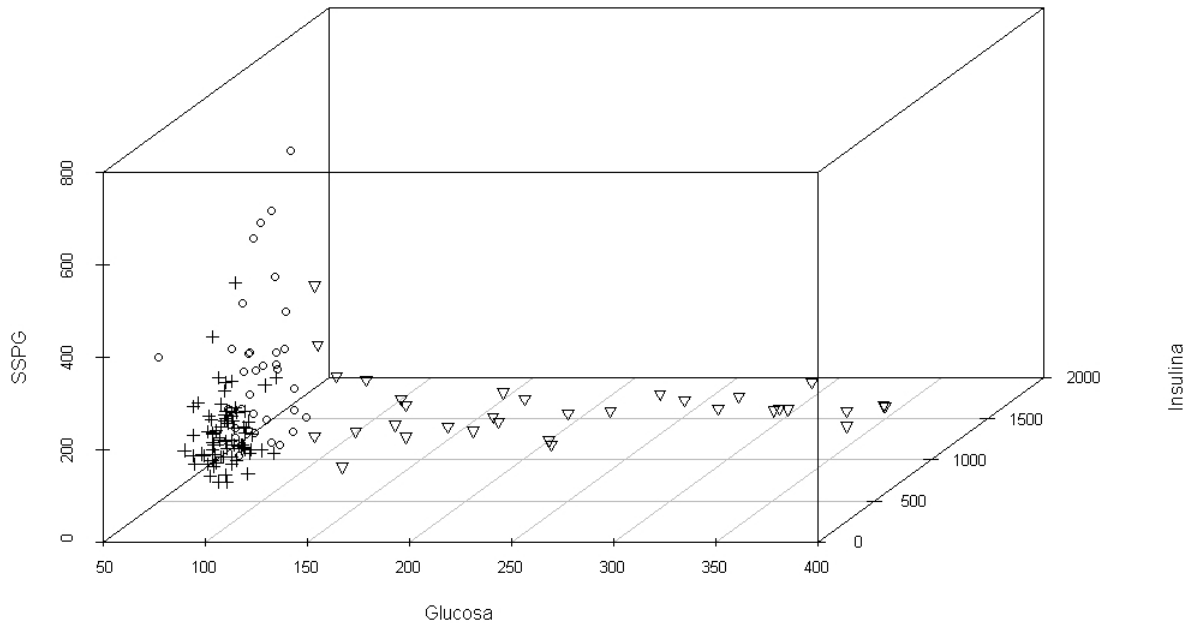

Figura 4.2: Gráfica de dispersión de los datos clasificados del Ejemplo 4.1.
$+$ Normales; $\circ$ Prediabéticos; ∇ Diabéticos

Las observaciones se pueden ver en la Gráfica 4.2. Se nota una clara diferencia entre los pre-diabéticos y los diabéticos. Aunque se aprecia bien el grupo de los normales, sobre todo porque se encuentran muy juntas las observaciones, este grupo está muy cercano al de los pre-diabéticos, quizás en esta zona será más difícil discriminar.

En este caso ya se conoce el número de grupos, por lo tanto, el modelo de mezclas será de tres componentes. Las variables a estimar son los parámetros de cada componente. Una vez que se calculan estos valores estimados, se debe verificar que este modelo ayude a discriminar correctamente los datos, bajo alguna de las reglas de discriminación antes mencionadas.

Antes de empezar a usar este método, se eligen las observaciones que ayudarán a estimar los parámetros y las que servirán para analizar que tan bien está discriminando. Para esto se generó una muestra aleatoria bidimensional con 145 observaciones, de una distribución multinomial con parámetros n=1 y pr=(0.8,0.2). Con esto se asegura que cada observación será un vector donde una de las entradas será cero y la otra forzosamente tendrá que ser uno.

Manteniendo el orden original de las observaciones de la base de datos, se asocia la *i-ésima* observación de la base con la *i-ésima* observación que se generó anteriormente. Entonces las observaciones que se utilizarán para estimar serán las que tengan el valor uno en la primer entrada y las que servirán para predecir serán aquellas que tengan el valor cero en la segunda entrada. Los resultados se muestran en el Cuadro 4.1

	Completa	Entrenamiento	Prueba
Normales	76	59	17
Prediabéticos	36	26	10
Deabéticos	33	25	8
Total	145	110	35

Cuadro 4.1: Distribución de las observaciones para entrenamiento y para prueba.

Se aplica el modelo de mezclas finitas a las observaciones de entrenamiento con sólo tres componentes como se había mencionado antes. Es por esta razón que al analizar el BIC para elegir el modelo correcto sólo se analizará aquellos modelos donde sólo hay tres componentes. En el Cuadro 4.2 están los mejores modelos, como se puede observar el mejor modelo es el que da una forma, tamaño y orientación distinta para cada componente. Si se analiza un poco la Figura 4.2, este resultado es coherente, ya que la variabilidad del grupo de los diabéticos se ve mayor a lo del grupo de los normales, además la inclinación de la curva de nivel de la distribución, aparenta estar en diferente inclinación para el grupo de los prediabéticos y los diabéticos.

Modelo	VVI	EEE	EEV	VEV	VVV
BIC	-3689.741	-3723.836	-3690.131	-3660.007	-3624.417

Cuadro 4.2: Criterio de Información Bayesiana para el Ejemplo 4.1

Suponiendo que la densidad de la muestra sea como en la Ecuación 4.6. Donde $\Sigma_i = \lambda_i D_i A_i D_i'$, los parámetros estimados para cada componente son los que se muestran en el Cuadro 4.3.

$$f(\boldsymbol{x}) = \sum_{i=1}^{3} \pi_i \phi_i(\boldsymbol{x}; \boldsymbol{\mu}_i, \Sigma_i) \tag{4.6}$$

Las Figuras 4.3, 4.4 y 4.5, muestran las observaciones con el ajuste que se hace de cada una de las componentes. Las observaciones más obscuras corresponden a aquellas en las que hubo una mayor incertidumbre al momento de asignarlas a un grupo determinado. El hecho que en estos puntos hay una mayor incertidumbre en su clasificación, no significa que sean forzosamente datos atípicos para algún grupo, más que nada son observaciones que se encuentran en un punto intermedio entre los grupos. De hecho podrán existir datos que sean atípicos para un grupo, pero al momento de ser clasificado en un grupo la incertidumbre es muy pequeña.

Se observa que en las tres figuras, donde hay más problemas para discriminar es en el área donde se encuentran acumulados los pacientes normales y los prediabéticos. Esto es debido a que se encuentran muy juntas como se había visto en la Figura 4.2. Es importante analizar las tres en paralelo, ya que son perspectivas diferentes. Se observa que en las Figuras 4.3 y 4.5 hay muchas observaciones obscuras cercanas al grupo de los normales, pero en la Figura 4.4 no hay este tipo de observaciones. Se puede apreciar una observación obscura en particular, en las Figuras 4.3 y 4.4 se ve muy cercana al grupo de los diabéticos. Si se analizan sólo estas dos gráficas, no habría razón por lo cual

Grupo	Peso	$\boldsymbol{\mu}$	Σ
Normales	0.5581325	$\begin{pmatrix} 90.81658 \\ 355.72530 \\ 168.99570 \end{pmatrix}$	$\begin{pmatrix} 64.95955 & 106.9386 & 80.43817 \\ 106.9386 & 2127.2597 & 626.50990 \\ 80.43817 & 626.50990 & 2587.73788 \end{pmatrix}$
Prediabéticos	0.2246611	$\begin{pmatrix} 104.5570 \\ 509.1481 \\ 293.9225 \end{pmatrix}$	$\begin{pmatrix} 171.5692 & 976.622 & -473.9838 \\ 976.622 & 7759.947 & -1793.9545 \\ -473.9838 & -1793.9545 & 24062.6761 \end{pmatrix}$
Diabéticos	0.2172064	$\begin{pmatrix} 214.6457 \\ 1036.2562 \\ 101.7866 \end{pmatrix}$	$\begin{pmatrix} 6098.337 & 27435.38 & -5150.268 \\ 27435.38 & 136324.04 & -27466.923 \\ -5150.268 & -27466.923 & 7133.534 \end{pmatrix}$

Cuadro 4.3: Parámetros de cada componente del Ejemplo 4.1

hubiera una mayor incertidumbre ya que a vista se asignaría al grupo mas cerano, pero al observar la Figura 4.5 se nota que desde esa perspectiva esta observación si se encuentra un poco más cercana al grupo de los prediabéticos. Recuerde que este modelo asigna cada observación a un grupo con base en la Regla Discriminante de Bayes y no a la distancia como tal. En general no hay muchas observaciones con gran incertidumbre, de hecho para el grupo de los diabéticos no hay tanto problema para clasificar.

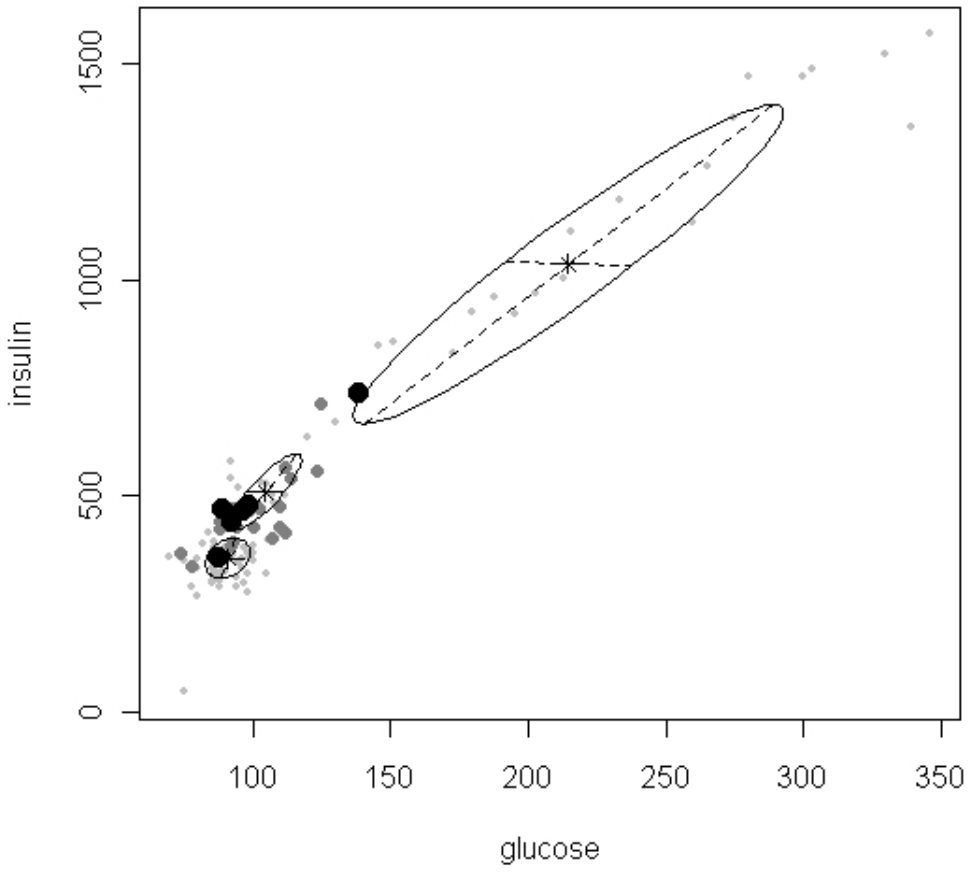

Figura 4.3: Glucosa - Insulina. Gráfico del ajuste de los datos al modelo de mezclas finitas

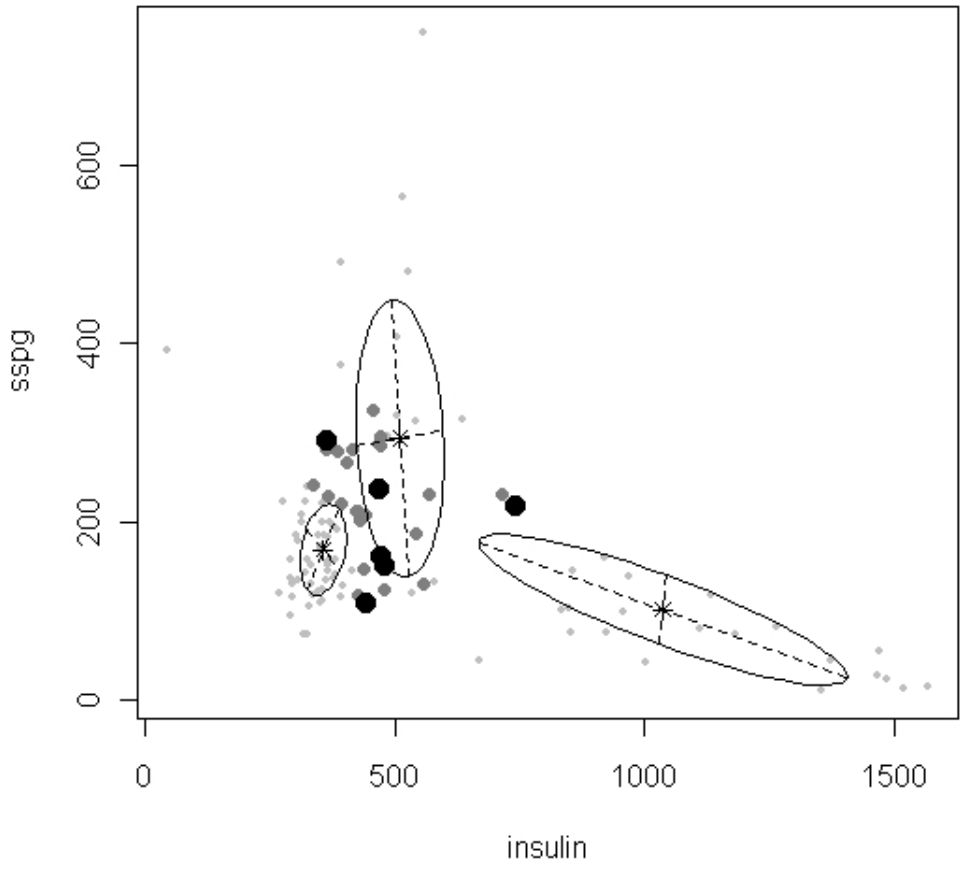

Figura 4.4: Insulina - SSPG. Gráfico del ajuste de los datos al modelo de mezclas finitas

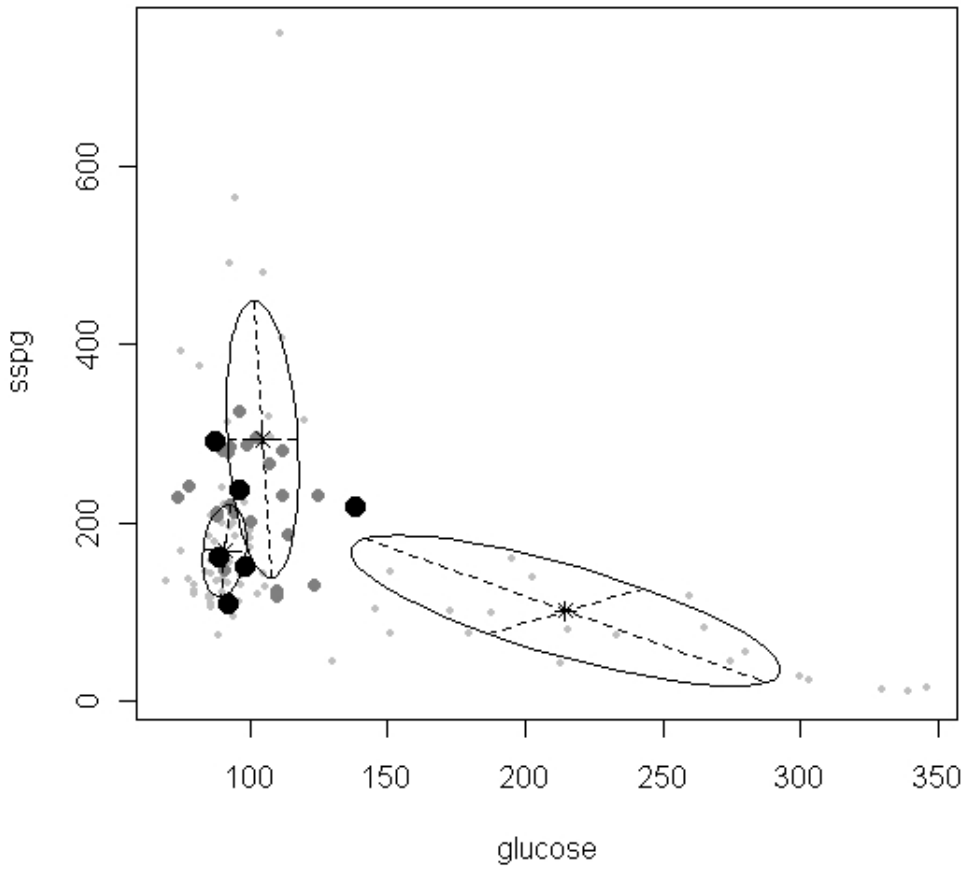

Figura 4.5: SSPG - Glucosa. Gráfico del ajuste de los datos al modelo de mezclas finitas

Obs.	Glucosa	Insulina	SSPG	Grupo Original	Grupo Asignado	Incertidumbre
62	88	439	208	2	1	0.1694
63	100	429	201	2	1	0.1387
65	89	472	162	2	1	0.3655
66	91	436	148	2	1	0.1365
69	82	390	375	1	2	0.0026
77	94	426	213	2	1	0.1355
82	93	393	490	1	2	0.0000
104	75	45	392	2	3	0.0000
105	92	442	109	2	1	0.2992
107	92	580	132	2	3	0.0480
110	88	423	212	2	1	0.1163
115	125	714	232	3	2	0.1633
134	123	557	130	3	2	0.1052
135	130	670	44	3	2	0.0073
136	120	636	314	3	2	0.0002

Cuadro 4.4: Observaciones mal clasificadas de las observaciones prueba

Con base en la muestra que se tomó para obtener los parámetros estimados de cada componente, se hace un análisis de aquellas observaciones bien clasificadas y de las que no lo fueron.

De las 110 observaciones tomadas, 15 fueron mal clasificadas, es decir, el 13.63 % de estas observaciones. Estos datos se muestran en el Cuadro 4.4. Existe un mayor error en los casos de los grupos de Normales y Prediabéticos, en ambos sentidos, esto se debe a la cercanía entre ellos. Si se observa un poco la columna de la incertidumbre, se nota que estos valores son pequeños, lo que hace pensar que el modelo tuvo una mala clasificación debido a que estas observaciones están muy alejadas de la media de su grupo.

Pero el análisis anterior es con base en las observaciones que sirvieron para determinar los parámetros del modelo, es decir, la muestra de entrenamiento. Para verificar que tan bueno es el modelo, se usarán los datos que no se utilizarón anteriormente, la muestra de prueba. Para esto se usará la Regla Discriminante de Máxima Probabilidad y la Regla Discriminante de Bayes. Haciendo el análisis, ambas reglas arrojan el mismo resultado, este se ve en el Cuadro 4.5. En este caso sólo 2 observaciones de las 35 fueron mal clasificadas, lo que representa un 5.71428 % . Si se analizan estas observaciones se nota que estas son cercanas a la media del grupo donde el modelo las asignó. Por ejemplo la observación 131, la cual pertenece al grupo de los diabéticos y fue asignada al grupo de los prediabéticos, el valor en la variable que mide la glucosa es pequeña comparada con los otros individuos de este grupo, y además la variable que mide la insulina es muy grande, estos valores la alejan de la media de su grupo. Esto justifica un poco el por qué fueron mal clasificadas.

Obs.	Glucosa	Insulina	SSPG	Grupo Original	Grupo Asignado
83	85	425	143	2	1
131	124	538	460	3	2

Cuadro 4.5: Observaciones mal clasificadas de las observaciones de entrenamiento

En general, este modelo si esta clasificando bien los datos, tomando en cuenta que para el grupo de los normales y los prediabéticos siempre habrá problemas.

Ejemplo 4.2. La hemofilia es un trastorno de la sangre hereditario en el que la sangre no se puede coagular normalmente en el lugar donde hay una herida o lesión. Está causada por un gen anormal en el cromosoma X. Si una madre es portadora del gen anormal en uno de sus cromosomas, ella no tendrá hemofilia, pero será portadora del trastorno. Eso quiere decir que puede pasar el gen de la hemofilia a sus hijos. Hay un 50 % de probabilidades de que cualquiera de sus hijos pueda heredar dicho gen y nacerá con hemofilia. También hay un 50 % de probabilidades de que cualquiera de sus hijas sea portadora del gen, sin que tengan hemofilia ellas. Es muy raro para una niña el nacer con hemofilia, pero puede pasar si el padre tiene hemofilia y la madre es portadora del gen de la hemofilia. Entonces su hija tendrá el gen anormal en sus dos cromosomas X. La hemofilia tipo A, es causada por una deficiencia del factor VIII, una de las proteínas que ayuda a la sangre a formar los coágulos.

En 1974 Habbema, Hermans y van den Broken [13], hicieron un estudio sobre este padecimiento genético. Se realizaron estudios a 75 mujeres, 30 de ellas eran mujeres sanas, es decir, no padecían ni eran portadoras de dicho trastorno. El resto (45) son mujeres portadoras de la hemofilia tipo A. Las variables que se midieron fueron, $y_1 = log_{10}(AHFactivity)$ y $y_2 = log_{10}(AHFlikeantigen)$. Lo importante en este problema es poder tener una manera de discriminar, en base a estos datos, entre mujeres portadoras y mujeres sanas.

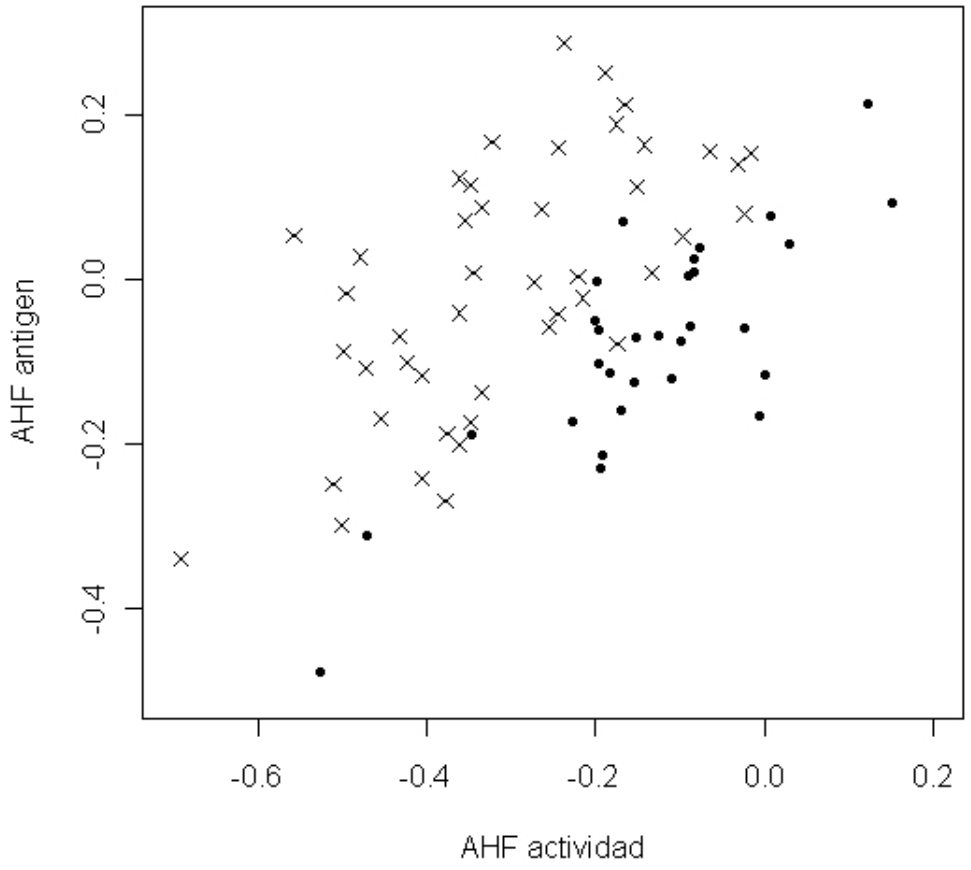

Figura 4.6: Gráfica de dispersión de las observaciones del Ejemplo 4.2. • - Mujeres sanas; × - Mujeres portadoras

Grupo	Completa	Entrenamiento	Prueba
Sanas	35	25	5
Portadoras	45	39	6
Total	75	64	11

Cuadro 4.6: Distribución de las observaciones muestra y prueba.

La Figura 4.6, muestra estos datos. Si se analiza un poco esta figura, se observa que no existe una separación muy grande entre los grupos. De hecho si en la gráfica no hubiera distinción entre los grupos, a simple vista no sería fácil ver la existencia de varios grupos. En el grupo de las mujeres sanas se ven algunas observaciones un tanto alejada de las demás, estas se encuentran en la esquina inferior izquierda. Recordando que esta base cuenta con sólo 75 datos, se puede pensar que las observaciones antes mencionadas son datos atípicos, pero no se puede asegurar esto ya que estas son un porcentaje no muy pequeño. Estas situaciones tendrán un efecto para el cálculo de los parámetros del modelo.

Con base en los datos anteriores, ase ajustará un modelo de mezclas finitas a estos datos. El número de componentes será dos, correspondientes a cada grupo (Ecuación 4.7). Las observaciones que se utilizarán para ajustar los datos, fueron escogidas de manera aleatoria.

De nuevo se generó una muestra de 75 observaciones bidimensionales, de una distribución multinomial con parámetros n=1 y pr=(0.8,0.2), al igual que en el ejemplo anterior. Las observaciones que servirán para la muestra de entrenamiento serán aquellas que tengan el valor de uno en la primer entrada y las observaciones que servirán para probar la eficacia del modelo, serán las que tengan el valor de cero en la segunda entrada. El resultado de esta simulación se muestra en el Cuadro 4.6. La densidad de la muestra es:

$$f(\boldsymbol{x}) = \sum_{i=1}^{2} \pi_i \phi_i(\boldsymbol{x}; \boldsymbol{\mu}_i, \Sigma_i) \tag{4.7}$$

Una vez establecidas las observaciones, se realiza el ajuste. En este caso se fija el número de componentes, pero no hay restricción sobre el supuesto de la matriz de covariazas. El Cuadro 4.7 muestra los cinco principales valores del BIC, de esto se puede deducir que el mejor modelo es el "VVI". En la Figura 4.7 se ve el ajuste, si la se compara con la Figura 4.6 este ajuste está muy lejos de apegarse a los datos. Sin embargo, si se elige el supuesto en donde cada componente tiene una matriz de covarianzas diferente, los datos son mejor ajustados como se ve en la Figura 4.8.

Modelo	EEI	VEI	EVI	VVI	VVV
BIC	98.80192	95.19756	95.81218	91.64923	94.84358

Cuadro 4.7: Criterio de Información Bayesiana para el Ejemplo 4.2

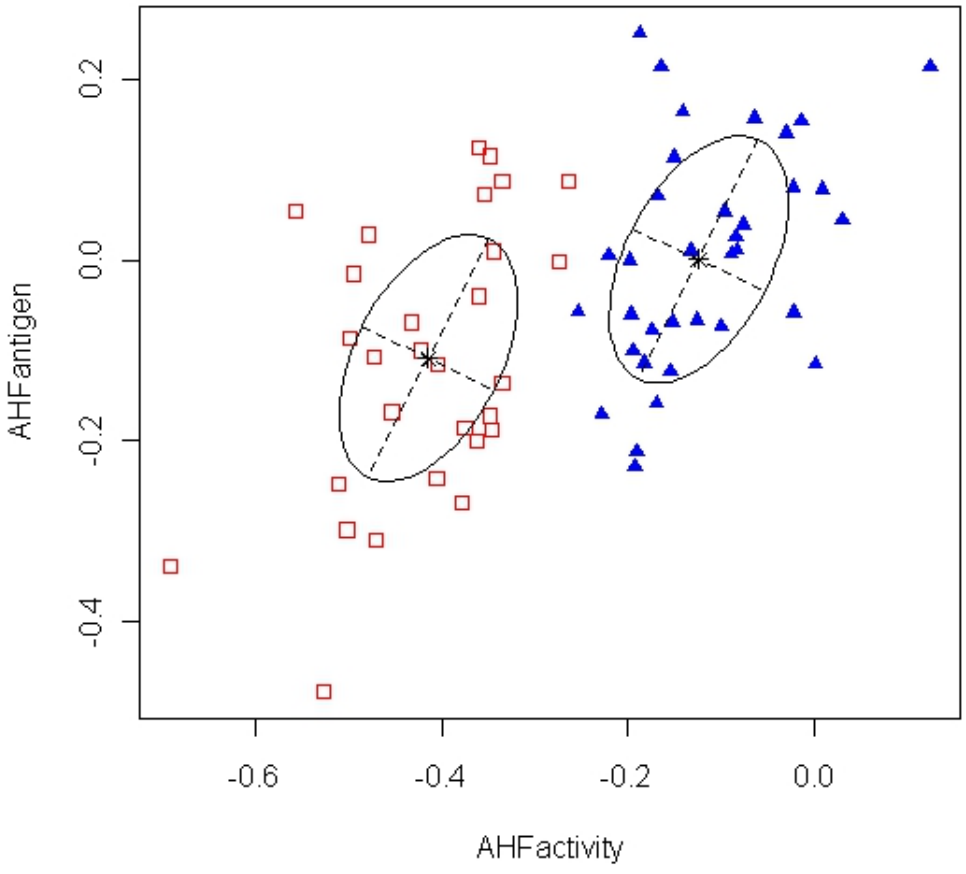

Figura 4.7: Clasificación de las observaciones bajo el modelo VVI

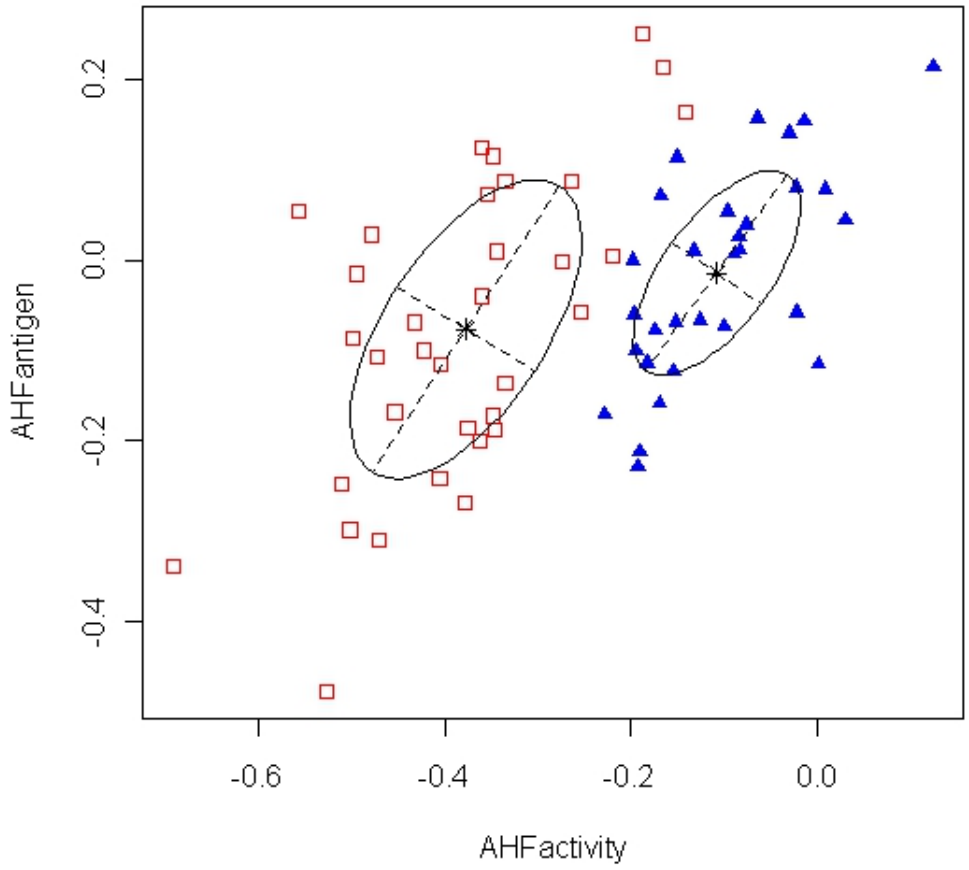

Figura 4.8: Clasificación de las observaciones bajo el modelo VVV

Bajo el modelo en donde la matriz de varianza y covarianza es totalmente diferente para cada componente, se obtienen los parámetros estimados del Cuadro 4.8.

Grupo	Peso	$\boldsymbol{\mu}$	Σ
Sanas	0.4608466	$\begin{pmatrix} -0.10824927 \\ -0.01426894 \end{pmatrix}$	$\begin{pmatrix} 0.008055731 & 0.00656918 \\ 0.00656918 & 0.0128033 \end{pmatrix}$
Portadoras	0.5391534	$\begin{pmatrix} -0.37731829 \\ -0.072621467 \end{pmatrix}$	$\begin{pmatrix} 0.01536417 & 0.01246155 \\ 0.01246155 & 0.02741867 \end{pmatrix}$

Cuadro 4.8: Parámetros de cada componente del Ejemplo 4.2

Con base en el ajuste del modelo de mezclas finitas, el 17.1875 % de las 64 observaciones de la muestra fueron mal clasificadas. Sin embargo cuando se usa la regla discriminante de máxima probabilidad y la regla discriminante de Bayes (en el conjunto prueba), el 9.09 % de las observaciones fueron mal clasificadas. Lo que significa que sólo una de las once observaciones fue mal clasificada. Esto nos dice que el modelo ajusta bien los datos, a pesar de no tener un gran número de observaciones.

Ejemplo 4.3. Los datos de este ejemplo fueron tomados de la Referencia [19]. Las pacientes de esta base de datos fueron sometidas a una aspiración con aguja fina (PAAF), el cual es un procedimiento para detectar cáncer de mama. Se utiliza a menudo cuando un bulto sospechoso en el pecho se encuentra al tacto o si se detecta una anomalía en una prueba de imagen como la radiografía, ecografía o mastografía. El procedimiento implica colocar una aguja muy delgada dentro de la masa y la extracción de células o el líquido de un quiste o una masa sólida para una evaluación microscópica. La muestra en general consta de 569 observaciones de las cuales 357 son benignos y 212 son malignos, son 32 los atributos de la muestra, los cuales son:

1. Número de Identificación
2. Diagnóstico (B-Benigno, M-Maligno)
   (Por cada célula examinada los siguientes 10 atributos)
   - *a*) radio (media de las distancias desde el centro hacia los puntos en el perímetro)
   - *b*) textura (desviación estándar de los valores de escala de grises)
   - *c*) perímetro
   - *d*) superficie
   - *e*) suavidad (variación local en longitudes de radio)
   - *f*) compacidad (perímetro $^2 \div$ área - 1.0)
   - *g*) concavidad (la gravedad de las porciones cóncavas del contorno)
   - *h*) puntos cóncava (número de porciones cóncavas del contorno)
   - *i*) simetría
   - *j*) dimensión fractal (" aproximación costa " - 1)

Para el caso de clasificación de la muestra en las dos clases, se encontró que se obtienen mejores resultados de discriminación usando las variables de Área, Suavidad y Textura. Es por esto que para el análisis discriminante usando el modelo de mezclas finitas se usará estas tres variables, de la tercer célula muestreada, es decir, los atributos 24, 26 y 27 del total.

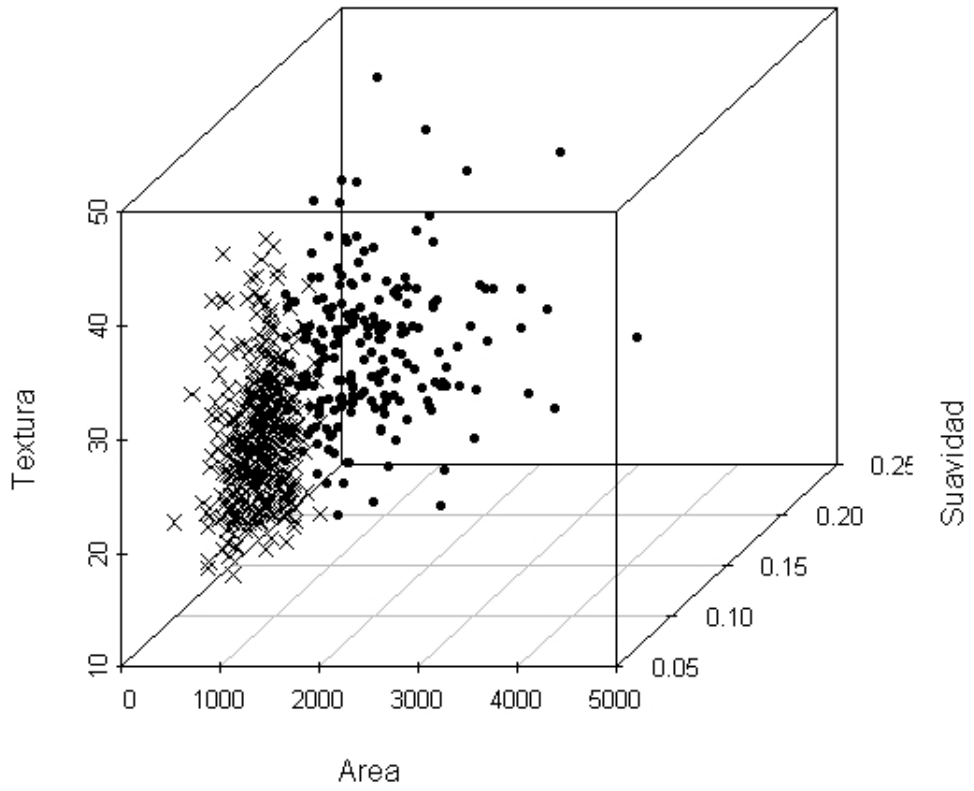

Figura 4.9: Gráfica en tres dimensiones del Ejemplo 4.3× - Benignos, • - Malignos

Dichas varibles son gráficadas y se muestran en las Figuras 4.9 y 4.10 . Existe una separación entre los grupos, sobre todo para las gráficas de textura vs areas y area vs suavidad. El hecho de que las medias de las variables Área y Textura son mas pequeñas en el caso del grupo de los Benignos, permite apreciar mejor esta separación. Otra diferencia entre los grupos que se puede apreciar en estas gráficas, es el hecho de que el grupo de malignos tiene una mayor dispersión. Esto hace que el modelo se pueda ajustar bien,

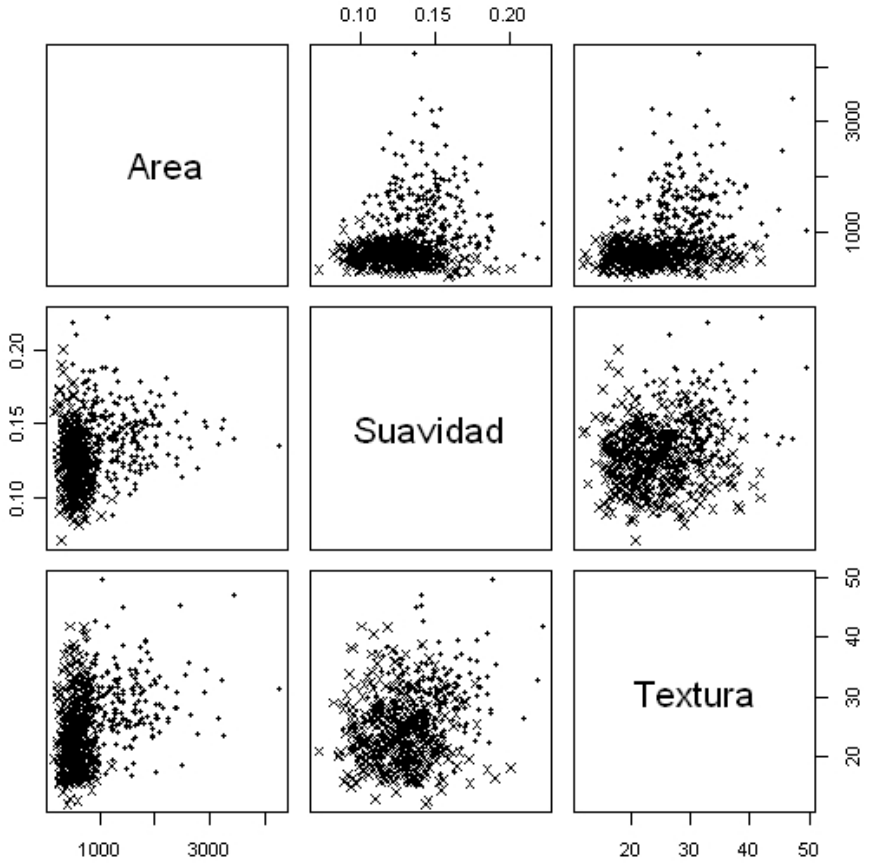

Figura 4.10: Gráficas en dos dimensiones del Ejemplo 4.3× - Benignos, • - Malignos

teniendo en cuenta que no se puede suponer una misma matriz de covarianzas. Además no se ven datos atípicos que puedan perjudicar en gran medida los resultados.

De nuevo antes de comenzar el análisis se elegieron las observaciones servirán para describir el modelo y las observaciones que se utilizarán para ver que tan bueno es este modelo, es decir la muestra de entrenamiento y la muestra de prueba. Para esto se utilizó el mismo método de generar una muestra aleatoria bidimensional con 569 observaciones, de una distribución multinomial con parámetros n=1 y pr=(0.8,0.2), como en los Ejemplos 4.1 y 4.2. El resultado de la distribución de la muestra de entrenamiento y la muestra prueba se ve en el Cuadro 4.9.

Entonces se aplicará el modelo de mezclas finitas bajo 4.8, en una muestra de 454 observaciones y se verificará que tan bien son clasificados las observaciones con una muestra de 115 observaciones.

Grupo	Completa	Entrenamiento	Prueba
Benignos	357	286	71
Malignos	112	168	44
Total	569	454	115

Cuadro 4.9: Distribución de las observaciones en las muestras prueba y entrenamiento.

Modelo	VEI	EVI	VVI	EEE	VVV
BIC	-7587.196	-7486.026	-7446.479	-7585.116	-7464.815

Cuadro 4.10: Criterio de Información Bayesiana para el Ejemplo 4.3

$$f(\boldsymbol{x}) = \sum_{i=1}^{2} \pi_i \phi_i(\boldsymbol{x}; \boldsymbol{\mu}_i, \Sigma_i) \tag{4.8}$$

Restringiendo el modelo a dos componentes, se obtienen los siguientes resultados mostrados en el Cuadro 4.10 para el Criterio de Información Bayesiana con respecto a cada uno de los supuestos en la matriz de covarianza. Se observa que el mejor modelo es cuando se supone diferencia en las matrices y no hay inclinación en las elipses de las curvas de nivel de los diferentes cortes. Es decir, el modelos donde $\Sigma_k = \lambda_k A_k$ con $k = 1, 2$.

Con base al modelo antes mencionado, los parámetros de dicho análisis se muestran él en Cuadro 4.11.

Grupo	Peso	$\boldsymbol{\mu}$	Σ
Malignos	0.3798649	$\begin{pmatrix} 1337.49 \\ 0.14453 \\ 28.846 \end{pmatrix}$	$\begin{pmatrix} 367959.5 & 0 & 0 \\ 0 & 0.0004903797 & 0 \\ 0 & 0 & 26.65496 \end{pmatrix}$
Benignos	0.6201351	$\begin{pmatrix} 575.919 \\ 0.12510 \\ 23.617 \end{pmatrix}$	$\begin{pmatrix} 29080.56 & 0 & 0 \\ 0 & 0.000385867 & 0 \\ 0 & 0 & 28.54535 \end{pmatrix}$

Cuadro 4.11: Parámetros de cada componente del Ejemplo 4.3

Haciendo un análisis más gráfico de los resultados, las Gráficas 4.11, 4.12 y 4.13 muestran la clasificacion de las observaciones bajo el modelo. Si se compara con la Gráfica 4.10, se observa que son muy parecidas, lo que nos hace pensar en una buena clasificación. Corraborando un poco lo analizado gráficamente la clasificación de cada observación en cuanto al modelo y la real. De las 454 observaciones de la muestra prueba sólo el 5.28 % fueron mal clasificadas.

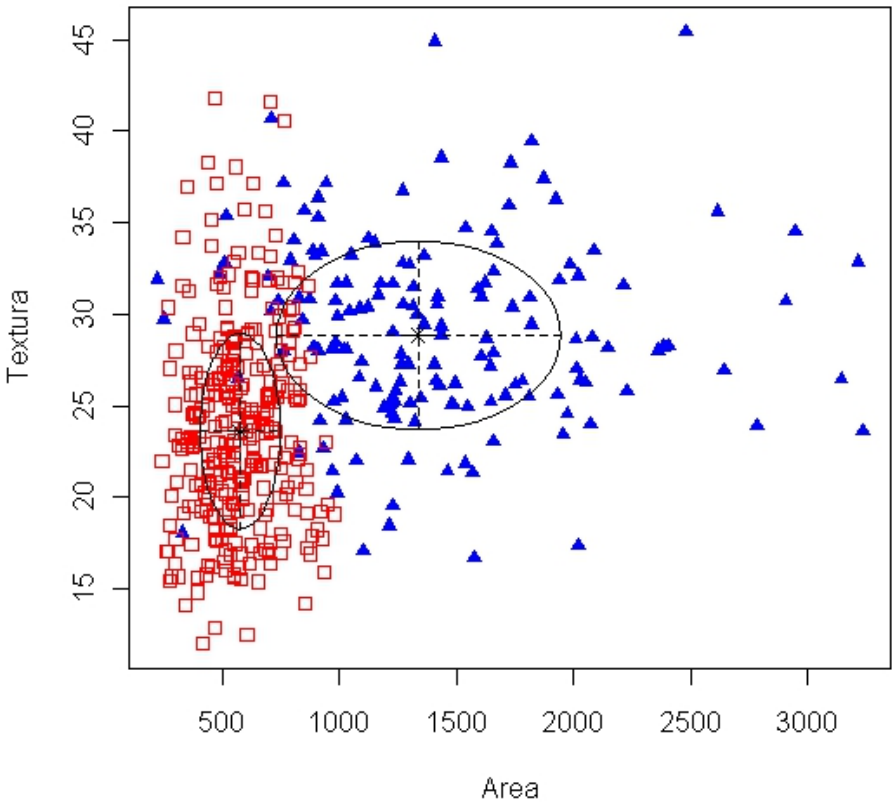

Figura 4.11: Proyección (Area vs Textura) de la clasificación en base al modelo de mezclas finitas para el Ejemplo 4.3. Cuadrados-Benignos; Triángulos-Malignos

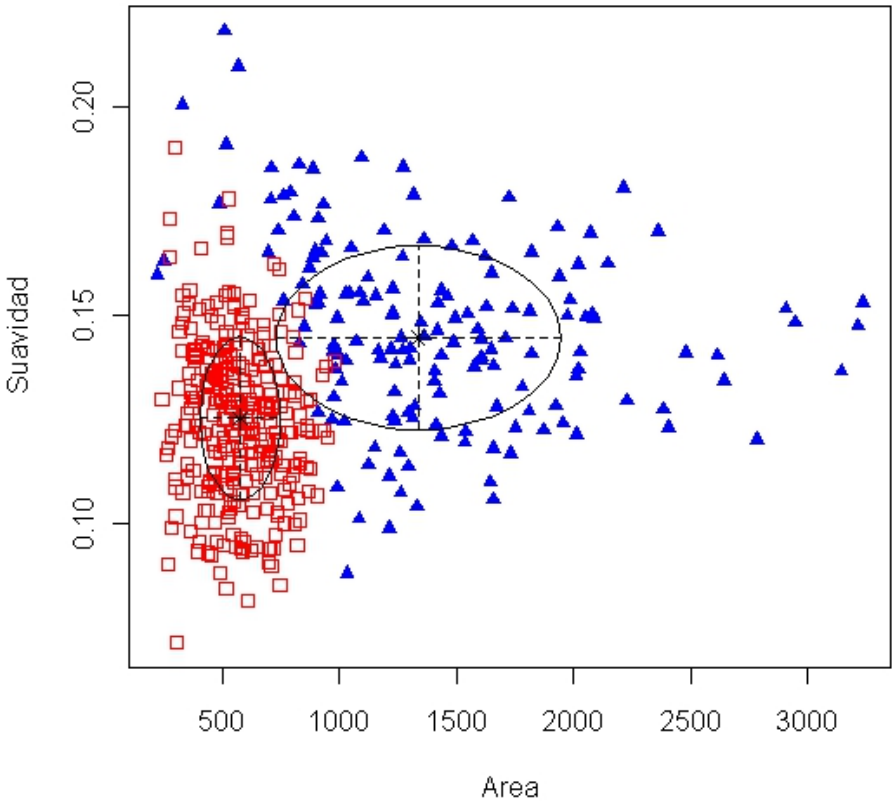

Figura 4.12: Proyección (Area vs Suavidad) de la clasificación en base al modelo de mezclas finitas para el Ejemplo 4.3 Cuadrados-Benignos; Triángulos-Malignos

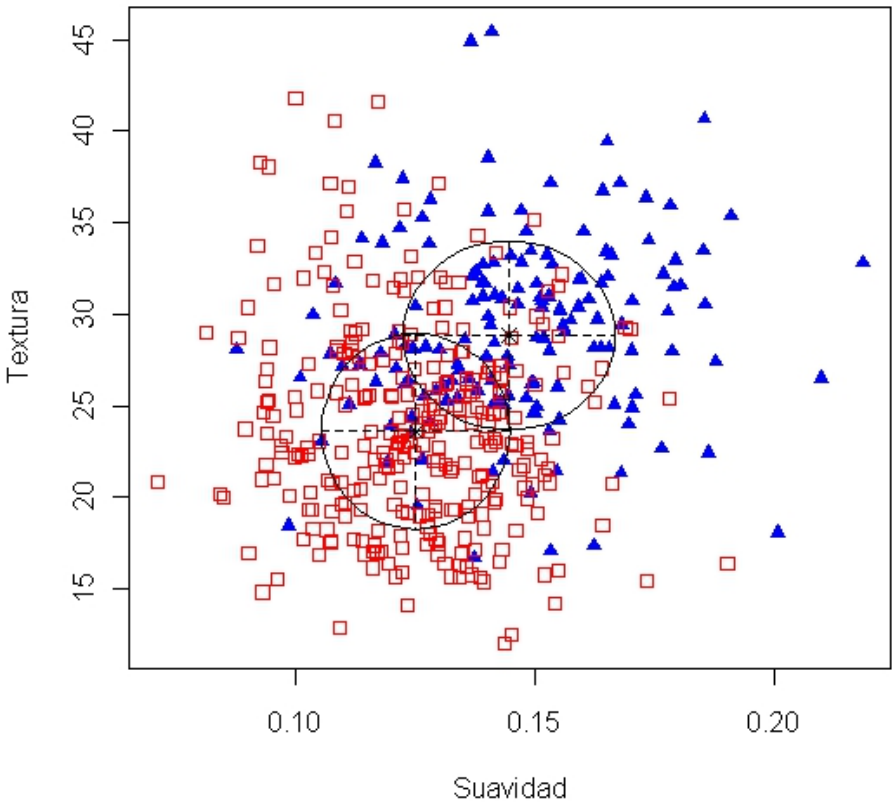

Figura 4.13: Proyección (Suavidad vs Textura) de la clasificación en base al modelo de mezclas finitas para el Ejemplo 4.3 Cuadrados-Benignos; Triángulos-Malignos

Ahora, enfocando el análisis en la muestra prueba con 115 observaciones, se usa la regla discriminante de máxima probabilidad y la regla discriminante Bayes.

Suponiendo que cada grupo sigue la distribución asociada a cada componente y usando la regla discriminante de máxima verosimilitud el 3.47 % de la muestra de entrenamiento es mal clasificada. Si se usa la regla discriminante bayes el 2.608 % de las observaciones son mal clasificadas. Estas pocas observaciones se exponen en la Grafica 4.14, se nota que son observaciones centradas, por ende no tan fácil de clasificar.

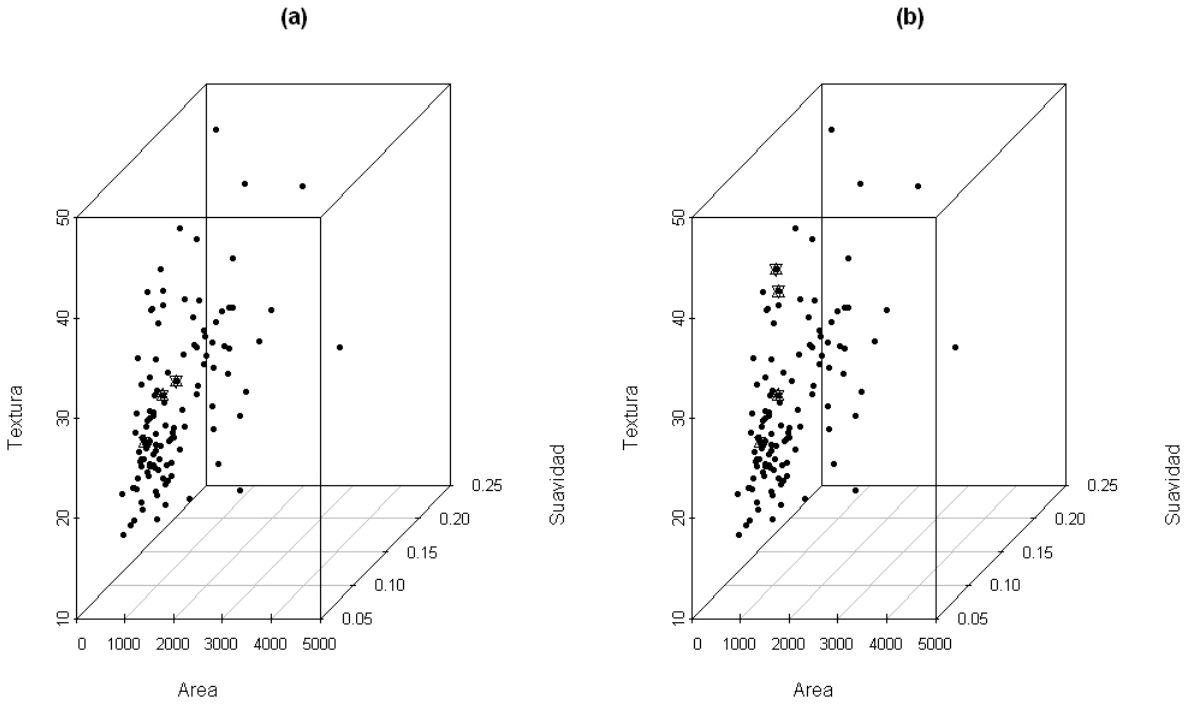

Figura 4.14: Observaciones mal clasificadas, encerradas en una estrella. (a) Regla discriminante bayes. (b) Regla discriminante máxima probabilidad

En general, usando este modelo se encuentra una buena clasificación de la muestra.

Conclusiones

Cuando se tiene una base de datos que está ya clasificada en cierto número de grupos, lo que se intenta es encontrar una manera de clasificar en algún grupo una nueva observación. Un método para encontrarla, es usando el Discriminante lineal de Fisher, pero no siempre trazando una línea recta se puede encontrar una forma de discriminar. Para estos casos existen diferentes reglas de discriminación, pero para usarlas hay que hacer supuestos sobre la distribución de los cada uno de los grupos.

Una manera en particular de encontrar esta regla, es haciendo el supuesto de que las observaciones en general siguen un modelo de mezclas finitas. Cada componente corresponde a un grupo, entonces el problema se limita a encontrar el valor de los parámetros de cada componente para inferir la distribución de cada grupo. Una vez encontrada se usa algunas de las reglas de discriminación. Tanto la regla de discrimante de máxima probabilidad como la de bayes son coherentes con la realidad.

De nuevo aquí nos se puede encontrar con casos donde el discriminante lineal sea mejor, pero esto depende mucho las observaciones, sobre todo si siguen el supuesto de normalidad.

Conclusiones Generales

El modelo de mezclas finitas es una suma ponderada de un número finito de densidades. Aunque pareciera que es un modelo exclusivo para el problema de estimación de densidad, no es así. Ya que por su construcción, se puede asociar a cada componente un grupo dentro de la población en general. Esto es, si en la base de datos hubiera comportamientos distintos entre las observaciones, de tal manera que puedan ser agrupadas las observaciones un tanto similares. Cada grupo compartiría ciertas características, de hecho se puede suponer que estas observaciones pertenecen a una misma distribución.

Considerando lo anterior, es factible usar este modelo para hacer una análisis discriminante y además para un análisis de conglomerados. En el primer caso se conoce el número total de grupo y lo importante es conocer una manera de catalogar en un grupo, alguna nueva observación. Para el caso de análisis de conglomerados, lo importante es identificar el número de grupos, quienes son y poder dar una interpretación de los mismos.

Una diferencia cuando se usa un análisis de discriminante y un análisis de conglomerados, además del conocimiento o no del número de grupos. Es la manera en que se empieza a atacar el problema, para el caso de discriminante ya no es tan necesario un análisis descriptivo, pero para el caso de conglomerados si lo es. Si no se hace antes un análisis descriptivo en conglomerados, no es fácil percatarse del número de grupos.

Ejemplos del uso de este modelo hay muchos, solo se mencionaron algunos. Cabe señalar que este modelo no siempre es el mejor, como se vió en algunos de los ejemplos descritos. Esto es porque no siempre las observaciones siguen una distribución normal. Además para que esto funcione mejor estas los grupos deben estar ligeramente separados. Sin embargo, en ocasiones esto no se da, ni a simple vista se podría asegurar que hay grupos.

Es por esto que es bueno utilizar un modelo de mezclas finitas, cuando se puedan distinguir un poco la existencia de grupos. En este trabajo sólo hay ejemplos del caso cuando suponemos normalidad, pero existen muchos artículos en la actualidad donde se están trabajando con otro tipo de distribuciones, de tal manera que se ajusten mejor.

Apéndice A

Anexo técnico

Ejemplo 1.1

```
obs=100
p=20
y=matrix(nrow=obs,ncol=p)
for (i in 1:obs)
{
for (j in 1:p)
{
y[i,j]=rnorm(1,0,1)
}
}
m=Mclust(y,modelNames=''EEE'',G=2)
mu1=as.vector(m$parameters$mean[,1])
mu2=as.vector(m$parameters$mean[,2])
sigma=m$parameters$variance$Sigma
a=solve(sigma)%*%(mu1-mu2)
ay=matrix(nrow=obs,ncol=1)
for (i in 1:obs)
{
ay[i,]=t(a)% *% y[i,]
}
hist(ay,main='' '' ,xlab='' '',ylab='' '')
```

Ejemplo 2.1

```
obs=500
mul=rmultinom(obs,size=1,prob=c(1/2,1/2))
comp=cbind(rnorm(obs,-3/2,(1/2)),rnorm(obs,3/2,(1/2)))
simu7=as.vector(100)
for ( i in 1:obs)
```

```
{
simu7[i]=comp[i,1]*mul[1,i]+comp[i,2]*mul[2,i]
}
Entonces simu7 contiene las observaciones finales.
```

Ejemplo 2.2

```
obs=500
mul=rmultinom(obs,size=1,prob=c(9/20,9/20,1/10))
comp=cbind(rnorm(obs,-6/5,(3/5)),rnorm(obs,6/5,(3/5)),rnorm(obs,0,(1/4)))
simu9=as.vector(100)
for ( i in 1:obs)
{
simu9[i]=comp[i,1]*mul[1,i]+comp[i,2]*mul[2,i] + comp[i,3]*mul[3,i]
}
Entonces simu9 contiene las observaciones finales.
```

Ejemplo 2.3

```
obs=500
m1=c(1,1)
m2=c(4,4)
sigma1=matrix(c(1,0, 0, 1),ncol=2)
sigma2=matrix(c(1,0, 0, 1),ncol=2)
mul=rmultinom(obs,size=1,prob=c(1/2,1/2))
comp1=rmvnorm(obs,m1, sigma1)
comp2=rmvnorm(obs,m2, sigma2)
simu7b=matrix(ncol=2,nrow=obs)
for ( i in 1:obs)
{
simu7b[i,1]=comp1[i,1]*mul[1,i]+comp2[i,1]*mul[2,i]
simu7b[i,2]=comp1[i,2]*mul[1,i]+comp2[i,2]*mul[2,i]
}
Entonces simu7b contiene las observaciones finales.
```

Ejemplos 4.1, 4.2 y 4.3

```
total=Total de observaciones
base=Base de datos completa
fat=rmultinom(total,size=1,prob=c(0.80,0.20))
fat=t(fat)
tamaño=sum(fat[,1])
for (i in 1:total)
{
if (fat[i,1]==1) prueba=rbind(prueba,base[i,])
```

```
else entrenamiento=rbind(entrenamiento,base[i,])
}
prueba=prueba[-1,]
entrenamiento=entrenamiento[-1,]
```

Finalmente las matrices prueba y entrenamiento continen las muestras Prueba y entrenamiento respectivamente

Bibliografía

[1] Akaike Hirotugu " A new look at the statistical model identification ", *IEEE Transactions on Automatic Control* Vol. 19 No.6, 716-723, 1974

[2] Banfield, J.D. and Raftery, A.E. " Model-based Gaussian and non-Gaussian clustering" . *Biometrics*, 49, 803-821, 1993

[3] Bozdogan, H. and Sclove, S.L. " Multi-sample cluster analysis using Akaike's information criterion " . *Ann. Inst. Statist. Math.* 36, 163-180, 1984.

[4] Butler, R.W. "Predictive likelihood inference with applications (whith discussion)" . *J. R. Statist. Soc. B*48, 1-38, 1986.

[5] Charlie, C. V. L. y Wicksell, S.D. " On the dissection of frequency functions " . *Arkiv för Mathematik Astronomi och Fysik* 18, No. 6, 1924.

[6] Day, N.E. " Estimating the components of a mixture or two normal distributions ", *Biometrika* No.56, 463-474, 1969

[7] Dempster, A.P., Laird, NM., y Rubin, D.B. " Maximum likelihood from incomplete data via the EM algorithm (whith discussion) " . *Journal of the Royal Statistical Society B* 39, 1-38, 1977.

[8] Doetsh, G. " Die elimination des dopplereffekts auf spektroskopische feinstrukturen und exakte bestimmung der komponenten " . *Zeitschrift für Physik* 49, 705-730, 1928.

[9] Fraley C, Raftery A.E . " mclust Version 3 for R: Normal Mixture Modeling and Model-based Clustering " . Technical Report 504, University of Washington, Department of Statistics, 2006

[10] Fraley C. and Raftery A. E. " Model-Based Clustering, Discriminant Analysis, and Density Estimation"*Journal of American Statistical Association* Vol. 97 No.458, 611-632

[11] Fryer, J.G. y Robertson, C.A. " A comparison of some methods for estimating mixed normal distributions " . *Biometrika* 59, 639-648, 1972.

[12] Gibbons J.D. and Chakraborty S., *Nonparametric Statistical Inference.*Marcel Dekker, 1985

[13] Habbema, J.D.F., Hermans, J., and van den Broek, K. " A stepwise discriminant analysis program using density estimation". *Compstat 1974, Proceedings Computational Statistics* Viena: Physica-Verlag, 101-110, 1974.

[14] Haughton, D. M. A. *The Annals of Statistics* " On the Choice of a Model to Fit Data From an Exponential Family " 16, 342-355,1988.

[15] Hartigan John A. *Clustering Algorithms.* John Wiley & Sons. 1975

[16] Mardia, K. V., J. T. Kent y J. M. Bibby *Multivariate Analysis.* Academic Press. 1979

[17] McLachlan G. J. and D. Peel, *Finite Mixture Models.* New York: Wiley, 2000.

[18] McLachlan G. J. and K. E. Basford, *Mixture Models: Inference and Applications to Clustering.* New York: M. Dekker, 1988.

[19] Merz J., and Murphy, P.M., UCI Repository of Rachine Learning Databases. http://www.ics.uci.edu/- learn/MLRepository.html, 1996.

[20] Reaven G.M. and Miller R.G. ," An attempt to define the nature of chemical diabetes using a multidimensional analysis " *Diabetologica* 16:17-24, 1979.

[21] Schwarz Gideon E . " Estimating the dimension of a model ", *Annals of Statistics* Vol. 6 No.2, 461-464, 1978

[22] Schott, James R. , *Matrix analysis for statistics.* New York: Wiley, 1997.

[23] Sclove, S.L. " Application of model-selection criteria to some problems in multivariate analysis" . *Psychometrika* 52, 333-343, 1987.

[24] Scott D.W., *Multivariate Density Estimation: Theory, Practice and Visualization.* New York: Wiley, 1992

[25] Tan, W. Y. y Chang, W.C. " Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities" . *Journal of the American Statistical Association* 67, 702-708, 1972.

[26] Titterington, D.M., Smith, A.F.M., and Markov, *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley. 1985